

Training Manual on

DATA ANALYTICS AND VISUALISATION

for Urban Practitioners



Training Manual on

DATA ANALYTICS AND VISUALISATION

for Urban Practitioners

TITLE

Training Manual on Data Analytics and Visualisation

PUBLISHER

National Institute of Urban Affairs (NIUA)

NIUA TEAM

Ruchi Gupta
Raman Kumar Singh
Paritosh Goel
Vignesvar J
Shreyas Chorgi

GIZ TEAM

Monika Bahl
Shriman Narayan Sai Raman

TECHNICAL CONSULTANTS

Swastik Harish, Director, Swastik Harish and Associates, India
Teja Malladi, Co-founder and CEO, MAPSolve AI, India
Dr. Chaya Degaonkar, Director, Public Affair Center, India
Mrinalini Kabbur, Head - Data Analytics, Public Affairs Centre, India
Dr. Annapoorna Ravichander, Head - Policy Engagement and Communication, Public Affair Centre, India, and
Vyas Vashisth, Creative Consultant, PAF Global, India

DESIGN

Deep Pahwa
Devender Singh Rawat
Bhavnes Bhanot
Tehan Katar
Preeti Shukla

SUPPORTED BY

Sustainable Urban Development in Smart Cities II (SUDSC II) project, which is jointly implemented by the Ministry of Housing and Urban Affairs (MoHUA), Government of India, and the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH commissioned by the German Federal Ministry for Economic Cooperation and Development (BMZ) as part of the Indo-German Development Cooperation.

DATE OF PUBLICATION

May 2024
Copyright © NIUA (2024)

DISCLAIMER

National Institute of Urban Affairs (NIUA) its employees, and advisors make no representation or warranty and shall have no liability to any person, under any law, statute, rules or regulations or tort, principles of restitution for unjust enrichment or otherwise for any loss, damages, costs or expenses which may arise from or be incurred or suffered on account of anything contained in this document or otherwise, including the accuracy, adequacy, correctness, completeness or reliability of the document and any assessment, assumption, statement or information contained therein or deemed to form part of this document.

This is an open-source document and can be used for reference purposes. Text from this document can be quoted if proper references are provided and source is acknowledged.

ABOUT THE DOCUMENT

Cities are dynamic ecosystems, constantly evolving to meet the needs of their residents. As urbanisation accelerates, the challenges faced by the municipal administrators grows inevitably, from traffic management to waste disposal, housing affordability to climate resilience, and much more. Data acts as paramount tool to develop better understanding of the intricacies of the urban system. Embracing data and cultivating analytical & visualisation capabilities, empowers the concerned officials to make better informed decision making for sustainable and resilient urban habitat.

Under the strategic partnership on Sustainable Urban Development - Smart Cities (SUDSC II) which is jointly being implemented by Ministry of Housing and Urban Affairs (MoHUA) and Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), NIUA is rendering an end-to-end ecosystem support for mainstreaming risk- informed, integrated, resilient, and sustainable urban development in cities. The project involves four components: Policy Advocacy, Capacity Building (CB), Mainstreaming Innovations, and Instituting Communities of Practice.

As part of the Capacity building component, a CB module on Data Analytics and Visualisation was developed by with technical support from Swastik Harish and Associates (SHA) and Public Affairs Centre (PAC), India as consultants. This course will empower urban practitioners and municipal administrators to transform data into actionable insights, improve decision-making, and contribute to the development of more efficient, sustainable, and resilient cities.

This Training Manual is developed based on the 'Instructional Design Framework (IDF)' developed by NIUA under the strategic partnership of SUDSC-II. This framework aids in developing learning modules with the goal of engaging, encouraging, and motivating learners to gain deeper and meaningful level of knowledge. It provides the formats/ templates for undertaking need-assessments, development of the course structure and related contents, and checklists necessary for the creation of the capacity building module.

The State Administrative Training Institutes and other training institutes conduct various training programs for the government officials, and other urban practitioners who work with the state or local governments. This manual is an easy reference for taking-up such courses as part of their regular capacity building programmes.

TABLE OF CONTENTS

INTRODUCTION	1
RATIONALE FOR DEVELOPING THE MODULES	2
STRUCTURE OF THE DOCUMENT	3
MODULES OF THE COURSE	4
MODULE 1: WORKING WITH TABULAR DATA	6
Session 1: Overview of data types, formats and key terminology	6
Session 2: Simple Statistical Analysis of Tabular Data	24
Session 3: Basics of Data Visualisation	54
Session 4: Integrating and Summarizing Tabular Datasets	78
MODULE 2: WORKING WITH SPATIAL DATA	86
Session 1: Introduction to Spatial Data	86
Session 2: Visualising and Map-making	96
Session 3: Integrating Tabular Data with Spatial Datasets	106
MODULE 3: DATA INTEGRATION, DASHBOARDS AND DECISION-SUPPORT SYSTEMS	110
Session 1: Data Dashboards	110
Session 2: Results-based Framework	124
Session 3: Group Exercise - based on gender related datasets	136
ANNEXURES	142
Annexure 1 - Reference Agenda: Capacity building on Data Analytics and Visualisation	143
Annexure 2 - Feedback form	146
Annexure 3 - Room Typology - seating	152



INTRODUCTION

The urban landscape is awash in data, from traffic patterns and energy consumption to public safety statistics and demographic trends. Yet, many urban practitioners lack the skills to effectively harness this data to inform their decision-making. This is where a training manual on data analytics and visualisation for urban practitioners comes in.

- **Bridge the Data Gap:**
 - **Current Scenario:** Many urban professionals, like planners, policymakers, and community development workers, come from non-technical backgrounds. They may not have the statistical fluency or analytical tools to navigate complex datasets.
 - **Bridging the Gap:** A training manual can equip these professionals with the fundamental skills in data analysis and visualisation, empowering them to confidently extract insights from data and translate them into actionable plans.
- **Improve Decision-Making:**
 - **Data-Driven Decisions:** Urban challenges are multifaceted and require nuanced solutions. Data analysis can reveal patterns and relationships that might otherwise be missed, leading to more informed and effective decision-making.
 - **From Intuition to Evidence:** The manual can guide practitioners through the process of identifying relevant data sources, cleaning and analysing data, and using data visualisation tools to communicate insights to stakeholders.
- **Enhance Project Efficiency and Impact:**
 - **Optimizing Resource Allocation:** Data analysis can help in identifying areas of need, optimizing resource allocation, and tracking the progress of projects. This can lead to improved efficiency and cost savings.
 - **Measuring Impact:** By measuring the impact of interventions through data analysis, urban practitioners can demonstrate the effectiveness of their work and secure funding for future initiatives.
- **Foster Collaboration and Innovation:**
 - **Shared Language of Data:** A common understanding of data analysis and visualisation can create a shared language among diverse urban stakeholders, facilitating collaboration and knowledge sharing.
 - **Data-Driven Innovation:** Data can spark new ideas and approaches to urban challenges. The manual can equip practitioners with the skills to identify trends, anticipate future needs, and develop innovative solutions.
- **Prepare for the Future of Cities:**
 - **Smart Cities:** The rise of smart cities makes data literacy an essential skill for urban professionals. The manual can prepare practitioners for the data-driven future of urban planning and management.
 - **Sustainable Development:** Data analysis can play a crucial role in achieving sustainable development goals, such as reducing emissions, managing resources efficiently, and creating equitable and liveable cities.

In conclusion, capacity building on data analytics and visualisation can empower urban practitioners to transform data into actionable insights, improve decision-making, and contribute to the development of more efficient, sustainable, and resilient cities. By equipping urban professionals with these critical skills, we can pave the way for a data-driven future that benefits all urban residents.

Building the capacity of urban practitioners in the public sector on working with data more confidently, across the spectrum of data collection and management, data analysis and visualisation, and approaches to use data for risk-informed decision-making.

RATIONALE FOR DEVELOPING THE MODULES

Cities are complex ecosystems, pulsating with life and brimming with challenges. From managing traffic flow to optimizing public services, ensuring the well-being of residents demands a deep understanding of their intricate dynamics. This is where data analytics and visualisation emerge as powerful tools for urban practitioners.

Traditionally, urban planning and management relied heavily on intuition and experience. While valuable, these approaches often lacked the granularity and objectivity needed to navigate the complexities of modern cities. This training manual bridges this gap, equipping urban practitioners with the skills and knowledge to harness the power of data for informed decision-making.

Why is this manual essential?

- **Data is everywhere:** From sensor networks to social media, cities generate a constant stream of data. This manual equips you to extract insights from this vast ocean of information.
- **Evidence-based decisions:** Gut feelings and anecdotal evidence are no longer enough. This manual empowers you to make data-driven decisions that are effective, sustainable, and responsive to the needs of your community.
- **Enhanced transparency and communication:** Data visualisation allows you to communicate complex issues clearly and compellingly to stakeholders, fostering collaboration and trust.
- **Empowering urban practitioners:** This manual puts the power of data analysis directly in your hands, enabling you to become a data-driven agent of change in your city.
- **Gender dis-aggregated data:** The capacity building programme also allows us to address specific thematic gaps in urban data and decision-making, such as on gender. The course therefore makes specific reference and use of data that is gender dis-aggregated to enable thinking on aspects of social justice that have hitherto not found adequate representation.

STRUCTURE OF THE DOCUMENT

This manual for the course on Data Analytics and Visualisation has been designed and developed keeping in mind two different but overlapping aims. Firstly, the manual will allow practitioners who are interested in working with data to gain specific insights about approaches, methods and skills through the insights presented in each of the slide decks and teaching material. Secondly, the manual is an easy reference for faculty in institutions and organizations that want to take up teaching such courses as part of their regular capacity building programmes.

The manual is divided into the following sections and sub-sections for ease of use:

- **Modules of the course (three numbers) along with key learning outcomes and a suggested 3-day course schedule.**
- **Details of individual sessions within each Module, including:**
 - Session titles, descriptions, and ideal duration.
 - Learning outcomes.
 - Resource requirements in terms of faculty expertise, classroom arrangements and equipment required, case studies and practice datasets.
 - Learner prerequisites, if any.
 - Session slides/ presentations along with teaching notes
- **Spot feedback process, including a suggested feedback form (Refer Annexure 2)**

Faculties may use this manual to design and develop programmes according to their specific needs. They may mix and match sessions according to the training needs analysis of their respective cohorts of learners.

MODULES OF THE COURSE

.....

Working with tabular data (1 to 1.5 days duration)

Through presentations and hands-on work, participants will be able to differentiate between tabular data types, describe basic statistical analysis and then recreate methods for visualisation and joining different datasets.

- Various types of tabular data
 - Statistical analyses such as central tendency, dispersion and relationships; making projections
 - Data visualisation basics
 - Summarizing and joining datasets using pivot tables, joining and look up

Working with spatial data (1 to 1.5 days duration)

Through presentations and hands-on work, participants will learn about the absolute fundamentals of spatial data, and then work their way through to basic visualisation on simple tools.

- Types of spatial data, what they mean and their technical limits
 - Sources and tools for spatial data acquisition, including geotagging
 - Simple online and offline tools to use spatial data
 - Joining with other datasets, including with existing tabular data
 - Visualisation-assisted analysis of spatial data

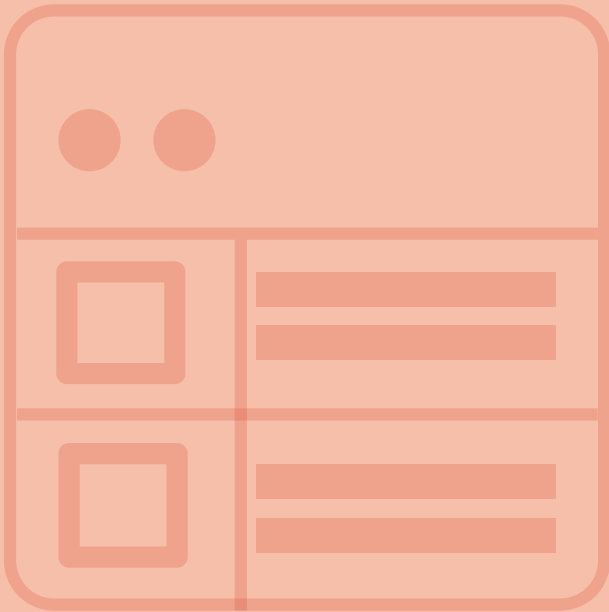
Data analysis for decision-making (0.5 to 1 day duration)

Through case studies, participants will learn how to integrate datasets on spatial or other platforms, and be able to describe how to make narratives and dashboard for decision-making

- Data Dashboarding; Case study on a data 'dashboard' and how can it help in making decisions
 - Result-based Framework/ Result-based Management; Case study on Policy responses to diminishing participation of women in the workforce in India.
 - Group exercise based on gender disaggregated datasets

These modules can be covered in three (3) or four (4) days. A suggested course programme spanning 3 days is presented in Annexure 1.





MODULE 1

WORKING WITH TABULAR DATA

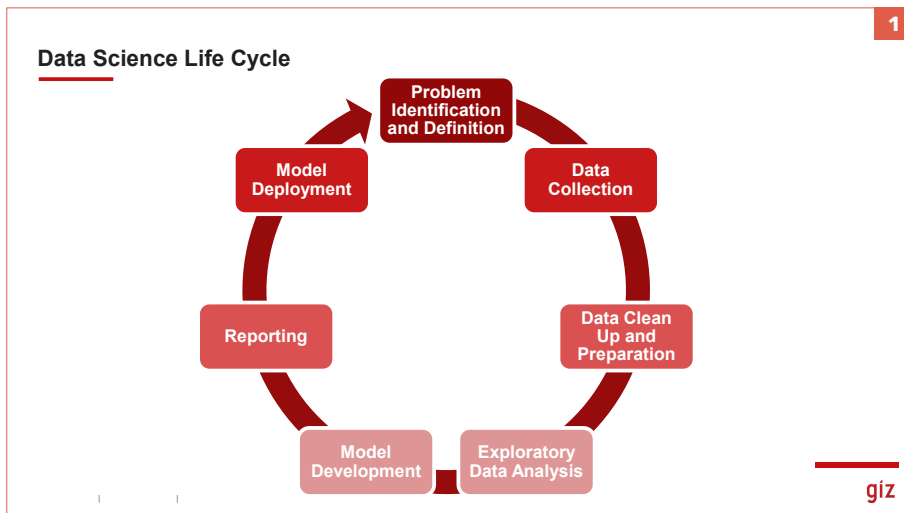
Session 1: Overview of data types, formats and key terminology

Duration: (Ideal) 1 hour

Session 1: Overview of data types, formats and key terminology

SESSION SUMMARY

1	BRIEF SESSION DESCRIPTION	<p>The session covers the basics of data and its types. The topics include</p> <ul style="list-style-type: none">• Data, Information and Knowledge• Types of Data - Qualitative Vs Quantitative, Structured, Unstructured and Semi-Structured• What is Metadata• What is Big Data• Which are the popular data tools
2	LEARNING OUTCOMES	<p>At the end of the session participants will be able to differentiate between data types and load different formats of data into excel for analysis</p>
3	CASE STUDIES (IF ANY)	<p>NA</p>
4	FACULTY REQUIREMENT	<p>No particular qualification or experience is required but some exposure to how data is being used would help. Also, some familiarisation with shared datasets is advisable</p>
5	PRACTICE DATASETS	<p>Folder: Day 1- Data sets/ Session 1 Access via https://drive.google.com/drive/folders/1NIEnGDtiT14akAIQMAsHgXhXt34umSvq?usp=sharing</p>
6	LEARNER PREREQUISITES	<p>None</p>
7	CLASSROOM ARRANGEMENT	<p>Traditional Classroom (Refer Annexure 3)</p>
8	TECHNICAL REQUIREMENTS	<p>MS Excel</p>

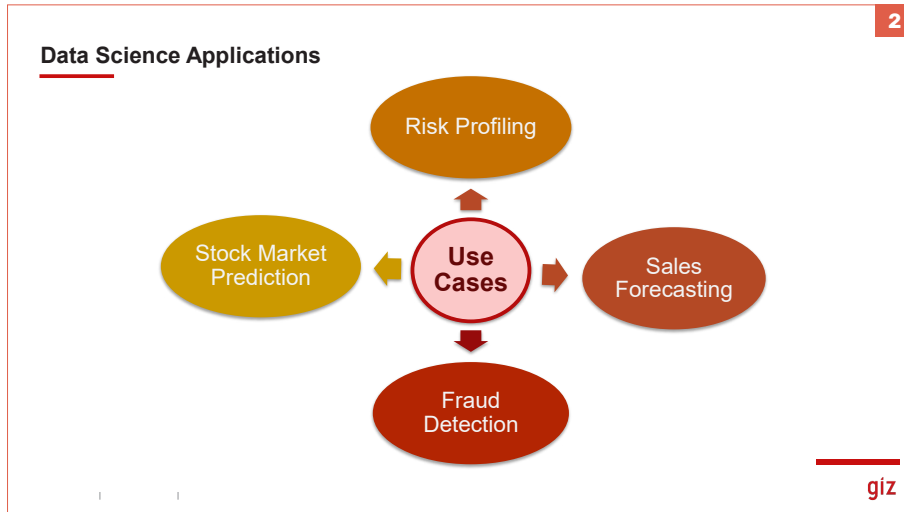


Data science is a multidisciplinary field that extracts knowledge and insights from structured and unstructured data through statistical analysis, machine learning, and domain expertise. It aids in informed decision-making, predictions, and pattern recognition.

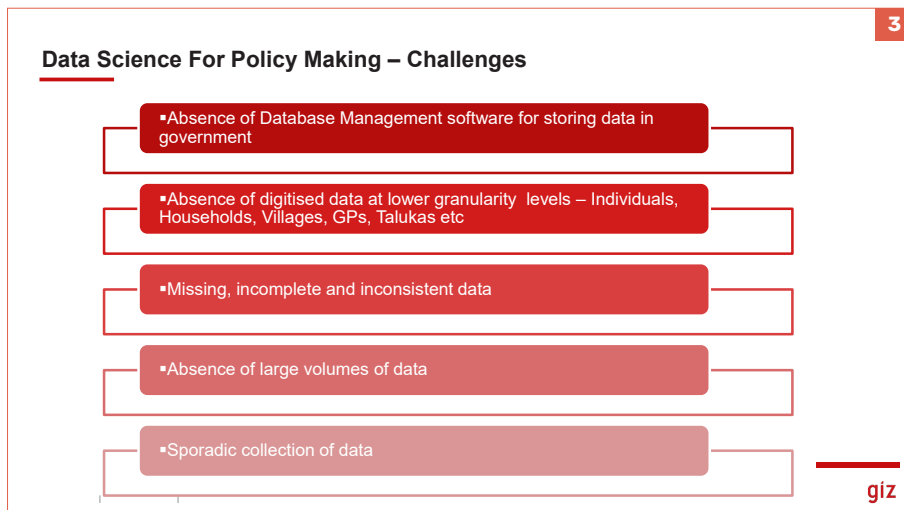
A data science life cycle is an iterative set of steps

- In this stage, the data scientist works with stakeholders and domain experts to understand the business problem or the pain points. Together they define the goals and objectives of the project. The concepts being studied and analysed are clearly defined. An indicator framework is developed to make these concepts measurable for analysis.
- For the indicators, the data is collected, either through primary or secondary sources. The data extracted from various sources can be stored in simple excel, csv files or more sophisticated data management software.
- Once the data is collected, it needs to be cleaned and transformed prior to analysis. Data cleaning involves addressing missing data, Duplicate Data, Irrelevant Data, Outliers, Structural Errors (“N/A” and “Not Applicable” appear in the same column but they are the same) etc, Data Preparation includes feature scaling using standardization or normalization to bring all the data in the same range. Sometimes, the data such as gender, educational qualification etc is categorical in nature. Since most of the statistical techniques, including ML algorithm work on quantitative data, the data may have to be transformed to a numerical value. In such cases, dummy variables are created for the categorical data. These variables take values of 0 and 1, where the values indicate the presence or absence of a category.
- Once the data is cleaned and prepared, it is ready for analysis. Exploratory data analysis perform the initial investigation of data to understand the patterns. It includes application of descriptive statistics including univariate and bivariate analysis, various visualisation techniques etc. The Exploratory Analysis is the main focus of the training.
- Once the initial analysis is performed, models can be developed for predictions, pattern recognition etc
- The results of the analysis and models can be published using dashboards, articles, reports etc.
- Final step is the deployment of an application that automates the whole process from data collection to models building and reporting.

A data scientist needs to specialize in all the aspects of life cycle.



The data science applications are developed extensively in some sectors such as finance, retail, healthcare etc. This is facilitated by the use of online transaction systems or data management softwares for all the transaction in the last couple of decades. This has enabled storage of large amount of data leading to development of data science models for applications such as stock predictions, sales forecasting, fraud detection etc.



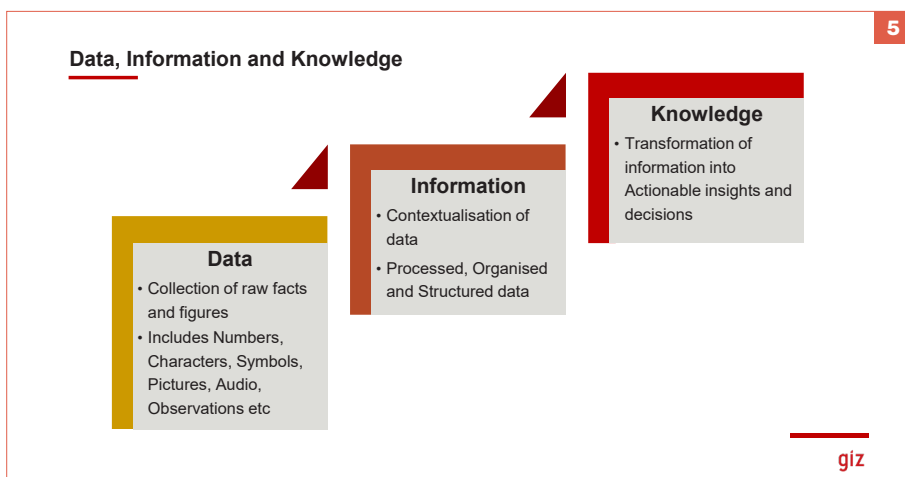
However, development of data science application for policy making has several challenges.

- **Absence of digitised data at lower granularity** - The transactions in last mile delivery institutions such as primary health centres, anganawadis, primary schools, is still manually done, usually in the form of registers. Only aggregated data at the GP, taluka or mostly at the district level is entered into the system.
- **Missing, incomplete and inconsistent data** - Since the data is collected manually, it is prone to mistakes.
- **Absence of large volumes of data** – Since data at the lower granularity is absent, the volume of data is small.
- **Sporadic collection of data** – There is no predictability in the collection of data by the departments. It is done once a while.



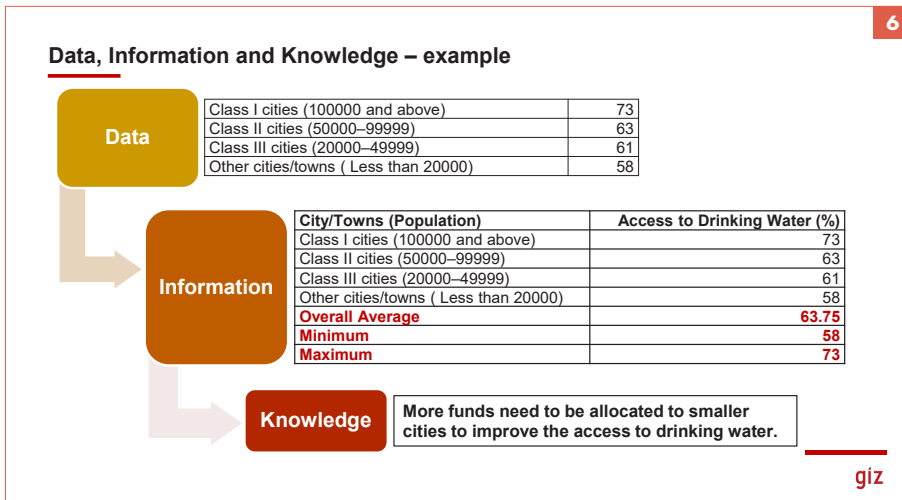
Even with the availability of limited data, there are many use cases for the application of data science in public policy.

- **Evidence base decision making** – Effective decisions are based on the analysis of data and information, rather than guess-work or instinct. Sometimes, the issues or problems in governance may be known. But data science helps to quantify this.
- **Resource Optimisation** – The available resources, financial, physical, human are limited in supply to meet the growing requirement of the population. Data Science helps to identify the areas of focus for optimal allocation of resources.
- **Measuring Governance** – Data science can also be used to measure the governance. In the recent years, various index are generated to measure the governance in different areas.
- **Effective Policy Implementation** – Data Science is extensively used in program monitoring and evaluation.



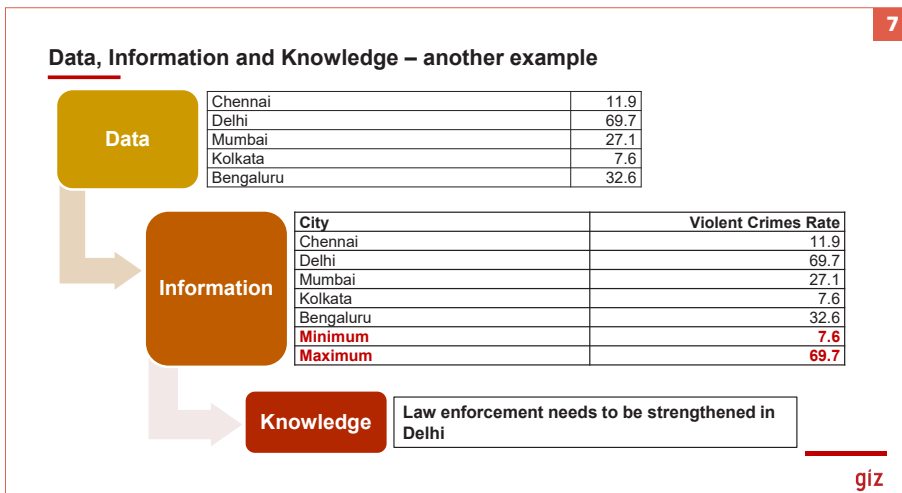
In this slide and the next, we will learn about Data, Information and Knowledge, through both the definition and examples.

- Let us begin with the definition of data. **What is data?** Data is collection of raw facts and figures. It can be numbers, characters, symbols, observations, pictures, audio, video etc. It is unorganized in nature and without processing, it is useless to the humans.
- **Information** is processed and organized data presented in a given context and is beneficial to the humans.
- In simple terms, **knowledge** is the understanding gained from information, through organization, analysis and interpretation. It gives actionable insights and enables decision-making and problem-solving.



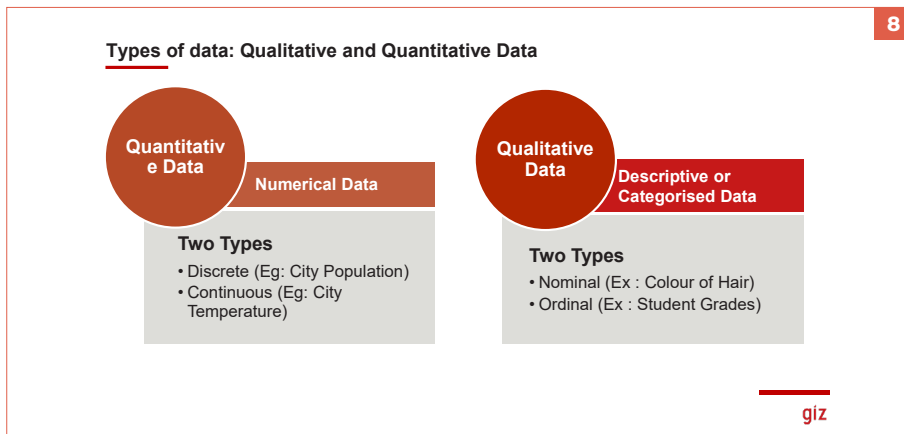
Let us learn more with examples.

- In the top most figure, there is some text and numbers. Do you understand anything from these? No.
- Now, let us process this data. In the second figure, the data is given a context through column description. With this, we understand that the data is about access to drinking water in various categories of cities.
- The data is further processed by calculating average, minimum and maximum access to drinking water among various categories of cities.
- This processed data leads to knowledge i.e the access to drinking water is poor among smaller cities and more allocation of funds is necessary in improving access to drinking water facilities.



Let us learn more with examples.

- In the top most figure, there is some text and numbers. Do you understand anything from these? No.
- Now, let us process this data. In the second figure, the data is given a context through column description. With this, we understand that the data is about various crime rates in the top 5 most populous cities.
- The data is further processed by calculating average, minimum and maximum violent crime rates in Indian cities.
- This processed data leads to knowledge i.e the city of Delhi has the maximum crime rate and the law enforcement needs to be strengthened there, when compared to other most populous cities.



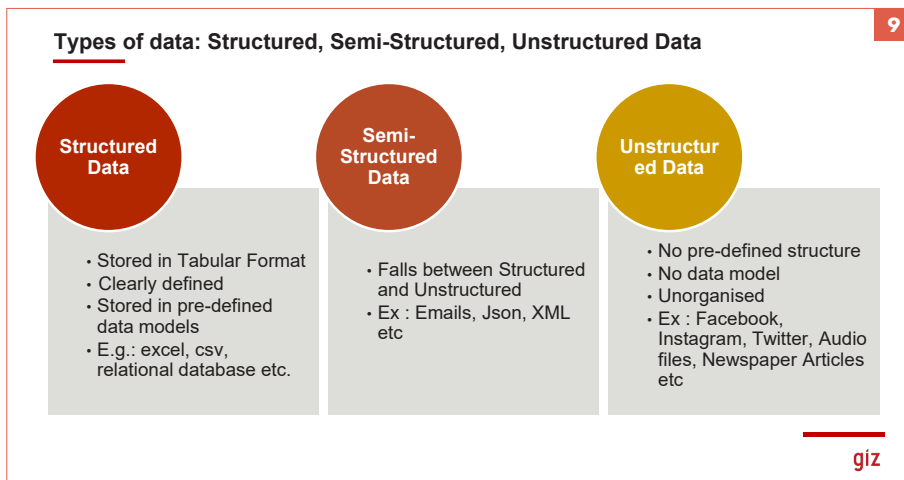
The data can be of two types: Quantitative and Qualitative

Qualitative data is descriptive, categorical and conceptual. For example, your favourite movie, colour of your pen, age category etc.

Further, the qualitative data can be nominal or ordinal. Nominal data does not have an order or a rank while ordinal data has some predetermined or natural order. country, gender, hair colour etc are nominal data. student grades, customer feedback (good, average, bad) etc are ordinal data.

Quantitative data can be counted, measured and expressed in numbers.

Here also, there are two types. Discrete data is only take certain values. It usually comes in the form of whole numbers or integers. Continuous data, on the other hand, has values that are not fixed and have an infinite number of possible values. It comes in the form of fractions, decimals etc. Population of a city, number of students in classroom are examples of discrete data while height/weight of a person, temperature of a city are examples of continuous data.



The data can also be classified based on the organization and format of the data. There are 3 classifications: Structured, Unstructured and Semi-Structured

- **Structured data** is stored in tabular format, in the form of rows and columns. It is clearly defined and stored in a pre-defined data model. Examples are Excel, Database Management Systems like MySQL, PostgreSQL, Oracle etc
- **Unstructured data** does not have a pre-defined model or structure. It is unorganised, irregular and ambiguous. This includes a combination of audio, video, messages, images, texts etc. Facebook, Instagram, Twitter, Youtube, Newspaper Articles etc are examples of unstructured data. This is a very useful data and provides a lot of information.
- **Semi-Structured data** falls between structured and unstructured data. Examples include JSON files, Emails, XML, HTML etc

10

Metadata

Metadata – data about data

- Metadata: "data that provides information about other data", but not the content of the data. Examples: metadata for a document might include a collection of information like the author, file size, the date of creation etc and keywords to describe the document.
- Helps to understand how the data is structured, definitions of terms used, how it was collected, and how it should be read.
- Makes the document searchable.

giz

- Metadata is formed through two words: Meta and Data. The literal meaning of meta is referring to itself or explicit awareness of itself. Metadata is also data that gives information about other data. It provides details such as the source, type, owner, title, size of data and also its relationships to other data sets. It helps to understand a particular data set and also guides on how to use it.
- Search engines, such as Google, optimise search by seeking out information from defined metadata fields rather than reading the whole content.

11

Popular Software and Tools

Data Storage	Data Analysis	Models & Algorithms	Data visualisation	Cloud based platforms
<ul style="list-style-type: none"> • Postgres • MySQL • Excel • CSV 	<ul style="list-style-type: none"> • R • Python • SAS • MATLAB • SPSS • Stata • Excel 	<ul style="list-style-type: none"> • SciKitLearn • Tensor Flow • Keras 	<ul style="list-style-type: none"> • Tableau • Power BI • GGPlot • Rshiny • Plotly 	<ul style="list-style-type: none"> • Teradata • Snowflake • AWS • Databricks

giz

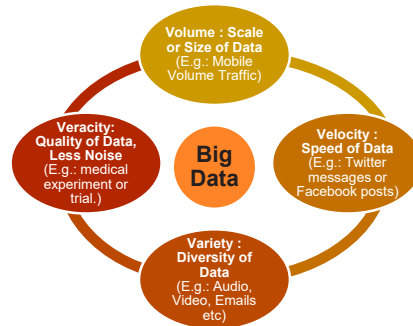
The software and tools for data analytics can be classified under the following categories:

- **Data Storage:** This includes excel and data base management softwares such as Postgres and MySQL
- **Data Analysis:** This includes open source statistical programming languages such as R and Python. There are licensed software tools such as SAS, SPSS, MATLAB, STATA etc. They come with easy to use User Interfaces. Microsoft Excel also provides powerful data analysis functionality.
- **Models and Algorithms:** scikit-learn is a free software machine learning library for the Python programming language. TensorFlow is a free and open-source software library for machine learning and artificial intelligence. Keras is an open-source library that provides a Python interface for artificial neural networks.
- **Data Visualisation and Dashboard:** Tableau and Power BI are very popular tools for data visualisation. They require licenses. Shiny is an R package that enables building interactive web applications that can execute R code on the backend. Similarly, plotly is an interactive graphing library for python.
- **Cloud based data platforms:** A cloud data platform allows organisation to move their data from traditional storage such as data warehouse, data management softwares to the cloud. It is an integrated software solution that enables businesses to collect, store, process, and use data in the cloud while offering an integrated view to manage and secure the data. The most popular platforms include Teradata, Snowflake, AWS, Microsoft Azure, Google Cloud and Databricks.

Big Data

A data which is very large in size and hard to measure

This large amount of data is hard to analyse using traditional data handling systems such as MS Excel, R, Python etc



giz

Big data is a data which is very large in size and hard to measure.

This large amount of data is hard to analyse using traditional data handling systems such as MS Excel, R, Python etc.

For data to be qualified as big data, it needs to possess the following characteristics

- **Volume:** The data needs to be huge, from various data sources.
- **Velocity:** The data is accumulated at a high speed, sometimes every milliseconds. The examples include mobile traffic data, stock exchange data etc.
- **Variety:** Big data is a combination of structured, semistructured and unstructured data. This examples include youtube videos, emails, facebook messages etc.
- **Veracity:** Veracity refers to the quality, accuracy, integrity and credibility of data.
 - The data should not have too many missing and incomplete values.
 - It should be devoid of abnormalities such as outliers.
 - The data should be accurate and gathered from reliable and trustworthy sources.
 - The veracity of data is extremely important in medical data so that judgements can be made with knowledge.

Sources of Big Data

Social Media	Platforms like Facebook, Twitter, Instagram, and LinkedIn generate vast amounts of data in the form of text, images, videos, and user interactions.
IoT (Internet of Things) Devices	Devices such as sensors, smart appliances, industrial machines, and wearable technology generate real-time data about their environment and usage.
Websites and E-commerce	Online transactions and customer reviews on websites and e-commerce platforms produce significant data.
Financial Transactions	Banking and financial institutions generate vast amounts of data through transactions, stock market activities, and customer interactions.

giz

- **Social Media:** Platforms like Facebook, Twitter, Instagram, and LinkedIn generate vast amounts of data in the form of text, images, videos, and user interactions. They use big data to power content recommendation systems. These systems suggest content to users based on their past interactions and interests.
- **IoT (Internet of Things) Devices:** The Internet of Things (IoT) refers to physical objects connected through shared networks. A variety of sensors gather information and share it across systems that can store, manage, filter, and analyse the data. An IoT device can refer to everything from wearables to medical devices to industrial equipment.
- **Websites and E-commerce:** Online transactions and customer reviews on websites and e-commerce platforms produce significant data. E-commerce businesses collect this data and use it to increase customer experience.
- **Financial Transactions:** Banking and financial institutions generate vast amounts of data through transactions. The institutions can use this big data to detect

- fraudulent activities such as money laundering or identity theft. The other source of big data in finance include stock market activities. This data can be used make decisions related to buying and selling.

Sources of Big Data (Contd..)	
Healthcare	Electronic health records, medical imaging, and wearable health devices contribute to big data in the healthcare industry.
Scientific Research	Scientific experiments, simulations, and observations generate extensive datasets in fields like astronomy, genomics, and climate research.
Energy Usage	Utilities collect data on energy consumption patterns, helping optimize energy distribution.
Satellite Imagery	Remote sensing data from satellites is used for environmental monitoring, agriculture, and disaster management.


- **Healthcare:** Electronic health records, medical imaging, and wearable health devices contribute to big data in the healthcare industry. This data can be analysed to guide decision-making, improve patient outcomes, and decrease health care costs, among other things.
- **Scientific Research:** Scientific experiments, simulations, and observations generate extensive datasets in fields like astronomy, genomics, and climate research. Astronomical data mainly include images, spectra, time-series data, and simulation data. Most of the data are saved in catalogues or databases. Big data analytics in genomics can be used to decode DNA sequences.
- **Energy Usage:** Big Data in the energy sector can bring numerous benefits, including improving operational efficiency, decreasing costs, boosting customer satisfaction, and optimizing energy production.
- **Satellite Imagery:** Remotesensingdatafromsatellites is used for environmental monitoring, agriculture, and disaster management. Data is collected by sensors from around the globe and instruments on satellites that are used to assess the Earth’s conditions and help to predict climate events. Remote

sensing technologies, coupled with satellite imagery and ground-based sensors, provide real-time data on deforestation, land cover changes, ocean temperature, air quality, and more.

Tools in Big Data Analytics


•An open-source software utility to process gigabytes and terabytes of data. It provides distributed storage (Hadoop Distributed File System - HDFS) for storing data.

Hadoop




•Apache Spark is an open-source, distributed computing system designed for big data processing and analytics

Apache Spark




•An open source tool that supports data storage. A popular NoSQL database suitable for storing unstructured and document-oriented data.

MongoDB




•A distributed NoSQL database designed for high scalability, often used for time-series data and large-scale applications

Cassandra




•A tool to integrate and process data. Stored in cloud after processing

Integrate.io



•A data visualization tool that allows users to create interactive and shareable dashboards and reports.

Tableau



giz

All the big data characteristics discussed earlier impact the tools and techniques that are used to handle big data.

Big data techniques are simply methods, algorithms, and approaches to process, analyse, and manage big data. On the surface, they are the same as in regular data. However, the big data characteristics call for different approaches and tools.

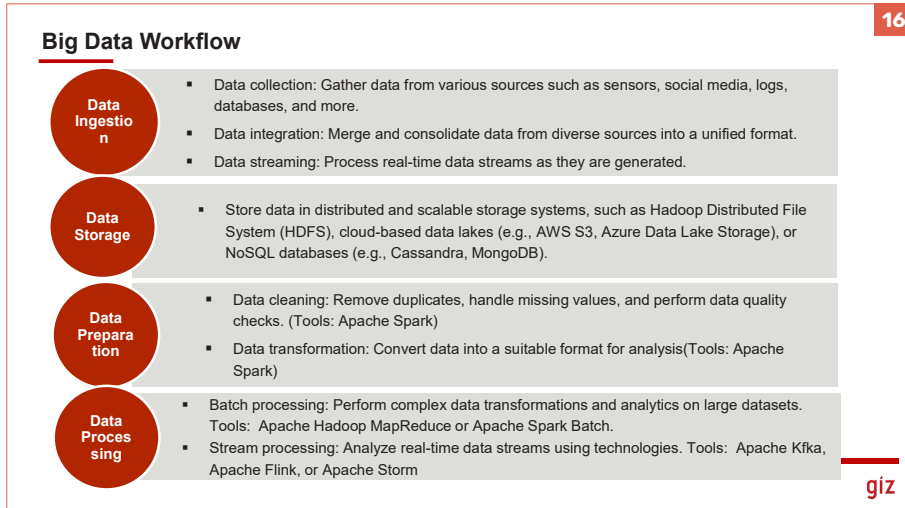
Here are some prominent tools and techniques used in the big data domain.

- All the big data characteristics discussed earlier impact the tools and techniques that are used to handle big data.
- Big data techniques are simply methods, algorithms, and approaches to process, analyse, and manage big data. On the surface, they are the same as in regular data. However, the big data characteristics call for different approaches and tools.

Here are some prominent tools and techniques used in the big data domain.

- **Apache Hadoop** - Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage (Hadoop Distributed File System - HDFS) and processing of big data using the MapReduce programming model.
- **Apache Spark** - Apache Spark is an open-source, distributed computing system designed for big data processing and analytics.
- **MongoDB** - MongoDB is an open source tool that supports data storage. It is a popular NoSQL database suitable for storing unstructured and document-oriented data.
- **Cassandra** - Cassandra is a distributed NoSQL database designed for high scalability, often used for time-series data and large-scale applications.
- **Integrate.io** - Integrate.io provides visual approach with minimal hand coding data pipeline platform specializing in Extracting, Transforming and Loading data from various sources. This helps to automate business processes and manual data preparation.

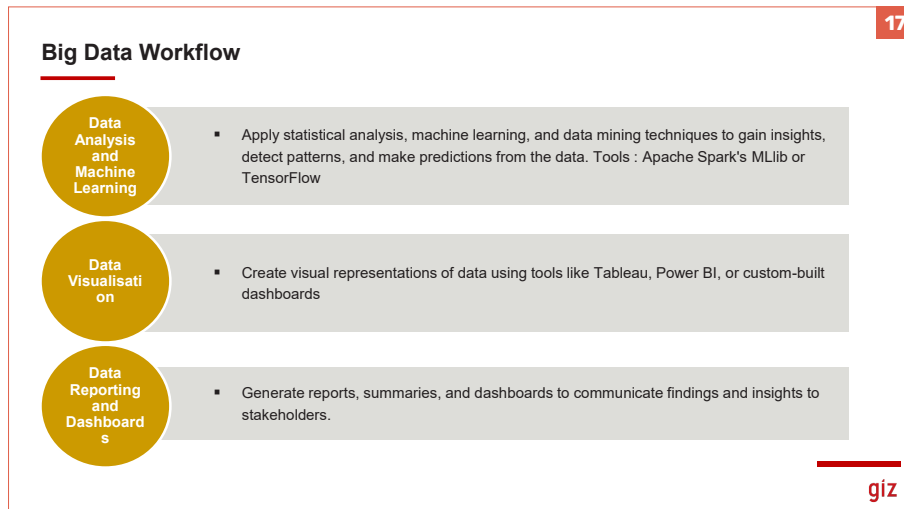
- **Tableau** - Tableau is a data visualisation tool that allows users to create charts, graphs and interactive and shareable dashboards and reports.



A big data workflow typically involves a series of steps aimed at collecting, processing, analysing, and interpreting large volumes of data to extract valuable insights. Here's a general outline of a big data workflow.

- **Data Ingestion** - This includes collecting, integrating and transferring big data from source into a storage system. The data is collected from various source such as sensors, databases, social media, web logs, etc. Data may be structured or unstructured. If data is collected from multiple diverse sources, it may have to be merged and consolidated into a unified format. After this, data will have to be transferred to the storage system (data lake, data warehouse, etc.)
- **Data Storage** - Data needs to be stored in a way that makes it accessible for processing and analysis. This often involves storing data in distributed storage systems like Hadoop Distributed File System (HDFS), cloud-based storage solutions (AWS S3, Azure Blob Storage), or NoSQL databases such as MongoDB.
- **Data Processing** - In this step, data is cleaned, transformed, and prepared for analysis. This may involve filtering out irrelevant data, handling missing values, and converting data into a format suitable for

analysis. Technologies like Apache Spark, Apache Flink, or MapReduce are commonly used for large-scale data processing.



Batch and stream processing are two different approaches to processing data in big data systems, each suited to different types of applications and requirements. Batch processing involves collecting and processing data in chunks or batches. Data is collected over a period of time, stored, and then processed offline in large volumes. Stream processing involves continuously processing data in real-time as it is generated. Data is processed as individual events or small micro-batches as soon as it arrives.

- **Data Analysis** - Once the data is prepared, various analytical techniques are applied to derive insights from the data. This may include descriptive analytics to summarize the data, predictive analytics to forecast future trends, or prescriptive analytics to suggest actions based on analysis.
- For more complex analysis, machine learning algorithms and advanced analytical techniques may be applied to big data. This can include clustering, classification, regression, and other machine learning tasks to uncover hidden patterns and relationships in the data.
- **Data Visualisation** - Visualising data in the form of charts, graphs, or dashboards helps in understanding the patterns and trends present in the data. Tools like Tableau, Power BI, or matplotlib in Python are commonly used for data visualisation.
- **Data Reporting and Dashboards** - Data reporting involves summarizing the insights derived from data analysis into understandable and actionable formats. This often includes generating reports,

visualisations, and dashboards that communicate key metrics and trends to stakeholders. Tools like Tableau, Power BI, and Google Data Studio are commonly used for creating reports and dashboards.

18

Big Data Analytics

It is the process to analyse huge datasets

Example : Spotify has 551 million users. Based on the shares, likes and search history, the data analytics engine recommends songs for the user.

Based on search history, Google News gives customised news feed

Based on items purchased, items mentioned as owned, and items rated, Amazon recommends items the person maybe interested in

Most of the e-commerce companies such as Amazon, Flipkart etc rely on information from the big data analytics

giz

The analysis of big data is termed as “Big Data Analytics”. Some of the examples of analysis are given in the slide.

- Spotify has 551 million users. Based on the shares, likes and search history, the data analytics engine recommends songs for the user.
- Based on search history, Google News gives customised news feed.
- Based on items purchased, items mentioned as owned, and items rated, Amazon recommends items the person maybe interested in .

19

Type of Big Data Analytics

Descriptive Analytics : Answers the question “What happened?”	Example : By gathering data on users’ in-platform behaviour, Netflix analyses the data to determine which TV series and movies are trending at any given time
Predictive Analytics : Answers the question “What might happen in the future?”	Example : Using historical data from previous financial statements, as well as data from the broader industry, projecting sales, revenue, and expenses
Diagnostic Analytics : Answers the question, “Why did this happen?”	Example, identify why sales decreased during a specific time period
Prescriptive Analytics : Answers the question “What should we do next?”	Example : if at least 50 percent of customers in a dataset selected that they were “very unsatisfied” with your customer service team, the algorithm may recommend additional training.

giz

Type of analytics possible for big data is similar to analytics for any data.

- **Descriptive Analytics** - Descriptive big data analytics refers to the process of analyzing large volumes of data to uncover patterns, trends, and insights about past events or behaviours. It primarily aims to summarize and understand historical data.

It gives answers to “What happened?”

For Example, by gathering data on users’ in-platform behaviour, Netflix analyses the data to determine which TV series and movies are trending at any given time.

- **Predictive Analytics** - Predictive big data analytics involves using large volumes of data to forecast future outcomes or trends. It leverages advanced statistical techniques and machine learning algorithms to make predictions based on past patterns and relationships in the data.

It gives answer to the question “What might happen in the future?”

For Example, using historical data from previous financial statements, as well as data from the broader industry, projecting sales, revenue, and expenses.

- **Diagnostic Big Data Analytics** - Diagnostic big data analytics involves analysing large volumes of data to identify the root causes of specific events, issues, or problems. It focuses on understanding why certain outcomes occurred by examining patterns, correlations, and anomalies within the data. Diagnostic analytics often serves as a crucial step in problem-solving and decision-making processes.

It gives answer to the question “Why did this happen?”

For Example, identify why sales decreased during a specific time period.

- **Prescriptive Analytics** - Prescriptive big data analytics takes insights from descriptive, diagnostic, and predictive analytics a step further by not only forecasting what is likely to happen but also recommending actions to achieve desired outcomes. It leverages advanced algorithms and techniques to provide actionable recommendations that guide decision-making processes.

It gives answer to the question “What should we do next?”

For Example, if at least 50 percent of customers in a dataset selected that they were “very unsatisfied” with your customer service team, the algorithm may recommend additional training.

20

Applications of Big Data Analytics

Amazon	Collects huge amounts of data through its services such as amazon shopping, amazon web services, amazon pay etc. It uses big data technologies to store and process this data. It uses the data to discover customer behaviour and decision making on the sale of different products
Netflix	Collects data from their more than 150 million subscribers and applies big data analytics models to discover customer behaviour and buying patterns. Then, using that information to recommend movies and TV shows based on their subscribers' preferences.
Satellite Data	The satellite images can be used for monitoring crop health, growth stages, pest and disease detection for timely action to control, soil quality assessment to enable use for maximum productivity.

giz

Few applications Applications of Big Data Analytics are detailed in this slide.

- The e-commerce giant Amazon collects huge amounts of data through its services such as amazon shopping, amazon web services, amazon pay etc. It uses big data technologies to store and process this data. It uses the data to discover customer behaviour and decision making on the sale of different products.
- The streaming service Netflix collects data from their more than 150 million subscribers, and applies

big data analytics models to discover customer behaviour and buying patterns. Then, using that information to recommend movies and TV shows based on their subscribers' preferences.

- The satellite images can be used for monitoring crop health, growth stages, pest and disease detection for timely action to control, soil quality assessment to enable use for maximum productivity.

21

Applications of Big Data Analytics (Contd..)

Health Care	Medical imaging helps radiologists and clinicians identify abnormalities more accurately. This information assists in diagnosis and treatment planning.
Financial Trading	Large volumes of data are analysed in real-time to make split-second trading decisions
Fraud Detection and Prevention	Big Data Analytics allows financial institutions to monitor transactions in real-time, identifying suspicious patterns or anomalies that may indicate fraudulent activities.

- In Health Care, medical imaging helps radiologists and clinicians identify abnormalities more accurately. This information assists in diagnosis and treatment planning.
- In Financial Trading, large volumes of data are analyzed in real-time to make split-second trading decisions.
- The Fraud Detection and Prevention Big Data Analytics allows financial institutions to monitor transactions in real-time, identifying suspicious patterns or anomalies that may indicate fraudulent activities.

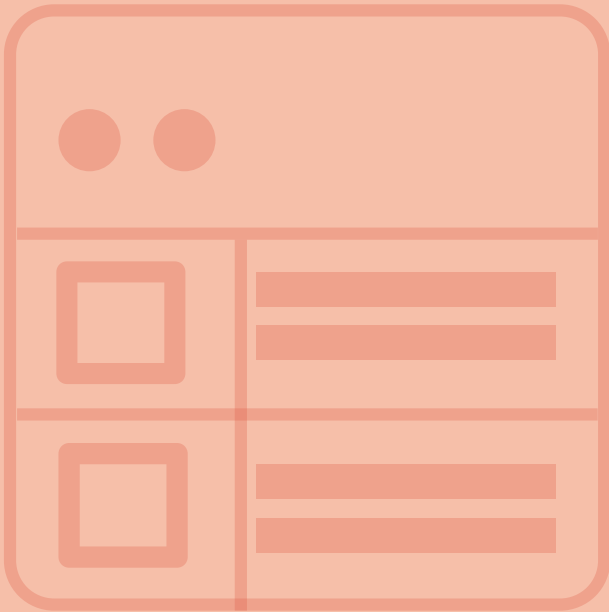
22

Big Data in Public Policy

Challenges	Lack of Digital Data : Registers continue to be used. Digitised data is hard to find at the disaggregated (individual) level Data Integration issues : data silos in different departments and no mechanism for integrating them Dearth of trained professionals
Some initiatives	RCH Portal, Poshan Tracker, Open Government Data Platform, Smart Cities Open Data Portal

The above slide highlights the pertinent challenges and some initiatives taken by relevant agencies. The big data and analytics face numerous challenges in public policy:

- **Lack of Digital Data :** The online transaction processing systems or database management systems still do not exist to capture data at the last mile institutions. Registers continue to be used. Digitised data is hard to find at the disaggregated (individual) level.
- **Data Integration issues :** data silos in different departments and no mechanism for integrating them
- **Dearth of trained professionals** is another issue.



MODULE 1

WORKING WITH TABULAR DATA

Session 2: Simple Statistical Analysis of Tabular Data

Duration: (Ideal) 1 hour

Session 2: Simple Statistical Analysis of Tabular Data

SESSION SUMMARY

1	BRIEF SESSION DESCRIPTION	<p>The session covers the basics of statistics with hands on examples in excel. The topics includee</p> <ul style="list-style-type: none">• Descriptive Statistics – The descriptive statistics includes<ul style="list-style-type: none">▪ Measures of Central Tendency – Arithmetic and Geometric Mean, Median and Mode▪ Measures of Dispersion – Range, Variance, Standard Deviation and Coefficient of variation• Skewness and Normal Distribution• Projections using Geometric Mean• Correlation• Introduction to regressiaon• Which are the popular data tools?
2	LEARNING OUTCOMES	<ul style="list-style-type: none">• A working knowledge of traditional statistical techniques and the ability to apply these methods to datasets related to public policy and governance• Learn about making projections for time series data using statistical techniques• Understanding various steps and techniques in developing index
3	CASE STUDIES (IF ANY)	<ul style="list-style-type: none">• Ease of Living Index• SDG 11 Index
4	FACULTY REQUIREMENT	<p>Basic Understanding of MS Excel. Basic school level Arithmetic is necessary for understanding the statistical concepts</p>
5	PRACTICE DATASETS	<p>Folder: Day 1- Data sets/ Session 2-3-4 Access via https://drive.google.com/drive/folders/1NIEnGDtiT14akAIQMAsHgXhXt34umSvq?usp=sharing</p> <p>Ref: Sheet1, Sheet 2, Sheet3, Sheet4, Sheet5, Sheet 6 of DAV Module 1_ Practice datasets</p>
6	LEARNER PREREQUISITES	<p>Basic Mathematics</p>
7	CLASSROOM ARRANGEMENT	<p>Traditional Classroom (Refer Annexure 3)</p>
8	TECHNICAL REQUIREMENTS	<p>MS Excel</p>

1

Descriptive Statistics

- Organises, summarises and presents the data in an informative way
 Ex : Average test score of students in a class
 The preferred mode of transportation of the employees in an office

There are four major types of descriptive statistics

Measures of Frequency	Measures of Central Tendency	Measures of Dispersion	Measures of Position
<ul style="list-style-type: none"> • Count, • Percent, • Frequency 	<ul style="list-style-type: none"> • Mean, • Median, • Mode 	<ul style="list-style-type: none"> • Range • Variance, • Standard deviation 	<ul style="list-style-type: none"> • Quartile, • Percentile

- The descriptive statistics is the first step in any analysis. How is the performance of students in a class? How well has a batsman performed during the last 3 months in one day cricket matches? Which district had the highest outbreak of dengue in the last year? These are some the questions descriptive statistics helps to answer.
- Descriptive statistics summarises and presents the characteristics of data in a meaningful way. There are three types: central tendency, dispersion and distribution.
- Arithmetic Mean, Geometric Mean, Median and Mode are the main measures of central tendency. Among Measures of Dispersion, Range, Variance and Standard Deviation are widely used. Skewness measures the degree of asymmetry of the distribution, while Kurtosis measures the degree of peakiness and flatness of a distribution.

Measures of Frequency

Count - count refers to the number of data values in a dataset. It represents how many observations or occurrences are present.

Percent - A percentage is a number or ratio that indicates a portion of a whole, with 100% representing the entirety

Frequency - The number of times a value occurs in a set of data.

giz

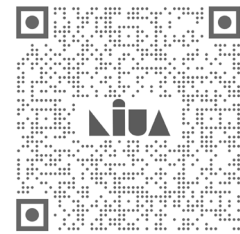
Count refers to the number of data values in a dataset. It represents how many observations or occurrences are present. For example, if we have a dataset of exam scores for a class of students, the count would be the total number of scores recorded.

Frequency refers to the number of times a particular value occurs in a dataset. For example, if we have a dataset of quiz scores, and the scores are 4, 6, 6, 9, and 6, the frequency of the score "6" is 3 because it appears three times in the dataset.

In statistics, a percentage is a relative value expressed as a fraction of 100. It represents hundredth parts of any quantity. For example:

- 1% (symbolized as 1%) is one hundredth of a quantity.
- 100% represents the entire quantity.
- 200% specifies twice the given quantity.

Sum: The sum of a set of values is the total obtained by adding all those values together. Summation refers to the operation of adding up a sequence of numbers to find their total or sum. It is denoted by the Greek letter capital sigma (Σ).



Sum and Percentage

Measures of Central Tendency – Arithmetic Mean

- Measure of Central Tendency is used to determine the center of the distribution of data.
- Arithmetic Mean is the most popular measure of central tendency
- It is the average of all the observations
- It is calculated as the sum of all the observations divided by the total number of observations

$$\bar{X} = \frac{\sum_i^n x_i}{N}$$

Example: Find the arithmetic mean of 4, 8, 12, 16, 20.

Solution:

- Sum of observations = 4+ 8+12+16+20 = 60
- Number of observations = 5
- Arithmetic Mean = $\frac{60}{5} = 12$

giz

Measure of Central Tendency is used to determine the center of the distribution of data.

Arithmetic mean is the most widely used measure of central tendency. It is calculated as the sum of all the observations divided by the number of observations. It is very simple to calculate and works for both positive and negative numbers.

Calculation of Arithmetic Mean in Excel

Please refer

<https://support.microsoft.com/en-au/office/average-function-047bac88-d466-426c-a32b-8f33eb960cf6#:~:text=Description,the%20average%20of%20those%20numbers.>

The examples shows the calculation of Arithmetic Mean

Arithmetic mean is the most widely used measure of central tendency. It is calculated as the sum of all the observations divided by the number of observations. It is very simple to calculate and works for both positive and negative numbers.

Please refer

<https://support.microsoft.com/en-au/office/average-function-047bac88-d466-426c-a32b-8f33eb960cf6#:~:text=Description,the%20average%20of%20those%20umbers>

4

Measures of Central Tendency – Geometric Mean

- Geometric mean is defined as the nth root of the product of n observations
- It is calculated by taking the product of all the observations and taking nth root of the product.

$$GM = \sqrt[n]{x_1 * x_2 * x_3 \dots X_n}$$

Example: Find the geometric mean of 4, 8, 12, 16, 20.

Solution:

- Product of observations = $4 \times 8 \times 12 \times 16 \times 20 = 122880$
- Number of observations = 5
- Geometric Mean = $\sqrt[5]{122880} = 10.421$

giz

Geometric mean is a better measure when there are large fluctuations in data. It is the nth root of the product of n numbers. The geometric mean can only be used for positive numbers.

Calculation of Geometric Mean in Excel

<https://support.microsoft.com/en-us/office/geomean-function-db1ac48d-25a5-40a0-ab83-0b38980e40d5>

5

Measures of Central Tendency – Median

- Median is the middle value of the given distribution of data when arranged either in ascending order or descending order
- The Median is calculated as follows:
 - Arrange the observations (n) from smallest to largest or largest to smallest.
 - **Median = $(\frac{n+1}{2})^{\text{th}}$ observation**
 - If the number of observations is odd, the median is the middle data point in the distribution.
 - If the number of data points is even, the median is the average of the two middle observations in the distribution.

Find the Median of 4, 8, 12, 16, 20.

Solution:

- Here, $n = 5$ -> the number of observations is odd
- Median = $\frac{5+1}{2} = 3^{\text{rd}}$ value = 12

Find the Median of 4, 8, 12, 16, 20, 20

Solution:

- Here, $n = 6$ -> the number of observations is even
- Median = $\frac{6+1}{2} = 3.5$
- The average of 3rd and 4th value = $\frac{12+16}{2} = 14$

giz

The median is the middle value in a set of data. The data needs to be first organized in order, either smallest to largest or largest to smallest before calculating the middle value. Once the data is organized, to find the midpoint value, divide the number of observations by two. If there are an odd number of observations, round that number up, and the value in that position is the median. If the number of observations is even, take the average of the values found above and below that position.

Calculation of Median in Excel

<https://support.microsoft.com/en-us/office/median-function-ea66592d-8df3-4e66-945c-75740408ddf0>

The examples show the calculation of median for odd and even number of observations.

Measures of Central Tendency – Mode

- The mode is the **most frequently occurring observation** in a distribution.
- The mode is useful when there are a lot of repeated values in a dataset.
- There can be no mode, one mode, or multiple modes in a dataset.
- The mode is the only measure of central tendency for categorical data.

Example : Find the Mode of 4, 8, 12, 16, 20, 20
Solution: 20

Example : Find the Mode of 4, 8, 12, 16, 20, 24
Solution: No Mode

Example : Find the Mode of 4, 8, 12, 12, 20, 20
Solution: 12 and 20

giz

Mode is the value that appears maximum number of times in a data set. It is used mostly for qualitative data. The mode is used when there are repeated values in a dataset and also for qualitative data. A distribution can have one mode, multiple modes or no mode. The examples show the case for each type.

Calculation of Mode in excel:

<https://support.microsoft.com/en-us/office/mode-function-e45192ce-9122-4980-82ed-4bdc34973120>

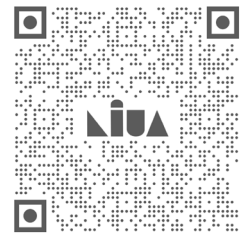
Measures of Central Tendency (Mean, Median and Mode Example)

Find the mean, median and mode of the following numbers:
23, 29, 20, 32, 23, 21, 33, 25

Arithmetic Mean	$\frac{23+29+20+32+23+21+33+25}{8} = 25.75$
Median	Arrange it in ascending or descending order : 20, 21, 23, 23, 25, 29, 32, 33 Middle value = $\frac{8+1}{2} = 4.5$ Median = $\frac{23+25}{2} = 24$
Mode	23 is the mode since it shows up the greatest number of times (2 times)
Geometric Mean	$\sqrt[8]{23 * 29 * 20 * 32 * 23 * 21 * 33 * 25} = 25.34$

giz

This slide we calculate all the measures of central tendency i.e., mean, median, and mode for the given same dataset and find the statistical difference between all the measures.



Mean Median Mode

8

Which is the Best Measure of Central Tendency?

- For quantitative data, arithmetic mean is the best measure of central tendency
- However, the arithmetic mean is affected by the extreme values as it includes all the data in the distribution for calculation.
- Median is the preferred measure of central tendency for extreme values. It is not influenced by the extreme values. It is only sensitive to values at the middle.

Example:

- The arithmetic mean of 4,8,12,16,2 is 12 (calculation in previous slide)
- Now add 100 to the distribution and find the arithmetic mean of 4, 8, 12, 16, 20, 100
- Arithmetic Mean = $\frac{160}{6} = 26.66$
- The arithmetic mean increased from 12 to 26.66 with just the addition of one number
- The median for the same distribution would be $\frac{6+1}{2} = 3.5^{\text{th}}$ value i.e. the average of 3rd and 4th value.
- Median = $\frac{12+16}{2} = 14$. This is more representative of the distribution

giz

For the quantitative data, arithmetic mean is the most used measure of central tendency. However, arithmetic mean does not work well if there are extreme values in the distribution. In this context, median is the most preferred measure. It is not influenced by the extreme values. It is only sensitive to the values in the middle.

The example shows the working of mean and median when extreme values are present in the distribution.

9

Which is the Best Measure of Central Tendency? (Contd..)

For qualitative data, mode is the only measure of central tendency

Example:

Find the most preferred mode of transportation?
 The following is the preferred mode of transportation of 10 employees in a company.
 Bus, Car, Bus, Bus, Bike, Bike, Bike, Bus, Car, Bus

Solution:

- Bus occurs 5 times
- Car occurs 2 time
- Bike occurs 3 times
- The mode is 5. Hence, the most preferred mode of transportation is Bus

giz

When there is categorical or qualitative data, mean and median cannot be used. Since mode is a measure that is based on counting, it can easily be used on such data. Mode is an observation in a distribution that appears the most often.

The example shows the working of mode on a categorical data.

10

Which is the Best Measure of Central Tendency? (Contd..)

- Geometric Mean is the best measure of central tendency for calculating rate of change
- The steps in calculating projections or rate of returns using Geometric Mean
 - Add 1 (100 in case of percentage) is to each number to offset the issue of negative numbers
 - Multiply all the numbers
 - Take the root of count of the numbers in the series.
 - Subtract one from the result

Example: The housing price of a city increased by 10% in year 1 and decreases by 10% in year 2 at the current price. What is the average change in the housing price?

Suppose 100 is the original price. A 10% increase will yield 110 and 10% decrease will be 90

<p>Arithmetic Mean: $\frac{10+(-10)}{2} = 0 \rightarrow$ There is no change</p>	<p>Geometric Mean: $\sqrt[2]{110 * 90} = 99.5 \rightarrow$ 0.5% decrease</p>
---	--

Geometric Mean is used while calculating average rate of change. One of the most important reasons for this is because it takes into account the effects of compounding. The steps for calculating average rate of change is given in the steps below:

- Add 1 (100 in case of percentage) to the rate of change to offset the issue of negative numbers. This happens when there is a negative change year on year.
- Calculate Geometric Mean of the numbers calculated in step #2.
- Subtract one 100 in case of percentage) from the result.

An example of Arithmetic Mean Vs Geometric Mean while calculating average rate of change is shown with an example.

11

Application of Geometric Mean - Projections

The Geometric Mean takes into account the compounding that occurs from period to period whereas the Arithmetic Mean does not.

Geometric Mean is used for calculating projections, rate of returns etc

The steps in calculating projections or rate of returns using Geometric Mean

1. Calculate the percentage of change for the successive periods (years)..
2. 1 is added to each percentage change calculated above to offset the issue of negative numbers.
3. Take the Geometric Mean of the numbers calculated in step 2
4. Subtract one from the result
5. Now, calculate the population multiplier as 1 + (result in step 4 /100)
6. Last available population number multiplied by population multiplier times the time gives the projected population. Time is the difference between year for population is to be projected and the last year for which population numbers are available

Population Projections consist of mathematical models which are used to analyse changes in population numbers. One of the basic models for estimating population is Geometric Mean as it is the preferred measure for calculating average rate of change. The geometric mean taken into account the compounding that occurs from period to period. Hence it is preferred over arithmetic mean.

Application of Geometric Mean – Projections (Contd..)

This forecasts the population for 2030 based on the population data between 2011 and 2023 using Geometric Mean

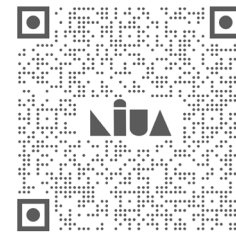
A. Year	B. Population	C. Change %	D. Add 1
2011	33406		
2012	33642	0.7065	1.7065
2013	33875	0.6926	1.6926
2014	34109	0.6908	1.6908
2015	34344	0.6890	1.6890
2017	34578	0.6813	1.6813
2017	34761	0.5292	1.5292
2018	34943	0.5236	1.5236
2019	35125	0.5208	1.5208
2020	35307	0.5181	1.5181
2021	35489	0.5155	1.5155
2022	35633	0.4058	1.4058
2023	35776	0.4013	1.4013

Geometric Mean	1.56908
Geometric Mean - 1	0.5691
Time period	7
2030 Population Projection	37225.7

giz

The steps for forecasting the population for 2030 based on the population data between 2011 and 2023 are shown.

- Calculate the annual percentage change as (population for the given year - population for previous year) divided by population of previous year times 100. The results are given in column C.
- Add 100 to the percentage change calculated above.
- Take Geometric mean of numbers given in column D.
- Subtract 100 from result above
- Calculate the population multiplier as 1 plus result from step 4 divided by 100
- Take the difference between 2030 (projection year) and 2023 (last population available year)
- Multiply $35776 * (1 + 0.5691/100)^7$

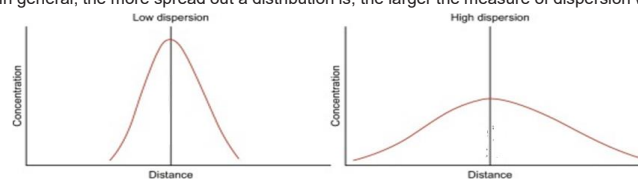


Projection Techniques

Descriptive Statistics – Measures of Dispersion

Measures of dispersion are descriptive statistics that describe how similar a set of data in a distribution are to each other or what is the spread of the data in a distribution

- The more similar the data are to each other, the lower the measure of dispersion will be and vice versa
- In general, the more spread out a distribution is, the larger the measure of dispersion will be



Example: {1,1,2,2,4} and {2,2,2,2,2} both give an arithmetic mean of 2. Then how to differentiate the two distributions? Measure of dispersion helps in this.

giz

The measures of dispersion calculate the deviation of data from the central tendency, usually the mean. Similar the data, lower is the dispersion and vice versa. In case of larger dispersion, the curve is wider, as seen in the figure on the right hand side.

The importance of measure of dispersion is clearly shown with the help of an example. The two distributions, {1,1,2,2,4} and {2,2,2,2,2} have the exact same arithmetic mean of 2. The measure of dispersion helps to differentiate the two distributions. The first distribution has more variation while the second distribution has no variation.

Descriptive Statistics – Measures of Dispersion (Contd..)

Range	The difference between maximum and minimum value
Variance	The average of the square deviations from the mean $\sigma^2 = \frac{\sum_1^N (X-\mu)^2}{N}$
Standard Deviation	The square root of variance

- The Standard Deviation is the most popular measure of dispersion
- The larger the variance is, the more the scores deviate, on average, away from the mean and vice versa

giz

- Range, Variance and Standard deviation are the most widely used measures of dispersion.
- Range is the difference between the maximum and minimum value.
- Calculation of Range in Excel: There is no direct function to calculate the range. Find the maximum value and the minimum value among the data points and subtract to calculate range.
- Calculating Minimum value: <https://support.microsoft.com/en-us/office/max-function-e0012414-9ac8-4b34-9a47-73e662c08098>
- Calculating Maximum value: <https://support.microsoft.com/en-us/office/calculate-the-smallest-or-largest-number-in-a-range-45fe249f-96c3-443b-8e9f-87f16c48462c>
- Calculation of variance involves squaring the difference between each data point and the mean, adding up those squares.
- The standard deviation is calculated by taking the square root of the variance.

Measures of Dispersion – Variance

- Variance is the average of the square deviations from mean
- The **population variance is denoted by σ^2** and **sample variance by S^2**

$$\sigma^2 = \frac{\sum_1^N (X-\mu)^2}{N}$$

$$S^2 = \frac{\sum_1^n (X-\bar{x})^2}{n-1}$$

Example: The variance for two distributions discussed in the previous slide is given below

x	1	1	2	2	4	Mean = 2
x-2	-1	-1	0	0	2	
(x-2) ²	1	1	0	0	4	Variance = 1.5

x	2	2	2	2	2	Mean = 2
x-2	0	0	0	0	0	
(x-2) ²	0	0	0	0	0	Variance = 0

The larger the variance is, the more the data deviates, on average, away from the mean and vice versa

giz

- Range, Variance and Standard deviation are the most widely used measures of dispersion.
- Range is the difference between the maximum and minimum value.
- Calculation of Range in Excel: There is no direct function to calculate the range. Find the maximum value and the minimum value among the data points and subtract to calculate range.
- Calculating Minimum value: <https://support.microsoft.com/en-us/office/max-function-e0012414-9ac8-4b34-9a47-73e662c08098>
- Calculating Maximum value: <https://support.microsoft.com/en-us/office/calculate-the-smallest-or-largest-number-in-a-range-45fe249f-96c3-443b-8e9f-87f16c48462c>
- Calculation of variance involves squaring the difference between each data point and the mean, adding up those squares.
- The standard deviation is calculated by taking the square root of the variance.

16

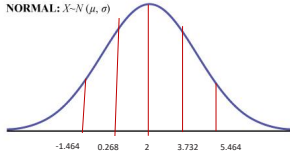
Measures of Dispersion – Standard Deviation

- Standard Deviation is the square root of variance
- It is the most popular measure of dispersion
- The **population variance is denoted by σ** and **sample variance by S**

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X - \mu)^2}{N}}$$

$$S = \sqrt{\frac{\sum_{i=1}^n (X - \bar{x})^2}{n-1}}$$

NORMAL: $X \sim N(\mu, \sigma)$



Example: The standard deviation for two distributions discussed in the previous slide is given below

$\sigma \{1, 1, 2, 2, 4\} = \sqrt{3} = 1.783$

$\sigma \{2, 2, 2, 2, 2\} = \sqrt{0} = 0$

giz

Standard Deviation is the measure of the dispersion of statistics. Standard deviation formula is used to find the deviation of the data value from the mean value i.e. it is used to find the dispersion of all the values in a data set to the mean value. The standard deviation is calculated by taking the square root of the variance.

Calculating Standard Deviation

- <https://support.microsoft.com/en-au/office/stdev-function-51fecaaa-231e-4bbb-9230-33650a72c9b0>

17

Measures of Dispersion – Coefficient of Variation

Coefficient of Variation

- The ratio of the standard deviation to the mean.
- $CV = \frac{\sigma}{\mu}$
- Indicates the size of a standard deviation in relation to its mean
- The higher the co-efficient of variation, the greater the dispersion level around the mean
- Helps to compare the degree of variation from one data series to another.

City	Housing Prices (₹)					Mean	Standard Deviation	CV
Bangalore	74,05,000	64,05,000	45,06,000	77,00,000	93,69,000	70,77,000	17,89,595.07	0.25
Trivandrum	54,02,000	34,02,000	45,01,000	57,00,000	27,69,000	43,54,800	12,59,975.67	0.29

giz

- Coefficient of variation is a type of relative measure of dispersion. It is expressed as the ratio of the standard deviation to the mean. It indicates the size of standard deviation in relation to its mean. Higher the coefficient of variation, greater is the dispersion around the mean.
- The coefficient of variation helps to compare two data sets on the basis of the degree of variation.
- The example compares the housing prices between Bangalore and Trivendrum. The standard deviation is more in Bangalore. However, coefficient of variation is more in Trivendrum. This indicates the variations in housing prices is more in Trivendrum and higher standard deviation in Bangalore is mainly due to larger numbers.
- Calculation of coefficient of variation in excel.
- There is no direct function to calculate coefficient of variation in excel.
- First calculate mean : <https://support.microsoft.com/en-au/office/average-function-047bac88-d466-426c-a32b-8f33eb960cf6>
- Calculate standard deviation: <https://support.microsoft.com/en-au/office/stdev-function-51fecaaa-231e-4bbb-9230-33650a72c9b0>
- Divide mean by standard deviation to get the coefficient of variation.



SD Var CV

18

Descriptive Statistics - Measures of Position

- A measure of position determines the position of a value in relation to other values in a distribution
- Quartiles** divide the data into four equal parts
- Percentiles** divide it into hundredths, or 100 equal parts
- Percentiles and Quartiles are important measures of position
- The median is also the 50th percentile.

giz

The slide is self explanatory. The measure of position determines the position of a value in relation to other values in a distribution. Percentiles, Quartiles and Ranks are important measures of position. Quartiles divide the data into 4 equal parts while percentiles divide the data into 100 equal parts. Ranks order the data from lower to higher or higher to lower.

19

Measures of Position - Percentile

- Percentile is a value indicates the percent of a distribution that is equal to or below it
- Percentiles are commonly used to report scores in tests, like the SAT, GRE etc
- Percentile rank of a value say x is calculated as

$$\frac{\text{Number of observations below } x}{n} * 100, n \text{ is the size of the distribution}$$

Example: 7,3,12,15,14,47,3,12,15,14,4,20, Determine the percentile ranking of 14

Solution:

- The distribution needs to be arranged first : 3,3,4,7,12,12,14,14,15,15,20,47
- Percentile rank of 14 = $\frac{\text{Number of values below 14}}{12} * 100 = \frac{6}{12} * 100 = 50$
- This implies that 50% of the values are below 14

giz

- While there is no universal definition of percentile, it is commonly expressed as the percentage of values in a set of data set that fall below a given value.
- The calculation of percentile ranks in excel.
- <https://support.microsoft.com/en-us/office/percentrank-function-f1b5836c-9619-4847-9fc9-080ec9024442>

Measures of Position – Percentile (Contd..)

20

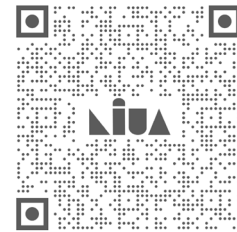
Example: Determine the 35th percentile of the distribution 7,3,12,15,14,47,3,12,15,14,4,20

Solution:

- The distribution needs to be arranged first : 3,3,4,7,12,12,14,14,15,15,20,47
- The index of the value at the 35th percentile rank = $\frac{\text{Percentile} * (n+1)}{100} = \frac{35}{100} * 13 = 4.55$
- The average of the values at the 4th and 5th position = $\frac{7+12}{2} = 9.5$

giz

- Determines the value at a certain percentile in a given dataset.
- In the given example, the value at 90th percentile is 88. This means that all the students who scored above 88 would be above 90% of the students.
- Calculation of Percentile in excel.
- <https://support.microsoft.com/en-us/office/percentile-function-91b43a53-543c-4708-93de-d626debdddca>



Percentile

Measure of Position - Quartile (Contd..)

21

There are three quartile values — a **lower quartile**, **median**, and **upper quartile** - to divide the distribution into four intervals, each containing 25% of the data points

The lower quartile, or first quartile, is denoted as Q1 is the value at the 25 th percentile	$Q1 = \frac{n+1}{4}$ th observation
The median is the middle value which is at the 50 th percentile	$\text{Median} = \frac{n+1}{2}$ th observation
The upper quartile, or third quartile, is denoted as Q3 is the value at the 75 th percentile	$Q3 = \frac{3(n+1)}{4}$ th observation

Where n is the number of observations

giz

Quartiles are three values that divide sorted data into four parts, each with an equal number of observations.

First quartile: Also known as Q1, or the lower quartile. This is the number halfway between the lowest number and the middle number.

Second quartile: Also known as Q2, or the median. This is the middle number halfway between the lowest number and the highest number.

Third quartile: Also known as Q3, or the upper quartile. This is the number halfway between the middle number and the highest number.

The calculation of Q1 and Q3 is similar to median.

- Arrange the data, either smallest to largest or largest to smallest.
- For Q1, divide the number of observations by 4. If there are an odd number of observations, round that number up, and the value in that position is the Q1. If the number of observations is even, take the average of the values found above and below that position.
- For Q3, multiply the number of observations by 3 and then divide the result by 4. If there are an odd number of observations, round that number up, and the value in that position is the Q3. If the number of observations is even, take the average of the values found above and below that position.

22

Measure of Position - Quartile (Contd..)

Example: 5, 7, 4, 4, 6, 2, 8
Solution:
 Arrange the distribution in an order 2, 4, 4, 5, 6, 7, 8

$Q1 = \frac{7+1}{4} = 2$ The 2 nd observation is 4	$Q2 = \frac{7+1}{2} = 4$ 4 th observation is 5	$Q3 = \frac{3+(7+1)}{4} = 6$ 6 th observation is 7
--	--	--

Example: 5, 7, 3, 4, 6, 2, 8, 10
Solution:
 Arrange the distribution in an order 2, 3, 4, 5, 6, 7, 8, 10

$Q1 = \frac{8+1}{4} = 2.25$ The average of 2 nd and 3 rd observation is 3.5	$Q1 = \frac{8+1}{2} = 4.5$ The average of 4 th and 5 th observation is 5.5	$Q3 = \frac{3+(8+1)}{4} = 6.75$ The average of 6 th and 7 th observation is 7.5
--	---	--

giz

- The calculation of quartiles in excel is given in the link below:
- <https://support.microsoft.com/en-us/office/quartile-function-93cf8f62-60cd-4fdb-8a92-8451041e1a2a>
- The examples in the slide show the calculation of quartiles.

23

Outliers

- A lower fence and higher fence need to be defined to identify the outlier
- All the values below lower fence and above higher fence are outliers
- Formulas for lower and higher fence are as below:
 - **Lower Fence = $Q1 - 1.5 (IQR)$**
 - **Higher Fence = $Q3 + 1.5 (IQR)$**
 - **Inter Quartile Range (IQR) = $Q3 - Q1$**

Example: Identification of outlier can be explained with the following example
 {1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9,27} is the distribution where n = 19

$Q1 = \frac{19+1}{4} = 5$; 5 th observation is 3	IQR = 7 - 3 = 4	Lower Fence = $3 - (1.5 * 4) = 3 - 6 = -3$
$Q3 = \frac{3(19+1)}{4} = 15$; 15 th observation is 7		Higher Fence = $7 + (1.5 * 4) = 7 + 6 = 13$

So all the values below -3 and higher than 13 is an outlier. So, only 27 is the outlier

giz

An outlier is an observation that lies an abnormal distance from other observations in the dataset. To calculate outlier, two metrics, a higher fence and lower fence need to be calculated. These are derived using first quartile(Q1) and second quartile(Q3).

- The steps for identifying outliers is given below:
- Calculate Q1 and Q3. The calculation of Q1 and Q3 are given in the previous slide.
 - Calculate Inter Quartile Range (IQR) difference between Q3 and Q1.
 - Lower Fence is Q1 minus 1.5 times IQR.
 - Higher Fence is Q3 plus 1.5 times IQR.
 - All the values below Lower Fence and all the values above Higher Fence are outliers.

The identification of an outlier is shown with an example. There is no direct formula to calculate an outlier in excel. It has to be identified using the steps described above. The lower fence is calculated as Q1 minus 1.5 times.

Statistical Data Distributions

In simple terms, a distribution is a collection of data, or scores, on a variable.

Usually, these scores are arranged in order from smallest to largest and then they can be presented graphically

Scores or data can be either discrete or continuous type.

Depending on the type of data, the distribution are also referred as discrete or continuous

Important discrete distributions include Bernoulli Distribution, Binomial distribution, Poisson distribution

Continuous distribution includes Gaussian or normal distribution. Many data from various fields, including social science conforms to this distribution and hence the focus of much of the field of statistics

Many tests are designed normally distributed populations

giz

- All along the term 'distribution' has been used. In simple terms, it is a collection of data or scores or observations on a variable or a topic under study.
- Scores or data can be either discrete or continuous type. If the data is discrete, the corresponding distribution is referred as discrete distribution. If the data is continuous, it is a continuous distribution.
- Important discrete distributions include Bernoulli Distribution, Binomial distribution, Poisson distribution.
- Continuous distribution includes Gaussian or normal distribution. Many data from various fields, including social science conforms to this distribution and hence the focus of much of the field of statistics.
- The continuous distributions are the focus of discussion here. Many tests are designed normally distributed populations.

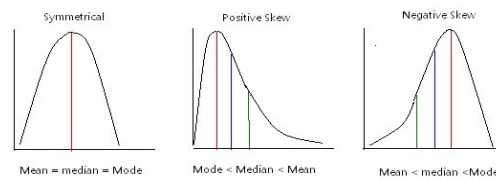
Measures of Shape- Skewness

Skewness is a measure of symmetry.

A distribution, or data set, is symmetric if it looks the same to the left and right of the centre point

A symmetrical distribution is also referred as Normal Distribution

The mean, median and mode are exactly the same.



giz

- Skewness is a measure of asymmetry of a distribution.
- If the long tail is on the right, then the skewness is rightward or positive; if the long tail is on the left, then the skewness is leftward or negative.
- In case of positive skewness, there are higher number of data points with lower values. For example, scores on an hard exam are likely to have right skewness, with most scores close to the minimum.
- In case of negative skewness, there are higher number of data points with higher values. For example, scores on an easy exam are likely to have left skewness, with most scores close to the maximum,
- The skewness for a normal distribution is zero indicating symmetry. Here the values for mean, median and mode are exactly the same.
- Calculating skewness in excel: <https://support.microsoft.com/en-au/office/skew-function-bdf49d86-b1ef-4804-a046-28eaea69c9fa>

Z-Score

Z-SCORE measures exactly how many standard deviations above or below the mean a value is

The formula for calculating z-score is given by $\frac{x-\mu}{\sigma}$,

where x is the value, μ is the mean and σ the standard deviation

Here are some important facts about z-scores:

- A **positive z-score** says the value is **above average**.
- A **negative z-score** says the value is **below average**.
- A **z-score close to 0** says the value is **close to average**.
- When the **z-score is greater than 3 or less than -3**, it indicates that the value is significantly different from the average and can be considered **unusual or an outlier**.

giz

Z-score is a statistical measurement that describes a value's relationship to the mean in a distribution. Z-score is a dimensionless quantity that is measured in terms of standard deviations from the mean.

The z-score of a value in a distribution is calculated by taking the difference between the value and mean subtracted by standard deviation.

If a Z-score is 0, it indicates that the value is identical to the mean.

- A positive z-score says the value is above the mean.
- A negative z-score says the value is below the mean.
- A value can be considered unusual if its z-score is above or below 3 standard deviation.

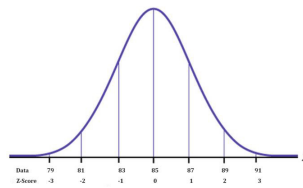
There is no built-in function to calculate z-score in excel.

- Calculate mean: <https://support.microsoft.com/en-us/office/average-function-047bac88-d466-426c-a32b-8f33eb960cf6>
- Calculate standard deviation: <https://support.microsoft.com/en-us/office/stdev-p-function-6e917c05-31a0-496f-ade7-4f4e7462f285>
- Subtract mean from the value and divide it by standard deviation to get the Z-score.

27

Z-Score – An Example

The grades on a history midterm at a school have a mean of 85 and a standard deviation of 2. It can be shown through the diagram as below.



- If a student scores 86, the z-score can be calculated as $\frac{86-85}{2} = 0.5$
- The student's score is **0.5σ above the mean**

giz

- An example showing the calculation of Z-score along with the graphical representation is shown in this slide.

28

Applications of Z-Score

Z-scores are often used for feature scaling to bring different features to a common scale.

*Suppose there is a dataset with Age, Height and Weight of a child. The Age is in years, weight in Kgs and Height is in inches or feet. The application of z-scores makes the features unit less and brings them to a common scale with $\mu = 0$ and $\sigma = 1$. This process is known as **standardization**. Feature scaling is commonly used before application of an ML algorithm.*

Z-scores can be used to identify outliers in a dataset. Data points with Z-scores beyond a certain threshold (usually 3 standard deviations from the mean) may be considered outliers.

When working with regression models, Z-scores of residuals can be analyzed to check for homoscedasticity (constant variance of residuals)

giz

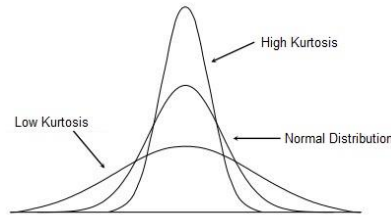
- The different applications of Z-score are described in the slide. They are self explanatory.
- The two most important applications of Z-score in statistics is the identification of outliers and feature scaling to bring different features to a common scale.
- Data points with Z-scores beyond a certain threshold may be considered outliers. The Empirical Rule states that 99.7% of data observed following a normal distribution lies within 3 standard deviations of the mean. Hence any data outside 3 standard deviation can be considered as an outlier.
- The application of z-scores makes the features unit less and brings them to a common scale with $\mu = 0$ and $\sigma = 1$. This process is known as standardisation. Feature scaling is commonly used before application of an ML algorithm.

Measures of Shape- Kurtosis

Kurtosis refers to the pointedness of a peak in the distribution curve. It helps to identify outliers present in the distribution.

High kurtosis in a data set is an indicator that data has heavy tails or outliers.

Low kurtosis in a data set is an indicator that data has light tails or lack of outliers.



giz

Kurtosis is a statistical measure that quantifies the shape of a probability distribution. It provides information about the tails and peakedness of the distribution compared to a normal distribution. Here are some key points about kurtosis:

- **Positive Kurtosis:** When a distribution has positive kurtosis, it means that the tails are heavier (fatter) than those of a normal distribution. The peak of the distribution is more pronounced or “pointy.” This indicates that extreme values (outliers) are more likely to occur.
- **Negative Kurtosis:** Conversely, negative kurtosis indicates lighter tails than a normal distribution. The peak is flatter, and extreme values are less likely.
- **Mesokurtic Distribution:** A distribution with kurtosis close to zero is called mesokurtic. It resembles a normal distribution.
- **Leptokurtic Distribution:** A leptokurtic distribution has positive kurtosis. It is more peaked and has heavier tails than a normal distribution. Examples include financial returns and stock market data.
- **Platykurtic Distribution:** A platykurtic distribution has negative kurtosis. It is flatter and has lighter tails than a normal distribution. Examples include uniform distributions and some biological measurements.

Normalisation

Normalisation is another popular feature scaling method. The process of converting values in a distribution between 0 to 1 is known as Normalisation

Min- Max scaler is the most popular Normalisation technique. It is calculated as $\frac{x - \text{Min}}{\text{Max} - \text{Min}}$

where x is the value, Min is the minimum value in the distribution and max is the maximum value in the distribution

The other technique is to divide the values of the distribution by the maximum value in the distribution.

In deep learning, while doing Image classification, each pixel ranges between 0 to 255. Before training the image, these pixels are converted in the range 0 to 1 by dividing each cell by 255

giz

- Normalisation is another popular feature scaling method. Through Normalisation, the values in the distribution are converted between 0 and 1.
- Min- Max scaler is the most popular Normalisation technique. The calculation is given in the slide. The normalized score of a value is the difference between value and the minimum value in the distribution divided by the difference between maximum and minimum value in the distribution.
- Normalisation is the first step in implementation of many machine learning algorithms. This is also the step while developing an index.

Normalisation calculation methods

Normalization transforms the data into a standard scale which is unit less

Popular Normalisation Methods

•Min-Max Method

▪Positive Indicator $\rightarrow X_{i_Normalised} = \frac{X_i - Min(X)}{Max(X) - Min(X)}$

◦Negative Indicator $\rightarrow X_{i_Normalised} = \frac{Max(X) - X_i}{Max(X) - Min(X)}$

Z-Score Method

Positive Indicator $\rightarrow X_{i_Normalised} = \frac{X_i - Mean(X)}{std.dev(X)}$

◦Negative Indicator $\rightarrow X_{i_Normalised} = \frac{Mean(X) - X_i}{std.dev(X)}$

giz

- The min-max method, z-score are some of the popular normalization techniques. The min-max transforms the data into values between 0 and 1.
- The z-score calculates values that measure deviation away from mean. The calculation for these techniques for both positive and negative indicators are given.

Correlation

A statistical technique used to determine the degree of relationship between two variables;
Ex : Price and Demand of a product

Its primary goal is to identify the existence or lack of a relationship between two variables. In general, its main goal is to determine a numerical value that demonstrates the connection between the variables and how they move together.

Correlation can be positive, negative, or no correlation.	Positive Correlation is a relationship in which of one variable increases, the other one also increased and vice-versa
	Negative Correlation is a relationship in which one variable increases, the other one decreases and vice-versa
	No correlation indicates there is no apparent relationship

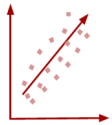
giz

- Correlation is a bi-variate statistical technical technique. It is used to understand the relationship between two variables. For example. The price of a product and its demand, Education and Income etc. Its main aim is to see if there's a connection between two things. Basically, it tries to find a number that shows how the two things are related and how they change together.
- The correlation can be positive, negative or no relationship.
- A positive correlation is a relationship in which, if one variable increases, the other variable also increases and vice-versa. Education and income are perceived to have a positive relationship as higher education leads to higher income.
- Negative Correlation is a relationship in which one variable increases, the other one decreases and vice-versa. For example, if the price of product increases, the demand reduces.
- No correlation indicates that there is no relationship.

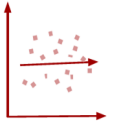
33

Correlation (Contd..)

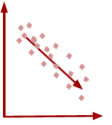
The value of the correlation coefficient varies between +1 and -1 with ± 1 indicating a perfect relationship.



Positive
Correlation



Zero
Correlation



Negative
Correlation

Strong Negative	No relationship	Strong Positive
-1	0	+1

The coefficient of correlation is the measure of strength and direction of the relationship between two variables. It is denoted by r .

The Pearson correlation technique and Spearman correlation are popular techniques for measuring the relationship between two variables.

giz

- The strength and direction of the relationship is measured through correlation coefficient. It is denoted by r . The value of r varies between +1 and -1. Closer the values are to ± 1 , higher is the strength of the relationship.
- The Pearson correlation coefficient is the most common way of measuring a linear correlation between two variables. The Spearman's rank correlation coefficient is the other popular correlation technique.
- Running Pearson's correlation in excel: <https://support.microsoft.com/en-au/office/correl-function-995dcef7-0c0a-4bed-a3fb-239d7b68ca92#:~:text=The%20CORREL%20function%20returns%20the,the%20use%20of%20air%20conditioners>

34

Correlation (Contd..)

Ex : Finding the relation between electricity consumption and temperature recorded by using Pearson's coefficient

Temperature	Electricity (Units Consumed)
24	80
27	82
30	84
31	101
34	110
35	115
38	140
40	142
42	156
45	157

Correlation = 0.9736

giz

Running Pearson's correlation in excel: <https://support.microsoft.com/en-au/office/correl-function-995dcef7-0c0a-4bed-a3fb-239d7b68ca92#:~:text=The%20CORREL%20function%20returns%20the,the%20use%20of%20air%20conditioners>

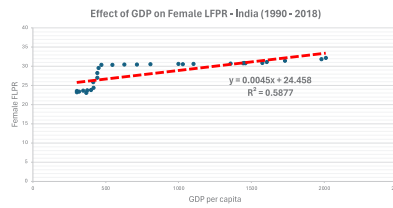


Correlation

Linear Regression

A Statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables

Example: Effect of Per Capita Income on Female Labour Force Participation Rate in India (F-LFPR) (1990-2018)



Types of Linear Regression	Simple linear regression	one independent variable
	Multiple linear regression	more than one independent variable

Regression lines can be used as a way of visually depicting the relationship between the independent (x) and dependent (y) variables in the graph.

giz

- Correlation describes the strength and direction of a relationship between two or more variables. However, it does not measure the change in one variable as a result of the change in the values of the other variable.
- Regression is the best technique to measure the impact of one variable on the other.
- The impacted variable is called target or dependent variable. The impacting variables are called predictor or independent variables.
- The independent variables can be one or many while there is only one dependent variable.
- In simple regression, there is one independent variable.
- If there are more than one independent variables, it is called multi-variate regression.

38

Linear Regression (Contd..)

The mathematical equation for Linear regression:

$Y = b + aX$

Y - Dependent Variable
X - Independent Variable
b - the intercept
a - the linear coefficient

The intercept b is the value of dependent variables when value of independent variable is 0.

The regression coefficient value represents the change in the dependent variable given a one unit change in the independent variable

- A positive sign indicates that as the independent variable increases, the dependent variable also increases.
- A negative sign indicates that as the independent variable increases, the dependent variable decreases

giz

One of the examples of regression is predicting the sale price of a house based on age, size, number of bedrooms, location etc.

- The mathematical equation for Linear regression: $Y = b + aX$.
- The intercept 'b' is the value of dependent variables when value of independent variable is 0.
- The regression coefficient 'a' measures the change in the dependent variable 'Y' given a one unit change in the independent variable 'X'.
 - A positive sign indicates that as the independent variable increases, the dependent variable also increases.
 - A negative sign indicates that as the independent variable increases, the dependent variable decreases.

Running Linear Regression in Excel:

- Load the Analysis tool pack in Excel : <https://support.microsoft.com/en-au/office/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4>
- Follow these steps to run regression analysis.
 - Select "Data" from the toolbar. The "Data" menu displays.
 - Select "Data Analysis". The Data Analysis - Analysis Tools dialog box displays.
 - From the menu, select "Regression" and click "OK".
 - In the Regression dialog box, click the "Input Y Range" box and select the dependent variable data (Visa (V) stock returns).
 - Click the "Input X Range" box and select the independent variable data (S&P 500 returns).
 - Click "OK" to run the results.

39

Construction of Index

An index is a composite measure that aggregates multiple indicators or data points

It is a way of measuring a construct—like ease of living—using more than one data item.

It is used to summarise and rank specific observations.
 Examples of index include Municipal Performance Index, Ease of Living Index, SDG Index etc

Steps in the construction of Index

```

            graph LR
            A[Developing Indicator Framework] --> B[Collection of Raw Data for each indicator from various data sources]
            B --> C[Normalisation of raw data]
            C --> D[Developing an Index]
            D --> E[Categorisation of geographic areas based on the index]
            
```

giz

An index is a composite measure that aggregates multiple indicators or data points. It facilitates to measure a construct or a subject of interest such as poverty, governance, ease of living etc that are difficult to measure otherwise. Often, such constructs are defined by multiple variables (indicators) and hence the requirement to aggregate them to arrive at a score, known as an index.

Steps in the construction of an index

- Identifying the variables or indicators that define the construct.
- Collection of data for the identified indicators.
- Normalisation of data.
- Developing an index involves aggregating the indicator data.
- The last step is optional. This is particularly used when index is developed for geographic areas. Based on the index, the geographies can be classified.

40

Construction of Index – Ease of Living Index

The Ease of Living Index Index is an annual report published by the Ministry of Housing and Urban Affairs

It is an approach used to assess and rank cities or regions based on the quality of life and overall ease of living experienced by their residents.

The methodology was developed by the Ministry of Housing and Urban Affairs (MoHUA) in India as part of their Ease of Living Index initiative

The Index examines liveability of 114 Indian cities

The Ease of Living Framework is organized in a tree structure under 4 pillars that include 14 categories and 77 indicators on the subjects of Quality of Life, Economic Ability and Sustainability.

The index is constructed at the category and pillar level along with the overall index

giz

- Ease of Living Index is an assessment framework launched in 2017 by the Ministry of Housing and Urban Affairs to assess the quality of life in cities. The main objective of the index is to enable data driven approach in urban planning and management and promote healthy competition among cities.
- The Index examines liveability of 114 Indian cities.

Construction of Index – Developing Indicator Framework

41

Indicator is a data element, a variable or a measure of something

Framework is the Structure

An indicator Framework is a tool for organising indicators to measure the specified goals and objectives

Ease of Living Index
 Cycle 1: 2018
 Cycle-2: 2020

35% Weightage
Quality of Life

- Education
- Health
- Housing and Shelter
- WASH and SWM
- Mobility
- Safety and Security
- Recreation

15% Weightage
Economic Ability

- Level of Economic Development
- Economic Opportunities

20% Weightage
Sustainability

- Environment
- Green Space and Buildings
- Energy Consumption
- City Resilience

30% Weightage
Citizen Perception Survey

- Citizen Perception Survey

giz

The Ease of Living Framework is organized in a tree structure under 4 pillars that include 14 categories and 77 indicators on the subjects of Quality of Life, Economic Ability and Sustainability.

- Measuring Ease of Living in various cities in the final objective.
- This is measured through the pillars, Quality of Life, Economic Ability and Sustainability,
- The citizen perception survey is undertaken that captures perception of citizens with respect to quality of life in their cities. The results of this survey along with Quality of Life, Economic Ability and Sustainability scores to arrive at the final index.
- The Quality of Life is defined by 7 categories, Economic Ability by 2 categories and Sustainability by 4 categories.
- The Ease of Living, 3 pillars and 14 categories are constructs that cannot be measured directly.
- 77 indicators are defined to directly measure something. These 77 indicators are distributed across the 14 categories.
- The indicators are aggregated to first arrive at the category index. The category index are then aggregated to arrive at the pillar index. The pillar index along with the results of citizen perception survey are aggregated to arrive at the Ease of Living Index.

Construction of Index – Developing Indicator Framework (Contd..)

42

Ease Of Living index

giz

The diagram shows part of the indicator framework. The categories and indicators under Quality of Life are depicted in the diagram.

The indicator can be positive or negative. The positive indicators are those where higher the value, better the performance and vice versa. The negative indicators are those where lower the value, better the performance and vice versa.

Literacy rate is a positive indicator whereas drop out rate is a negative indicator. Similarly, availability of hospital beds is a positive indicator whereas infant mortality rate is a negative indicator.

Construction of Index – Data Collection

43

Identification of required data guided by the data framework

Urban data can be sourced from primary surveys or secondary datasets

Main Sources of Secondary data sets

- Open data initiatives from the government (mydata.gov.in)
- Internal data with the government
- Data from various surveys conducted by the government departments and private entities
- Satellite imagery data
- Web extracted data

giz

- The data for the identified indicators can be sourced from various secondary or primary data sets. In the development or governance sector, the secondary sources include:
 - Open data initiatives from the government (mydata.gov.in)
 - NITI Aayog’s NDAP platform
 - Internal data with the government
 - Data from various surveys conducted by the government departments and private entities
 - Satellite imagery data
 - Web extracted data

The primary datasets include surveys and census.

Construction of Index – Aggregation

44

Various Methodologies exist. Simple and Weighted Average are the most popular. This methodology is used for calculating urban indices such as municipal Performance Index, Ease of Living Index, SDG Urban India Index etc.

Category Score : A simple average of the normalised indicator value

$$\text{E.g.: } CS_{\text{Health}} = \frac{\sum_1^6 WX_i}{6}$$

Pillar Score = This is the simple average of category scores

Methodology for Ease of Living Index

$$\text{E.g.: } PS_{\text{Quality of Life}} = \frac{W \cdot CS_{\text{Health}} + W \cdot CS_{\text{Education}} + W \cdot CS_{\text{Wash}} + W \cdot CS_{\text{Mobility}} + W \cdot CS_{\text{Housing}} + W \cdot CS_{\text{Safety}} + W \cdot CS_{\text{Recreation}}}{7}$$

Overall Score = This is the weighted average of Pillar Score

$$\text{Ease of Living Index} = \frac{0.35 \cdot PS_{\text{Quality of Life}} + 0.15 \cdot PS_{\text{Economic Ability}} + 0.2 \cdot PS_{\text{Sustainability}} + 0.3 \cdot PS_{\text{Citizen Perception Survey}}}{100}$$

giz

Data normalization is a technique used to transform the values of a dataset into a common scale. The data is transformed into a standard scale that is unit less.

For example, a data set may include population and area of barren land of a district. While population is usually in lakhs, the barren land is in hundred or thousands. The data science techniques work better if the variables are on a similar scale. Hence, normalization of data is important.

The min-max method, z-score are some of the popular normalization techniques. The min-max transforms the data into values between 0 and 1. The z-score calculates values that measure deviation away from mean. The calculation for these techniques for both positive and negative indicators are given.

Construction of Index – Ranking and Categorisation

Rank	Million + City	Score	Rank	Less than Million City	Score
1	Bengaluru	66.70	1	Shimla	60.90
2	Pune	66.27	2	Bhubaneswar	59.85
3	Ahmedabad	64.87	3	Silvassa	58.43
4	Chennai	62.61	4	Kakinada	56.84
5	Surat	61.73	5	Salem	56.40
6	Navi Mumbai	61.60	6	Vellore	56.38
7	Coimbatore	59.72	7	Gandhinagar	56.25
8	Vadodara	59.24	8	Gurugram	56.00
9	Indore	58.58	9	Davanagere	55.25
10	Greater Mumbai	58.23	10	Tiruchirappalli	55.24

The final ranking is given separately for cities with more than one million population and cities with less than one million population.

Construction of Index – SDG 11 Index

The India SDG index is an annual report published by NITI Aayog. It computes index for SDG 1 to 16. The unit of index is state

The main mission of SDG 11 titled "sustainable cities and communities" is to "Make cities inclusive, safe, resilient and sustainable".

The indicator framework for SDG 11 (2020-21) consists of 8 indicators. The indicators and corresponding data sources are given below:

Indicator	Data Source
Percentage of wards with 100% door to door waste collection	SBM –U, Ministry of Housing and Urban Affairs
Percentage of individual household toilets constructed against target	
Percentage of wards with 100% source segregation	
Percentage of MSW processed to the total MSW generated	Ministry of Statistics and Program Implementation
Percentage of urban households living in katcha houses	
Percentage of urban households with drainage facility	Ministry of Home Affairs
Deaths due to road accidents in urban areas	
Installed sewage treatment capacity as a % of sewage generated in urban	Ministry of Environment, Forest & Climate Change

SDG 11 Index – Normalisation of Raw Data

The SDG index methodology for all the goals uses distance to target methodology to normalise the data

State-wise data values of each of the Indicators is rescaled from its raw form into a score ranging from 0 to 100— with 0 denoting lowest performer and 100 indicating that the target has been achieved.

For some values, increasing value means worse performance. These are called as negative indicators. Otherwise, indicators are positive.

In SDG 11 indicator framework, Percentage of urban households living in katcha houses and Deaths due to road accidents in urban areas are negative indicators. The other indicators are positive indicators.

SDG 11 Index – Normalisation of Raw Data (Contd..)

To make data comparable across indicators, State-wise data values of each of the National Indicators was rescaled from its raw form into a score ranging from 0 to 100— with 0 denoting lowest performer and 100 indicating that the target has been achieved

Positive Indicator	Negative Indicator
$x' = \frac{x - \min(x)}{T(x) - \min(x)} \times 100$	$x' = \left[1 - \frac{x - T(x)}{\max(x) - T(x)} \right] \times 100$
x = Raw data value; x'= normalized score after rescaling min(x) = minimum observed value of the indicator in the dataset T(x) = National target value of the indicator	x = Raw data value; x'= normalized score after rescaling T(x) = National target value of the indicator max(x) = maximum observed value of the indicator in the dataset

Construction of Index – Aggregation

Various Methodologies exist. Simple and Weighted Average, Principal Component Analysis, Factor Analysis are some of the popular methodologies.

SDG India Index score is computed for each State/UT as the arithmetic mean of the normalised values of all the Indicators within the Goal. In calculating the average, equal weights are assigned to each indicator and the arithmetic mean is rounded off to the nearest whole number.

Based on the Index, States and UTs are classified into 4 categories under each of the SDGs including SDG 11

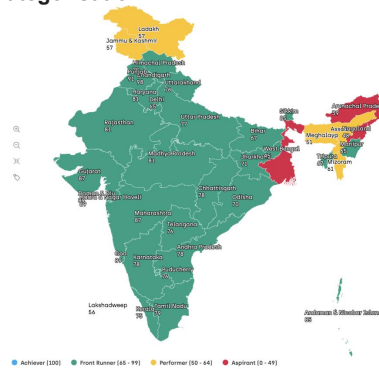
- **Achiever** – when SDG India Index score is equal to 100
- **Front Runner** – when SDG India Index score is less than 100 but greater than or equal to 6
- **Performer** – when SDG India Index score is less than 65 but greater than or equal to 50
- **Aspirant** – when SDG India Index score is less than 50

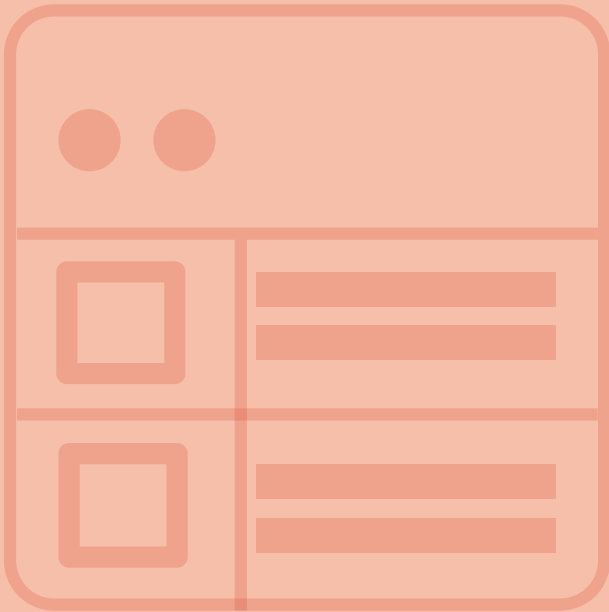
Construction of Index – Ranking and Categorisation

The aspirant states are Arunachal Pradesh, West Bengal and Nagaland

The performing states are Jammu and Kashmir, Ladakh, Assam, Meghalaya, Mizoram and Lakshadweep

The other states are Front runners.





MODULE 1

WORKING WITH TABULAR DATA

Session 3: Basics of Data Visualisation

Duration: (Ideal) 1 hour

Session 3: Basics of Data Visualisation

SESSION SUMMARY

1	BRIEF SESSION DESCRIPTION	This sessions covers the recent history of data visualisation, that grew in Europe during the early modern period. It showcases several examples of how early forms of visualisation were used to make strategic decisions on warfare, public health and mortality. Thereafter, it uses a 'visualisation of visualisations' to outline and clarify what are the different forms of visualisation and what their purposes are. It describes how visualisations can be used to show, 1) comparisons, 2) composition, 3) distributions and 4) relationships between data points and datasets
2	LEARNING OUTCOMES	At the end of the session, participants will be able to use various built- in charts and graphs for data visualisation tools in MS Excel, and participants will be able to create, format colors, legends to improve the readability of the charts
3	CASE STUDIES (IF ANY)	Several small case studies and examples are integrated into the presentation itself
4	FACULTY REQUIREMENT	No particular qualification or experience is required, but faculty should have a good understanding of why and how data is to be visualized; they should have a good command over typical visualisation tools such as MS Excel or Google Sheets (recommended). They should familiarize themselves with the practice datasets beforehand, and be aware of basic principles of aesthetic aspects of colors, icons, lines and fills, etc., as well as overall composition
5	PRACTICE DATASETS	Folder: Day 1- Data sets/ Session 2-3-4 Ref: Sheet7, Sheet8, Sheet9 of DAV Module 1_Practice datasets Access via https://drive.google.com/drive/folders/1NIEngDtIT14akAlQMAsHgXhXt34umSvq?usp=sharing
6	LEARNER PREREQUISITES	Basic understanding of tabular data (expected to be learned from the previous two sessions in the module).
7	CLASSROOM ARRANGEMENT	Traditional Classroom (Refer Annexure 3)
8	TECHNICAL REQUIREMENTS	MS Excel OR Google Sheets (preferred) - participants will have to sign in on their google accounts to use the same

**STATISTICAL TABLE OF
 HINDOOSTAN.**

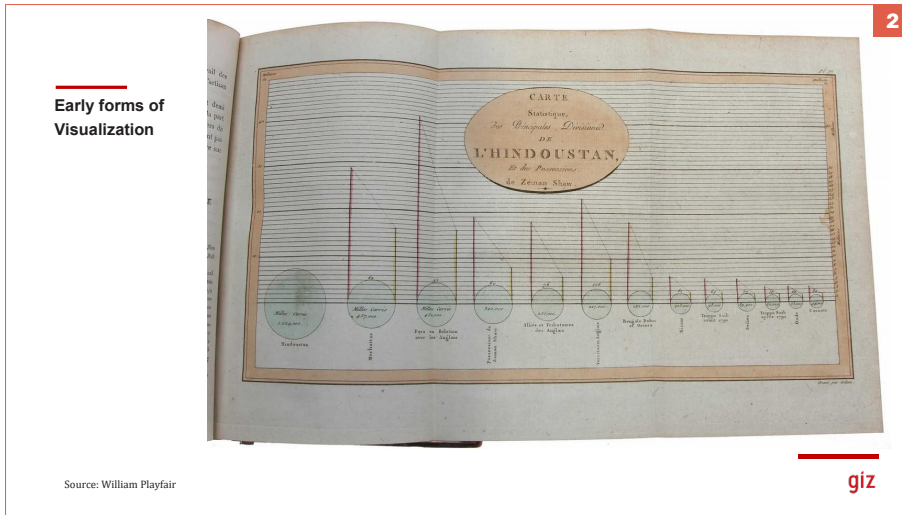
Hindoostan 1801

Extent of Hindoostan in square miles	1,024,800
Number of inhabitants	77,986,818
Number of persons to a square mile, in different provinces	62. 80. 114. 125
Number of English acres	655,872,000
Number of acres to each person about	8½
Revenue in pounds sterling	30,000,000
Commercial exports, about	7,000,000
Imports	3,000,000
Extent of sea-coast in leagues, about	1,200
Peninsula of India in square miles	167,911
Extent of the Merhatta empire in square miles	457,144
British possessions	217,185
British allies	235,497
British interests in India in square miles	452,652
Number of inhabitants in ditto	41,062,890
Revenue of ditto	15,459,000
Nizam's territories	103,690
Revenues in pounds sterling	2,600,000
Military strength, cavalry 40,000 infantry 30,000 70,000	
Dominions of the late Tippoo Sultan before the partition in 1792 in square miles	98,000
Revenue	2,380,000
Dominions of Tippoo after the partition in 1792 in square miles,	62,000
Revenue	1,425,000

Source: William Playfair

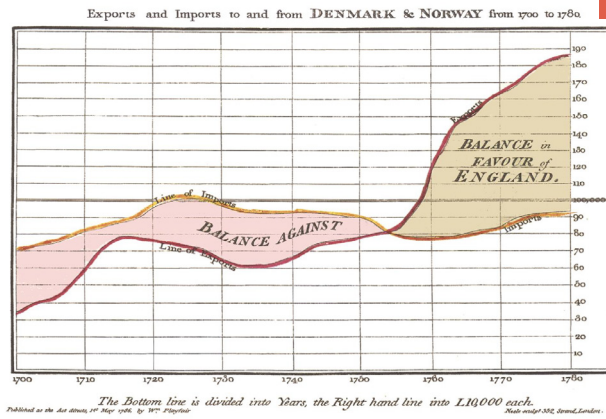
giz

- This slide and the next slide are meant to work together.
- The image on this slide is a statistical summary of some aspects of the area of ‘Hindoostan’ or present-day India, as put together by British interests in the early 19th century. Key points/takeaways from this slide.
 - As can be noticed, there are statistics about the extent of the area of the dominion (all of Hindoostan as well as other kingdoms such as Tipu Sultan, Nizam, ‘Merhatta’ empire, and so on)
 - Population in these dominions, as well as exports, revenues, etc. are also calculated and presented in this table/list form.
 - The trainer can illustrate these key points that are tabulated and move to the next slide.



- The central point of this slide is that a person by the name of William Playfair took the statistical table presented in the previous slide and made what is perhaps one of the first forms of data visualisations that we know of. As can be seen in the image.
- Different 'categories' such as Hindoostan, Merhatta, Nizam, Tipu, etc. have been laid out on the X-axis
- At each point where the categories have been placed, a circle is made to depict the total area of the dominion, thereby giving an immediate view of the physical extent of the particular category/dominion
- While it is not mentioned in the graph itself, the red vertical line on the left of each circle indicates the total population of that dominion and the yellow line indicates its total revenue.
- In this way, a few takeaways emerge:
 - In one glance, a dominion's key statistics can be compared.
 - The lines between the red and yellow vertical lines indicate the number of people per unit of revenue allowing the viewer to broadly estimate the level of wealth of the dominion (this is not strictly true as the gap between the vertical lines also changes due to the size of the bubble, thus affecting the slope of the joining line as well).
 - The chart aids strategic decision-making in terms of geo-political priorities on trade, conquest, alliance, etc.
- However, it may be noted that this form of representation was not considered official or formal at the time, and William Playfair would go on to develop many visualisation ideas, which were not acknowledged as useful by the ruling classes of the time. Perhaps he was somewhat ahead of his time...

Early forms



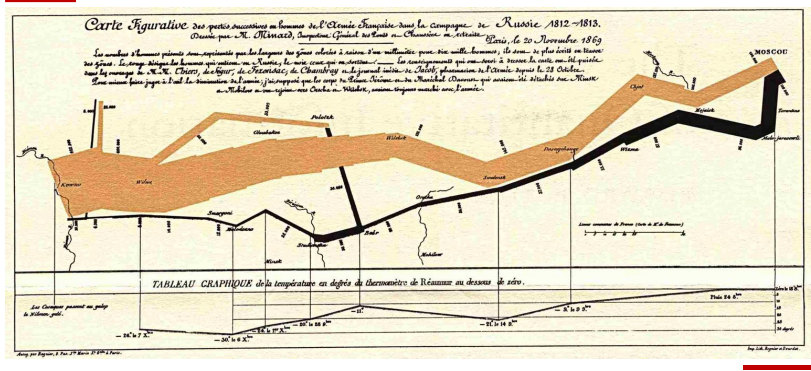
Source: William Playfair

giz

Another chart by William Playfair that would be considered quite useful by our current standards of visualisation. Here:

- The imports from Denmark and Norway to England are noted in the red line over the period of 1700 to 1780; the yellow line shows the exports to the same countries over the same period.
- The resultant area between the lines is the difference between the imports and exports, otherwise called the balance of trade, a very important indicator of a country's financial position - a positive difference is better.
- He has also colored the trade balance portions in red (when negative) and yellow (when positive) which could imply a 'bad' or 'good' scenario. However, this may be speculative as there is no clear evidence that red was associated with bad in that context. In any case, the differentiation between positive and negative is useful for the eye to focus on the point of inflection around 1755.
- Interestingly, William Playfair has made a strong notation on what the axes represent and their units, and also annotated on the lines of actual data, in a way making a proper legend for the chart, so no one can be confused. This was missing in his previous chart we saw.

Early forms

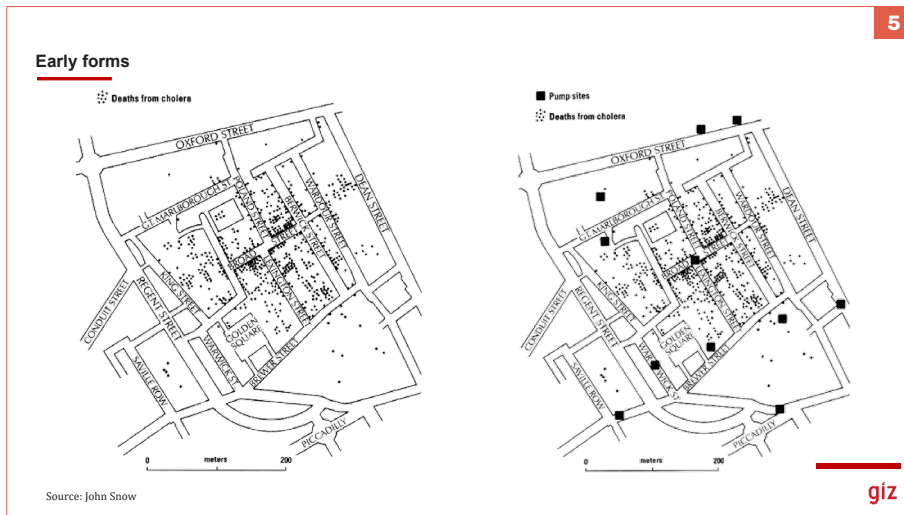


Source: Minard



A truly multivariate visualisation from the early 19th century here, and one of the favorites of many data viz enthusiasts in the current day. This one viz traces the context, action and results of Napoleon’s entire campaign against Russia. A few descriptive points and takeaways from this image:

- The two color bands (yellow for the advancing forces and black for the retreating forces) represent the army of Napoleon. The movement is approximately plotted on the actual geography of the campaign with important landmarks such as locations of cities and main battles, river crossings and rendezvous points, etc., marked out, signifying a spatial dimension to the data. A geographical scale is also presented.
- The width of the bands represent the size of the army at that point, plotted to a certain scale, illustrating quantitative data along with the spatial context. As can be seen, the massive advancing army reduces in size significantly even before reaching Moscow (Moscou). A few contingents separate out from the main force presumably for other battles and also rejoin the retreating forces later. This allows a viewer to ascertain the size of the army at different points at a glance.
- Since battles are only represented when the color changes from yellow to black, the reduction in the size of the advancing army can only be attributed to other reasons. Historically, it is well known that much of Napoleon’s advancing army was in fact decimated by bad weather, treacherous ground conditions and disease.
- This aspect is highlighted in the case of the retreating army. As the main battle in Moscow took place just before winter, the retreating army had to come back to France at the peak of winter. Lines from the retreating army’s band vertically drop down to a temperature graph at the bottom of the viz. As can be seen, severe winter conditions had swept in by the time the army made it back to France, and the deaths resulting from these conditions reduced the size of the army that finally returned to a fraction of what had left. This creative but factual approach to the viz enlivens the data that in textual form would have been much longer, showing how a visualisation can summarize different data dimensions in a single graphic.
- In all, 6 variables have been depicted in this viz: size, location and direction of the army, various geographical points of interest, a timeline and temperature data.



One last example of a historical data viz, that was used to save several lives in a pandemic situation. John Snow was a doctor and not a statistician (but it is interesting to note that now biostatistics is a major and growing field now). About the map and its use:

- During a cholera outbreak in London in the mid-19th century, Dr. John Snow decided to plot the number of deaths in every house on a map of the area. It emerged quite clearly that there was a clustering of deaths in certain streets (map on the left).
- On a hunch, he decided to also include the location of street water pumps (there was no piped supply at the time in London) on the same map (on the right on the slide). A clear relationship seemed to emerge between the location of a particular pump and the number of deaths due to cholera. This led to the closure of the pump by the authorities and there was an immediate reduction in the number of deaths.
- As the relationship between water and cholera was established, it was discovered from further research that cholera was in fact transmitted through water.

What is Data Visualization?

- Visual communication of data
About 25% of content in a report will be graphs and charts
- 3 Principles

Complexity → **Simplicity** Content and information don't care what they are

User ↔ **Viewer** Quality, relevance and integrity

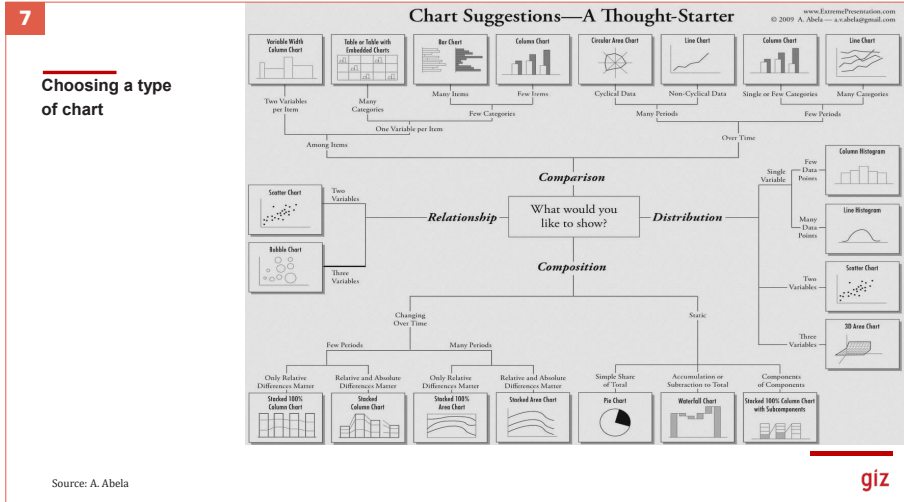
Approximately right rather than perfectly wrong Empirical-statistical, mathematical and visual-artistic

Source: Edward Tufte

giz

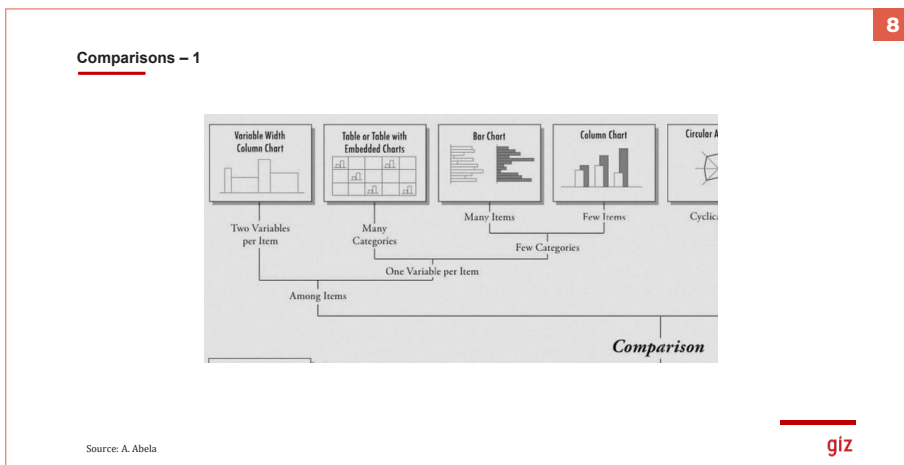
Edward Tufte is largely recognized as one of the pioneers and leading minds on data viz in the modern world. His practical and theoretical work on data analytics and viz is considered groundbreaking and yet easily accessible for learners. Here a few points that he mentions are summarized. Some notes on how to think about them:

- First of all, it is estimated that about a quarter of the content of all reports and publications will be graphs and other visualisations, making familiarity with this skill a key to professional growth.
- Edward Tufte outlines 3 principles of data visualisation. The first is that complex data must be represented simply. In this statement, there is a principle that content or data are agnostic to their representation method or approach - they will only say the same thing whether they are reported in words, visuals or any other medium - the medium doesn't change the message, in other words. Practically, this means that creators of visualisations should try out many different methods to see the data and bring its message to their viewers.
- The second principle is that the creator of a visualisation must be able to allow the user to switch to becoming a viewer. The difference between the two is simply one who has a direct use case for the data, almost like an automated function or algorithm, and another who can see the data from different perspectives and draw conclusions in many different ways. While the creator allows users to become viewers (with broader aims), they must maintain the quality, relevance and integrity of the data in question through their viz. In other words, when creating visualisations, creators must be able to communicate straightforward messages (as in point 1) but also allow viewers to interact more richly with the data.
- The last principle is difficult to communicate but can be simply said to be the point of view of an experienced expert such as Tufte. His view is that from a statistical approach, a creator has to shift to, or at least include in their method a creative angle. When including a creative angle, we cannot compromise on the two principles above, but rather, aim to not avoid the complexity of the data. This means that even if some of the data may be confounding or 'not adding up to a story', it must be presented in an accessible way, and not hidden away or discarded (for eg, in the case of outliers in data). Thus the story may end up being approximately right, that is, complex but tending toward some meaning, rather than truncated at the point where it fails to make sense (which would be completely wrong).



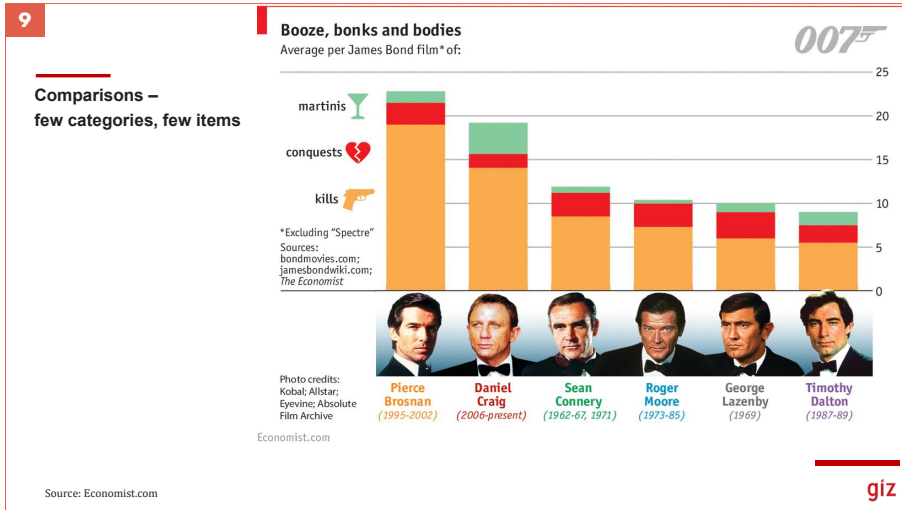
Now having covered some historical background and cases, we move to the formal part of the session. This slide is simple to understand and explain:

- Through visualisation of tabular data, you can show four (somewhat overlapping) things: a comparison between categories and items, a relationship between two or more variables, the composition of a dataset (what is it made up of) and the distribution of variables in a dataset (usually used for larger datasets).
- This ‘visualisation of visualisations’ by Abela is a very useful summary of these representations. The next sections will delve into each of these types of representations in some detail with examples.



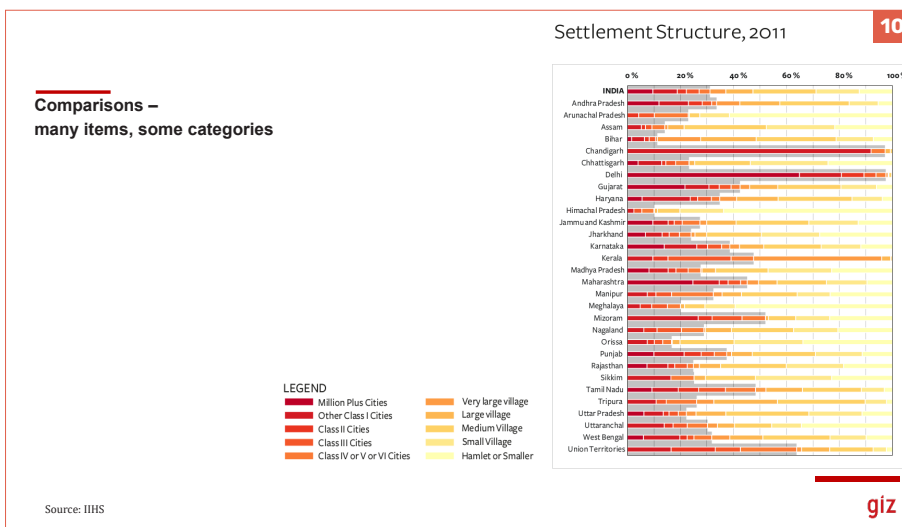
First, we look at the broadest use of visualisation - comparing datasets. Within this, we will first look at comparisons between items and categories. A brief note on how to define category and item (for the purpose of this session):

- For ease of understanding, it can be said that categories are higher level of data classification than ‘items’. In most charts, including the ones you are seeing in the diagram above, the categories are usually on the X-axis and represent the most important things that the creator wants the viewer to compare.
- Items are usually on the Y-axis and represent the questions or inquiries that we want to conduct across the categories. So, for example, if we want to compare level of urbanization in 5 select states of India in 2001 and 2011, we will make a column chart with the states in the X-axis and the percentage of urbanization on the Y-axis and two columns per state representing the percentage values of urbanization in that state.



This may look like a trivial graph but serves to lightly explain the concepts of categories, items and also the concept of stacking.

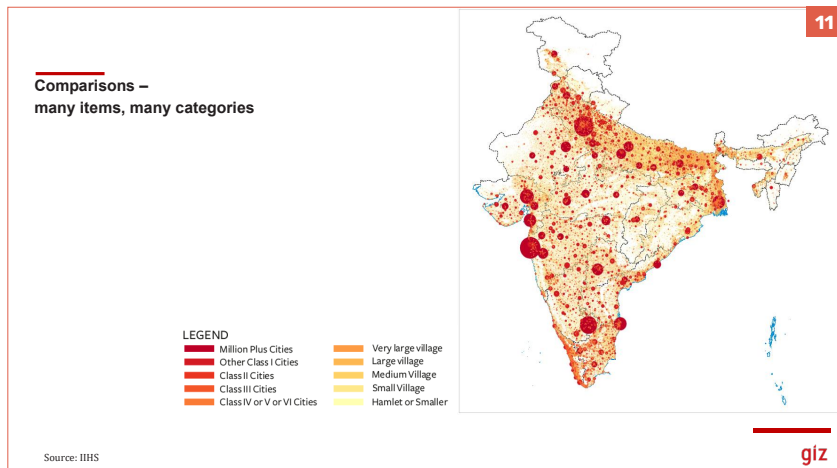
- The X-axis has the categories that we want to compare - in this case, the different actors who have played James Bond in the film series.
- On the Y-axis are the items - in this case, kills, romantic conquests and martini drinks. The items are stacked for each category because they are small quantities in the same unit of numbers.
- The design and presentation of this chart is also worth noting - how photos and icons have been used along with bold colors.



In practical terms, with the kind of data we use in this sector, there are probably going to be many categories and items. In such a case (and it is discretionary what is the limit) a bar chart may be more suitable. In a bar chart we are comparing vertically rather than horizontally.

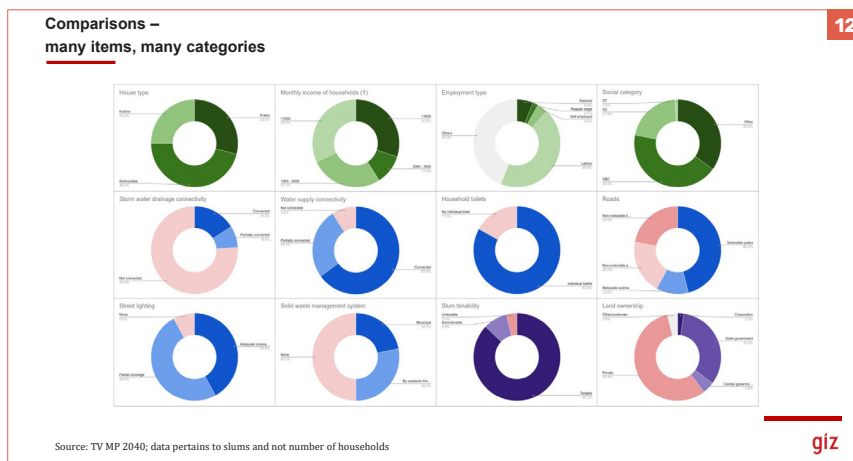
- Here the states of India are the categories as shown in the Y-axis and the type of settlement (classified as per size by the Census) are in the X-axis, represented by different colors.
- One thing to note is that the colors here are along a spectrum of a single color (in this case, yellow to red). This kind of a spectrum is called a monochromatic spectrum, and is useful for several reasons:
 - Too many different colors distract the eye.
 - Emphasis on certain items can be laid by using darker shades of the colors.
 - This will work well even in a black and white situation, such as when it is printed on a b/w printer.
- An interesting dimension has been added to the viz by bounding colors (items) up to a certain point in grey - which represents urban settlements.





Here the categories are all of India's settlements and the items are their populations and Census classification. However, since there are more than 6 lakh settlements in the country, it is not possible to put them on a column or bar chart. So, the visualisation has been done on a map. In many ways, map platforms are the best location to visualize large or 'big' datasets.

- Now each settlement is identified by its geographical coordinates.
- The size of the circle at these coordinates is reflective of the population of that settlement.
- The color of the circle represents the Census classification of that settlement.

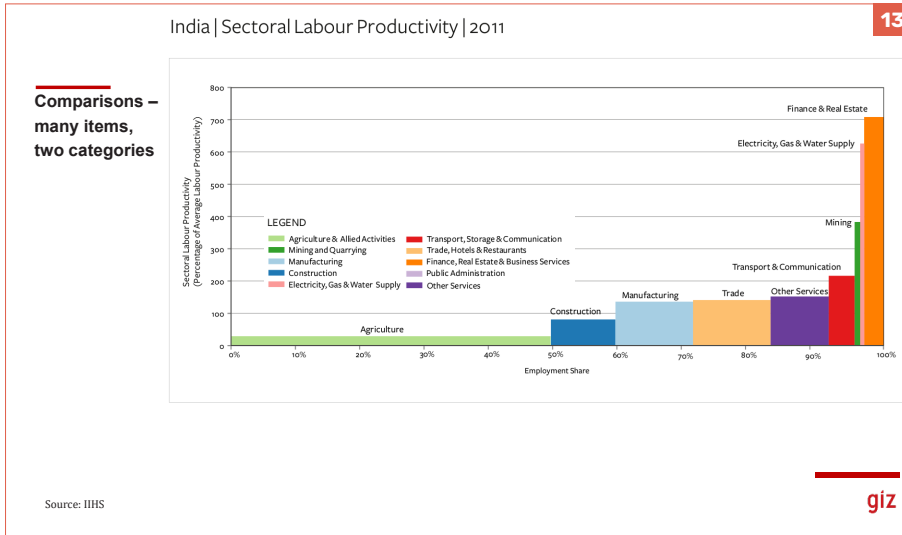


Another way to visualize data that has many categories and many items is to use an array of charts. Pi charts or bar charts or column charts can be composed graphically as shown in the example on this slide. This is data on slums identified in Thiruvananthapuram, Kerala for the purpose of making a master plan for the city.

- The creator must ensure there is some relation between various charts in order to aid and encourage comparison. For example here, the green charts are largely talking about households and housing, the blue-pink charts are talking of basic services, and the pink-purple charts are talking of institutional questions. For a viewer, it can be expected that they will find some patterns in the data when they are able to easily discern some higher level relation between different datasets such as these. So, someone working on social sciences may focus on the green charts, engineers may focus on the next set, and institutional actors may focus on the last set.
- An important limitation of such data is that it pertains to slums as distinct categories but not on the number of households within them. If some of the large slums have no water supply, for instance, this kind of data and viz will not give a correct picture about the actual need for investment in water supply. Can the class explain why?

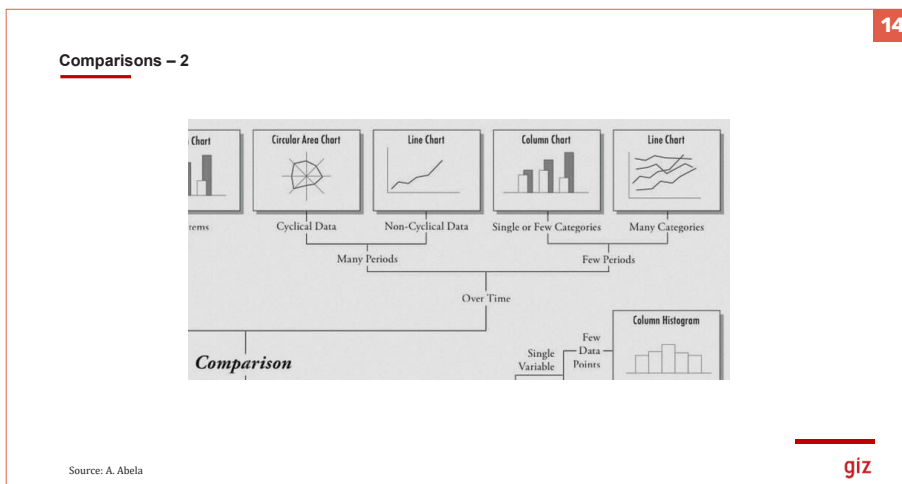


Pie chart

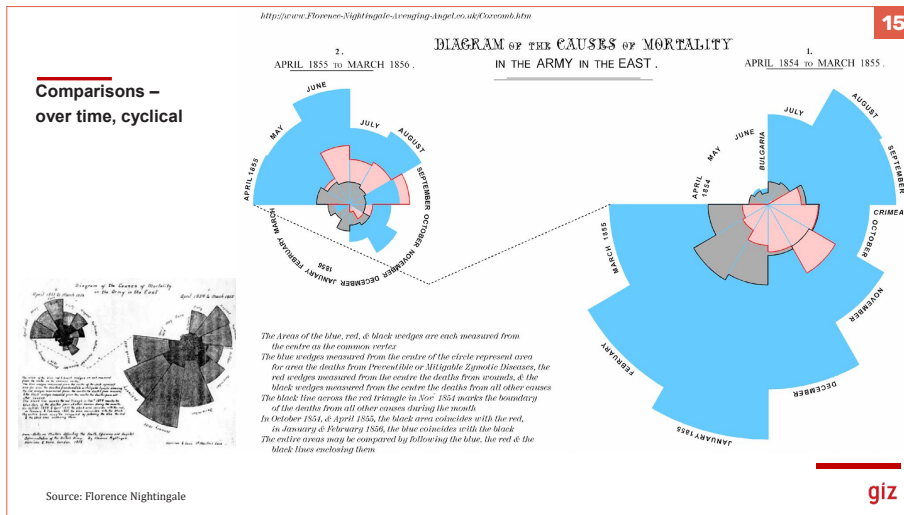


Usually the width of the column in a column chart or the thickness of the bar in a bar chart have no relevance. This is a slightly different chart in which each category itself has data relevant on both axes - in this case, the employment share and the labour productivity.

- In this case, the X-axis denotes one category, that is, employment share. The total adds up to 100% but could also be a numeral actual value.
- The Y-axis represents another category, that is, labour productivity as a ratio of input (investment) vs output (production).
- The data itself is about different sectors of the economy. The chart therefore gives immediate insight into the productivity vs employment share of these sectors. It is easy to ascertain that while agriculture employs nearly 50% of the working population (X-axis), its productivity is much less than 100%. Even construction has productivity less than 100%. Of the sectors that have high productivity, finance and real estate top the list but employ barely 5% of the workforce.
- This kind of chart takes some time to understand so will be best positioned as a high level summary and as a stand alone feature on a presentation or report. The distinct colors to the data add to the readability of the graphic.

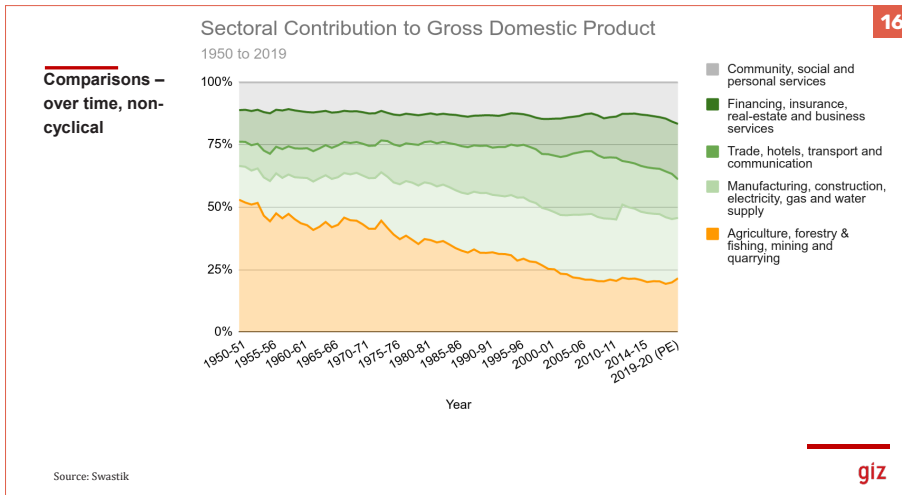


Now we move to showing comparisons over time. Time itself can be seen as cyclical (hours of a day, seasons, etc.) or non-cyclical (years or months).



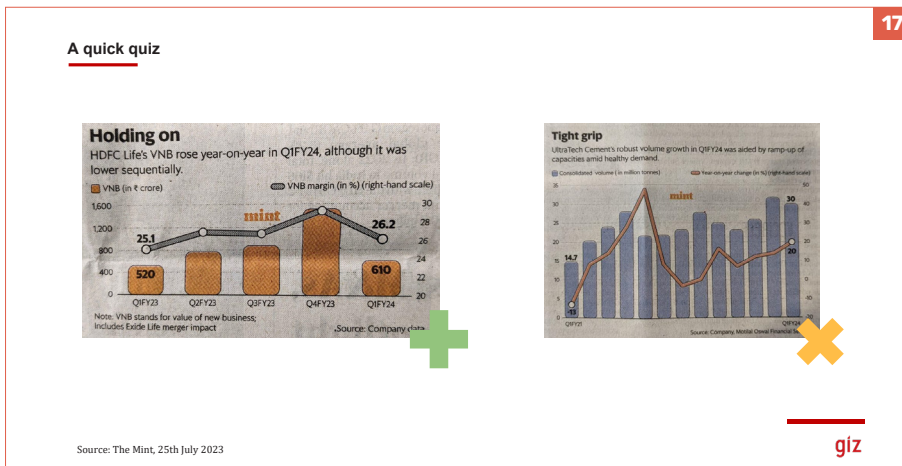
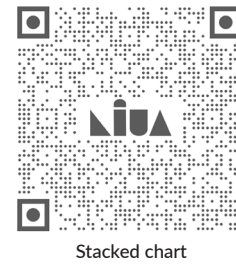
Florence Nightingale, the person responsible for giving the profession of nursing a substantive standing in the field of clinical medicine, was also a statistician. In the mid-19th century, when England was waging a war in Crimea, she collected data on deaths of soldiers with a view to prevent them. She represented her data to decision makers in England and action was taken based on her reports. The chart on the left is the original made by Ms. Nightingale and on the right is a replica made for better clarity. These show:

- Over the cycle of a year, the number of deaths and injuries among soldiers fighting in this war, in the form of a circular chart that has 'sectors' the depth of which denote the actual numerical data. Red denotes the deaths due to war injuries, black the deaths due to other causes and blue the deaths due to disease.
- As she was able to show, the deaths from diseases (in this case, mitigable zygotic diseases) far outweigh deaths due to other causes, including war injuries for most of the year. These are also most pronounced during the winter months.
- This proved that if the decision makers in England wanted to save more soldiers from death, they would need to focus on disease and not war. In a way, she was able to change the war effort to send more doctors and nurses rather than soldiers.



Non-cyclical time series data is commonly available and easy to use. The chart in this slide shows:

- The relative share of different sectors of the Indian economy to the overall GDP as a percentage, changing over time (in this case, 1950-51 to 2019-20).
- A clear pattern emerges in the shift in the economy, that is further highlighted by applying a different color range (greens) to sectors that are largely urban in nature, as compared to sectors that are rural in nature (orange). Government contribution to the economy (categorized as community, social and personal services) are colored grey to reduce the emphasis on them.
- So, the key takeaway of the chart is to show that the urban sectors have started to dominate the economy.



'Column and line charts' are often used to show numerical and percentage change data simultaneously in one chart. Popular publications often use such charts, and the slide here shows one such set from The Mint. At this stage, this slide should be used to test the participants' learning a little bit, because one of the charts is factually incorrect. The question is, which one?

- It is the chart on the right that is incorrect. Logically, the year a particular variable drops in value, its percentage growth line should slope down. But we can see this logical rule is violated in the chart on the right.

18

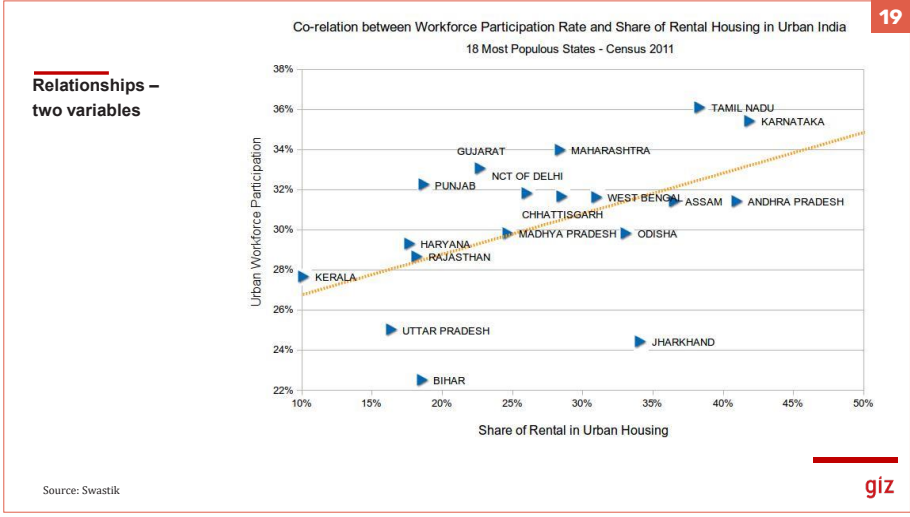
Relationships

The diagram illustrates the relationship between different types of charts and the number of variables they represent. A 'Scatter Chart' is associated with 'Two Variables', and a 'Bubble Chart' is associated with 'Three Variables'. Both are connected to a central box labeled 'Relationship'.

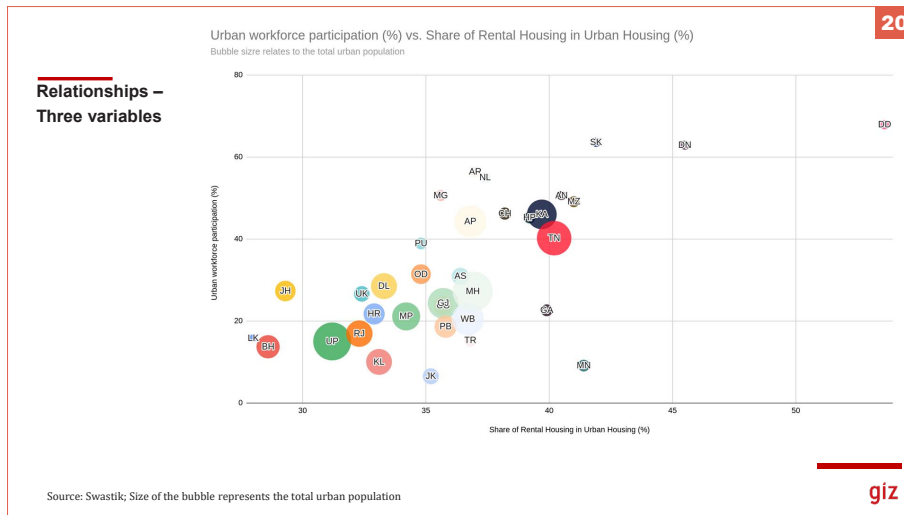
Source: A. Abela

giz

- We now show how charts can be used to show relationships between 2 or 3 variables.

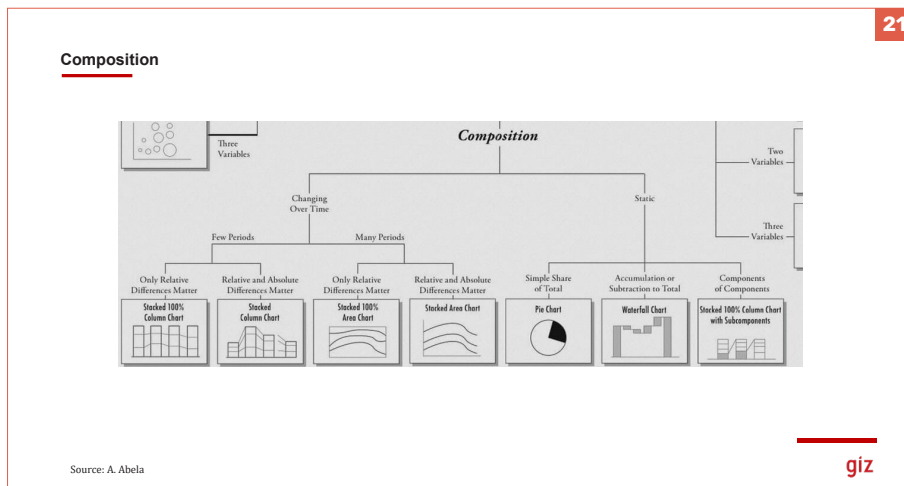


Advanced statistical analysis includes finding the 'correlation' between two variables. A high correlation implies a greater degree of association between two variables, implying that if one of the variables was changed or modified through some action, the other would also likely change to some degree. This phenomenon can also be visualized graphically using a scatter plot between these two variables. In the image on this slide, such a relationship is explored between the two variables - share of rental housing and workforce participation rate (both in percentage). As can be seen, there appears to be a correlation between the two variables, as signified by the slope of the regression line. A regression line of 45 degrees slope implies a high correlation, given the frame of the chart is a square.



This chart is nearly the same as the previous one, but with more updated data. There is also a third variable added, that is, the level of urbanization of the population of the state/UT. This last variable is visualized using the size of the bubble. In this way, up to 3 variables can easily be visualized on one chart. There is the possibility of adding the data of one more variable, even though it is not done in this chart. For example, if the creator wanted to show the broad regions of the country to which these states/UTs belong (South, North-east, North, West, etc.) the colour of the bubble could be used. So, all the bubbles from the North could have one color, and so on.

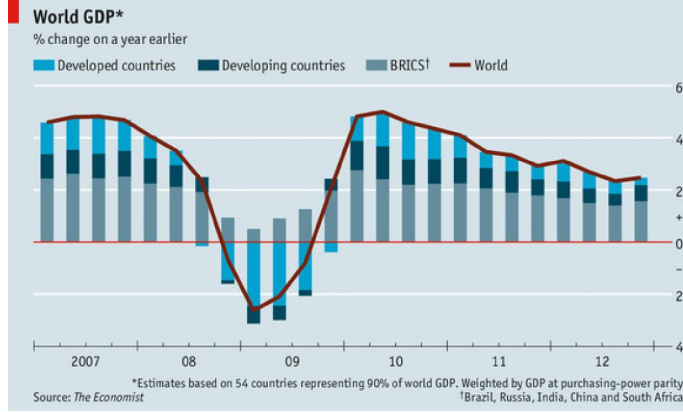
It must again be mentioned that charts that have a lot of variables take time to understand, even by advanced viewers, and so such visualisations must be used sparingly and strategically.



Now we move to a section on how to use visualisations to understand composition of categories and variables.

22

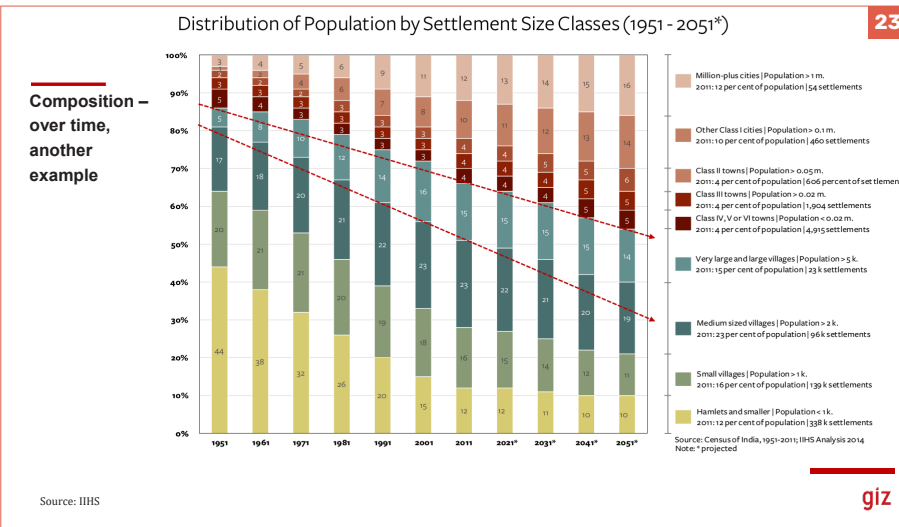
Composition – over time, absolute difference matters



Source: Economist.com

giz

Static, or one-time composition is easy to show on a pi chart or a stacked column. The latter approach can also be extended to show changing composition over time, such as in this chart on this slide. The data is simple: between 2007 and 2012 what was the contribution of various categories of countries to the growth (in percentage) of the world's GDP? What is interesting about this chart is that it captures data around the 2008-09 global financial crisis which pulled the world's GDP growth down into negative numbers (as depicted by the red line). This negative growth was 'pulled down' by the category of developed countries and to some extent by the developing countries, and 'pulled up' by the BRICS nations. In this chart, the red line represents only the total of the stacked columns, and so may seem redundant. However, in the negative growth period, it represents the net growth which is not easily discernible using the columns - and therefore serves a particular purpose for this viz.

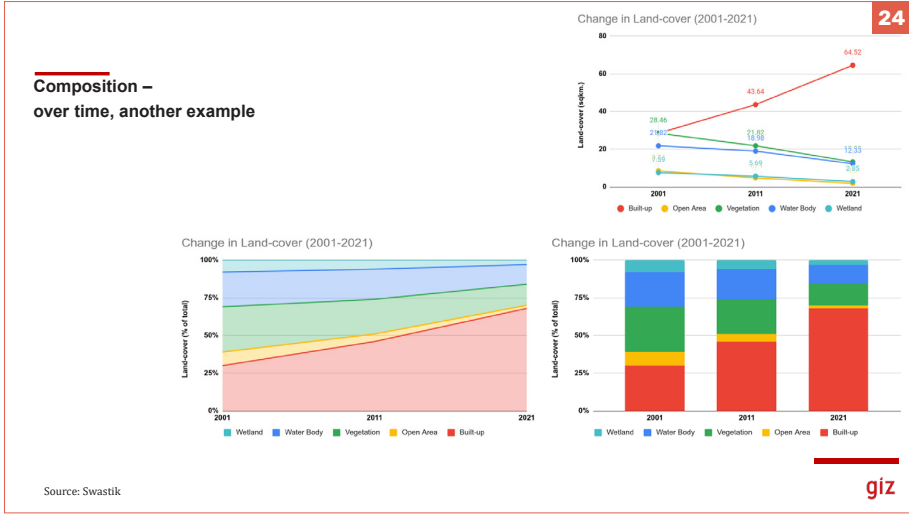


Source: IHHS

giz

A stacked column chart over time is a useful tool to explain changes in the values/properties of various categories. The chart on this slide shows how there has been a substantial increase in settlements in India categorized as urban (pink-maroon colours), including changes projected till 2051. It highlights, using bright red lines, settlements categorized by the Census as very large and large villages, perhaps to draw attention to the fact that these kinds of settlements may already be considered urban. In this, the creator of this chart has taken a creative approach, superimposing their analysis on factual data.

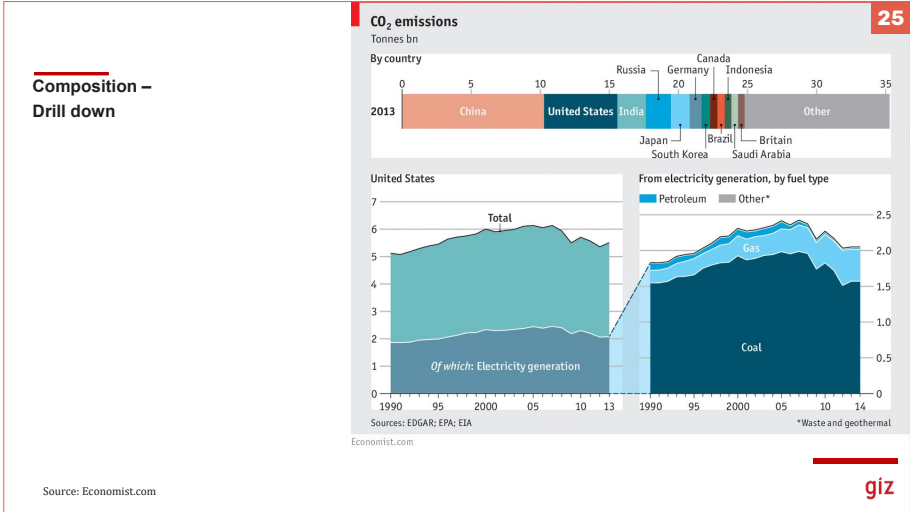
**Composition –
over time, another example**



A simple slide to illustrate and summarize 3 different ways to show change over time.

- Points and lines of actual numerical data
- Stacked composition using percentages
- Stacked columns using percentages

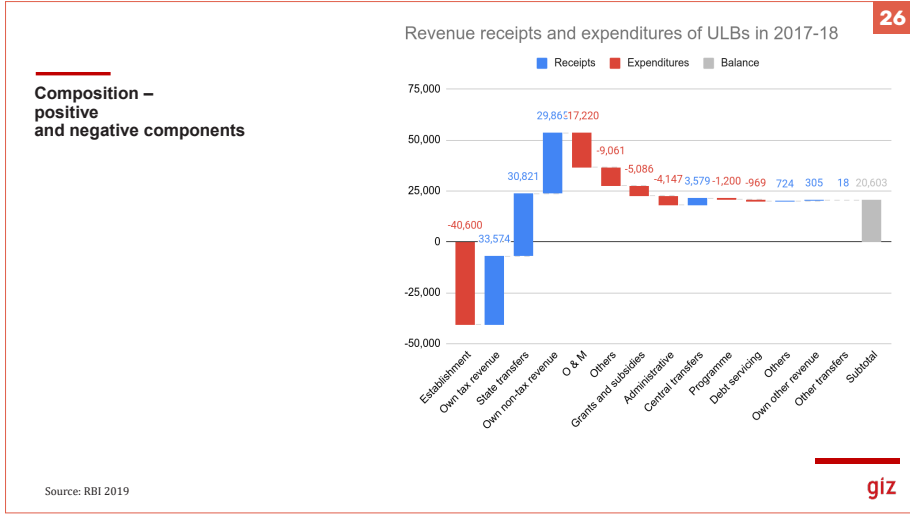
Which one do the participants find more compelling? Why?



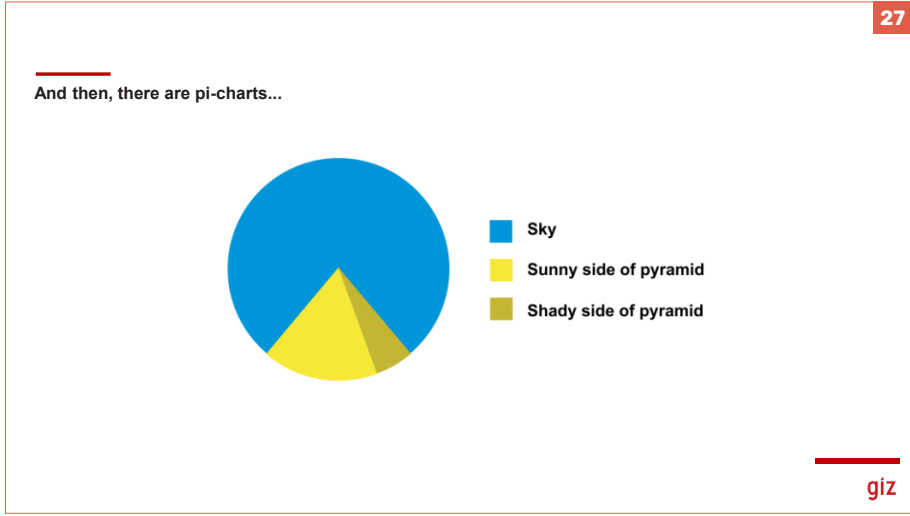
Visualisation, especially ones hosted on servers and accessible on digital devices, can help drill down into data. On a digital device, a viewer could possibly tap or click to find out more about a particular category or item. It is not so easy to replicate on a static medium such as a report or presentation. However, this chart on this slide gives us some clues on how it could be done. It shows us:

- The total carbon dioxide emissions in billion tonnes in 2013 by country in the bar chart on top. We can easily see that the US contributes about 5 bn T out of the total of 35.
- The charts at the bottom break this down further, and over time.
 - On the left, we see how this total 5 bn T in 2013 was reached, and the chart highlights the role of electricity generation in this change.
 - On the right, we can see that of the total emission due to electricity generation, a vast majority was due to coal fired power generation.

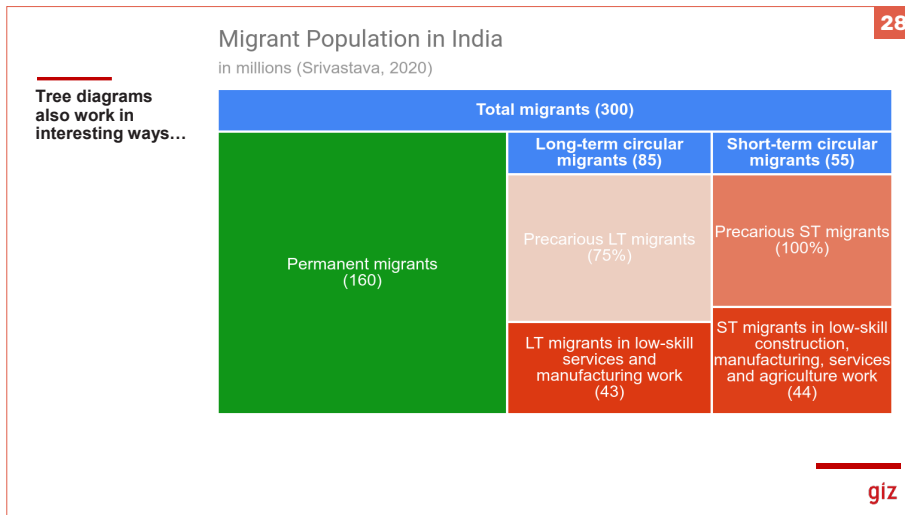
Thus, the chart leads us into details that would be interesting and useful. In the end, since all the units remain the same, we can safely conclude that in 2013-14 about 1.5 of 35 billion tonnes of carbon dioxide released in the atmosphere was due to coal-power generation of electricity in the US. The first bar chart tells us that this is comparable to the total carbon emission of India, Russia, Japan, etc.



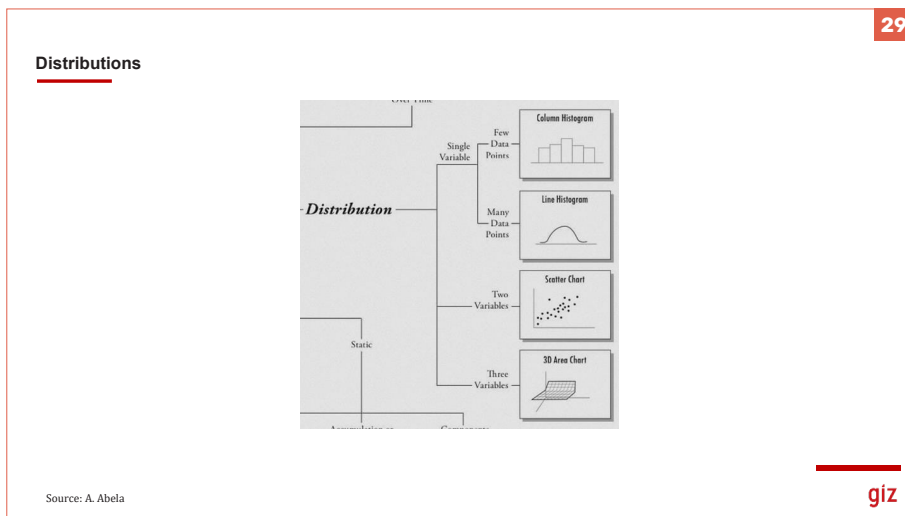
Composition can have positive and negative values. For example, the final balance in a finance book is a combination of credits and debits. This chart shows how to use the 'waterfall chart' option available in various popular visualisation software to generate a chart on what are the positive and negative contributions to a final balance. In the typical template available on popular platforms like Google Sheets or MS Office, negative components are usually shown in red/orange and positive components are shown in blue/green, and the final balance may be shown in grey.



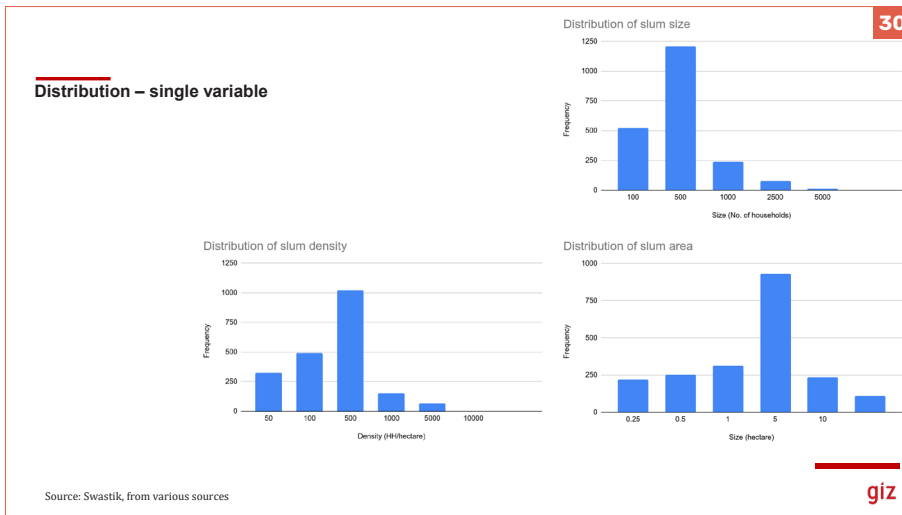
This is a light-hearted humorous slide. The purpose of this slide is to lighten the atmosphere of the class after the technical session. It is assumed that most, if not all of the participants know, understand are already aware on how to make simple pi charts to show composition of a dataset.



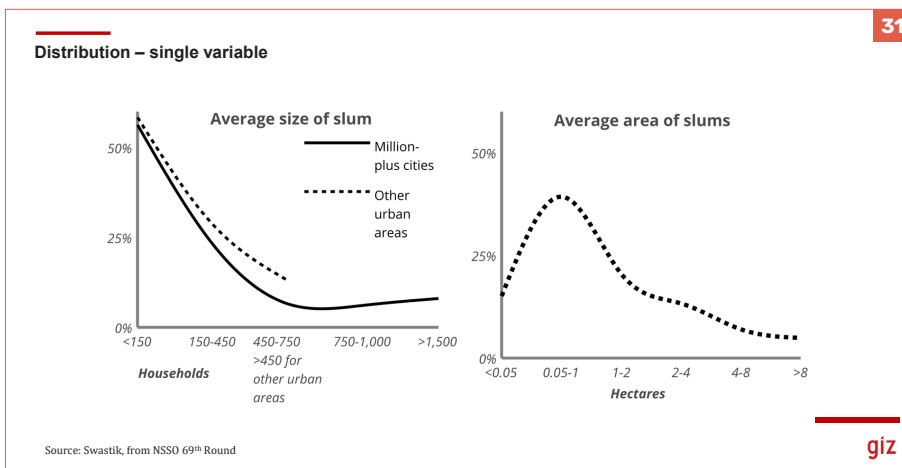
A Tree Diagram illustrates composition of datasets, within sections and/or hierarchies. It is available now as an option in software such as Google Sheets. The advantage of a tree diagram is that it can show relative compositions of connected datasets. In the diagram on this slide, for example, the total number of migrants in India is estimated to be 300 million (top blue bar), of which 160 million are permanent migrants (left green bar), 85 and 55 million are long-term circular migrants and short-term circular migrants, respectively. In the last two categories, further disaggregation is shown in the red bars.



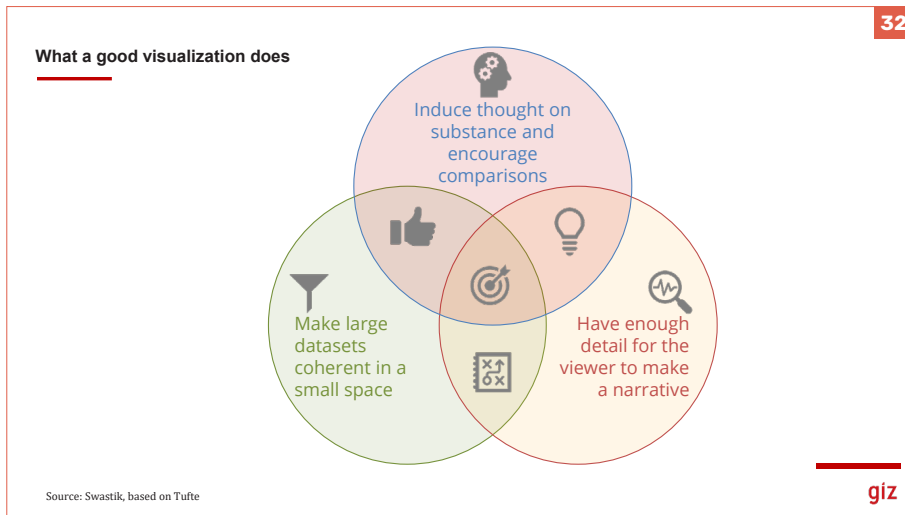
We now move to the last category of visualisations - showing distributions. Please note we have already shown previously how to show distributions between 2 and 3 variables on bubble charts or scatter plots. Now we will learn how to show distributions in column charts or histograms.



Distributions can be shown using continuous variables or discrete categories. When we have large datasets, of thousands of values, distributions are a very useful way to quickly summarize them. For example, suppose we have data of all the slums in Indian cities, and this data consists of their size (number of households) and area (in hectares). One way to show this distribution of any of these variables would be to create a column chart with all the slums on the X-axis and the value of the data (size or area) on the Y-axis. We would then achieve a histogram of the distribution. However, with large data, this approach may limit the human readable understanding of the data, because there are simply too many data points, as each slum would have a different value for each of the variables. In order to simplify the data, the creator may choose to classify the variables into some discrete ‘buckets’ of values, as is shown in the graphs on this slide. After classifying the values into distinct and discrete buckets (size value <100 households, between 100 and 500, from 500 to 1000, and so on) the viewer is able to make meaningful insights from the data. It becomes quickly clear from the visualisation that the majority of slums have a size of between 100 and 500 households, and area between 1 and 5 hectares and a resultant density of between 100 to 500 households per hectare. These kind of distributions help create probability curves for further analysis.



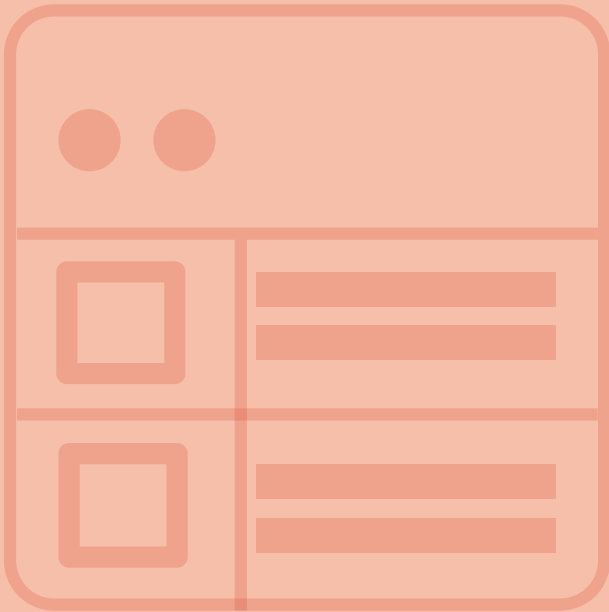
In the graph above on the left, the distribution is further disaggregated into data from million plus cities and other urban areas, thus allowing a viewer to see the distribution of slum size across two different size-class of cities simultaneously. In this case, the distribution has been shown as a continuous line in order to clearly communicate the trends in the distribution. So, it is obvious for the viewer that (in the case of the area of slums, for example) there is a decreasing trend in slum areas, implying that there are far fewer slums of large areas, and the highest number of slums (around 40%) are between 0.05 and 1 hectare in area.



This animated slide summarizes the actions and goals of good data visualisation. The following describes each of the components:

- The pink circle with the mind working icon conveys the message that a good visualisation must induce thought on substance and encourage comparisons in the data.
- The green circle with the filter icon indicates that the visualisation must be able to make large datasets accessible and coherent.
- The yellow circle with the lens icon illustrates that the visualisation should have adequate details so that a viewer can make a narrative story in their minds.
- When the visualisation achieves the narrative and the substance goals, it leads to good ideas, as denoted by the bulb icon.
- When the visualisation achieves the substance and coherence goals, it leads to better decision making
- When the visualisation achieves the coherence and narrative goals, it allows viewers to navigate through the data confidently.
- It is only when you have achieved all 3 goals that you have hit the target.

Notes



MODULE 1

WORKING WITH TABULAR DATA

Session 4: Integrating and Summarizing Tabular Datasets

Duration: (Ideal) 1 hour

Session 4: Integrating and Summarising Tabular Datasets

SESSION SUMMARY

1	BRIEF SESSION DESCRIPTION	This session covers topics such as need for data integration, various data integration methods and how to integrate data in excel using vertical look up (VLookUp) function. Data Summarisation using pivot tables is demonstrated in excel.
2	LEARNING OUTCOMES	At the end of the session, participants will be able to identify and perform suitable data integration techniques to combine data from different sources into one dataset using excel.
3	CASE STUDIES (IF ANY)	Several small case studies and examples are integrated into the presentation itself
4	PRACTICE DATASETS	Folder: Day 1- Data sets/ Session 2-3-4 Ref: Sheet10, Sheet11 of DAV Module 1_Practice datasets Access via https://drive.google.com/drive/folders/1NIEnGDtiT14akAIQMAsHgXhXt34umSvq?usp=sharing
5	FACULTY REQUIREMENT	Basic Understanding of MS Excel. Knowledge about reading excel or csv datasets
6	LEARNER PREREQUISITES	None
7	CLASSROOM ARRANGEMENT	Traditional Classroom (Refer Annexure 3)
8	TECHNICAL REQUIREMENTS	NA

1

Data Integration - Merging Multiple Datasets

Merging data from different sources into a single dataset is necessary for a comprehensive analysis

- Data required for analysis usually resides in multiple data sources
- Each source may only have small chunks of data or only on some features

Horizontal Merge

- The data from multiple data sets with identical columns is combined into a single dataset.

Vertical Merge

- The data from multiple data sources is combined into a single dataset based on a common field or attribute. This is commonly referred as 'Join'

- The data required for analysis seldom exists in one data source. Sometimes, small chunks of data on the same features is spread across multiple sources. Similarly, each source may contain all the required data but only for some features. The data for other features maybe in some other dataset or datasets. Hence merging data from difference sources into a single dataset is necessary for a comprehensive analysis.
- For example, integrating data on population density, demographics, and migration patterns helps in understanding the growth of urban areas. This information is crucial for planning housing, healthcare, and educational facilities to meet the needs of a growing population. The required data is spread across difference departments. Hence, it is important to bring the data together.
- The data merging maybe horizontal or vertical.
- In Horizontal merge, the data from multiple data sets with identical columns is combined into a single dataset.
- In Vertical Merge, the data from multiple data sources is combined into a single dataset based on a common field or attribute. This is commonly referred as 'Join'.

2

Common Data Joining Techniques

State Code	District Code	Sub District Code	Town-Village Code	Town-Village Name	State Code	Town Code	Town Name	Population
32	588	05631	803256	Kanhangad (M + OG)	32	803256	Kanhangad (M+OG)	1,25,564
32	591	05640	803267	Kozhikode (M Corp. + OG)	32	803267	Kozhikode (M Corp.+OG)	5,50,440
32	592	05641	803269	Malappuram (M + OG)	32	803269	Malappuram (M+OG)	1,01,386
32	594	05654	803280	Thrissur (M Corp.)	32	803275	Palakkad (M)	1,30,955
32	595	05659	803288	Kochi (M Corp. + OG) (Part)	32	803280	Thrissur (M Corp.)	3,15,957
32	598	05674	803299	Alappuzha (M + OG)	32	803288	Kochi (M Corp.+OG)	6,33,553
32	600	00000	803306	Kollam (M Corp. + OG)	32	803299	Alappuzha (M+OG)	2,40,991
32	601	05689	803310	Attingal (M)	32	803306	Kollam (M Corp.+OG)	3,67,107
32	601	00000	803312	Thiruvananthapuram (M Corp. + OG)	32	803312	Thiruvananthapuram (M Corp.+OG)	7,88,271
32	601	05692	803313	Neyyattinkara (M)				

↓

Dataset 1 : Towns/villages directories and geo codes

↓

Dataset 2: Towns/villages Population

To show the vertical merge or join, two datasets are taken as an example. The first datasets contains, the details about the town such as state, district, sub-district it resides in and the second dataset contains the population of each town. Using these two datasets, difference data joining techniques such as inner join, outer join, left join and right join are shown in the subsequent slides.

3

Common Data Joining Techniques – (Inner Join)

An inner join returns only the rows that have matching values in both datasets. It combines the rows from both datasets based on the common field. Rows that do not have a match in the other dataset are excluded from the result.

Inner join on dataset 1 and dataset 2 with common field (State Code + Town/Village Code) results in the following

State Code	District Code	Sub District Code	Town-Village Code	Town-Village Name	Population
32	588	05631	803256	Kanhangad (M + OG)	1,25,564
32	591	05640	803267	Kozhikode (M Corp. + OG)	5,50,440
32	592	05641	803269	Malappuram (M + OG)	1,01,386
32	594	05654	803280	Thrissur (M Corp.)	3,15,957
32	595	05659	803288	Kochi (M Corp. + OG) (Part)	6,33,553
32	598	05674	803299	Alappuzha (M + OG)	2,40,991
32	600	00000	803306	Kollam (M Corp. + OG)	3,67,107
32	601	00000	803312	Thiruvananthapuram (M Corp. + OG)	7,88,271

giz

- An inner join returns only the rows that have matching values in both datasets. It combines the rows from both datasets based on the common field.
- In the given example, (State Code + Town/Village Code) are the common field between two datasets. Rows that do not have a match in the other dataset are excluded from the result.
- The fields (32, 803310) and (32, 803313) in dataset 1 do not have a matching field in dataset 2. Hence they are excluded. Similarly, some records in dataset 2 also do not have a matching record in dataset 1. Hence they are also excluded.

4

Common Data Joining Techniques – (Outer Join)

An outer join combines all the rows from both datasets. It includes the matching rows as well as the non-matching rows from both datasets.

Outer join on dataset 1 and dataset 2 with common field (State Code + Town/Village Code) results in the following

State Code	District Code	Sub District Code	Town-Village Code	Town-Village Name	Population
32	588	05631	803256	Kanhangad (M + OG)	1,25,564
32	591	05640	803267	Kozhikode (M Corp. + OG)	5,50,440
32	592	05641	803269	Malappuram (M + OG)	1,01,386
32	594	05654	803280	Thrissur (M Corp.)	3,15,957
32	595	05659	803288	Kochi (M Corp. + OG) (Part)	6,33,553
32	598	05674	803299	Alappuzha (M + OG)	2,40,991
32	600	00000	803306	Kollam (M Corp. + OG)	3,67,107
32	601	05689	803310	Attingal (M)	
32	601	00000	803312	Thiruvananthapuram (M Corp. + OG)	7,88,271
32	601	05692	803313	Neyyattinkara (M)	
32			803275	Palakkad (M)	1,30,955

giz

- An outer join combines all the rows from both datasets. It includes the matching rows as well as the non-matching rows from both datasets.
- The records for the fields (32, 803310) and (32, 803313) in dataset 1 are included after the merge even though there is no matching record in dataset 2.

Common Data Joining Techniques – (Left Join)

A left join returns all the rows from the left dataset and the matching rows from the right dataset.

Left join on dataset 1 and dataset 2 with common field (State Code + Town/Village Code) results in the following

State Code	District Code	Sub District Code	Town-Village Code	Town-Village Name	Population
32	588	05631	803256	Kanhangad (M + OG)	1,25,564
32	591	05640	803267	Kozhikode (M Corp. + OG)	5,50,440
32	592	05641	803269	Malappuram (M + OG)	1,01,386
32	594	05654	803280	Thrissur (M Corp.)	3,15,957
32	595	05659	803288	Kochi (M Corp. + OG) (Part)	6,33,553
32	598	05674	803299	Alappuzha (M + OG)	2,40,991
32	600	00000	803306	Kollam (M Corp. + OG)	3,67,107
32	601	05689	803310	Attینگال (M)	
32	601	00000	803312	Thiruvananthapuram (M Corp. + OG)	7,88,271
32	601	05692	803313	Neyyattinkara (M)	

giz

- A left join returns all the rows from the left dataset and the matching rows from the right dataset.
- The records for the fields (32, 803310) and (32, 803313) are in dataset 1. Hence, they are included in the merged dataset, though there are no corresponding entries for the field in dataset 2.
- The field (32, 803275) in dataset 2 does not have a matching record in dataset 1. After left join, this record is not included in the merged dataset.

Common Data Joining Techniques – (Right Join)

A right join is similar to a left join, but it returns all the rows from the right dataset and the matching rows from the left dataset.

Right join on dataset 1 and dataset 2 with common field (State Code + Town/Village Code) results in the following

State Code	District Code	Sub District Code	Town-Village Code	Town-Village Name	Population
32	588	05631	803256	Kanhangad (M + OG)	1,25,564
32	591	05640	803267	Kozhikode (M Corp. + OG)	5,50,440
32	592	05641	803269	Malappuram (M + OG)	1,01,386
32	594	05654	803280	Thrissur (M Corp.)	3,15,957
32	595	05659	803288	Kochi (M Corp. + OG) (Part)	6,33,553
32	598	05674	803299	Alappuzha (M + OG)	2,40,991
32	600	00000	803306	Kollam (M Corp. + OG)	3,67,107
32	601	00000	803312	Thiruvananthapuram (M Corp. + OG)	7,88,271
32			803275	Palakkad (M)	1,30,955

giz

- A right join is similar to a left join, but it returns all the rows from the right dataset and the matching rows from the left dataset.
- The field (32, 803275) in dataset 2 does not have a matching record in dataset 1. However, this record is included in the merged dataset.
- The records for the fields (32, 803310) and (32, 803313) are in dataset 1. Hence, they are not included in the merged dataset.

Common Data Joining Function in Excel – VLOOKUP

Vertical Lookup is a function commonly used in spreadsheet programs to search for a specific value in a table or range of data and retrieve a corresponding value from a specified column.

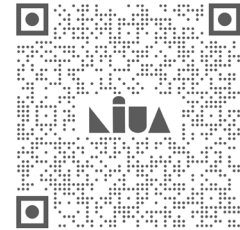
`=VLOOKUP(E2,Sheet2!C2:D7,2)` → Inserts population data from sheet2 into sheet1 for the town code

	A	B	C	D	E	F		A	B	C	D
1	State Code	District Code	Sub District Code	Town Code	Town-Village Name	Population	1	State Code	Town Code	Town Name	Population
2	32	588	5631	803256	Kanhangad (M + OG)	125564	2	32	803256	Kanhangad (M+OG)	1,25,564
3	32	591	5640	803267	Kozhikode (M Corp. + OG)	550440	3	32	803267	Kozhikode (M Corp.+OG)	5,50,440
4	32	592	5641	803269	Malappuram (M + OG)	101386	4	32	803269	Malappuram (M+OG)	1,01,386
5	32	594	5654	803280	Thrissur (M Corp.)	315957	5	32	803280	Thrissur (M Corp.)	3,15,957
6	32	595	5659	803288	Kochi (M Corp. + OG) (Part)	633553	6	32	803288	Kochi (M Corp.+OG)	6,33,553
7	32	598	5674	803299	Alappuzha (M + OG)	240991	7	32	803299	Alappuzha (M+OG)	2,40,991

Sheet 1 Sheet 2

giz

- Vertical Lookup is a function commonly used in spreadsheet programs, such as Microsoft Excel for integrating data from multiple sources. It searches for a specific value in a table or range of columns and retrieve a corresponding value from a specified column.
- Please refer this source to execute the function in excel: <https://support.microsoft.com/en-us/office/vlookup-function-0bbc8083-26fe-4963-8ab8-93a18ad188a1>



VLookup function

Data Summarisation

Presentation of key information from a dataset in a concise and informative manner

Allows extracting important patterns, or insights from raw data

Descriptive Statistics, Visualisation, Aggregation etc are popular methods of summarization

Pivot tables are also a powerful tool for summarizing data

- Aggregates and summarises data across multiple variables by grouping rows and columns and applying aggregate functions to specific variables.
- Popular tool with Microsoft Excel

giz

- Data Summarisation allows presentation of key information from a dataset in a concise and informative manner. It extracts important patterns and insights from data.
- Descriptive statistics and visualisation are popular methods of summarisation.
- Pivot table is a powerful data analysis tool available in spreadsheet applications such as Microsoft Excel and Google Sheets for summarizing large and complex datasets.
- A Pivot Table can summarise, sort, reorganise, group, count, total or average data stored in a table. It allows us to transform columns into rows and rows into columns. It allows grouping by any field (column), and using advanced calculations on them.

Data Summarisation – Pivot Table

Ex:

Sl.No	Location	Price	Area	No. of Bedrooms
69	Banashankari	2.3E+07	2480	3
351	Banashankari	2.9E+07	3033	4
352	Banashankari	2.1E+07	2378	3
353	Banashankari	2.4E+07	2522	3
354	Banashankari	3E+07	3205	4
702	Banashankari	3E+07	3205	4
10	Electronic City	3506000	660	1
28	Electronic City	3506000	660	1
48	Electronic City	3506000	660	1
2661	Jayanagar	6400000	925	2
2663	Jayanagar	7175000	1285	3
5273	JP Nagar	7840000	1500	3
5704	JP Nagar	1.9E+07	1500	3
3500	JP Nagar	4900000	1150	2
4619	JP Nagar	2288000	1050	2
5232	JP Nagar	7900000	1111	2
5233	JP Nagar	9318000	1111	2
3857	Raja Rajeshwari N	5661000	1500	3
3973	Raja Rajeshwari N	8400000	1105	2

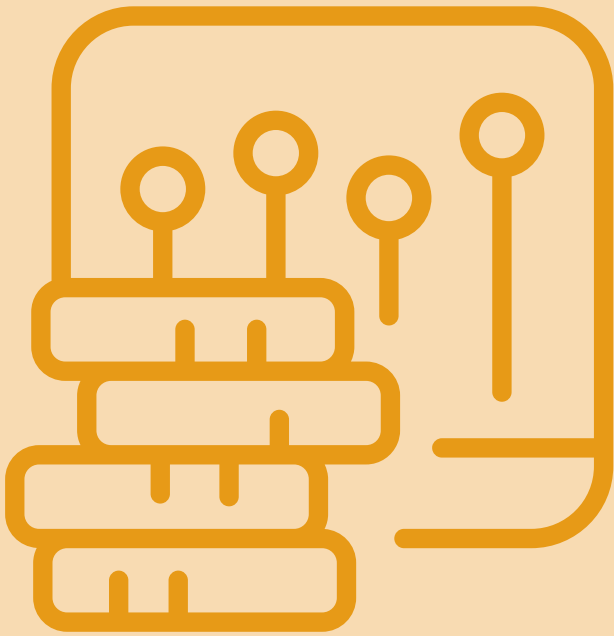
Average of Price	Column Label				
Row Labels	1	2	3	4	Grand Total
Banashankari		78,42,965	1,72,19,436	2,42,24,308	1,49,86,667
Electronic City	54,08,700	60,20,488	82,39,335	1,11,23,054	74,10,575
Jayanagar		2,01,00,000	1,01,70,111	2,60,00,000	1,41,37,929
JP Nagar	58,16,750	77,00,955	1,54,06,741	2,61,04,707	1,50,83,552
Raja Rajeshwari Nagar		62,66,500	43,69,667		51,28,400

giz

- This slide shows an example of pivot table on a housing dataset. The pivot table determines the average price, are and number of bed rooms for each locality.
- The steps for creating pivot table in excel is given in this link:
- <https://support.microsoft.com/en-au/office/create-a-pivottable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99134456576>
- Expand Create a PivotTable in Excel for Windows for creating a basic pivot table. Other sections can also be expanded for advanced pivot table functionality.



Pivot table summary



MODULE 2

WORKING WITH SPATIAL DATA

Session 1: Introduction to Spatial Data

Duration: (Ideal) 1 hour

Session 1: Introduction to Spatial Data

SESSION SUMMARY

1	BRIEF SESSION DESCRIPTION	This session covers on the concepts of spatial thinking, GIS and remote sensing. It covers on the basic spatial data types, file formats and extensions. Participants will explore Google Earth Pro during the hands on session.
2	LEARNING OUTCOMES	At the end of the sessions participants will be able to identify and differentiate between vector and raster data and their formats. Participants will be able to create some points, lines and polygons over satellite imagery in google earth and export a simple map.
3	CASE STUDIES (IF ANY)	Several small case studies and examples are integrated into the presentation itself
4	PRACTICE DATASETS	None
5	FACULTY REQUIREMENT	Familiarity with concepts of spatial thinking, GIS and Remote Sensing and its application. Experience of working with Google Earth Pro.
6	LEARNER PREREQUISITES	None
7	CLASSROOM ARRANGEMENT	Traditional Classroom (Refer Annexure 3)
8	TECHNICAL REQUIREMENTS	Access to Internet and installing Google Earth Pro. An external wired/ bluetooth mouse for easy use of the software

Before we learn more about GIS and Remote Sensing, we need to understand the importance of the spatial thinking first. Spatial thinking is essential to create precise, relevant, and logical maps, and to help understand, communicate and answer some of the complex questions using data and data visualisation.

Spatial Thinking

1

"...is thinking that finds meaning in the shape, size, orientation, location, direction or trajectory, of objects, processes or phenomena, or the relative positions in space of multiple objects, processes or phenomena."

It is defined by its three main components.

Concepts of Space

- Measurements, calculations – scales, coordinate system
- Understanding of its properties – dimensionality, proximity, continuity and separation

Tools of Representation

- any medium – maps, drawings

Process of Reasoning

- Spatial relationships leading to decision making

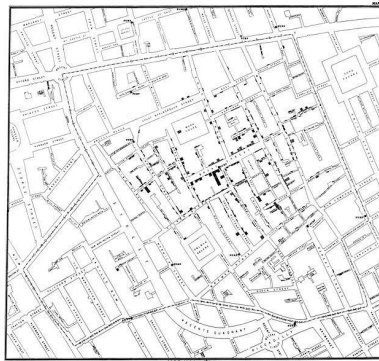
Source: National Academies of Sciences, Engineering, and Medicine. 2006. Learning to Think Spatially. Washington, DC: The National Academies Press. <https://doi.org/10.17226/11019>.

giz

- The definition of the spatial thinking and its components as per the book Learning to Think Spatially (2006) published by National Academies of Sciences, Engineering, and Medicine. <https://doi.org/10.17226/11019>
- In this context, we are specifically talking about physical space on earth in three dimensions - x,y,z and also across time. So we find meaning, relationships, and ways to define problems, figure out answers by measuring the space and the elements in its space, we can understand their relationships such as how big or small they are, how close or far they are, etc.
- How to best represent and visualize these relationships using various tools, mediums, such as maps etc
- And the last and the most important is how read and perceive these to make decisions.

John Snow's cholera map, 1854

2



Map of the book "On the Mode of Communication of Cholera" by John Snow, originally published in 1854 by C.F. Cheffins, Lith, Southampton Buildings, London, England. Image downloaded from <https://en.wikipedia.org/wiki/File:Snow-cholera-map-1.jpg#file>

giz

- This is one of the first applications of spatial thinking approaches, through spatial data on cholera cases, water distribution maps, representing it on a map to identify the source of the cholera outbreak.
- Please read more about the map and its origins from here https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak to explain the full back story of the map
- GIS aids in spatial thinking by measuring, overlaying and representing geographical data and analysis.

Geographic Data

Geographic data is data and information with specific location or position information on earth. This can be

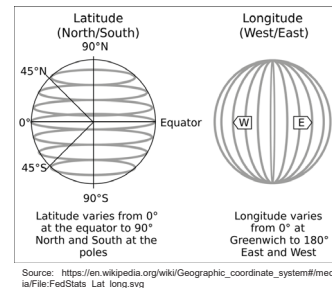
Geographic Coordinates - Longitude (x) , Latitude (y)

- 79.080598° E 21.149850° N (Decimal Degrees)
- 79° 4'50.14"E 21° 8'59.45" N (Degrees Minutes Seconds)

Textual Information Address information, Place name, Landmarks, Pin Code etc.

- Near Zero mile stone, Nagpur, Maharashtra, India

Geocoding, is the process of taking a text-based description of a location to geographic coordinates.



Source: https://en.wikipedia.org/wiki/Geographic_coordinate_system#/media/File:FedStats_Lat_long.svg

giz

- Geographic data is data and information with specific location / position information on earth. Geographic data is also referred commonly as spatial data, geoinformation, geodata.
- Image shows the Latitude and Longitude lines. Longitude lines are the lines parallel to meridian and Latitude are the lines, parallel to the equator, referred to as x, y and the height information is z value, the third dimension. Geographic Coordinate System (GCS) is the system for measuring and positioning on the earth surface.
- While the addresses, pin codes have some form of the geographic information that represent a specific place on the earth, they have to be converted to the location information of Longitude and Latitude to accurately locate them. This process of converting from place names, address to location information is called on geocoding.
- Determining, calculating the location of entity easier with Global Positioning System (GPS)
- World Geodetic System 1984 (DD MM SS / DD.DD) and Universal Transverse Mercator coordinate system (Meters) are two commonly used systems for measuring location.

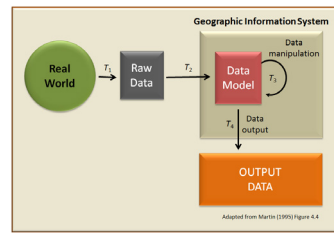
Geographical Information Systems (GIS)

4

GIS consists of integrated computer **hardware** and **software** that **store, manage, analyse, edit, output,** and **visualise geographic data.**

The two other important part of a system are **People** and **Feedback**

Geographic data is data and information with specific location / position information on earth.



Source: https://en.wikipedia.org/wiki/Geographic_information_system

giz

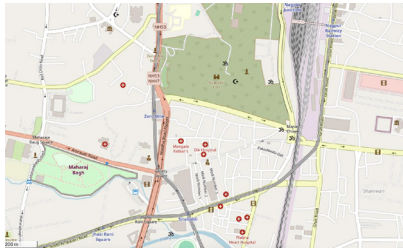
- A definition of GIS and Geographic data.
- GIS like any other information system consists of Hardware (Computer systems: Servers, Mobiles, GPS), Software (OS, GIS software like Quantum GIS), Processes that are used for capturing, storing, retrieving, analyzing, managing, presenting and sharing Geographic data. The other two main components of GIS or any other information system are People – who are users, consumers/audience of the information, maps etc, and the Feedback – to consistently improve the processes and systems etc.
- As represented in the figure, the raw data, which is captured/representing the real world is used to model/analyze in GIS to create data and information for various purposes. Examples of GIS based master plans, GIS for weather monitoring and prediction and other relevant examples can be used to explain about the applications and use of GIS based on audience in the room.
- This is not be confused with Geo Information Science which is related the discipling of studying methods and techniques of analysis geographic data.
- Google maps is one example of GIS system that is of everyday use.

Geographic Data Types: Vector Data

5

Vector data

- is stored as a series of X, Y coordinate pairs
- is used to represent points, lines and areas/polygons
- features have attributes.
- is generated through data creation process such as digitisation, GPS trackers etc.



Source: OpenStreetMap

Source: <https://docs.qgis.org>

giz

- Also referred to as data models – as in modelling real world data in GIS.
- There are two types of geographic data – vector data and raster data.
- When a feature’s geometry consists of only a single vertex, it is referred to as a point feature – Location of hospitals and other public amenities.
- Where the geometry consists of two or more vertices and the first and last vertex are not equal, a polyline feature is formed. Roads and Rail way networks for example.
- Where three or more vertices are present, and the last vertex is equal to the first, an enclosed polygon feature is formed. Buildings, Parks, Administrative Boundaries for example.
- Vector features have attributes, which consist of text or numerical information that describe the features. The name, address, photographs etc that is attached to a feature on google maps in an example. Try searching for zero mile point on google maps and look for its attribute information.

Vector File Formats

6

Name	Extension
Esri Shapefile	.shp
Geopackage	.gpkg
Geographic JavaScript Object Notation (GeoJSON)	.geojson
Google Keyhole Markup Language (KML/KMZ)	.kml/.kmz
GPS eXchange Format	.gpx
Open Street Map	.osm

giz

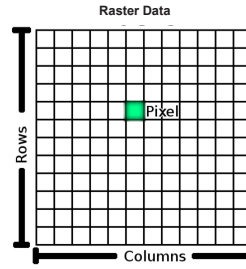
- Commonly used file formats
- Esri Shapefile is linked / dependent on .shp, .shx, .dbf, and .prj files for the softwares to read the vector files.
- Geopackage is an open file format.
- KML is the format for Google Earth Pro software.

Geographic Data Types: Raster Data

7

Raster data

- are stored as a grid of values.
- are made up of a matrix of pixels (also called cells), each containing a value that represents the conditions for the area covered by that cell.
- are generated commonly aerial photography and satellite imagery (Remote Sensing).
- can be continuous or discrete data types



Source: <https://docs.qgis.org>

giz

- Raster data is used in a GIS application when we want to display information that is continuous across an area and cannot easily be divided into vector features.
- Examples of continuous raster's can be Elevation data, Weather Data etc.
- Examples of Discrete raster's can be Land Cover Maps.

Raster File Formats

8

Name	Extension
Geographic Tagged Image File Format (GeoTiff/TIFF)	.tif/.tiff
Digital Elevation Model	.dem
Joint Photographic Experts Group (JPEG)	.jpeg/.jpg

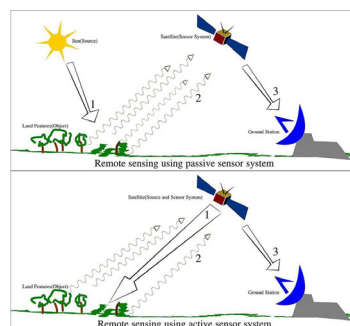
giz

Remote Sensing

Remote Sensing generally refers to the use of satellite- or aircraft-based sensor technologies to detect and classify objects on Earth. The two types of remote sensing are

- **Passive** - Using Electro Magnetic Radiation (Sun)
- **Active** – Using RADAR sensor etc.

Resolution is key to understand its uses, applications and limitations of any type of remote sensing data. The four main types are **Spectral, Spatial, Temporal and Radiometric resolution**,



Source: https://en.wikipedia.org/wiki/Remote_sensing#/media/File:Remote_Sensing_illustration.jpg

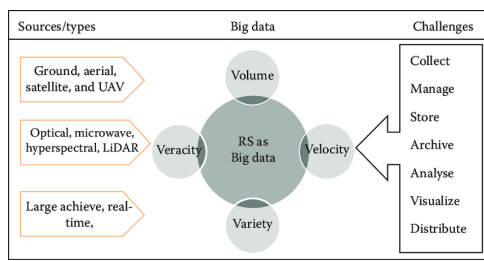
Source: https://en.wikipedia.org/wiki/Remote_sensing



- Definition of Remote Sensing.
- Most common type of remote sensing is passive remote sensing, where sensors onboard satellites measuring in the reflected radiation from objects on earth. Different elements absorb and reflect electro magnetic radiation differently, which is used for identifying features. Most humans are only able to see within the visible spectrum of the electro magnetic radiation, the sensor are designed to also capture other wavelengths.
- We will not be going into the details of the resolutions in this course, but to read more info about remote sensing. <https://www.earthdata.nasa.gov/learn/backgrounders/remote-sensing>

Remote Sensing ‘Big Data’

- Continuous acquisition from space and air borne sensors of various spatial, spectral resolution and for various applications
- Requirement of storing, processing, analysing, disseminating and archiving large volumes of Remote Sensing data is a challenge.
- Recent advances in High performance computing, cloud computing, cloud storage are handling Remote Sensing Big data



Source: https://en.wikipedia.org/wiki/Remote_sensing



Reference Reading:

- <http://repository.uwl.ac.uk/id/eprint/1732/1/fgcs2.pdf>
- https://www.researchgate.net/publication/274252933_Processing_Remote-Sensing_Data_in_Cloud_Computing_Environments

Further Reading and Reference Material:

- https://webapps.itc.utwente.nl/librarywww/papers_2009/general/principlesgis.pdf
- [https://geo.libretexts.org/Bookshelves/Geography_\(Physical\)/Essentials_of_Geographic_Information_Systems_\(Campbell_and_Shin\)](https://geo.libretexts.org/Bookshelves/Geography_(Physical)/Essentials_of_Geographic_Information_Systems_(Campbell_and_Shin))
- <https://www.e-education.psu.edu/geog468/>
- https://saylordotorg.github.io/text_essentials-of-geographic-information-systems/

Open-source data - Vector format

11

Downloading vector data
from OpenStreetMap
(OSM)



giz

- What is Open street map: <https://www.openstreetmap.org/about>
- Working with Open Street Map <https://welcome.openstreetmap.org/working-with-osm-data/>
- For other prospects of Open street maps, refer the following link <https://www.openstreetmap.org/help>

Open-source data - Raster format

12

Downloading remote
sensing indices from
Sentinel hub.



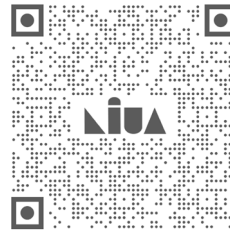
giz

- Generic information about USGS <https://www.usgs.gov/about>
- Additional reference links for downloading data from USGS <https://www.usgs.gov/the-national-map-data-delivery/gis-data-download>

Spatial data format conversion

13

Hands-on session on Map Shaper to convert data from one format to another



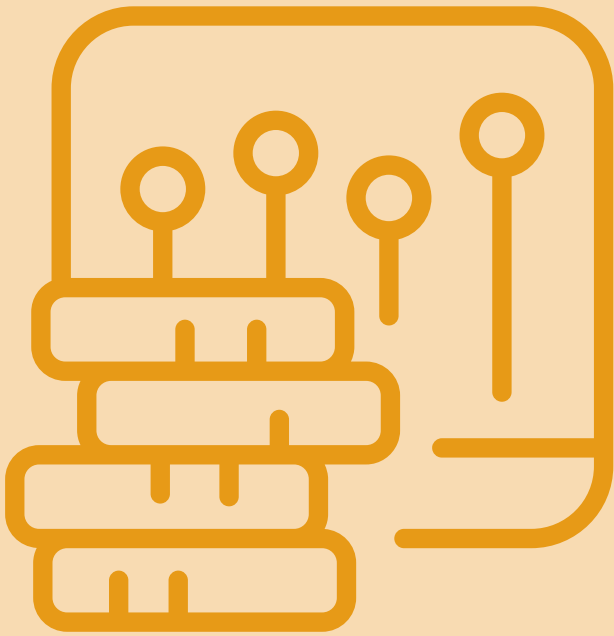
giz

MAP SHAPER

While Map Shaper offers a wide range of functionalities for spatial data manipulation, this guide focuses on its core strength: converting spatial data between different formats.

Learning Resources:

- This YouTube video demonstrates the conversion process step-by-step: <https://www.youtube.com/watch?v=Vk4-iSPVXgE>
- The video showcases a specific format conversion. However, the general steps remain consistent across different formats. When exporting your files, simply choose the desired output format from the available options.



MODULE 2

WORKING WITH SPATIAL DATA

Session 2: Visualising and Map-making

Duration (ideal) 1.5 Hour

Session 2: Visualising and Map-making

SESSION SUMMARY

1	BRIEF SESSION DESCRIPTION	This session covers on the approach to spatial data visualisation, key elements of making a map and reading a map. They will understand the vector and raster visualisation types. Using the datasets provided, they will use datawrapper and ArcGIS online to visualise vector data and share interactive maps online
2	LEARNING OUTCOMES	At the end of the session participants will be able to read and differentiate between good map from a bad map with the help of colors, classification types and map elements. Participants will create and share an online interactive map using sample vector data and its attributes using various classification techniques
3	CASE STUDIES (IF ANY)	Several small case studies and examples are integrated into the presentation itself
4	PRACTICE DATASETS	Folder: Day 2- Data sets/ Session 2 1. CSV File of Trivandrum PCA (2011) ward level data 2. Census 2011: 1 lakh + cities table that can be used for geo coding and then importing into visualisation software. Access via https://drive.google.com/drive/folders/1srtRCWEI-ZrTe8Nyg1ZbMQL313w4PEeP?usp=sharing
5	FACULTY REQUIREMENT	Working knowlege of importing and visualising spatial Data on Datawrapper and ArcGIS online.
6	LEARNER PREREQUISITES	None
7	CLASSROOM ARRANGEMENT	Traditional Classroom (Refer Annexure 3)
8	TECHNICAL REQUIREMENTS	Access to Internet, Signing up to the free / public account on ArcGIS online and Data Wrapper. An external wired/bluetooth mouse for easy use of the software. Sign-in or sign up for google sheets and Geocode by Awesome Tables for Geocoding.

Steps for making a Map or Visualisation

1

- Purpose, objective, audience and the context of the map
- Selection and organization of geographic information – Available data and Cost
- Analysis
- Map creation and design
- Map interpretation and integration with other existing information

giz

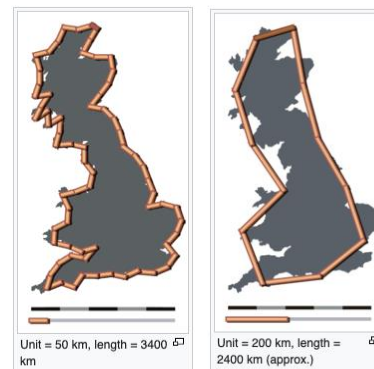
This outlines a fundamental approach preceding the creation of maps or any form of geographic data visualisation. It is also applicable to other forms of data visualisation.

- It's crucial to determine the purpose, objectives, audience for the map, and its intended context before initiating the process. This clarity is essential at the beginning.
- Based on the map's intended context and objectives, we can then identify the available data for its creation. Sometimes, maps or visualisations utilize existing available data for representation. However, if the relevant data isn't accessible, consideration of costs becomes necessary for either data creation or acquisition.
- Once the objectives and data have been identified, the process of map design and creation is started. The subsequent slides present various methods for classifying and representing the data.
- Understanding how users will interpret the map and its analysis is equally significant. Accompanying text or descriptions, either on the map or separately, aid in this comprehension.

Representation with Scale

2

An example of the coastline paradox. If the coastline of Great Britain is measured using units 100 km (62 mi) long, then the length of the coastline is approximately 2,800 km (1,700 mi). With 50 km (31 mi) units, the total length is approximately 3,400 km (2,100 mi), approximately 600 km (370 mi) longer.



Source: https://en.wikipedia.org/wiki/Coastline_paradox

giz

- Understanding how representations change with different scales and their impact on generalization is important.
- The amount of information to be visualized and displayed on a map depends upon factors such as scale, the map's purpose, and the intended audience.

Key elements

3

The acronym **DOGSTAILS** makes it easy to remember the Key elements:

D	Date	When was the map made? Is it still reliable?
O	Orientation	Principle Directions - Compass Rose / North Arrow
G	Grid	Set of lines forming grids representing the Coordinates on the map Latitude & Longitude
S	Scale	Is there a map scale?
T	Title	What is the title of the map? What is the time and place of the map?
A	Author	Who made the map?
I	Index	Does the map have an index where specific information can be found?
L	Legend	Does the map have a legend that explains the map symbols?
S	Source	Source of the Data

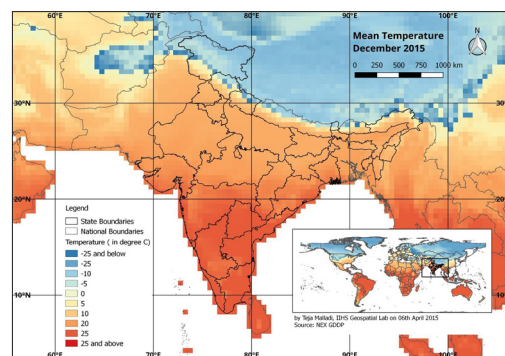
Source: <https://education.nationalgeographic.org/resource/map/>

giz

- DOGSTAILS is an easy way to remember the key elements for making a map and reading map.
- The image of the next slide is an example of how to read a map with the help of DOGSTAILS.

Key elements

4



Source: Author

D Date

O Orientation

G Grid

S Scale

T Title

A Author

I Index

L Legend

S Source

giz

Color schemes

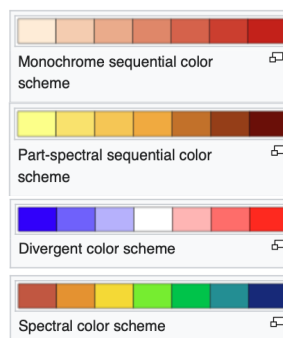
5

Continuous or Graduated Data

- Sequential
- Divergent

Discrete or Categorized Data

- Spectral



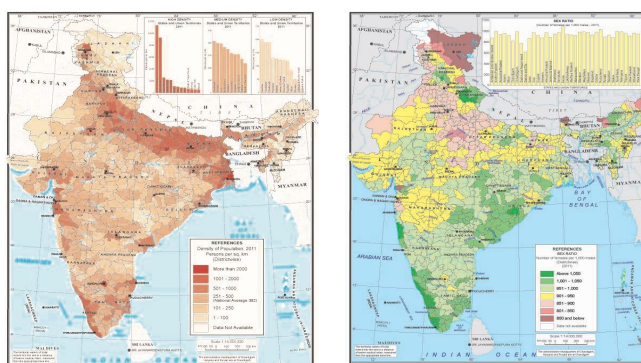
Source: https://en.wikipedia.org/wiki/Color_scheme

giz

- Selecting the right color scheme is important for effectively communicating the intended message and creating attractive visualisations. These color schemes are applicable for both vector and raster data visualisation.
- For data that is continuous or graduated, such as the visualisation of population density or temperature, Sequential and Divergent color schemes can be used. For example, temperature data can be visualized from minimum to maximum using a sequential scheme, where lower values are represented with lighter intensity and higher temperatures with darker intensity. Similarly, variations from the long-term average can be visualized using a divergent scheme, where data closer to the average uses the middle color, usually white, while data diverging from the average can be represented by other hues, depending on the map type.
- For categories and discrete values, such as land use maps, spectral color scheme is used that differentiates each of the categories.
- Some of the thematic maps are shown in the next few slides, that uses various color schemes.

Vector Data | Choropleth

6



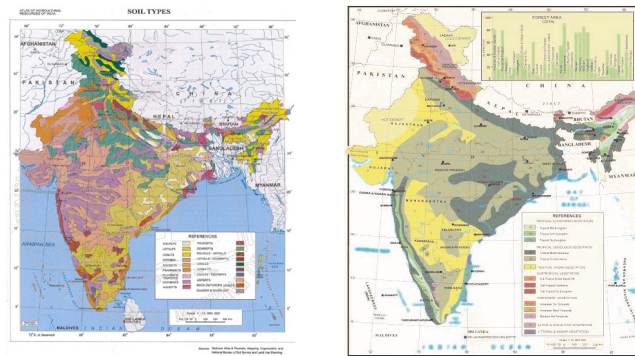
Source: 2023 National Atlas & Thematic Mapping Organisation | Government of India

giz

- Choropleth maps are the thematic maps that show a specific variable across different areas – commonly administrative units such as state, district, ward boundaries etc.
- The left map shows a sequential color scheme and the map on the right shows a divergent color scheme.

Vector Data | Chorochromatic

7



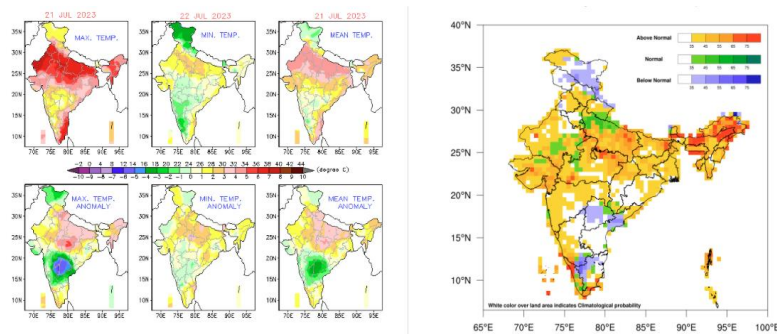
Source: 2023 National Atlas & Thematic Mapping Organisation | Government of India

giz

- Chorochromatic maps are thematic maps that show various categories across regions. The two maps shows spectral or different colors to visualize various categories.

Raster Data | Continuous Raster

8



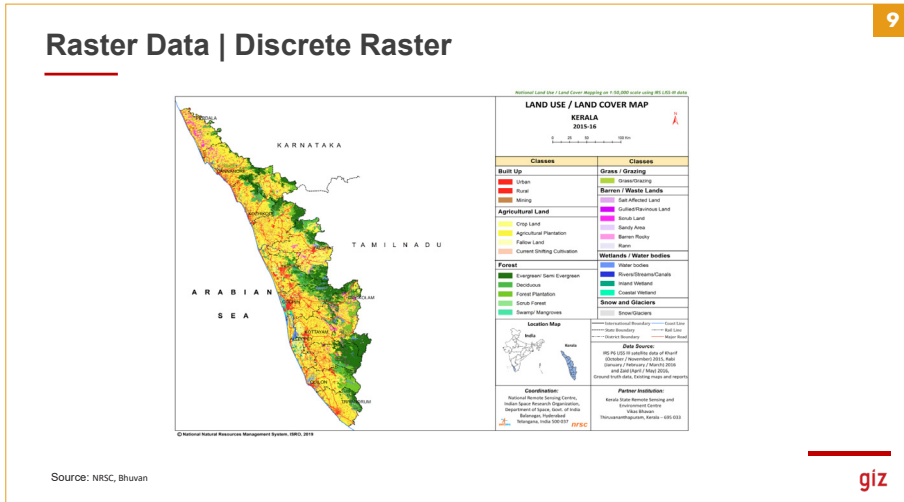
Source: IMD

giz

- These maps are the examples of using sequential or divergent colors scale to visualize continuous data.

Raster Data | Discrete Raster

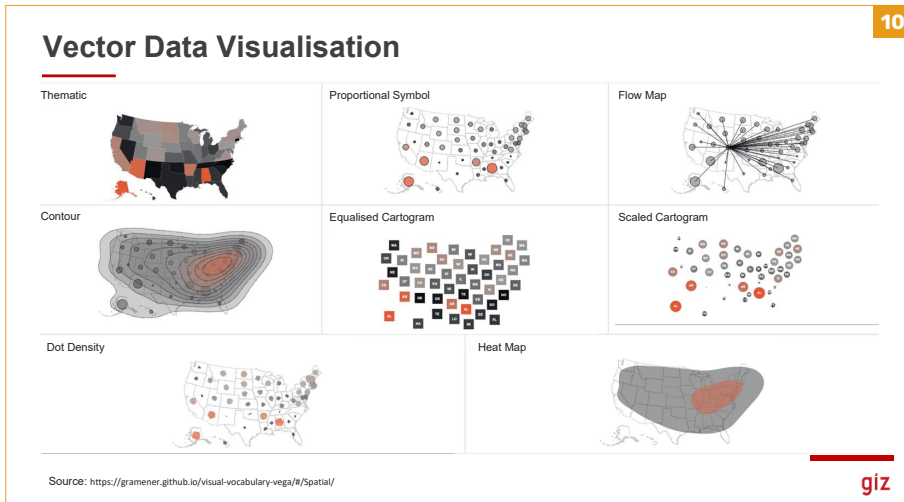
9



- This map is an examples of using discrete colors scheme to visualize different categories.

Vector Data Visualisation

10



- Different types of spatial data visualisation. The choice of these will depend on the data available, the object and the audience. In addition to these there are also 3D visualisations.

Spatial Map making

11

Hands-on session on Google Earth Pro to explore



Creation of spatial data
using Google earth pro



Time series satellite
imageries_Google earth pro

giz

For getting started, download the desktop version of Google Earth Pro from <https://www.google.com/earth/about/versions/> Please note that desktop version is different from the web and mobile versions and has more advanced features and tools

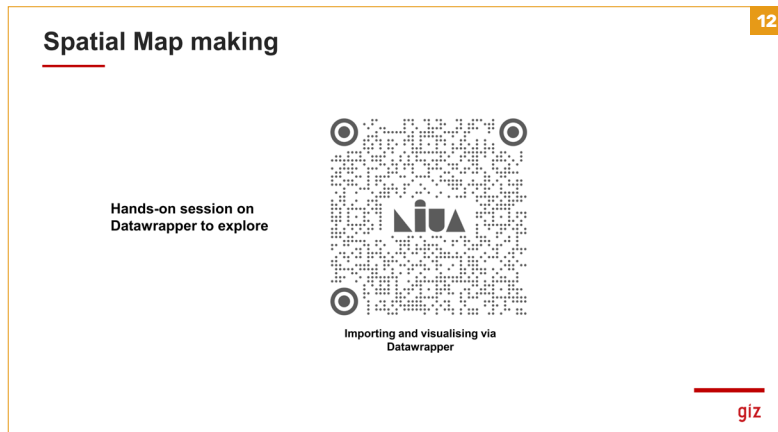
Learning Resources

- For detailed information, tutorials and learning resources on all features and tools, refer to the official Google Earth Pro user guide at <http://earth.google.com/intl/ar/userguide/v4/index.html>

Additional Resources

In addition to the user guide, the following links from the GeoDelta Labs provide step-by-step video demonstrations of Google Earth Pro tools and features:

- Google Earth Pro - A Complete Beginner's Guide by GeoDelta Labs - <https://youtube/31Gl1VZjtg4?feature=shared>
- Google Earth Pro Advanced Tutorial (Part 1) by GeoDelta Labs - <https://www.youtube.com/watch?v=9G99E4kDIQ>
- Google Earth Pro Advanced Tutorial (Part 2) by GeoDelta Labs - [https://www.youtube.com/watch?v=BKrssHG\]w7k](https://www.youtube.com/watch?v=BKrssHG]w7k)



Before you begin

- For getting started, please create an account at <https://www.datawrapper.de> There is also an option of creating visualisations on Data Wrapper without signing up. However, users will be able to share interactive maps and charts with a trial account.

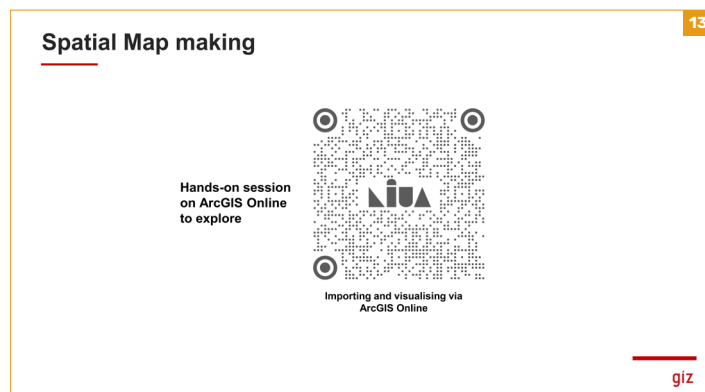
Learning Resources

- For detailed information, tutorials and learning resources on all features and tools, refer to Data Wrapper Academy <https://academy.datawrapper.de>. The academy has many tutorials on using Datawrapper for creating visualisations using spatial and tabular data.

Creating Maps using Data Wrapper

The following links provide tutorials for creating choropleth and symbol maps.

- Choropleth Maps - <https://academy.datawrapper.de/category/93-maps>
- Symbol Maps - <https://academy.datawrapper.de/category/278-symbol-maps>



Before you begin

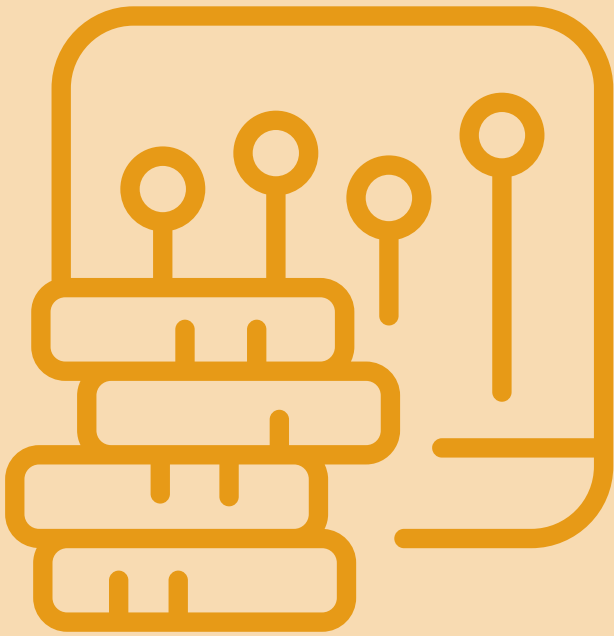
- Please sign up for an ArcGIS Online Public Account for free at https://www.arcgis.com/sharing/rest/oauth2/signup?client_id=arcgisonline&redirect_uri=http://www.arcgis.com&response_type=token

Learning Resources

- For detailed instructions on adding layers from files and visualising them, refer to [https:// doc.arcgis.com/en/arcgis-online/create-maps/add-layers-from-file.htm](https://doc.arcgis.com/en/arcgis-online/create-maps/add-layers-from-file.htm)

Additional Resources

- In addition to the user guide, the following link from the ESRI provide step-by-step video demonstrations of ArcGIS Online features: [https://www.youtube.com/playlist?list=PLZ-9T\[cleAUweMUclDQyOVY7hVsEIA07AI](https://www.youtube.com/playlist?list=PLZ-9T[cleAUweMUclDQyOVY7hVsEIA07AI)



MODULE 2

WORKING WITH SPATIAL DATA

Session 3: Integrating Tabular Data with Spatial Datasets
Duration (ideal) 1 Hour

Session 3: Integrating Tabular Data with Spatial Datasets

SESSION SUMMARY

1	BRIEF SESSION DESCRIPTION	In this session participants will be able to cover geocode the addresses using Google sheets online and then visualise the point dataset online. Participants will also understand and practice on how to integrate tabular datasets with spatial data using table join.
2	LEARNING OUTCOMES	At the end of the session participants will be able to convert tabular data with spatial attributes to a spatial format using geocoding. They will also be able to join tabular data with spatial data and also combine different spatial data based on their location.
3	CASE STUDIES (IF ANY)	Several small case studies and examples are integrated into the presentation itself.
4	PRACTICE DATASETS	Folder: Day 2- Data sets/ Session 3 01. Sample ward boundary - Thiruvananthapuram_Census_Ward_Boundary_2011 - Shapefile 02. OSM Data sets Access via https://drive.google.com/drive/folders/1srtRCWEI-ZrTe8Nyg1ZbMQL313w4PEeP?usp=sharing
5	FACULTY REQUIREMENT	Experience of Geocoding, Georeferencing. Working knowledge of importing and visualising spatial Data on Datawrapper and ArcGIS online
6	LEARNER PREREQUISITES	None
7	CLASSROOM ARRANGEMENT	Traditional Classroom (Refer Annexure 3)
8	TECHNICAL REQUIREMENTS	Access to Internet, Signing up to the free / public account on ArcGIS online and Data Wrapper. An external wired/bluetooth mouse for easy use of the software.

DATA WRAPPER

Before you begin

For getting started, please create an account at <https://www.datawrapper.de>. There is also an option of creating visualisations on Data Wrapper without signing up. However, users will be able to share interactive maps and charts with a trial account.

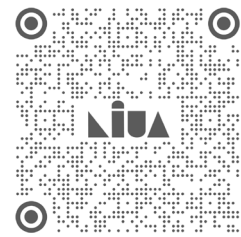
Learning Resources

For detailed information, tutorials and learning resources on all features and tools, refer to Data Wrapper Academy <https://academy.datawrapper.de>. The academy has many tutorials on using Datawrapper for creating visualisations using spatial and tabular data.

Creating Maps using Data Wrapper

The following links provide tutorials for creating choropleth and symbol maps.

- Choropleth Maps - <https://academy.datawrapper.de/category/93-maps>
- Symbol Maps - <https://academy.datawrapper.de/category/278-symbol-maps>



Spatial joining of datasets via Datawrapper

GEOCODING BY AWESOME TABLE

Before you begin

This add-on works for Google Sheets and requires a Google account. If you don't have one, sign up for free at <https://support.google.com/accounts/answer/27441?hl=en>.

Installing Awesome Table on Google Sheets

For detailed instructions on installing the Awesome Table on Google Sheets, refer to the following link: <https://support.awesome-table.com/hc/en-us/articles/4415781652370-Install-the-Awesome-Table-add-on-for-Google-Sheets>.

Learning resources for Geocoding using Awesome Table:

For detailed instructions on using the Geocode feature, refer to the Awesome Table how to guide: <https://support.awesome-table.com/hc/en-us/articles/360000112449--Part-2-Geocode-addresses>



Creation of geotag locations using Geocode

ARCGIS ONLINE

Before you begin

Please sign up for an ArcGIS Online Public Account for free at https://www.arcgis.com/sharing/rest/oauth2/signup?client_id=arcgisonline&redirect_uri=http://www.arcgis.com&response_type=token

Learning Resources

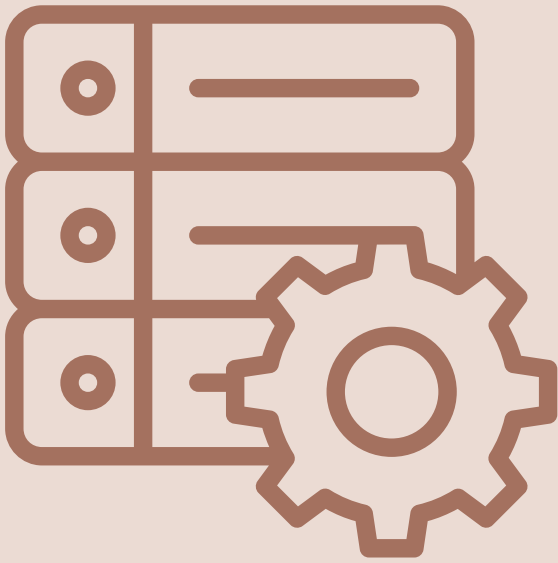
For detailed instructions on adding layers from files and visualising them, refer to <https://doc.arcgis.com/en/arcgis-online/create-maps/add-layers-from-file.htm>

Additional Resources

In addition to the user guide, the following link from the ESRI provide step-by-step video demonstrations of ArcGIS Online features: <https://www.youtube.com/playlist?list=PLZ9TJcleAUweMUcIDQy0VY7hVsEIA07AI>



Spatial joining of datasets via ArcGIS Online



MODULE 3

DATA INTEGRATION, DASHBOARDS AND DECISION-SUPPORT SYSTEMS

Session 1: Data Dashboards

Duration (ideal) 1 Hour

Session 1: Data Dashboards

SESSION SUMMARY

1	BRIEF SESSION DESCRIPTION	The session covers on the objectives and approaches to building data dashboards. Participants will also explore multiple online dashboards to evaluate their advantages and limitations. The idea that dashboards can be aimed towards decision-making (real-time) and information dissemination (public portals) is elucidated Participants will be taken through case studies on coimbatore SDG dashboard and NIUA smart city dashboards
2	LEARNING OUTCOMES	At the end of the session, participants will be able to describe the purpose and audience of different types of data dashboards, cite good examples of the same and be able to describe the salient features of a good data dashboard that can support evidence-based decision-making
3	CASE STUDIES (IF ANY)	National Gati Shakti program data dashboard
4	PRACTICE DATASETS	None
5	FACULTY REQUIREMENT	Familiarity with dashboards, for both decision-making and for information dissemination
6	LEARNER PREREQUISITES	None
7	CLASSROOM ARRANGEMENT	Traditional Classroom (Refer Annexure 3)
8	TECHNICAL REQUIREMENTS	None

Dashboards

1

Defined as a “a visual display of data used to monitor conditions and/or facilitate understanding.”

Dashboards are

- majorly web-based and interactive. Some of them are also built on excel and other tools
- dynamic in nature, with regular updates and provides a single page or summary at a glance
- a combination of interactive or static visualisations

Dashboards are developed for achieving one of the following purposes based on audience.

- Decision support
- Communication

Source: Sarikaya, A., Correll, M., Bartram, L., Tory, M., & Fisher, D. (2018). What do we talk about when we talk about dashboards?. *IEEE transactions on visualization and computer graphics*, 25(1), 682-692.

giz

- Definition is from S. Wexler, J. Shaffer, and A. Cotgreave. *The Big Book of Dashboards: Visualising Your Data Using Real-World Business Scenarios*. John Wiley & Sons, 2017 accessed from Sarikaya, A., Correll, M., Bartram, L., Tory, M., & Fisher, D. (2018). What do we talk about when we talk about dashboards?. *IEEE transactions on visualisation and computer graphics*, 25(1), 682-692.
- Dashboards are increasingly become the go to data visualisation platforms for communication and also decision making purposes.

Dashboards

2

Decision support dashboards are developed for either strategic or operational decision making. These are majorly dynamic in nature with regular updates in data.

Communication dashboards are targeted for general audience or public consumption purposes for building awareness, education purposes. These do not have to be dynamic in nature and could be built on static datasets.

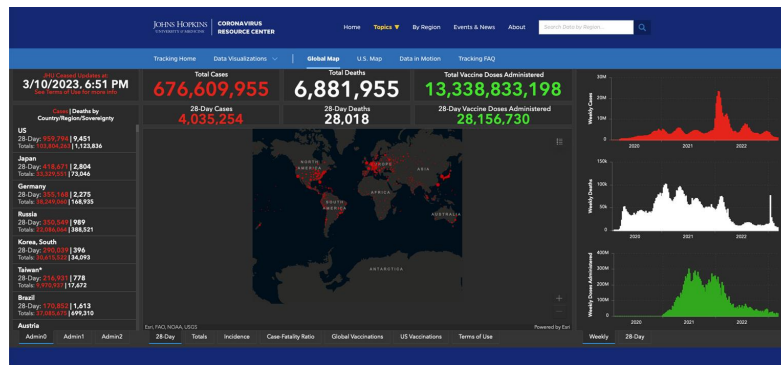
Source: Sarikaya, A., Correll, M., Bartram, L., Tory, M., & Fisher, D. (2018). What do we talk about when we talk about dashboards?. *IEEE transactions on visualization and computer graphics*, 25(1), 682-692.

giz

- This presentation is based on Sarikaya, A., Correll, M., Bartram, L., Tory, M., & Fisher, D. (2018). What do we talk about when we talk about dashboards?. *IEEE transactions on visualisation and computer graphics*, 25(1), 682-692.

Communication Dashboard Example

3



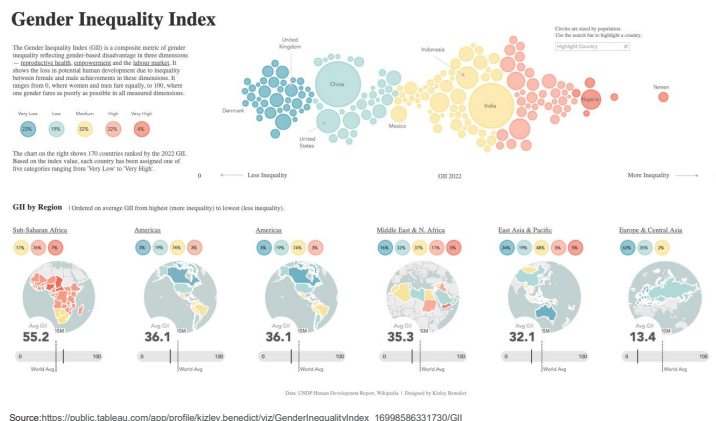
Source: <https://coronavirus.jhu.edu/map.html>

giz

- The coronavirus dashboard is one of the most visited dashboards during the pandemic. This is mostly for public information on coronavirus cases reported across the globe and their trends. It is built on ArcGIS Dashboards.

Communication Dashboard Example

4



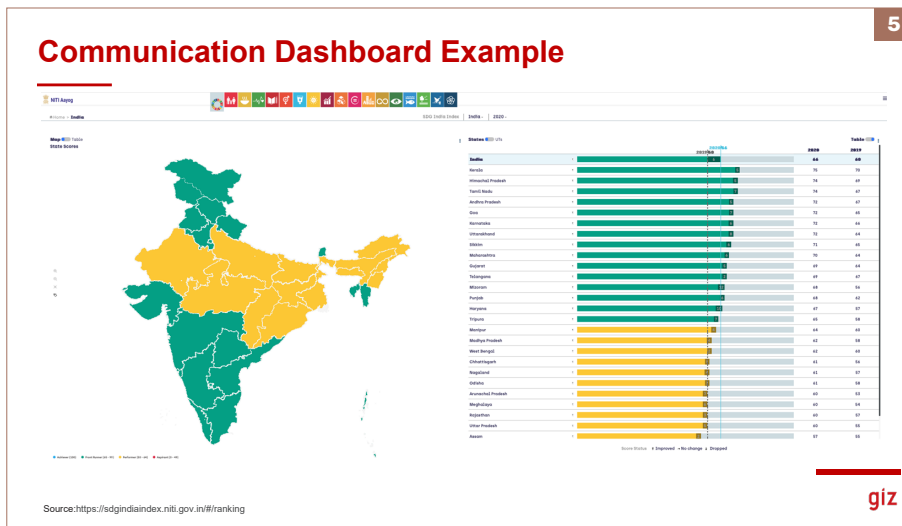
Source: https://public.tableau.com/app/profile/kizley_benedict/viz/GenderInequalityIndex_16998586331730/GII

giz

- This is an example of an interactive dashboard with static data built on Tableau. This compares gender inequality across various countries and regions across bubble charts and choropleth maps.

Communication Dashboard Example

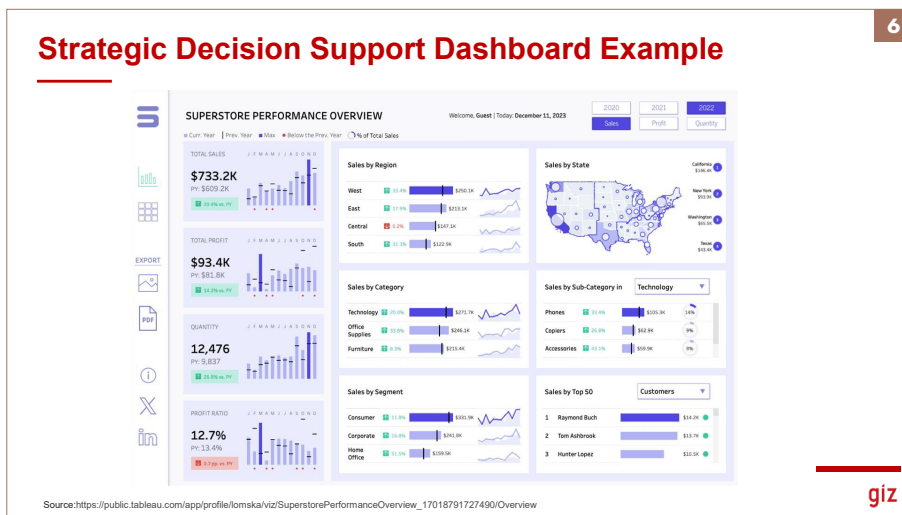
5



- An example of a SDG dashboard, developed by NITI Aayog to track the status of the SDG indicators at the state level. This is a combination of choropleth map and bar chart.

Strategic Decision Support Dashboard Example

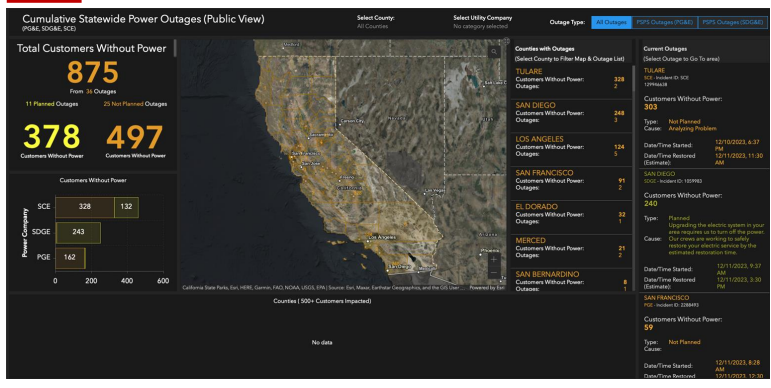
6



- This is an example of strategic decision making dashboard built on Tableau, that tracks sales across the country and across time series.

Operational Decision Support Dashboard Example

7



Source: <https://www.arcgis.com/apps/dashboards/7edefc1970444b839ebbf7b45e51e2d>

giz

- This is an example of operational decision support dashboard built on ArcGIS Dashboards. This is used for monitoring power supply for the region by the agency on live basis and respond to the outages.

Introduction

Few challenges being faced by the logistic sector of India:

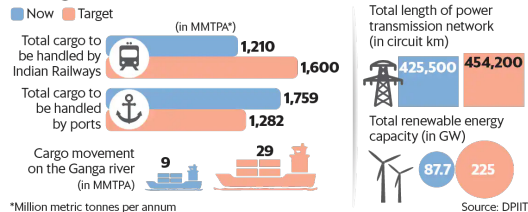
- **High Logistics Cost:** Logistics cost approx. 14% of GDP. Transportation and Warehousing are more than 60% of the cost.
- **Performance in Global Rankings:** India's Ease of Doing Business ranking (2019): 63. Logistics Performance Index (2018): 44
- **Standalone Digital Systems:** Logistics sector's digitization lacks seamless data flow, complicating documentation and approval.

Source: DPIIT

The master plan

The PM Gati Shakti aims to break inter-ministerial silos in infrastructure development. It will be achieved through integrated planning and coordinated implementation between different government departments.

Key targets by FY25



giz

The logistics industry, in India faces hurdles that impede its effectiveness and ability to compete. A key obstacle is the logistics expenses amounting to around 14% of the GDP with transportation and warehousing making up than 60% of this expenditure. Moreover India's standing in rankings concerning logistics and business friendliness is unsatisfactory ranking 44th in the Logistics Performance Index and 63rd in the Ease of Doing Business index. Another challenge lies in the absence of systems within the logistics sector resulting in complexities, during documentation and approval procedures. It is imperative to tackle these issues for enhancing the efficiency, competitiveness and overall performance of India's logistics field.

9

Aim

- Integrated Planning and Implementation
- Expediting Works on the Ground
- Saving Costs
- Creating Jobs
- Improving Competitiveness

ON A ROLL

Projects assessed on PM Gati Shakti principles

No. of projects		TOTAL 100
Ministry of Road Transport and Highways	40	
Ministry of Railways	40	
Ministry of Housing and Urban Affairs	8	
National Industrial Corridor Development Corporation	5	
Ministry of Petroleum and Natural Gas	4	
Ministry of Ports, Shipping & Waterways	2	
Ministry of New and Renewable Energy	1	

100 projects worth ₹5.89 trillion assessed on the principles of PM Gati Shakti in 54 Network Planning Group meetings since the launch of the National Master Plan. Each of these projects is worth over ₹500 crore

Some of these projects are: Pune-Bengaluru Expressway; Indo-Nepal Border-Haldia Corridor; Green Energy Corridor Phase-II; Inter-State Transmission System for 13 Gw renewable energy projects in Ladakh; Chennai-Trichy-Tuticorin Expressway

Source: business-standard.com

giz

This slide explains the Aim of the PM Gati Shakti.

10

Pillars of the project

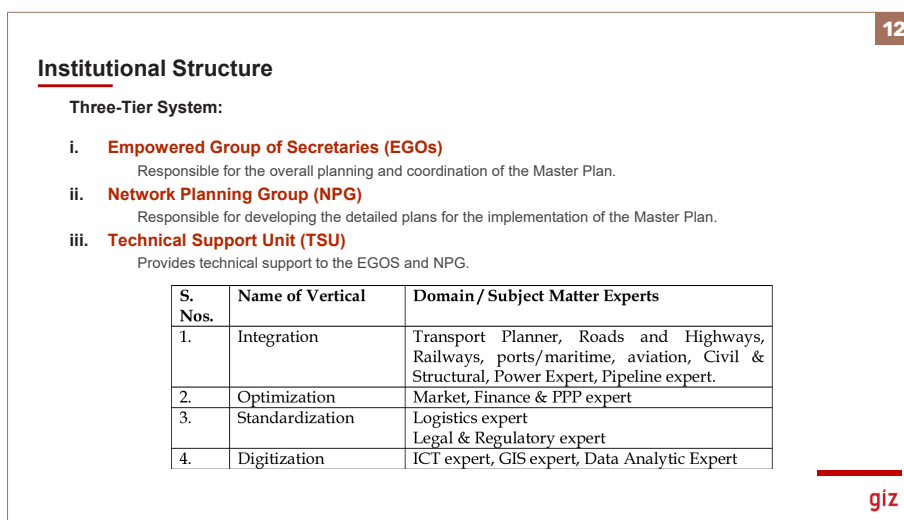
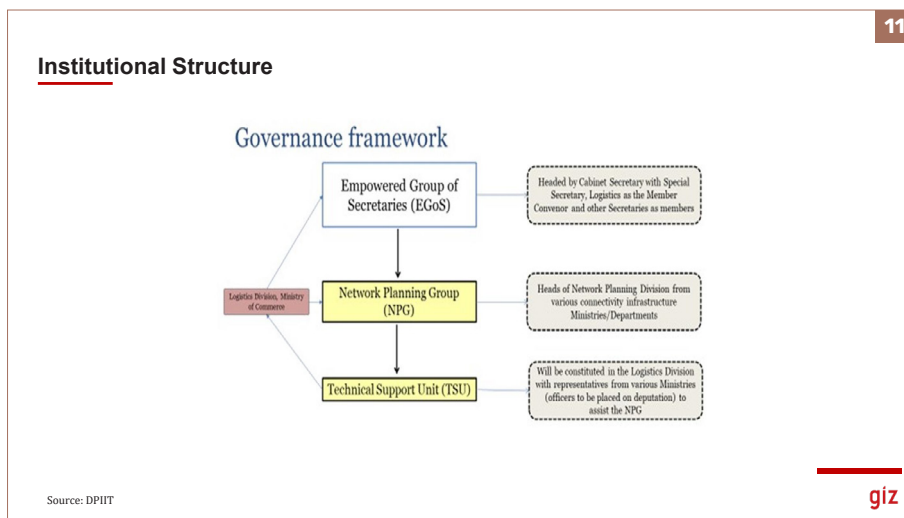
Source: pmgatishakti.gov.in

giz

The PM Gati Shakti Master Plan is an ambitious initiative by the Indian government aimed at revolutionizing the country's infrastructure landscape. Here's an updated summary of its six foundational pillars:

- **Comprehensiveness:** This pillar ensures that various ongoing and future infrastructure projects are brought together under a single, cohesive framework. This integration facilitates comprehensive development across sectors.
- **Prioritisation:** Projects are carefully selected and prioritised based on their expected economic and societal benefits. This strategic focus helps in directing efforts and resources towards initiatives with the highest potential for positive impact.
- **Optimisation:** The plan employs cutting-edge technology to maximize resource efficiency. This approach is designed to minimize logistics costs, thereby enhancing the overall effectiveness of infrastructure development.
- **Synchronisation:** A key aspect of the plan is the synchronised planning and implementation across different governmental departments and ministries. This coordination is crucial in preventing duplicative efforts and ensuring optimal resource utilization.
- **Analytical:** Advanced analytical tools, including GIS-based spatial planning, are utilized for informed decision-making and meticulous project monitoring. This data-driven approach aids in maintaining high standards of precision and accountability.

- **Dynamic:** Recognizing the fluid nature of developmental needs, the plan is built to be flexible and responsive. It can adapt to evolving priorities, ensuring that the infrastructure growth remains aligned with the nation's objectives.
- These pillars collectively form the backbone of the PM Gati Shakti Master Plan, supporting India's vision for a seamless and robust infrastructure network that drives comprehensive growth and development.



The PM Gati Shakti National Master Plan has a three-tier institutional structure designed to ensure better decision-making and coordination among various Central Ministries/Departments and State Governments. Here's a summary of the three tiers:

Empowered Group of Secretaries (EGoS): This group is responsible for providing strategic direction, planning, and enabling policy decisions to ensure the comprehensive and integrated development of infrastructure connectivity projects.

Network Planning Group (NPG): The NPG is tasked with undertaking the detailed planning of infrastructure projects, ensuring the elimination of silos in project planning and execution.

Technical Support Unit (TSU): The TSU provides technical inputs to the NPG and EGoS, ensuring that the planning and execution of projects are based on sound technical advice.

This structure is replicated at both the Central and State levels to promote synergy and coherence in the implementation of infrastructure projects across the country. The PM Gati Shakti initiative aims to enhance multi-modal connectivity and expedite the implementation of infrastructure projects, thereby boosting economic growth and improving the ease of living for citizens.

13

Key Features of the project

- **Multi-modal connectivity:** Different modes of transport will be integrated to create a seamless network.
- **Digital platform:** The data collection and analysis will be done digitally.
- **Integrated planning:** Both government agencies and private sector partners will be involved.
- **Coordinated implementation**
- **Geospatial technology**
- **Economic zones and industrial corridors**

The key features of this program are:

- **Multi-modal connectivity:** Different modes of transport will be integrated to create a seamless network.
- **Digital platform:** The data collection and analysis will be done digitally.
- **Integrated planning:** Both government agencies and private sector partners will be involved.
- **Coordinated implementation** between various ministries.
- **Geospatial technology** and its advanced features.
- **Economic zones and industrial corridors.**

14

The Seven Engines

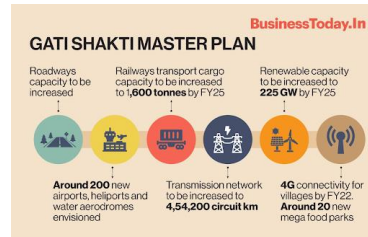
The seven engines of Gati Shakti are:	Supporting infrastructure:	Powered by:
<ul style="list-style-type: none"> • Roads • Railways • Airports • Ports • Mass transport • Waterways • Logistics infrastructure 	<ul style="list-style-type: none"> • Energy Transmission • IT Communication • Bulk Water & Sewerage • Social Infrastructure 	<ul style="list-style-type: none"> • Clean Energy • Sabka Prayas

The PM Gati Shakti program comprises of seven engines. They are Roads, Rails, Airports, Ports, Mass transport, Waterways and Logistic Infrastructure. These seven engines will pull forward the economy in unison.

Specific initiatives

The PM Gati Shakti plan includes several specific initiatives, including:

- 75 economic zones
- 11 industrial corridors
- 2 new defense corridors
- 25,000 kilometers of new roads
- 4G connectivity to all villages
- 17,000 kms of gas pipeline
- Expansion of the national highway network to 2 lakh kilometers
- Doubling of the railways network
- Development of 100 cargo terminals
- Around 200 new airports, heliports and water aerodromes
- Renewable capacity to be increased to 225 GW
- 20 new mega food parks
- 202 fishing clusters/harbors/landing centres
- Creation of a single digital platform for planning and monitoring of infrastructure projects



Source: BusinessToday.in

giz

The few initiatives of the Gati Shakti initiatives are detailed out in this slide.

What the scheme covers

The PM Gati Shakti plan will incorporate the following infrastructure schemes:

- **Bharatmala:** Development of 65,000 kilometers of new roads and upgrade 51,000 kilometers of existing roads.
- **Sagarmala:** This project aims to develop 100 ports and shipping terminals.
- **Inland waterways:** Development of 1,300 km of waterways.
- **UDAN:** This regional air connectivity scheme aims to promote air travel in tier-2 and tier-3 cities.
- **Dry Ports and Logistics Hubs:** The DPLH scheme will develop 100 dry ports and logistics hubs.



Source: pmgatisakti.gov.in

giz

The Gati Shakti initiative led by the Prime Minister includes a range of projects focused on enhancing India's transportation and logistics sectors. These projects consist of initiatives, like Bharatmala for road development Sagarmala for port enhancement inland waterways for river transport utilization, UDAN, for air connectivity promotion and the establishment of ports and logistics hubs. The goal of these efforts is to boost connectivity streamline the movement of goods and individuals and foster economic progress nationwide.

Process Reforms

17

- **Ministry of Finance circular:** All ministries must update their EFC Memos to indicate PM GatiShakti compliance.
- **Deletion or inclusion of projects:** Any deletion or inclusion of projects from the PM GatiShakti National Master Plan must be approved by the EGOS.

Source: DPIIT

giz

Digitization Efforts

18

- Single GIS platform maps economic zones and multimodal connectivity infrastructure for 3 time periods.
- Data updated continuously through APIs created by individual ministries with BISAG-N.
- LDB synchronizes physical infrastructure for multimodal transportation.
- ULIP integrates softer infrastructure of LDB and other platforms.

giz

A single GIS platform has been created to map all existing and proposed economic zones, along with the multimodal connectivity infrastructure for three time periods: 2014-15, 2020-21, and 2024-25. This will help to track the progress of infrastructure development and identify areas where improvement is needed.

The data on the platform is updated on a continuous basis through APIs created by individual ministries with BISAG-N. This ensures that the data is always up-to-date and accurate.

The Logistic Data Bank (LDB) synchronizes physical infrastructure to promote a comprehensive and integrated multimodal national network of transportation and logistics. This will help to ensure that goods and services can be transported efficiently and effectively across the country.

The Unified Logistic Interface Platform (ULIP) will integrate the softer infrastructure of LDB and other platforms under PM GatiShakti. This will create a single, unified platform for managing logistics operations.

Implementation in Goa

- Goa government to create cultural map using PM Gati Shakti portal.
- The map will include information on all cultural assets in Goa, promote tourism and preserve heritage.
- Map will be in a digital platform for users to search by location, type, event, which will also include info on infrastructure at locations, such as parking, food, accommodation.
- State government working to ensure future projects better planned, involve local stakeholders.

DRAWING BENEFITS

- Create cultural atlas with list of festivals, events and fairs and marked on a GIS map
- Make a time series map and analysis of all tourist footfalls in state
- Map all libraries, religious sites and museums
- Plan pump house along the banks of Kalna
- Evacuation plan for flood-prone regions of Sanquelim and Amona
- Mapped all fair price shops (FPS) and godowns in state
- Map all agri clusters in Goa

Source: timesofindia.indiatimes.com

giz

The government of Goa plans to utilize the PM Gati Shakti portal to develop a cultural map that will showcase all cultural assets within the state. This initiative aims to boost tourism and safeguard the region's heritage. The digital map will offer users the ability to search for cultural sites by location, type, or event. Additionally, it will provide information on infrastructure facilities available at these locations, such as parking, food, and accommodation options. The state government is committed to ensuring that future projects are well-planned and inclusive, with active involvement from local stakeholders.

Digital Master Planning Tool

The Digital Master Planning tool is a web-based application that has been developed by BISAG-N.

The comprehensive database of the ongoing and future projects of various Ministries has been integrated with **200+ GIS layers**.

The tool is based on a **Geographic Information System (GIS) platform**, which allows for the visualization and analysis of spatial data.

The maps use satellite imagery available from **ISRO** and base maps from **Survey of India**.

The tool incorporates **data on specific action plans** of all the Ministries/Departments within a comprehensive database.

The maps are built on **open-source technologies** and hosted securely on **MEGHRAJ, the cloud platform of the Government of India**.

This data is used to create dynamic maps of all infrastructure projects with real-time updation.

giz

The plan has been developed as a Digital Master Planning tool by BISAG-N (Bhaskaracharya National Institute for Space Applications and Geoinformatics) and has been prepared in dynamic Geographic Information System (GIS) platform wherein data on specific action plan of all the Ministries/Departments have been incorporated within a comprehensive database. Dynamic mapping of all infrastructure projects with real-time updation will be provided by way of a map developed by BISAG-N. The map will be built on open-source technologies and hosted securely on MEGHRAJ i.e. cloud of Govt. of India. It will use satellite imagery available from ISRO and base maps from Survey of India. The comprehensive database of the ongoing & future projects of various Ministries has been integrated with 200+ GIS layers thereby facilitating planning, designing and execution of the infrastructure projects with a common vision.

Benefits of using the Digital Master Planning Tool

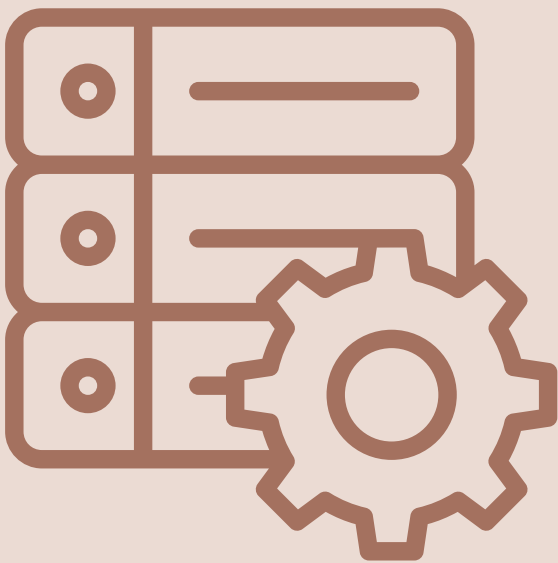
- The digital system is a software where individual Ministries will be given separate user identification (login ids) to update their data on a periodic basis.
- The data of all the individual Ministries will be integrated in one platform which will be available for planning, review and monitoring.
- The Logistics Division, Ministry of Commerce & Industry (MOCI) will further assist all the stakeholders through BISAG-N for creating and updating their required layers in the system and update their database through Application Programming Interface (APIs).

The government of Goa plans to utilize the PM Gati Shakti portal to develop a cultural map that will showcase all cultural assets within the state. This initiative aims to boost tourism and safeguard the region's heritage. The digital map will offer users the ability to search for cultural sites by location, type, or event. Additionally, it will provide information on infrastructure facilities available at these locations, such as parking, food, and accommodation options. The state government is committed to ensuring that future projects are well-planned and inclusive, with active involvement from local stakeholders.

Ministries involved

The following 21 ministries are involved in the PM Gati Shakti National Master Plan for Multi-modal Connectivity

- Ministry of Railways
- Ministry of Road, Transport & Highways
- Ministry of Ports, Shipping and Waterways
- Ministry of Civil Aviation
- Ministry of Petroleum & Natural Gas
- Department of Chemicals & Petro-Chemicals
- Department of Fertilizers
- Ministry of Steel
- Ministry of Coal
- Ministry of Mines
- Ministry of Power
- Ministry of Electronics and Information Technology
- Department of Telecommunications
- Department for Food and PDS
- Ministry of Agriculture and Farmer Welfare
- Ministry of Fisheries, Animal Husbandry & Dairying
- Ministry of Housing and Urban Affairs
- Department of Expenditure
- Ministry of Tourism
- Ministry of Commerce and Industry
- NITI Aayog



MODULE 3

DATA INTEGRATION, DASHBOARDS AND DECISION-SUPPORT SYSTEMS

Session 2: Results-based Framework
Duration (ideal) 30 min

Session 2: Results-based Framework

SESSION SUMMARY

1	BRIEF SESSION DESCRIPTION	Results-based Framework or Results-based Management (RBM) is gaining increasing traction in development sector as a way to track activities and outcomes. Primarily, a project or programmes's processes can be divided into input, activities, output, outcome and impact aspects. Key concepts like SMART (Specific Measureable Achievable Relevant Time-bound) results are defined. This session focuses on the link between these three aspects and how to prepare an RBM for the participants' own work in programme/ scheme planning and implementation.
2	LEARNING OUTCOMES	At the end of the session participants will be able to explain the concept of RBM and develop an output outcome framework for monitoring the results. They gain skills in application of the tool in developing/ analysing the municipal performance index.
3	CASE STUDIES (IF ANY)	Policy Responses to Deminishing Participation of Women in the Workforce in India
4	PRACTICE DATASETS	None
5	FACULTY REQUIREMENT	Familiarity with development sector terminology and frameworks for results-based monitoring
6	LEARNER PREREQUISITES	None
7	CLASSROOM ARRANGEMENT	Traditional Classroom (Refer Annexure 3)
8	TECHNICAL REQUIREMENTS	None

Why Result Based Management (RBM)? 1/2

1

- Greater accountability and transparency, improved performance, and generation of knowledge.
- A move from output to towards outcome-based planning in governance.
- This considers not just the inputs that are reaching our citizens, but also the outcomes that these investments have over a period of time.
- This helps us realistically assess the gaps between the actual outcomes and the desired goals.
- Data- informed decision-making- fixing targets,-allocations across sub sectors- optimization of resources and results, ensuring end outcomes for the beneficiaries.(end-to-end approach)

giz

- An evidence-based approach to policymaking is now accepted as an effective tool to improve policy outcomes. It is built around the assumption that better quality decisions will be made if the process is informed by robust evidence. Evidence-based policy provides an effective mechanism to establish, in a scientifically valid way, what works or does not work, and for whom it works or does not work. With this structured approach to evaluation, knowledge can be used to improve practice, allowing successful programs to develop iteratively over time. Rigorous evaluation can end the spinning of wheels and bring more effective social policy outcomes. At global level, J-PAL has shown that how complex development challenges can be solved with data analytics and evidence.

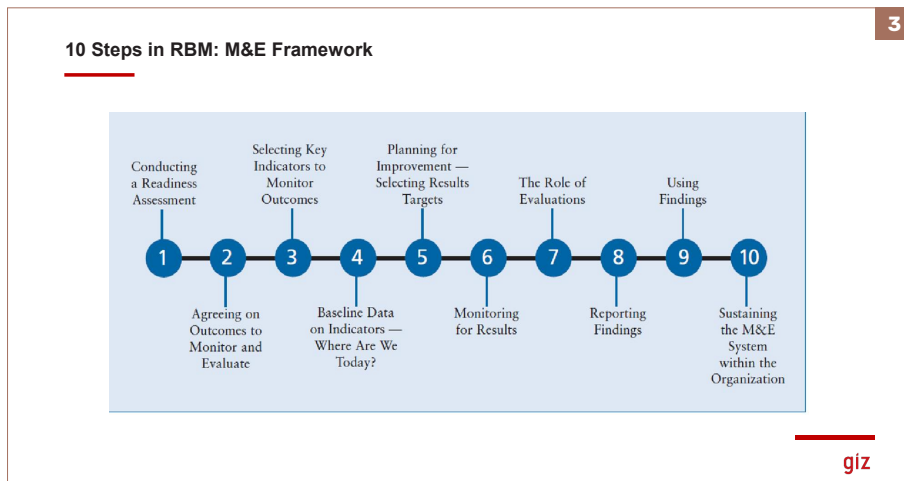
Why Result Based Management (RBM)? 2/2

2

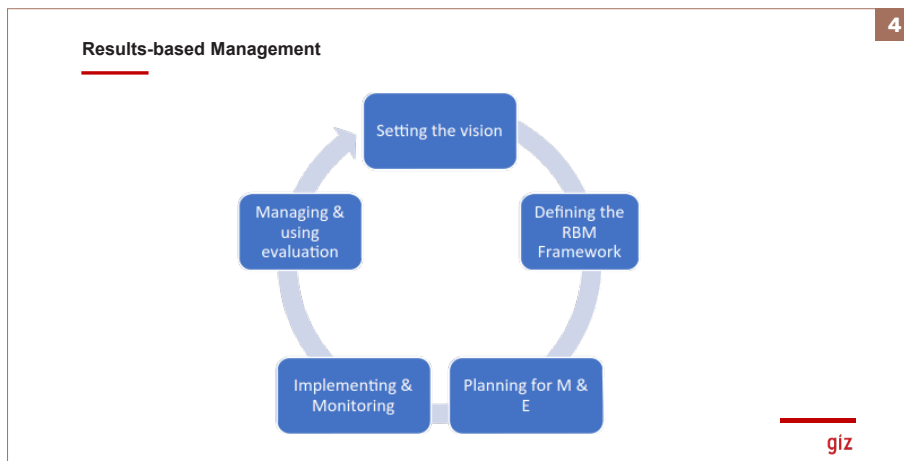
- Greater efficiency in planning and delivery of services.
- Efficient monitoring of targets, financial allocations, outputs and outcomes.
- Formulation of State/ national policies/reforms for improvement in Ease of Living and Municipal Performance of all cities in the State
- Formulate contextually relevant policies based on successful case studies

giz

- Good planning, combined with effective monitoring and evaluation, can play a major role in enhancing the effectiveness of development programmes and projects. Good planning helps us focus on the results that matter, while monitoring and evaluation help us learn from past successes and challenges and inform decision making so that current and future initiatives are better able to improve people's lives and expand their choices.



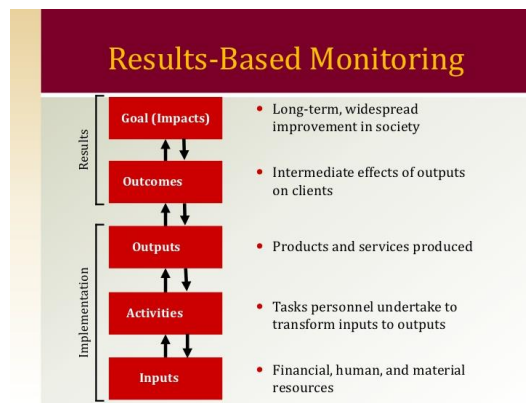
- The Ten Steps to designing, building, and sustaining a Results-Based Monitoring and Evaluation System (Jody and Rist, 2004) can be used for projects, programs, and policies for effective monitoring and enhancing Programme outcomes. The use of such results-based M&E systems can help bring about major cultural changes in the ways that organizations and governments operate. When built and sustained properly, such systems can lead to greater accountability and transparency, improved performance, and generation of knowledge.



- Integration of Planning, monitoring and evaluation together is defining Results-Based Management (RBM). RBM is defined as “a broad management strategy aimed at achieving improved performance and demonstrable results.” (UNDP,2009). Impact evaluations of social programmes have emerged as an important tool to guide social policy in developing policies as they allow for accurate measurement and attribution of impact can help policymakers identify programmes that work and those that do not work, so that effective and performing programmes can be promoted, others may require modifications in design & structures to produce better results and a few may be ineffective which need to be discontinued for resource optimization.
- Basic Concepts.
- What is Result?
 - A result is a describable or measurable change in state that is derived from a cause and effect relationship.
 - This means that a result is a change that can be observed, described and measured in some way.
- SMART Results
 - The results are SMARTS.
 - S= Specific M= Measurable A= Achievable R= Relevant T= Time bound.

Result Chain

5



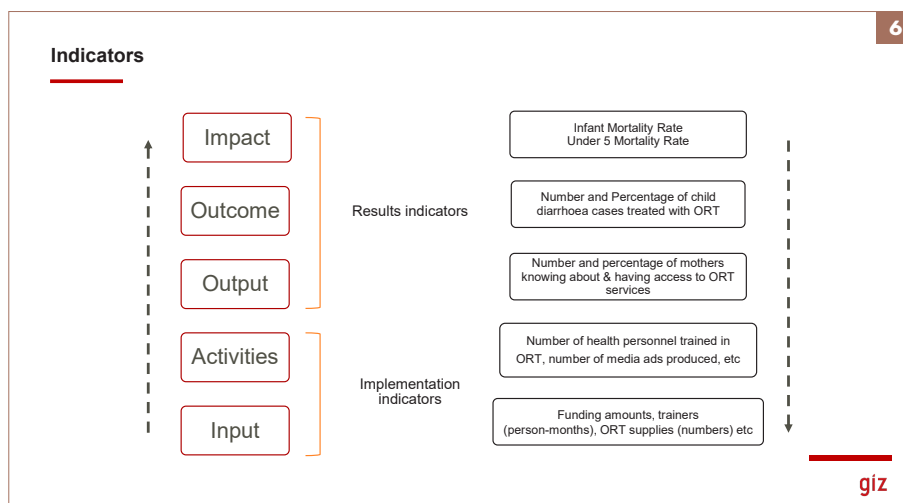
giz

The Result Chain is the link between Inputs- Activities-Outputs-Outcomes and the Impact

- Inputs: The financial, human and material resources used to implement activities.
- Activities: Work performed through which inputs such as funds, human, TA, material and other types of resources are mobilized to produce specific outputs (process indicator).
- Output: short-term consequence of activities (almost entirely under the control of manager)
- Outcome: medium-term consequence of an intervention's outputs: it shows the effects outputs have at the level of the beneficiaries (only partly under the control of the manager).
- Impact: long-term higher-order result to which a programme is intended to contribute.

Indicators

6



giz

- This is a worked out example of an RBM framework.

Challenges faced in application of RBM

- Stimulate Data Demand:
- Assess Data Requirements
- Data Validation
- Data Integration and Data visualization
- Define Use Cases
- Cross-Cutting Data Sets: - drop out- learning outcomes-school density
- Identify Data Challenges & Solutions etc. (Data Smart Cities)
- Disaster and climate-resilient urban development.
- Capacity building – data understanding and application

The Result Chain is the link between Inputs- Activities-Outputs-Outcomes and the Impact

- **Inputs:** The financial, human and material resources used to implement activities.
- **Activities:** Work performed through which inputs such as funds, human,TA, material and other types of resources are mobilized to produce specific outputs (process indicator).
- **Output:** short-term consequence of activities (almost entirely under the control of manager).
- **Outcome:** medium-term consequence of an intervention's outputs: it shows the effects outputs have at the level of the beneficiaries (only partly under the control of the manager).
- **Impact:** long-term higher-order result to which a programme is intended to contribute.

Case-study on Policy Responses to Diminishing Participation of Women in the Workforce in India

1. Background and context - FLFPR in India and the World
2. Evidence of decreasing participation of women in the workforce
3. Identification of key reasons behind the trends - context for policy
4. Measures adopted recently and qualitative evaluation of trade-offs in political economy

This slide outlines the contents of the case-study.

Female Labor Force Participation Rate (FLFPR)

Labor Force Participation Rate (LFPR) is defined as the proportion of the working-age population (15-59 years) of a country or region that is working or actively seeking work. It therefore signifies the total 'supply' of labor in a geographic area.

LFPR can be broken down by age groups, for example **youth** (15-29 years of age).

LFPR can also be broken down by gender, such as **Female Labor Force Participation Rate or FLFPR**.

LFPR can also be analysed across categories of geographic locations, such as **rural** and **urban**.

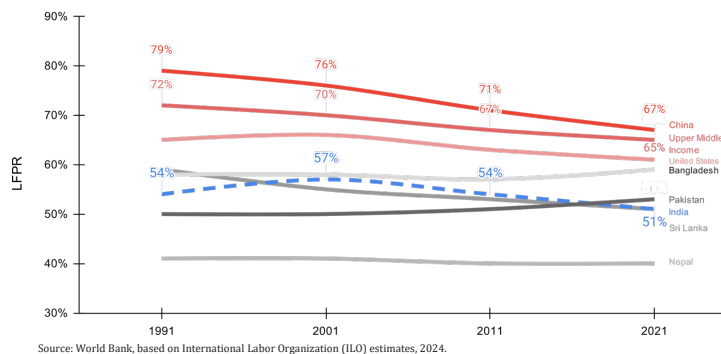
Source: Various Government of India documents and institutional reports from ILO, etc.

giz

- This slide is to lay out the key definitions that will be used in the case-study. The definitions are self-explanatory. We will be focusing on the FLFPR in this case-study.

Total Labor Force Participation Rate (LFPR)

Overall LFPR

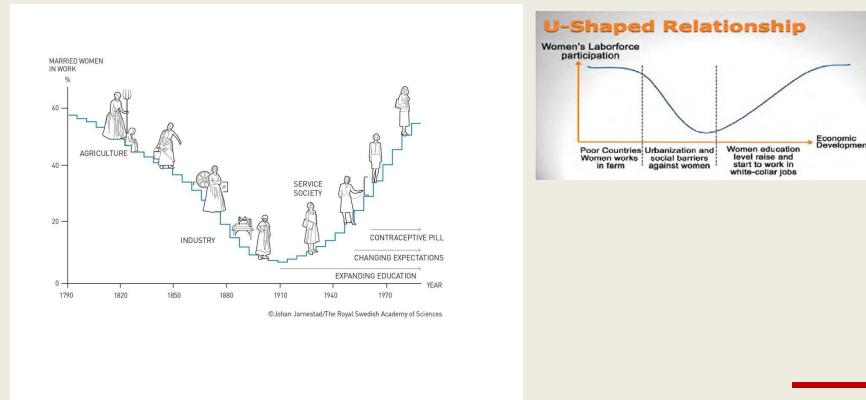


giz

On this slide is a simple time-series visualisation of LFPR in select countries and regions, from 1991 to 2021. The values for India and Upper Middle Income countries are highlighted. A few takeaways:

- As can be seen, overall LFPR in India has not shown significant improvement. In fact, in recent decades, it seems to be declining further, which will be a concern for policy given that Gross Domestic Product (GDP) has been growing quite well (jobless growth?).
- Of special concern to Indian policymakers will be that we have a fast-growing young/youth population, and even so, LFPR is declining.
- Upper middle income countries demonstrate a higher LFPR even with aging populations.

Women and the Economy - Theory by Nobel-winning Economic Historian Claudia Goldin

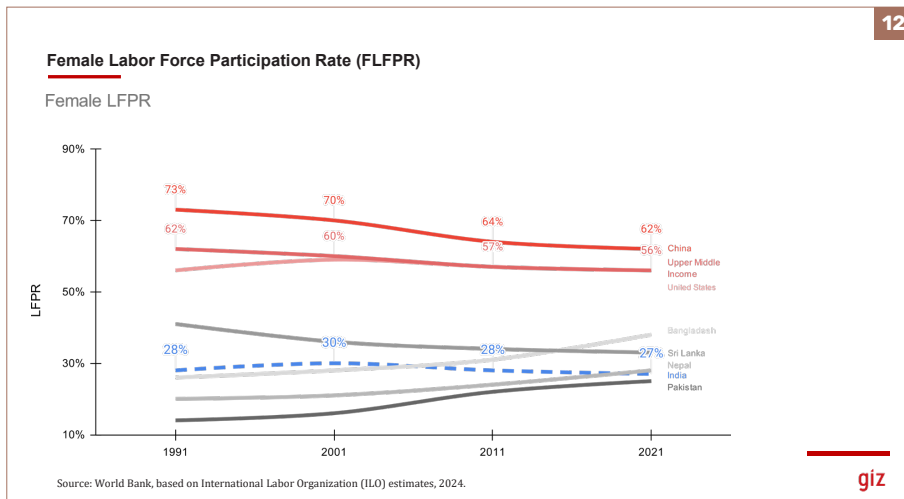


Source: Royal Swedish Academy of Sciences, based on the work of Claudia Goldin & The Hindu October 10, 2023

giz

On this slide, Claudia Goldin's (2023 Nobel Prize winner) hypothesis on women and the economy is illustrated. In the illustration/visualisation on the left, note:

- The y-axis represents the % of married women in work. The x-axis represents major time periods since the beginning of industrialisation.
- In an agricultural economy, women would work on farms and within houses which would require little travel, which Claudia (and many other economists) emphasize should be called labor force participation. If counted, the FLFPR would be quite high in pre-industrial times. As the economy industrialized, women's labor force participation reduced, due to a combination of reasons that "include opportunities for combining paid work and a family, decisions (and expectations) related to pursuing education and raising children, technical innovations, laws and norms, and the structural transformation in an economy." In other words, women were able to access industrial jobs much lesser due to societal, economic and policy reasons. However, as the economy become more service oriented, and better education options emerged, women re-enter the labor force. This is also supported by changing societal expectations and the ability to plan families using contraception.
- A simpler version of the same illustration is presented alongside on the right. Essentially, women's labor force participation is expected to form a U-shape with increasing modernization and economic development.



On this slide, we focus on Female Labor Force Participation Rate (FLFPR) in India. In contrasting with the previous two slides, some takeaways emerge:

- FLFPR in India is much lower, almost half of the LFPR. In fact, since the previous data included women, it is obvious that the overall LFPR is brought down by the FLFPR.
- FLFPR in India is one of the lowest in the world, just marginally above the Arab states and countries with dire economic issues such as Pakistan.
- It is much lower than developed countries such as the United States, as well as upper middle income countries, and also much lower than China.
- Also, FLFPR does not seem to follow the U-shaped trend that is expected in theory. India has massively industrialized even earlier and from the 1990s onward, the GDP is driven increasingly by the service sector and white collar jobs. Continuously declining FLFPR therefore is a concern for research and action, that we will speak about in the following slides.

Why is the FLFPR consistently low in India?

Economists and researchers attribute India's consistently low FLFPR to several overlapping reasons, some of which are listed below:

1. Patriarchy* pushes women into low-quality inferior work related to household operations, caregiving (children, elders, animals, etc.). Many women therefore do not enter the labor market due to the pressures of domestic duties.
2. Women tend to prefer work that is close to home, part time or flexible work, work aligned with their skills, and which has a safe work environment, reducing their job/work options.
3. Lack of tactical support to balance various domestic and professional responsibilities.

Source: Various authors, The Hindu, 2023-24



* Patriarchy has been defined as a social structural phenomenon in which males have the privilege of dominance over females. This supremacy is manifested: in values, attitudes, and customs in the society; in ownership of assets, incomes and wealth; and in institutes and organisations that govern our society and economy. With economic growth and increasing education, the strength of patriarchy has perhaps declined in some ways. However, the overall culture of male dominance over women has not changed much in our traditional society. – Indira Hirway, The Hindu 24 March 2024.

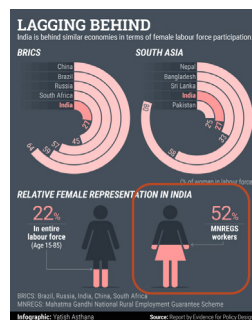
The points related to low FLFPR in India can be explained in the following manner. With a structural societal patriarchy, it is likely that women will not be able to find the work/job options that can balance domestic and professional responsibilities. In other words, in order to overcome the constraints of patriarchy in Indian society, policy would have to look at measures at different levels of socio-economic phenomena. One would be at a very high, strategic level, that could aim to change society's attitude towards women wanting to work, irrespective of their domestic circumstances such as marriage, children, caregiving, etc. This is likely to take a long time. Policy may be better positioned to provide support at a tactical level - helping women balance their responsibilities better. What could some of these measures be, given the broad reasons stated for the low FLFPR?

The employment situation of rural and urban women both look bleak (2022)

	Rural			Urban			Total		
	Male	Female	Total	Male	Female	Total	Male	Female	Total
Youths									
Own-account worker	40.2	25.4	35.2	57.1	60.0	57.8	43.1	29.7	38.7
Employer	1.6	0.5	1.2	5.0	0.9	3.9	2.1	0.6	1.6
Unpaid family worker	58.3	74.1	63.6	37.9	39.1	38.2	54.8	69.8	59.6
Total	100	100	100	100	100	100	100	100	100
Adults									
Own-account worker	86.6	39.1	69.6	79.8	67.3	77.0	85.1	42.7	71.0
Employer	5.5	1.1	3.9	14.2	2.4	11.6	7.4	1.2	5.4
Unpaid family worker	7.9	59.9	26.5	6.0	30.2	11.4	7.5	56.1	23.6
Total	100	100	100	100	100	100	100	100	100

Source: Periodic Labour Force Survey data for 2022.

Source: India Employment Report 2024: Youth employment, education and skills.



This data illustrates that in rural areas, women are mostly engaged in unpaid family work, whether they are youth (15-29) or adults (30-59). In urban areas, the situation is better but not great, as more women are at least own-account workers.

In rural areas, the Mahatma Gandhi Rural Employment Guarantee Scheme (MNREGS) has proved to be a game-changer for women to access paid employment. More than half the workers in MGNREGS are women. But what about in urban areas that are attracting more migrants, including women?

15


The Government Wants to Find Out More, Focus on Women in Urban Areas

Centre starts survey to assess women participation in workforce

The survey is being taken to assess the spread of women employee-friendly practices in the country like creche facilities for children and equal pay for equal work

January 30, 2024, 09:51 pm | Updated 09:58 pm | 105M views

THE HINDU BUREAU



Women workers employe... of a government factory in Trichy district of Tamil Nadu. (Photo Credit: The Hindu)

Source: The Hindu, 30 January 2024

Women Migrating from Rural to Urban Areas

Census 2011



Reason for migration	Total women migrants
Work	1,743,520
Business	177,887
Marriage	21,963,994
Moved with HH	14,417,680
Others	2,607,217
Subtotal	38,936,311

The government is seeking details such as formation of internal complaints committee (ICC) for Prevention of Sexual Harassment at Workplace (POSH), creche facilities for children, equal pay for equal work, flexible or remote working hours for women and transportation facilities during late hours.

giz

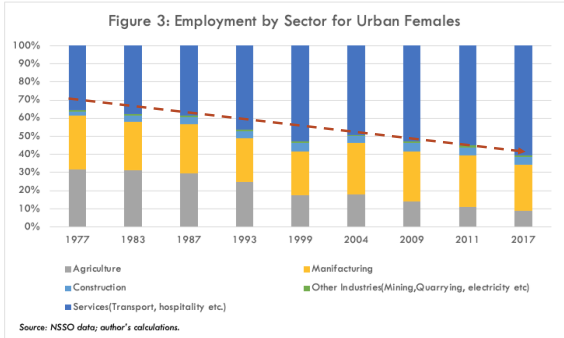
*Along with various researchers and multilateral institutions such as the World Bank and ILO, the government (at the union and state levels) has started inquiring deeper into the question of FLFPR. As can be seen from the newspaper screenshots above, the government is interested in finding out about incidence of sexual harassment, availability of creches, equal pay, flexible working possibilities and transportation options. While these do not clearly address the deeper societal questions of patriarchy, it is fairly clear that the government has a say in these themes, through policies and programs. What are some of these policies and what premise were they based on? We shall discuss this in the following slides.

It is important to point out that a lot of the new jobs and work will emerge in urban areas (which have much higher production based on service sectors), and therefore policy would need to consider support to women in urban areas more closely. In other words, while jobs in rural areas will largely remain agricultural or agri-based, in which women are participating in any case, increasing women's participation in the urban economy is going to be the deeper and more difficult challenge. This approach is critical due to increasing migration into urban areas (graph on the right, which shows that nearly 40 million women were migrants from rural areas to urban areas), due to various social, economic and environmental factors.

16

The Nature of Work that Women in Urban Areas do is Changing

Figure 3: Employment by Sector for Urban Females



Sector	1977	1983	1987	1993	1999	2004	2009	2011	2017
Agriculture	~30%	~28%	~25%	~22%	~18%	~15%	~12%	~10%	~8%
Manufacturing	~25%	~25%	~25%	~25%	~25%	~25%	~25%	~25%	~25%
Construction	~10%	~10%	~10%	~10%	~10%	~10%	~10%	~10%	~10%
Services (Transport, hospitality etc)	~35%	~37%	~40%	~43%	~47%	~50%	~53%	~55%	~57%
Other Industries (Mining, Quarrying, electricity etc)	~0%	~0%	~0%	~0%	~0%	~0%	~0%	~0%	~0%

Source: NSSO data, author's calculations.

Original Source: National Sample Survey Office Reports

giz

In this chart, based on NSSO data, we see that the kind of work done by women in urban areas has been undergoing a transformation, that is in line with the changing nature of work in urban areas of India. Essentially, work has become more service oriented and less based on agriculture. The share of manufacturing has remained more or less constant. Now, in a city or town, service (and manufacturing) jobs/work would not necessarily be close to home, and also require a dedicated amount of time spent on the work itself. In other words, women in the service or manufacturing sectors would need to, a) spend dedicated time away from home and thus not be able to contribute to daily domestic work as much, and b) require commuting to the place of work. Both concern policy and can result in interventions to lessen the burden of domestic chores, as well as lessen the financial burden and increase safety while commuting.

Government Schemes and Programs to Support Women's Work in Urban Areas

Based on surveys, analysis and synthesis of various research studies conducted by economists and the government itself, the Government of India and State Governments have come up with a variety of schemes to address the issue of low FLFPR. The following table illustrates some of these schemes, as responses to assertions from research.

Evidence-based Finding	Government Response
Need to remove the drudgery of domestic work such as cooking with inefficient fuels	Extension of LPG cylinder reach and subsidies
Reduce or remove the burden of compensating for lack of basic services such as water and sanitation	AMRUT and SBM schemes; state-level water supply and sanitation schemes
Reduce burden of child-care and caregiving in general	National Creche Scheme for The Children of Working Mothers
Increase safety and safe, affordable transport options	Free bus schemes by state governments (eg, Telangana and Karnataka)
Lack of appropriate job markets and skill pathways	Skill India Mission with a focus on women; Initial discussions on MNREGS for urban women

Source: Various sources

giz

This is a critical slide. As can be seen, the government, at the union and state level is seized of the matter of low FLFPR. The table in this slide lists some of the main schemes and programs that the government has initiated to, a) ease the typical burdens of women in India, in order to enable them to find the time and energy to look for more productive, paying work, and b) financial incentives such as subsidized fuel and transport to encourage more women to be mobile and find better employment opportunities. In this way, the main takeaways are:

- The government may not be able to easily intervene or impact societal norms and patriarchal mindsets. However, it is focusing on tactical interventions that could help women to participate in the labor force in greater numbers.
- Governments at different levels are able to respond as per their jurisdictions and legislative lists.

At this point, it is critical to note that the connection between evidence and policy is not a linear cause-effect connection. This is due to several reasons, including but not limited to, a) pervasive, persistent mindsets about women working outside the house that can only be dislodged slowly, b) political economy and election manifestos and resulting priorities, and c) competition for resources and trade-offs between various stakeholders. This last point is illustrated in the next slide.

The trade-off Between Resources and Stakeholders in our Political Economy

A spoke in TSRTC's wheel of fortune

Zero-fare travel on State-run buses, courtesy the Congress government's Mahalakshmi scheme, has led to an unprecedented surge in occupancy but, at the same time, underscored the inadequacy of the current fleet. Even as passengers enjoy savings, they stress the need for more buses to address issues of overcrowding and timely services, especially with Sankranti round the corner, finds Syed Mohammed

January 17, 2024 11:24 am | Updated April 09, 2024 02:08 am IST

BY SYED MOHAMMED



A TSRTC white, female-friendly Hyderabad-based bus is seen at a government-sponsored Mahalakshmi scheme, allowing zero-fare travel for girls, women and transgender persons on State-run buses. (Photo Credit: SAMAKRISHNA.G)

Source: The Hindu newspaper, March-April 2024.

Mahalakshmi free bus scheme saves Telangana women ₹ 1,177 crores in four months

April 08, 2024 07:42 pm | Updated April 09, 2024 02:08 am IST - HYDERABAD

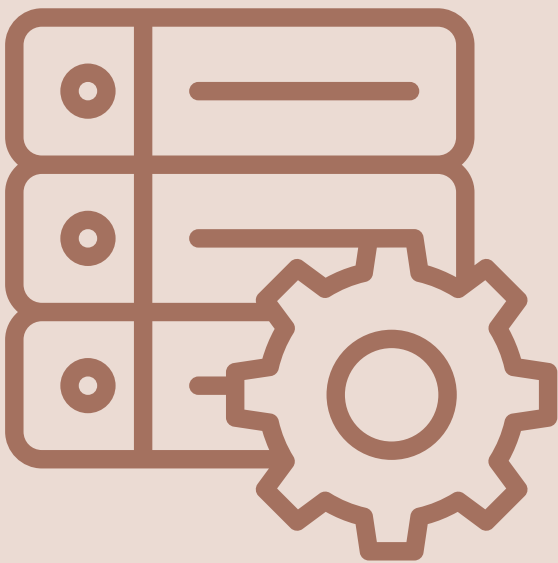
THE HINDU BUREAU



The Mahalakshmi free bus scheme has emerged as a beacon of financial relief for women and transgender persons across Telangana representative photo (Photo Credit: SAMAKRISHNA.G)

giz

On the left, the headline speaks of the resource crunch that the Telangana State Road Transport Corporation (TSRTC) faces due to the free bus scheme of the Government of Telangana. On the right, however, the scheme has saved individual women a massive amount of money, besides providing them with a safe, exclusive option to travel. Newspaper reports (that cannot be considered scheme evaluations) speak of how many more women are now able to access better healthcare services, jobs and options for socio-economic mobility thanks to the scheme, which is something intangible and cannot be valued in simple financial terms.



MODULE 3

DATA INTEGRATION, DASHBOARDS AND DECISION-SUPPORT SYSTEMS

Session 3: Group Exercise - based on gender related datasets
Duration (ideal) 1.5 Hours

Session 3: Group Exercise - based on gender related datasets

SESSION SUMMARY

1	BRIEF SESSION DESCRIPTION	Using a provided dataset (or participants can use their own dataset), participants can be divided into groups of between 4 and 6 people. They have to use the dataset to create a narrative regarding the topic of the data.
2	LEARNING OUTCOMES	At the end of the group exercise, the participants will be able to confidently navigate through given datasets, convert and join them as required and develop the ability to create a data story using tabular and spatial datasets. They will gain confidence in generating maps and charts for the given dataset after appropriate and adequate 'modeling' of the datasets together.
3	CASE STUDIES (IF ANY)	None
4	PRACTICE DATASETS	Folder: Day 3 Group exercises_Datasets <ol style="list-style-type: none">1. Census 2011 House listing Women Headed Households Material of Wall and Roof2. Same as above, Assets3. Same as above, Quality of House4. Female Labour Force Participation Rate; All countries 1991-2023 Access via https://drive.google.com/drive/folders/17vr7FnAZdldPYcukB7WejVV5fsrTd7Ed?usp=sharing
5	FACULTY REQUIREMENT	Experience of data modeling will be useful. That is, the faculty should be able to adequately guide participant groups to summarize and shape the data into easily comprehensible descriptions. Faculty should have a good understanding of visualisation of tabular and spatial data. They should also have the ability to provide feedback in real-time to the efforts being made by the participant groups
6	LEARNER PREREQUISITES	None, if the participants have gone through and grasped the concepts and practices of the previous modules and sessions
7	CLASSROOM ARRANGEMENT	U-Shape or Conference (Refer Annexure 3)
8	TECHNICAL REQUIREMENTS	Laptops/desktops with internet connectivity (at least 1 MBps per machine); Basic software like MS Office and Google Earth Pro; Google accounts if participants choose to use Google apps

Teach the rest of the class !!!

15

Choose the Audience

- Local government elected representatives.
- Local government administrators and technical staff.
- Students at graduate or post-graduate levels.
- Researchers and practitioners
- Others

Choose a concept or tool

- **Concepts:**
 - Any concept from Day 1 (tabular data analysis)
 - Any concept from Day 2 (working with spatial data)
- **Tools:**
 - Vlookup, weighting or any other commands
 - Visualisation of tabular data
 - Google Earth Pro or Datawrapper

giz

Activity Overview:

- **Group Formation:** Divide participants into groups of equal size, ensuring diverse representation in each group.
- **Target Audience Selection:** Each group selects a target audience from the following options:
 - Local government elected representatives
 - Local government administrators and technical staff
 - Graduate or post-graduate students
 - Researchers and practitioners
- **Concept and Tool Selection:** Groups choose a concept and appropriate tools discussed during the training program to explain to their selected audience.

Requirements

15

PLEASE DEFINE

- Learner Prerequisites
- Scope and Learning Outcomes
- Make a quick note on why you chose a certain pedagogical method for bridging:

Content <> Audience

giz

- **Presentation Preparation:** Groups will create a presentation or demonstration customized for their selected audience. This presentation should illustrate the comprehension and application of the chosen concept.
- The presentation must include the following elements:
 - Definition of Learner Prerequisites
 - Scope and Learning Outcomes
 - A brief explanation of the chosen pedagogical method and the rationale behind its selection.

Datasets

PROVIDED (USE OTHERS IF YOU WANT)

- Census 2011 House listing Women Headed Households Material of Wall and Roof
- Same as above, Assets
- Same as above, Quality of House
- Female Labour Force Participation Rate; All countries 1991-2023



giz

- At the end of the group exercise, the participants will be able to confidently navigate through given datasets, convert and join them as required and develop the ability to create a data story using tabular and spatial datasets.
- They will gain confidence in generating maps and charts for the given dataset after appropriate and adequate 'modelling' of the datasets together.

Feedback and Spot Assessment of Learning Outcomes

Feedback is an important step in any capacity building programme, in order to understand whether the planned learning outcomes have been delivered or not, and how well. In this section we lay out the rationale of collecting spot feedback on the last day or immediately after the course.

The key to gathering useful spot feedback is to design a data input mechanism that is easy to understand and intuitive to fill up. Therefore, the form needs to be simple, to the extent that participants, who may have a short amount of time, can easily understand the questions and give quick quantitative responses to them. Keeping this time constraint in mind, the suggested feedback form includes just about 25 questions, including basic identification details.

The questions are structured around trying to understand if the participants feel that they have learned something useful and interesting, and if they perceive that it will be useful to their work lives, professions and positions. This section is referred to as 'General Applicability of the Training Programme'. Responses to this section will be useful for institutions and organizations to understand if and how the programme has relevance for their institutional goals, albeit from an individual's point of view. There is a further section that solicits feedback on specific modules and sessions of the course, that is useful for faculty to understand how the sessions were received and if the overall trajectory of the course was successful in delivering the intended learning outcomes.

However, it is important to note a certain binding imitation of such a spot feedback mechanism. This is simply that spot feedback can be influenced by a number of factors, ranging from the imperative to get back to daily work routines, to interpersonal and social equations, and even a tendency to misrepresent responses deliberately due to institutional pressures. It is therefore highly recommended that spot feedback be supplemented at a later stage with a more in-depth study of changes in professional practice to understand whether there has been any sustained change therein.

For the immediate use of faculty conducting this course, a generic spot feedback form is available in Annex 2.



ANNEXURES

ANNEXURE 1: Capacity Building on Data Analytics and Visualisation

AGENDA

Day 1

Time of the session	Session Name
09:30 – 10:30	Registration and Welcome Address
10:30 – 11:00	Group Photo and/or Tea Break
11:00 – 12:00	<ul style="list-style-type: none">• Introduction to data – types, formats, key terminologies, including an introduction to ‘Big Data’: concepts and methods• Popular software and tools for storing, analyzing, and presenting data Practical <ul style="list-style-type: none">• Familiarization with Microsoft (MS) Excel
12:00 – 13:00	Simple statistical analysis of single datasets <ul style="list-style-type: none">• Descriptive statistics (measures of central tendency, measures of dispersion, skewness, normal distribution, correlation)• Construction of index<ul style="list-style-type: none">• Normalisation• Indicators• Methodologies• Categorisation• Case studies’ outlines and principles<ul style="list-style-type: none">• State Urban Indices (for eg, TN Urban Liveability Index 2023)• National Urban Outcomes Framework by NIUA• SDG indicators at city, state and national levels, with a focus on SDG 11 (cities) and SDG 5 (gender) Practical <ul style="list-style-type: none">• Overview of making an index in excel
13:00 – 14:00	Lunch
14:00 – 15:15	Visualisation-assisted analysis of tabular data <ul style="list-style-type: none">• Choosing the right visualisation<ul style="list-style-type: none">• Compositions• Relationships• Comparisons• Distributions Various visualisation techniques <ul style="list-style-type: none">• Pie charts• Scatter plots• Bubble charts• Bar and line charts Practical <ul style="list-style-type: none">• Application of visualisation techniques on sample dataset using MS Excel
15:15 – 15:45	Tea Break
15:45 – 17:00	Integrating and summarizing datasets <ul style="list-style-type: none">• Data integration overview• Requirement for data integration (keys, shared attributes, etc.)• Merge and join (horizontal vs. vertical)• Data integration in excel (lookup, pivot table) Practical <ul style="list-style-type: none">• Application of data integration on sample datasets using MS Excel

Day 2

Time of the session	Session Name
09:30 – 11:00	<p>Introduction to GIS, remote sensing, types of geospatial data</p> <p>Presentations on</p> <ul style="list-style-type: none">• What is GIS?• What is remote sensing?• What is a vector?• What is a raster?• Spectral, spatial, and temporal resolution• What are attributes and spatial data formats?• How representation differ with scales• Remote sensing and ‘big data’ <p>Hand on session on Google Earth Pro to explore</p> <ul style="list-style-type: none">• Time series satellite data• Creating point, line, and polygon and exporting them• Exploring terrain tool and measurement tools• Visualisation
11:00 – 11:30 Tea Break	
11:30 – 13:00	<p>Spatial data visualisation</p> <p>Presentations on</p> <ul style="list-style-type: none">• Map reading• Symbology and elements of a map: Vector data classification types (Graduated, Categorized, Size, Color schemes, Raster data visualisation) <p>Hands on session on using Datawrapper and ArcGIS online to import, visualize vector data and sharing them as interactive web maps</p>
13:00 – 14:00 : Lunch	
14:00 - 15:30	<p>Working with vector data: integrating tabular datasets with spatial data and spatial data creation</p> <p>Presentations on</p> <ul style="list-style-type: none">• Data creation process – geocoding, georeferencing and digitization• Joining tabular data with vector data• Spatial data join (attributes by location) <p>Hands on session on geocoding tabular datasets using Google Sheets, joining tabular data with spatial data, and spatial join of two vector datasets</p>
15:15 - 15:45 : Tea Break	
15:45 - 17:00	<p>Integrating and summarizing datasets</p> <ul style="list-style-type: none">• Data integration overview• Requirement for data integration (keys, shared attributes, etc.)• Merge and join (horizontal vs. vertical)• Data integration in excel (lookup, pivot table) <p>Practical</p> <ul style="list-style-type: none">• Application of data integration on sample datasets using MS Excel
15:30 – 16:00 Tea Break	
16:00 - 17:00 or	<p>Open data</p> <ul style="list-style-type: none">• Overview of various open data sources for vector and raster data <p>Demonstration</p> <ul style="list-style-type: none">• Downloading vector data from OpenStreetMap (OSM) and downloading remote sensing indices from Sentinel hub
16:00 - 17:00	<p>Catch-up session with assisted exercises on mapping platforms</p>

Day 3

Time of the session	Session Name
09:30 – 11:00	Data ‘dashboards’ <ul style="list-style-type: none">• Designing data dashboards for various audiences• Common methods and tools available for making data dashboards, focus on GIS platforms• Case 1: Leveraging map-based integrated data visualisation platforms for national level master plans of infrastructure such as the Gati Shakti program• Case 2: Coimbatore SDG Dashboard
11:00 – 11:30 TEA BREAK	
11:30 – 13:00	Data analysis for evidence-based policy <ul style="list-style-type: none">• Evidence-based policy – Result Based Management (RBM)• Gap analysis, situation analysis, distance frontier approach
13:00 – 14:00 LUNCH	
14:00 - 15:15	Focused and assisted group exercise session and hands on experience on platforms and tools covered in the course <ul style="list-style-type: none">• Creating/understanding indices• Data integration• Data cleaning• Design and presentation
15:15 - 15:45 : Tea Break	
15:45 - 16:30	Presentations by groups on data analysis and visualisation
16:30 – 17:00	Summary, Valedictory and Vote of Thanks

Note: At the end of each day, a conclusion session (preferably of 15 mts duration) must be included to summarise the proceedings of the whole day training.

ANNEXURE 2

Feedback Form - Course on Data Analytics and Visualization

* Indicates required question

1. Name *

2. Gender *

Mark only one oval.

Female

Male

Prefer not to say

Other

3. Age *

4. Email address *

5. Mobile no. *

6. Designation *

7. Organization *

8. District (you work in) *

General Applicability of Training Content

9. How satisfied are you with the overall quality of the training content? *

Mark only one oval.

1 2 3 4 5

Highly Highly satisfied

10. How well did the training content align with your expectations and needs? *

Mark only one oval.

1 2 3 4 5

Not Highly aligned

11. Has the training resulted in improved understanding of the subject? *

Mark only one oval.

1 2 3 4 5

No i Significant improvement

12. To what extent do you believe the training content is relevant and applicable *
to your specific job role and responsibilities.

Mark only one oval.

1 2 3 4 5

Not Highly relevant

13. To what extent will the training content be able to help you in addressing the *
challenges you face in your job.

Mark only one oval.

1 2 3 4 5

Not Extremely helpful

14. Did the training provide you with the necessary tools and resources to apply *
the knowledge and test your knowledge effectively?

Mark only one oval.

1 2 3 4 5

Not Completely

15. How satisfied are you with the support and guidance provided by the trainers *
during the training dissemination?

Mark only one oval.

1 2 3 4 5

Very Very satisfied

16. How frequently do you think you will apply knowledge and skills gained from the training in your professional work? *

Mark only one oval.

1 2 3 4 5

Never Apply frequently

17. How confident do you feel in applying the newly acquired skills and knowledge in real life situations at work? *

Mark only one oval.

1 2 3 4 5

Not Very confident

18. Do you anticipate any improvements in your performance or productivity because of the training received? *

Mark only one oval.

1 2 3 4 5

No improvement Significant improvement

19. Do you provide your consent for training providers to reach out to you after 6-months for a survey to assess applicability of training content? *

Mark only one oval.

Yes

No

Feedback on specific modules of the program

20. After the training, how confident do you feel about doing basic statistical analytics and visualization with tabular data? *

Mark only one oval.

1 2 3 4 5

Not Quite confident

21. After the training, how confident you feel to describe different types of tabular and spatial data? *

Mark only one oval.

1 2 3 4 5

Not Quite confident

22. After the training, how confident do you feel about using spatial data for some of your work? *

Mark only one oval.

1 2 3 4 5

Not Quite confident

23. After the training, how confident do you feel about joining different types of datasets? *

Mark only one oval.

1 2 3 4 5

Not Quite confident

24. After the training, how confident do you feel about describing ways in which data can be used for policy analysis and decision-making? *

Mark only one oval.

1 2 3 4 5

Not Quite confident

25. Describe a few more concepts, methods and tools you would like to learn in the future, in order to help you work with data in your role.

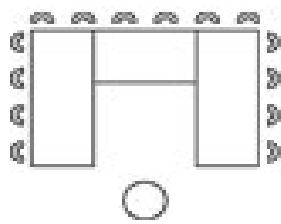
This content is neither created nor endorsed by Google.

Google Forms

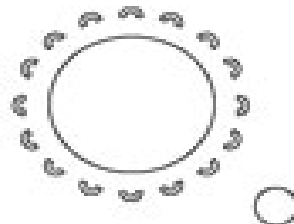
ANNEXURE 3

Room Typology - Seating

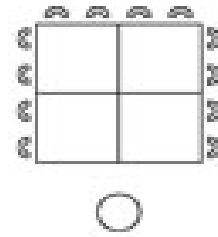
One of the critical aspects of this training model is the seating arrangement, which plays a pivotal role in promoting interaction, maintaining focus, and ensuring effective communication among the participants. Here are some options for the seating arrangement in a Training of Trainers session:



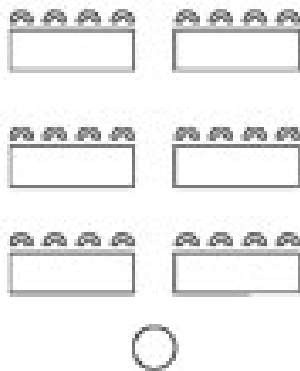
U-shape



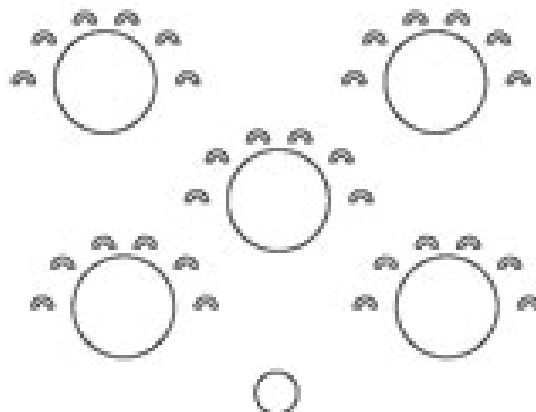
Single square or round



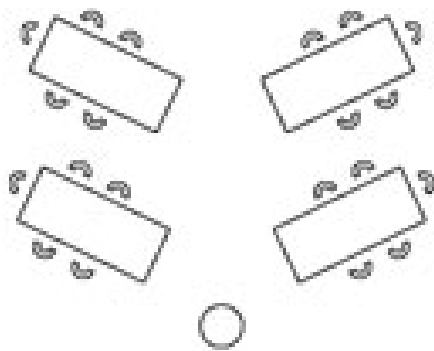
Conference



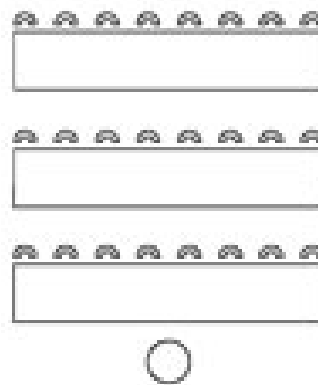
Classroom



Clusters



V-shape



Traditional classroom

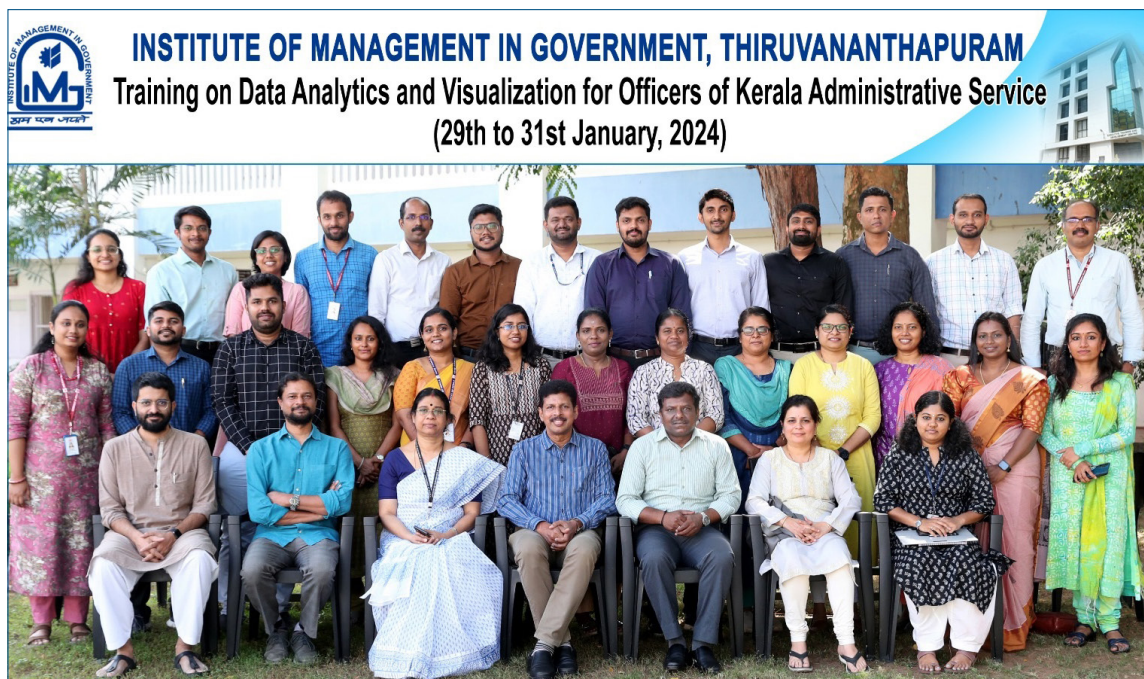
Capacity Building training programs on Data Analytics and Visualisation – conducted based on this course modules

The first capacity development training on “Data Analytics and Visualisation”, was organised at the Institute of Management in Government (IMG), Thiruvananthapuram, Kerala from 25th-27th July 2023, for 28 officials representing various departments of the State.



Images from the first capacity development training organised at the Institute of Management in Government (IMG), Thiruvananthapuram, Kerala from 25th-27th July 2023

The second capacity building training program was conducted on Institute of Management (IMG), Kerala for a total of 90 KAS Officers in 3 batches, starting from 29th January 2024 -3rd February 2024



The second capacity building training program was conducted on Institute of Management (IMG), from 29th January 2024 -3rd February 2024

The third capacity building training on “Data Analytics and Visualisation”, was organised at the Anna Administrative Staff College (AASC), Chennai, Tamil Nadu from 20th - 22nd February 2024 for 25 officials (approx.) representing various departments of the State



Images from the third capacity building training organised at Anna Administrative Staff College (AASC), Chennai, Tamil Nadu from 20th-22nd February 2024



National Institute of Urban Affairs

NATIONAL INSTITUTE OF URBAN AFFAIRS

1st Floor, Core 4b, India Habitat Centre, Lodhi Road, New Delhi - 110003, India
Phone: (+91 11) 24634971, 24643576 | Website: <https://niua.in/>