Fundamentals of Data Science (24BTELY107)
Semester - I/II

Module I

Introduction to Data Science

Definition—Big Data and Data Science Hype—Datafication—Data Science Profile—Meta Data—Definition—Data Scientist—Statistical Inference—Populations and Samples—Populations and Samples of Big Data—Modelling-Data Warehouse—Philosophy of Exploratory Data Analysis—The Data Science Process—A Data Scientist's Role in this Process Case Study: Real Direct—Housing Market Analysis

1.1 Introduction: Definition

- Data science is a deep study of the large amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.
- Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems. It is the future of artificial intelligence.

- Data Science is about data gathering, analysis and decision-making.

- Data Science is about finding patterns in data, through analysis, and make future predictions.

- By using Data Science, companies are able to make:

- Better decisions (should we choose A or B)
- Predictive analysis (what will happen next?)
- Pattern discoveries (find pattern, or maybe hidden information in the data)
- Data science is the combination of: statistics, mathematics, programming, and problem-solving; capturing data in ingenious ways; the ability to look at things differently; and the activity of cleansing, preparing, and aligning data. This umbrella term includes various techniques that are used when extracting insights and information from data.

❖ Need for Data Science

- Data Science is used in many industries in the world today, e.g. banking, consultancy, healthcare, and manufacturing.
- For route planning: To discover the best routes to ship
- To foresee delays for flight/ship/train etc. (through predictive analysis)
- To create promotional offers
- To find the best suited time to deliver goods
- To forecast the next years revenue for a company
- To analyze health benefit of training

- To predict who will win elections

- Data Science can be applied in nearly every part of a business where data is available. Examples are:

- Consumer goods
- Stock markets
- Industry
- Politics
- Logistic companies
- E-commerce

- A Data Scientist requires expertise in several backgrounds:

- Machine Learning
- Statistics
- Programming (Python)
- Mathematics
- Databases

- Data science is all about:

- Asking the correct questions and analyzing the raw data.
- Modeling the data using various complex and efficient algorithms.
- Visualizing the data to get a better perspective.
- Understanding the data to make better decisions and finding the final result.

Fig 1.1 Data Science basics

❖ **Example:**

- Let suppose we want to travel from station A to station B by car.
- Now, we need to take some decisions such as which route will be the best route to reach faster at the location, in which route there will be no traffic jam, and which will be cost-effective.
- All these decision factors will act as input data, and we will get an appropriate answer from these decisions, so this analysis of data is called the data analysis, which is a part of data science.

1.2 Big Data

- Big data refers to significant volumes of data that cannot be processed effectively with the traditional applications that are currently used.
- The processing of big data begins with raw data that isn't aggregated and is most often impossible to store in the memory of a single computer.
- Big data is used to analyze insights, which can lead to better decisions and strategic business moves.
- Big data is high-volume, and high-velocity or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.
- Big data is a combination of structured, semi-structured and unstructured data that organizations collect, analyze and mine for information and insights.
- Companies use big data in their systems to improve operational efficiency, provide better customer service, create personalized marketing campaigns and take other actions that can increase revenue and profits.
- Businesses that use big data effectively hold a potential competitive advantage over those that don't because they're able to make faster and more informed business decisions.
- Medical researchers use big data to identify disease signs and risk factors.
- Doctors use it to help diagnose illnesses and medical conditions in patients.
- In addition, a combination of data from electronic health records, social media sites, the web and other sources gives healthcare organizations and government agencies up-to-date information on infectious disease threats and outbreaks.
- Big data helps oil and gas companies identify potential drilling locations and monitor pipeline operations. Likewise, utilities use it to track electrical grids.
- Financial services firms use big data systems for risk management and real-time analysis of market data.
- Manufacturers and transportation companies rely on big data to manage their supply chains and optimize delivery routes.

- Government agencies use bug data for emergency response, crime prevention and smart city initiatives.
    - Big data refers to significant volumes of data that cannot be processed effectively with the traditional applications that are currently used. The processing of big data begins with raw data that isn't aggregated and is most often impossible to store in the memory of a single computer.
    - Data analytics is the science of examining raw data to reach certain conclusions. Data analytics involves applying an algorithmic or mechanical process to derive insights and running through several data sets to look for meaningful correlations. It is used in several industries, which enables organizations and data analytics companies to make more informed decisions, as well as verify and disprove existing theories or models. The focus of data analytics lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows.
    - Data science enables companies not only to understand data from multiple sources but also to enhance decision making. As a result, data science is widely used in almost every industry, including health care, finance, marketing, banking, city planning, and more.
    - The term data science is so widely used today that its definition has become blurry. Some associate it with computer science and some with statistics; most frequently, it is linked to machine learning (ML) and data mining.

Big Data:

Fig 1.2 Big Data

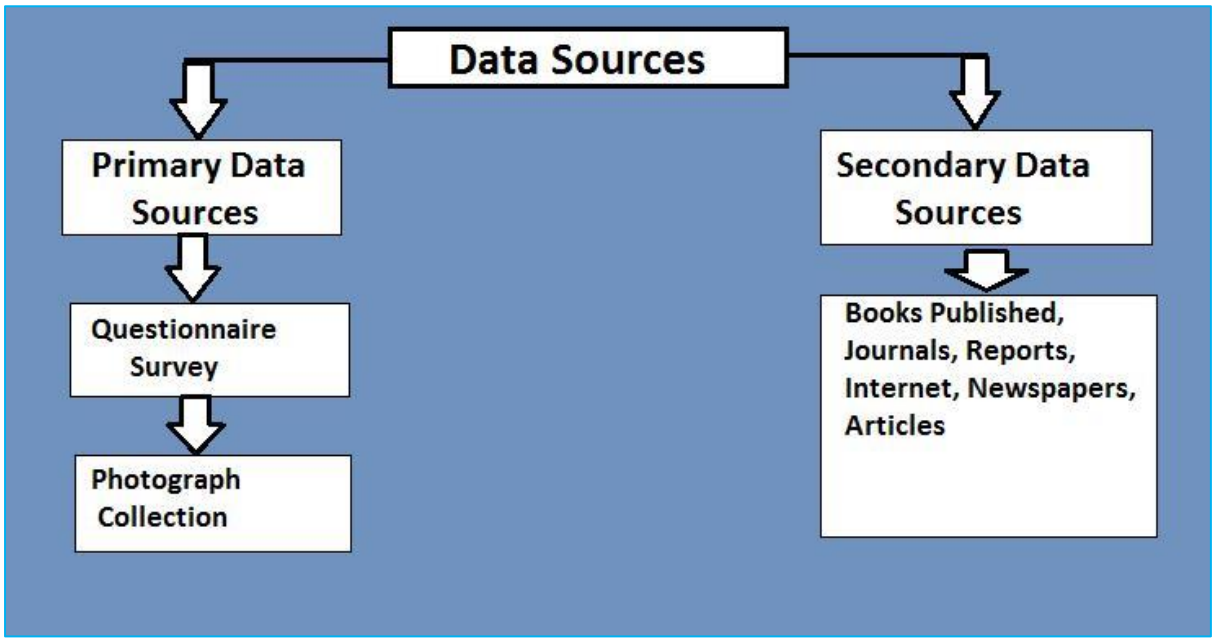


Fig 1.3 Data Sources

- Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis.
- In the process of big data analysis, “Data collection” is the initial step before starting to analyse the patterns or useful information in data. The data which is to be analysed must be collected from different valid sources.

The actual data is then further divided mainly into **two types** known as:

1. Primary data
2. Secondary data

❖ Applications of Big Data

Big Data for Financial Services

- Credit card companies, retail banks, private wealth management advisories, insurance firms, venture funds, and institutional investment banks all use big data for their financial services.
- The common problem among them all is the massive amounts of multi-structured data living in multiple disparate systems, which big data can solve.
- Big data is used in several ways, including:

➢ Customer analytics
➢ Compliance analytics
➢ Fraud analytics
➢ Operational analytics

❖ **List of types of Big Data**:

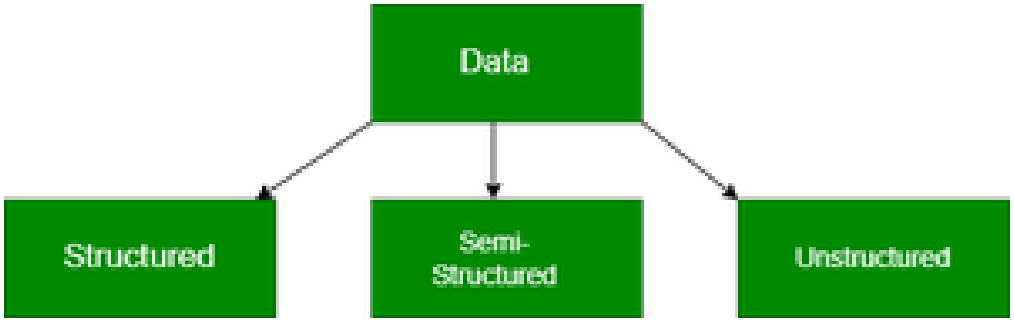


Fig 1.4 Types of Data

- **Structured Data**:

  - This type of data is highly organized and easily searchable by basic algorithms. It is often stored in relational databases and can be represented in tables with rows and columns.
  - Examples: SQL databases, spreadsheets, and data from CRM systems.

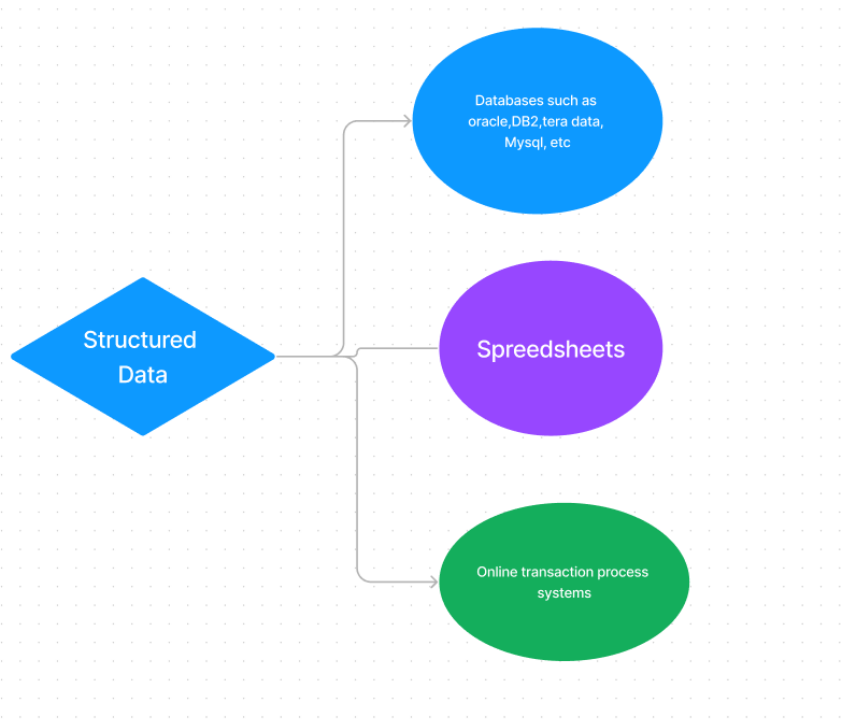


Fig 1.5 Structured Data

• **Unstructured Data**:

- This type of data does not have a pre-defined data model or is not organized in a pre-defined manner. It is more challenging to collect, process, and analyze.
- Examples: Text documents, emails, social media posts, videos, images, and sensor data.

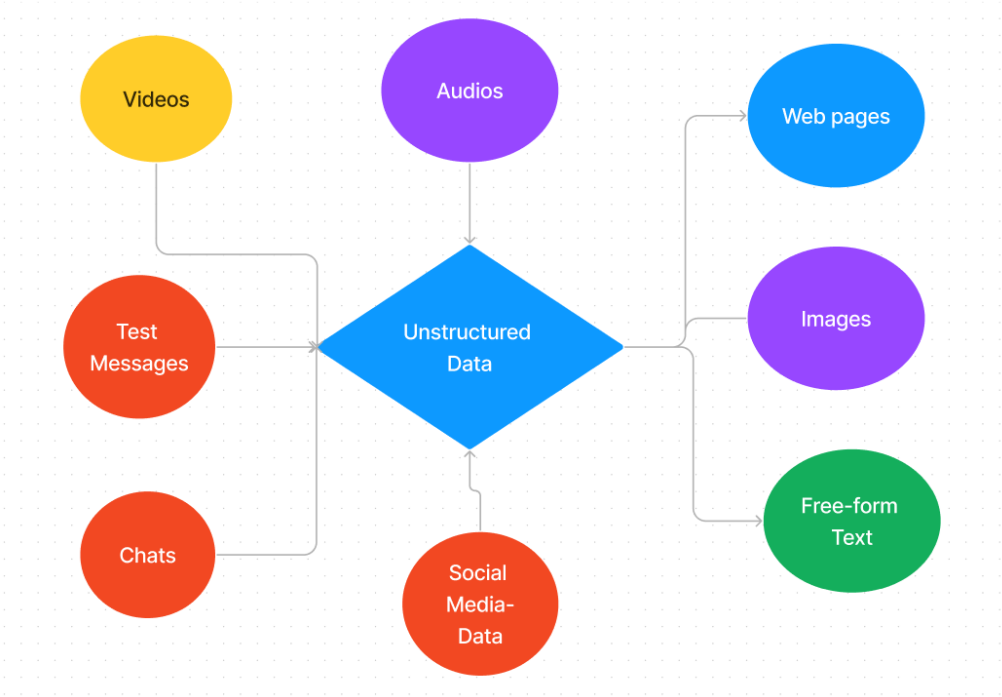


Fig 1.6 Unstructured data

• **Semi-structured Data**:

- This type of data does not conform to the formal structure of data models but contains tags or other markers to separate semantic elements. It is more flexible than structured data but easier to organize than unstructured data.
- Examples: XML files, JSON documents, and NoSQL databases.

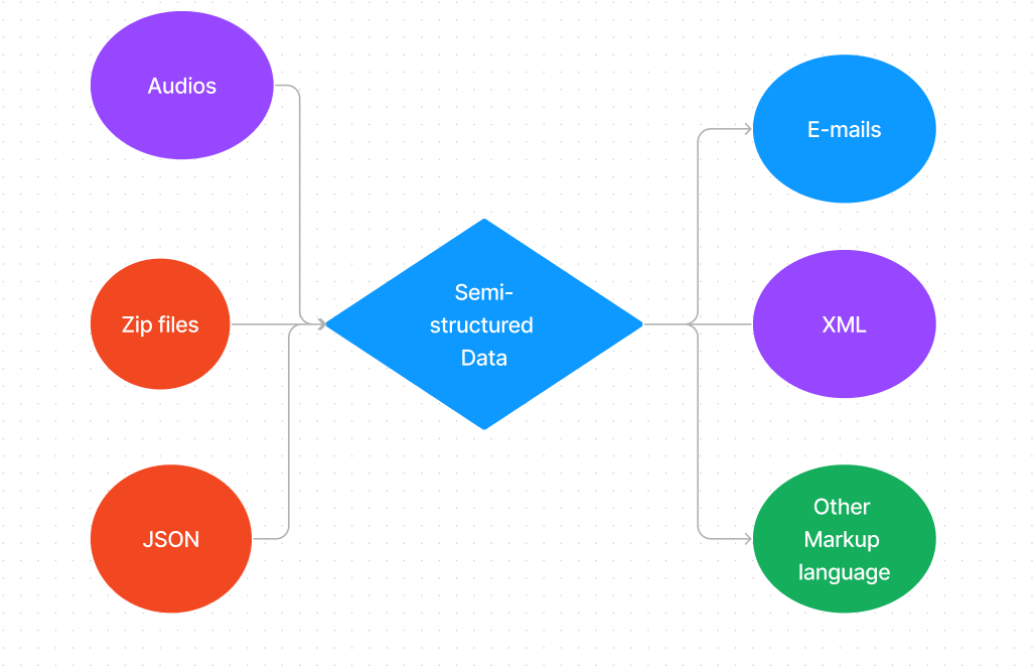
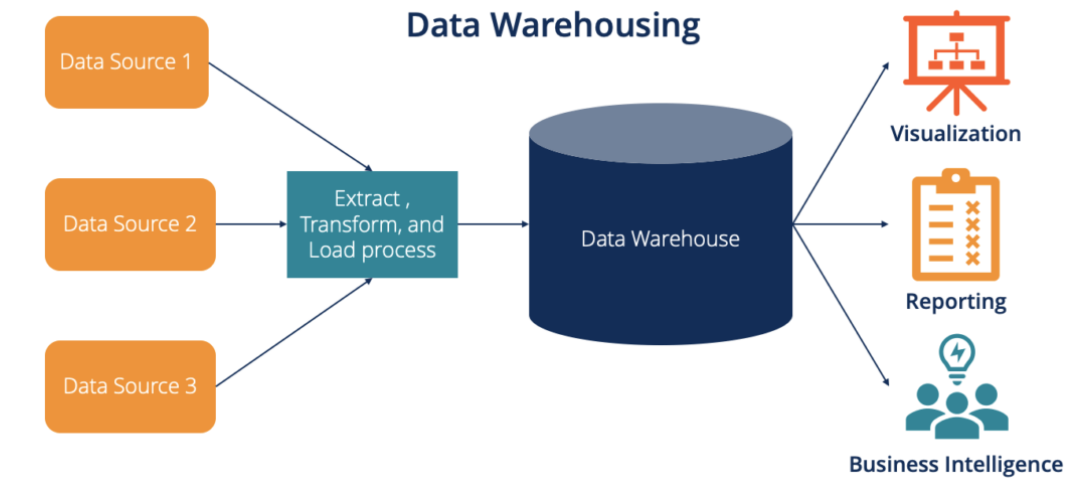


Fig 1.7 Semi structured Data

❖ Data Warehouse:



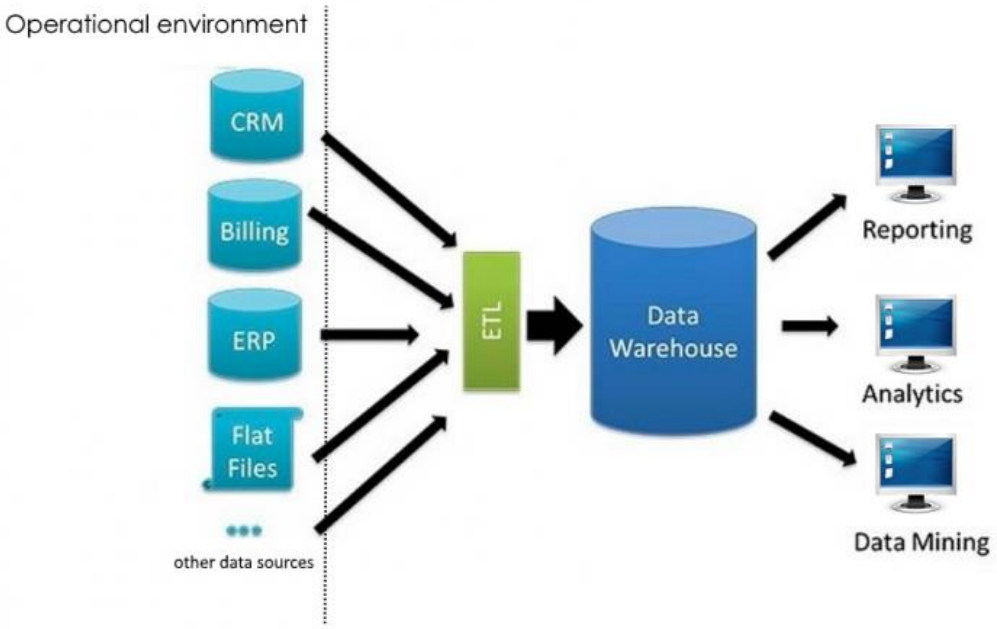


Fig 1.8 Data warehouse and application

A data warehouse is an enterprise system used for the analysis and reporting of structured and semi-structured data from multiple sources, such as point-of-sale transactions, marketing automation, customer relationship management, and more. A data warehouse is suited for ad hoc analysis as well custom reporting.

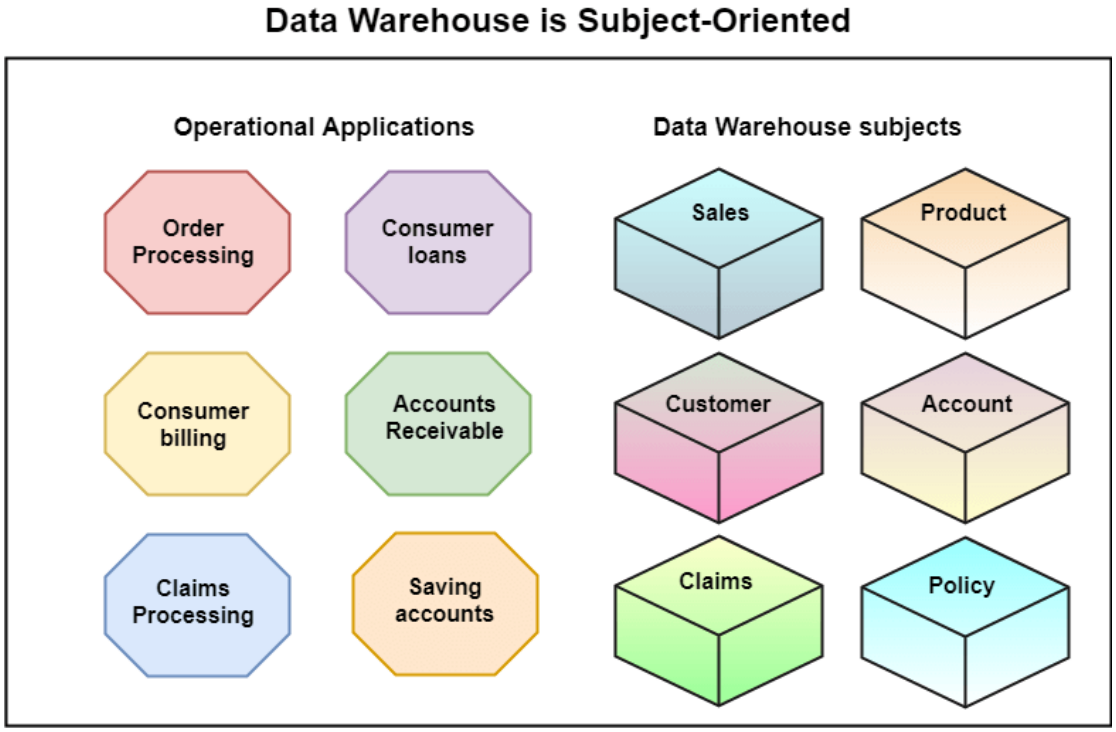


Fig 1.9 Data warehouse and applications

Goals of Data Warehousing

- To help reporting as well as analysis
- Maintain the organization's historical information
- Be the foundation for decision making

Example:

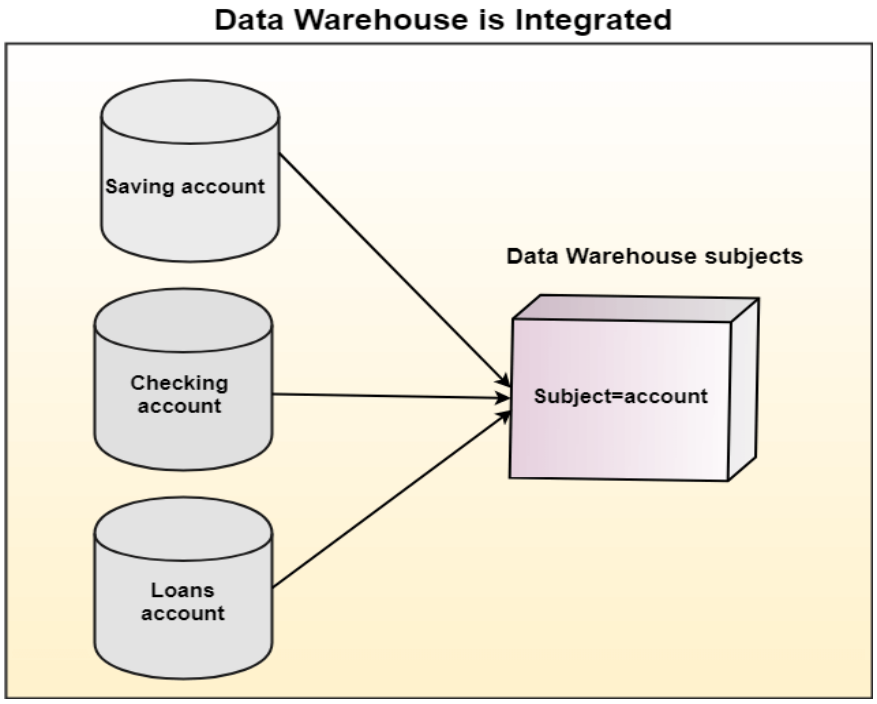


Fig 1.10 Integrated Data warehouse

❖ Definition of Data Preparation:

Data preparation follows a series of steps that starts with collecting **the right data, followed by cleaning, labeling, and then validation and visualization**.

Data preparation is an important step in data analytics as well as in business intelligence. It's also a core function of business analysts.

Raw data is usually collected from multiple sources. For example, if you want to analyze customer behavior, your raw data might come from your company's sales database and customer relationship management (CRM) system. In this case, sales records would be stored in a sales table while customer information would be stored in a customer table. These two tables would probably have identical fields (or columns), but they would contain different values. It's your job as a data scientist to combine these two tables into one big table so you can determine which customers bought what products and how much they paid for those products.

The data preparation process is often overlooked in the rush to find insights in new data sources. It's a tedious, time-consuming task, but it's important to implementing a data-driven business strategy.

Data preparation is the process of collecting, transforming and enriching raw data to make it suitable for analysis. This can include cleaning, normalizing and consolidating raw data sets, as well as adding new dimensions or attributes to the data.

Data preparation is sometimes referred to as "data wrangling."

1.3 Data Science Hype

- Data science enables companies not only to understand data from multiple sources but also to enhance decision making.
- As a result, data science is widely used in almost every industry, including health care, finance, marketing, banking, city planning, and more.
- Data Science has been recognized as one of the most exciting IT domains in the world. Recent reports suggest that leading countries like USA and China are investing billions of dollars to integrate their industries with Artificial Intelligence (AI).
- AI is emerging and being applied across various areas including finance, healthcare, manufacturing etc.
- The hype is crazy—people throw around tired phrases straight out of the height of the pre-financial crisis era like “Masters of the Universe” to describe data scientists.

- Statisticians already feel that they are studying and working on the “Science of Data.”
- The media often describes data science in a way that makes it sound like as if it’s simply statistics or machine learning in the context of the tech industry.

Finance

- Each day, many financial institutions handle multi-billion monetary transactions including ATM withdrawals, debit/credit card payments, deposits, online

payments, and so on. It is also evident that considerable amount of fraud and corruption take place hindering financial growth of the country.

- Modern ML (Machine Learning) techniques and analytics in combination with human-based skills are being adopted to curb mishaps. The system recognizes potential threats and flag them as fraudulent in order to minimize losses beforehand. AI techniques not only create early warning signs but also reduce human errors, thereby increasing efficiency.

Healthcare

- Presence of Artificial Intelligence in the healthcare domain helps health managers to analyze simple to most complicated medical conditions. It allows them to examine symptoms, diagnose diseases, and even suggest medical treatments.
- The medical industry is applying AI concepts to enhance their accuracy and bringing improvements. Apart from diagnosing diseases and suggesting treatments, both AI and ML algorithms are being utilized to improve the healthcare quality as well as cutback on high-end medical costs.

Manufacturing

- Manufacturing is one of the most vital industries in our country. However, they constantly face challenges while maintaining logistics, product forecasting, and supply chain management as well.
- Manufacturing companies can enhance their efficiency through automation with AI and machine learning.

Agriculture

- A noticeable presence of AI has also been seen in the field of agriculture.
- Agriculture takes the help of artificial intelligence to improve production and minimize wastages.
- Farmers are integrating traditional farming practices with AI to automate the processes.

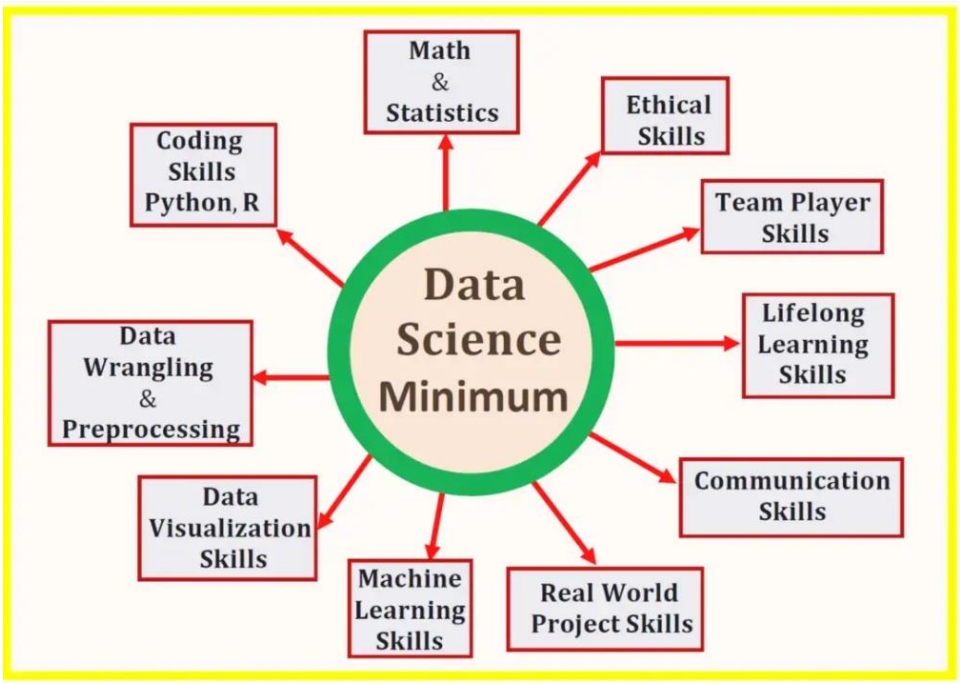


Fig 1.11 Data Science Minimum

❖ Summary of **DATA SCIENCE APPLICATIONS AND EXAMPLES**

- **Healthcare:** Data science can identify and predict disease, and personalize healthcare recommendations.
- **Transportation:** Data science can optimize shipping routes in real-time.
- **Sports:** Data science can accurately evaluate athletes' performance.
- **Government:** Data science can prevent tax evasion and predict incarceration rates.
- **E-commerce:** Data science can automate digital ad placement.
- **Gaming:** Data science can improve online gaming experiences.
- **Social media:** Data science can create algorithms to pinpoint compatible partners.

- **Fintech:** Data science can help create credit reports and financial profiles, run accelerated underwriting and create predictive models based on historical payroll data.

1.4 Datafication

- Datafication as a process of “taking all aspects of life and turning them into data.” As examples, they mention that “Google’s augmented-reality glasses datafy the gaze. Twitter datafies stray thoughts. LinkedIn datafies professional networks.”
- Datafication is an interesting concept and led us to consider its importance with respect to people’s intentions about sharing their own data.
- Datafication refers to the collective tools, technologies, and processes used to transform an organization into a data-driven enterprise.
- An organizational trend of defining the key to core business operations through a global reliance on data and its related infrastructure.
- Datafication refers to the fact that daily interactions of living things can be rendered into a data format and put to social use.

Benefits of Datafication

- Datafication is a technique that is financially advantageous to pursue since it provides great opportunity for streamlining corporate procedures.
- Datafication is a cutting-edge process for creating a futuristic framework that is both secure and inventive.

1.5 Data Science Profiles/Data scientist

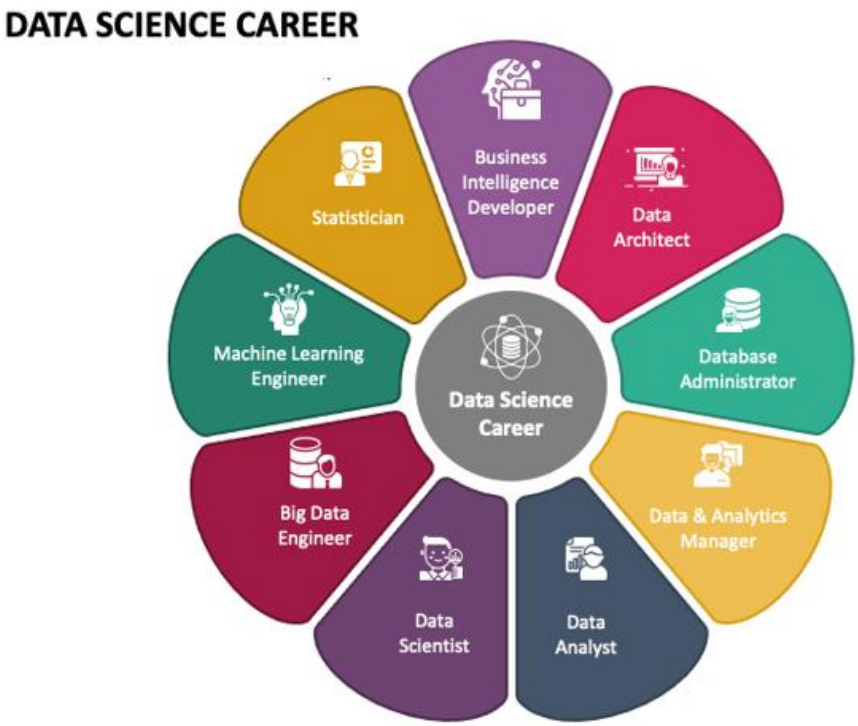


Fig 1.12 Data Science Career

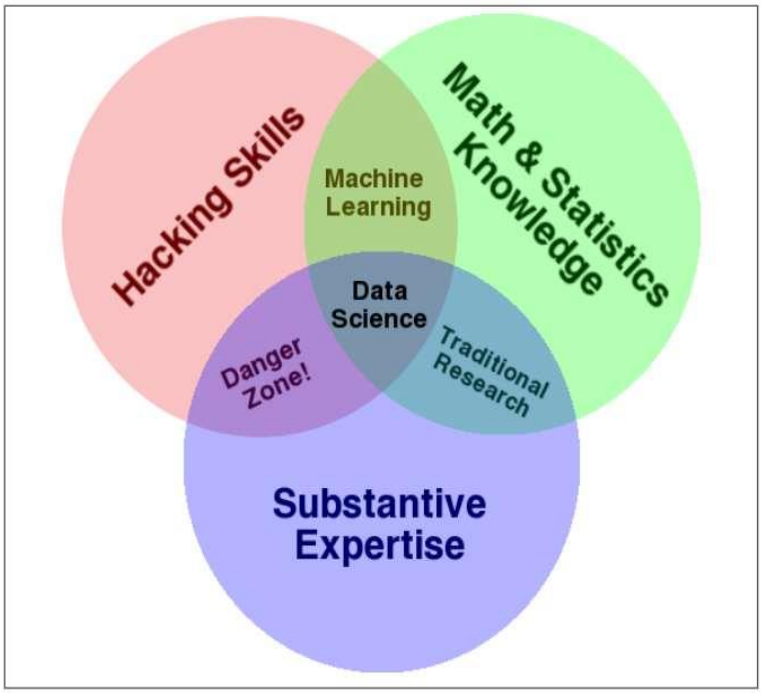


Fig 1.13 Drew Conway's Venn diagram of Data Science

- Data Science is such a broad field that includes several subdivisions like data preparation and exploration; data representation and transformation; data visualization and presentation; predictive analytics; machine learning, etc.

- The principal purpose of Data Science is to find patterns within data. It uses various statistical techniques to analyze and draw insights from the data. From data extraction, wrangling and pre-processing.

- Data Science Job Profiles:

1. Data Analyst

   Data Analysts are the individuals who are responsible for reviewing the data so that they can identify the key information in the businesses of customers. Therefore, it is the process of collecting, processing, and analyzing the data to extract meaningful insights and also data analyst support in decision-making processes.

   - Key Responsibilities

- To maintain the collected data in a simple form and prepare the data for business communication.
- They use a statistical approach to visualize and produce the reports.
- Data analysts assess and understand the trends and patterns, and also evaluate the big datasets.

2. Data Scientist

   Data Scientist are the individual who uses the data to understand it. Therefore these data scientist are responsible to collect, analyze and interpret the data to help to drive the decision making.

- Key Responsibilities

- Data scientist is used to discover the data sources, analyze the information which based on the patterns and trends.
- They automate the procedure of data collection and works on the data pre-processing on the structured and unstructured data.
- Data scientist generates the predictive models and builds the machine learning algorithm.

3. Data Engineer

Data Engineer refers to experts who are responsible for maintaining, designing and optimizing the data infrastructure for the data management and transform them.

- Key Responsibilities

- Data Engineers are responsible for creating and optimizing the data sets for data business and scientists.
- They suggest improvements to enhance reliable and quality of the models and dataset.
- Data engineers develop the algorithms and the prototypes to convert those data into some useful insights.

4. Business Analyst

Business Analyst are the people's who help the organization to fulfil their goals and also assess the organization, analyze the data and improve the systems and processes for the future.

- Key Responsibilities

  - Business analyst conduct several researches to evaluate in the business models and
  - Business analyst develop innovative solutions for the difficult business problems.
  - They are expert in allocating forecasting, budgeting and resources in the businesses.

5. Machine Learning Engineer

   Machine learning Engineer refers to the critical members of the data science team. These engineer tasks are building, researching and designing the AI which are further responsible for the machine learning and improving and maintaining the existing the systems of artificial intelligence.

   - Key Responsibilities

     - Machine learning engineer helps in creating and designing the machine learning systems.
     - They develop the data pipelines and the effective ML models and datasets.

1.6 Definitions: Meta-data, Statistical inference, Populations and Samples:

- **Metadata** means "data about data". Metadata is defined as the data providing information about one or more aspects of the data; it is used to summarize basic information about data that can make tracking and working with specific data easier.

- There are three main **types of metadata**: Descriptive, Administrative, and Structural.

1. **Descriptive metadata** enables discovery, identification, and selection of resources. It can include elements such as title, author, and subjects.

Examples of Descriptive Metadata

- Library Catalogs: Metadata about books, including title, author, publication date, subject headings, and ISBN.
- Digital Repositories: Descriptive information about digital objects, such as datasets, images, and documents, including titles, creators, descriptions, and formats.
- Archives: Metadata for archival materials, such as personal papers, photographs, and historical documents, including descriptions, dates, and contributors.

2. **Administrative metadata** facilities the management of resources.
   - Administrative metadata is a type of metadata that helps manage and support the use of a resource, typically digital objects, throughout its lifecycle. It encompasses information needed for managing, preserving, and providing access to the resource. Here are the main components and purposes of administrative metadata:

(i) Technical Metadata: Describes the technical characteristics of the digital resource, such as file format, creation date, software and hardware used, and technical requirements for accessing and rendering the file.

(ii) Preservation Metadata: Contains information necessary for the long-term preservation of the resource, including details about the provenance (origin and history), any actions taken to preserve it (e.g., migrations, format conversions), and conditions or requirements for maintaining its usability over time.

(iii) Rights Management Metadata: Includes details about the legal and access rights associated with the resource, such as copyright status, intellectual

property rights, access permissions, restrictions, and any licensing information.

- Examples of Administrative Metadata
  - Digital Libraries: Metadata about digitized books, manuscripts, and other resources, including technical details, preservation actions, and rights information.
  - Archives: Metadata for archival collections, detailing provenance, custodial history, and access permissions.
  - Repositories: Metadata for datasets, software, and other digital objects, including technical specifications, usage statistics, and licensing information.

3. **Structural Metadata**: structural metadata is metadata that describes the structure, type, and relationships of data. For example, in a SQL database, the data is described by metadata stored in the Information Schema and the Definition Schema.

- Examples of Structural Metadata

  - **Books and Documents**: Information about chapters, sections, and sub-sections, as well as pagination and links between different parts of the document.
  - **Multimedia Objects**: Metadata describing scenes, segments, tracks, or frames in videos and audio files.
  - **Websites**: The organization of web pages, including navigation structures, links, and the relationship between different sections of the site.
  - **Digital Collections**: How items in a collection are grouped, ordered, and related to each other, such as collections of photographs, datasets, or archival materials.

1. **Statistical Inference**: Statistical inference is a method of making decisions about the parameters of a population, based on random sampling. It helps to assess the relationship between the dependent and independent variables. The purpose of statistical inference to estimate the uncertainty or sample to sample variation.

2. **Population and Sample in big data:**

Data: Both population and sample involve data. Population refers to the entire group or set of individuals, objects, or events being studied, while a sample is a subset of the population that is used for analysis. Descriptive Statistics: Descriptive statistics can be used to analyse both populations and samples.

Example: All the students in the class are population whereas the top 10 students in the class are the sample. All the members of the parliament is population and the female candidates present there is the sample.

- Population vs. sample

First, you need to understand the difference between a population and a sample, and identify the target population of your research.

- The **population** is the entire group that you want to draw conclusions about.
- The **sample** is the specific group of individuals that you will collect data from.

3. **Data Modelling:**

- Data modelling: Data modelling is the process of creating a visual representation of either a whole information system or parts of it to communicate connections between data points and structures.

- Data modelling is a process of creating a conceptual representation of data objects and their relationships to one another. The process of data modelling typically involves several steps, including requirements gathering, conceptual design, logical design, physical design, and implementation.

- Data Modelling in software engineering is the process of simplifying the diagram or data model of a software system by applying certain formal techniques. It involves expressing data and information through text and symbols. The data model provides the blueprint for building a new database or reengineering legacy applications.
- Data Modelling thus helps to increase consistency in naming, rules, semantics, and security. This, in turn, improves data analytics. The emphasis is on the need for availability and organization of data, independent of the manner of its application.

1.7 Philosophy of Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is an analysis approach that identifies general patterns in the data. These patterns include outliers and features of the data that might be unexpected. EDA is an important first step in any data analysis.
- **Important steps:**

Import Libraries
Configure Settings
Prepare Data
Import Data set files
Null Value Check

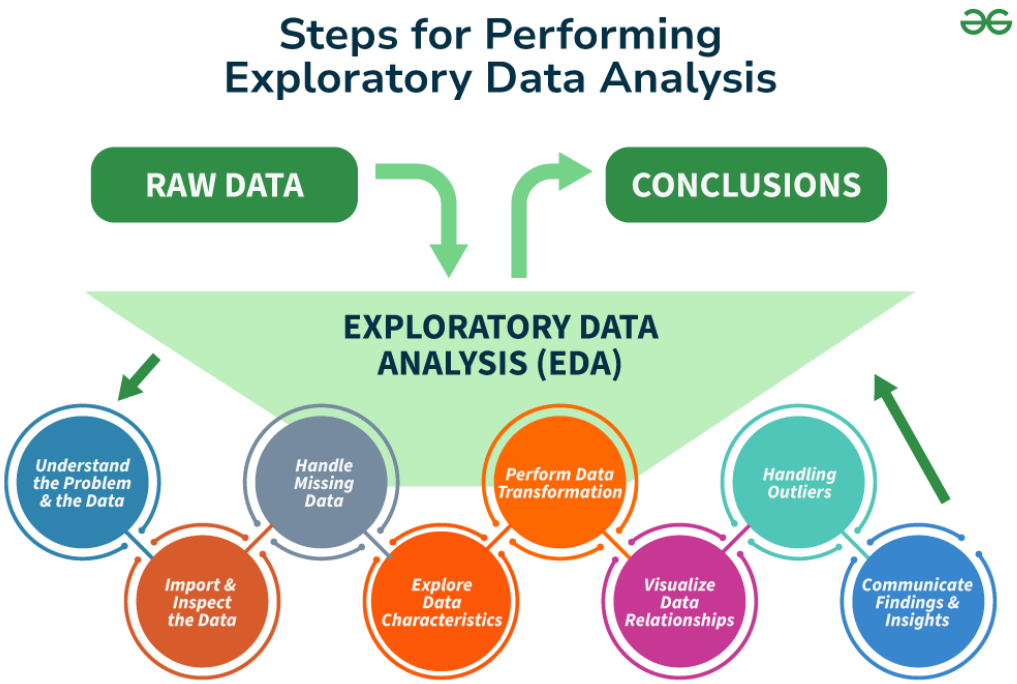


Fig 1.14 Exploratory Data Analysis(EDA)

- Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.
- EDA helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.
- EDA is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task. It provides a better understanding of data set variables and the relationships between them.
- It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.
- The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand

patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

- Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals.
- EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modelling, including machine learning.

1.8 Data Science Process

Step 1: Defining the problem. The first step in the data science lifecycle is to define the problem that needs to be solved. ...
Step 2: Data collection and preparation. ...
Step 3: Data exploration and analysis. ...
Step 4: Model building and evaluation. ...
Step 5: Deployment and maintenance.

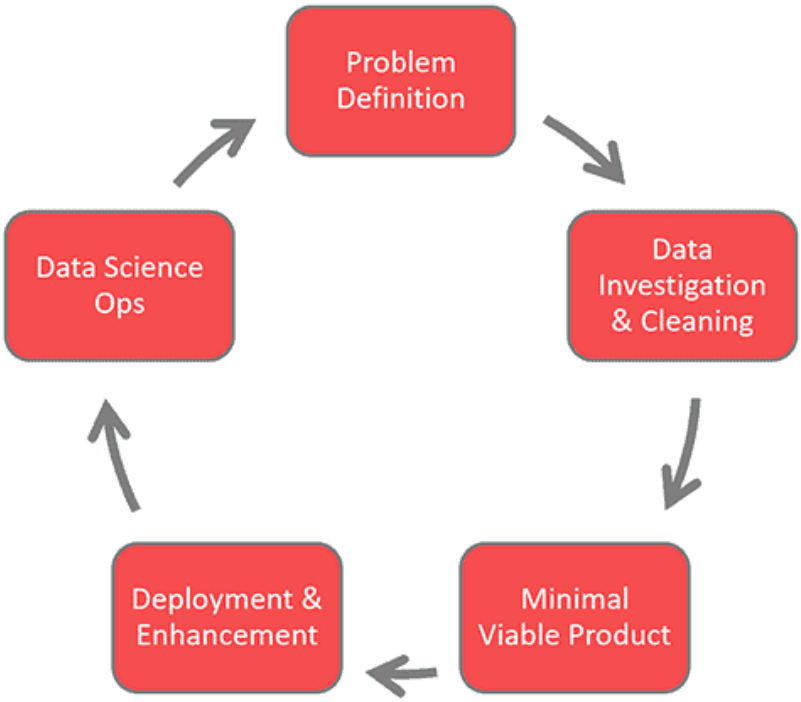


Fig 1.15 Data science process steps

- A data science life cycle is an iterative set of data science steps you take to deliver a project or analysis. Because every data science project and team are different, every specific data science life cycle is different. However, most data science projects tend to flow through the same general life cycle of data science steps.

❖ The Data science process

- Define the Problem and Set Objectives.
- Data Collection and Understanding.
- Data Preprocessing and Cleaning.
- Exploratory Data Analysis (EDA)
- Model Building and Machine Learning.
- Interpretation and Insights.
- Deployment and Monitoring.
- Deployment.

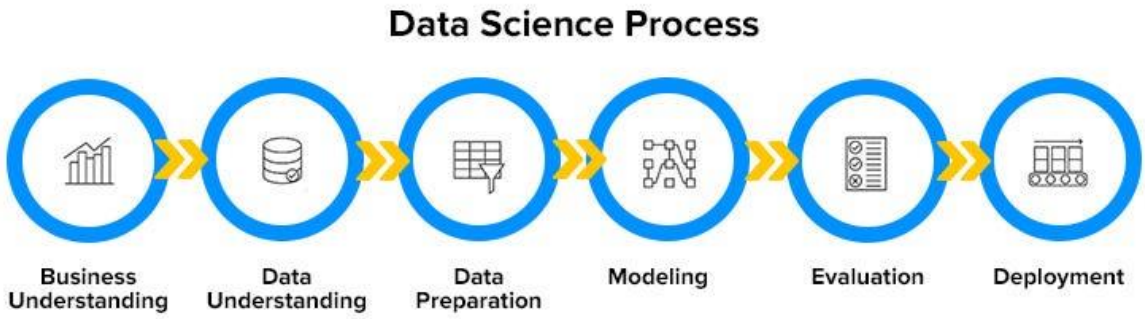


Fig 1.16 Data science process

1. **Problem Definition:**

Generally, the project lead or product manager manages this phase. Regardless, this initial phase should:

- State clearly the problem to be solved and why
- Motivate everyone involved to push toward this why
- Define the potential value of the forthcoming project
- Identify the project risks including ethical considerations
- Identify the key stakeholders
- Align the stakeholders with the data science team
- Research related high-level information
- Assess the resources (people and infrastructure) you'll likely need
- Develop and communicate a high-level, flexible project plan
- Identify the type of problem being solved*
- Get buy-in for the project

2. **Data Investigation and Cleaning:**

Without data, you've got nothing. Therefore, the team needs to identify what data is needed to solve the underlying problem. Then determine how to get the data:

- Is the data internally available? -> Get access to it
- Is the data readily collectable? -> Start capturing it
- Is the data available for purchase? -> Buy it

Once you have the data, start exploring it. Your data scientists or business/data analysts will lead several activities such as:

- Document the data quality
- Clean the data
- Combine various data sets to create new views
- Load the data into the target location (often to a cloud platform)
- Visualize the data

- Present initial findings to stakeholders and solicit feedback

- **DATA CLEANING & PREPARATION**

Before making analyzing the data, it is important to clean and prepare data. The methods used to clean and prepare the data are as listed below:

- Changing Data Types of Columns from object to Floats
- Filling in Missing Information
- Checking For Duplicate Rows
- Splitting Long Strings
- Creating Various New Columns

3. Minimal Viable Model/Product

All data science life cycle frameworks have some sort of modelling phase. However, I want to emphasize the importance of getting something useful out as quickly. This concept borrows from the idea of a Minimal Viable Product.

"The minimum viable product is that version of a new product which allows a team to collect the maximum amount of validated learning about customers with the least effort."

Or more simply — don't start out building a full-fledged product and then launch. Rather, get out something of value and receive feedback about whether it is on the right track. And if not, shift directions.

4. Deployment and Enhancements:

Deployment: Typically the more "engineering-focused" team members such as data engineers, cloud engineers, machine learning engineers, application developers, and quality assurance engineers execute this phase.

"No machine learning model is valuable, unless it's deployed to production."

Enhancements:

Extend the model to similar use cases (i.e. a new “Problem Definition” phase)

Add and clean data sets (i.e. a new “Data Investigation and Cleaning” phase)

Try new modelling techniques (i.e. developing the next “Viable Model”)

5. Data Science Ops:

As data science matures into mainstream operations, companies need to take a stronger product focus that includes plans to maintain the deployed systems long-term. There are three major overlapping facets of management to this.

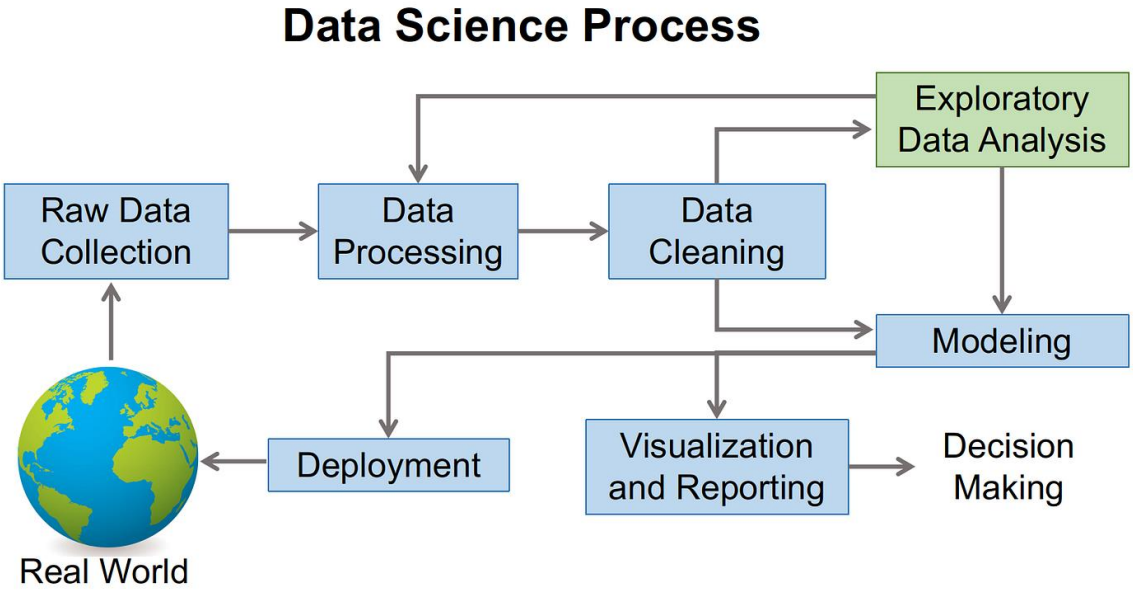


Fig 1.17 Block diagram of Data science process

- The data science process typically involves several key steps that guide the collection, processing, analysis, and interpretation of data to generate actionable insights. Here’s an overview of the typical stages:

1. **Problem Definition**:
    - **Understanding the Problem**: Define the problem you want to solve or the question you want to answer.
    - **Objectives and Goals**: Set clear objectives and goals for what you want to achieve with your analysis.

2. **Data Collection**:
    - **Data Sources**: Identify and gather data from various sources, which could include databases, APIs, web scraping, surveys, and more.
    - **Data Acquisition**: Collect the data, ensuring it's relevant and sufficient for the analysis.

3. **Data Cleaning and Preprocessing**:
    - **Data Cleaning**: Handle missing values, remove duplicates, correct errors, and deal with outliers.
    - **Data Transformation**: Normalize or standardize data, create new features, and convert data types as needed.

4. **Exploratory Data Analysis (EDA)**:
    - **Descriptive Statistics**: Calculate summary statistics like mean, median, mode, and standard deviation.
    - **Data Visualization**: Use plots (e.g., histograms, scatter plots, box plots) to visualize the data and uncover patterns and relationships.

5. **Feature Engineering**:
    - **Feature Selection**: Identify the most relevant features for your model.
    - **Feature Creation**: Create new features that may enhance model performance.

6. **Modeling**:
    - **Model Selection**: Choose the appropriate algorithms and models based on the problem type (e.g., regression, classification, clustering).
    - **Model Training**: Train the model using the training dataset.
    - **Hyperparameter Tuning**: Optimize model parameters to improve performance.

7. **Model Evaluation**:
    - **Validation**: Evaluate the model using a validation set to test its performance.
    - **Metrics**: Use appropriate metrics (e.g., accuracy, precision, recall, F1-score, RMSE) to assess model performance.
    - **Cross-Validation**: Perform cross-validation to ensure the model's robustness and generalizability.

8. **Model Deployment**:
    - **Implementation**: Deploy the model into a production environment where it can be used to make predictions on new data.
    - **Integration**: Integrate the model with existing systems and workflows.

9. **Monitoring and Maintenance**:
    - **Monitoring**: Continuously monitor the model's performance to ensure it remains accurate and relevant.
    - **Updating**: Regularly update the model as new data becomes available or as the underlying problem changes.

10. **Communication and Reporting**:
    - **Insights**: Communicate findings and insights to stakeholders through reports, dashboards, and presentations.
    - **Decision-Making**: Use the insights to inform decision-making and drive business strategies.

11. **Iterative Improvement**:
    - **Feedback Loop**: Use feedback from stakeholders and model performance to iteratively improve the process and models.

1.9 A Data Scientist's role: involves leveraging data to derive actionable insights and support data-driven decision-making. Their responsibilities can be summarized into several key areas:

1. **Problem Definition**
    - Collaborate with stakeholders to understand business objectives and translate them into data science problems.
2. **Data Collection**
    - Identify and gather relevant data from various sources using techniques like web scraping, APIs, and database querying.
3. **Data Cleaning and Preprocessing**
    - Clean and preprocess data by handling missing values, removing duplicates, and transforming data into a suitable format for analysis.
4. **Exploratory Data Analysis (EDA)**
    - Perform descriptive statistics and create visualizations to uncover patterns and insights within the data.

5. **Feature Engineering**
    - Create and select important features to enhance model performance and reduce dimensionality.
6. **Model Building**
    - Choose appropriate algorithms, train models, and fine-tune parameters to optimize performance.
7. **Model Evaluation**
    - Evaluate models using relevant metrics and validate them to ensure they generalize well to new data.
8. **Model Deployment**
    - Develop and implement a strategy for deploying models into production environments, ensuring seamless integration with existing systems.
9. **Monitoring and Maintenance**
    - Continuously monitor model performance, update and retrain models as necessary, and perform error analysis to refine models.
10. **Communication and Reporting**
    - Present findings and insights through clear and compelling storytelling, creating reports and dashboards for stakeholders.
11. **Ethical Considerations**
    - Ensure data privacy, compliance with regulations, and mitigate biases to promote fairness and ethical use of data.
12. **Continuous Learning**
    - Stay updated with the latest advancements in data science and continuously experiment with new techniques and tools.

In essence, a Data Scientist combines technical skills with business acumen to transform data into valuable insights that drive strategic decisions and operational improvements.

1.10 Case Study:

❖ Case Study in Data Science - Urban Planning and Smart Cities

1. Singapore

- Singapore is pioneering the smart city concept, using data science to optimize urban planning and public services.
- They gather data from various sources, including sensors and citizen feedback, to manage traffic flow, reduce energy consumption, and improve the overall quality of life in the city-state.

Singapore - Efficient Urban Planning using Data Science:

- Singapore's real-time traffic management system, powered by data analytics, has led to a 25% reduction in peak-hour traffic congestion, resulting in shorter commute times and lower fuel consumption.
- Singapore has achieved a 15% reduction in energy consumption across public buildings and street lighting, contributing to significant environmental sustainability gains.
- Citizen feedback platforms have seen 90% of reported issues resolved within 48 hours, reflecting the city's responsiveness in addressing urban challenges through data-driven decision-making.
- The implementation of predictive maintenance using data science has resulted in a 30% decrease in the downtime of critical public infrastructure, ensuring smoother operations and minimizing disruptions for residents.

2. Barcelona

- Barcelona has embraced data science to transform into a smart city as well.
- They use data analytics to monitor and control waste management, parking, and public transportation services.
- Barcelona improves the daily lives of its citizens and makes the city more attractive for tourists and businesses.
- Data science has significantly influenced Barcelona's urban planning and the development of smart cities, reshaping the urban landscape of this vibrant Spanish metropolis by:
- Barcelona's data-driven waste management system has led to a 20% reduction in the frequency of waste collection in certain areas, resulting in cost savings and reduced environmental impact.

- The implementation of smart parking solutions using data science has reduced the average time it takes to find a parking spot by 30%, easing congestion and frustration for both residents and visitors.
- Public transportation optimization through data analytics has improved service reliability, resulting in a 10% increase in daily ridership and reduced waiting times for commuters.
- Barcelona's efforts to become a smart city have attracted 30% more tech startups and foreign investments over the past five years, stimulating economic growth and job creation in the region.

❖ Case Study in Data Science - Housing Market Analysis

- Predicting or estimating the selling price of a property can be of great help when making important decisions such as the purchase of a home or real estate as an investment vehicle.
- It can also be an important tool for a real estate sales agency, since it will allow them to estimate the sale value of the real estate that for them in this case are assets.

1. Analysing Data to Predict Market Trends

- Data science in real estate helps to forecast property market trends and any risks that might exist in the investment.
- By using data that consists of a combination of different variables and predictive analysis implemented to that, data scientists understand and analyse how-
  - consumer groups have been behaving overtime
  - what type of properties have been in demand
  - the kind of leisure activities consumers are involving themselves
  - facilities that can be integrated with residential spaces to enhance consumer experience
  - evolution in the rents being charged

2. Formulating the Property Price Indices

- One of the most significant applications of data science in real estate is to collect and leverage information relating to the adjoining local areas.

- These include, supermarkets in the vicinity, educational institutes, business and commerce hubs, traffic in the neighbourhood, crime rates, cafes and restaurants, and physical infrastructure.
- These qualitative and quantitative variables play in to influence the pricing of individual properties.
- Additionally, through data science in real estate, a system can be deployed wherein the individual variables work as additions.
- For example, there can be an average price set for the properties in one specific building.
- Now, the variables affected by the floor number, size of rooms and the view from the window, work as additions that are charged for additionally.
- Therefore, the internal variables of the property alongside the hyperlocal variables work to formulate the property price indices and help real estate agents to cater better to the needs of the clients.

3. Understanding Investment Performance

- In the field of real estate, no two properties can ever be identical.
- Variables differ even with properties in the same building, not to mention the changing value of properties with time.
- Understanding individual sub-market performance is therefore a difficult problem to deal with.
- As a solution to this issue, the changing price of an asset can be tracked over time by using data science in real estate.
- In the world of real estate, each property is unique, with factors varying even among those situated within the same building.
- Adding to the complexity, property values change constantly due to fluctuations in the market and evolving infrastructure.

4. Estimating Profitability of Investment and Construction

- Whether one invests in a commercial real estate space or a residential one, location intelligence acts as a very important aspect to gauge whether the investment would be able to yield the expected profits in the future.

- With the proper information about the geography of a particular property, accessibility of services around it, land ownership, zoning, regional laws, etc. an investor or a real estate consultant can make a more informed decision by visualizing and analyzing prospects.

5. Managing Finances of Properties

- Let's assume that you manage a diverse pool of properties across various localities in Mumbai.
- All through the work is the same, you need to evaluate the reasons why one property is draining more resources in comparison to another.
- This could be in terms of losses incurred due to higher vacancy rates or systems malfunctions.
- Fortunately, data science in real estate management helps you in identifying the root cause.
- This is done by gathering data such as, receivables and budget, profitability and cost analysis, planning for tenant build outs from different properties.
- The data can then be evaluated based on various metrics, you can zone down to the bottom of the problem, and formulate solutions for the same.
- By gathering data on receivables, budgets, profitability, tenant build-outs, and even vacancy rates across your properties, data science lets you evaluate them based on various metrics.

6. Trimming Down Energy Consumption

- With the incorporation of data science in real estate, identifying the root cause of energy wastage has now become possible.
- Nowadays, there are a plethora of apps and software available that gather and assess energy data from smart meters and sensors, and can also detect faults in the heating, ventilation, and air conditioning (HVAC) systems.
- Based on the weather changes and the usage pattern, these apps offer a holistic understanding of energy spendings.
- This can help property managers, homeowners, and tenants to alter their lifestyles and change energy consumption patterns.

7. Simplifying Home Searching or Buying Process

- Data science usage in real estate not only benefits the investor and broker class, but it also streamlines the home searching, buying and renting process.
- It is very much possible that real estate property prices vary drastically across different cities.
- This can be attributed to factors that range from how well it is connected to the areas around, the commercial centers present in the area, and the modes of transportation and commutation.
- When these are effectively analyzed through data science, it helps buyers decide upon a living location or understand the expenses involved if they have made up their mind to shift to another city.
- By examining user behavior, their lifestyle preference, budget range, amenities preference and other such factors, you can offer property suggestions that match the requirements of the users.
- This will therefore save customers' time in scooping through multiple property listings.
- Data science in real estate also simplifies the process of finding, purchasing, or renting homes for individuals and families.
- Property prices can fluctuate significantly between cities, influenced by factors such as connectivity to nearby areas, proximity to commercial centers, and availability of transportation options.
- In addition, examining user behavior, preferences in terms of lifestyle, budget constraints, desired amenities, and other relevant aspects can lead to personalized property recommendations tailored to meet customer needs.

8. Revamping the Marketing Strategy

- Data science in real estate aids in collecting and examining information through multiple sources.
- This can help agencies in understanding the behavior and preferences of the consumers, assessing the competition, and marketing their services in a more creative way.

- Once user preference is understood, virtual staging, 3D rendering and visualization, Google or Facebook ads, and listings can be optimized in order to attract the target audience.

9. Identifying and Segregating Leads

- A very interesting way to harness the power of data science in real estate is in the field of lead nurturing and segregation.
- With the help of data science-backed applications and softwares, giving a “seller or buyer score” to leads which are most likely to sell/buy properties has now become possible.
- This assessment is made by evaluating factors like demographics, income changes and purchasing behavior.

Prospects for Growth

- Modern technologies have revolutionized the real estate market.
- Many companies have already shifted to big data-machine learning powered software for analyzing data, calculating the profitability of an apartment purchase, portfolio management, and estimating property rentals.
- The study of how customers, groups, or organizations select, buy, use, and dispose of ideas, goods, and services can impact, inform and govern the decision-making process of the producing firms and organizations to a large extent.

Module II

Mathematical Preliminaries

Module II
Mathematical Preliminaries
Probability—Descriptive Statistics—Correlation Analysis and Regression.

Data Munging: Properties of Data—Collecting Data—Cleaning Data—Crowdsourcing.

2.1 Probability:

Probability is the chance that something will happen — how likely it is that some event will happen.

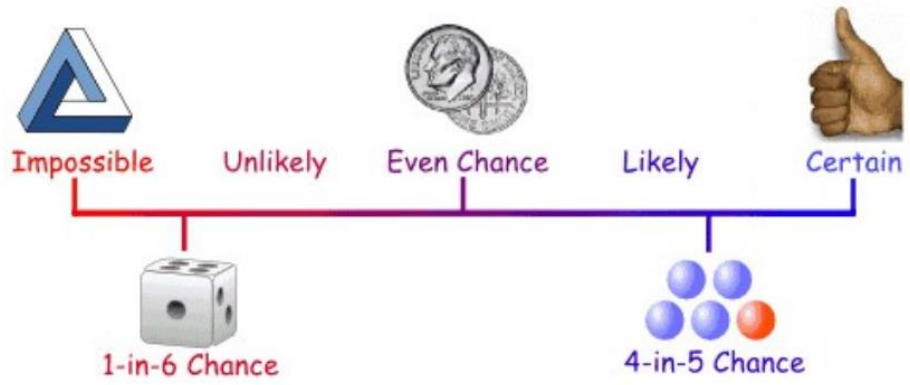


Fig 2.1 Probability

- Probability of an event happening P(E) = Number of ways it can happen n(E)/ Total number of outcomes n(T)
- Probability is the measure of the likelihood that an event will occur. Probability is quantified as a number between 0 and 1, where 0 indicates impossibility and 1 indicates certainty shown in Figure
- **Importance of Probability:**

  In certain areas of our everyday lives, confusion and randomness exist and having a strong knowledge of probability allows us to make sense of these uncertainties. Knowing about chance allows us to make educated judgments on what is likely to happen, based on a trend of previously collected data or estimation.

❖ How to use Probability in Data Science?

Data Science also makes use of statistical inferences to forecast or interpret computer patterns, while statistical inferences use data distribution of probabilities. Therefore, it is important to know the likelihood and its implementations to work effectively on data science problems.

❖ **Events in Probability**

- In probability theory, an event is a set of outcomes of an experiment or a subset of the sample space.
- If P(E) represents the probability of an event E, then, we have,
  - P(E) = 0 if and only if E is an impossible event.
  - P(E) = 1 if and only if E is a certain event.
  - 0 ≤ P(E) ≤ 1.
- Suppose, we are given two events, "A" and "B", then the probability of event A, P(A) > P(B) if and only if event "A" is more likely to occur than the event "B".
- Sample space(S) is the set of all of the possible outcomes of an experiment and n(S) represents the number of outcomes in the sample space.

➢ P(E) = n(E)/n(S)

➢ P(E') = (n(S) - n(E))/n(S) = 1 - (n(E)/n(S))

E' represents that the event will not occur.

- Therefore, now we can also conclude that,

$$P(E) + P(E') = 1$$

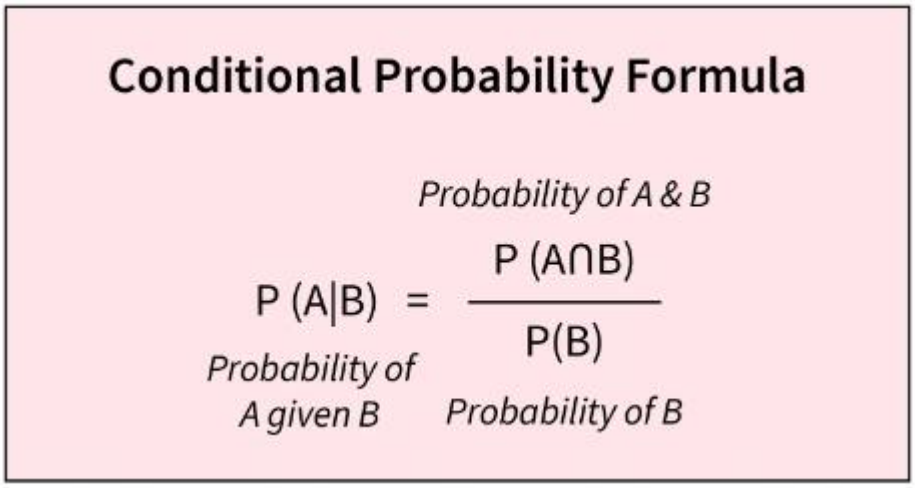
❖ Terms in Probability

Term	Definition
Sample Space	Set of all possible outcomes in a probability experiment. For instance, in a coin toss, it's "head" and "tail".
Sample Point	One of the possible results in an experiment. For example, in rolling a fair six-sided dice, sample points are 1 to 6.
Experiment	A process or trial with uncertain results. Examples include coin tossing, card selection, or rolling a die.
Event	A subset of the sample space representing certain outcomes. Example: getting "1" when rolling a die.
Favorable Outcome	An outcome that produces the desired or expected consequence.

❖ Conditional Probability:

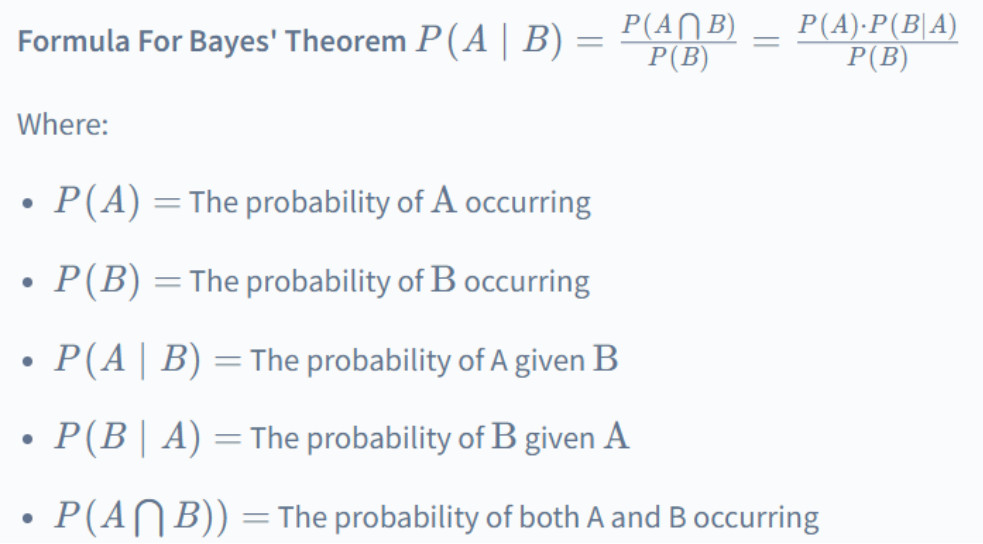
- Conditional probability is the probability or chance of occurrence of an event, given that another event has already occurred. It is used to calculate the probability of conditional events. For example, what is the probability of drawing a red card from a deck if one spade has already been drawn? This can be calculated using conditional probability.

- The condition probability of an event A given B is denoted by P(A/B), and it can be calculated using the below formula –



- Bayes Theorem

In probability theory, The Bayes Theorem is used to describe the probability of an event based on the prior knowledge of the other related conditions or events. So, if we know the conditional probabilities P(B/A) and prior probabilities P(A) and P(B), then we can calculate P(A/B) using the Bayes Theorem mentioned in below figure –



It is one of the most important concepts of Probability Theory.

Problems on Probability with solutions:

Example 1: A coin is thrown 3 times .what is the probability that atleast one head is obtained?

Sol: Sample space = [HHH, HHT, HTH, THH, TTH, THT, HTT, TTT]

Total number of ways = 2 × 2 × 2 = 8. Fav. Cases = 7

P (A) = 7/8

OR

P (of getting at least one head) = 1 – P (no head)⇒ 1 – (1/8) = 7/8

Example 2: Find the probability of getting a numbered card when a card is drawn from the pack of 52 cards.

Sol: Total Cards = 52. Numbered Cards = (2, 3, 4, 5, 6, 7, 8, 9, 10) 9 from each suit

4 × 9 = 36

P (E) = 36/52 = 9/13

2.2 Descriptive statistics

- Descriptive statistics are the informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population.
- Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).
- Measures of central tendency include the mean, median, and mode.
- Measures of variability include standard deviation, variance, minimum and maximum variables.

❖ Types of Descriptive Statistics

1 Measures of Central Tendency

2 Measure of Variability

3 Measures of Frequency Distribution

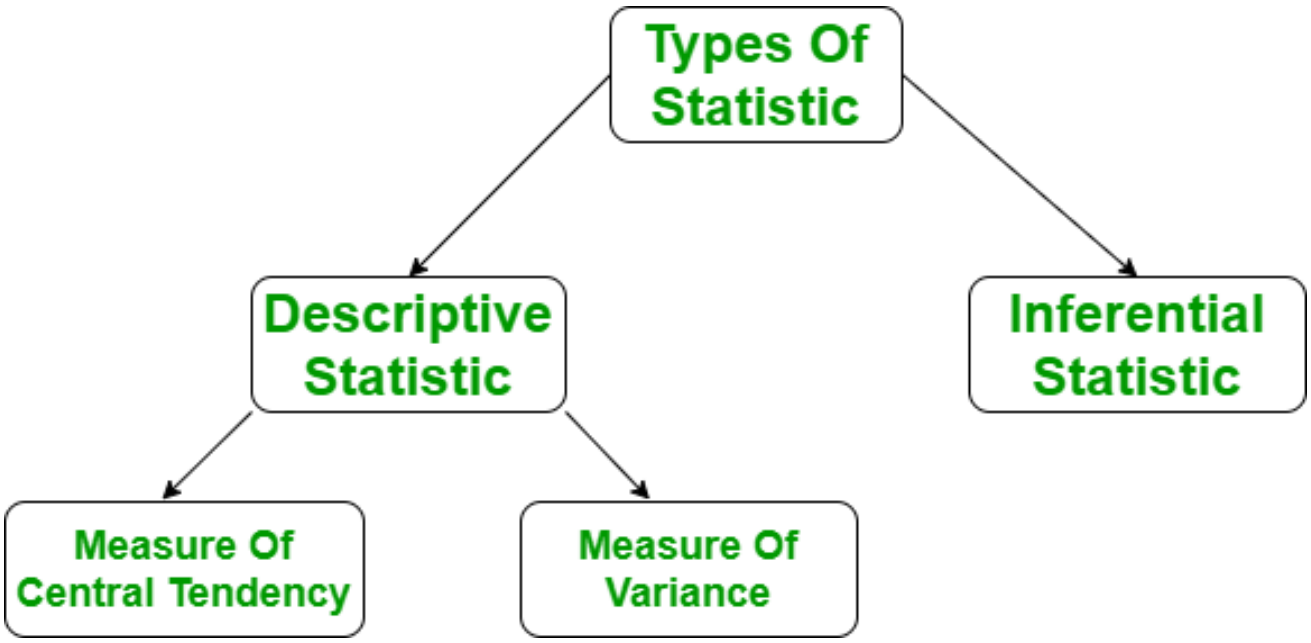


Fig 2.2 Types of Statistic

1. Measures of Central Tendency

It represents the whole set of data by a single value. It gives us the location of the central points. There are three main measures of central tendency:

Mean

Mode

Median

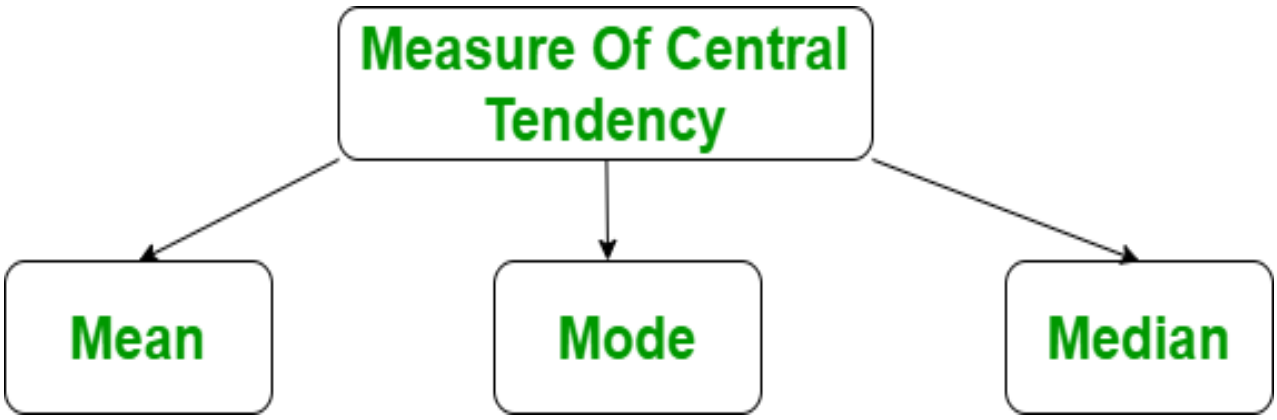


Fig 2.3 Main measures of central tendency

- Mean

  It is the sum of observations divided by the total number of observations. It is also defined as average which is the sum divided by count.

$$\bar{x} = \frac{\sum x}{n}$$

where,

- x = Observations
- n = number of terms

Let's look at an example of how can we find the mean of a data set using python code implementation.

Python3

```python
import numpy as np

# Sample Data
arr = [5, 6, 11]

# Mean
mean = np.mean(arr)

print("Mean = ", mean)
```

Output :

```
Mean =  7.333333333333333
```

- ***Mode***

It is the value that has the highest frequency in the given data set. The data set may have no mode if the frequency of all data points is the same. Also, we can have more than one mode if we encounter two or more data points having the same frequency.

Python3

```
from scipy import stats

# sample Data
arr = [1, 2, 2, 3]

# Mode
mode = stats.mode(arr)
print("Mode = ", mode)
```

Output:

```
Mode =  ModeResult(mode=array([2]), count=array([2]))
```

➢ Median

It is the middle value of the data set. It splits the data into two halves. If the number of elements in the data set is odd then the center element is the median and if it is even then the median would be the average of two central elements.

Python3

```
import numpy as np

# sample Data
arr = [1, 2, 3, 4]

# Median
median = np.median(arr)

print("Median = ", median)
```

Output:

```
Median =  2.5
```

❖ **Summary: Mean, Median, and Mode**

What can we learn from looking at a group of numbers?

In Machine Learning/Data Science (and in mathematics) there are often three values that interests us:

- **Mean** - The average value

- **Median** - The mid point value
- **Mode** - The most common value

2. **Measure of Variability**

Measures of variability are also termed measures of dispersion as it helps to gain insights about the dispersion or the spread of the observations at hand. Some of the measures which are used to calculate the measures of dispersion in the observations of the variables are as follows:

- Range
- Variance
- Standard deviation

3. **Measures of Frequency Distribution**

Measures of frequency distribution help us gain valuable insights into the distribution and the characteristics of the dataset. Measures like,

- Count
- Frequency
- Relative Frequency
- Cumulative Frequency

2.3 Correlation Analysis & Regression

- The most commonly used techniques for investigating the relationship between two quantitative variables are correlation and linear regression.

- Correlation refers to a mutual relationship mutual relationship or association between quantitative variables.

- It can help in predicting one quantity from another.

- Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation.

- Regression analysis is a set of statistical processes for estimating the relationships estimating the relationships among variables

- Correlation in correlation and regression can be defined as a numeric value that determines whether variables are linearly related and give a numeric value to the corresponding strength. Regression is an equation that checks how a change in one variable will result in a change in another variable.

- **The Formula for Correlation and Regression:**

The formula for correlation and regression is given as follows

- Correlation: $r_{xy} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_1^n (x_i - \bar{x})^2 \sum_1^n (y_i - \bar{y})^2}}$
- Regression line equation: $y = \alpha + \beta x$, where $\beta = r_{xy} \frac{\sigma_y}{\sigma_x}$ and $\alpha = \bar{y} - \beta\bar{x}$

The Similarity Between Correlation and Regression?

- The similarity between correlation and regression is that if the correlation coefficient is positive (or negative) then the slope of the regression line will also be positive (or negative).
- The best way to find the correlation and regression between two variables is by using Pearson's correlation coefficient and by employing the ordinary least squares method respectively.
- **Correlation Example:**

- Correlation refers to the statistical relationship between the two entities. It measures the extent to which two variables are linearly related. For example, the height and weight of a person are related, and taller people tend to be heavier than shorter people. You can apply correlation to a variety of data sets.
- Questions and method to understand the correlation:
  What Is Correlation?
  What is Correlation Coefficient?
  Types of Correlation Coefficient
  Calculate Correlation Using Excel

- **There are three types of correlation:**

1) Positive Correlation: A positive correlation means that this linear relationship is positive, and the two variables increase or decrease in the same direction.

2)Negative Correlation: A negative correlation is just the opposite. The relationship line has a negative slope, and the variables change in opposite directions, i.e., one variable decreases while the other increases.

3)No Correlation: No correlation simply means that the variables behave very differently and thus, have no linear relationship.

❖ **Types of Correlation:**

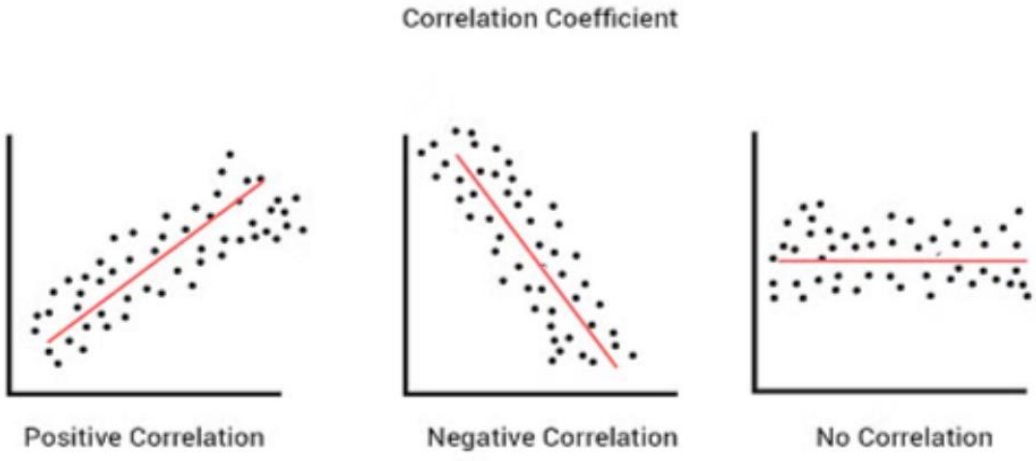


Fig 2.4 Types of Correlation

2.4 Regression:

❖ Regression is a popular statistical technique used in Data Science for the prediction of unknown values in a data set based on the known features. It is used when there is a missing value in a data set.

❖ In the context of machine learning and data science, regression specifically refers to the estimation of a continuous dependent variable or response from a list of input variables, or features.

❖ For example, it can determine which marketing channels or advertising strategies influence sales most, allowing businesses to allocate resources more effectively.

- **An easy example of regression:**
  - We could use the equation to predict weight if we knew an individual's height.
  - In this example, if an individual was 70 inches tall, we would predict his weight to be: Weight = 80 + 2 x (70) = 220 lbs = 99 kg
  - In this simple linear regression, we are examining the impact of one independent variable on the outcome.

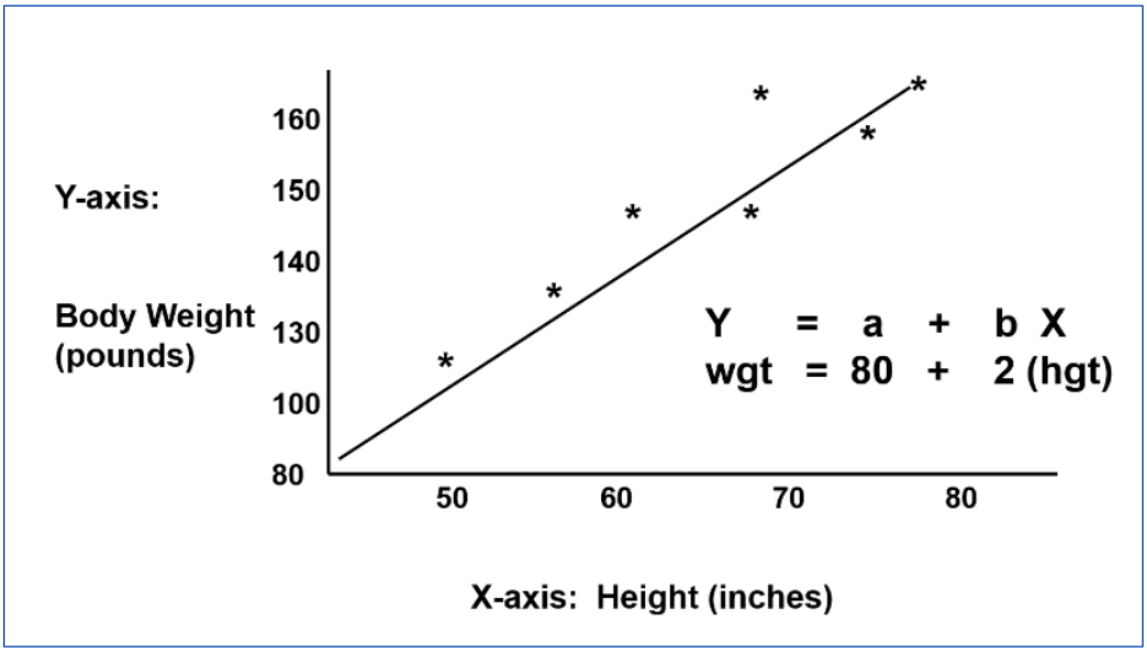


Fig 2.5 Regression example

➢ Difference between Correlation and Regression

Correlation	Regression
Correlation is used to determine whether variables are related or not.	Regression is used to numerically describe how a dependent variable changes with a change in an independent variable
Correlation tries to establish a linear relationship between variables.	It finds the best-fitted regression line to estimate an unknown variable on the basis of the known variable.
The variables can be used interchangeably	The variables cannot be interchanged.
Correlation uses a signed numerical value to estimate the strength of the relationship between the variables.	Regression is used to show the impact of a unit change in the independent variable on the dependent variable.
The Pearson's coefficient is the best measure of correlation.	The least-squares method is the best technique to determine the regression line.

2.5 Data Munging:

- Data munging, also known as data wrangling, is the process of transforming and cleaning raw data into a format suitable for analysis. In the realm of data science, it's often said that 80% of the work involves preparing and cleaning data, and data munging is at the heart of this process.
- The raw data collected from various sources can be messy, inconsistent, or incomplete, making it challenging to derive meaningful insights. Data munging involves several key steps:

1. **Data Collection:** Gathering data from diverse sources such as databases, spreadsheets, APIs, or web scraping.
2. **Data Cleaning:** Identifying and handling missing values, correcting errors, removing duplicates, and standardizing formats.
3. **Data Transformation:** Converting data into a consistent format, reshaping it for analysis, and creating new variables or features.
4. **Data Integration:** Combining data from multiple sources while ensuring consistency and coherence.
5. **Data Enrichment:** Augmenting the dataset with additional information or derived features to enhance analysis.
6. **Quality Assurance:** Verifying the integrity and accuracy of the transformed data through validation checks.

2.6 Properties of Data

In machine learning (ML) and data science, understanding the properties of data before conducting Exploratory Data Analysis (EDA) is crucial.

Here are the key properties to consider:

- **Type:** Understanding the data types (numerical, categorical, ordinal, binary, etc.) helps in selecting the appropriate statistical methods and models.
- **Quality:** Assessing the quality of data, including completeness (missing values), accuracy, and consistency, is vital for reliable analysis.
- **Distribution:** Knowing the distribution of the data (normal, skewed, uniform, etc.) helps in selecting the right statistical tests and transformations.
- **Range:** The minimum and maximum values in the dataset provide insights into the scale and potential outliers.
- **Central Tendency:** Measures such as mean, median, and mode give an idea of the typical value in the dataset.
- **Variance and Standard Deviation:** These measures indicate the spread or dispersion of the data, which is crucial for understanding variability.
- **Skewness and Kurtosis:** These metrics describe the asymmetry and peakedness of the data distribution, respectively.
- **Correlations:** Identifying relationships between variables helps in understanding dependencies and multicollinearity.

2.7 Collecting Data

- The process of gathering and analyzing accurate data from various sources to find answers to research problems, trends and probabilities to evaluate possible outcomes is known as Data Collection.
- The two methods used to gather information for research or analysis purposes are:

➢ Primary Data Collection methods

➢ Secondary Data Collection methods

1. Primary Data Collection Methods

- Primary data or raw data is a type of information that is obtained directly from the first-hand source through experiments, surveys or observations.
- The primary data collection method is further classified into two types. They are:
  - ➢ Quantitative Data Collection Methods
  - ➢ Qualitative Data Collection Methods

Quantitative Data Collection Methods

- It is based on mathematical calculations using various formats like close-ended questions, correlation and regression methods, mean, median or mode measures.
- This method is cheaper than qualitative data collection methods and it can be applied in a short duration of time.

Qualitative Data Collection Methods

- It does not involve any mathematical calculations.
- This method is closely associated with elements that are not quantifiable.
- This qualitative data collection method includes interviews, questionnaires, observations, case studies, etc.

❖ The various methods to collect the qualitative data are:

1. Observation Method

- Observation method is used when the study relates to behavioral science.
- This method is planned systematically. It is subject to many controls and checks.

- The different types of observations are:
  - Structured and unstructured observation
  - Controlled and uncontrolled observation
  - Participant, non-participant and disguised observation

2. Interview Method

- The method of collecting data in terms of verbal responses.
- It is achieved in two ways, such as:

- Personal Interview – In this method, a person known as an interviewer is required to ask questions face to face to the other person. The personal interview can be structured or unstructured, direct investigation, focused conversation, etc.
- Telephonic Interview – In this method, an interviewer obtains information by contacting people on the telephone to ask the questions or views, verbally.

3. Questionnaire Method

- In this method, the set of questions are mailed to the respondent.
- They should read, reply and subsequently return the questionnaire.
- The questions are printed in the definite order on the form.
- A good survey should have the following features:
  - Short and simple
  - Should follow a logical sequence
  - Provide adequate space for answers
  - Avoid technical terms
  - Should have good physical appearance such as colour, quality of the paper to attract the attention of the respondent

❖ Key steps in Data Collection Process:

- Well-designed data collection processes include the following steps:
  - ❖ Identify a business or research issue that needs to be addressed and set goals for the project.
  - ❖ Gather data requirements to answer the business question or deliver the research information.
  - ❖ Identify the data sets that can provide the desired information.

- ❖ Set a plan for collecting the data, including the collection methods that will be used.
- ❖ Collect the available data and begin working to prepare it for analysis.

2. Secondary Data Collection Methods:

Secondary data is data collected by someone other than the actual user. It means that the information is already available, and someone analyses it. The secondary data includes magazines, newspapers, books, journals, etc. It may be either published data or unpublished data.

Published data are available in various resources including

Government publications
Public records
Historical and statistical documents
Business documents
Technical and trade journals
Unpublished data includes

Diaries
Letters
Unpublished biographies, etc.

2.8 Data Cleaning in Data Science

- ❖ Data cleaning:
- ➢ Data collected in raw form may not be in usable form for analysis.

- ➢ Before we start analysis, a proper cleaning is required.
- ➢ Cleaning of data may include
- ➢ Distinguishing errors
- ➢ Data compatibility / unification
- ➢ Imputation of missing values
- ➢ Estimating unobserved (zero) counts

- Outlier detection

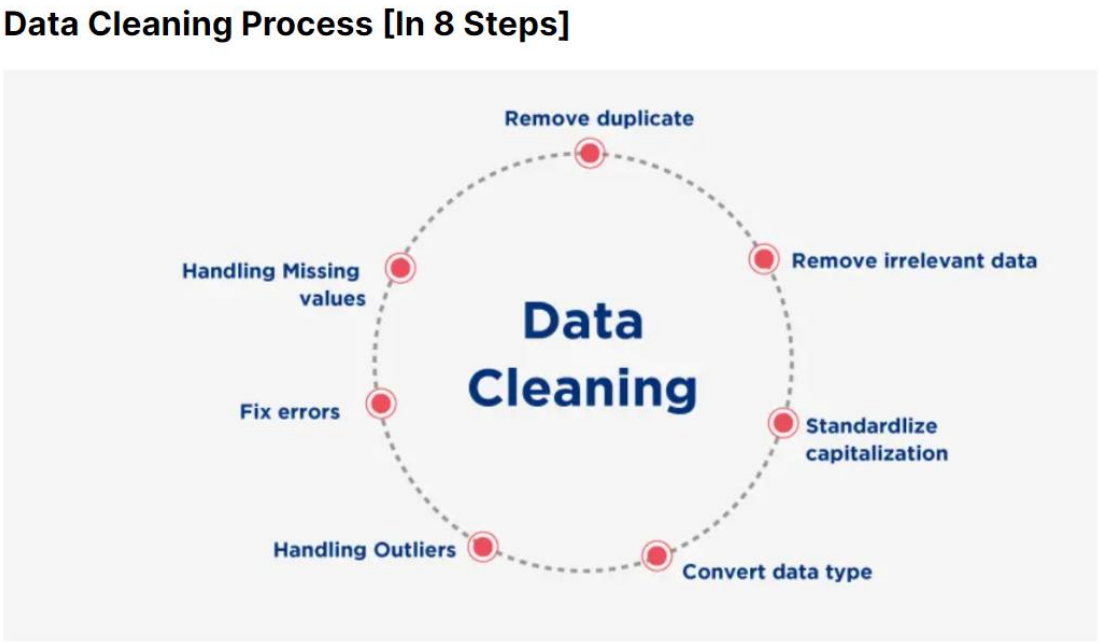


Fig 2.6 Basic steps of Data cleaning

- The data cleaning process typically involves several tasks, such as:

1. Handling missing data: Strategies like imputation or deletion are employed to address missing values in the dataset, ensuring that analyses are not compromised by incomplete information.

2. Removing duplicates: Identifying and removing duplicate records helps prevent redundancy and ensures that each observation in the dataset is unique.

3. Standardizing formats: Consistency in data formatting, such as dates, currencies, and units of measurement, is crucial for accurate analysis and interpretation.

4. Detecting and correcting errors: Automated techniques or manual inspection may be used to detect and correct errors in the data, such as typos, incorrect values, or anomalies.

5. Dealing with outliers: Outliers can skew statistical analyses and machine learning models. Data cleaning may involve identifying and either removing or transforming outliers to mitigate their impact.

6. Validating data integrity: Checking for data integrity issues, such as referential integrity constraints or logical inconsistencies, ensures that the dataset accurately represents the underlying phenomena.

2.9 Crowdsourcing in data science

- Crowdsourcing refers to the practice of obtaining data, insights, or solutions to problems by soliciting contributions from a large group of people, typically via an online platform. This approach leverages the collective intelligence and diverse perspectives of a crowd to achieve tasks that might be challenging or time-consuming for traditional methods or individual experts.

- Crowdsourcing involves obtaining work, information, or opinions from a large group of people who submit their data through the Internet, social media, and smartphone apps.

  In the context of data science, crowdsourcing can involve tasks such as:

  1. Data annotation and labeling: Crowdsourcing can be used to label or annotate large datasets for tasks like image classification, sentiment analysis, or text categorization.

  2. Data collection: Gathering large volumes of data from diverse sources, such as user-generated content, sensor data, or public records, by engaging a crowd of contributors.

  3. Problem-solving: Presenting a data science problem or challenge to a crowd and soliciting solutions, insights, or algorithms from participants.

4. Model validation and evaluation: Using crowdsourcing to validate the performance of machine learning models or to evaluate the quality of predictions or classifications made by algorithms.

❖ Types of Crowdsourcing:

- Wisdom - Wisdom of crowds is the idea that large groups of people are collectively smarter than individual experts when it comes to problem-solving or identifying values.
- Creation - Crowd creation is a collaborative effort to design or build something. Open-source software is the good example.
- Voting - Crowd voting uses the democratic principle to choose a particular policy or course of action by "polling the audience."
- Funding - Crowdfunding involved raising money for various purposes by soliciting relatively small amounts from a large number of funders.

❖ **Examples of Crowdsourcing**

1. Doritos:

- It is one of the companies which is taking advantage of crowdsourcing for a long time for an advertising initiative.
- They use consumer-created ads for one of their 30-Second Super Bowl Spots (Championship Game of Football).

2. Starbucks:

- Another big venture which used crowdsourcing as a medium for idea generation.
- Their white cup contest is a famous contest in which customers need to decorate their Starbucks cup with an original design and then take a photo and submit it on social media.

3. Airbnb:

- A very famous travel website that offers people to rent their houses or apartments by listing them on the website.
- All the listings are crowdsourced by people.

❖ Applications of Crowdsourcing:

- Crowdsourcing is widely used in various sectors from education to health.
- It is not only accelerating innovation but democratizing problem-solving methods.
- Crowdsourcing can be used in the following fields:

1. Enterprise
2. Information Technology
3. Marketing
4. Education
5. Finance
6. Science and Health

❖ Advantages Of Crowdsourcing

- Evolving Innovation: Crowdsourcing helps in getting innovative ideas from people belonging to different fields and thus helping businesses grow in every field.
- Save costs: There is the elimination of wastage of time of meeting people and convincing them. Only the business idea is to be proposed on the internet and you will be flooded with suggestions from the crowd.
- Increased Efficiency: Crowdsourcing has increased the efficiency of business models as several expertise ideas are also funded.

❖ Disadvantages Of Crowdsourcing

- Lack of confidentiality: Asking for suggestions from a large group of people can bring the threat of idea stealing by other organizations.
- Repeated ideas: Often contestants in crowdsourcing competitions submit repeated, plagiarized ideas which leads to time wastage as reviewing the same ideas is not worthy.

❖ **Example:**

Using the computation formula for the sum of squares, calculate the population standard deviation and sample deviation for the score?
a) 1,3,7,2,0,4,3,7,
b) 10,8,5,0,1,7,9,2,1

- **Population Standard Deviation (σ):**
-
  - $\sigma = \sqrt{(\Sigma(x - \mu)^2 / N)}$
  - where:
    - Σ is the sum of
    - x is each individual value
    - μ is the population mean
    - N is the population size
    -
- **Sample Standard Deviation (s):**
  - $S = \sqrt{(\Sigma(x - \bar{x})^2 / (n - 1))}$
  -
  - where:
    - Σ is the sum of
    - x is each individual value
    - $\bar{x}$ is the sample mean
    - n is the sample size

Calculations

a: 1, 3, 7, 2, 0, 4, 3, 7

Sample Standard Deviation

- Calculate the mean ($\bar{x}$): (1+3+7+2+0+4+3+7)/8 = 3.5
- Calculate the sum of squared deviations: 42 (same as before)
- Calculate the variance: 42/7 = 6
- Calculate the standard deviation: $\sqrt{6} \approx 2.45$

Population Standard Deviation

1. Calculate the mean (μ): (1+3+7+2+0+4+3+7)/8 = 3.5
2. Calculate the sum of squared deviations:
   $(1-3.5)^2 + (3-3.5)^2 + (7-3.5)^2 + (2-3.5)^2 + (0-3.5)^2 + (4-3.5)^2 + (3-3.5)^2 + (7-3.5)^2 = 42$
3. Calculate the variance: 42/8 = 5.25
4. Calculate the standard deviation: $\sqrt{5.25} \approx 2.29$

b: 10, 8, 5, 0, 1, 7, 9, 2, 1

Sample Standard Deviation

1. Calculate the mean ($\bar{x}$): (10+8+5+0+1+7+9+2+1)/9 = 4.67
2. Calculate the sum of squared deviations:
   $(10-4.67)^2 + (8-4.67)^2 + (5-4.67)^2 + (0-4.67)^2 + (1-4.67)^2 + (7-4.67)^2 + (9-4.67)^2 + (2-4.67)^2 + (1-4.67)^2 = 121.56$
3. Calculate the variance: 121.56/8 = 15.195
4. Calculate the standard deviation: $\sqrt{15.195} \approx 3.89$

Population Standard Deviation

- Calculate the mean (μ): (10+8+5+0+1+7+9+2+1)/9 = 4.67
- Calculate the sum of squared deviations: 121.56 (same as before)
- Calculate the variance: 121.56/9 = 13.51
- Calculate the standard deviation: $\sqrt{13.51} \approx 3.67$

Results

- **Dataset a:**
  - Population Standard Deviation ≈ 2.29
  - Sample Standard Deviation ≈ 2.45
- **Dataset b:**
  - Population Standard Deviation ≈ 3.67
  - Sample Standard Deviation ≈ 3.89

Q. Compute the mean, median and mode for the following data sets
a) 45, 55, 60, 60, 63, 63, 63, 63, 65, 65, 70
b) 26.9, 26.3, 28.7, 27.4, 26.6, 27.4, 26.9, 26.9
a: 45, 55, 60, 60, 63, 63, 63, 63, 65, 65, 70

Mean

- **Formula:** Mean = (Sum of all values) / (Number of values)
- **Calculation:** Mean = (45 + 55 + 60 + 60 + 63 + 63 + 63 + 63 + 65 + 65 + 70) / 11 = 61.09

Median

- **Formula:** Median = Middle value when data is ordered
- **Calculation:**
  - Arrange data in ascending order: 45, 55, 60, 60, 63, 63, 63, 63, 65, 65, 70
  - Median = 63

Mode

- **Formula:** Mode = Most frequent value
- **Calculation:** Mode = 63

b: 26.9, 26.3, 28.7, 27.4, 26.6, 27.4, 26.9, 26.9

Mean

- **Formula:** Mean = (Sum of all values) / (Number of values)
- **Calculation:** Mean = (26.9 + 26.3 + 28.7 + 27.4 + 26.6 + 27.4 + 26.9 + 26.9) / 8 = 27.14

Median

- **Formula:** Median = Middle value when data is ordered
- **Calculation:**
  - Arrange data in ascending order: 26.3, 26.6, 26.9, 26.9, 26.9, 27.4, 27.4, 28.7
  - Since there are an even number of values, the median is the average of the two middle values: (26.9 + 26.9) / 2 = 26.9

Mode

- **Formula:** Mode = Most frequent value
  Calculation: Mode = 26.9

Module III

Scores—Ranking and Statistical Analysis

Module III

Scores—Ranking and Statistical Analysis:

Scores and Rankings: Developing Scoring Systems—Z-scores and Normalization

Statistical Analysis: Sampling from Distributions—Statistical Distributions—Statistical Significance—Permutation Tests and P-values

- **Scores and Rankings:**

3.1 Developing Scoring Systems:

Developing scoring systems in data science involves creating a model or algorithm that assigns a score or rating to individuals, entities, or events based on various attributes or features. These scoring systems are commonly used in fields such as credit risk assessment, fraud detection, marketing analytics, and healthcare.

Here's a brief overview of the **steps involved in developing scoring systems**:

1. **Define the Objective**: Clearly define the purpose of the scoring system. Determine what you want to predict or assess, such as creditworthiness, likelihood of fraud, customer churn, etc.

2. **Data Collection and Preprocessing**: Gather relevant data from various sources. This may include demographic information, transaction history, behavioral data, etc. Clean and preprocess the data to handle missing values, outliers, and inconsistencies.

3. **Feature Selection and Engineering**: Identify the features (variables) that are most predictive of the outcome. This involves selecting relevant variables and creating new features through transformations, binning, or other methods to enhance predictive power.

4. **Model Selection**: Choose an appropriate modeling technique based on the nature of the problem and data characteristics. Commonly used models include logistic regression, decision trees, random forests, gradient boosting, and neural networks.

5. **Model Training**: Split the data into training and validation sets. Train the selected model on the training data, optimizing model parameters to maximize predictive performance.

6. **Evaluation and Validation**: Evaluate the performance of the model using appropriate metrics such as accuracy, precision, recall, F1 score, ROC AUC, etc. Validate the model on unseen data to ensure generalizability.

7. **Calibration and Thresholding**: Calibrate the model to ensure that predicted probabilities align with observed frequencies. Determine an appropriate threshold for converting probabilities into binary outcomes (e.g., class labels or risk categories).

8. **Deployment and Monitoring**: Deploy the scoring system in production environment, integrating it into decision-making processes or workflows. Continuously monitor model performance and update the system as needed to maintain accuracy and relevance.

9. **Ethical Considerations**: Pay attention to ethical considerations such as fairness, transparency, and privacy throughout the development and deployment process. Ensure that the scoring system does not discriminate against certain groups or violate privacy regulations.

10. **Documentation and Communication**: Document the entire process including data sources, preprocessing steps, model selection, performance metrics, and deployment details. Communicate findings and recommendations to stakeholders in a clear and understandable manner.

Developing scoring systems requires a combination of domain expertise, statistical knowledge, and technical skills in data manipulation, modeling, and programming. It's an iterative process that involves experimentation, evaluation, and refinement to create effective and reliable predictive models.

- **Scoring Systems Affect Rankings:**
  - The system of scores heavily affects standings by choosing the order using the set worth.

- Higher numbers mean high ranks, highlighting top accomplishments or features. A solid score system guarantees justness and precision, showing a place's genuine place against others in a given situation or contest.

- When ranking places or entities, the scoring system determines their standings. The system assigns a value to each entity based on a set of guidelines and criteria, including factors such as importance, performance, or features.

- These values are then used to rank the entities in order, with higher scores indicating a higher rank.

- A well-designed score system ensures fairness and accuracy in ranking, providing an objective measure of each entity's performance or achievement in a given situation or competition.

- By comparing the scores of different entities, we can determine their relative positions and identify the top performers or features in a particular category or context.

3.2 Z-Score:

Z-score is a statistical measure that quantifies the distance between a data point and the mean of a dataset.

Z-score, also known as a standard score, is a statistical measurement that quantifies how many standard deviations a data point is from the mean of a dataset. It's calculated using the formula: $Z = \frac{x-\mu}{\sigma}$

Where:

- x is the value of the data point.
- μ is the mean of the dataset.
- σ is the standard deviation of the dataset.

- Z-scores are used for various purposes, including

- **Normalization**: Transforming data so that it has a mean of 0 and a standard deviation of 1. This is helpful for comparing variables with different scales.
- **Outlier Detection**: Identifying data points that are unusually far from the mean of the dataset.
- **Comparisons**: Allowing for the comparison of data points across different datasets or variables.
- **Standardizing Features**: Pre-processing data before feeding it into machine learning algorithms, particularly those that are sensitive to the scale of the input variables.

Example: Let's say you have a dataset of exam scores where the mean score is 75 and the standard deviation is 10. You want to find the Z-score for a student who scored 85 on the exam.

- Mean (μ) = 75
- Standard Deviation (σ) = 10
- Data point (x) = 85

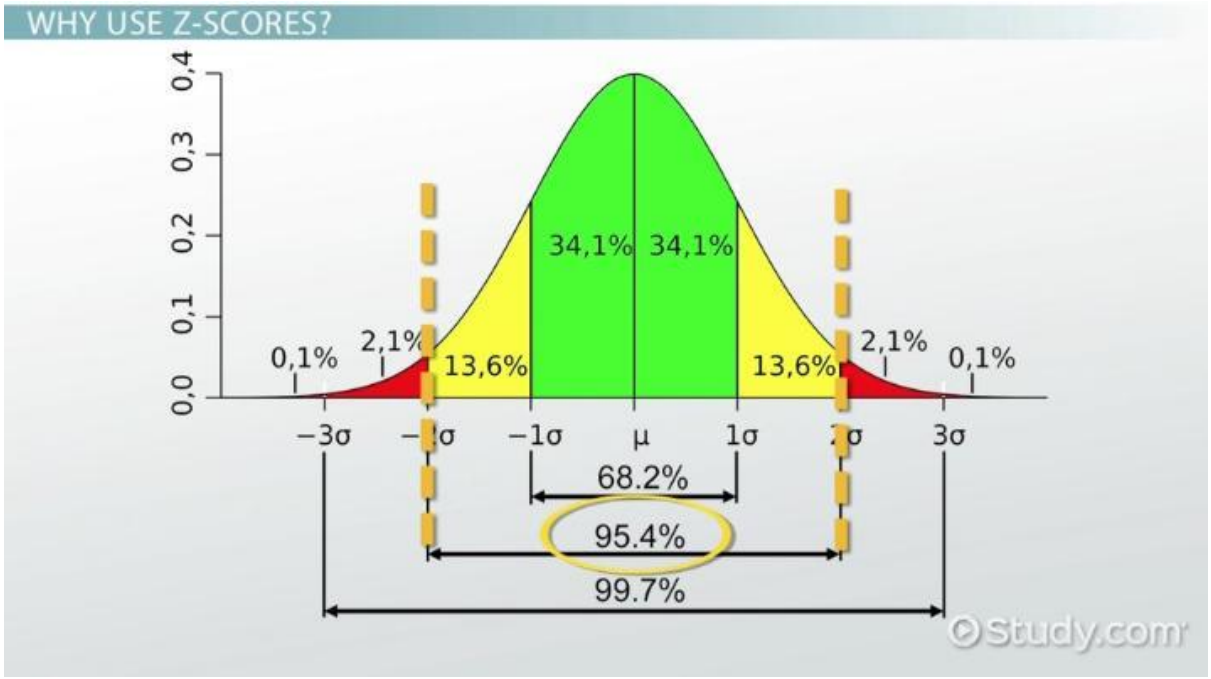
Substituting the values into the formula:

$$Z = \frac{85-75}{10} = \frac{10}{10} = 1$$

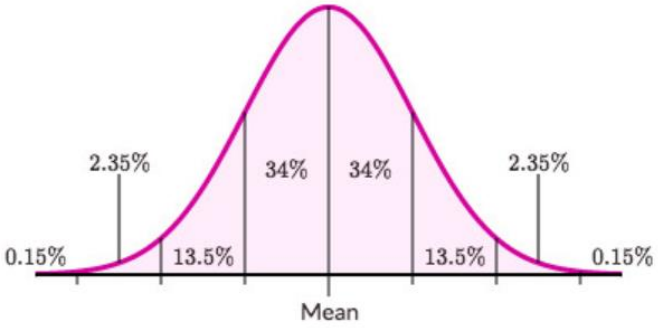
So, the Z-score for the student who scored 85 is 1. This means the student's score is 1 standard deviation above the mean.

❖ Characteristics of Z-Score

- The magnitude of the Z-score reflects how far a data point is from the mean in terms of standard deviations.
- An element having a z-score of less than 0 represents that the element is less than the mean.
- Z-scores allow for the comparison of data points from different distributions.
- An element having a z-score greater than 0 represents that the element is greater than the mean.
- An element having a z-score equal to 0 represents that the element is equal to the mean.
- An element having a z-score equal to 1 represents that the element is 1 standard deviation greater than the mean; a z-score equal to 2, 2 standard deviations greater than the mean, and so on.
- An element having a z-score equal to -1 represents that the element is 1 standard deviation less than the mean; a z-score equal to -2, 2 standard deviations less than the mean, and so on.
- If the number of elements in a given set is large, then about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2; about 99% have a z-score between -3 and 3.
- This is known as the Empirical Rule, and it states the percentage of data within certain standard deviations from the mean in a normal distribution as demonstrated in the image below



Here is an example of a normal distribution:



A normal distribution following empirical rule.

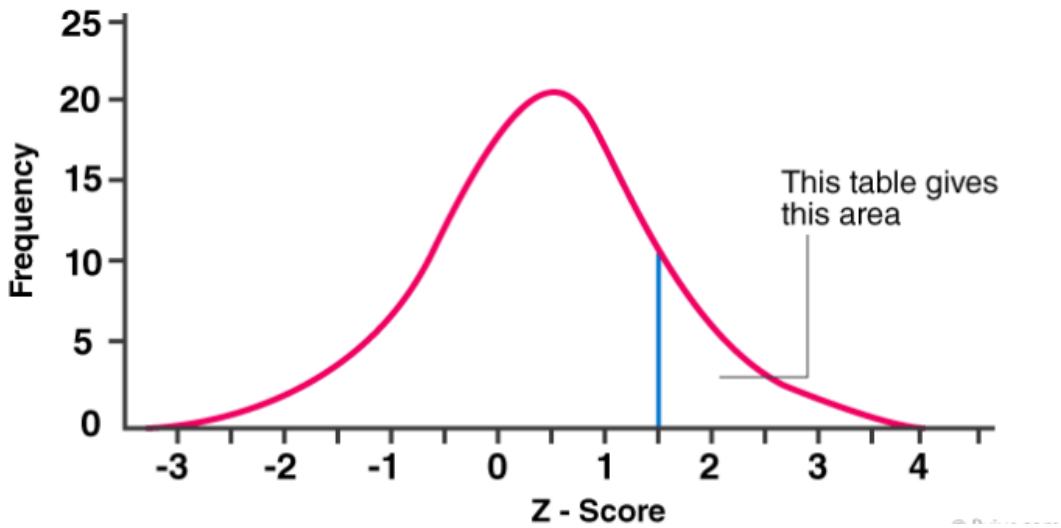
❖ Z-Scores vs. Standard Deviation

Z- Score	Standard Deviation
Transform raw data into a standardized scale.	Measures the amount of variation or dispersion in a set of values.
Makes it easier to compare values from different datasets because they take away the original units of measurement.	Standard Deviation retains the original units of measurement, making it less suitable for direct comparisons between datasets with different units.
Indicate how far a data point is from the mean in terms of standard deviations, providing a measure of the data point's relative position within the distribution	Expressed in the same units as the original data, providing an absolute measure of how spread out the values are around the mean

❖ Summary - how to **interpret z-scores**:

- A z-score of less than 0 represents an element less than the mean.
- A z-score greater than 0 represents an element greater than the mean.
- A z-score equal to 0 represents an element equal to the mean.
- A z-score equal to 1 represents an element, which is 1 standard deviation greater than the mean; a z-score equal to 2 signifies 2 standard deviations greater than the mean; etc.
- A z-score equal to -1 represents an element, which is 1 standard deviation less than the mean; a z-score equal to -2 signifies 2 standard deviations less than the mean; etc.

- If the number of elements in the set is large, about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2 and about 99% have a z-score between -3 and 3.



In data science, the Z-score is a fundamental tool that helps simplify and enhance data analysis. Here's a beginner-friendly overview of its uses:

1. **Standardizing Data**:
    - **Normalizing Data**: Z-scores transform different data sets to a common scale, with a mean of 0 and a standard deviation of 1. This is particularly useful when combining or comparing data from different sources.
2. **Identifying Outliers**:
    - **Detecting Anomalies**: A Z-score indicates how many standard deviations a data point is from the mean. Data points with Z-scores above 3 or below -3 are usually considered outliers. Identifying these outliers can highlight errors or significant variations in your data.
3. **Comparing Data**:
    - **Cross-Dataset Comparisons**: Z-scores allow you to compare values from different distributions. For instance, if you want to compare test

scores from two different exams with different scoring systems, converting the scores to Z-scores allows for an apples-to-apples comparison.

4. **Feature Scaling for Machine Learning**:
   - **Improving Model Performance**: Many machine learning algorithms, like K-nearest neighbors (KNN) and support vector machines (SVM), perform better when the input data is scaled. Z-scores standardize the features, ensuring that each feature contributes equally to the distance computations.
5. **Assessing Probability**:
   - **Understanding Distributions**: Z-scores relate to the standard normal distribution. This means you can use Z-scores to calculate probabilities and percentiles, helping you understand the likelihood of different outcomes.

How to Calculate Z-scores

To calculate a Z-score for a data point X:

$$Z = \frac{X - \mu}{\sigma}$$

where:

- X is the data point,
- μ is the mean of the data set,
- σ is the standard deviation of the data set.

Example

Imagine you have a dataset of students' test scores:

$$[70, 75, 80, 85, 90, 95, 100]$$

1. **Calculate the Mean (μ):**
   - $\mu = \frac{70+75+80+85+90+95+100}{7} = 85$
2. **Calculate the Standard Deviation (σ):**
   - $\sigma = \sqrt{\frac{(70-85)^2+(75-85)^2+(80-85)^2+(85-85)^2+(90-85)^2+(95-85)^2+(100-85)^2}{7}} \approx 10.21$
3. **Calculate the Z-score for a score of 90:**
   - $Z = \frac{90-85}{10.21} \approx 0.49$

A Z-score of 0.49 means that a score of 90 is 0.49 standard deviations above the mean.

By understanding and utilizing Z-scores, you can better analyze and interpret your data, making your data science work more effective and insightful.

❖ **Importance of Z-Score**: Why are z-scores useful in data analysis?

➢ What is a z-score?

A z-score tells you how many standard deviations away your score is from the average score.

➢ **Average score (mean):** This is like finding the middle point where most students' scores are.

➢ **Standard deviation:** This tells you how spread out all the scores are from the average. If the standard deviation is small, most scores are close to the average. If it's large, the scores are more spread out.

➢ **Why are z-scores useful?**

Comparison:

Z-scores let you compare your score to others easily. If you have a **positive z-score**, it means you scored above average. If you have a **negative z-score**, you scored below average.

Understanding Performance: It helps you understand how well you did compared to the entire class, not just whether you passed or failed.

Finding Outliers: Z-scores help identify scores that are very high or very low compared to the rest of the data, which can be interesting to investigate.

➢ **Example**

Let's say the average score on the test is 75, and the standard deviation is 10.

If you scored 85, your z-score would be +1. This means you scored one standard deviation above the average.
If you scored 65, your z-score would be -1. This means you scored one standard deviation below the average.
By looking at the z-score, you can quickly see how well you did compared to everyone else in a way that's easy to understand.

- **Positive z-score**: Data point above the mean.
- **Negative z-score**: Data point below the mean.
- **Magnitude of z-score**: Indicates how many standard deviations the data point is from the mean.

- **Meaning of - if z-score is 0:**

A z-score of 0 means that a data point is exactly at the mean (average) of the dataset. In other words, it indicates that the value is neither above nor below average; it's right in the middle. Z-scores measure how far away a data point is from the mean, in terms of standard deviations, so a z-score of 0 indicates no deviation.

3.3 Normalization

- Normalization is the process of scaling numerical data to a standard range.
- This is often done to ensure that different features or variables are comparable and to prevent some features from dominating others in machine learning algorithms.
- Normalization typically involves scaling data to have a mean of 0 and a standard deviation of 1, or to a range between 0 and 1.

- There are several common techniques for normalization:

1. **Min-Max Scaling**: This method scales the data to a fixed range, usually between 0 and 1. The formula for min-max scaling is:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

2. **Z-score Standardization**: Also known as standardization, this method scales the data to have a mean of 0 and a standard deviation of 1. The formula for z-score standardization is:

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

where μ is the mean of the data and σ is the standard deviation.

3. **Robust Scaling**: This method is similar to min-max scaling but uses the median and interquartile range (IQR) instead of the minimum and maximum values. This makes it more robust to outliers.

4. **Unit Vector Transformation**: This method scales each sample's vector to have a length of 1, which is useful for algorithms that rely on the magnitude of the feature vector.

- **Example**:

Let's consider a simple example of normalizing a dataset using the min-max scaling method.

Suppose we have a dataset consisting of exam scores ranging from 60 to 90. Here are the scores of five students:

{60,70,75, 80,90}

We want to normalize these scores to a range between 0 and 1 using the min-max scaling method.

First, we determine the minimum and maximum values in the dataset:

- Minimum score (Xmin) = 60
- Maximum score (Xmax) = 90

Next, we apply the min-max scaling formula to each score:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

For each score:

1. For $X = 60$:
   $$X_{\text{norm}} = \frac{60-60}{90-60} = \frac{0}{30} = 0$$
2. For $X = 70$:
   $$X_{\text{norm}} = \frac{70-60}{90-60} = \frac{10}{30} = \frac{1}{3}$$
3. For $X = 75$:
   $$X_{\text{norm}} = \frac{75-60}{90-60} = \frac{15}{30} = \frac{1}{2}$$
4. For $X = 80$:
   $$X_{\text{norm}} = \frac{80-60}{90-60} = \frac{20}{30} = \frac{2}{3}$$
5. For $X = 90$:
   $$X_{\text{norm}} = \frac{90-60}{90-60} = \frac{30}{30} = 1$$

So, after normalization, the scores would be:

$$\{0, \tfrac{1}{3}, \tfrac{1}{2}, \tfrac{2}{3}, 1\}$$

- This means that the lowest score in the dataset is mapped to 0, the highest score is mapped to 1, and the other scores are linearly scaled between 0 and 1 based on their relative positions in the original range.

- **Statistical Analysis**

3.4 Sampling from Distributions

- **Definition:**

❖ Sampling:

Sampling is the process of selecting a subset of data from a larger dataset or a population. This smaller subset is called a sample. In data science, we often work with samples because it's usually impractical or impossible to analyze an entire population.

Types of Sampling:

Sampling is a way of selecting a smaller group from a larger population to study and understand the bigger group. Here are some simple explanations of different types of sampling in data science:

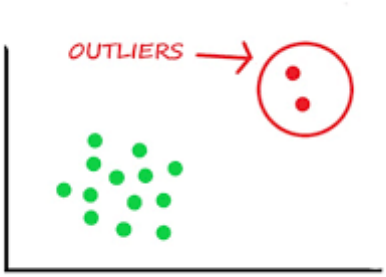
❖ Distribution:

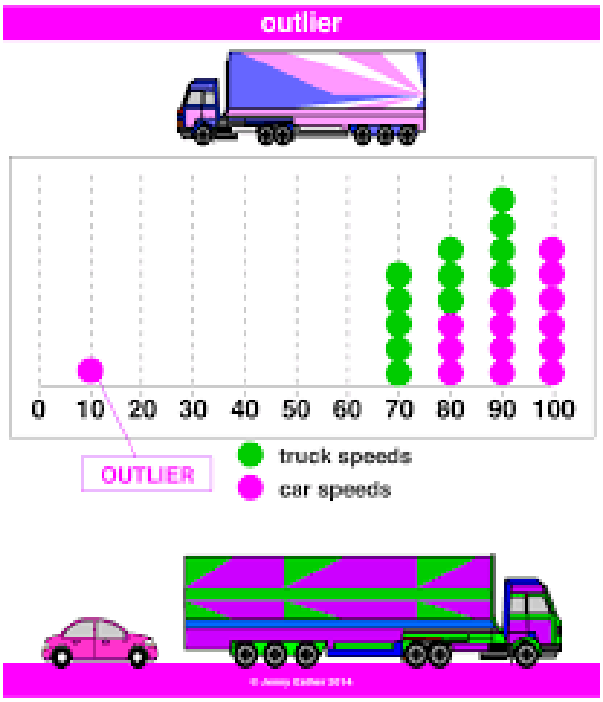
A distribution is a mathematical function that describes how data points are spread out or distributed. It tells us the probability of different outcomes.

❖ Outliers

- Definition of outliers: An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

  An outlier is a single data point that goes far outside the average value of a group of statistics.



- For example, in a group of 5 students the test grades were 9, 8, 9, 7, and 2. The last value seems to be an outlier because it falls below the main pattern of the other grades.
- An outlier is an extreme value in a data set that is either much larger or much smaller than all the other values.
- Example:
- 

1. **Types of Sampling from Distributions**

1. **Normal Distribution**: Also known as the bell curve, it's symmetric around the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. The normal distribution appears as a "bell curve" when graphed.

```
import numpy as np
# Sampling from a normal distribution with mean 0 and standard deviation 1
sample = np.random.normal(0, 1, 1000)
```

2. **Uniform Distribution**: Every outcome is equally likely, such as when you roll a die one time, the probability that it falls on a number between 1 and 6 follows a uniform distribution because each number is equally likely to occur.

```
import numpy as np
# Sampling from a uniform distribution between 0 and 1
sample = np.random.uniform(0, 1, 1000)
```

3. **Binomial Distribution**:

- Sampling from a binomial distribution, which represents the number of successes in a fixed number of independent Bernoulli trials.
- Describes the number of successes in a fixed number of binary (yes/no) trials. Binomial distribution is a common discrete distribution used in statistics, as opposed to a continuous distribution, such as normal distribution. This is because binomial distribution only counts two states, typically represented as 1 (for a success) or 0 (for a failure), given a number of trials in the data.
- For example, if we toss a coin, there could be only two possible outcomes: heads or tails, and if any test is taken, then there could be only two results: pass or fail. This distribution is also called a binomial probability distribution. There are two parameters n and p used here in a binomial distribution.

```
import numpy as np
```

```
# Sampling from a binomial distribution with 10 trials and success probability 0.5
sample = np.random.binomial(10, 0.5, 1000)
```

4. **Poisson Distribution**: Sampling from a Poisson distribution, which represents the number of events occurring in a fixed interval of time or space.

```
import numpy as np
# Sampling from a Poisson distribution with mean 3
sample = np.random.poisson(3, 1000)
```

5. **Exponential Distribution**: Sampling from an exponential distribution, which represents the time between events in a Poisson process

```
import numpy as np
# Sampling from an exponential distribution with rate parameter 0.5
sample = np.random.exponential(1/0.5, 1000)
```

3.5 Statistical distributions with examples:

They provide mathematical descriptions of the probability of different outcomes in various scenarios. Understanding these distributions is essential for tasks such as hypothesis testing, parameter estimation, and uncertainty modeling. Here are some common statistical distributions used in data science:

- **Normal Distribution (Gaussian Distribution)**:
  - **Description**: Bell-shaped distribution characterized by its mean and standard deviation.
  - **Example**
  - In finance, stock returns often follow a normal distribution, enabling risk assessment and portfolio optimization.

- In quality control, measurements like product weights or lengths might be normally distributed, aiding in process monitoring and control.

- **Binomial Distribution**:
  - **Description**: Models the number of successes in a fixed number of independent Bernoulli trials with the same probability of success.
  - **Example**
    - A/B testing in marketing to determine the effectiveness of different website designs or marketing strategies by measuring conversion rates.
    - Predicting the outcome of binary classification tasks in machine learning, such as whether an email is spam or not.
    -

- **Poisson Distribution**:
  - **Description**: Models the number of events occurring in a fixed interval of time or space given a constant average rate.
  - **Example**
    - Analyzing website traffic, where the number of page visits in a given time period can be modeled using a Poisson distribution.
    - Predicting the number of customer arrivals at a service point, such as a call center, in a given time frame.

- **Exponential Distribution**:
  - **Description**: Describes the time between consecutive events in a Poisson process, where events occur at a constant average rate.
  - **Example**
    - Modeling the time between customer arrivals at a store or website for optimizing staffing or server capacity.
    - Predicting the time until failure of a machine or equipment component in reliability engineering.

- **Uniform Distribution**:
  - **Description**: All outcomes within a specified range are equally likely.
  - **Example**
    - Generating random numbers for simulations or random sampling tasks.
    - Initializations in machine learning algorithms, such as random weights in neural networks.

3.6 Statistical significance

- Statistical significance refers to the claim that a set of observed data are not the result of chance but can instead be attributed to a specific cause.
- Statistical significance is important for academic disciplines or practitioners that rely heavily on analysing data and research, such as economics, finance, investing, medicine, physics, and biology.
- Statistical significance can be considered strong or weak.
- When analysing a data set and performing the necessary tests to discern whether one or more variables affect an outcome, strong statistical significance helps support the fact that the results are real and not caused by luck or chance.
- Simply stated, if a p-value is small, then the result is considered more reliable.

Key points:

1. Statistical significance refers to the claim that a result from data generated by testing or experimentation is likely to be attributable to a specific cause.
2. A high degree of statistical significance indicates that an observed relationship is unlikely to be due to chance.
3. The calculation of statistical significance is subject to a certain degree of error.
4. Statistical significance can be misinterpreted when researchers do not use language carefully in reporting their results.
5. Several types of significance tests are used depending on the research being conducted.

3.7 Types of Statistical Significance Tests

- **Null Hypothesis Testing: Permutation Tests and P-values**

- **Null Hypothesis**: Start with the idea that nothing special is happening.

- The **null hypothesis (H0)** is a fundamental concept in statistics and data science used to test whether there is a significant effect or relationship present in the data.
- It generally represents a statement of no effect or no difference.
- For example, if you are testing whether a new drug is effective, the null hypothesis might be that the drug has no effect on patients compared to a placebo.

1. **Permutation Test**:
   - Mix things up many times to see what would happen by random chance.
   - **Permutation tests** are a non-parametric method used to test the null hypothesis.
   - They involve repeatedly rearranging the data to create a distribution of the test statistic under the null hypothesis.

   - Here's how permutation tests work:
   - **Calculate the observed test statistic**:
   - This could be the difference in means, correlation coefficient, or another relevant statistic.
   - **Shuffle the data**: Randomly permute the data labels or values to generate a new dataset under the assumption that the null hypothesis is true.
   - **Calculate the test statistic for the permuted data**: Compute the same test statistic for this new, shuffled dataset.
   - **Repeat**: Repeat the shuffling and test statistic calculation many times (e.g., 10,000 permutations) to build a distribution of the test statistic under the null hypothesis.
   - **Compare**: Determine where the observed test statistic falls in this distribution to assess the significance.
   -

     ❖ **Definition and Interpretation**

2. **P-value**: Check if your findings are surprising enough to say, "Yes, there's something special happening!"

- A **p-value** is the probability of obtaining test results at least as extreme as the observed results, assuming the null hypothesis is true. It quantifies the evidence against the null hypothesis:

- **Low p-value (typically < 0.05)**: Indicates strong evidence against the null hypothesis, suggesting that the observed effect is unlikely to have occurred by chance.
- **High p-value (typically ≥ 0.05)**: Indicates weak evidence against the null hypothesis, suggesting that the observed effect could plausibly have occurred by chance.

In the context of permutation tests, the p-value is calculated by:

1. Counting how many times the permuted test statistic is as extreme or more extreme than the observed test statistic.
2. Dividing this count by the total number of permutations.

For example, if you performed 10,000 permutations and in 300 of them, the permuted test statistic was as extreme or more extreme than the observed test statistic, the p-value would be = 300/10,000 = 0.03

Module IV

Data Visualization and Mathematical Models

Module IV

Data Visualization: Basic principles—ideas and tools for data visualization
Visualizing Data: Exploratory Data Analysis—Developing a Visualization Aesthetic—Chart Types
Mathematical Models: Philosophies of Modelling—A Taxonomy of Models—Baseline Models in ML—Evaluating Models

4.1 Data visualization:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. The goal is to make complex data more understandable and usable by transforming it into a visual context. This can involve:

- **Charts** (e.g., bar charts, line charts, pie charts)
- **Graphs** (e.g., scatter plots, histograms)
- **Maps** (e.g., heat maps, geographical maps)
- **Infographics** (combining graphics and text to tell a story)
- **Dashboards** (interactive platforms that display multiple visualizations)

Effective data visualization helps in decision-making, identifying correlations, and communicating insights clearly to a broad audience, regardless of their data analysis proficiency.

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

❖ **Steps in data visualization?**

- Step 1 — Be clear on the question. ...
- Step 2 — Know your data and start with basic visualizations. ...
- Step 3 — Identify messages of the visualization, and generate the most informative.
- Step 4 — Choose the right chart type. ...
- Step 5 — Use color, size, scale, shapes and labels to direct attention to the key.

4.2 Basic principle of data visualization:

- It is a way to communicate complex information in a visual and intuitive manner, making it easier for people to understand and analyze the data. By transforming raw data into visual representations, data visualization allows patterns, trends, and insights to be easily identified and interpreted.

Data visualization is a powerful tool that helps to communicate information clearly and effectively through graphical representations. Here are some key points and simple examples to get beginners started with data visualization:

Key Points:

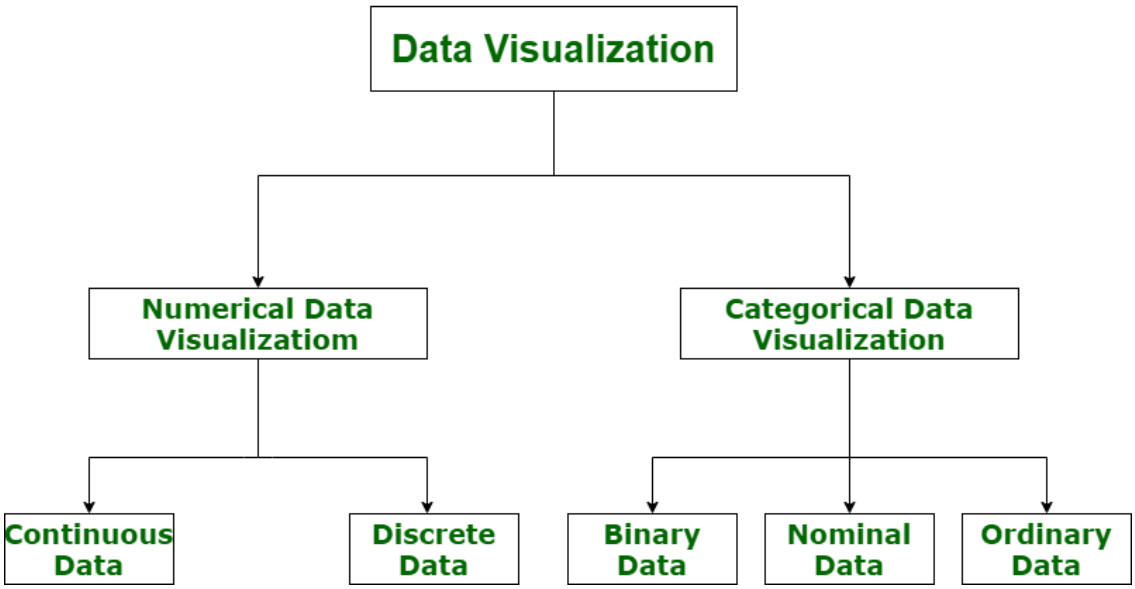
1. **Understand Your Data**: Before creating any visualizations, get familiar with your data. Understand its structure, types of variables, and key metrics.
2. **Choose the Right Chart**: Different types of data require different types of charts. Choosing the right chart helps to convey your message more effectively.
3. **Keep It Simple**: Avoid clutter. Simple, clean visualizations are often the most effective.
4. **Use Color Wisely**: Colors can highlight key points, but too many colors can be distracting. Use a consistent color palette.
5. **Label Clearly**: Ensure all axes, legends, and data points are clearly labelled so the viewer can easily understand the chart.

❖ **Types of Data for Visualization**

Performing accurate visualization of data is very critical to market research where both numerical and categorical data can be visualized, which helps increase the impact of insights and helps in reducing the risk of analysis paralysis. So, data visualization is categorized into the following categories:

- Numerical Data
- Categorical Data

Let's understand the visualization of data via a diagram with its all categories.



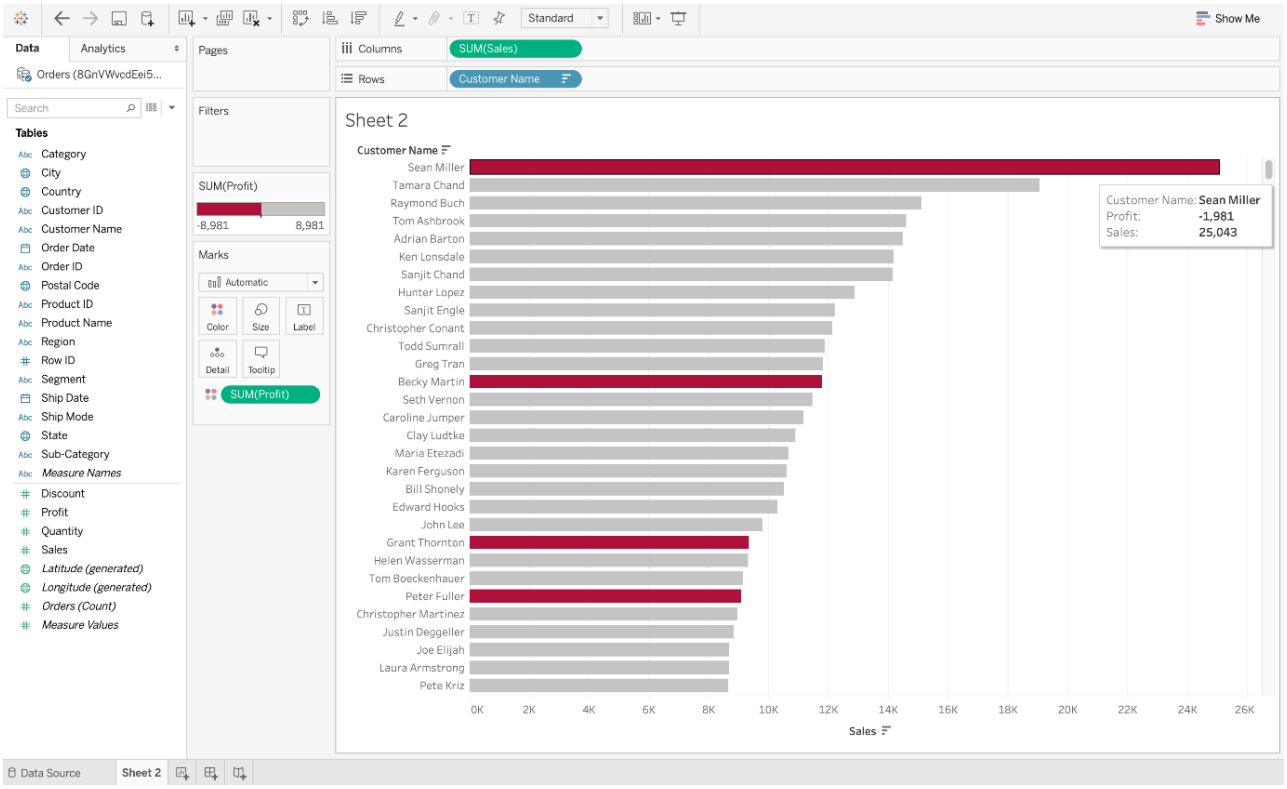
❖ Why is Data Visualization Important?

Let's take an example. Suppose you compile visualization data of the company's profits from *2013 to 2023* and create a line chart. It would be very easy to see the line going constantly up with a drop in just 2018. So you can observe in a second that the company has had continuous profits in all the years except a loss in 2018.

It would not be that easy to get this information so fast from a data table. This is just one demonstration of the usefulness of data visualization. Let's see some more reasons why visualization of data is so important.

1. **Data Visualization Discovers the Trends in Data**

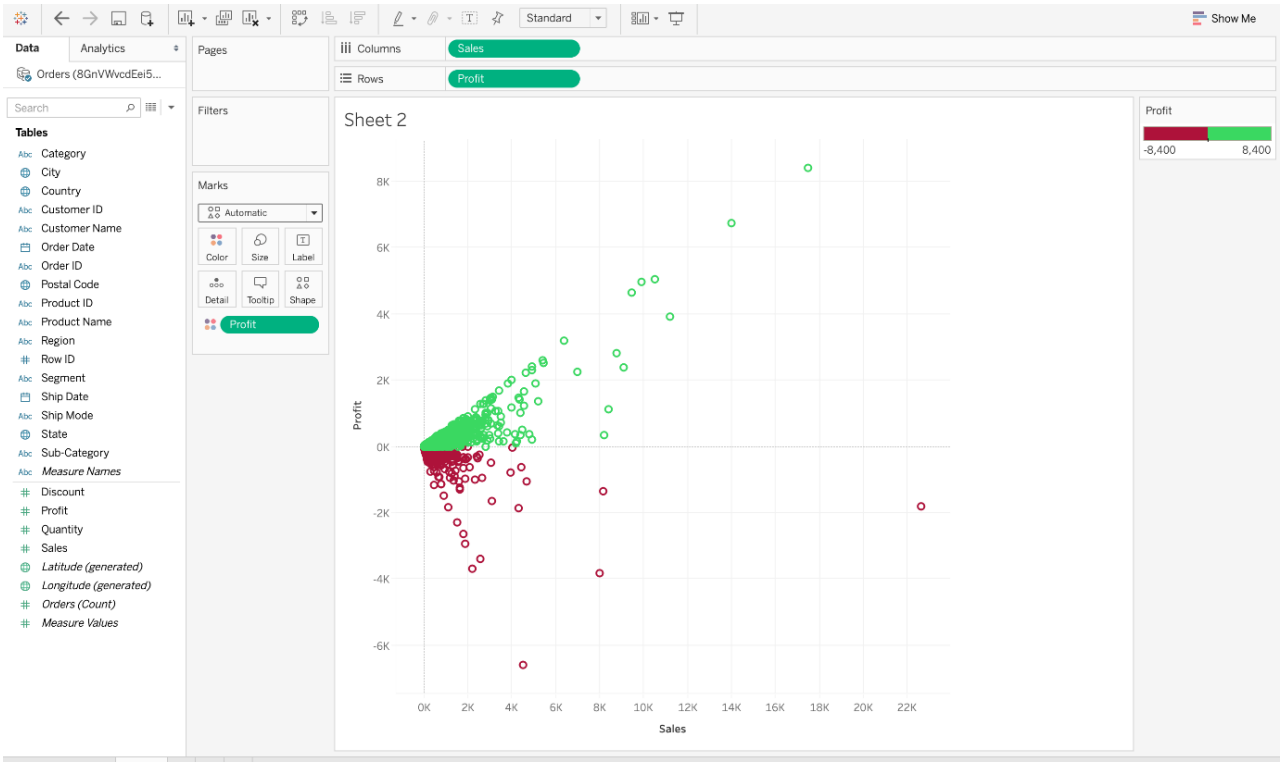
The most important thing that data visualization does is discover the trends in data. After all, it is much easier to observe data trends when all the data is laid out in front of you in a visual form as compared to data in a table. For example, the screenshot below on visualization on Tableau demonstrates the sum of sales made by each customer in descending order. However, the color red denotes loss while grey denotes profits. So it is very easy to observe from this visualization that even though some customers may have huge sales, they are still at a loss. This would be very difficult to observe from a table.



2.**Data Visualization Provides a Perspective on the Data**

Visualizing Data provides a perspective on data by showing its meaning in the larger scheme of things. It demonstrates how particular data references stand concerning

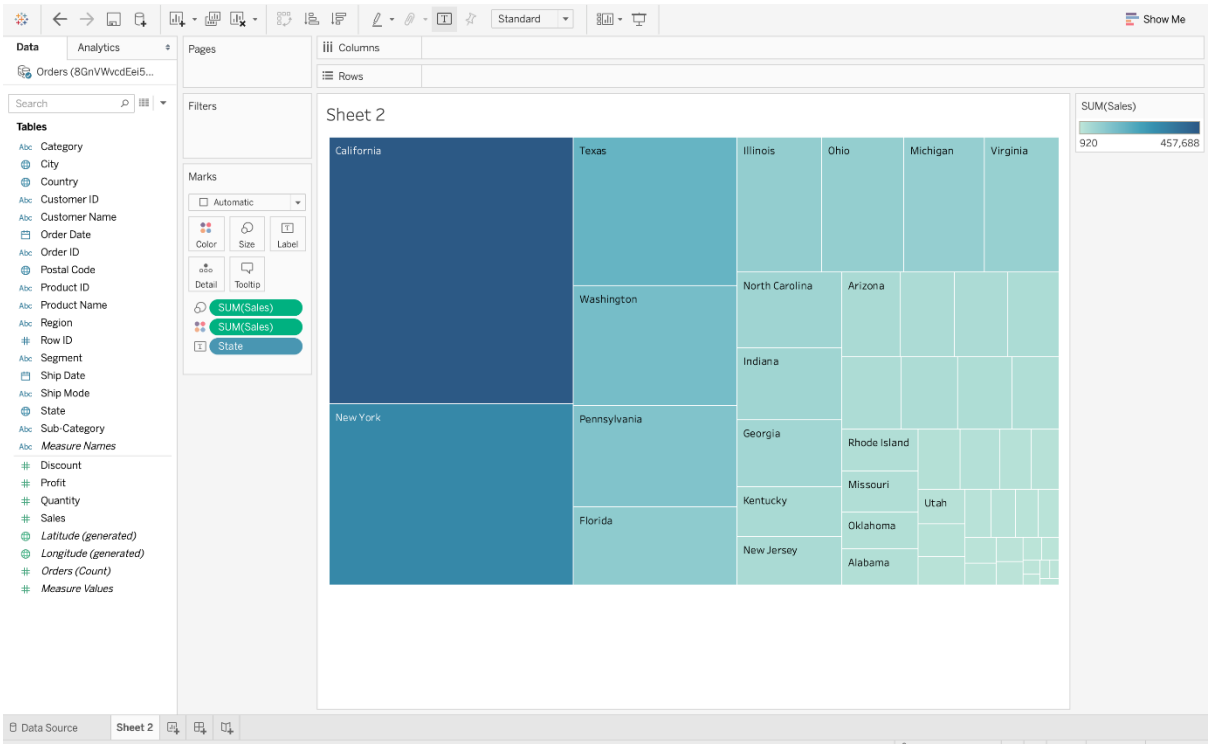
the overall data picture. In the data visualization below, the data between sales and profit provides a data perspective concerning these two measures. It also demonstrates that there are very few sales above 12K and higher sales do not necessarily mean a higher profit.



3.Data Visualization Puts the Data into the Correct Context

It isn't easy to understand the context of the data with data visualization. Since context provides the whole circumstances of the data, it is very difficult to grasp by just reading numbers in a table. In the below data **visualization on Tableau**,

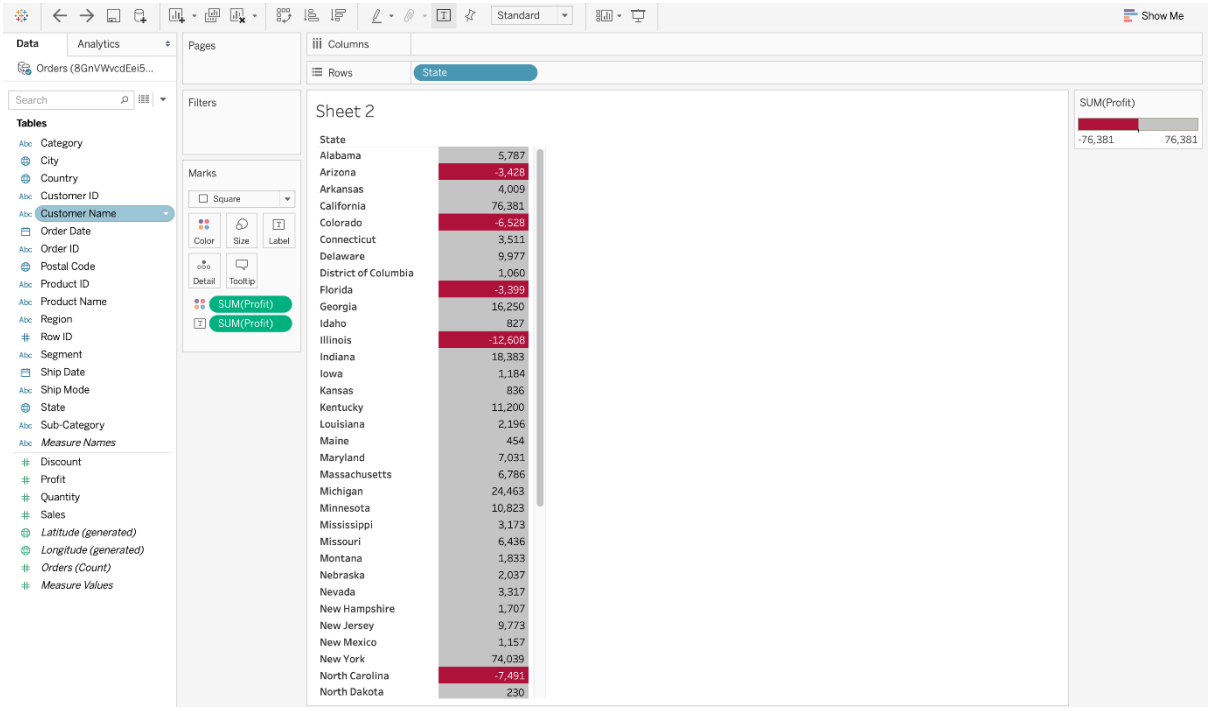
a **TreeMap** is used to demonstrate the number of sales in each region of the United States. It is very easy to understand from this data visualization that California has the largest number of sales out of the total number since the rectangle for California is the largest. But this information is not easy to understand outside of context without visualizing data.



4.**Data Visualization Saves Time**

It is definitely faster to gather some insights from the data using data visualization rather than just studying a chart. In the screenshot below on Tableau, it is very easy to identify the states that have suffered a net loss rather than a profit. This is because all the cells with a loss are colored red using a heat map, so it is obvious states have suffered a loss. Compare this to a normal table where you would need to check each

cell to see if it has a negative value to determine a loss. Visualizing Data can save a lot of time in this situation.



5.**Data Visualization Tells a Data Story**

Data visualization is also a medium to tell a data story to the viewers. The visualization can be used to present the data facts in an easy-to-understand form while telling a story and leading the viewers to an inevitable conclusion. This data story, like any other type of story, should have a good beginning, a basic plot, and an ending that it is leading towards. For example, if a data analyst has to craft a data visualization for company executives detailing the profits of various products, then

the data story can start with the profits and losses of multiple products and move on to recommendations on how to tackle the losses.

Data Visualization can effectively communicate complex information, engage the audience, support decision-making, and provide an excellent user experience, thereby maximizing the impact of the data being presented.

- Effective data visualization relies on **12 key design principles** that help convey information accurately and efficiently.

1. Clarity

The visualization should be clear and easily understood by the intended audience.

2. Simplicity

Keep the visualization simple and avoid unnecessary complexity.

3. Purposeful

Understand what message or insight you want to communicate and design for that purpose.

4. Consistency

Maintain consistency in the design elements throughout the visualization.

5. Contextualization

Provide context for the data being presented.

6. Accuracy

Ensure the visualization accurately represents the underlying data.

7. Visuals Encoding

Choose appropriate visual encodings for the data types you are visualizing.

8. Intuitiveness

Design the visualization to be intuitive and easy to comprehend.

9. Interactivity

Consider adding interactive elements to the visualization, such as tooltips, zooming, filtering, or highlighting.

10. Aesthetics

Although aesthetics are subjective, a visually appealing design can engage viewers and increase their interest in the data.

11. Accessibility

Accessibility is key; if users can't read the data, it's useless.

12. Hierarchy

Work out hierarchy of information early on and always remind yourself of what the purpose of representing the data is.

4.3 Ideas & Tools of Data visualization:

1. **Python Libraries:**
   - **Matplotlib:** A versatile plotting library for creating static, animated, and interactive visualizations in Python.
   - **Seaborn:** Built on top of Matplotlib, Seaborn provides a high-level interface for drawing attractive and informative statistical graphics.
   - **Pandas Plotting:** Pandas, a data manipulation library, offers simple plotting capabilities for quick visualizations directly from DataFrames.
   - **Plotly:** A powerful library for creating interactive plots and dashboards in Python. Plotly supports a wide range of chart types and is particularly useful for web-based visualizations.
   - **Bokeh:** Another library for creating interactive visualizations, Bokeh is designed to produce elegant and concise graphics with minimal effort.
   - **Altair:** Declarative statistical visualization library in Python, which creates simple and intuitive visualizations with a concise syntax.
2. **R Programming:**
   - **ggplot2:** A popular package in R for creating elegant and complex visualizations based on the grammar of graphics principles.
   - **Plotly:** Besides Python, Plotly also has an R interface, providing similar capabilities for creating interactive plots.

- **Lattice:** Offers a powerful set of functions for creating conditioned plots, which are particularly useful for exploring relationships in multivariate data.

3. **JavaScript Libraries:**
   - **D3.js (Data-Driven Documents):** A JavaScript library for producing dynamic, interactive data visualizations in web browsers. D3.js provides unparalleled flexibility but requires more coding compared to other libraries.
   - **Chart.js:** A simple yet versatile JavaScript library for creating responsive and visually appealing charts on web pages.
   - **Highcharts:** Another JavaScript library for creating interactive charts and graphs. Highcharts is known for its wide range of chart types and ease of use.
4. **Business Intelligence (BI) Tools:**
   - **Tableau:** A popular BI tool that offers intuitive drag-and-drop functionality for creating interactive dashboards and visualizations.
   - **Power BI:** Microsoft's BI tool, which integrates seamlessly with Excel and other Microsoft products, allowing users to create powerful visualizations and reports.
5. **Specialized Tools:**
   - **Plotly Dash:** A Python framework for building analytical web applications, Dash enables the creation of interactive dashboards using Plotly visualizations.
   - **QlikView/Qlik Sense:** Business intelligence and data visualization platforms that provide powerful analytics capabilities and interactive visualizations for exploring data.

   - **Excel**: Widely used and accessible. Good for basic data visualization and analysis. Supports various chart types and basic interactive elements.

4.4 Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often using visual methods. It is a crucial step in the data analysis process as it helps in understanding the data's underlying structure, identifying patterns, spotting anomalies, testing hypotheses, and checking assumptions with the help of summary statistics and graphical representations.

- Key Steps in EDA

1. **Data Collection and Cleaning**
    - **Data Collection**: Gathering the raw data from various sources.
    - **Data Cleaning**: Handling missing values, removing duplicates, correcting errors, and dealing with outliers.
2. **Data Summary and Descriptive Statistics**
    - **Summary Statistics**: Calculating measures such as mean, median, mode, standard deviation, variance, skewness, and kurtosis.
    - **Data Types**: Understanding the types of variables (categorical, numerical, ordinal).
3. **Data Visualization**
    - **Univariate Analysis**: Examining each variable individually. Common plots include histograms, box plots, and bar charts.
    - **Bivariate Analysis**: Analyzing the relationship between two variables. Common plots include scatter plots, correlation matrices, and pair plots.
    - **Multivariate Analysis**: Investigating more than two variables simultaneously. Techniques include using scatter plot matrices and heatmaps.
4. **Identifying Patterns and Relationships**
    - **Correlation Analysis**: Checking for relationships between variables using correlation coefficients.
    - **Group Comparisons**: Using techniques like t-tests or ANOVA to compare groups.

5. **Data Transformation and Feature Engineering**
    - **Data Transformation**
    - : Applying transformations like normalization or scaling to bring data to a common scale.
    - **Feature Engineering**: Creating new features from existing ones to improve model performance.

➢ Tools and Techniques

- **Summary Statistics**: Measures of central tendency and variability.
- **Visualization Tools**: Matplotlib, Seaborn, Plotly for Python; ggplot2 for R.
- **Data Manipulation Libraries**: Pandas for Python; dplyr for R.
- **Statistical Tests**: Hypothesis testing, ANOVA, chi-square tests.

➢ Importance of EDA

1. **Understanding Data**: Provides a clear understanding of the data structure, distribution, and relationships.
2. **Data Quality Issues**: Helps in identifying and correcting data quality issues such as missing values, outliers, and anomalies.
3. **Hypothesis Generation**: Allows the formulation of hypotheses that can be tested in further analysis.
4. **Informing Further Analysis**: Guides the selection of appropriate statistical techniques and models.
5. **Communicating Insights**: Effective visualization can communicate findings and insights to stakeholders clearly and concisely.

➢ Example Workflow of EDA

1. **Load the data**: Import the data into a data frame.
2. **Inspect the data**: Check the first few rows, data types, and summary statistics.
3. **Clean the data**: Handle missing values and correct errors.
4. **Visualize the data**: Use plots to understand distributions and relationships.

5. **Analyse the data**: Calculate correlations, group statistics, and perform hypothesis testing.
6. **Feature engineering**: Create new variables if necessary.

- EDA is a fundamental step in the data analysis pipeline that lays the groundwork for more complex analyses. By summarizing the main characteristics of the data, identifying patterns, and visualizing key relationships, EDA provides essential insights that inform decision-making and further statistical modelling.

❖ **Handling missing values in EDA**

First step we do in exploratory data analysis when use a new dataset is to check the values , which can be replacing it with mode simply, some using very complex methods Some of the methods are

1. Replacement with Mean/Median/Mode
2. Random Sample Imputation
3. Capturing NAN with new feature
4. End of Distribution Imputation
5. Arbitrary Imputation
6. Frequent Category Imputation (only used with Categorical feature)
7. Make 'NAN' a new Category(only used with Categorical feature)

Types of missing data.

There are three types of missing data

- MCAR: means Missing completely at random. The probability of a missing data value is independent of any observation in the data set means it has no relation with the data, it just miss at random places.
- MNAR: means Missing Not at Random. For example, when most of the missing people from work are sickest people, people with the lowest education are missing on education, this kind of missing is referred as Missing Not at Random (MNAR).
- MAR: means Missing at Random

Checking for missing values during Exploratory Data Analysis (EDA) is crucial for several reasons:

1. **Data Quality Assessment**: Identifying missing values helps in assessing the quality and completeness of the dataset. A high percentage of missing data might indicate issues with data collection or storage processes.
2. **Bias Prevention**: Missing data can introduce bias if not handled properly. For instance, if certain values are missing systematically (e.g., more missing values in one category than another), this can lead to skewed analyses and incorrect conclusions.
3. **Informed Decision-Making**: Understanding the pattern and extent of missing data can inform decisions about data cleaning strategies, such as whether to impute missing values or to remove certain rows or columns entirely.
4. **Impact on Analysis**: Many statistical analyses and machine learning algorithms require complete data without missing values. Identifying and addressing missing data is necessary to ensure that these analyses are accurate and reliable.
5. **Model Performance**: Missing values can affect the performance of predictive models. Handling them appropriately can improve model accuracy and robustness.
6. **Insight into Data Collection**: Patterns of missing data can provide insights into the data collection process, helping to identify potential areas for improvement in data acquisition and storage.

4.5 Developing a Visualization Aesthetic:

- Aesthetics is obvious, it's the visual appeal of a graphic, but functionality is less obvious.
- Graphics have a functional purpose, which is to highlight patterns and trends in data in a visual way.
- A graphic is functionally successful when it is easy to understand the patterns in the data.
- The fundamental components of aesthetic visualization:
- Commonly used aesthetics in data visualization: position, shape, size, color, line width, line type.

- Some of these aesthetics can represent both continuous and discrete data (position, size, line width, color) while others can usually only represent discrete data (shape, line type).

Aesthetics describe every aspect of a given graphical element. A few examples are provided in Figure below. A critical component of every graphical element is of course its position, which describes where the element is located. In standard 2d graphics, we describe positions by an x and y value, but other coordinate systems and one- or three-dimensional visualizations are possible. Next, all graphical elements have a shape, a size, and a color. Even if we are preparing a black-and-white drawing, graphical elements need to have a color to be visible, for example black if the background is white or white if the background is black. Finally, to the extent we are using lines to visualize data, these lines may have different widths or dash–dot patterns. Beyond the examples shown here, there are many other aesthetics we may encounter in a data visualization. For example, if we want to display text, we may have to specify font family, font face, and font size, and if graphical objects overlap, we may have to specify whether they are partially transparent.

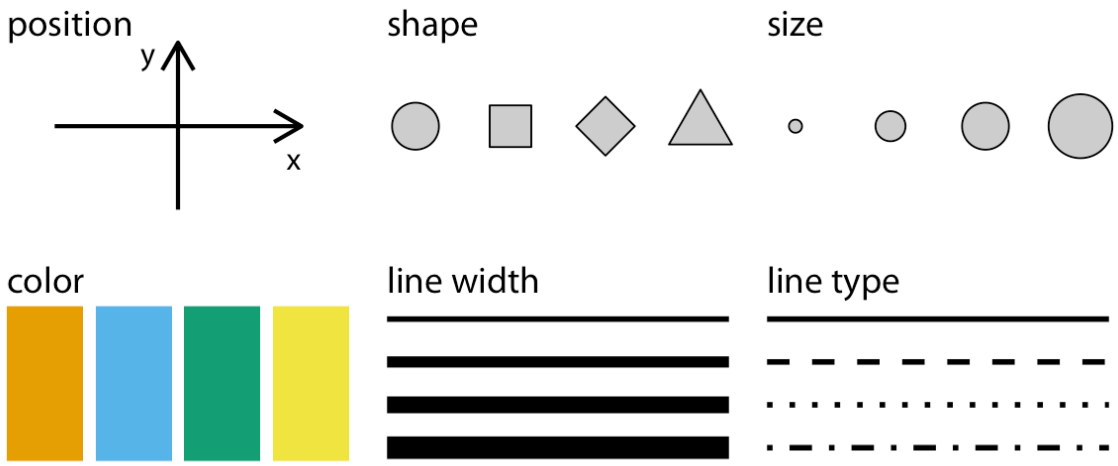


Fig: Commonly used aesthetics in data visualization: position, shape, size, color, line width, line type.

- Map the temperature onto the y axis, day of the year onto the x axis, location onto color, and visualize this aesthetics with solid lines. The result is a standard line plot showing the temperature normal at the four locations as they change during the year (Figure).

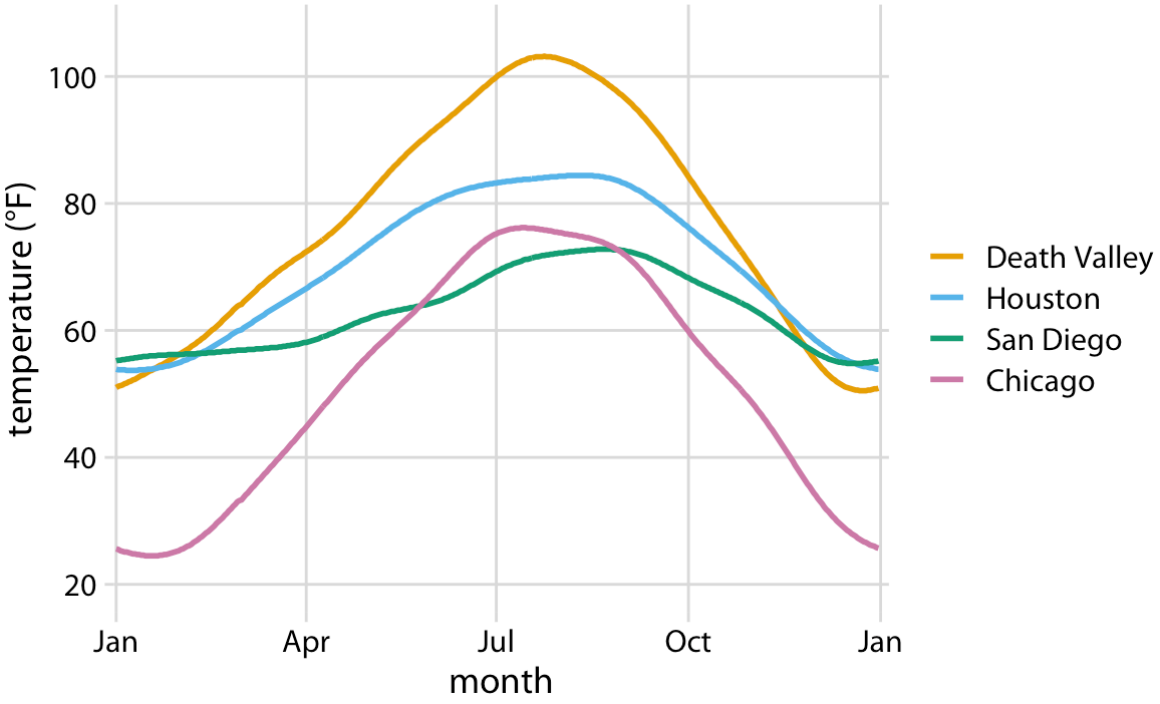


Fig. Daily temperature normals for four selected locations in the U.S. Temperature is mapped to the *y* axis, day of the year to the *x* axis, and location to line color.

4.6 Chart types:

1. Bar Chart

- Description: A bar chart displays data using rectangular bars, where the length of each bar represents the value.
- Use Case: Comparing different categories or groups.
- Example: Sales by product category.

- Key properties of Bar Graph are mentioned below:

• Every bar graph has a uniform width which is used to analyze data according to different points.

• It can be either horizontal or vertical.

• Every bar graph has two axes, one for the Graph and the other for the quantity of the data.

• The graph shows the comparison of data over a particular time.

Parts of a Bar Graph

The main parts of a bar graph include:

• Title: Describes the purpose or subject of the graph.

• X-axis (horizontal axis): Represents the categories or groups being compared.

• Y-axis (vertical axis): Displays the values or quantities corresponding to each category.

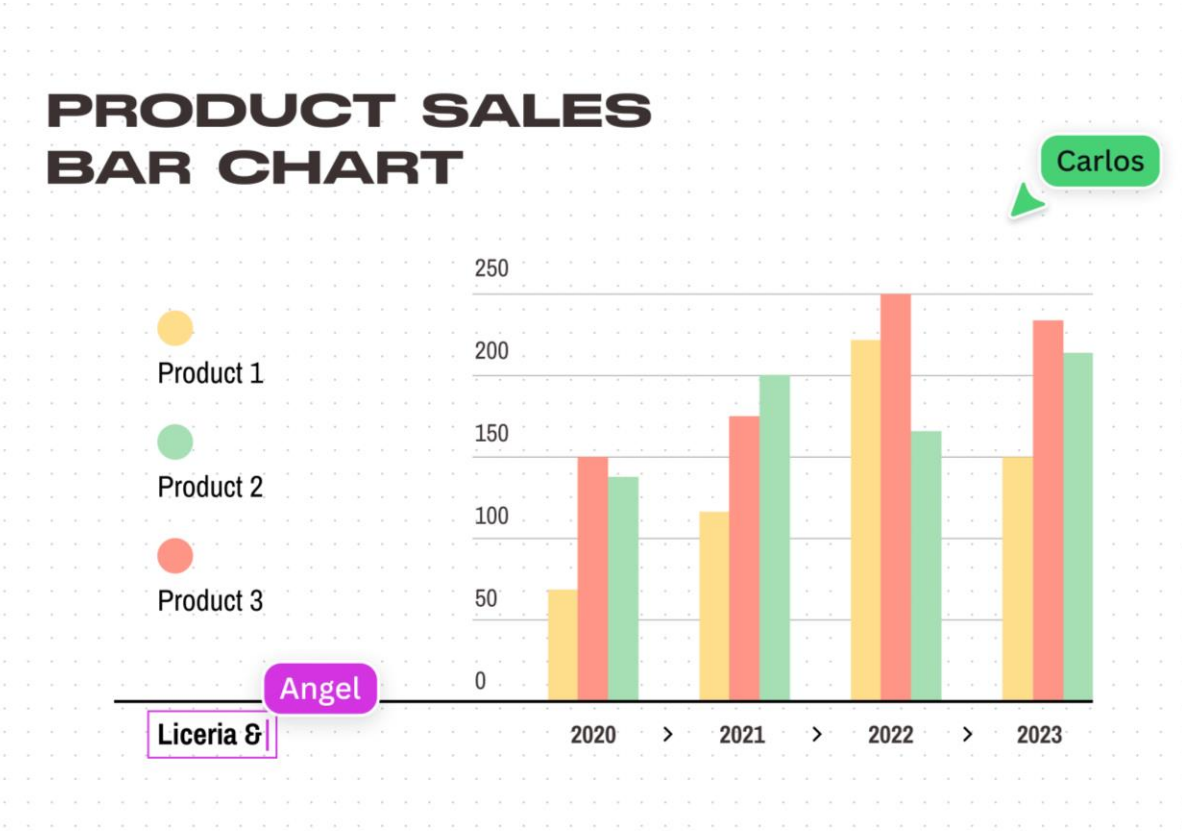
• Bars: Vertical or horizontal rectangles representing the data values for each category.

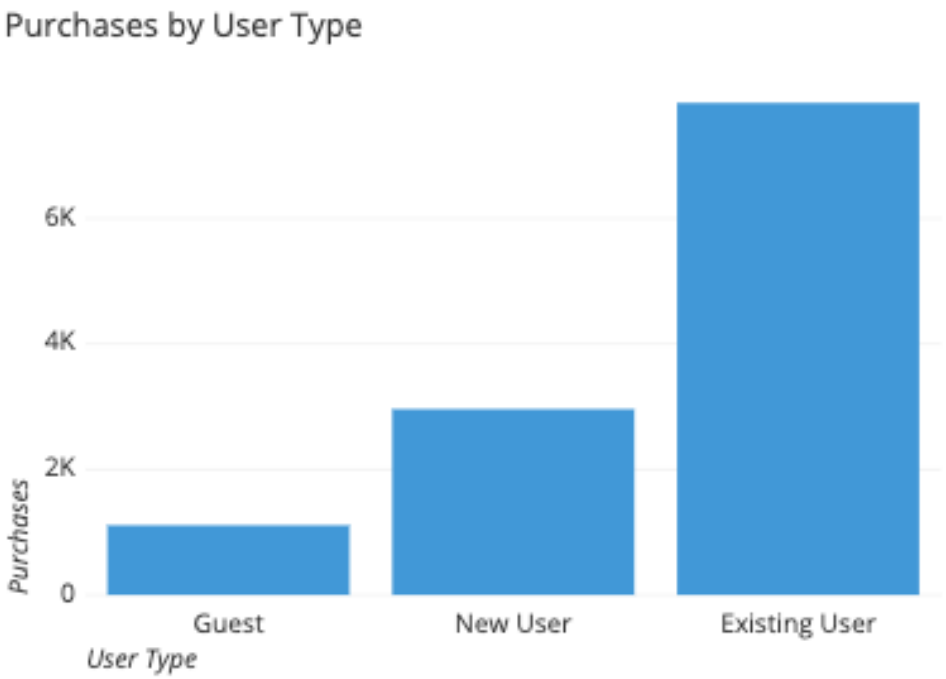
• Data labels: Numerical values attached to the bars to show the exact measurement.

• Legend: Explains the meaning of different colours or patterns if multiple data sets are presented.

• Scale: The units or intervals used on the axes to measure and represent the data accurately.

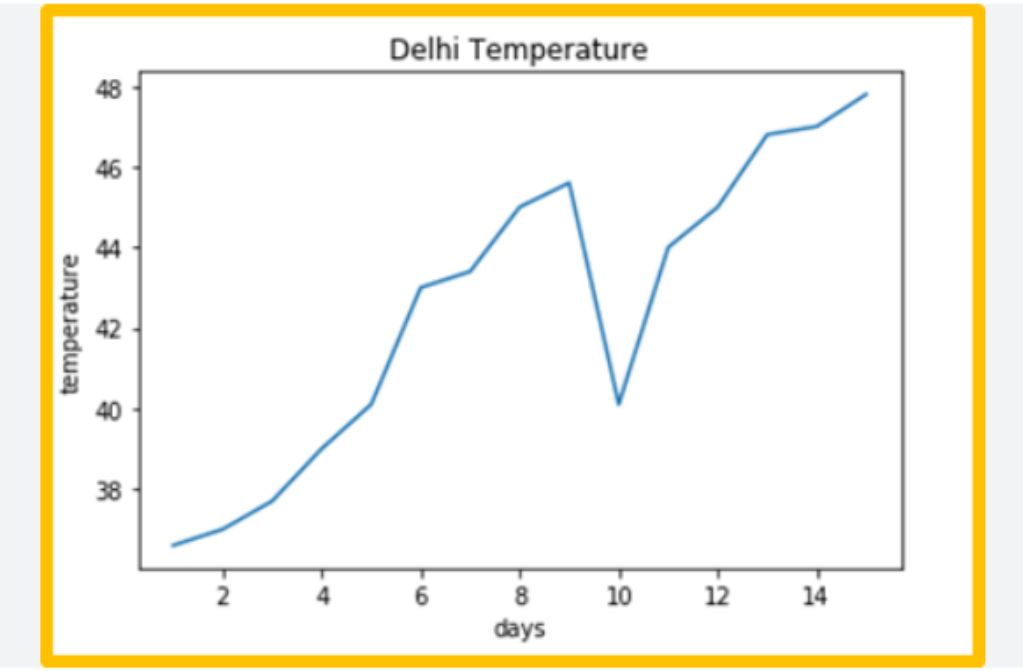
Few examples:


PRODUCT SALES
BAR CHART
Carlos
Product 1
Product 2
Product 3
Angel
Liceria &
250
200
150
100
50
0
2020
2021
2022
2023


Purchases by User Type
6K
4K
2K
0
Purchases
Guest
New User
Existing User
User Type

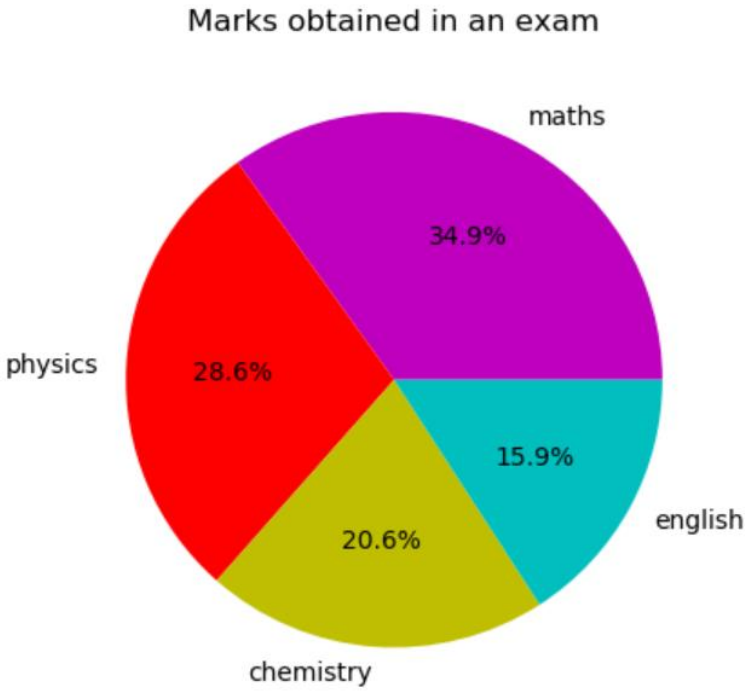
2. Line Chart

- Description: A line chart connects data points with a continuous line, often used to show trends over time.
- Use Case: Tracking changes over periods of time.
- Example: Monthly revenue over a year, temperature, your heart rate throughout the day and a company's daily sales.



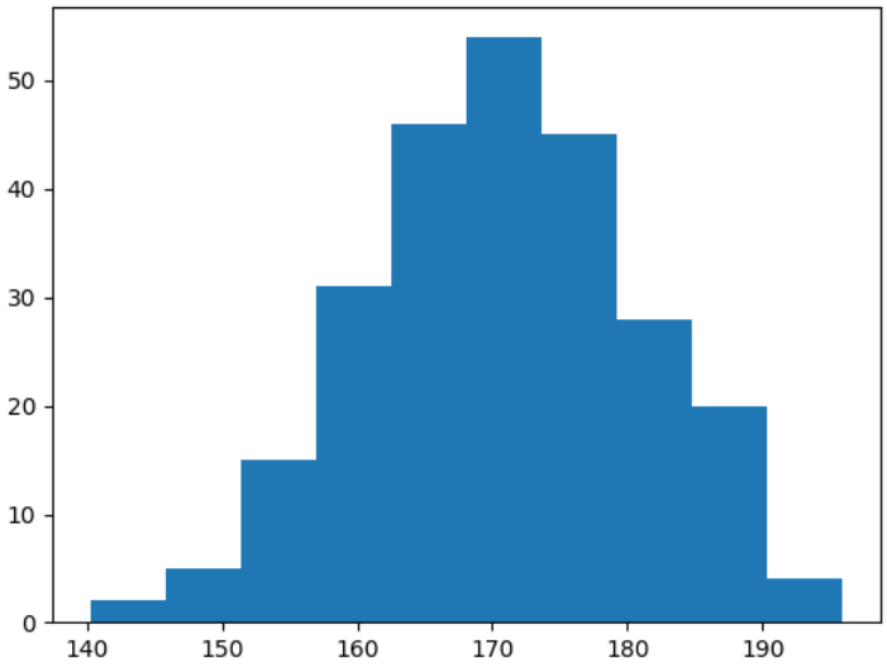
3. Pie Chart

- Description: A pie chart shows data as slices of a circle, where each slice represents a proportion of the whole.
- The pie chart is an important type of data representation. It contains different segments and sectors in which each segment and sector of a pie chart forms a specific portion of the total(percentage). The sum of all the data is equal to 360°. The total value of the pie is always 100%.
- Use Case: Showing percentage or proportional data.
- Example: Market share of different companies.



4. Histogram

- Description: A histogram displays the distribution of a dataset, showing the frequency of data points within specified ranges.
- A histogram is a graph showing frequency distributions.
- It is a graph showing the number of observations within each given interval.
- Example: Say you ask for the height of 250 people, you might end up with a histogram like this:



You can read from the histogram that there are approximately:

2 people from 140 to 145cm

5 people from 145 to 150cm

15 people from 151 to 156cm

31 people from 157 to 162cm

46 people from 163 to 168cm

53 people from 168 to 173cm

45 people from 173 to 178cm

28 people from 179 to 184cm

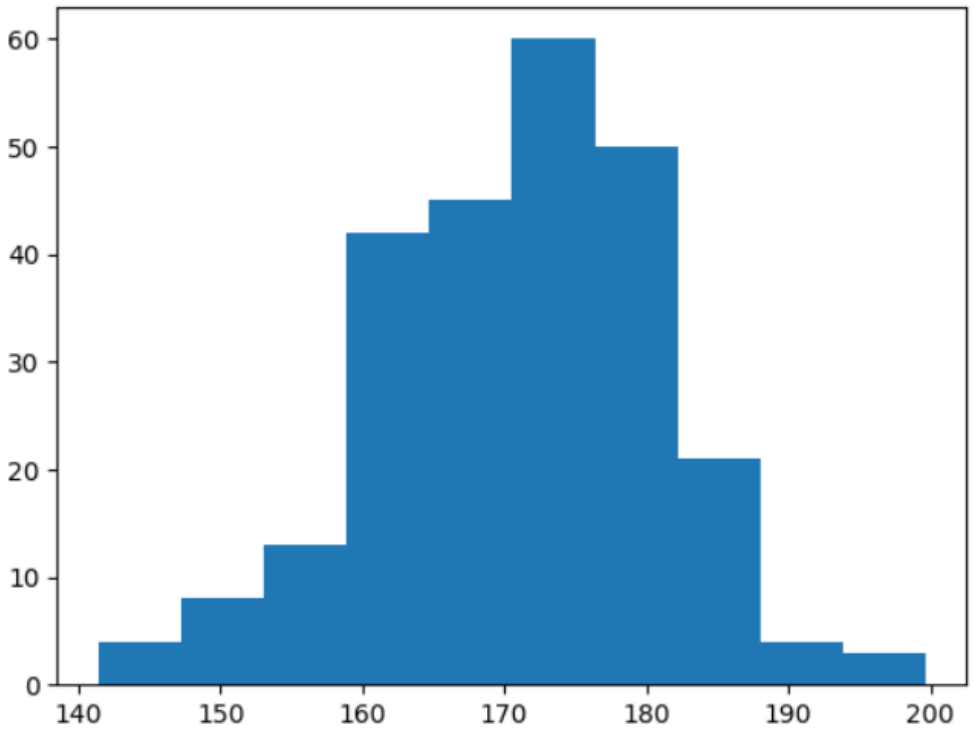
21 people from 185 to 190cm

4 people from 190 to 195cm

Use Case: Understanding the distribution of data.

Example: Distribution of test scores in a class.

```
1 import matplotlib.pyplot as plt
2 import numpy as np
3
4 x = np.random.normal(170, 10, 250)
5
6 plt.hist(x)
7 plt.show()
```



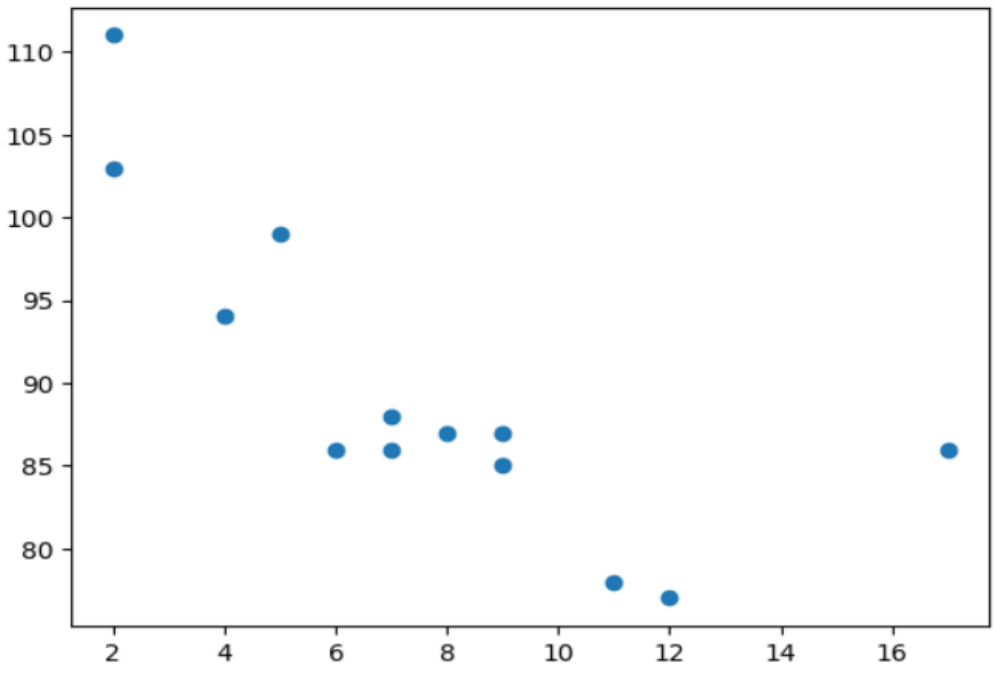
4. Scatter Plot

- Description: A scatter plot shows individual data points on a two-dimensional graph, with one variable on each axis.
- Scatter plots are the graphs that present the relationship between two variables in a data-set.
- The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.
- Use Case: Identifying relationships or correlations between variables.
- Example: Height vs. weight of individuals.
- The collected data of the temperature and humidity can be presented in the form of a scatter plot.

```python
#scatter plot
import matplotlib.pyplot as plt

x = [5,7,8,7,2,17,2,9,4,11,12,9,6]
y = [99,86,87,88,111,86,103,87,94,78,77,85,86]

plt.scatter(x, y)
plt.show()
```

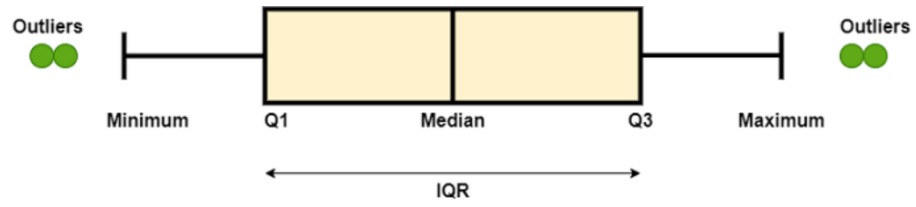


6. Box Plot (Box-and-Whisker Plot)

- **Description:** A box plot summarizes data distribution using a five-number summary: minimum, first quartile, median, third quartile, and maximum.

- A **Box Plot** is also known as **Whisker plot** is created to display the summary of the set of data values having properties like minimum, first quartile, median, third quartile and maximum. In the box plot, a box is created from the first quartile to the third quartile, a vertical line is also there which goes through the box at the median. Here x-axis denotes the data to be plotted while the y-axis shows the frequency distribution.
- **Use Case:** Displaying the spread and skewness of data.
- **Example:** Distribution of exam scores.

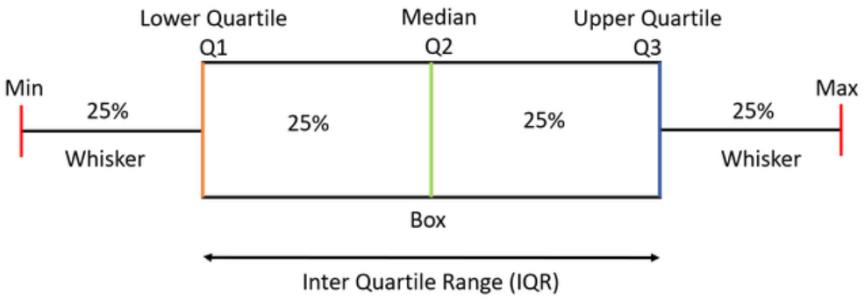
Box Plot visualization

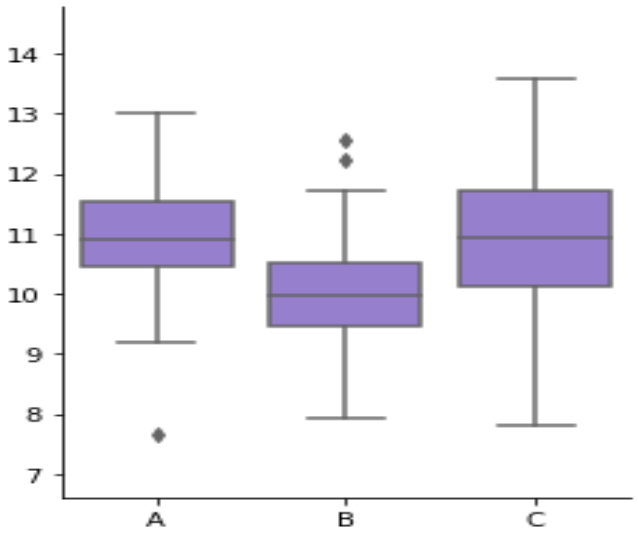


Elements of Box Plot

A box plot gives a five-number summary of a set of data which is-

- **Minimum** – It is the minimum value in the dataset excluding the outliers.
- **First Quartile (Q1)** – 25% of the data lies below the First (lower) Quartile.
- **Median (Q2)** – It is the mid-point of the dataset. Half of the values lie below it and half above.
- **Third Quartile (Q3)** – 75% of the data lies below the Third (Upper) Quartile.
- **Maximum** – It is the maximum value in the dataset excluding the outliers.


Lower Quartile
Q1
Median
Q2
Upper Quartile
Q3
Min
Max
25%
25%
25%
25%
Whisker
Whisker
Box
Inter Quartile Range (IQR)


14
13
12
11
10
9
8
7
A
B
C

7. **Heatmap**

Description: A heatmap represents data in a matrix format where values are indicated by color.

Use Case: Visualizing data density or intensity.

Example: Correlation matrix.

We can use a Heatmap to Visualize the Correlation Between Variables:

A heat map is a two-dimensional representation of data in which various values are represented by colors.

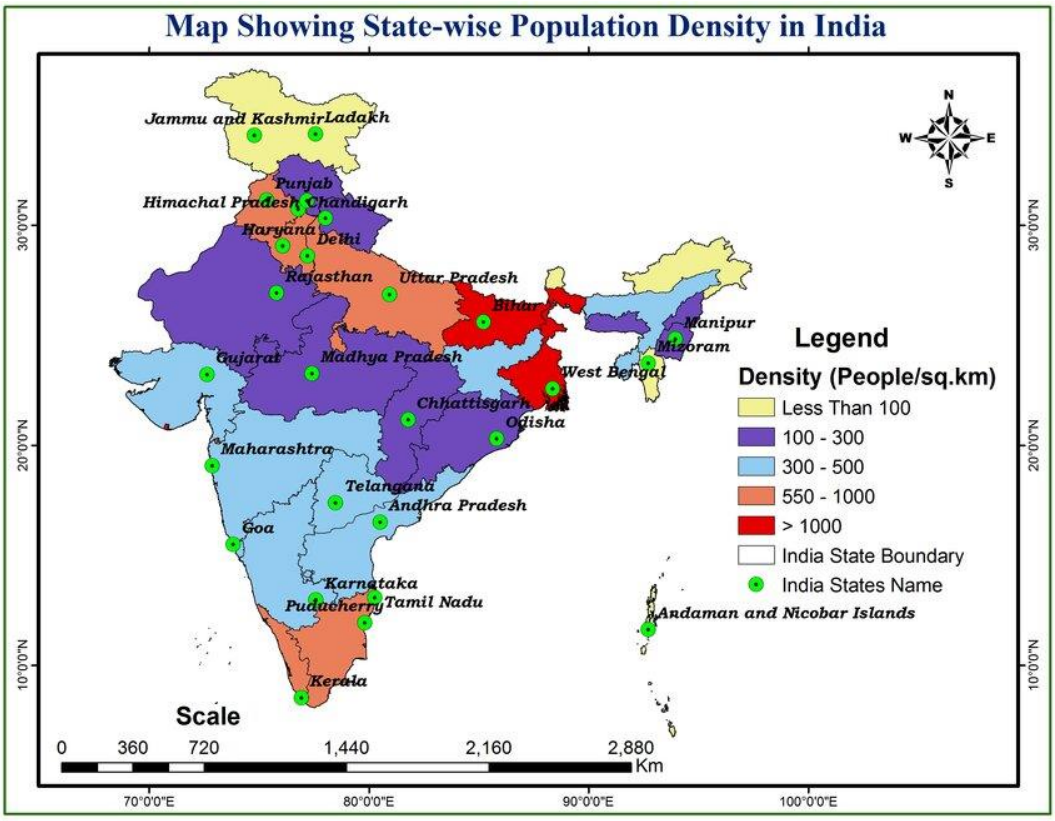
An example of a heat map is one that represents population

density. In this type of map, different colors represent different densities of population within a particular country or area.

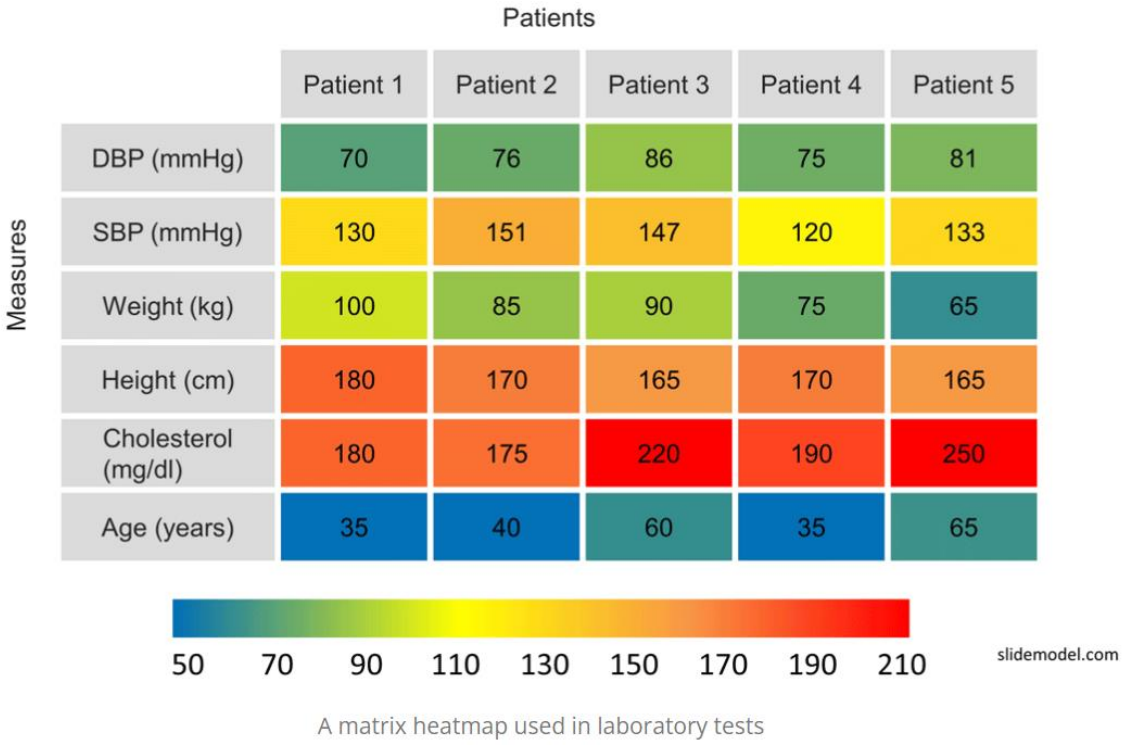
Geographical Heatmap:

A geographical heatmap is a spatial map to visualize data according to geographical location.

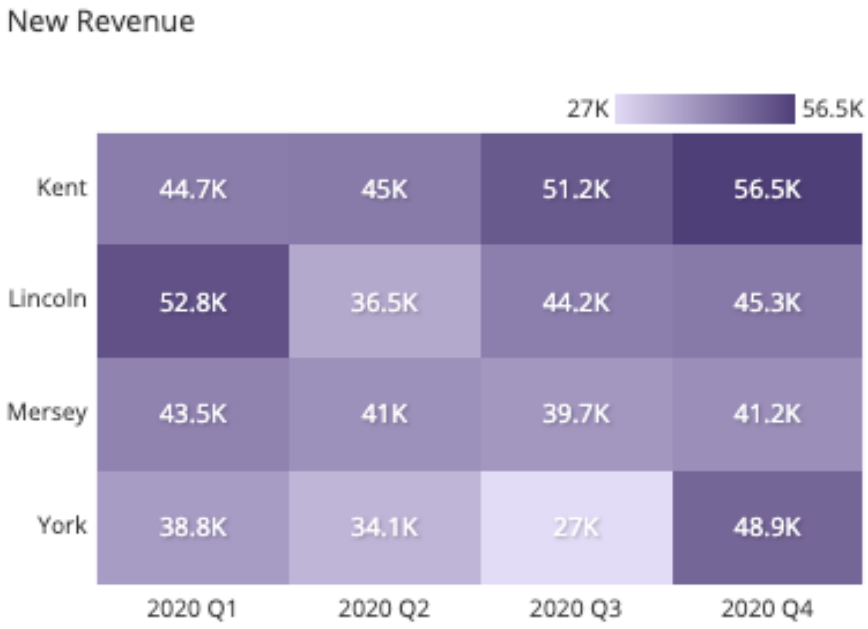
This can be done to show the phenomenon's intensity, such as weather trends/state population.



Matrix heatmap example:



A matrix heatmap used in laboratory tests



➢ Tips for Choosing the Right Chart:

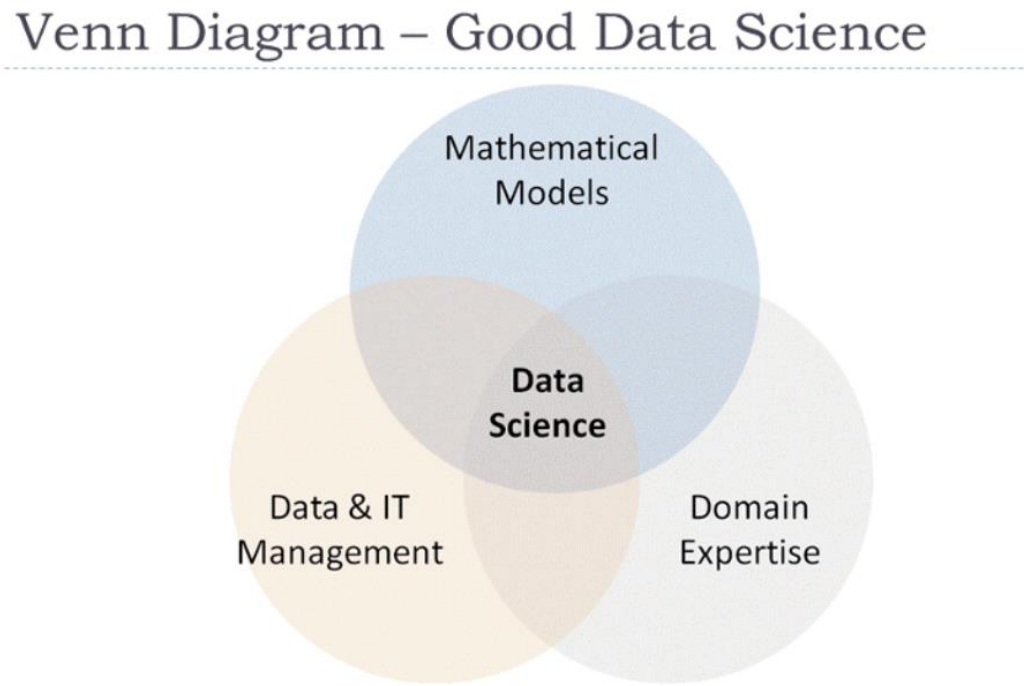
1. **Know Your Data:** Understand the type of data you have and what you want to convey.
2. **Audience:** Consider who will be viewing the chart and their familiarity with data visualization.
3. **Clarity:** Ensure the chart is not cluttered and the information is clear.
4. **Simplicity:** Avoid overcomplicating; use the simplest form that conveys the message effectively.

❖ **Role of a heatmap in multivariate analysis:**

A heatmap is a powerful visualization tool used in multivariate analysis to represent the relationship between multiple variables. It provides a graphical representation of data where individual values contained in a matrix are represented as colors. Here are some key roles that heatmaps play in multivariate analysis:

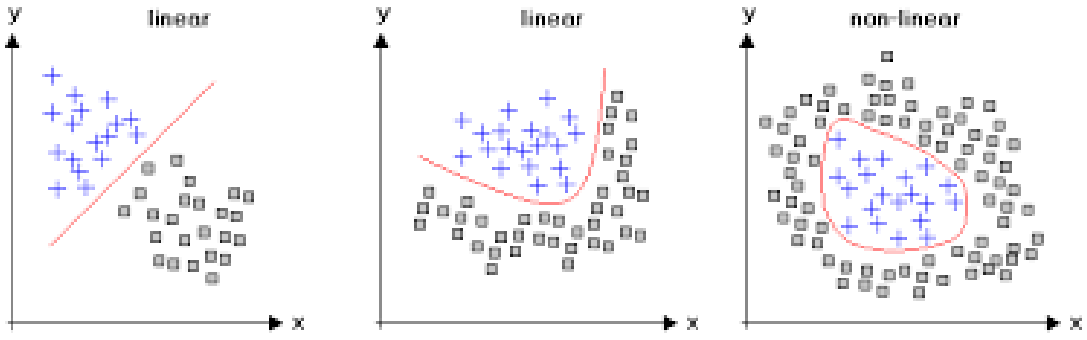
1. **Visualizing Correlations**: Heatmaps are often used to display correlation matrices, allowing analysts to quickly identify patterns, trends, and relationships between variables. Different colors represent different levels of correlation, making it easy to spot strong or weak correlations.
2. **Identifying Clusters**: By visualizing similarities and differences between data points, heatmaps can help identify clusters or groups within the data. Clustering is useful for segmenting data into meaningful groups, which can be further analyzed.
3. **Detecting Outliers**: Heatmaps can highlight outliers or anomalies in the data. Unusual patterns or colors that deviate from the rest of the dataset can indicate potential outliers that may require further investigation.
4. **Dimensionality Reduction**: Heatmaps can help in reducing the complexity of data by providing a clear visual summary of large datasets. They allow for a quick assessment of the overall structure of the data without getting bogged down by details.
5. **Comparing Multiple Variables**: Heatmaps enable comparison across multiple variables simultaneously, making them ideal for multivariate analysis. This is particularly useful in fields like genomics, finance, and social sciences, where datasets often contain many variables.
6. **Highlighting Relationships**: They can reveal relationships between different variables that might not be immediately apparent through numerical analysis alone. This can lead to new insights and hypotheses for further investigation.
7. **Enhancing Data Interpretation**: By providing a visual representation, heatmaps make it easier for analysts to interpret complex data, identify trends, and communicate findings effectively to others who may not be as familiar with the dataset.

4.7 Mathematical Models:



1. Linear vs. Non-Linear Models

- Linearity refers to the property of a system or model where the output is directly proportional to the input, while nonlinearity implies that the relationship between input and output is more complex and cannot be expressed as a simple linear function.



2. Deterministic vs Stochastic models

- Deterministic models are based on precise inputs and produce the same output for a given set of inputs. These models assume that the future can be predicted with certainty based on the current state. On the other hand, stochastic models incorporate randomness and uncertainty into the modeling process.

Stochastic Models	Deterministic Models
Accounts for uncertainty and randomness	No uncertainty or randomness
Random variation in the inputs	No random inputs
Estimate of probability of various outcomes	One solution to a specific set of values
Different results every time	Results are always the same

Deterministic vs. Stochastic Model:

Deterministic models are constantly offers certain output for a given set of fixed input variables. Hence the output continually falls with in a given specified range. Stochastic models might also not usually offers the identical output for a given set of enter variables, because it incorporates some randomness. A deterministic model means settled model. Deterministic model are often described by differential equations.

Stochastic models are random analysis because randomness is present, and Multiple runs are used to estimate probability distribution.

The benefit of stochastic models are they can predict the patterns comparable to practical patterns. considering the fact that most of the real structures frequently surprises us through special outcome, this may also be due we do not understand them completely. so the use of deterministic strategy to find out about the real or complicated structures are no longer really useful for most of the time.

Stochastic models are random analysis because randomness is present, and Multiple runs are used to estimate probability distribution.

The benefit of stochastic models are they can predict the patterns comparable to practical patterns. considering the fact that most of the real structures frequently surprises us through special outcome, this may also be due we do not understand them completely. so the use of deterministic strategy to find out about the real or complicated structures are no longer really useful for most of the time.

3. Static Vs. Dynamic Models

3. Static vs. Dynamic Model:

Static Model	Dynamic Model
Static means fixed.	Dynamic means change.
Output is determined only by the current input, reacts instantaneously.	Output takes time to react.
Relationship does not change.	Relationship changes with time, depend on past inputs and initial conditions
Relationship is represented by an algebraic equation.	Relationship is represented by an Differential equation.
	We require future input or past input.

4. Discrete vs. Continuous Model:

- Discrete models do not take into account the function of time and usually uses time-advance methods". A discrete model is that if the values belonging to the set are distinct and separate. Age, Height, Shoe size change in pocket number of books in a bag pack.
- Continuous models typically are represented with f(t) and the changes are reflected over continuous time intervals. A continuous model is that if the values belonging to the given set of data can take on any value within a finite or infinite interval. True height, true weight, time, speed, temperature, volume etc.

Continued….

- Attributes or Discrete
  - Counts of Names or labels; # good, # bad …# red, # white, # blue…
  - Counts Rank ordered; #good, #better, #best…

- Variables or Continuous
  - Measurements; Volts, distance, time, price…
  - Ratios, proportions; RPM, MPG, % butterfat…

5. <u>Qualitative Vs Quantitative Model</u>

- The difference between qualitative and quantitative data in data science:

Quantitative data is numbers-based, countable, or measurable.

Qualitative data is interpretation-based, descriptive, and relating to language.

Quantitative data tells us how many, how much, or how often in calculations.

Examples

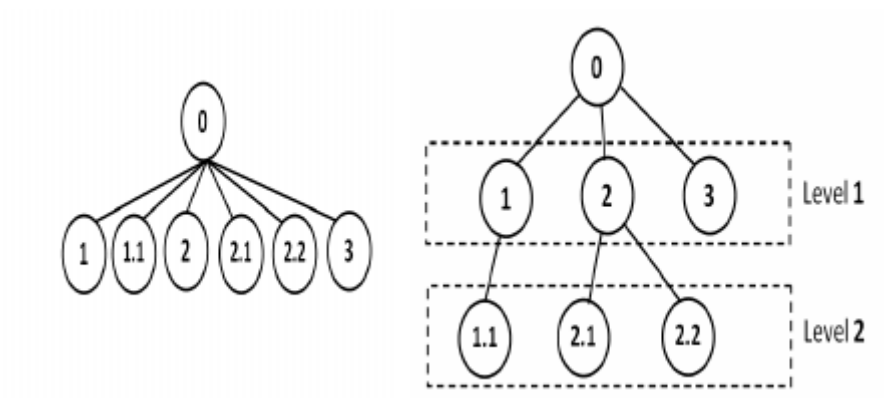
- Number of rabbits in a field. (D)
- Size of cars gas tank. (C)
- Number of text messages sent today. (D)
- Number of doughnuts eaten this week. (D)
- The level of lead in water. (C)
- Number of goals scored in a soccer match. (D)
- Length of a leaf. (C)
- No of languages spoken. (D)
- No of books on a shelf. (D)
- No of people in a family. (D)

Continued….

Quantitative	Qualitative
Level of occurrence.	Depth of understanding
Asks how many or how much?	Asks why?
Studies events	Studies motivation
Objective	Subjective
Discovery and proof	Enables discovery
More definitive	Exploratory in nature
Describes	Interprets

	Qualitative model	*Quantitative model*
Description	Concerned with meaning rather than measurement	Measurement expressible in quantity or numbers
Advantage	Describes important concepts that are not easy to quantify	Simplifies situations so that tasks can be managed realistically and system performance can be predicted precisely
Drawback	Lacks objective numeric analysis to estimate a system precisely	Mathematical models that are difficult to understand and interpret Mathematical results and structures of optimal equation might be difficult to apply in actual use
Applications	Accident/incident investigation Analysis of equipment breakdown record Simulation study on human behaviour under various conditions	Prediction for probability of error occurring at various events in the form of relative or absolute data

6. Flat vs. Hierarchical Models



Flat data model:

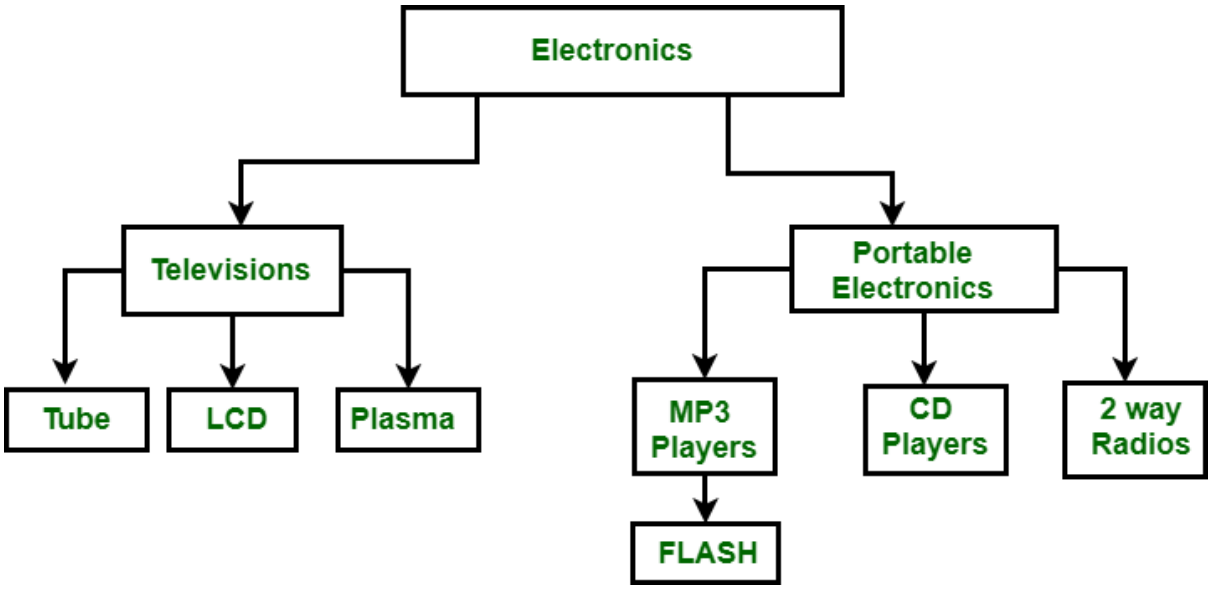
Ex:1-A simple flat data model for a retail store can include a single file with the following fields: Order ID. Customer name.

Ex:2-For instance, consider a database for an online bookstore. It has two columns titled "Book Title" and "Author". Each row is used to record different books and their respective authors. In the flat model, no two entries are the same.

Hierarchical Data Model:

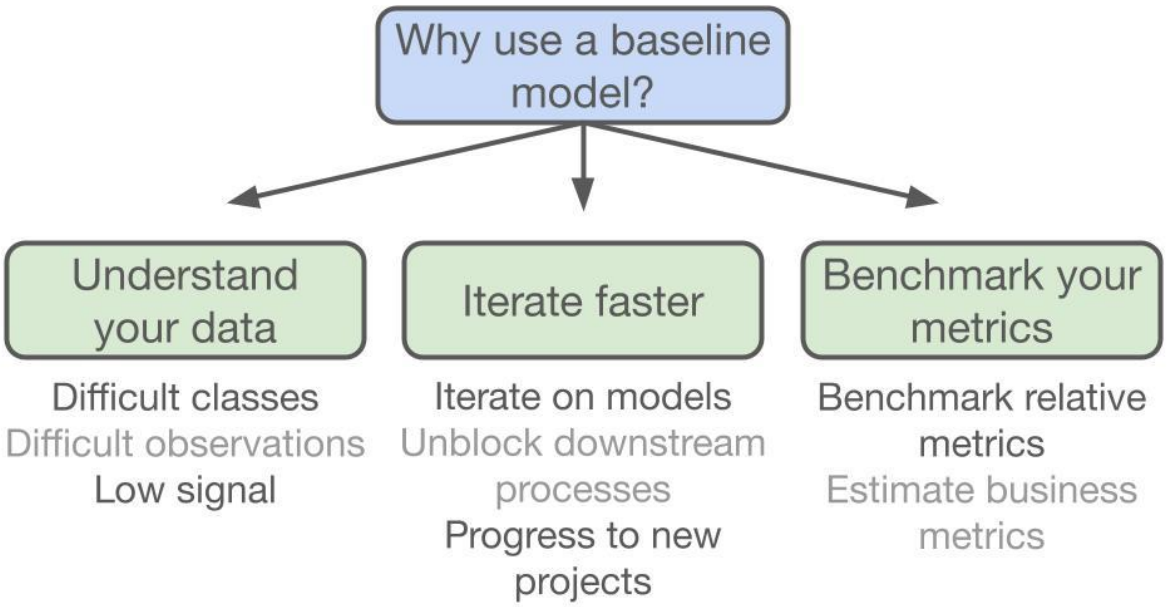
Hierarchical data model is the oldest type of the data model. It was developed by IBM in 1968. It organizes data in the tree-like structure. Hierarchical model consists of the following:

- It contains nodes which are connected by branches.
- The topmost node is called the root node.
- If there are multiple nodes appear at the top level, then these can be called as root segments.
- Each node has exactly one parent.
- One parent may have many child.



4.8 Base line models in ML:

- A baseline model is a simple model used to predict the outcome of data. It serves as a starting point for analysis, allowing us to assess the performance of more complex models and the impact of additional features.



- There are two common tasks for data science models: clasification and value prediction.

1. Baseline model for classification:

- In classification tasks, we are given a small set of possible labels for any given item, like (spam or not spam), (man or woman), or (bicycle, car, or truck).

(i) Zero Rule (ZeroR):

- **Description:** Predicts the majority class in the training data.

- The zero rule algorithm provides a baseline that reflects the minimum predictive power required for an algorithm to be considered useful.
- The zero rule, often referred to as the ZeroR classifier, is a simple baseline algorithm used in machine learning.

- **Usage:** Suitable for checking if a classification model performs better than random guessing.

  Ex: ZeroR:

  ZeroR for Classification: In classification tasks, ZeroR predicts the majority class. It ignores all input features and simply predicts the most frequent class in the training dataset. For instance, if a dataset has 70% of instances belonging to class A and 30% to class B, ZeroR will always predict class A.

➢ Advantages:

1. **Baseline Performance**: It provides a baseline performance metric. Any model that performs worse than ZeroR is considered ineffective.
2. **Simplicity**: It's simple to implement and provides a quick check to ensure that the data preprocessing and model training processes are working correctly.
3. Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods.

➢ Algorithm

Construct a frequency table for the target and select its most frequent value.

➢ Example:

"Play Golf = Yes" is the ZeroR model for the following dataset with an accuracy of 0.64.

Predictors / Target

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Play Golf	
Yes	No
9	5

- **Model Evaluation**

The following confusion matrix shows that ZeroR only predicts the majority class correctly. As mentioned before, ZeroR is only useful for determining a baseline performance for other classification methods.

(ii) **Random Classifier:**

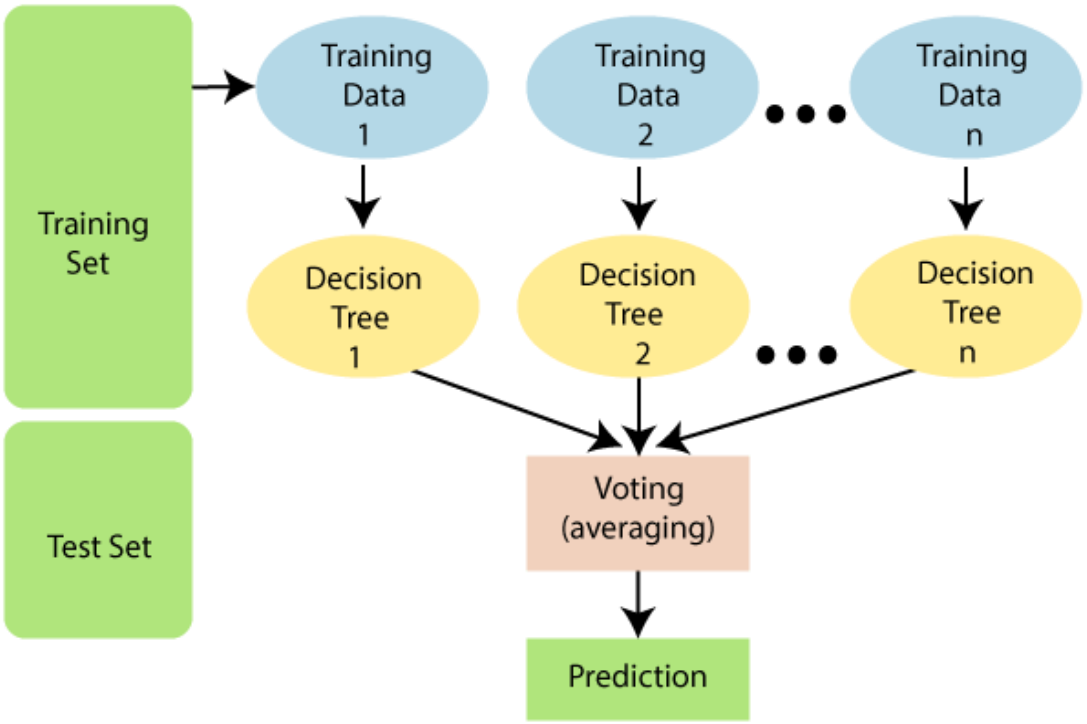
- **Description:** Assigns class labels randomly according to the class distribution in the training data.
- **Usage:** Helps to check if a model's performance is above random chance.
- **Random Forest** is a popular and powerful machine learning algorithm that is used for both classification and regression tasks.
- we can use Python and basic libraries to create a random classifier.

- **Classifier:**

A classifier is a type of machine learning model that assigns labels to data. For example, it can look at a picture and decide if it's a dog or a cat.

➢ Random Classifier:

A random classifier is a very simple model that makes random guesses. It doesn't learn from the data but just assigns labels at random.



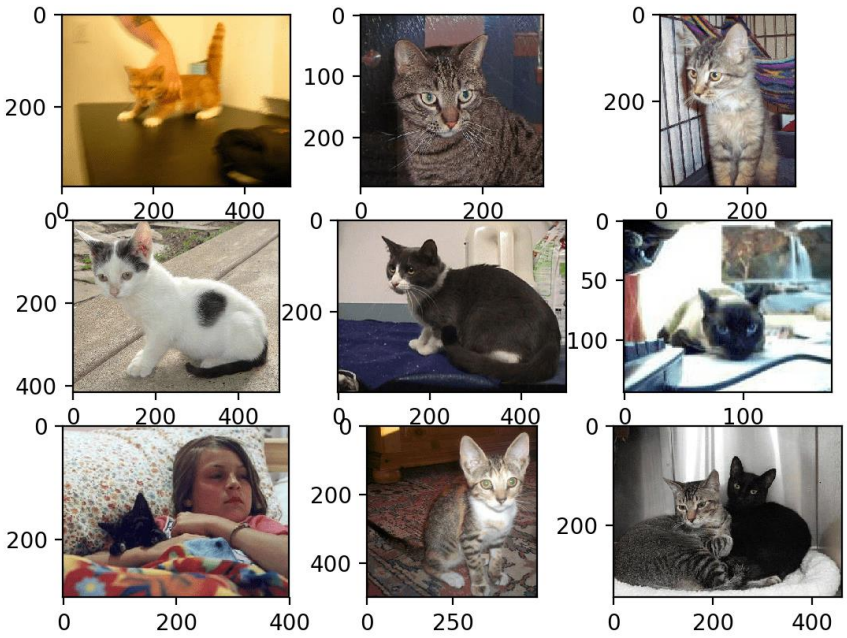
Steps to Create a Random Classifier

1. **Define the Problem**: Suppose we have images of animals and we want to classify them as either "Cat" or "Dog".
2. **Prepare the Data**: Normally, you would have some data to train your model. Here, we will just simulate some data.
3. **Make Random Predictions**: For each data point, randomly assign a label "Cat" or "Dog".
4. **Evaluate the Model**: Check how often the random classifier is correct.

   Ex: Dogs vs. Cats Prediction Problem

Step 1: Prepare the Data

Let's assume we have a simple dataset:



2. Baseline model for prediction:

(i) Mean Predictor (MeanR):

Description: Predicts the mean of the target values from the training data.

Usage: Simple to implement and provides a benchmark for regression models.
Implementation: DummyRegressor with strategy='mean' in scikit-learn.

Let's assume we have a simple dataset:

Input (X)	Output (Y)
1	3
2	4
3	2
4	5
5	6

The mean predictor will predict the mean of the output values (Y) for any given input.

$$\text{Mean}(Y) = \frac{\sum Y}{n}$$

Where:

- ∑Y is the sum of all output values.
- n is the number of data points.

$$\text{Mean}(Y) = \frac{3+4+2+5+6}{5} = \frac{20}{5} = 4$$

$$\text{MSE} = \frac{1}{n}\sum(Y_i - \hat{Y})^2$$

Where:

- Y_i is the actual output value.
- $\hat{Y}$ is the predicted value (mean in this case).

Let's calculate the MSE for our dataset:

Actual (Y)	Predicted ($\hat{Y}$)	Error ($Y - \hat{Y}$)	Squared Error $(Y - \hat{Y})^2$
3	4	-1	1
4	4	0	0
2	4	-2	4
5	4	1	1
6	4	2	4

Sum of Squared Errors = $1 + 0 + 4 + 1 + 4 = 10$

MSE = $\frac{10}{5} = 2$

(ii) Median Predictor:

Description: Predicts the median of the target values from the training data.
Usage: Useful especially in the presence of outliers as it is more robust than the mean.
Implementation: DummyRegressor with strategy='median' in scikit-learn.

Ex:

➢ **Median:**

First, let's explain what a median is:

- **Median** is the middle number in a sorted list of numbers.
- If the list has an odd number of items, the median is the number right in the middle.
- If the list has an even number of items, the median is the average of the two middle numbers.

(iii) Time Series Baselines

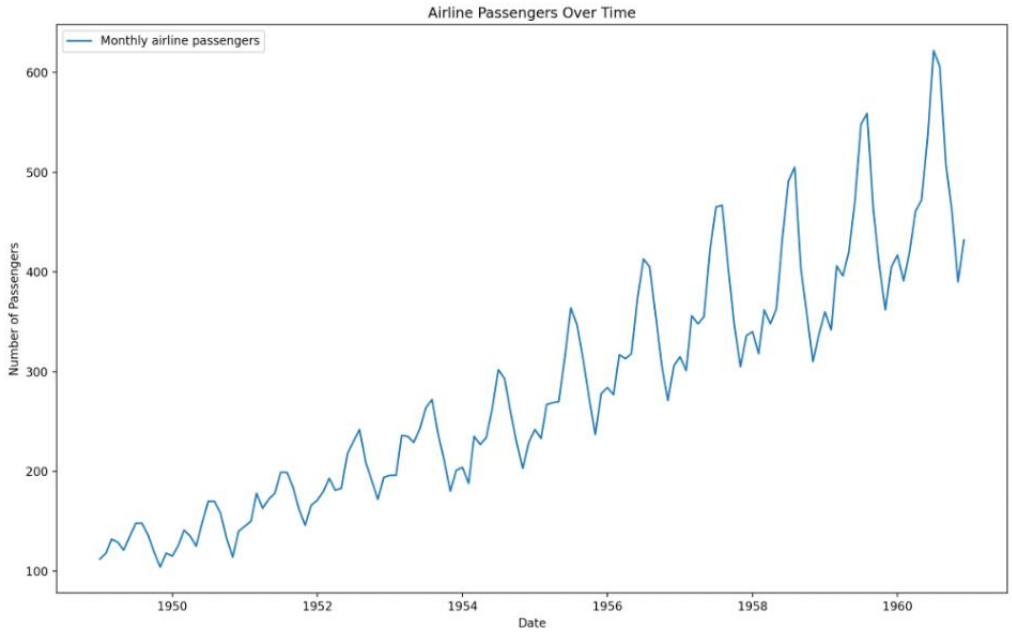
Time Series Data: Time Series Data is basically a specific data points that are recorded over a regular interval of time. This data is used to anayzae trends and patterns and make predictions. Examples of such data include weather measurements during weather forecasting, economic indicators, stock prices, and many more.

Objective: Develop a forecasting model for airline passenger numbers using time series data and linear regression.
Data: Historical airline passenger data, collected monthly.

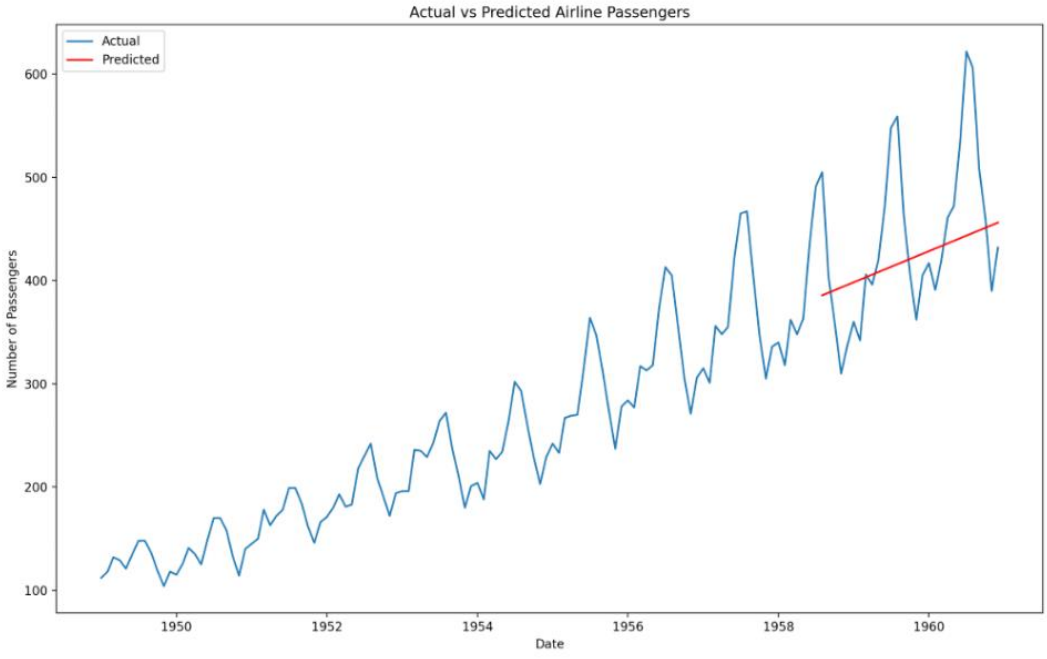
Step 1: Import Libraries and Load Data
Step 2: Visualize the Time Series Data
Step 3: Data Preparation for Linear Regression
Step 4: Fitting the Linear Regression Model
Step 5: Make Predictions
Step 6: Model Evaluation
Step 7: Visualize the Results

Finally plot the actual vs predicted values to visually understand the performance of the model.

https://www.geeksforgeeks.org/step-by-step-guide-to-modeling-time-series-data-using-linear-regression/



Airline Passengers



Actual vs Predicted Passengers Over time

The actual passenger numbers are consistently higher than the predicted passenger numbers. This suggests that the forecasting model used to create the predicted passenger numbers was not very accurate. There could be a number of reasons for this, such as the model not taking into account all of the relevant factors that affect airline passenger numbers. Hence, there is not use of forecasting.

4.9 Evaluating Models:

In data science, evaluation models are crucial for assessing the performance of predictive models. They help determine how well a model will perform on unseen data and guide improvements.

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses.

Here are some key points:

1. **Purpose**: Evaluation models measure the accuracy, efficiency, and robustness of predictive models. They ensure that the models provide reliable and valid results.
2. **Metrics**: Common evaluation metrics include:

- **Accuracy**: Proportion of correct predictions (used for classification).
- **Precision and Recall**: Precision measures the accuracy of positive predictions, while recall measures how well the model captures all positive instances.
- **F1 Score**: Harmonic mean of precision and recall, useful for imbalanced datasets.

F1 score

The F1 score is the harmonic mean of precision and recall. It is seen that during the precision-recall trade-off if we increase the precision, recall decreases and vice versa. The goal of the F1 score is to combine precision and recall.

F1 score = (2×Precision×Recall)/(Precision+Recall)

- **Mean Absolute Error (MAE)**: Average of absolute errors between predicted and actual values (used for regression).

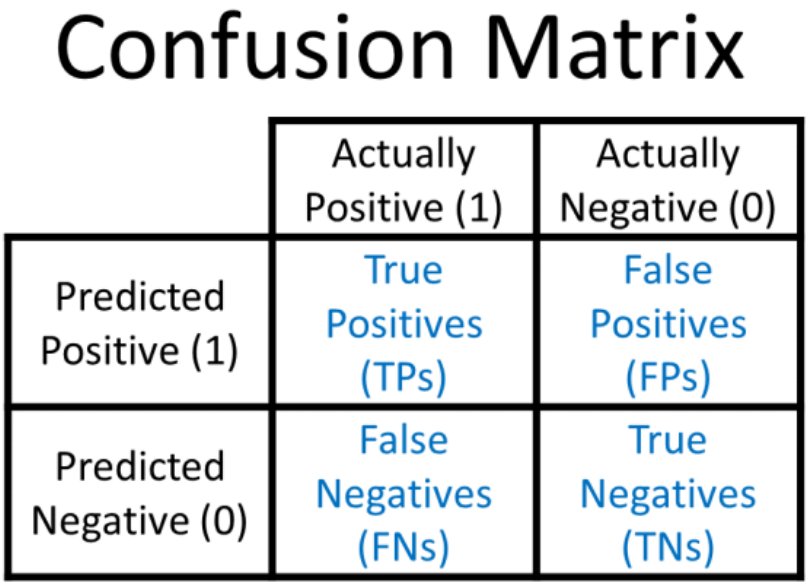
- **Root Mean Squared Error (RMSE)**: Square root of the average squared differences between predicted and actual values, giving more weight to larger errors.

The most commonly used metric is Mean Square error or MSE. It is a function used to calculate the loss. We find the difference between the predicted values and the truth variable, square the result and then find the average over the whole dataset. MSE is always positive as we square the values. The small the MSE better is the performance of our model. The formula of MSE is given:

$$MSE = \sum(y_{pred} - y_{actual})^2 / N$$

❖ Confusion Matrix

A confusion matrix is an N x N matrix where N is the number of target classes. It represents the number of actual outputs and the predicted outputs. Some terminologies in the matrix are as follows:



- True Positives: It is also known as TP. It is the output in which the actual and the predicted values are YES.

- True Negatives: It is also known as TN. It is the output in which the actual and the predicted values are NO.
- False Positives: It is also known as FP. It is the output in which the actual value is NO but the predicted value is YES.
- False Negatives: It is also known as FN. It is the output in which the actual value is YES but the predicted value is NO.

➢ **Definition/Basic concept of TRAING DATA & TESTING DATA**

- **Training Data**

  Imagine you have a robot dog named Robo, and you want to teach Robo how to fetch a ball. To do this, you need to practice with Robo many times. The more you practice, the better Robo gets at fetching the ball.

  In data science, this practice is called "training," and the examples you use to teach Robo are called training data. Training data is like a big book of examples that helps the robot (or a computer program) learn how to do a specific task.

- **Testing Data**

  Now, after Robo has practiced a lot, you want to see how well he has learned to fetch the ball. So, you take Robo to a new park and ask him to

fetch a ball again. This time, you don’t help Robo; you just watch to see how well he does on his own.

In data science, this part is called "testing," and the new examples you use to check how well Robo has learned are called testing data. Testing data is like a quiz or a test that helps you see if the robot (or a computer program) can do the task correctly without any help.

Summary

- Training Data: Examples used to teach a computer program how to do a task. It's like practicing with Robo to fetch the ball.
- Testing Data: New examples used to see how well the computer program has learned. It's like taking Robo to a new park to test if he can fetch the ball by himself.

  This way, by using both training and testing data, we can make sure that our computer programs (or Robo) learn well and can perform tasks correctly in different situations.

Module V

Social Network Graphs

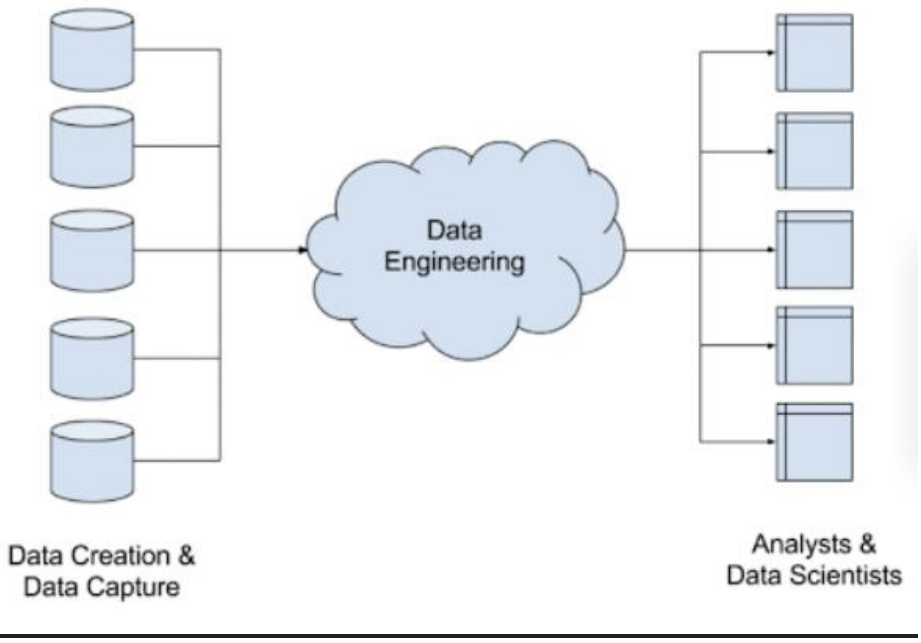
Module V

Social Network Graphs

Data Engineering—Map reduce—Word Frequency Problem—Map Reduce Solution with Example
Social Network Graphs: Social networks as graphs—Clustering of graphs—Partitioning of graphs

5.1 Data engineering:

Data engineering is the process of designing and building systems that let people collect and analyze raw data from multiple sources and formats. These systems empower people to find practical applications of the data, which businesses can use to thrive.



Data engineering is a skill that is in increasing demand. Data engineers are the people who design the system that unifies data and can help you navigate it. Data engineers perform many different tasks including:

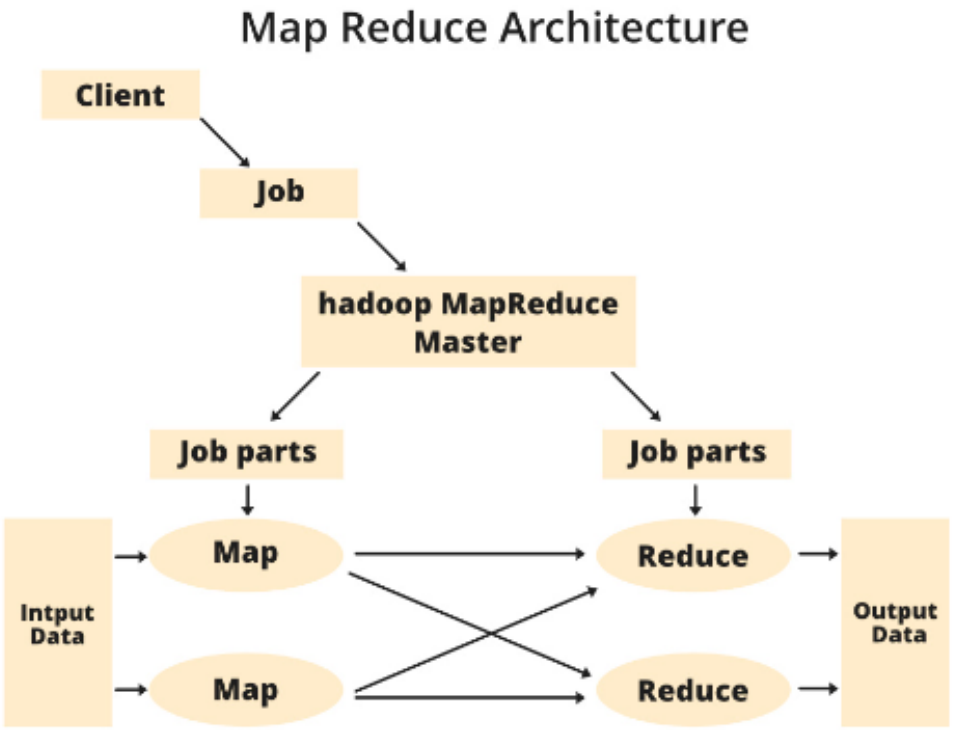
- Acquisition: Finding all the different data sets around the business
- Cleansing: Finding and cleaning any errors in the data
- Conversion: Giving all the data a common format
- Disambiguation: Interpreting data that could be interpreted in multiple ways
- Deduplication: Removing duplicate copies of data

5.2 Map Reduce

MapReduce and HDFS are the two major components of Hadoop which makes it so powerful and efficient to use. MapReduce is a programming model used for efficient

processing in parallel over large data-sets in a distributed manner. The data is first split and then combined to produce the final result. The libraries for MapReduce is written in so many programming languages with various different-different optimizations. The purpose of MapReduce in Hadoop is to Map each of the jobs and then it will reduce it to equivalent tasks for providing less overhead over the cluster network and to reduce the processing power. The MapReduce task is mainly divided into two phases Map Phase and Reduce Phase.

MapReduce Architecture:



Components of MapReduce Architecture:

1. **Client:** The MapReduce client is the one who brings the Job to the MapReduce for processing. There can be multiple clients available that continuously send jobs for processing to the Hadoop MapReduce Manager.
2. **Job:** The MapReduce Job is the actual work that the client wanted to do which is comprised of so many smaller tasks that the client wants to process or execute.

3. **Hadoop MapReduce Master:** It divides the particular job into subsequent job-parts.
4. **Job-Parts:** The task or sub-jobs that are obtained after dividing the main job. The result of all the job-parts combined to produce the final output.
5. **Input Data:** The data set that is fed to the MapReduce for processing.
6. **Output Data:** The final result is obtained after the processing.

In **MapReduce**, we have a client. The client will submit the job of a particular size to the Hadoop MapReduce Master. Now, the MapReduce master will divide this job into further equivalent job-parts. These job-parts are then made available for the Map and Reduce Task. This Map and Reduce task will contain the program as per the requirement of the use-case that the particular company is solving. The developer writes their logic to fulfill the requirement that the industry requires. The input data which we are using is then fed to the Map Task and the Map will generate intermediate key-value pair as its output. The output of Map i.e. these key-value pairs are then fed to the Reducer and the final output is stored on the HDFS. There can be n number of Map and Reduce tasks made available for processing the data as per the requirement. The algorithm for Map and Reduce is made with a very optimized way such that the time complexity or space complexity is minimum.

Let's discuss the MapReduce phases to get a better understanding of its architecture:
The MapReduce task is mainly divided into **2 phases** i.e. Map phase and Reduce phase.

1. **Map:** As the name suggests its main use is to map the input data in key-value pairs. The input to the map may be a key-value pair where the key can be the id of some kind of address and value is the actual value that it keeps. The *Map()* function will be executed in its memory repository on each of these input key-value pairs and generates the intermediate key-value pair which works as input for the Reducer or *Reduce()* function.

2. **Reduce:** The intermediate key-value pairs that work as input for Reducer are shuffled and sort and send to the *Reduce()* function. Reducer aggregate or group

the data based on its key-value pair as per the reducer algorithm written by the developer.

How Job tracker and the task tracker deal with MapReduce:

1. **Job Tracker:** The work of Job tracker is to manage all the resources and all the jobs across the cluster and also to schedule each map on the Task Tracker running on the same data node since there can be hundreds of data nodes available in the cluster.

2. **Task Tracker:** The Task Tracker can be considered as the actual slaves that are working on the instruction given by the Job Tracker. This Task Tracker is deployed on each of the nodes available in the cluster that executes the Map and Reduce task as instructed by Job Tracker.

There is also one important component of MapReduce Architecture known as **Job History Server**. The Job History Server is a daemon process that saves and stores historical information about the task or application, like the logs which are generated during or after the job execution are stored on Job History Server.

Simple example of MapReduce using a word count problem.

Example: Word Count using MapReduce

Problem:

Given a collection of documents, count the frequency of each word.

Map Function:

The **Map** function takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key, value) pairs.

For our word count example:

- **Input**: A list of documents.
- **Output**: Key-value pairs where the key is a word and the value is 1 (indicating the word has been encountered once).

Python code:

```
def mapper(document):
    # document is a string (a document)
    words = document.split()
    word_count = {}
    for word in words:
        word_count[word] = 1
    return word_count.items()
```

➢ *Reduce Function:*

The **Reduce** function takes the output from the Map function and combines those data tuples into a smaller set of tuples.

5.3 For our word count example: Word Frequency problem & solution:

- **Input**: Key-value pairs where the key is a word and the value is a list of counts.
- **Output**: Key-value pairs where the key is a word and the value is the total count of that word across all documents.

```
python
Copy code
from collections import defaultdict

def reducer(word_counts):
    word_count = defaultdict(int)
    for word, counts in word_counts:
        word_count[word] += sum(counts)
    return word_count.items()
```

Example Data:

Let's assume we have two documents:

1. "Hello world"
2. "Hello again world"

Applying MapReduce:

1. **Map Phase**:
   - Apply the mapper function to each document.

   For "Hello world":

   css
   Copy code
   [("Hello", 1), ("world", 1)]

   For "Hello again world":

   css
   Copy code
   [("Hello", 1), ("again", 1), ("world", 1)]

2. **Shuffle Phase**:
   - Group the intermediate key-value pairs from the Map phase by key. This is typically done automatically in distributed systems.
3. **Reduce Phase**:
   - Apply the reducer function to each group of key-value pairs.

   After reducing:

   css
   Copy code
   [("Hello", 2), ("world", 2), ("again", 1)]

Result:

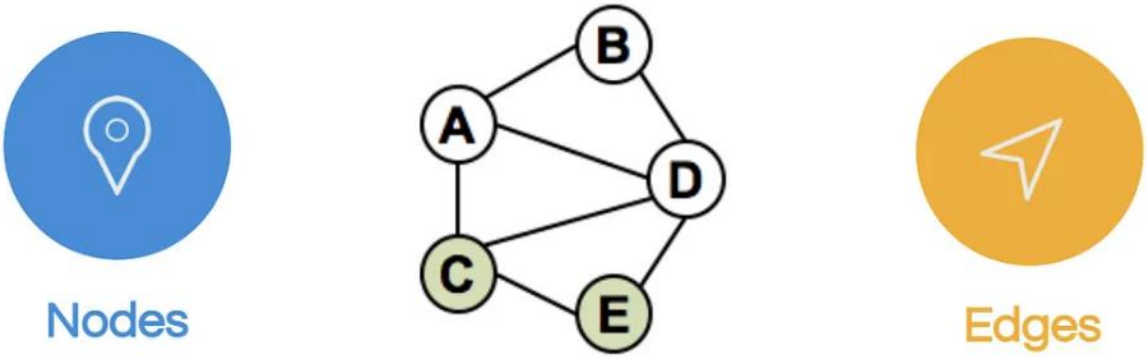
- The word "Hello" appears 2 times.
- The word "world" appears 2 times.
- The word "again" appears 1 time.

This example simplifies the MapReduce process to illustrate its basic principles. In real-world scenarios, MapReduce is used for processing large datasets in parallel across distributed systems, such as Apache Hadoop or Spark, to handle massive amounts of data efficiently.

5.4 Social networks as graphs:

Network Theory:

Network's basic components: nodes and edges.



Example Network

- **Nodes (or Vertices)**:

- Nodes (A, B, C, D, E in the example) are usually representing entities in the network, and can hold self-properties (such as weight, size, position and any other attribute) and network-based properties (such as *Degree*- number of neighbours or *Cluster*- a connected component the node belongs to etc.).

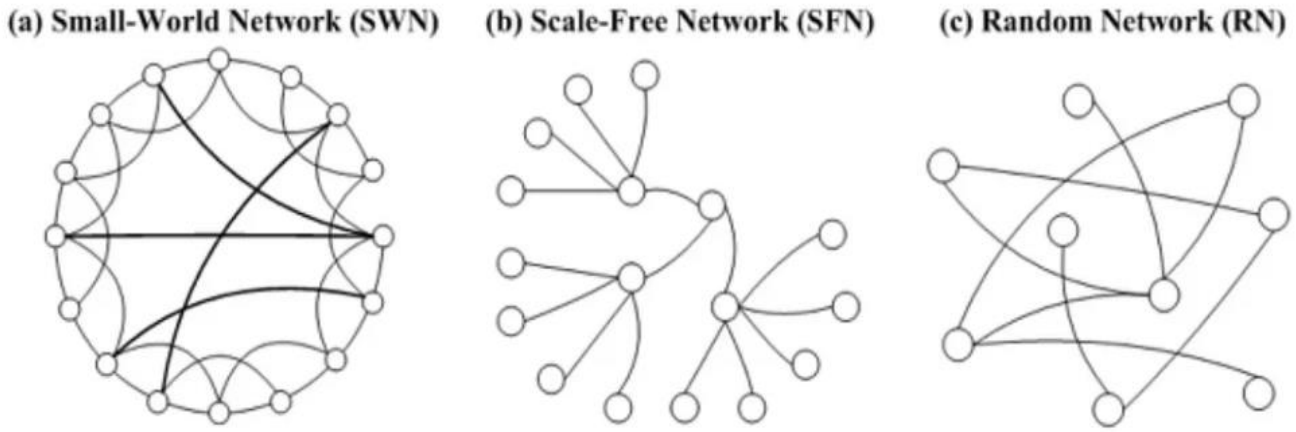
- These represent entities in the network, like people in a social network.

❖ **Edges (or Links)**:

- Edges represent the connections between the nodes, and might hold properties as well (such as weight representing the strength of the connection, direction in case of asymmetric relation or time if applicable).
- These represent connections or relationships between nodes, like friendships or interactions.

Graph: Graphs can be **directed** (where relationships have a direction, e.g., Twitter follows) or **undirected** (where relationships are mutual, e.g., Facebook friendships).

❖ Understanding Real-world networks



❖ **Building a Network**

- Networks can be constructed from various datasets, as long as we're able to describe the relations between nodes.
- We'll read the data from excel file to a pandas dataframe to get a tabular representation
- Then, we will build a directed graph using networkx from the edgelist we have as a pandas dataframe. Finally, we'll try the generic method to visualize.

❖ **Key Concepts in Graphs:**

1. Degree: The number of edges connected to a node. In directed graphs, we differentiate between in-degree (incoming edges) and out-degree (outgoing edges).
2. Path: A sequence of edges connecting two nodes.
3. Components: Subsets of a graph where there is a path between any two nodes within the subset.
4. Centrality: Measures the importance of a node within a network. Common centrality measures include:
5. Degree Centrality: Number of direct connections a node has.
6. Betweenness Centrality: Frequency at which a node appears on the shortest paths between other nodes.
7. Closeness Centrality: Measure of how close a node is to all other nodes in the network.
8. Eigenvector Centrality: Influence of a node in the network, considering the influence of its neighbours.

5.5 Social Network as Graph:

- Social networks can be modelled as graphs, where nodes represent entities and edges represent relationships. If relationships have degrees, these are labelled on the edges. Social graphs are often undirected, like Facebook friends, but can also be directed, like Twitter followers.

For example, In a small social network graph with nodes A through G:

- Edges: Represent friendships, e.g., B is friends with A, C, and D.

Locality Check: The graph has 9 edges out of a possible 21 pairs.

❖ Applications of Social Network Analysis:

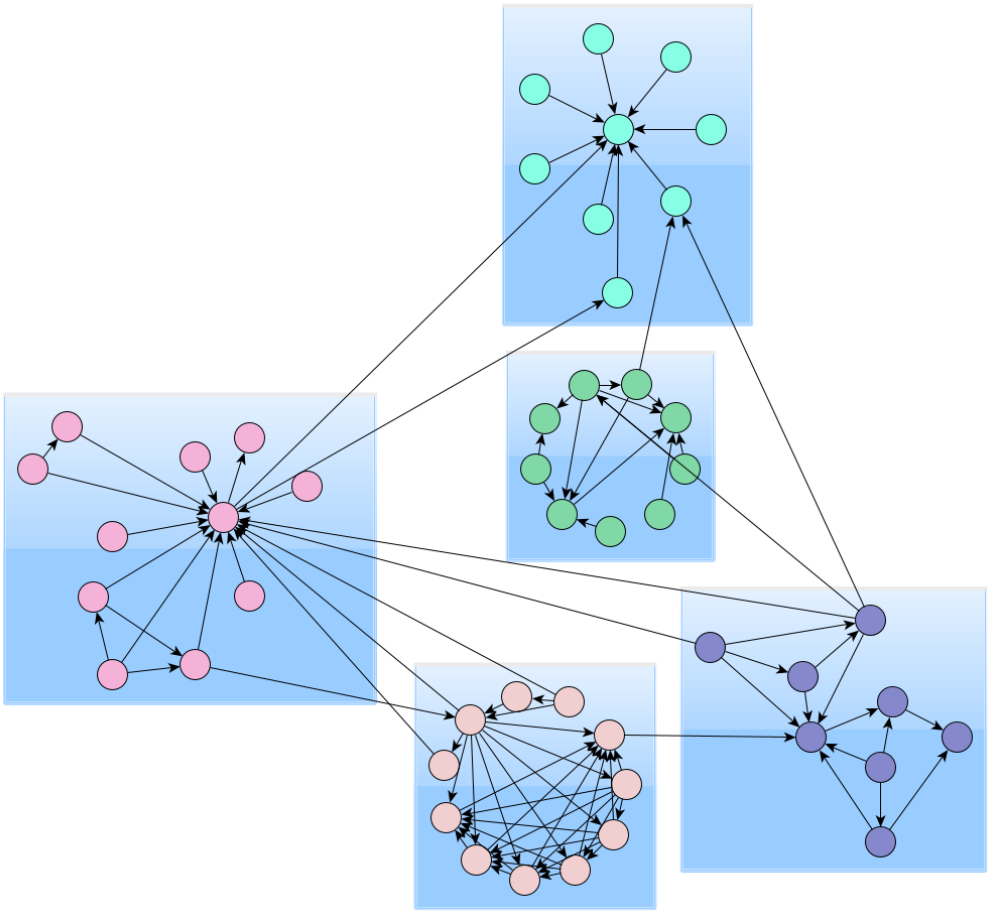
- Community Detection: Identifying groups of nodes that are more densely connected internally than with the rest of the network.
- Influence Analysis: Finding key influencers within a network, useful in marketing and information dissemination.
- Network Robustness: Analyzing the stability of a network under node or edge failures, important for infrastructure and communication networks.
- Epidemiology: Studying how diseases spread through populations, aiding in public health strategies.

❖ Tools for Social Network Analysis

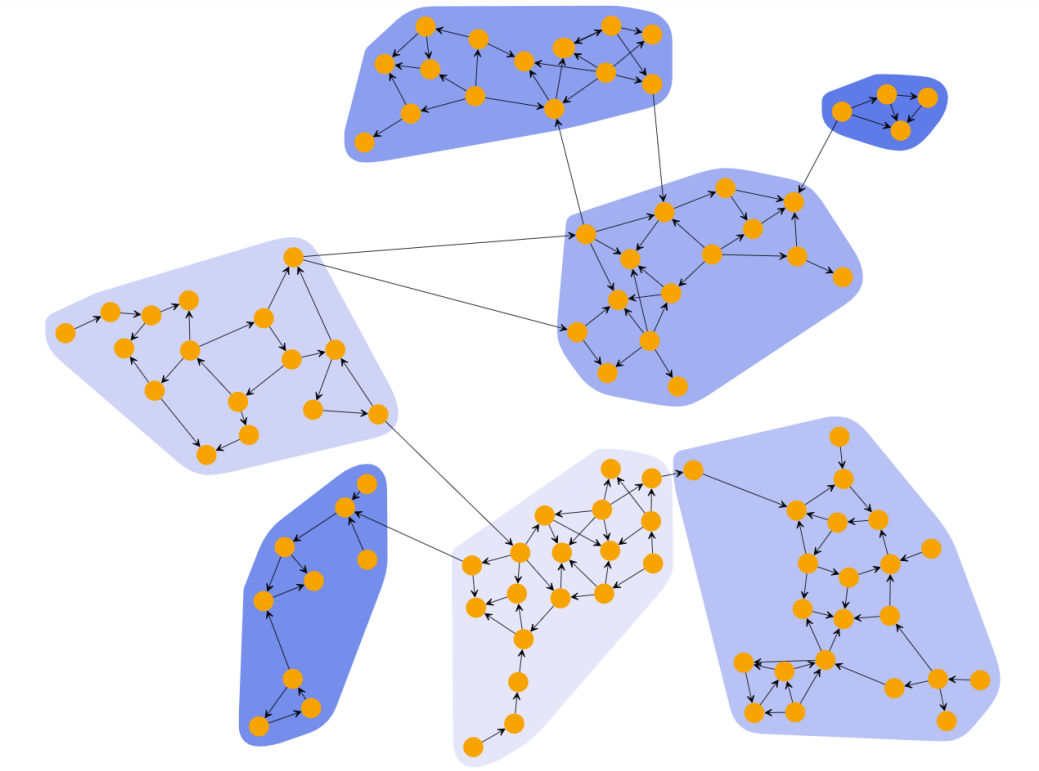
1. Gephi: A popular open-source software for visualizing and analyzing large networks.
2. NetworkX: A Python library for the creation, manipulation, and study of complex networks.
3. Pajek: Software for analysis and visualization of large networks.

5.6 Clustering of graphs:

- Clustering is used to identify groups of similar objects in datasets with two or more variable quantities.
- Nodes that based on this similarity value are considered to be similar are grouped into the so-called clusters. In other words, each cluster contains elements that share common properties and characteristics.



- **Edge Betweenness Clustering**



6 A real-life example of clustering in a social network graph can be observed on platforms like Facebook, Twitter, or LinkedIn.

Example: Facebook Friendship Network

- *Scenario:*

Consider a simplified version of Facebook where users are nodes, and friendships are edges connecting these nodes.

➢ *Clusters:*

1. **Family Cluster**:
    - Members of a family are more likely to be friends with each other.
    - Parents, siblings, cousins, and close relatives often form a tight-knit cluster.
2. **College Friends Cluster**:
    - People who attended the same college or university tend to form a community.
    - Alumni groups or classmates often stay connected and frequently interact within this group.
3. **Work Colleagues Cluster**:
    - Employees of the same company or colleagues in the same industry form another cluster.
    - Professional connections and work-related friendships are more common within this group.
4. **Hobby Group Cluster**:
    - Individuals who share the same hobbies or interests, such as a book club, sports team, or music band, form clusters.
    - These clusters are based on shared activities and common interests.

➢ *Visualization:*

Imagine a graph where:

- Nodes represent individuals.
- Edges represent friendships.
- Clusters are visually represented by dense regions within the graph where nodes are closely interconnected.

❖ **Practical Steps for Clustering**

1. Data Preparation

   - Collect and preprocess data to form the social network graph.
   - Represent the graph using an adjacency matrix or an edge list.

2. Algorithm Selection

   - Choose an appropriate clustering algorithm based on the network size and desired outcomes.

3. Implementation

   - Use graph analysis libraries such as NetworkX (Python), Gephi, or igraph to implement the clustering algorithm.

4. Evaluation

   - Evaluate the quality of the clusters using metrics like modularity, conductance, or silhouette score.

5. Visualization

Visualize the clusters to interpret the results and gain insights into the network structure.

5.7 Partitioning of graphs:

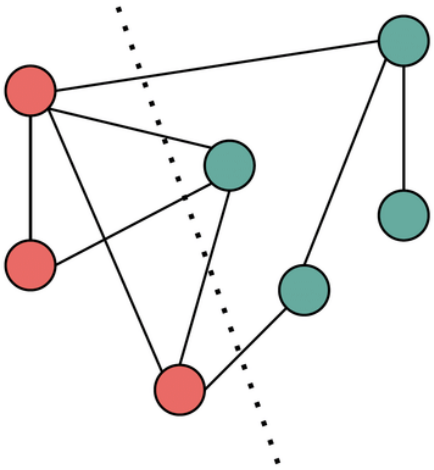
- Graph partitioning can be done by recursively bisecting a graph or directly partitioning it into k sets.
- There are two ways to partition a graph, by taking out edges, and by taking out vertices.
- Graph partitioning algorithms use either edge or vertex separators in their execution, depending on the particular algorithm.

- Using data partitioning techniques, a **huge dataset can be divided into smaller**, **simpler sections**.

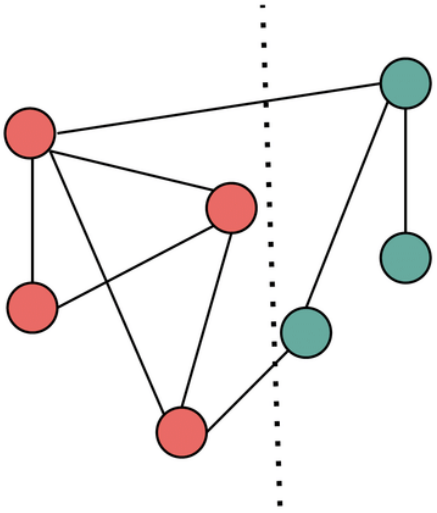
- Partitioning of a graph in data science involves dividing the graph into smaller, more manageable subgraphs while preserving the structure and properties of the original graph as much as possible. This is a crucial step in many applications such as parallel computing, clustering, and network analysis.

7 Applications of Graph Partitioning

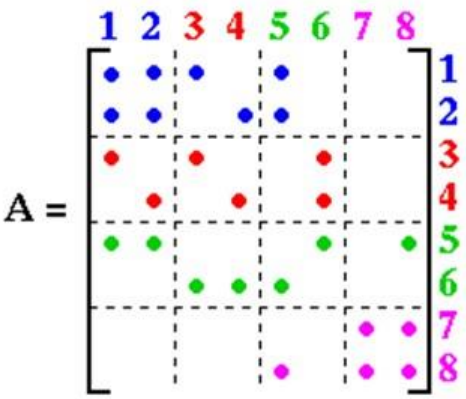
- Social Network Analysis: Identifying communities or clusters within social networks.
- Scientific Computing: Distributing large-scale computations across multiple processors.
- Parallel Computing: Load balancing and minimizing inter-processor communication.
- Bioinformatics: Analyzing protein-protein interaction networks or gene regulatory networks.
- Data Mining: Clustering data points represented as nodes in a graph.

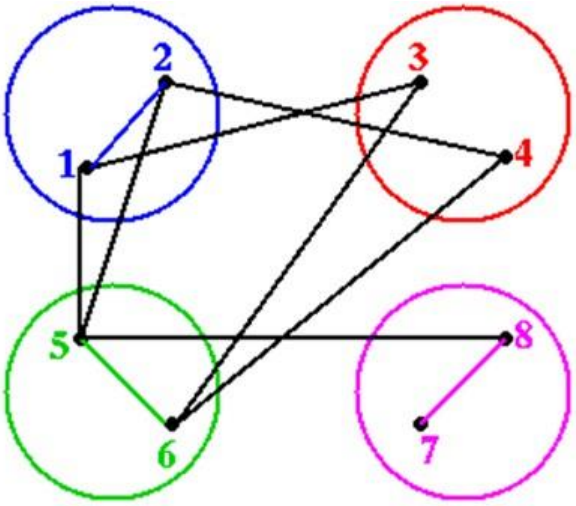


Bad partition



Good partition





- Representing this problem as a graph and using graph partitioning helps accomplish this task.