UNIT – 2 : INTRODUCTION TO DATA ANALYSIS

Introduction and Importance of Data Science:

What is Data Science?

Data Science is the area of study which involves extracting insights from vast amounts of data using various scientific methods, algorithms, and processes. It helps you to discover hidden patterns from the raw data. The term Data Science has emerged because of the evolution of mathematical statistics, data analysis, and big data.

Data Science is an interdisciplinary field that allows you to extract knowledge from structured or unstructured data. Data science enables you to translate a business problem into a research project and then translate it back into a practical solution.

Why Data Science?

Here are significant advantages of using Data Analytics Technology:

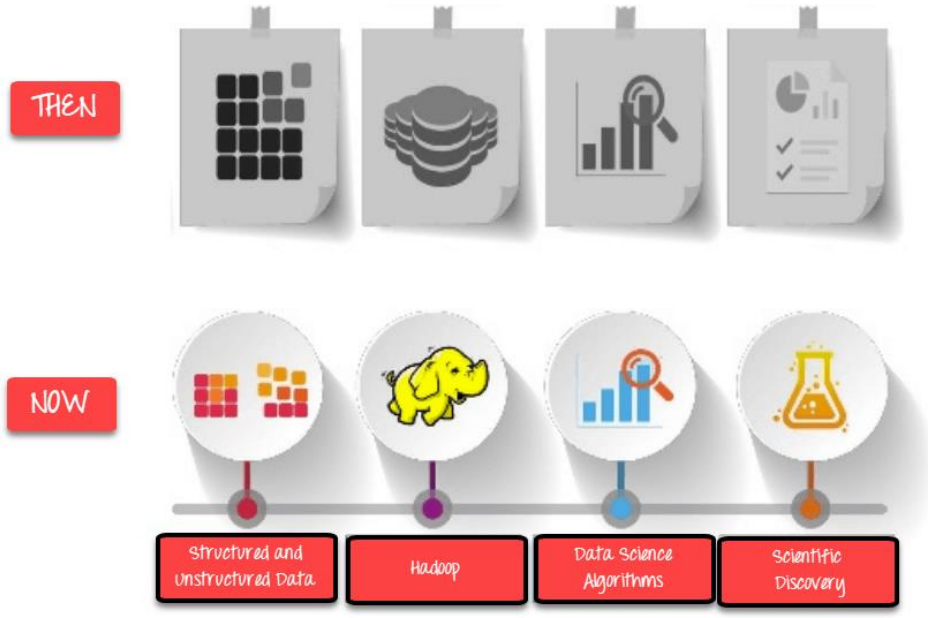
- Data is the oil for today's world. With the right tools, technologies, algorithms, we can use data and convert it into a distinct business advantage
- Data Science can help you to detect fraud using advanced machine learning algorithms
- It helps you to prevent any significant monetary losses
- Allows to build intelligence ability in machines
- You can perform sentiment analysis to gauge customer brand loyalty
- It enables you to take better and faster decisions
- It helps you to recommend the right product to the right customer to enhance your business

History Of Data Science:

The history of data science can be traced back to the early developments in statistics, computer science, and information theory. Here's a brief overview of key milestones in the history of data science:

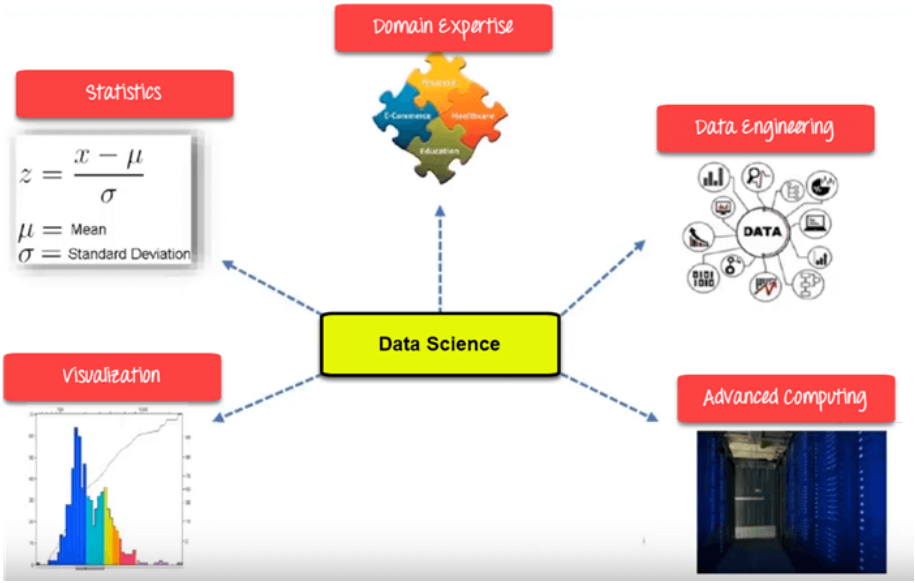
1. **Statistics Roots (18th Century):** The roots of data science can be traced back to the development of statistical methods. Statisticians like Sir Francis Galton and Karl Pearson made significant contributions to the field of statistics in the 19th century.

2. **Early Computers and Information Theory (20th Century):** The advent of computers in the mid-20th century, along with the development of information theory by Claude Shannon, provided the foundation for managing and processing large volumes of data.
3. **Spread of Databases (1960s):** With the development of database management systems (DBMS), organizations started to store and manage large amounts of structured data. This era marked the beginning of systematic data storage.
4. **Emergence of Data Warehousing (1980s):** The concept of data warehousing emerged, allowing organizations to integrate and store data from various sources in a central repository for analysis. This period laid the groundwork for centralized data storage and retrieval.
5. **Rise of Business Intelligence (1990s):** Business Intelligence (BI) tools gained popularity, enabling organizations to analyze historical data and generate reports for decision-making. Data warehouses and BI played a crucial role in shaping the early landscape of data-driven decision-making.
6. **The Era of Big Data (2000s):** The 21st century witnessed an explosion in the volume, velocity, and variety of data. This era, often referred to as the era of "Big Data," brought challenges and opportunities for handling massive datasets that traditional databases couldn't manage efficiently.
7. **Open-Source Tools and Hadoop (mid-2000s):** The development of open-source tools and frameworks, particularly Hadoop, played a pivotal role in the processing and analysis of large-scale data. Hadoop, with its distributed storage and processing capabilities, became a cornerstone in the world of big data.
8. **Introduction of the Term "Data Science" (2008):** The term "data science" gained prominence in 2008 when statisticians Jeff Wu and William Cleveland referred to it as a "fourth paradigm" of science. They highlighted the importance of extracting knowledge and insights from data.
9. **Proliferation of Machine Learning (2010s):** The 2010s saw a rapid growth in machine learning and predictive analytics. The availability of vast amounts of data and powerful computing resources contributed to the development and deployment of sophisticated machine learning models.
10. **Integration of Data Science in Various Industries (Present):** Today, data science has become an integral part of various industries, including finance, healthcare, marketing, and technology. Organizations leverage data science techniques to gain insights, make informed decisions, and drive innovation.



Evolution of Data Sciences

Data Science Components:



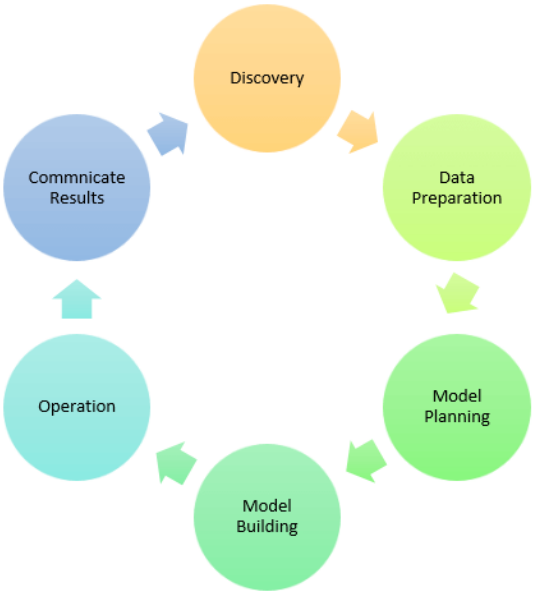
Statistics: Statistics is the most critical unit of Data Science basics, and it is the method or science of collecting and analyzing numerical data in large quantities to get useful insights.

Visualization: Visualization technique helps you access huge amounts of data in easy to understand and digestible visuals.

Machine Learning: Machine Learning explores the building and study of algorithms that learn to make predictions about unforeseen/future data.

Deep Learning: Deep Learning method is new machine learning research where the algorithm selects the analysis model to follow.

Data Science Process:



1. Discovery: Discovery step involves acquiring data from all the identified internal & external sources, which helps you answer the business question.

The data can be:

- Logs from webservers
- Data gathered from social media
- Census datasets
- Data streamed from online sources using APIs

2. Preparation: Data can have many inconsistencies like missing values, blank columns, an incorrect data format, which needs to be cleaned. You need to process, explore, and condition data before modelling. The cleaner your data, the better are your predictions.

3. Model Planning: In this stage, you need to determine the method and technique to draw the relation between input variables. Planning for a model is performed by using different statistical formulas and visualization tools. SQL analysis services, R, and SAS/access are some of the tools used for this purpose.

4. Model Building: In this step, the actual model building process starts. Here, Data scientist distributes datasets for training and testing. Techniques like association, classification, and clustering are applied to the training data set. The model, once prepared, is tested against the "testing" dataset.

5. Operationalize: You deliver the final baselined model with reports, code, and technical documents in this stage. Model is deployed into a real-time production environment after thorough testing.

6. Communicate Results: In this stage, the key findings are communicated to all stakeholders. This helps you decide if the project results are a success or a failure based on the inputs from the model.

Data Science Jobs Roles:

Most prominent Data Scientist job titles are:

- Data Scientist
- Data Engineer
- Data Analyst
- Statistician
- Data Architect
- Data Admin
- Business Analyst
- Data/Analytics Manager

Let's learn what each role entails in detail:

Data Scientist:

Role: A Data Scientist is a professional who manages enormous amounts of data to come up with compelling business visions by using various tools, techniques, methodologies, algorithms, etc.

Languages: R, SAS, Python, SQL, Hive, Matlab, Pig, Spark

Data Engineer:

Role: The role of a data engineer is of working with large amounts of data. He develops, constructs, tests, and maintains architectures like large scale processing systems and databases.

Languages: SQL, Hive, R, SAS, Matlab, Python, Java, Ruby, C + +, and Perl

Data Analyst:

Role: A data analyst is responsible for mining vast amounts of data. They will look for relationships, patterns, trends in data. Later he or she will deliver compelling reporting and visualization for analyzing the data to take the most viable business decisions.

Languages: R, Python, HTML, JS, C, C+ + , SQL

Statistician:

Role: The statistician collects, analyses, and understands qualitative and quantitative data using statistical theories and methods.

Languages: SQL, R, Matlab, Tableau, Python, Perl, Spark, and Hive

Data Administrator:

Role: Data admin should ensure that the database is accessible to all relevant users. He also ensures that it is performing correctly and keeps it safe from hacking.

Languages: Ruby on Rails, SQL, Java, C#, and Python

Business Analyst:

Role: This professional needs to improve business processes. He/she is an intermediary between the business executive team and the IT department.

Languages: SQL, Tableau, Power BI and, Python

Tools for Data Science:

Data Analysis	Data Warehousing	Data Visualization	Machine Learning
R, Spark, Python and SAS	Hadoop, SQL, Hive	R, Tableau, Raw	Spark, Azure ML studio, Mahout

Applications of Data Science:

Some applications of Data Science are:

Internet Search: Google search uses Data science technology to search for a specific result within a fraction of a second

Recommendation Systems: To create a recommendation system. For example, "suggested friends" on Facebook or suggested videos" on YouTube, everything is done with the help of Data Science.

Image & Speech Recognition: Speech recognizes systems like Siri, Google Assistant, and Alexa run on the Data science technique. Moreover, Facebook recognizes your friend when you upload a photo with them, with the help of Data Science.

Gaming world: EA Sports, Sony, Nintendo are using Data science technology. This enhances your gaming experience. Games are now developed using Machine Learning techniques, and they can update themselves when you move to higher levels.

Online Price Comparison: PriceRunner, Junglee, Shopzilla work on the Data science mechanism. Here, data is fetched from the relevant websites using APIs.

Challenges of Data Science Technology:

- A high variety of information & data is required for accurate analysis
- Not adequate data science talent pool available
- Management does not provide financial support for a data science team
- Unavailability of/difficult access to data
- Business decision-makers do not effectively use data Science results
- Explaining data science to others is difficult
- Privacy issues
- Lack of significant domain expert
- If an organization is very small, it can't have a Data Science team

Advantages of Data Science:

1. **Improved decision-making:** Data science can help organizations make better decisions by providing insights and predictions based on data analysis.
2. **Cost-effective:** With the right tools and techniques, data science can help organizations reduce costs by identifying areas of inefficiency and optimizing processes.
3. **Innovation:** Data science can be used to identify new opportunities for innovation and to develop new products and services.
4. **Competitive advantage:** Organizations that use data science effectively can gain a competitive advantage by making better decisions, improving efficiency, and identifying new opportunities.
5. **Personalization:** Data science can help organizations personalize their products or services to better meet the needs of individual customers.

Disadvantages of Data Science:

1. **Data quality:** The accuracy and quality of the data used in data science can have a significant impact on the results obtained.
2. **Privacy concerns:** The collection and use of data can raise privacy concerns, particularly if the data is personal or sensitive.
3. **Complexity:** Data science can be a complex and technical field that requires specialized skills and expertise.
4. **Bias:** Data science algorithms can be biased if the data used to train them is biased, which can lead to inaccurate results.
5. **Interpretation:** Interpreting data science results can be challenging, particularly for non-technical stakeholders who may not understand the underlying assumptions and methods used.

Importance of Data Science in Various Industries:

Large databases of structured and unstructured data must be mined using data science techniques to find hidden patterns and derive useful insights. Data science is crucial because of the numerous applications it may be used for, ranging from simple activities, such as asking Siri or Alexa for recommendations, to more sophisticated ones, such as operating a self-driving automobile.

Let's check out the importance of data science in society -

1. **Healthcare:**
   - **Personalized Medicine:** Data science analyzes patient data to tailor treatments based on individual genetics and health history.
   - **Predictive Analytics:** Helps forecast disease outbreaks, patient admission rates, and resource needs.
   - **Clinical Research:** Accelerates drug discovery and clinical trial processes through data-driven insights.
2. **Finance:**
   - **Risk Management:** Data science models assess and predict financial risks, improving decision-making in investments and lending.
   - **Fraud Detection:** Identifies and prevents fraudulent activities by analyzing patterns and anomalies in financial transactions.
   - **Algorithmic Trading:** Uses data-driven algorithms to make faster and more accurate trading decisions.
3. **Retail:**
   - **Customer Analytics:** Analyzing customer behavior data helps in personalized marketing, inventory management, and improving customer experience.
   - **Supply Chain Optimization:** Predictive analytics improves demand forecasting, inventory management, and logistics, enhancing overall efficiency.
   - **Dynamic Pricing:** Uses real-time data to adjust pricing based on demand, competition, and other factors.
4. **Manufacturing:**
   - **Predictive Maintenance:** Data science models predict equipment failures, reducing downtime and maintenance costs.
   - **Quality Control:** Analyzes production data to identify and rectify quality issues, improving product quality.
   - **Supply Chain Visibility:** Enhances supply chain management by optimizing inventory levels, reducing delays, and improving overall efficiency.
5. **Telecommunications:**
   - **Network Optimization:** Analyzes data to optimize network performance, reduce downtime, and improve service quality.
   - **Customer Churn Prediction:** Identifies factors leading to customer churn and helps in implementing retention strategies.

- **Fraud Detection:** Monitors for unusual patterns to detect and prevent fraudulent activities.

6. Marketing and Advertising:

- **Customer Segmentation:** Uses data to categorize customers based on behavior, demographics, and preferences for targeted marketing.
- **Campaign Optimization:** Analyzes marketing campaign performance to optimize strategies and improve return on investment.
- **Social Media Analytics:** Extracts insights from social media data to understand consumer sentiment and engagement.

7. Education:

- **Student Performance Analytics:** Helps identify factors influencing student success and retention.
- **Personalized Learning:** Tailors educational content and approaches based on individual student data.
- **Institutional Planning:** Analyzes data for resource allocation, budgeting, and strategic planning.

8. Energy:

- **Predictive Maintenance:** Monitors equipment performance to predict and prevent failures, reducing downtime.
- **Grid Management:** Analyzes data for optimizing energy distribution, improving efficiency, and reducing waste.
- **Renewable Energy Optimization:** Uses data for predicting renewable energy production and optimizing its integration into the grid.

9. Transportation and Logistics:

- **Route Optimization:** Uses data to optimize transportation routes, reducing fuel costs and delivery times.
- **Predictive Maintenance:** Analyzes data from vehicles and equipment to schedule maintenance proactively.
- **Supply Chain Visibility:** Improves visibility into the supply chain for better decision-making and efficiency.

10. Government:

- **Public Safety:** Uses data for predictive policing, disaster response planning, and public health initiatives.

- **Policy Decision Support:** Analyzes data to inform policy decisions and assess the impact of government programs.
- **Smart Cities:** Utilizes data for optimizing urban planning, transportation, and resource allocation.

Importance of Data Science in the Future:

Companies today have access to massive databases due to documenting every aspect of client engagement. Data science plays a crucial role in analyzing and developing these data-driven machine-learning models. As the industry grows, more jobs should become accessible because analysis needs more data scientists. A bright future is expected for those who are interested in a career in data science.

Artificial intelligence is a key component in the future of data science. In the future, AI will likely be the most powerful tool that data scientists will have to work with. Artificial intelligence is already being used by businesses to make decisions and run their operations. Artificial intelligence will be used in real-world scenarios to use automated solutions to screen through massive volumes of data to find patterns that help present firms make better decisions.

The need for data science in various fields highlights the importance of a data science course in today's world.

Big Data Analytics:

Big data analytics is the process of collecting, storing, analyzing, and visualizing large and complex datasets to gain insights and make informed decisions. It encompasses a wide range of techniques and tools, including data mining, machine learning, statistical analysis, and data visualization.

Key characteristics of big data analytics:

- **Volume:** Big data describes datasets that are too large to be processed by traditional data processing applications.
- **Velocity:** Big data is generated and processed at a rapid pace, often in real-time.
- **Variety:** Big data can be structured, semi-structured, or unstructured, meaning it can come from a variety of sources and formats.
- **Veracity:** The accuracy and trustworthiness of big data can be challenging to ensure.

Applications of big data analytics:

- **Customer relationship management (CRM):** Big data analytics can be used to analyze customer data to gain insights into customer preferences, behaviors, and demographics. This information can then be used to personalize marketing campaigns, improve customer service, and develop new products and services.
- **Fraud detection:** Big data analytics can be used to analyze financial transactions to identify patterns that suggest fraudulent activity. This can help banks, insurance companies, and other financial institutions to detect and prevent fraud.
- **Supply chain management:** Big data analytics can be used to optimize supply chains by tracking shipments, managing inventory, and predicting demand. This can help businesses to reduce costs, improve efficiency, and improve customer service.
- **Risk management:** Big data analytics can be used to analyze data from a variety of sources to identify and assess risks. This can help businesses to make informed decisions about investments, insurance premiums, and other risk-related activities.

Challenges of big data analytics:

- **Data acquisition and preparation:** Collecting and preparing big data can be a complex and time-consuming process. Data may need to be cleaned, normalized, and integrated from various sources before it can be analyzed.
- **Data storage and management:** Big data analytics requires storing and managing large datasets. This can be expensive and complex, and it requires specialized infrastructure and expertise.
- **Data analysis and interpretation:** Analyzing large and complex datasets can be computationally intensive and require advanced analytical techniques. It is also important to interpret the results of data analysis in a meaningful way.

The future of big data analytics:

Big data analytics is a rapidly growing field with the potential to transform many industries. As the volume, velocity, variety, and veracity of data continue to increase, there will be a growing demand for professionals with expertise in big data analytics.

Here are some trends in big data analytics:

- **Artificial intelligence and machine learning:** AI and machine learning are being used to automate data analysis tasks and make more sophisticated predictions.
- **Edge computing:** Edge computing is moving data processing closer to the source of data, which can improve performance and reduce latency.

- **Real-time analytics:** Real-time analytics is becoming more important as businesses need to make decisions in real time.
- **Explainable AI:** Explainable AI is making it possible to understand how AI models make decisions, which is important for trust and transparency.

Big data analytics is a powerful tool that can be used to gain insights from data and make informed decisions. As the field continues to grow, it is likely to have an even greater impact on businesses and organizations of all sizes.

Business Intelligence vs Big Data:

Aspect	Business Intelligence (BI)	Big Data
Data Size	Typically handles structured data in moderate volumes	Handles large volumes of structured and unstructured data
Data Sources	Mainly relies on internal, structured data sources	Ingests data from diverse sources, including internal and external, structured and unstructured
Processing Speed	Analyzes historical data with a focus on reporting and querying	Requires high-speed processing to handle real-time or near-real-time data
Data Structure	Works well with structured data stored in relational databases	Can handle structured, semi-structured, and unstructured data from various sources
Analytics Type	Primarily supports descriptive analytics and reporting	Supports a broader range of analytics including descriptive, diagnostic, predictive, and prescriptive
Tools and Technologies	Relies on traditional BI tools (e.g., Tableau, Power BI)	Uses specialized big data technologies (e.g., Hadoop, Spark) and machine learning tools
Scope	Generally focused on monitoring and analyzing past performance	Addresses a wider range of use cases, including predictive analytics, real-time analytics, and machine learning
User Base	Mainly used by business analysts, managers, and decision-makers	Used by data scientists, analysts, and engineers with expertise in big data technologies
Data Storage	Typically stored in data warehouses or relational databases	Stored in distributed file systems, NoSQL databases, and cloud storage
Data Processing Approach	Batch processing for historical data analysis	Supports both batch processing and real-time/streaming processing
Costs and Infrastructure	May require less infrastructure for moderate-sized datasets	Requires scalable infrastructure and may involve higher costs due to storage and processing demands

Goal	Aims to provide actionable insights for decision-making	Aims to extract valuable insights and knowledge from large, complex datasets
Example Scenario	Analyzing sales performance based on historical data	Analyzing social media data in real-time to detect trends and sentiments

Current Landscape of Analytics:

The analytics landscape is constantly evolving, with new technologies and techniques emerging all the time. However, there are a few key trends that are shaping the future of analytics.

Key trends in analytics:

- **The rise of artificial intelligence (AI) and machine learning (ML):** AI and ML are being used to automate many of the tasks involved in data analysis, and they are also being used to develop new and more powerful analytical techniques.
- **The increasing importance of data visualization:** Data visualization is becoming increasingly important as the amount of data that organizations collect continues to grow. Data visualization tools can help organizations to quickly and easily understand complex data sets.
- **The move to cloud-based analytics:** Cloud-based analytics is becoming increasingly popular as it offers a number of advantages, including scalability, flexibility, and cost-effectiveness.
- **The growing focus on real-time analytics:** Organizations are increasingly interested in using analytics to gain insights from data in real time. This allows them to make more informed decisions and respond to events more quickly.
- **The need for data governance:** As the amount of data that organizations collect continues to grow, it is becoming increasingly important to have effective data governance in place. Data governance helps to ensure that data is accurate, reliable, and secure.

Here are some of the key technologies that are shaping the future of analytics:

- **AI and ML:** AI and ML are being used to automate many of the tasks involved in data analysis, such as data cleaning, data preparation, and feature engineering. They are also being used to develop new and more powerful analytical techniques, such as deep learning and natural language processing.

- **Big data platforms:** Big data platforms, such as Hadoop and Spark, are being used to store and process large and complex datasets. These platforms are making it possible to analyze data that was previously too large or complex to analyze.
- **Data visualization tools:** Data visualization tools are becoming increasingly sophisticated, making it easier for organizations to create clear and concise visualizations of their data. These tools are also becoming more interactive, allowing users to explore data in new and innovative ways.
- **Cloud-based analytics platforms:** Cloud-based analytics platforms are making it easier for organizations to get started with analytics without having to invest in hardware and software. These platforms are also making it easier for organizations to scale their analytics deployments as their needs grow.
- **Real-time analytics streaming platforms:** Real-time analytics streaming platforms, such as Apache Kafka and Amazon Kinesis, are making it possible to analyze data in real time. This allows organizations to make more informed decisions and respond to events more quickly.

The future of analytics is bright:

Analytics is becoming increasingly important for organizations of all sizes. As the amount of data that organizations collect continues to grow, analytics will become even more essential for understanding and making informed decisions.

Here are some of the key challenges that organizations face when adopting analytics:

- **Data quality:** Data quality is a major challenge for organizations that are adopting analytics. Data may be inaccurate, incomplete, or inconsistent, which can make it difficult to get reliable insights from data.
- **Data silos:** Data silos are another common challenge for organizations that are adopting analytics. Data may be stored in different systems and formats, which can make it difficult to access and analyze.
- **Skills shortage:** There is a growing shortage of skilled data scientists and analysts. This can make it difficult for organizations to find the talent they need to implement and manage their analytics initiatives.

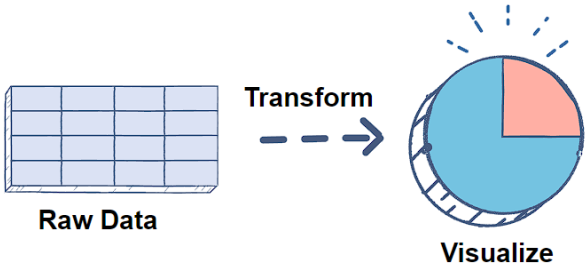
Despite these challenges, the benefits of analytics are clear. Organizations that are able to adopt analytics effectively can gain a competitive advantage by making more informed decisions, improving operational efficiency, and driving innovation.

The future of analytics is full of promise:

Analytics is a powerful tool that can be used to solve a wide variety of problems. As the field of analytics continues to evolve, we can expect to see even more innovative and powerful analytical techniques emerge. Organizations that are able to harness the power of analytics will be well-positioned for success in the years to come.

Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) is a way to investigate datasets and find preliminary information, insights, or uncover underlying patterns in the data. Instead of making assumptions, data can be processed in a systematic method to gain insights and make informed decisions.



Why Exploratory Data Analysis?

Some advantages of Exploratory Data Analysis include:

1. Improve understanding of variables by extracting averages, mean, minimum, and maximum values, etc.
2. Discover errors, outliers, and missing values in the data.
3. Identify patterns by visualizing data in graphs such as box plots, scatter plots, and histograms.

Hence, the main goal is to understand the data better and use tools effectively to gain valuable insights or draw conclusions.

The Foremost Goals of EDA:

1. Data Cleaning: EDA involves examining the information for errors, lacking values, and inconsistencies. It includes techniques including records imputation, managing missing statistics, and figuring out and getting rid of outliers.

2. Descriptive Statistics: EDA utilizes precise records to recognize the important tendency, variability, and distribution of variables. Measures like suggest, median, mode, preferred deviation, range, and percentiles are usually used.

3. Data Visualization: EDA employs visual techniques to represent the statistics graphically. Visualizations consisting of histograms, box plots, scatter plots, line plots, heatmaps, and bar charts assist in identifying styles, trends, and relationships within the facts.

4. Feature Engineering: EDA allows for the exploration of various variables and their adjustments to create new functions or derive meaningful insights. Feature engineering can contain scaling, normalization, binning, encoding express variables, and creating interplay or derived variables.

5. Correlation and Relationships: EDA allows discover relationships and dependencies between variables. Techniques such as correlation analysis, scatter plots, and pass-tabulations offer insights into the power and direction of relationships between variables.

6. Data Segmentation: EDA can contain dividing the information into significant segments based totally on sure standards or traits. This segmentation allows advantage insights into unique subgroups inside the information and might cause extra focused analysis.

7. Hypothesis Generation: EDA aids in generating hypotheses or studies questions based totally on the preliminary exploration of the data. It facilitates form the inspiration for in addition evaluation and model building.

8. Data Quality Assessment: EDA permits for assessing the nice and reliability of the information. It involves checking for records integrity, consistency, and accuracy to make certain the information is suitable for analysis.

Types of EDA:

EDA, or Exploratory Data Analysis, refers back to the method of analyzing and analyzing information units to uncover styles, pick out relationships, and gain insights. There are various sorts of EDA strategies that can be hired relying on the nature of the records and the desires of the evaluation. Here are some not unusual kinds of EDA:

1. Univariate Analysis: This sort of evaluation makes a speciality of analyzing character variables inside the records set. It involves summarizing and visualizing a unmarried variable at a time to understand its distribution, relevant tendency, unfold, and different applicable records. Techniques like histograms, field plots, bar charts, and precis information are generally used in univariate analysis.

2. Bivariate Analysis: Bivariate evaluation involves exploring the connection between variables. It enables find associations, correlations, and dependencies between pairs of variables. Scatter plots, line plots, correlation matrices, and move-tabulation are generally used strategies in bivariate analysis.

3. Multivariate Analysis: Multivariate analysis extends bivariate evaluation to encompass greater than variables. It ambitions to apprehend the complex interactions and dependencies among more than one variable in a records set. Techniques inclusive of heatmaps, parallel coordinates, aspect analysis, and primary component analysis (PCA) are used for multivariate analysis.

4. Time Series Analysis: This type of analysis is mainly applied to statistics sets that have a temporal component. Time collection evaluation entails inspecting and modeling styles, traits, and seasonality inside the statistics through the years. Techniques like line plots, autocorrelation analysis, transferring averages, and ARIMA (AutoRegressive Integrated Moving Average) fashions are generally utilized in time series analysis.

5. Missing Data Analysis: Missing information is a not unusual issue in datasets, and it may impact the reliability and validity of the evaluation. Missing statistics analysis includes figuring out missing values, know-how the patterns of missingness, and using suitable techniques to deal with missing data. Techniques along with lacking facts styles, imputation strategies, and sensitivity evaluation are employed in lacking facts evaluation.

6. Outlier Analysis: Outliers are statistics factors that drastically deviate from the general sample of the facts. Outlier analysis includes identifying and knowledge the presence of outliers, their capability reasons, and their impact at the analysis. Techniques along with box plots, scatter plots, z-rankings, and clustering algorithms are used for outlier evaluation.

7. Data Visualization: Data visualization is a critical factor of EDA that entails creating visible representations of the statistics to facilitate understanding and exploration. Various visualization techniques, inclusive of bar charts, histograms, scatter plots, line plots, heatmaps, and interactive dashboards, are used to represent exclusive kinds of statistics.

These are just a few examples of the types of EDA techniques that can be employed at some stage in information evaluation. The choice of strategies relies upon on the information traits, research questions, and the insights sought from the analysis.

Advantages of Using EDA:

Here are a few advantages of using Exploratory Data Analysis -

1. Gain Insights into Underlying Trends and Patterns

EDA assists data analysts in identifying crucial trends quickly through data visualizations using various graphs, such as box plots and histograms. Businesses also expect to make some unexpected discoveries in the data while performing EDA, which can help improve certain existing business strategies.

2. Improved Understanding of Variables

Data analysts can significantly improve their comprehension of many factors related to the dataset. Using EDA, they can extract various information such as averages, means, minimum and maximum, and more such information is required for preprocessing the data appropriately.

3. Better Preprocess Data to Save Time

EDA can assist data analysts in identifying significant mistakes, abnormalities, or missing values in the existing dataset. Handling the above entities is critical for any organization before beginning a full study as it ensures correct preprocessing of data and may help save a significant amount of time by avoiding mistakes later when applying machine learning models.

4. Make Data-driven Decisions

The most significant advantage of employing EDA in an organization is that it helps businesses to improve their understanding of data. With EDA, they can use the available tools to extract critical insights and make conclusions, which assist in making decisions based on the insights from the EDA.

Statistical Measures:

Statistical measures are numerical values that provide insights into various aspects of a dataset's distribution, central tendency, variability, and relationships between variables.

Statistical measures are a descriptive analysis technique used to summarise the characteristics of a data set. This data set can represent the whole population or a sample of it. Statistical measures can be classified as measures of central tendency and measures of spread.

Here are some common statistical measures used in Exploratory Data Analysis (EDA):

Measures of location (1-3):

1. **Mean:** The average value of a set of numbers. It is calculated by summing all the values and dividing by the number of observations.

$$\text{Mean } (\mu) = \frac{\Sigma x}{n}$$

2. **Median:** The middle value of a dataset when it is sorted in ascending or descending order. It is less sensitive to extreme values than the mean. In cases where the midpoint values are two (when the number of data points is even), you need to find the average of both middle values. When finding the median, it is appropriate to reorder your values in ascending order. Take the $\frac{n+1}{2}$ value if the number of data points is odd. When the number is even, take the $\frac{n}{2}$ and the **value** $\frac{n+2}{2}$**.**

3. **Mode:** The value that appears most frequently in a dataset. A dataset may have one mode, more than one mode, or no mode at all.

Measures of Spread (4-7):

4. **Variance** $(\boldsymbol{\sigma^2})$**:** A measure of the spread or dispersion of a set of values. It is calculated as the average of the squared differences from the mean.

$$\sigma^2 = \frac{\Sigma(x_1 - \mu)^2}{N}$$

5. **Standard Deviation** $(\boldsymbol{\sigma})$**:** The square root of the variance. It provides a measure of the average distance between each data point and the mean.

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

6. **Range:** The difference between the maximum and minimum values in a dataset.

$$\text{Range} = \text{Max} - \text{Min}$$

7. **Quartiles and the Interquartile Range (IQR):**

A **quartile** is a type of quantile that divides an ordered data set into four parts (quarters). A quartile is not the group of numbers that have been divided. It is the cut-off point in the division.

The interquartile range is the difference between the upper quartile and the lower quartile value.

To find the quartile of a given data set you can proceed as follows:

1. Order the values in ascending order.
2. Find the median. This is always labelled as the second quartile (Q_2).
3. Now find the median of both halves of the data set. The lowest half is labelled Q_1, and the highest half is labelled Q_3.
4. Find the interquartile range (IQR) by subtracting Q_1 from Q_3.

$$\text{IQR}=Q3-Q1$$

8. **Percentiles:** Values that divide a dataset into 100 equal parts. The median is the 50th percentile.

9. **Correlation Coefficient:** A measure of the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

$$Correlation(X,Y) = \rho_{XY} = \frac{Covariance(X,Y)}{\sigma_X \sigma_Y}$$

10. **Covariance:** A measure of how much two random variables change together. Positive values indicate a positive relationship, while negative values indicate a negative relationship.

$$Covariance(X,Y) = \sigma_{XY} = \frac{\sum(x_i - \mu_X)(y_i - \mu_Y)}{n-1}$$

1. **Skewness:** A measure of the asymmetry of a distribution. Positive skewness indicates a longer tail on the right, while negative skewness indicates a longer tail on the left.

2. **Kurtosis:** A measure of the "tailedness" of a distribution. High kurtosis indicates heavy tails and a peaked distribution, while low kurtosis indicates light tails and a flat distribution.

These statistical measures provide a quantitative summary of the characteristics of a dataset and play a crucial role in understanding its properties during the exploratory data analysis phase.

Basic Tools of EDA:

Programming Language Tools Used in EDA:

Some of the most common tools used to create an EDA are:

1. R: An open-source programming language and free software environment for statistical computing and graphics supported by the R foundation for statistical computing. The R language is widely used among statisticians in developing statistical observations and data analysis.

2. Python: An interpreted, object-oriented programming language with dynamic semantics. Its high level, built-in data structures, combined with dynamic binding, make it very attractive for rapid application development, also as to be used as a scripting or glue language to attach existing components together. Python and EDA are often used together to spot missing values in the data set, which is vital so you'll decide the way to handle missing values for machine learning.

Apart from these functions described above, EDA can also:

- **Perform k-means clustering:** Perform k-means clustering: it's an unsupervised learning algorithm where the info points are assigned to clusters, also referred to as k-groups, k-means clustering is usually utilized in market segmentation, image compression, and pattern recognition
- EDA is often utilized in predictive models like linear regression, where it's wont to predict outcomes.
- It is also utilized in univariate, bivariate, and multivariate visualization for summary statistics, establishing relationships between each variable, and understanding how different fields within the data interact with one another.

Plots & Graphs:

A) Univariate Analysis:

1. Histogram:

A histogram is a graphical representation of the distribution of a dataset. It provides a visual summary of the underlying frequency distribution of a set of continuous or discrete data. Histograms are particularly useful for understanding the shape, central tendency, and spread of the data.

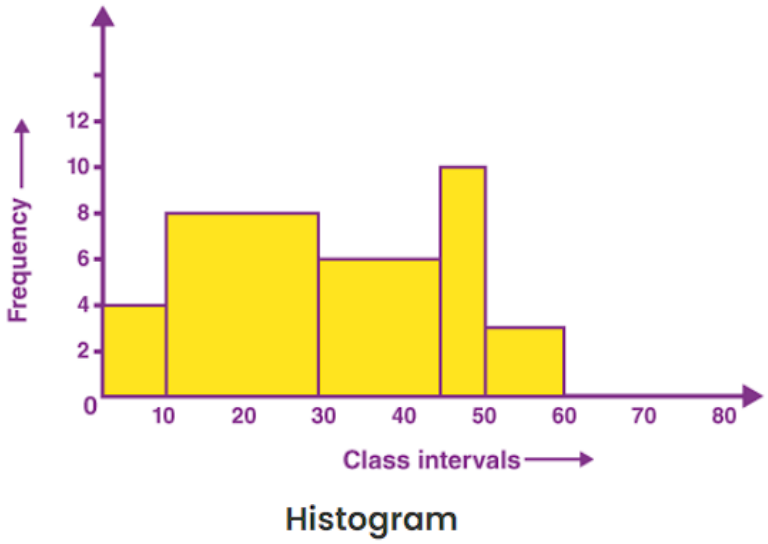
In other words, a histogram is a diagram involving rectangles whose area is proportional to the frequency of a variable and width is equal to the class interval.

Although histograms seem similar to graphs, there is a slight difference between them. The histogram does not involve any gaps between the two successive bars.

When to Use Histogram?

The histogram graph is used under certain conditions. They are:

- The data should be numerical.
- A histogram is used to check the shape of the data distribution.
- Used to check whether the process changes from one period to another.
- Used to determine whether the output is different when it involves two or more processes.
- Used to analyse whether the given process meets the customer requirements.



Histogram

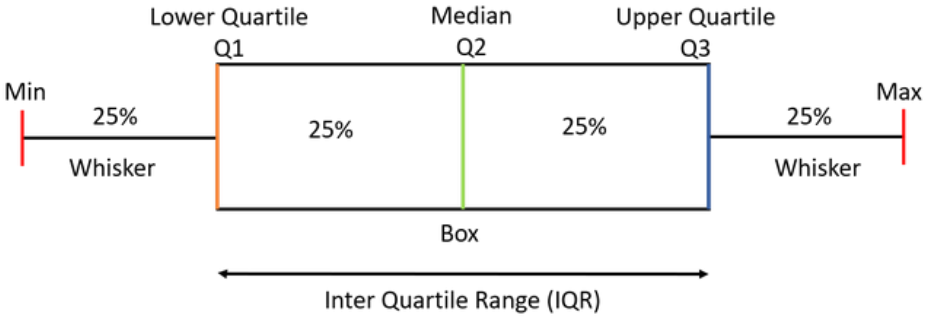
2. Box Plot (Box and Whisker Plot):

A box plot provides a visual summary of the distribution of a dataset. It shows the minimum, first quartile, median, third quartile, and maximum of a set of data.

It helps identify the presence of outliers and gives a sense of the overall distribution.

Parts of Box Plots: 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.

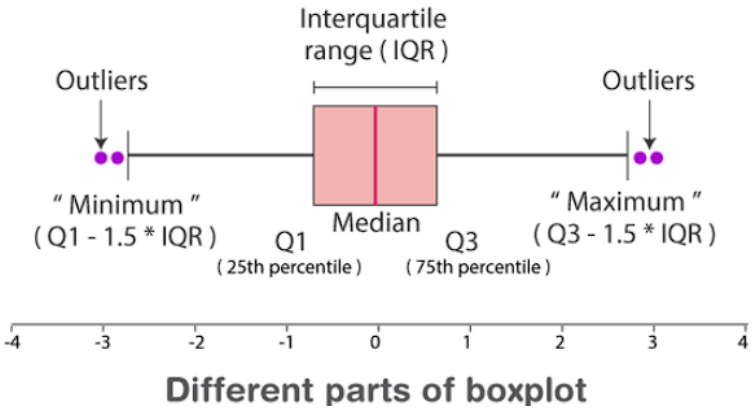


Note: The box plot shown in the above diagram is a perfect plot with no skewness. The plots can have skewness and the median might not be at the center of the box.

Interquartile Range (IQR): The difference between the third quartile and first quartile is known as the interquartile range. (i.e.) IQR = Q3-Q1

Outlier: The data that falls on the far left or right side of the ordered data is tested to be the outliers. Generally, the outliers fall more than the specified distance from the first and third quartile.

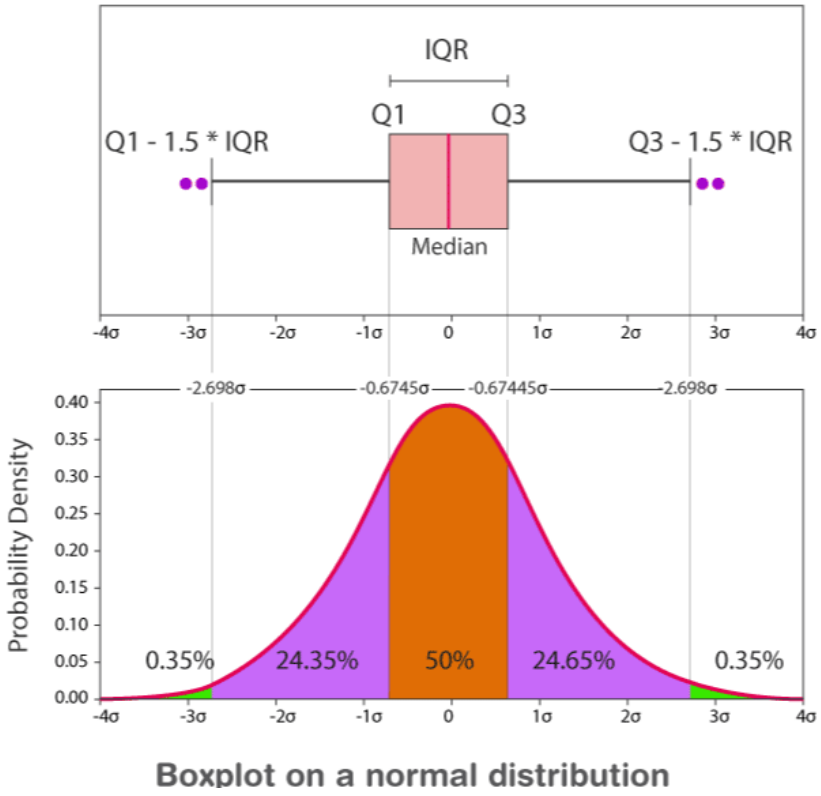
(i.e.) Outliers are greater than Q3+(1.5 . IQR) or less than Q1-(1.5 . IQR).



Different parts of boxplot

Boxplot Distribution:

The box plot distribution will explain how tightly the data is grouped, how the data is skewed, and also about the symmetry of data.



Boxplot on a normal distribution

Positively Skewed: If the distance from the median to the maximum is greater than the distance from the median to the minimum, then the box plot is positively skewed.

Negatively Skewed: If the distance from the median to minimum is greater than the distance from the median to the maximum, then the box plot is negatively skewed.

Symmetric: The box plot is said to be symmetric if the median is equidistant from the maximum and minimum values.

Applications:

It is used to know:

- The outliers and their values
- Symmetry of Data
- Tight grouping of data
- Data skewness – if, in which direction and how

3. Density Plot:

A density plot, also known as a **kernel density plot** or a **density trace graph**, is a powerful tool in data science for visualizing the distribution of a numerical variable. It provides a smooth and continuous representation of the data compared to the traditional histogram, making it ideal for understanding the underlying shape of the distribution and identifying areas of high or low density.

Here's a breakdown of density plots:

What it is:

- A visual representation of the probability density function (PDF) of a continuous variable.
- Unlike histograms that divide data into bins, density plots use kernel smoothing to create a continuous curve.
- This curve shows the probability of a data point falling within a specific range of values.

Why use it:

- **Reveal the true shape of the distribution:** Density plots are not affected by the choice of bin width, unlike histograms. This allows them to accurately capture the underlying shape of the distribution, even for complex or multimodal data.
- **Identify areas of high/low density:** The peaks of the curve indicate areas where data points are concentrated, while the valleys represent regions with fewer data points.
- **Compare multiple distributions:** Density plots can be overlaid on the same graph to compare the distributions of different variables or groups of data.

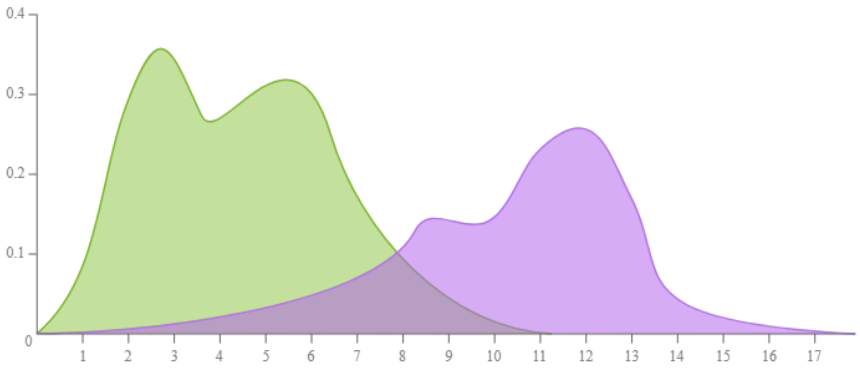
How it works:

- Kernel density estimation (KDE) is the underlying technique used to create density plots.
- KDE places a "kernel" (a smooth, bell-shaped curve) around each data point.
- The contributions of all the kernels are then summed up to create the overall density curve.
- The shape and width of the kernel can be adjusted to influence the smoothness and level of detail in the plot.

Examples of use cases:

- Analyzing the distribution of income levels in a population
- Visualizing the age distribution of customers for a particular product
- Comparing the height distribution of male and female athletes
- Identifying outliers in a dataset

Here's an example of a density plot:



4. Violin Plot: A violin plot, also known as a kernel density violin plot, is another valuable tool in data science for visualizing the distribution of numerical data. It combines the features of a box plot with a rotated kernel density estimation, providing a more comprehensive view of the data compared to either alone.

Here's a breakdown of violin plots:

What it is:

- A hybrid visualization combining a box plot and a kernel density plot.
- The box plot in the center displays the median, quartiles, and outliers.
- The mirrored kernel density plots on either side show the probability density of the data at different values, revealing the overall shape of the distribution.

Why use it:

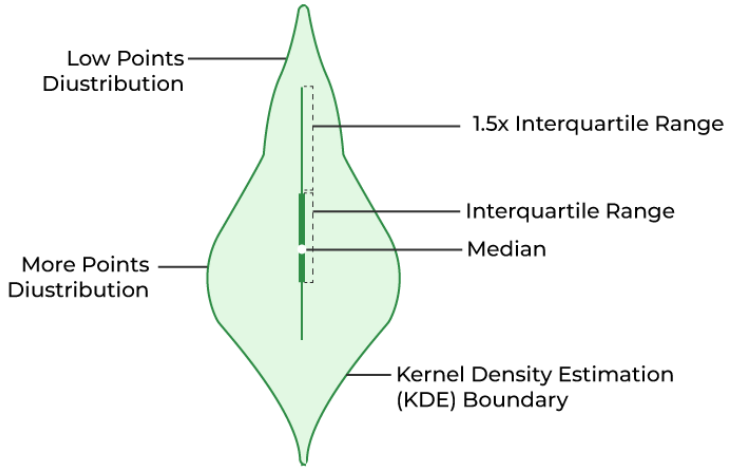
- **Provides more information than a box plot:** Violin plots reveal not only summary statistics like medians and quartiles but also the full distribution of the data, including skewness, tail behavior, and multimodality.
- **Easier to compare multiple distributions:** Overlaying violin plots for different groups allows for visual comparison of their distributions, highlighting differences in shape, spread, and central tendency.

- **Useful for multimodal data:** Violin plots excel at showcasing data with multiple peaks, providing insights into the underlying structure of the distribution.

How it works:

- The box plot portion is created using standard box plot calculations for quartiles, median, and outliers.
- The kernel density curves are generated using kernel density estimation (KDE), similar to a standard density plot.
- The mirrored curves are scaled by the data density and overlaid on either side of the box plot.

Here's an example of a violin plot:



Violin plot Distribution Explanation

A violin plot consists of four components.

- **A white Centered Dot at the middle of the graph** – The white dot point at the middle is the median of the distribution.
- **A thin gray bar inside the plot –** The bar in the plot represents the Quartile range of the distribution
- **A long thin line coming outside from the bar –** The thin line represents the rest of the distribution which is calculated by the formulae Q1-1.5 IQR for the lower range and Q3+1.5 IQR for the upper range. The point lying beyond this line are considered as outliers
- **A line boundary separating the plot-** A kDE plot is used for defining the boundary of the violin plot it represents the distribution of data points

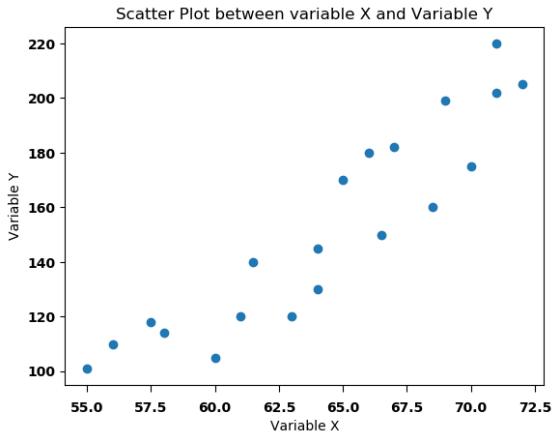
Examples of use cases:

- Comparing the income distribution of different age groups
- Investigating the height distribution of male and female athletes across different sports
- Analyzing the performance of different algorithms on a dataset
- Identifying outliers and potential subpopulations within a dataset

B) Bivariate Analysis:

1. Scatter Plot:

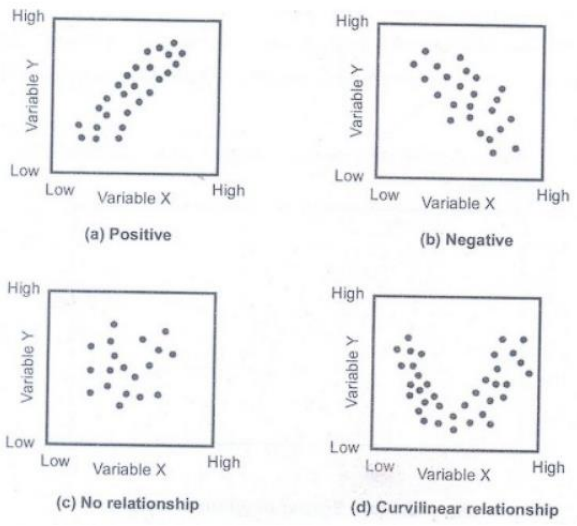
A scatter plot is a graphical representation of individual data points in a two-dimensional space. It displays the values of two continuous variables as points on a Cartesian plane, with one variable on the x-axis and the other on the y-axis. Each point on the plot represents a pair of values from the dataset.



The patterns or correlations found within a scatter plot will have a few different features.

- **Linear or Nonlinear:** A linear correlation forms a straight line in its data points while a nonlinear correlation might have a curve or other form within the data points.
- **Strong or Weak:** A strong correlation will have data points close together while a weak correlation will have data points that are further apart.
- **Positive or Negative:** A positive correlation will point up (i.e., the x- and y-values are both increasing) while a negative correlation will point down (i.e., the x-values are increasing while the corresponding y-values are decreasing).

However, if you don't see any of these features present within your graph, that means there's no correlation between your data.



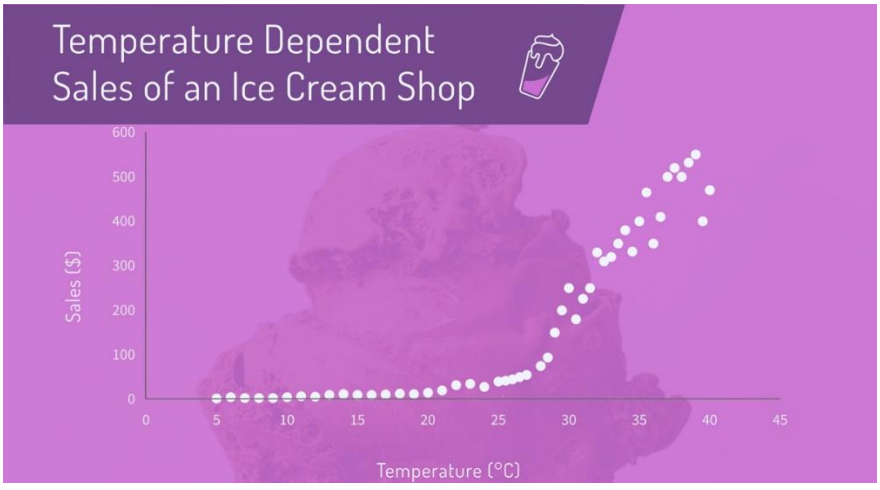
When To Use a Scatter Plot:

1. Use a scatter plot to determine whether or not two variables have a relationship or correlation.

Are you trying to see if your two variables might mean something when put together? Plotting a scattergram with your data points can help you to determine whether there's a potential relationship between them.

Let's say you're running an ice cream business, and you're curious to see if there's a pattern in why your sales have been low recently.

You might create a scatter plot to measure different factors, including outside temperature.



You always want to plot your scatter diagram with both the x-axis and the y-axis increasing as they go out so that you can determine correlation.

As we can see in the above example, people tend to buy ice cream – a cold dessert – less often when the temperature is cold outside.

2. Use a scatter plot when your independent variable has multiple values for your dependent variable.

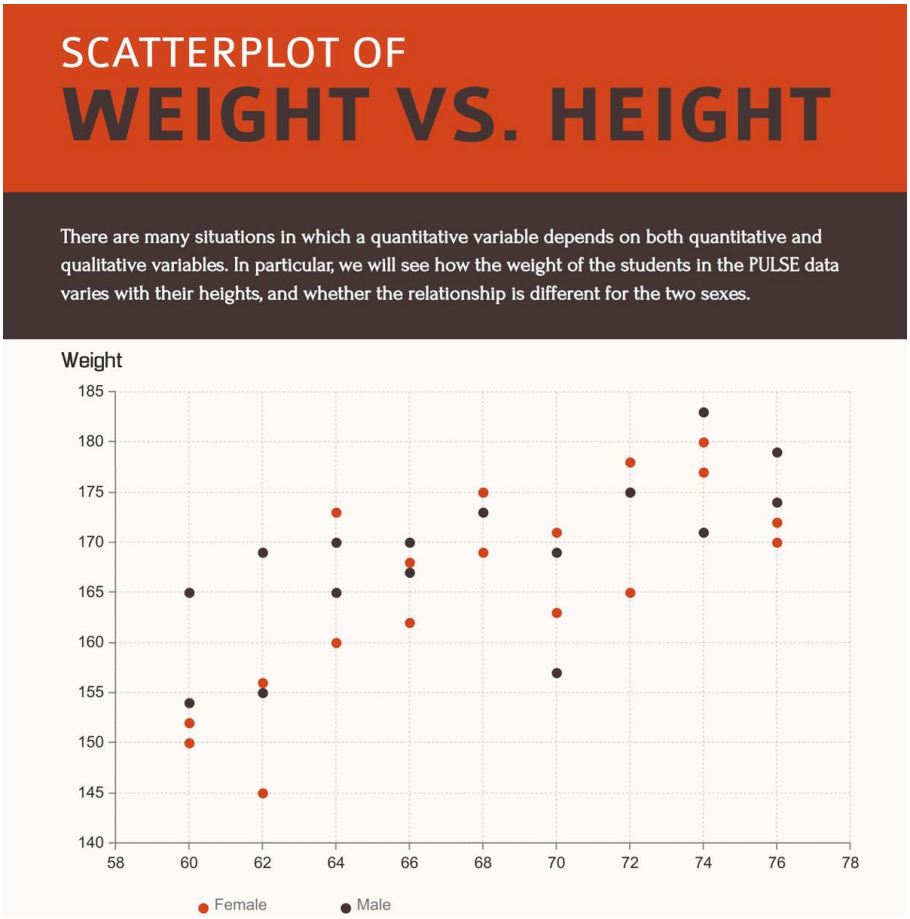
Okay, let's take it back to math class for a minute and go over what independent and dependent variables mean.

First of all, a variable is the thing you're trying to track or measure. Every graph has two variables – an independent variable that is typically graphed on the x-axis and a dependent variable that is typically graphed on the y-axis.

An independent variable is the controlled variable. This is what changes naturally, or what the person manipulating the experiment or graph changes.

A dependent variable is the variable that is being studied or measured. In the case of a scatter plot, it's the variable that we're looking to determine whether or not has a correlation with the independent variable.

If you're trying to determine if height and weight have a correlation, the height will be placed on the x-axis and weight will be placed on the y-axis, like in the example below.



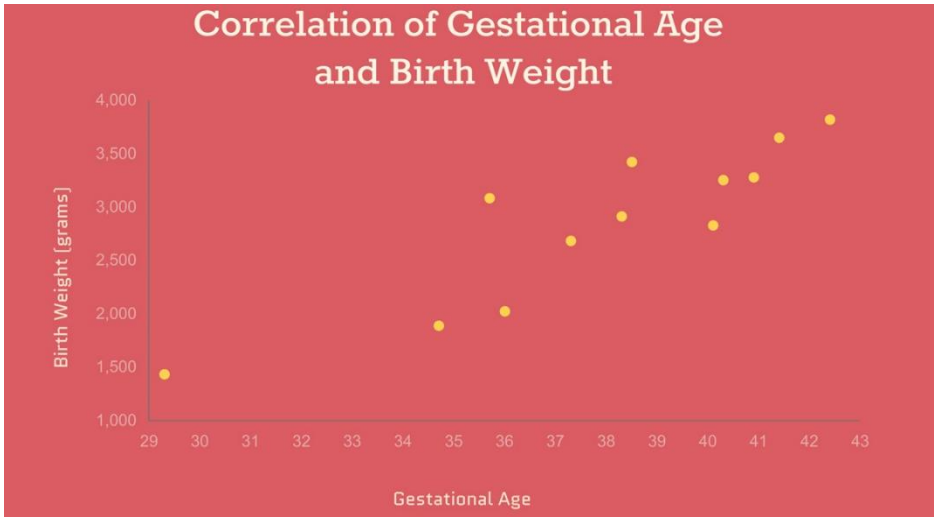
Because weight fluctuates much more than height, it's likely that you could have different weights for the same height in your data, giving you more than one dependent variable value for each independent variable.

3. Use a scatter plot when you have two variables that pair well together.

If you have two variables that pair well together, plotting them on a scatter diagram is a great way to view their relationship and see if it's a positive or negative correlation.

For example, think about birth weight versus gestational age (how long the baby has been in utero). It would make sense that a baby who was able to grow inside its mother for longer would be larger, and therefore weigh more, correct?

Let's take a look at this data on a scatter plot.



As we would expect, the longer a baby is able to "cook," the more it tends to weigh at birth.

Other examples of variables that appear to go hand in hand would be hours worked versus money made, time studied versus test grade or price versus diamond size.

When Not to Use a Scatter Plot:

1. Avoid a scatter plot when your data is not at all related.

There are certain variables that make it obvious that there's no correlation, therefore a scatter plot would be a useless way to visualize your information.

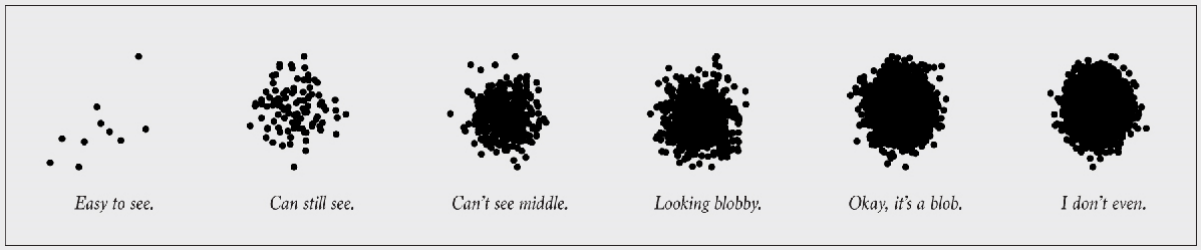
For example, if you're gathering a random survey on a classroom full of students, putting together the students' varying heights and the number of pets they have at home would make no sense on a scatter plot.

These two variables obviously have no relationship whatsoever, and while they can still be fun to graph, a bar chart (one for each data value) might be the better choice here.

2. Avoid a scatter plot when you have too large a set of data.

When you have so much data in your scatter plot that it clogs up the entire graph, this is the result of overplotting.

Statistician Nathan Yau sums up this phenomenon pretty well in the below graphic:



There are a few ways to counteract an overplotted scatter plot, though. First, consider using a heatmap that shows where the most point-heavy sections of your data are.

You could also color code various data sets, use translucent data points to create a heatmap-like effect and more.

However, your best bet is to avoid using a scatter plot when you have so much data that it becomes a large blob.

Things to Keep in Mind With a Scatter Plot:

1. Correlation is not always causation.

Just because you might see a strong positive or negative correlation in your data does not necessarily mean that your independent variable is the reason your dependent variable is measuring the way it is.

These are correlations, meaning that it appears that your independent variable does have some sort of effect on your dependent variable.

Let's jump back into our ice cream sales example.

While it may seem that the weather is the direct cause of a decrease in sales, there could be so many other factors that are leading to slower business.

Perhaps there was a natural disaster like a hurricane that led to a mandatory evacuation and therefore less business. A new ice cream shop could have opened down the street creating competition that wasn't there before.

Some days people just don't want to buy ice cream. And while, sure, the colder weather might be a factor, just because you see a correlation on a scatter plot does not mean you should take it as law.

2. You can have more than one dependent variable.

Your data set might include more than one dependent variable, and you can still track this on a scatter plot.

The only thing you'll want to change is the color of each dependent variable so that you can measure them against each other on the scatter plot.

Let's take a look back at our height versus weight example.

In that scatter plot, we added two different dependent variables – male and female – to see if there was also a difference between those factors. We colored female points orange and male points brown so that we could differentiate between the two.

This is another great way to avoid overplotting. Ensuring you're color coding your data helps to set it apart so that you can see more of your points.

Common issues when using scatter plots:

1. **Overplotting:**
   - **Issue:** When dealing with a large number of data points, they may overlap, making it difficult to discern individual points and patterns.
   - **Solution:** Consider using transparency (alpha blending) for points or using other techniques such as hexbin plots or 2D kernel density estimation to address overplotting.
2. **Scaling Issues:**
   - **Issue:** Differences in scale between the x and y axes can distort the appearance of the plot, making it challenging to accurately interpret the relationship between variables.
   - **Solution:** Ensure that the axes are appropriately scaled. Logarithmic scales or other transformations may be applied if necessary.
3. **Outliers:**
   - **Issue:** Outliers can disproportionately influence the interpretation of the scatter plot and may lead to inaccurate assessments of trends or correlations.
   - **Solution:** Consider identifying and addressing outliers before generating the scatter plot. Robust regression techniques or transforming the data may help mitigate the impact of outliers.
4. **Nonlinear Relationships:**
   - **Issue:** Assuming a linear relationship when the true relationship is nonlinear can lead to inaccurate interpretations.

- **Solution:** If a linear relationship is not apparent, consider exploring nonlinear relationships using alternative visualizations or fitting nonlinear regression models.

5. **Correlation vs. Causation:**

- **Issue:** A strong correlation between two variables in a scatter plot does not imply causation. It's important to recognize that correlation does not imply a direct cause-and-effect relationship.
- **Solution:** Always consider the context and additional information before making causal interpretations. Correlation may be coincidental or influenced by other variables.

6. **Limited to Two Variables:**

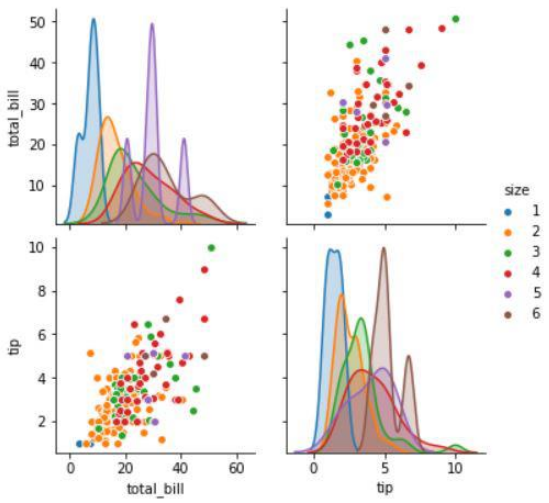
- **Issue:** Scatter plots are designed for visualizing the relationship between two variables. Exploring relationships involving more than two variables may require additional visualizations or statistical techniques.
- **Solution:** Use complementary visualizations, such as pair plots or 3D scatter plots, for exploring relationships involving more than two variables.

7. **Misleading Axis Labels:**

- **Issue:** Misleading or poorly labeled axes can lead to misinterpretation of the plot.
- **Solution:** Clearly label the x and y axes, include units, and provide a title that accurately describes the content of the scatter plot.

2. Pair Plot:

A pair plot (or pairs plot) is a grid of scatter plots that shows the relationships between all pairs of variables in a dataset. Along the diagonal, it usually displays histograms or kernel density plots for each variable.



When to Use a Pair Plot:

- Ideal for exploring relationships between multiple continuous variables in a dataset.
- Useful for identifying patterns, correlations, and potential outliers.
- Effective for gaining an overall understanding of the pairwise interactions in the data.

Advantages:

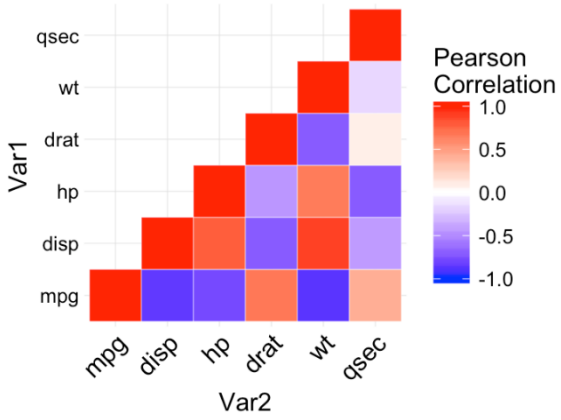
- Provides a comprehensive view of the relationships between multiple variables in one visualization.
- Useful for identifying potential areas of interest for further analysis.

Drawbacks:

- Can become visually cluttered with many variables
- may not be effective for categorical data.

3. Heatmap:

A heatmap is a graphical representation of data where values in a matrix are represented as colors. In the context of EDA, a heatmap is often used to visualize the correlation matrix between variables.



When to Use a Heatmap:

- Ideal for assessing the strength and direction of correlations between continuous variables.
- Useful for identifying multicollinearity (high correlations between independent variables) in regression analysis.

Advantages:

- Efficiently displays the correlation structure of a dataset.
- Clearly highlights areas of strong positive or negative correlation.

Drawbacks:

- Doesn't show detailed relationships like scatter plots
- limited to correlation information.

Choosing Between Pair Plot and Heatmap:

- **For Multivariate Exploration:** If you want to explore relationships between multiple variables simultaneously, a pair plot is more suitable as it provides a grid of scatter plots, allowing you to compare relationships across various pairs of variables.
- **For Correlation Assessment:** If your primary interest is in understanding the correlation structure between variables, a heatmap is more focused and efficient. It directly displays the correlation coefficients in a matrix form, making it easy to identify strong and weak correlations.
- **Combining Both:** In some cases, using both visualizations can be beneficial. Start with a pair plot for a broad overview of relationships, and then use a heatmap to specifically analyze the correlation matrix.

C) Categorical Data Analysis:

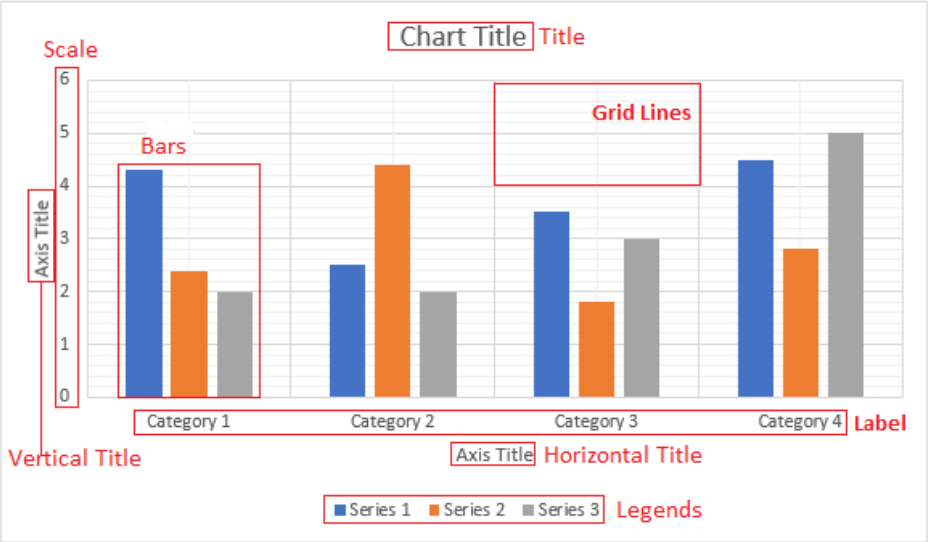
1. Bar Graph/Chart:

A **bar graph** or **bar chart** can be defined as a graph or chart that represents explicit data in the form of rectangular bars. In short, a bar graph is a graph with rectangular bars, either horizontal or vertical. A bar chart with vertical bars is also called a column chart. The length of the bars depends on the values because the bars are proportional to the values.

Components of a Bar Graph:

- **Chart Title:** It denotes the name of the bar chart. In this, we can write what the chart is representing.
- **Grid Lines:** The vertical and horizontal lines in gray color is called grid lines.
- **Bars:** A bar is corresponding to a value. It may be horizontal or vertical. The largest bar represents the largest value.

- **Axis Title:** A bar graph has two titles one is **vertical**, and the other is **horizontal**. Both the axis is related to each other. We can write the axis title for easy understanding. Suppose, the vertical axis represents expenses. So, we can write **Expenses (in rupees)** on the vertical axis. The expenses may be of different types, so we can write **types of expenses** on the horizontal axis.
- **Labels:** We can also categorize the horizontal axis title. For example, types of expenses can be categorized into **medical, transport, office**, etc.
- **Legends:** A legend specifies what a bar is representing. It is also known as the **key** of a chart. Consider the following graph; if we write **2019** in place of **Series 1**, it means the blue bars in the graph represent the data of the year 2019.
- **Scale:** The scale represents the **vertical values**. It may include rupees, population, size, etc.



Advantages of Bar Graph:

- It represents the data in pictorial form.
- It makes the analysis of data easy.
- We can easily compare data with other variables.
- It does not require too much effort to plot it.
- It is a widely used method of data representation.
- It is easy to understand.
- It is used in various fields such as retail, business, sports, etc.

When to Use a Bar Chart:

- When you need to compare a large set of categorical values
- When required to compare multiple categories or sub-categories simultaneously
- When you need to visualize two data sets on a single chart
- When you need to gather insights on deviations in data

Properties of Bar Graph:

- It may plot either vertical or horizontal.
- The space between bars and the width of the bars must be equal.
- The length of the bars represents the value of the variable being displayed.
- It must have a title, labels for each bar, and the scale for the length of the bar.

How to Plot a Bar Graph:

- **Collect Data:** The collection of data is the primary step to plot any graph. Without the data, we cannot plot a graph.
- **Draw the Axis:** A graph has two axes, **x-axis** (horizontal) and **y-axis** (vertical).
- **Label the Axis:** Label both the axis x and y. On the x-axis and y-axis, we represent the **categories** and **frequencies**, So, plot the points on the y-axis accordingly.
- **Draw Bars:** Draw the bars from the base and extend it to their corresponding frequency. If the value lies between the two plotted frequencies, take the approximate the value between these frequencies.
- **Interpret the Data:** After completing all the above steps, we can interpret the graph.

Types of Bar Graph:

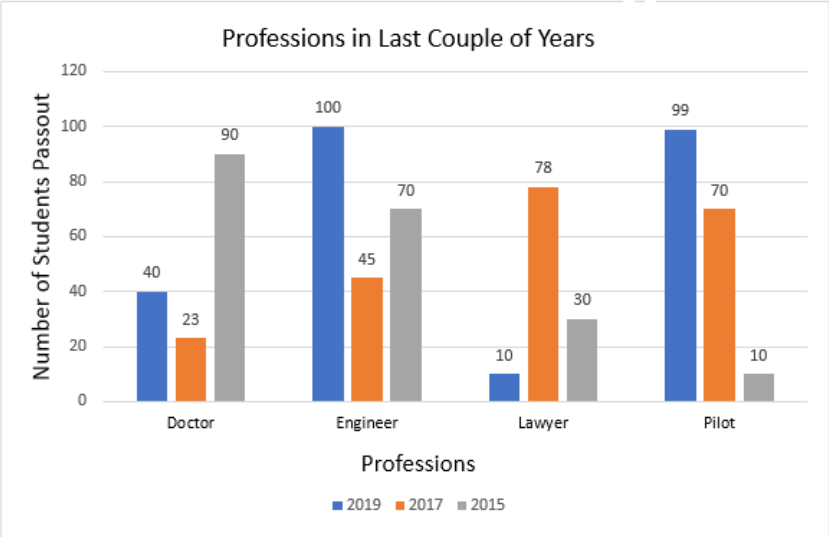
1. Vertical Bar Graph:

It is the most common type of bar graph. It is used when we want to present a series of data over time. In a vertical bar graph, categories appear at the horizontal axis. If a bar chart is arranged in highest to lowest frequency is called a **Pareto chart**. We can analyze and compare the data just by looking at the graph.

In the following table, we have taken the sample data of the years 2019, 2017, and 2015 of different professions.

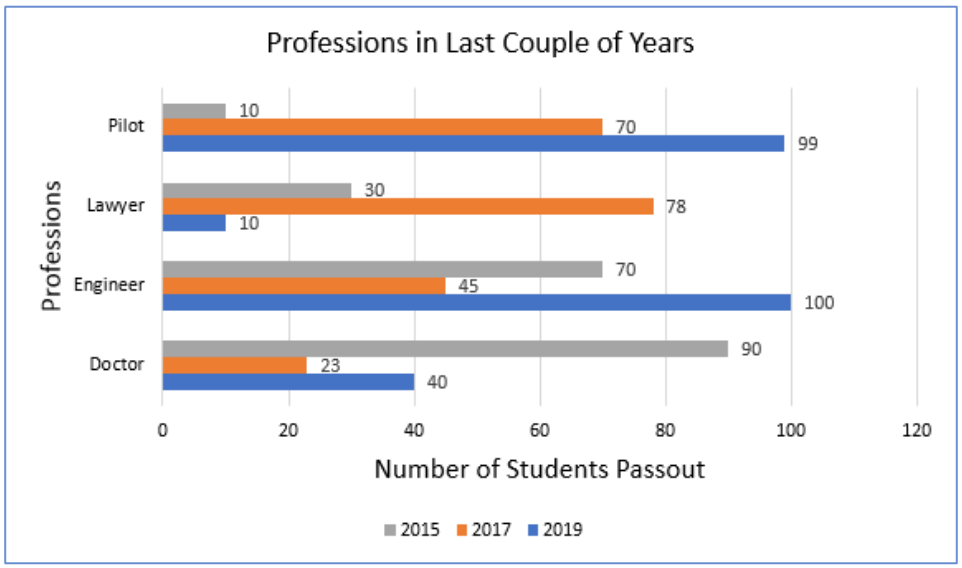
	2019	2017	2015
Doctor	40	23	90
Engineer	100	45	70
Lawyer	10	78	30
Pilot	99	70	10

Let's plot a vertical bar graph for the above data. The following graph shows the number of students who pass out in different professions.



2. Horizontal Bar Graph:

A horizontal bar graph is nothing but the representation of bars horizontally is called a horizontal bar graph. Let's plot a horizontal bar graph for the above data.

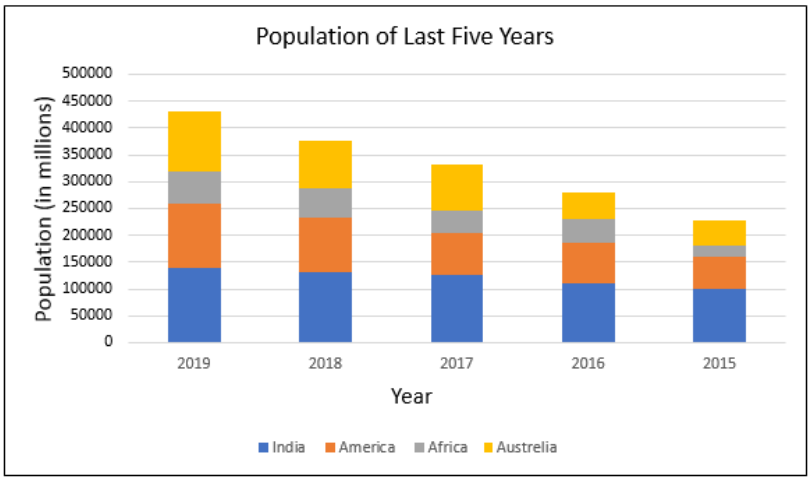


3. Stack Bar Graph:

The stack bar graph represents a lot of information about data. It can represent more than two data series within each category. Each bar shows the total for sub-groups within each individual category.

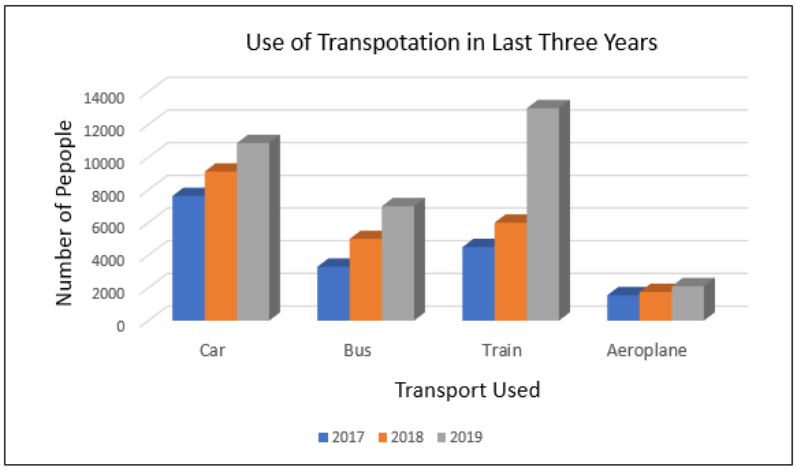
	India	America	Africa	Australia
2019	140000	120000	60000	110000
2018	132000	100000	55000	90000
2017	125000	80000	40000	86745
2016	110000	75000	44000	50987
2015	100000	59832	22000	44981

In the following bar graph, a bar represents a population of four different countries in a year.



4. 3D Bar Graph:

The following bar graph shows a 3D bar graph.



When Not to Use a Bar Chart:

1. When you need to represent and compare a continuous set of data

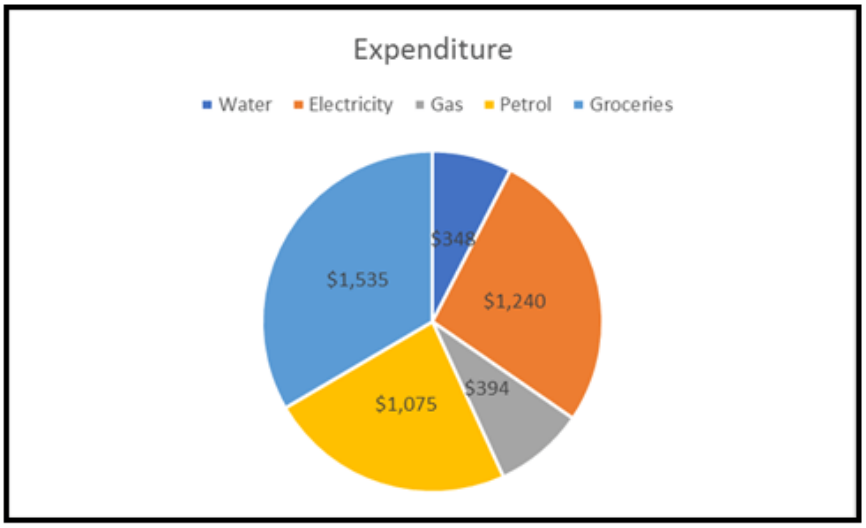
Do not use Bar Graphs when you need to represent continuous ordered quantities. In case of continuous data (such as a person's height) needs to be represented then use a Histogram. It is a best practice to leave gaps between the bars of a Bar Graph, so it doesn't look like a Histogram.

2. When you need to represent trends in time and other variables

Use Line graphs to represent trends in numerous quantities over time, by using multiple lines instead of using bar graphs which will make it more difficult to visualize multiple trends even with stacked or group charts. Line graphs have an advantage in that it's easier to see small changes on line graphs than bar graphs, and that the line makes the overall trends very clear.

2. Pie Chart:

A pie chart is a circular statistical graphic that is divided into slices to illustrate numerical proportions. Each slice represents a proportionate part of the whole, and the total sum of all slices forms a complete circle, representing 100%. Pie charts are commonly used to display the distribution of categories in a dataset and emphasize the relationship of each part to the whole.



Pie charts usually represent the data where the ***values can be added together or with only one data series (all the data points are positive).*** For instance, in the above example, we have created a pie chart to represent the expenditure list of an xyz company. The company has 5 different attributes, and therefore the pie chart shows 5 divisions (also known as slices of a Pie).

The ***larger the contribution of an attribute, the larger will be the size of the slice of the pie chart.*** A pie chart can be easily interpreted if data points are less (around 7). With more data points, you will have more slices, and some of them would be very small, making the graph difficult to read/interpret.

Components of a Pie Chart:

1. **Slices:** The main components of a pie chart are the slices, each representing a portion or percentage of the whole dataset. The size of each slice is proportional to the value it represents.
2. **Labels:** Labels are used to identify and describe each slice of the pie. They are often placed outside the pie chart, near the corresponding slice, and may include both the category name and the percentage or absolute value.
3. **Title:** A title provides context for the entire pie chart, summarizing the information being presented. It is usually placed at the center or top of the chart.

When to Use a Pie Chart:

1. **Part-to-Whole Relationships:** Pie charts are effective when you want to illustrate the composition of a whole in terms of its parts. Each slice represents a category, and the entire pie represents the whole dataset.
2. **Percentage Distribution:** Use pie charts to show the percentage distribution of categories within a dataset. It provides a visual representation of how each category contributes to the total.
3. **Limited Categories:** Pie charts work best when dealing with a small number of categories (typically less than seven). Too many slices can make the chart cluttered and difficult to interpret.

Advantages of Pie Charts:

1. Easy to create: Though most Excel charts are easy to create, trust me, Pie charts are the easiest of them all. With Pie charts, you need not worry about customization and formatting of your chart as most of the time, the default settings are good enough.

2. Easy to read: Pie charts are easy to see and therefore can be read easily if you only have a few data points.

3. Management is obsessed with Pie Charts: As per survey, it is reported that managers/clients love to have Pie charts in the official presentation.

Disadvantages of Pie Charts:

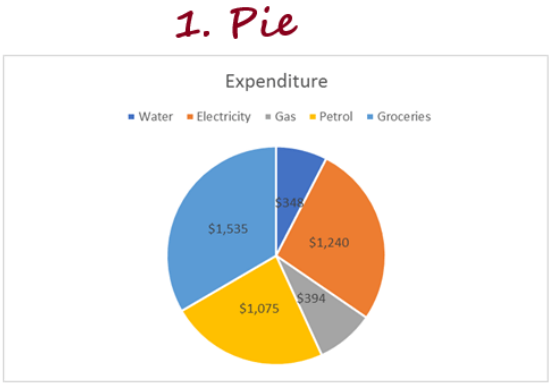
1. **Pie Charts are useful with fewer data points** and often becomes complex when adding more data points.
2. **Pie Charts can't be used to show a trend** as they are meant to present you the snap of the values at a certain point in time.
3. **Pie Charts can't show multiple types of data values.**
4. **Pie Charts the difference in data points is minor,** you may find it challenging to comprehend the pie chart visually.
5. **Pie Charts take more space and give less information.**

Types of Pie Charts:

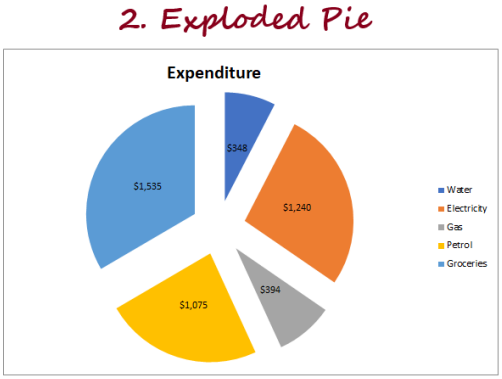
1. 2D Chart:

2D or 2-Dimensional charts are further divided into 3 types which are as follows:

- **Pie:** The other 2D pie is Normal Pie. It is used to show the contribution of each point to total value.

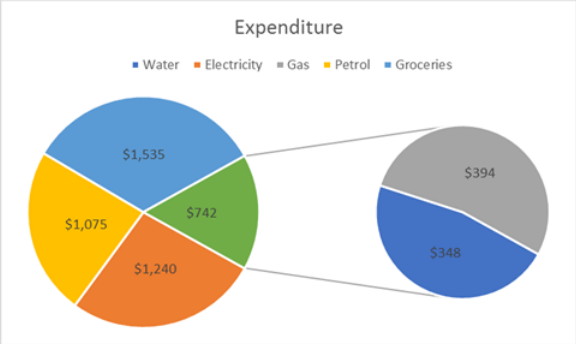


- **Exploded Pie:** Exploded Pie shows the contribution of each value to total value while emphasizing individual data values. You can also 'explode' a normal pie by first clicking on the pie and then selecting a slice and then dragging it away from the center.



- **PIE of PIE Chart**: Pie of Pie chart is used to extract some values from the main pie and combine them to the second pie. Using this chart you can make small percentages more readable or can emphasize more values.

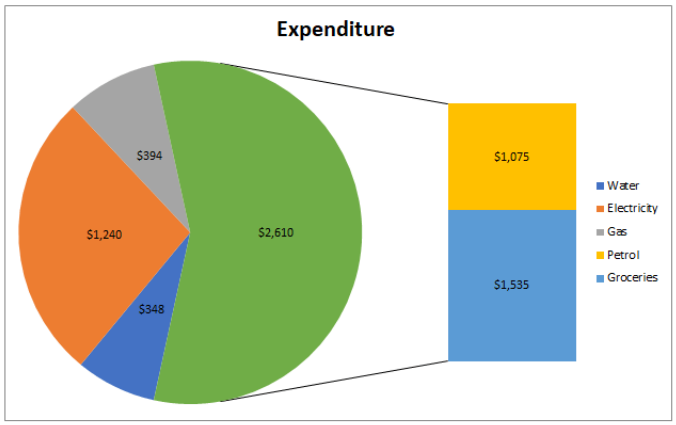
3. Pie of Pie Chart



As you can see in the above example, we have created two different pie charts where the first one is our original Pie chart, and the second pie chart represents the subset of the main pie chart.

- **Bar of PIE Chart**: Bar of Pie Chart is used to extract some values from the main pie and combine them into a stacked bar. Using this chart you can make small percentages more readable or can emphasize more values.

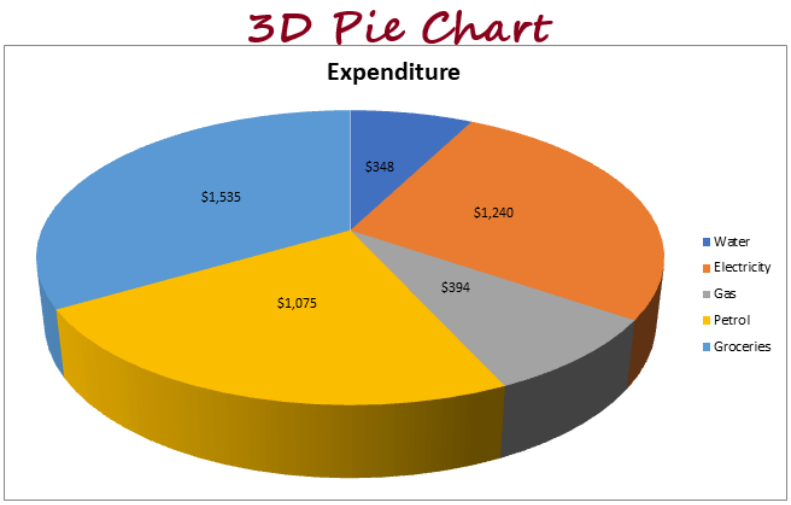
4. Bar of Pie Chart



As you can see in the example, the Bar of Pie chart is similar to Pie of the pie chart, but with the only difference that instead of a sub pie, a sub bar will be created.

2. 3D PIE Chart:

There is not much of a difference between 2D and 3D pie charts. Both are almost the same with a visual difference. That's because 3D charts have depth as well in addition to length and breadth (where 2D charts only have length and breadth).



Things to Remember:

1. You ***cannot use negative data points in Pie charts*** as they treat them the same as positive values and do not show any difference in the chart. Therefore, it dilutes the output causing unnecessary confusion.
2. ***Pie charts work best with fewer data points.*** It is also advisable only to ***use 7 or fewer chart slices.*** Adding more than 7 divisions will make the chart look cluttered and impact the visibility.
3. It is always advised to ***use different colours for each chart division*** as it helps the users to distinguish between different pie slices quickly.
4. ***PIE charts are commonly used to represent simple and ordinal data points.***

Summary Statistics:

Summary statistics are a vital part of Exploratory Data Analysis (EDA) in data science. They provide essential insights into the characteristics of your data, helping you understand its central tendency, spread, and other key properties.

Here are some commonly used summary statistics in EDA:

For Numerical Variables:

- **Measures of central tendency:**
  - **Mean:** Average value of all data points.
  - **Median:** Middle value when data is ordered from least to greatest.
  - **Mode:** Most frequent value in the data.

- **Measures of spread:**
  - **Standard deviation:** Average distance of data points from the mean.
  - **Variance:** Square of the standard deviation.
  - **Interquartile range (IQR):** Difference between the 75th and 25th percentiles (middle 50% of data).
- **Minimum and maximum values:** Lowest and highest data points.
- **Percentiles:** Divide data into 100 equal parts, indicating the values at each percentage point (e.g., 25th percentile is the value at the 25% mark).
- **Quantile-based measures:** Measures like skewness and kurtosis reveal the asymmetry and "peakedness" of the distribution.

For Categorical Variables:

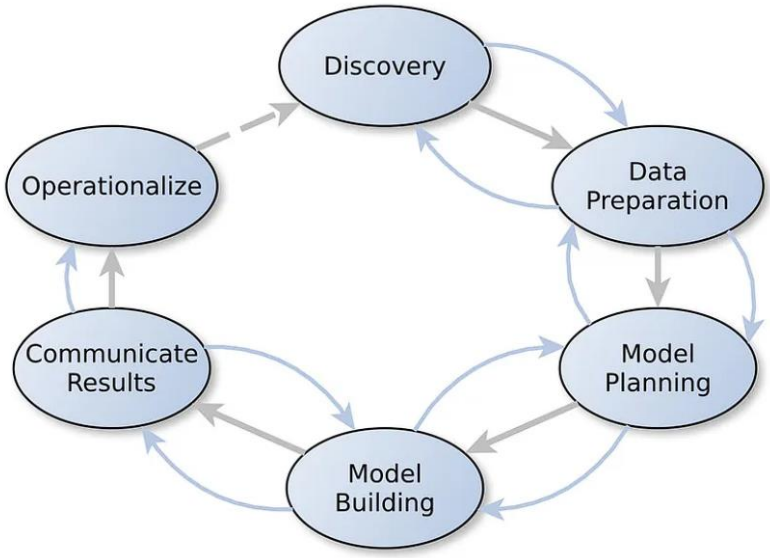
- **Frequency counts:** Number of data points in each category.
- **Proportions and percentages:** Ratio of data points in a category compared to the total.
- **Mode:** Most frequent category.
- **Cross-tabulations:** Frequencies of two variables together, revealing relationships between them.

Benefits of using summary statistics in EDA:

- **Quick overview of data:** Get a high-level understanding of the data's structure and characteristics.
- **Identify potential problems:** Detect missing values, outliers, or unexpected patterns.
- **Compare datasets:** Understand similarities and differences between different datasets.
- **Guide further analysis:** Inform decisions about which statistical tests or models to use.

Data Analytics Lifecycle:

Data Analytics Life Cycle defines the process of the how information is carried out in various phases for professionals working on a project. Its a step-by-step procedure that is arranged in a circular structure. Each phase has its own characteristics and importance.



Different Phases of Data Analytics Life Cycle

Phase 1: Discovery –

- The data science team learn and investigate the problem.
- Develop context and understanding.
- Come to know about data sources needed and available for the project.
- The team formulates initial hypothesis that can be later tested with data.

Phase 2: Data Preparation –

- Steps to explore, preprocess, and condition data prior to modeling and analysis.
- It requires the presence of an analytic sandbox, the team execute, load, and transform, to get data into the sandbox.
- Data preparation tasks are likely to be performed multiple times and not in predefined order.
- Data Collection methods that are used in this phase are:
  - **Data acquisition:** Collecting data through external sources.
  - **Data Entry:** Prepare data points through manual entry or digital systems.
  - **Signal reception:** Accumulating data from digital devices such as IoT devices and control systems.
- Several tools commonly used for this phase are – Hadoop, Alpine Miner, Open Refine, etc.

Phase 3: Model Planning –

- Team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
- In this phase, data science team develop data sets for training, testing, and production purposes.
- Team builds and executes models based on the work done in the model planning phase.
- An analytic sandbox is used to work with the data and to perform analytics throughout the project duration. Data can be loaded into the sandbox in three ways:
  - **Extract, Transform, Load (ETL)** - The data is transformed based on a set of business rules and then loaded into the sandbox.
  - **Extract, Load, Transform (ELT)** - The data is loaded into the sandbox and then transformed according to a set of business rules.
  - **Extract, Transform, Load, Transform (ETLT)** - It has two transformation levels and is a combination of ETL and ELT.
- Several tools commonly used for this phase are – Matlab, STASTICA.

Phase 4: Model Building –

- Team develops datasets for testing, training, and production purposes.
- Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
- Free or open-source tools – Rand PL/R, Octave, WEKA.
- Commercial tools – Matlab , STASTICA.

Phase 5: Communication Results –

- After executing model team need to compare outcomes of modeling to criteria established for success and failure.
- Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.
- Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.

Phase 6: Operationalize –

- The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.
- This approach enables team to learn about performance and related constraints of the model in production environment on small scale , and make adjustments before full deployment.
- The team delivers final reports, briefings, codes.
- Free or open source tools – Octave, WEKA, SQL, MADlib.

Key Benefits of Using the Data Analytics Lifecycle:

- **Structured Approach:** Provides a clear roadmap for data analysis projects, ensuring a logical and efficient flow of activities.
- **Business Alignment:** Ensures that data analytics efforts are focused on addressing specific business objectives and challenges.
- **Quality Control:** Emphasizes data preparation and cleaning to ensure the reliability and accuracy of analysis results.
- **Communication and Collaboration:** Facilitates communication and collaboration among different stakeholders involved in the data analytics process.
- **Continuous Improvement:** Encourages ongoing evaluation and refinement of the process to maximize its effectiveness and value.

Data Visualization:

Data visualization is the graphical representation of data to help people understand the patterns, trends, and insights within the data. It involves the creation of visual elements such as charts, graphs, and maps to communicate information effectively. Data visualization is a key component of data analysis and plays a crucial role in making complex data more accessible and understandable.

Here are some important aspects of data visualization:

1. Types of Visualizations:

- **Table:** A table is data displayed in rows and columns, which can be easily created in a Word document or Excel spreadsheet.
- **Chart or graph:** Information is presented in tabular form with data displayed along an x and y axis, usually with bars, points, or lines, to represent data in comparison.

An **infographic** is a special type of chart that combines visuals and words to illustrate the data.

  - **Gantt chart:** A Gantt chart is a bar chart that portrays a timeline and tasks specifically used in project management.
  - **Pie chart:** A pie chart divides data into percentages featured in "slices" of a pie, all adding up to 100%.

- **Geospatial visualization:** Data is depicted in map form with shapes and colors that illustrate the relationship between specific locations, such as a choropleth or heat map.
- **Dashboards:** Interactive displays, usually for business purposes, that consolidate and visualize multiple data points.

2. Data Representation:

- **Categorical Data:** Represented using bar charts, pie charts, or stacked bar charts.
- **Numerical Data:** Displayed through line charts, scatter plots, or bubble charts.
- **Temporal Data:** Time series data can be visualized using line charts, candlestick charts, or heatmaps.

3. Color and Design:

- **Color Coding:** Use of color to highlight and differentiate data points.
- **Contrast and Legibility:** Ensure that visual elements are easy to read and distinguish.
- **Consistency:** Maintain a consistent design to facilitate easier comprehension.

4. Interactivity:

- **Zooming and Panning:** Enables users to explore detailed data.
- **Filtering:** Allows users to focus on specific subsets of data.
- **Hover and Tooltip:** Display additional information when users interact with data points.

5. Tools and Technologies:

- **Programming Libraries:** Python libraries like Matplotlib, Seaborn, Plotly, and R libraries like ggplot2.
- **Business Intelligence Tools:** Tableau, Power BI, and QlikView.
- **Web-based Tools:** D3.js, Chart.js, and Google Charts.

6. Storytelling:

- **Narrative:** Use visualizations to tell a story or present a compelling case.
- **Annotations:** Add text or labels to provide context and explanation.

7. Best Practices:

- **Simplicity:** Avoid clutter and unnecessary complexity.
- **Accuracy:** Ensure that visualizations accurately represent the underlying data.
- **Audience Consideration:** Tailor visualizations to the target audience.

Data visualization examples:

These are a few examples of data visualization in the real world:

- **Data science:** Data scientists and researchers have access to libraries using programming languages or tools such as Python or R, which they use to understand and identify patterns in data sets. Tools help these data professionals work more efficiently by coding research with colors, plots, lines, and shapes.
- **Marketing:** Tracking data such as web traffic and social media analytics can help marketers analyze how customers find their products and whether they are early adopters or more of a laggard buyer. Charts and graphs can synthesize data for marketers and stakeholders to better understand these trends.
- **Finance:** Investors and advisors focused on buying and selling stocks, bonds, dividends, and other commodities will analyze the movement of prices over time to determine which are worth purchasing for short- or long-term periods. Line graphs help financial analysts visualize this data, toggling between months, years, and even decades.
- **Health policy:** Policymakers can use choropleth maps, which are divided by geographical area (nations, states, continents) by colors. They can, for example, use these maps to demonstrate the mortality rates of cancer or ebola in different parts of the world.

Jobs that use data visualization:

From marketing to data analytics, data visualization is a skill that can be beneficial to many industries. Building your skills in data visualization can help in the following jobs:

- **Data visualization analyst:** As a data visualization analyst (or specialist), you'd be responsible for creating and editing visual content such as maps, charts, and infographics from large data sets.
- **Data visualization engineer:** Data visualization engineers and developers are experts in both maneuvering data with SQL, as well as assisting product teams in creating user-friendly dashboards that enable storytelling.
- **Data analyst:** A data analyst collects, cleans, and interprets data sets to answer questions or solve business problems.

Benefits of data visualization:

Data visualization can be used in many contexts in nearly every field, like public policy, finance, marketing, retail, education, sports, history, and more. Here are the benefits of data visualization:

- **Storytelling:** People are drawn to colors and patterns in clothing, arts and culture, architecture, and more. Data is no different—colors and patterns allow us to visualize the story within the data.
- **Accessibility:** Information is shared in an accessible, easy-to-understand manner for a variety of audiences.
- **Visualize relationships:** It's easier to spot the relationships and patterns within a data set when the information is presented in a graph or chart.
- **Exploration:** More accessible data means more opportunities to explore, collaborate, and inform actionable decisions.

Principles of Data Visualisation:

1. Always know your Audience: Be aware of the people with which you're going to present the data with and tailor the **contents and visualisations** accordingly. People might prefer data that is: **Short, written, infographic, charts, etc**.

2. Use Annotations: Add **explanatory content** along with visuals to help the viewers to **clear any form of doubt** from the graphs presented in the report.

3. Select your charts with care: Select a chart that is best at **representing your data**. A pie chart is the most common and simple kind of chart but it can't be used in all kind of situations. You can take the help of other graphs like bar graphs, treemaps, or slope charts etc.

4. Try to avoid 3D graphs: Avoid using **3D graphs** unless you have a third variable to represent the data with. Using a 3d chart will **distort the perception** of data and they should be **avoided to prevent ambiguity.**

5. Start the column graphs with zero: The discrepancies between the data are **overemphasized** in bar and column charts that do not begin at **zero**. Consider illustrating the difference or change in values for minor changes in amounts.

6. Make labels easy to read: Whenever you're making a bar graph try to **rotate the graph charts** to make it horizontal just so that the **labels are clear**.

7. Break up complicated charts into smaller ones: By doing so you're bringing more **definition and clarity** to your content and further, it's an **effective way of visualizing data**.

8. Colour and font considerations:

- **Avoid Default colors and fonts**: Data is supposed to be **precise** and at the same time, it should be **attractive** for clearly conveying the message. For this same reason avoid using **default colors and fonts**.
- **Consider Colour blindness**: About **10%** of the population suffers from color blindness and modify your charts accordingly.

If Anyone Wants to Learn More

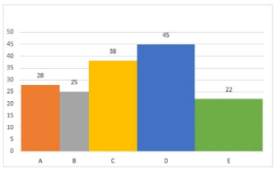
Chart Suggestions:

Charts are used as a means to present your **data visually** to enhance the user's understanding of the data. Here are a few graphs you can use to visualise your data.

Comparison:

For **comparing** between **two or more aspects of data** you can use charts like:

- **Variable Width Column Chart**
- **Table Embedded Charts**
- **Bar Charts Horizontal**
- **Bar Chart Vertical**



variable width column chart

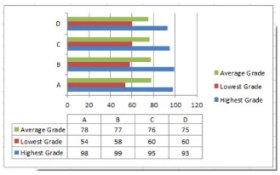
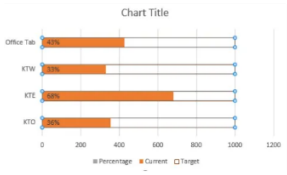
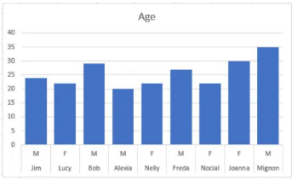


table embedded chart



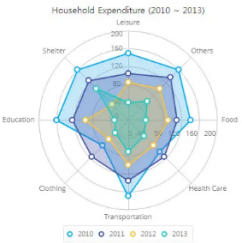
horizontal bar chart



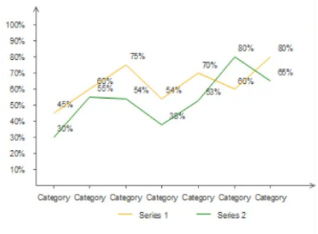
vertical bar chart

For **comparing timely** aspects of data you can get the help of charts such as:

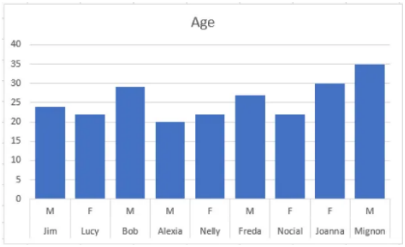
- **Circular Area Chart**
- **Line Chart**
- **Bar Chart Vertical**



Circular area chart



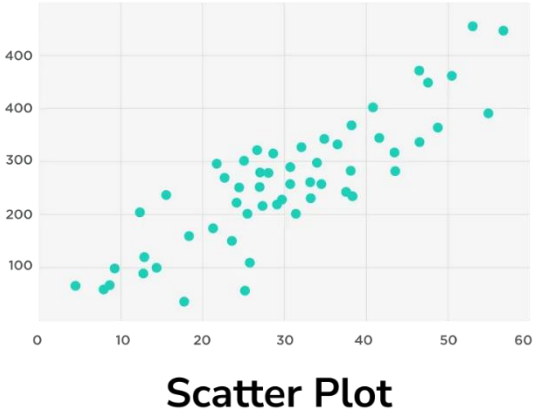
Line Chart



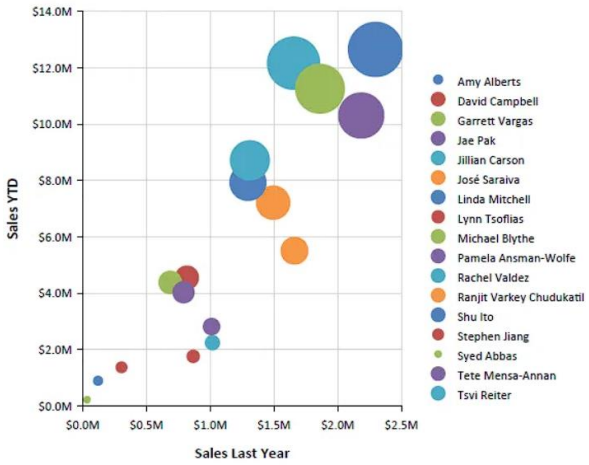
Bar chart

Relationship:

- For **comparing** between **two variables - Scatter Plot**
- For **comparing three or more** variables - **Bubble Scatter Plot**



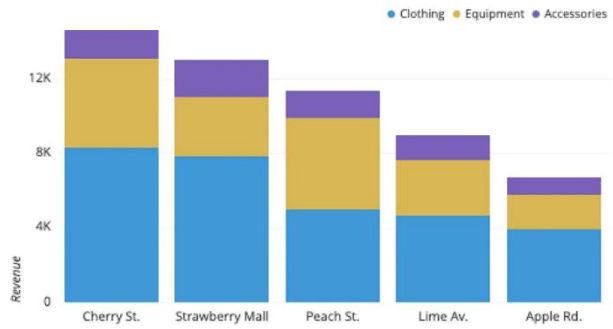
Scatter Plot



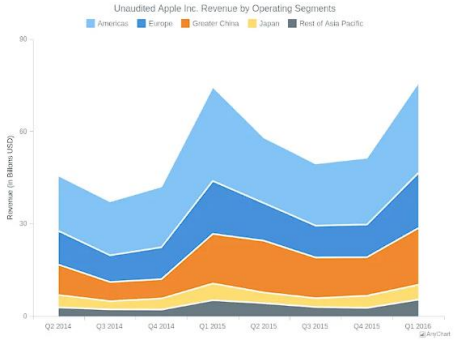
Bubble Scatter Plot

Composition:

- For **visualizing data** that **changes over time - Stacked bar chart, Stacked area chart**

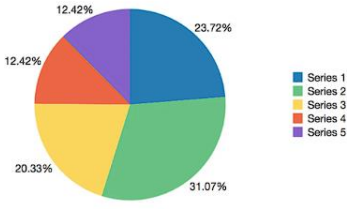


Stacked bar Chart



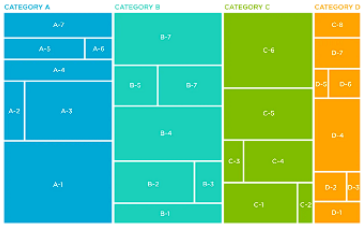
Stacked area Chart

- For **visualizing static data - Pie Chart, Waterfall Chart, Treemap**



Pie Chart

Waterfall Chart



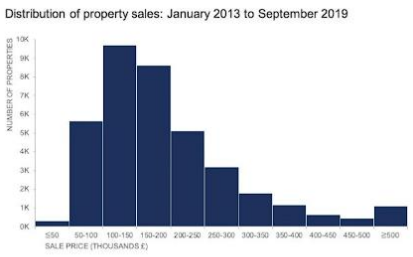
Treemap

Distribution:

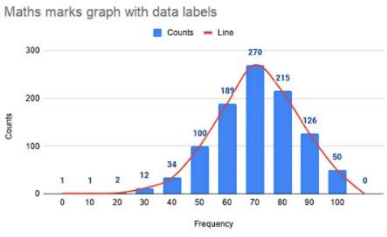
For **visualizing single variable - Bar Histogram**

For **visualizing multiple variables - Line Histogram**

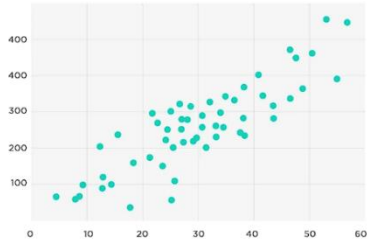
For **visualizing two variables - Scatter Plot**



Bar Histogram



Line Histogram



Scatter Plot