Chapter #4

“Data Analysis”

8 Hours | 9 Marks

Compiled By
Er. Sushil Dyopala

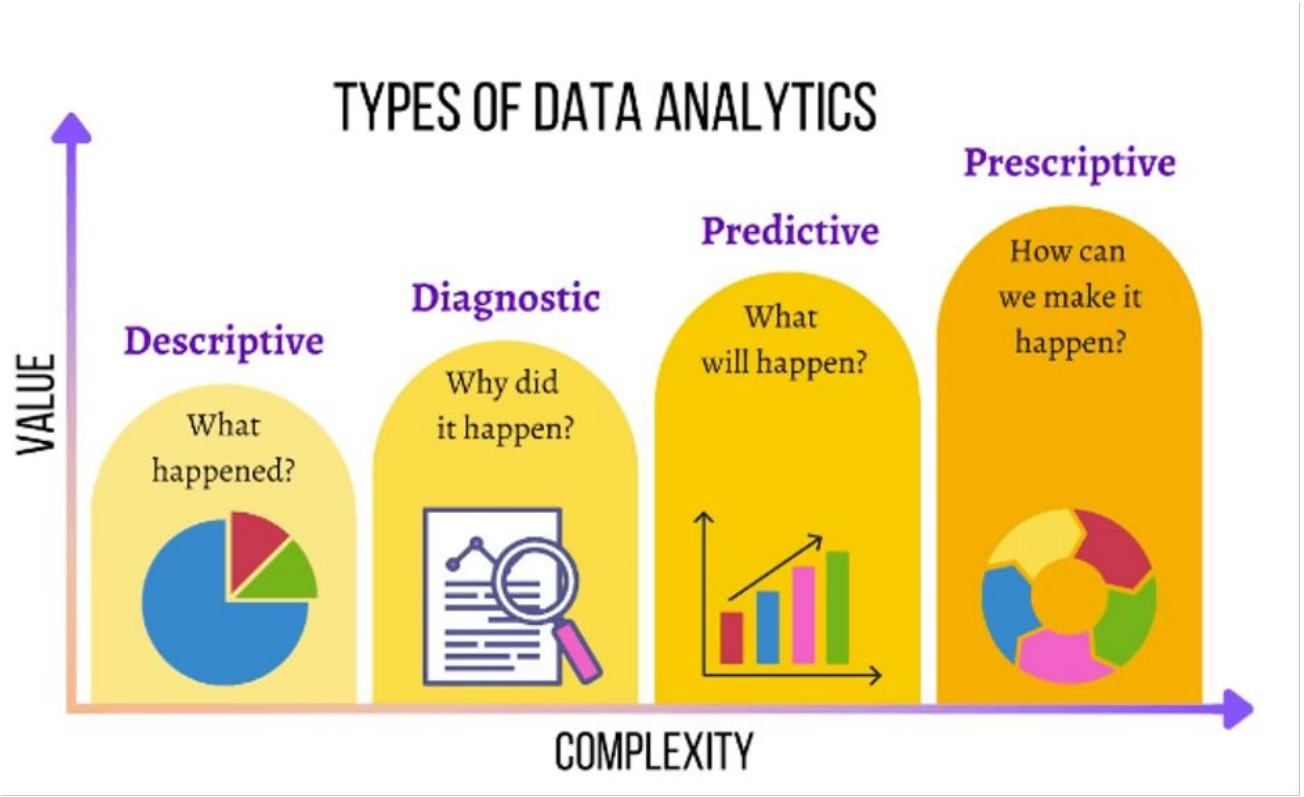
Topics

1. Data analytics: Descriptive, diagnostic, predictive and prescriptive analytics
2. Exploratory data analysis using descriptive statistics
3. Data visualization
4. Data visualization techniques
5. Principles of effective data visualization
6. Feature engineering and other aspects of data manipulation

1.1 Data Analytics

- Data Analytics is the process of examining, transforming, and interpreting raw data to uncover meaningful insights, patterns, and trends that support decision-making and strategic planning on business or research.
- It involves applying statistical, computational, and domain-specific techniques to derive actionable insights from data
- It provides data-driven or evidence-based insights, reducing reliance on intuition or guesswork
- dentifies bottlenecks, potential risks, optimizes business processes, improving productivity and efficiency.

1.1 Types of Data Analytics



1.1.1 Types of Data Analytics: Descriptive Analytics

- Descriptive Analytics: "What Happened?"
- Descriptive Analytics is the analysis of historical data to summarize and describe past or real-time events or trends.
- It does not explain the reasons behind the trends, but it provides a clear picture of past events.
- Key Features:
  - Answers the Question: "What happened or happening?"
  - Summarizes data with metrics like averages, growth rates, and proportions.
  - Highlights historical trends and anomalies.
  - Provides an aggregate view through dashboards and reports.

1.1.1 Types of Data Analytics: Descriptive Analytics

- Techniques:
  - Data Aggregation: Summarizes large datasets (e.g., SUM, AVG, COUNT).
  - Descriptive Statistics: Measures central tendencies (mean, median) and variability (standard deviation).
  - Data Visualization: Converts data into visuals (bar charts, heatmaps) for better interpretation.
  - Trend Analysis: Tracks changes over time using rolling averages or growth rates.
  - Clustering: Groups similar data points to uncover patterns (e.g., k-Means, DBSCAN).
  - Anomaly Detection: Identifies outliers in historical data using Z-scores or IQR.
  - OLAP (Online Analytical Processing): Allows multidimensional analysis of data cubes.
  - Frequency Analysis: Examines distribution or occurrence of values in a dataset.
- Example:
  - Social Media: Instagram analytics show the number of likes, shares, and comments on posts.
  - Healthcare: A hospital tracks patient admissions over the past year to identify seasonal trends.

1.1.2 Types of Data Analytics: Diagnostic Analytics

- Diagnostic Analytics: "Why Did It Happen?"
- Diagnostic analytics goes a step further than descriptive analytics by identifying the causes of trends and anomalies.
- It focuses on finding relationships and root causes using historical data.
- Key Features:
  - Answers the Question: "Why did it happen?"
  - Identifies causation and correlations between variables.
  - Investigates anomalies and unexpected trends.
  - Supports troubleshooting and scenario analysis.

1.1.2 Types of Data Analytics: Diagnostic Analytics

- Techniques
  - Drill-Down Analysis: Explores data at finer levels to find root causes.
  - Correlation Analysis: Measures relationships between variables (e.g., Pearson correlation).
  - Hypothesis Testing: Tests assumptions using t-tests, ANOVA, or Chi-square tests.
  - Regression Analysis: Identifies relationships between dependent and independent variables.
  - Time-Series Decomposition: Breaks down time-series data into trend, seasonal, and residual components.
  - Comparative Analysis: Compares metrics across groups or timeframes (e.g. pre/post-analysis).
  - Text and Sentiment Analysis: Analyzes text data to find insights (e.g., NLP techniques).
- Example:
  - Business: If sales declined in a particular month, diagnostic analytics would investigate why (e.g., increased competition, poor marketing).
  - Analyzing reasons behind a sudden drop in product sales.

1.1.3 Types of Data Analytics: Predictive Analytics

- Predictive Analytics: "What Will Happen?"
- Predictive Analytics uses historical data and statistical models to forecast future events or behaviors, enabling organizations to make proactive decisions.
- It helps in making proactive decisions based on probabilities.
- Key Features:
  - Answers the Question: "What will happen?"
  - Focuses on probability-based forecasts and trends.
  - Uses machine learning and statistical models for predictions.
  - Helps identify risks, opportunities, and uncertainties.

1.1.3 Types of Data Analytics: Predictive Analytics

- Techniques
  - Regression Analysis: Predicts outcomes using models like linear, logistic, and ridge regression.
  - Time-Series Forecasting: Predicts trends using ARIMA, SARIMA, or exponential smoothing.
  - Machine Learning: Builds predictive models (e.g., Random Forest, Gradient Boosting).
  - Neural Networks: Uses deep learning models like LSTMs for complex predictions.
  - Classification Models: Predicts categorical outcomes (e.g., Naive Bayes, Decision Trees).
  - Clustering: Identifies behavioral patterns for prediction (e.g., k-Means).
  - Anomaly Detection Models: Detects unusual patterns (e.g., Isolation Forest).
  - Market Basket Analysis: Predicts buying behaviors (e.g., Apriori).
- Example:
  - Healthcare: A hospital predicts which patients are likely to develop chronic diseases based on their medical history.
  - Weather Forecasting – Predicting rain, storms, and temperature.

1.1.4 Types of Data Analytics: Prescriptive Analytics

- Prescriptive Analytics: "What Should Be Done?"
- Prescriptive Analytics integrates predictions with optimization techniques to recommend the best course of action to achieve desired outcomes.
- Prescriptive analytics is the most advanced form of analytics.
- IKey Features:
  - Answers the Question: "What should we do?"
  - Combines optimization, simulation, and AI techniques.
  - Provides actionable strategies and decision recommendations.
  - Focuses on maximizing efficiency and minimizing risks.

1.1.4 Types of Data Analytics: Prescriptive Analytics

- Techniques
  - Optimization Models: Finds the best solution for problems (e.g., Linear Programming).
  - Simulation Modeling: Tests scenarios (e.g., Monte Carlo Simulations).
  - Heuristics and Metaheuristics: Solves complex optimization problems (e.g., Genetic
  - Algorithms).
  - Game Theory: Analyzes competitive decision-making scenarios.
  - Reinforcement Learning: Trains systems to make decisions (e.g., Q-Learning).
  - Decision Trees: Recommends actions by mapping out possible outcomes.
- Example:
  - Business: If sales are predicted to drop, prescriptive analytics suggests solutions like adjusting pricing or launching promotions.
  - Supply Chain Management – Optimizing logistics and inventory.

2.1 Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is a process of examining datasets to summarize their main characteristics, identify patterns, detect anomalies, and validate assumptions using visual and statistical techniques.
- It forms a critical step before applying advanced analytical methods and modeling, ensuring data readiness and reliability.
- Exploratory Data Analysis (EDA) is the first step in data analysis, where we examine and summarize a dataset before applying machine learning models or making decisions.
- Key Features of EDA
    - Summarization of Data: Provides an overview of central tendency, dispersion, and distribution.
    - Pattern Identification: Unveils trends and relationships within the data.
    - Anomaly Detection: Highlights outliers and inconsistencies.
    - Hypothesis Formation: Facilitates initial understanding to frame research questions.
    - Data Cleaning Insight: Identifies missing values, duplicates, and erroneous entries for rectification.

2.1 Exploratory Data Analysis

- Importance of Exploratory Data Analysis (EDA)
  - Ensures data quality by identifying and addressing missing, duplicate, or erroneous data.
  - Reveals key insights and patterns that inform decision-making and highlight data trends.
  - Prepares data for advanced analytics, including predictive and prescriptive models, by transforming and structuring it.
  - Validates assumptions about data distributions and relationships to ensure reliable model development.
  - Detects outliers and anomalies that could skew results or indicate underlying data issues.
- Steps in EDA: It is a systematic process to explore and understand data.
  - Data Inspection: Understand data structure, size, and types of variables.
  - Data Cleaning: Handle missing, inconsistent, or erroneous data entries.
  - Data Transformation: Normalize, scale, or encode data for further analysis.
  - Visualization: Create plots to identify patterns, trends, and correlations.
  - Statistical Summaries: Calculate measures like mean, median, variance, and correlation coefficients.

2.2 Descriptive Statistics for EDA

1. Measures of Central Tendency
    - Measures of central tendency describe the center of a dataset.
    - The three main measures are:
        i. Mean: The arithmetic average, representing overall data behavior but sensitive to outliers.
        ii. Median: The middle value in sorted data, useful for skewed distributions.
        iii. Mode: The most frequently occurring value, often applied to categorical data.
2. Measures of Dispersion
    - Measures of dispersion describe how spread out the data is.
    - Range: Difference between the highest and lowest values.
    - Variance: Measure of how far data points are from the mean.
    - Standard Deviation: Square root of variance, representing data spread.
    - Interquartile Range (IQR): The difference between the third quartile (Q3) and the first quartile (Q1), used to detect outliers.

2.2 Descriptive Statistics for EDA

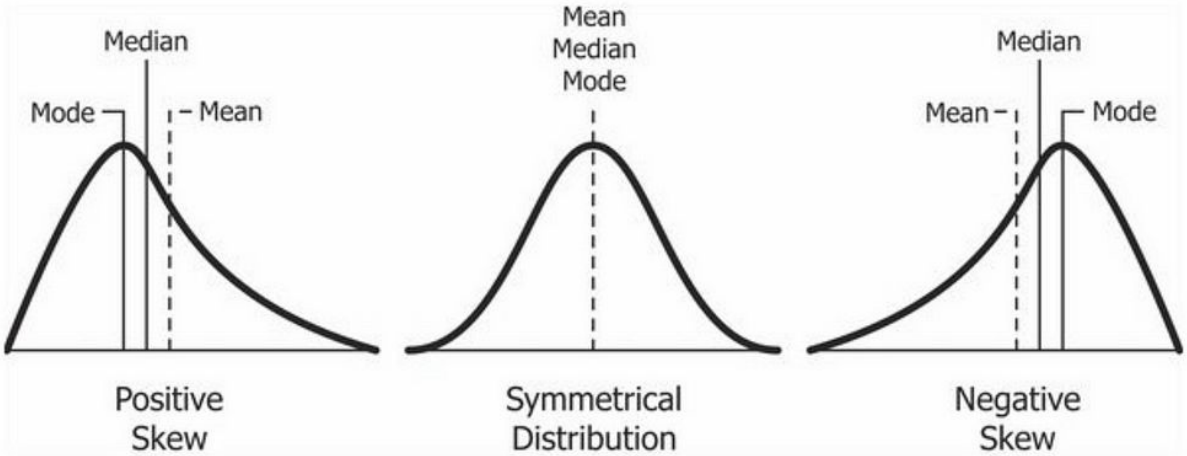
3. Data Distribution and Histograms

- Data distribution tells how values are spread in a dataset.
- A histogram is a bar graph showing data distribution.
- Types of Distributions
  - Normal Distribution (Bell Curve): Data is symmetrically distributed around the mean.
  - Skewed Distribution:
    - Right-skewed (positive skew): Tail on the right
    - Left-skewed (negative skew): Tail on the left
    - Bimodal Distribution: Two peaks in the histogram (e.g., data with two distinct groups).

2.2 Descriptive Statistics for EDA

3. Data Distribution and Histograms

Note: A histogram is used to visualize the distribution of continuous numerical data, where bars are grouped into ranges and touch each other, while a bar graph compares discrete categories with distinct bars separated by spaces.



2.2 Descriptive Statistics for EDA

4. Identifying Outliers and Missing Values

- Identifying Outliers
  - Outliers are extreme values that lie far from most of the data points.
  - Commonly used outlier detection methods:
    - Z-score method: If $|Z|>3$, it is an outlier.
    - Interquartile Range (IQR) Method: Data points below $Q1-1.5\times IQR$ or above $Q3+1.5\times IQR$ are outliers.
- Handling Missing Values
  - Null or incomplete entries in a dataset are missing values.
  - Handling Techniques:
    - Remove rows if missing values are minimal.
    - Impute using Mean/Median/Mode.

2.2 Descriptive Statistics for EDA

5. Basic Data Summaries and Visualizations

- Basic summaries and visualizations are the cornerstone of EDA, offering quick insights.
- Summaries:
  - Descriptive Statistics: Mean, median, standard deviation, and count.
  - Describe() function in Python gives count, mean, std, min, quartiles, and max.
  - Groupby() function helps summarize data by categories.
- Data Visualizations
  - Bar Charts – Compare categorical variables.
  - Histograms – Show frequency distribution.
  - Boxplots – Show median, quartiles, and outliers.
  - Scatter Plots – Show relationships between two variables.

2.2 Descriptive Statistics for EDA

6. Correlation Analysis and Data Relationships

- Correlation measures the relationship between two numerical variables.
- Correlation Coefficient (r) measures the strength and direction of linear relationships.
  - Positive Correlation (r>0) – Both variables increase together (e.g., height and weight).
  - Negative Correlation (r<0) – One variable increases, the other decreases (e.g., temperature and sweater sales).
  - No Correlation (r=0) – No relationship.
- Pearson Correlation Coefficient(r):

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

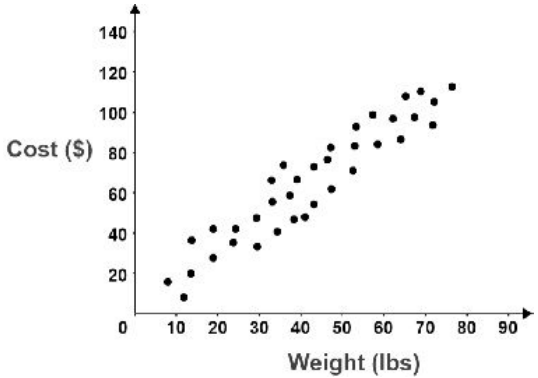
y_i = values of the y-variable in a sample

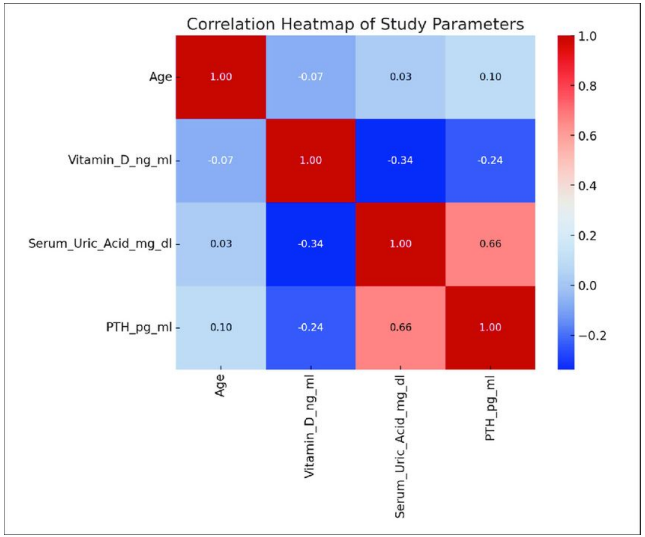
$\bar{y}$ = mean of the values of the y-variable

2.2 Descriptive Statistics for EDA

6. Correlation Analysis and Data Relationships

- Pearson Correlation: For linear relationships between continuous variables.
- Spearman Correlation: For non-linear monotonic relationships or ordinal data.
- Visualizing Correlation
  - Heatmaps – Show correlation matrix.
  - Scatter Plots – Show relationships visually.





3.1 Data Visualization

- Data visualization is the graphical representation of information and data.
- It transforms complex datasets into visual formats like charts, graphs, and maps, enabling users to identify patterns, trends, and insights quickly and effectively.
- The main objectives of visualization include:
  - Exploration of Data – Helps analysts explore data to identify hidden patterns.
  - Presentation of Findings – Communicates data-driven insights to stakeholders clearly.
  - Comparison of Data Points – Helps compare various datasets and track performance over time.
  - Summarization of Large Data Sets – Reduces complexity and makes large datasets understandable.
  - Decision Support – Assists businesses and researchers in making informed decisions.
  - Enhancement of User Experience – Makes reports and dashboards interactive and user-friendly.

3.1 Importance of Data Visualization

- Better Understanding of Data – Visualization helps in identifying patterns, trends, and outliers that may not be apparent in raw data.
- Efficient Decision-Making – By presenting data graphically, stakeholders can make data-driven decisions quickly and accurately.
- Improved Communication – Data visualizations simplify complex information, making it easier to share insights with a broad audience, including non-technical users.
- Identifying Relationships and Trends – Through visual representation, correlations and trends between different variables become clearer.
- Enhanced Data Storytelling – Visualization helps in narrating a compelling story behind data, making it more engaging and informative.
- Detecting Errors and Anomalies – Unusual patterns and outliers in data can be easily spotted using visual techniques.

3.2 Types Data Visualization

1. Exploratory Data Visualization
    - Exploratory data visualization is used to analyze and understand data, identify patterns, trends, correlations, and outliers before drawing conclusions.
    - It is mostly used during the data analysis phase and is typically interactive.
    - Characteristics:
        - Used for data analysis and discovery.
        - Often performed by data scientists, analysts, and researchers.
        - Helps in uncovering hidden insights within raw data.
        - Interactive dashboards or tools are often used (e.g., Tableau, Power BI, Matplotlib, Seaborn, Plotly).
        - Multiple visualizations may be created before deciding on the best one.

3.2 Types Data Visualization

1. Exploratory Data Visualization
    - Common Charts Used:
        - Scatter Plots: Exploring relationships between two continuous variables.
        - Histograms: Analyzing data distribution and frequencies.
        - Heatmaps: Highlighting intensity or correlation between variables.
        - Box Plots: Identifying outliers and understanding data spread.
        - Bubble Charts: Extending scatter plots with an additional data dimension (size).
        - Parallel Coordinates: Visualizing multi-dimensional data.
        - Geospatial Visualizations (Maps): Exploring data based on geography.
    - Example Use Cases:
        - Identifying anomalies or fraud detection in financial transactions.
        - Understanding correlations between different variables in a dataset.

3.2 Types Data Visualization

2. Explanatory Data Visualization

- Explanatory data visualization is used to communicate insights clearly and effectively.
- It is designed to tell a story using data, making it easy for decision-makers or a general audience to understand findings.
- Characteristics:
  - Focused on presenting key insights in a clear and engaging way.
  - Used in reports, business presentations, and storytelling.
  - Requires simplified and polished visual elements.
  - Often static but can also be interactive.
  - Aimed at a non-technical audience (e.g., executives, stakeholders).

3.2 Types Data Visualization

2. Explanatory Data Visualization

- Common Charts Used:
  - Line Charts (for trends over time)
  - Bar Charts (for comparisons)
  - Pie Charts (for proportions)
  - Bullet Charts (for goal vs. actual performance)
  - Infographics (to summarize findings visually)
- Example Use Cases:
  - A business report showing yearly revenue trends.
  - A COVID-19 dashboard highlighting daily cases and recoveries.
  - A presentation to investors explaining market growth.

3.2 Types Data Visualization

Aspect	Exploratory Visualization	Explanatory Visualization
Purpose	Focused on understanding the data, uncovering patterns, and forming hypotheses.	Designed to communicate specific insights or findings clearly and effectively.
Approach	Involves creating multiple, rough visuals (e.g., scatter plots, box plots) to explore data relationships.	Uses polished visuals (e.g., bar charts, dashboards) that are tailored to convey a clear narrative or message.
Stage	Conducted at the early stages of data analysis to investigate and explore.	Conducted after exploratory analysis to present the key findings effectively.
Complexity	Often complex and requires analytical knowledge to interpret patterns and insights.	Simplified for clarity, making it easy for non-technical audiences to understand.

3.3 Infographics and Visualization

- Infographics
  - Infographics combine text, images, and data to communicate information in a visually engaging way.
  - They are designed for storytelling and easy comprehension, often used in marketing, education, and media.
  - Characteristics:
    - Highly stylized with icons, illustrations, and structured layouts.
    - Includes both qualitative and quantitative data.
    - Designed for a broad audience (e.g., social media, blogs, news articles).
    - Focuses on making complex topics simple and visually appealing.
  - Example Use Cases:
    - A visual resume summarizing a job candidate’s skills and experience.
    - A social media post explaining “10 Benefits of Data Science.”

3.3 Infographics and Visualization

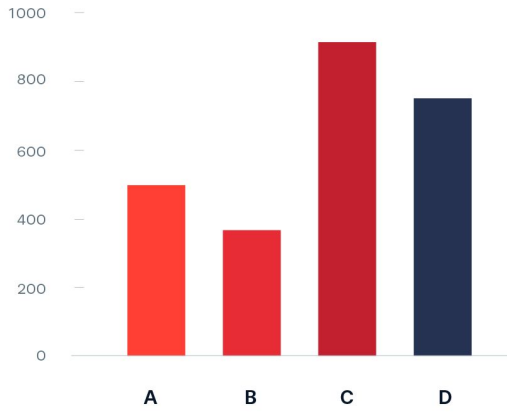
- Visualization
  - Data visualization refers to graphical representation of structured data to identify trends, patterns, and relationships.
  - Unlike infographics, it focuses on accurate representation of data rather than heavy styling.
  - Characteristics:
    - Data-focused and often interactive.
    - Used for analysis, reporting, and decision-making.
    - Based on structured data from databases or spreadsheets.
    - Found in dashboards, reports, and scientific research.
  - Example Use Cases:
    - A real-time dashboard showing stock market trends.
    - A heatmap visualizing population density.

3.3 Infographics and Visualization

Feature	Infographics	Data Visualization
Purpose	Storytelling & simplifying info	Analyzing & presenting data
Design	Highly stylized, includes icons & text	Focused on charts & graphs
Data Type	Can include both structured & unstructured data	Primarily structured data
Audience	General public, social media users	Analysts, decision-makers
Tools Used	Canva, Adobe Illustrator	Tableau, Excel, Python (Matplotlib, Seaborn)
Example	An image showing "The History of AI"	A bar chart of "AI Research Funding Over Time"

4.1 Data Visualization Techniques

- Data visualization techniques use charts, plots, and dashboards to represent data in a clear, insightful manner, making it easier to identify patterns, trends, and relationships.
- **Charts**
  - A chart is a graphical representation of data to simplify understanding and identify patterns, trends, or insights.
  - Typically summaries trends, comparisons, or summaries for non-technical audiences in presentations, reports, or general communication.
  - Key Characteristics:
    - Focus on simplicity and clarity for quick interpretation.
    - Effective for conveying high-level insights to broad audiences.
  - Focus: Charts target non-technical users
  - Examples: Bar, Line, Pie Charts



4.1 Data Visualization Techniques

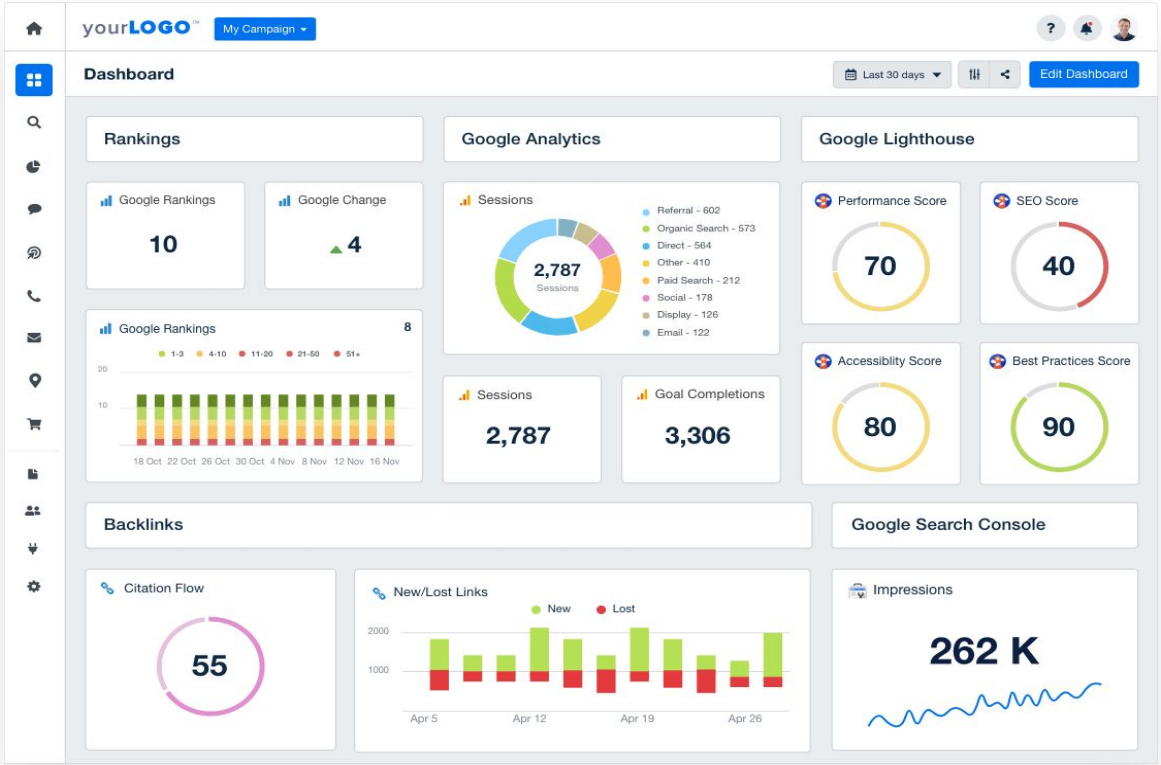
- **Plots**
  - Plots are specialized visualizations used in technical and scientific contexts to explore detailed data patterns, distributions, or relationships, aiding analysts and researchers in extracting deeper insights.
  - Key Characteristics:
    - Focus on depth and precision in representing data.
    - Suitable for multivariate, distributional, or correlation analysis.
    - Common in technical reports, research papers, or exploratory data analysis.
  - Focus: Plots are suitable analysts and researchers
  - Examples: Scatter, Box Plot, Heatmap

4.1 Data Visualization Techniques

- **Dashboards**
  - Dashboards are interactive platforms that consolidate multiple visualizations, metrics, and data sources into a single interface, offering real-time monitoring and decision-making tools for business or operational contexts.
  - Key Characteristics:
    - Integrates diverse charts, plots, and KPIs into one view.
    - Allows user interaction, such as filtering or drilling down for details.
    - Frequently used for strategic decision-making and operational monitoring.
  - Focus: dashboards are suitable for decision-makers.
  - Examples: Sales Dashboard, Website Analytics, Stock Monitoring Dashboard

4.1 Data Visualization Techniques

- **Dashboards**



4.2 Do It Yourself (DIY)

- Creating bar plots, histograms, line graphs, scatter plots.
- Heatmaps, box plots, Time Series plot and pie charts.

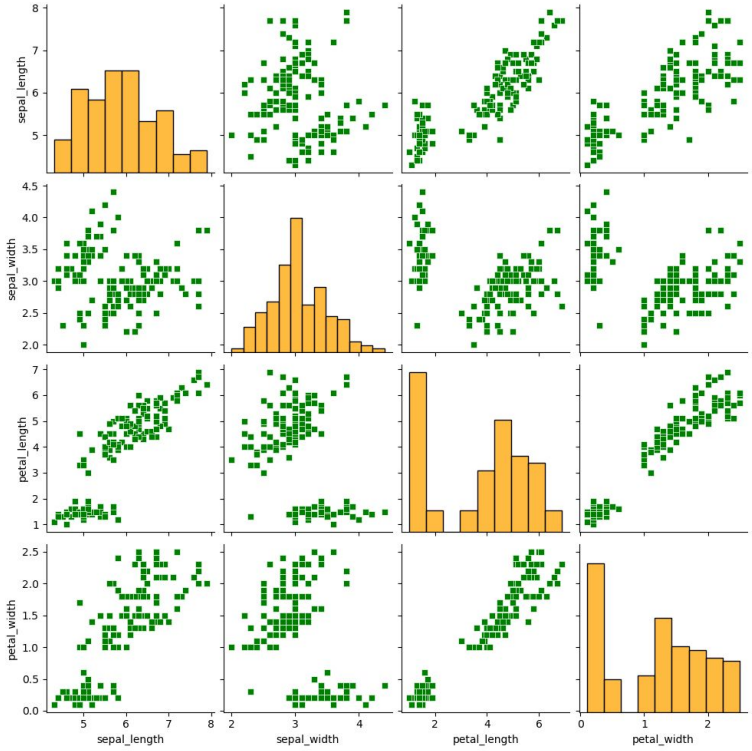
P.S. Draw using pen and paper from exam point of view and replicate it using python libraries.

4.3 Multi-Dimensional Visualizations

- Multi-dimensional visualizations are used to represent data with more than two variables, making it easier to analyze complex relationships and patterns.
- When dealing with high-dimensional data, traditional 2D charts like bar charts or scatter plots may not be sufficient.
- Multi-dimensional visualizations help reveal hidden insights by displaying multiple variables simultaneously.

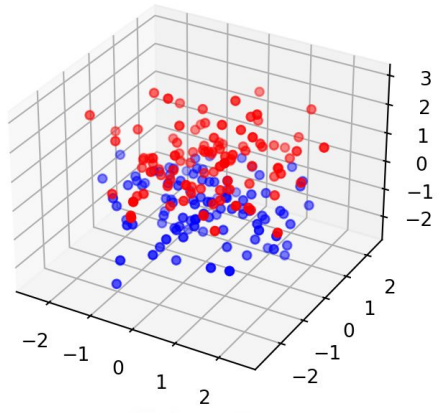
4.3 Multi-Dimensional Visualizations

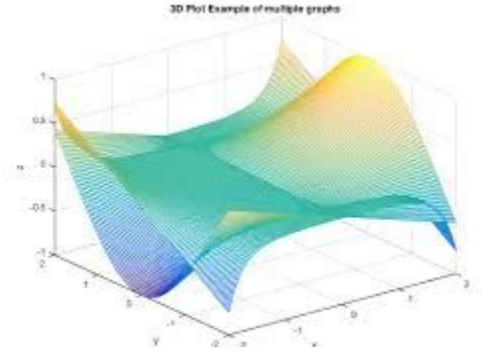
- Pair Plots
  - A pair plot (also known as a scatterplot matrix) is a collection of scatter plots arranged in a grid where each variable is plotted against every other variable.
  - Each cell in the grid represents a scatter plot between two variables.
  - The diagonal often contains histograms or KDE (Kernel Density Estimation) plots showing the distribution of individual variables.

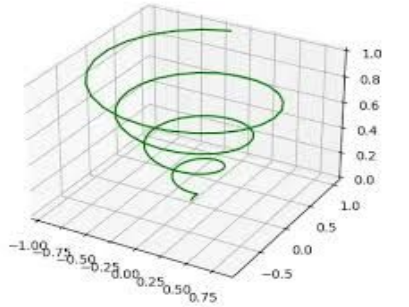


4.3 Multi-Dimensional Visualizations

- 3D Plots
  - A 3D plot extends traditional 2D scatter plots by adding a third axis (Z-axis) to visualize three variables at once.
  - Helps in visualizing data that has three continuous numerical variables.
  - Allows better insight into clustering, classification, and patterns in data.
  - Types of 3D Plots:
    - 3D Scatter Plot: Displays three numerical variables.
    - 3D Surface Plot: Represents relationships between three continuous variables.
    - 3D Line Plot: Shows trends across three dimensions







4.4 Common Visualization Tools

- Matplotlib
  - A Python library for creating static, interactive, and animated plots.
  - It provides fine control over chart appearance but requires more coding for advanced plots.
  - Customization: Fine control over plot elements.
  - Versatility: Supports static, animated, and interactive visualizations.
- Seaborn
  - Built on top of Matplotlib, Seaborn simplifies the creation of attractive, informative statistical graphics with built-in themes and color palettes.
  - Ease of Use: Simplifies complex visualizations with minimal code.
  - Aesthetics: Attractive default themes and color palettes.

4.4 Common Visualization Tools

- Tableau
  - A powerful tool for creating interactive and shareable dashboards with a user-friendly drag-and-drop interface.
  - Interactivity: Create dynamic, interactive dashboards.
  - Ease of Use: Drag-and-drop interface for quick visualizations.
- Power BI
  - A Microsoft tool for data analysis and visualization, seamlessly integrating with Microsoft services to build robust dashboards and reports.
  - Integration: Seamlessly connects with various data sources, especially Microsoft products.
  - Advanced Analytics: Built-in AI for deeper insights.
- Google Data Studio
  - A free tool for creating interactive reports and dashboards, ideal for those using Google services.
  - Integration: Connects easily with Google Analytics, Ads, and other Google products.
  - Collaboration: Enables real-time collaboration and easy sharing of reports.

4.4 Common Visualization Tools

- Tableau
  - A powerful tool for creating interactive and shareable dashboards with a user-friendly drag-and-drop interface.
  - Interactivity: Create dynamic, interactive dashboards.
  - Ease of Use: Drag-and-drop interface for quick visualizations.
- Power BI
  - A Microsoft tool for data analysis and visualization, seamlessly integrating with Microsoft services to build robust dashboards and reports.
  - Integration: Seamlessly connects with various data sources, especially Microsoft products.
  - Advanced Analytics: Built-in AI for deeper insights.
- Google Data Studio
  - A free tool for creating interactive reports and dashboards, ideal for those using Google services.
  - Integration: Connects easily with Google Analytics, Ads, and other Google products.
  - Collaboration: Enables real-time collaboration and easy sharing of reports.

4.4 Box Plot Numerical

1. Given the following dataset, construct a box and whisker plot:

   12, 18, 24, 30, 36, 42, 48, 54, 60, 66

2. Find the five-number summary for the dataset and construct a box plot:

   8, 15, 22, 25, 30, 35, 40, 45, 50, 55, 60

3. A dataset contains the following numbers:

   5, 10, 15, 20, 25, 30, 35, 40, 45, 100

If a box plot is drawn, will 100 be considered an outlier?

Note:

- While drawing box plot, **lower bound = Q1 - 1.5 * IQR** and **upper bound = Q3 + 1.5 * IQR**.
- Minimum value = max(lower bound, smallest data element) e.g. min_value = max(20, 10) = 20
- Maximum value = min(upper bound, largest data element) e.g. max_value = min(50, 100) = 50

4.5 Principles of Effective Data Visualization

- Misleading visualizations distort data insights, leading to incorrect interpretations.
- Effective visualization principles ensure clarity, accuracy, and unbiased communication.
- **Clarity**
  - Use simple, clear designs that emphasize the data's message. Avoid unnecessary decorations (chartjunk).
  - Readable Text: Use legible font sizes and clear labels for axes, legends, and titles.
  - Focus on the Data: Avoid visual distractions like excessive colors, 3D effects, or unnecessary elements.
  - Highlight Key Insights: Use emphasis, such as color or annotations, to guide the audience to the main message.

4.5 Principles of Effective Data Visualization

- **Accuracy**
  - Maintain proportionality in data representation.
  - Misleading scales, truncated axes, or skewed visuals can exaggerate trends.
  - It reflects the truthful representation of data without distortions or manipulations.
  - Proportional Representations: Ensure that visual elements like bar lengths or pie slice sizes match the data values.
  - Avoid Truncated Axes: Starting axes at non-zero values can mislead viewers by exaggerating trends.
  - Appropriate Scales: Use consistent scales to represent the data correctly.

4.5 Principles of Effective Data Visualization

- **Consistency**
  - Consistency ensures uniformity in the design to avoid misinterpretation and promote understanding.
  - Standardized Scales and Units: Use the same scales, units, and baselines across multiple charts for comparisons.
  - Unified Style: Maintain consistent colors, fonts, and chart types within a dataset.
  - Repetition for Patterns: Consistency helps establish recognizable patterns for the audience.

4.5 Principles of Effective Data Visualization

- **Simplicity**
  - Simplicity removes unnecessary complexity to make the visualization accessible to a broader audience.
  - Minimal Design: Include only elements essential to the story the data tells.
  - Avoid Overloading: Present limited data per visualization to prevent overwhelming the viewer.
  - Streamlined Layout: Use clean layouts with for better readability.

4.5 Principles of Effective Data Visualization

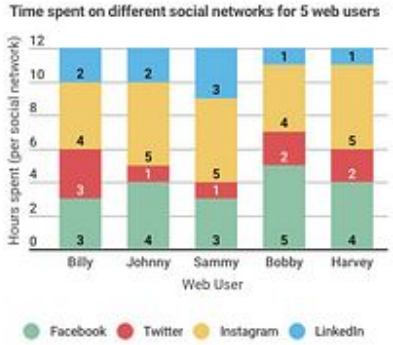
- Best Practices for Data Visualization in EDA
  - Choose the Right Chart Type: Match the visualization to the data type and analysis goal.
  - Simplify Visuals: Avoid unnecessary clutter; keep visuals focused and clean.
  - Highlight Key Insights: Use color, annotations, and labels to emphasize important patterns.
  - Ensure Accuracy: Avoid misleading scales, distorted proportions, or incomplete data.
  - Iterate: Update visuals as new insights emerge during analysis.

4.5 Principles of Effective Data Visualization

1

The Good

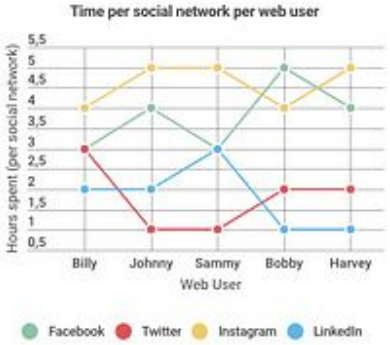
You show data clearly to make the most out of it and clearly state your point to your audience.



2

The Bad

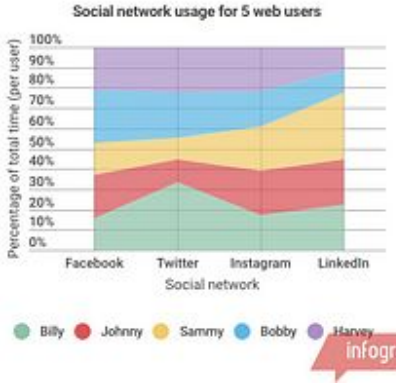
It is hard to create good representations of your data. Sometimes, it's just not right and explains nothing.



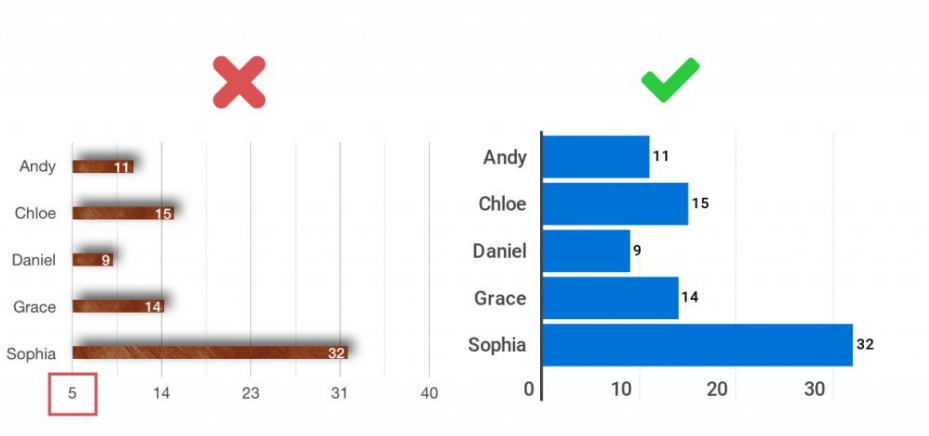
3

The Ugly

Even worse: some people intentionally play with our natural biases and use it to their advantage!

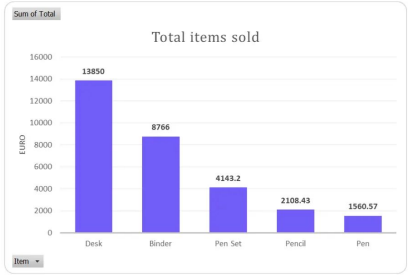


4.5 Principles of Effective Data Visualization



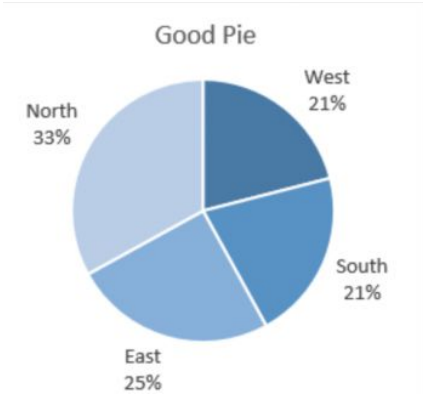
Cluttered chart

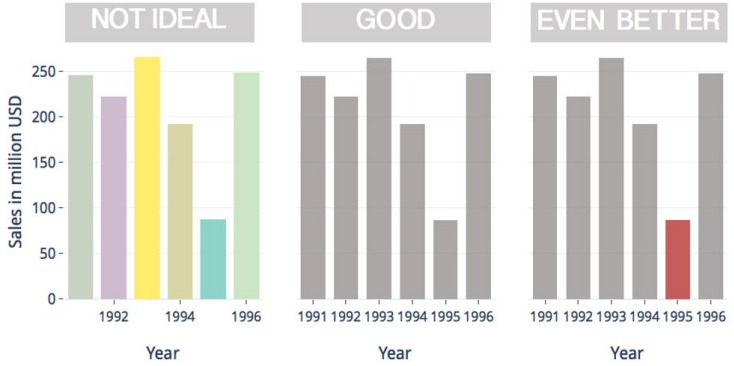
(bad data visualization example)



Clean chart

(good data visualization example)





4.6 Choosing the Right Chart for the Data

- Selecting the appropriate chart type is crucial for effectively communicating insights from data.
- The choice depends on the type of data (categorical, numerical, time series) and the purpose of visualization (comparison, distribution, relationship, composition).
- **Charts for Comparison**
  - Used when comparing categories or groups.
  - Bar Chart: Best for comparing categorical data (e.g., sales by product category).
  - Grouped/Stacked Bar Chart: Useful for sub-category comparisons within categories.
  - Column Chart: Similar to a bar chart but with vertical bars.
  - Example: Comparing monthly revenue across different products.

4.6 Choosing the Right Chart for the Data

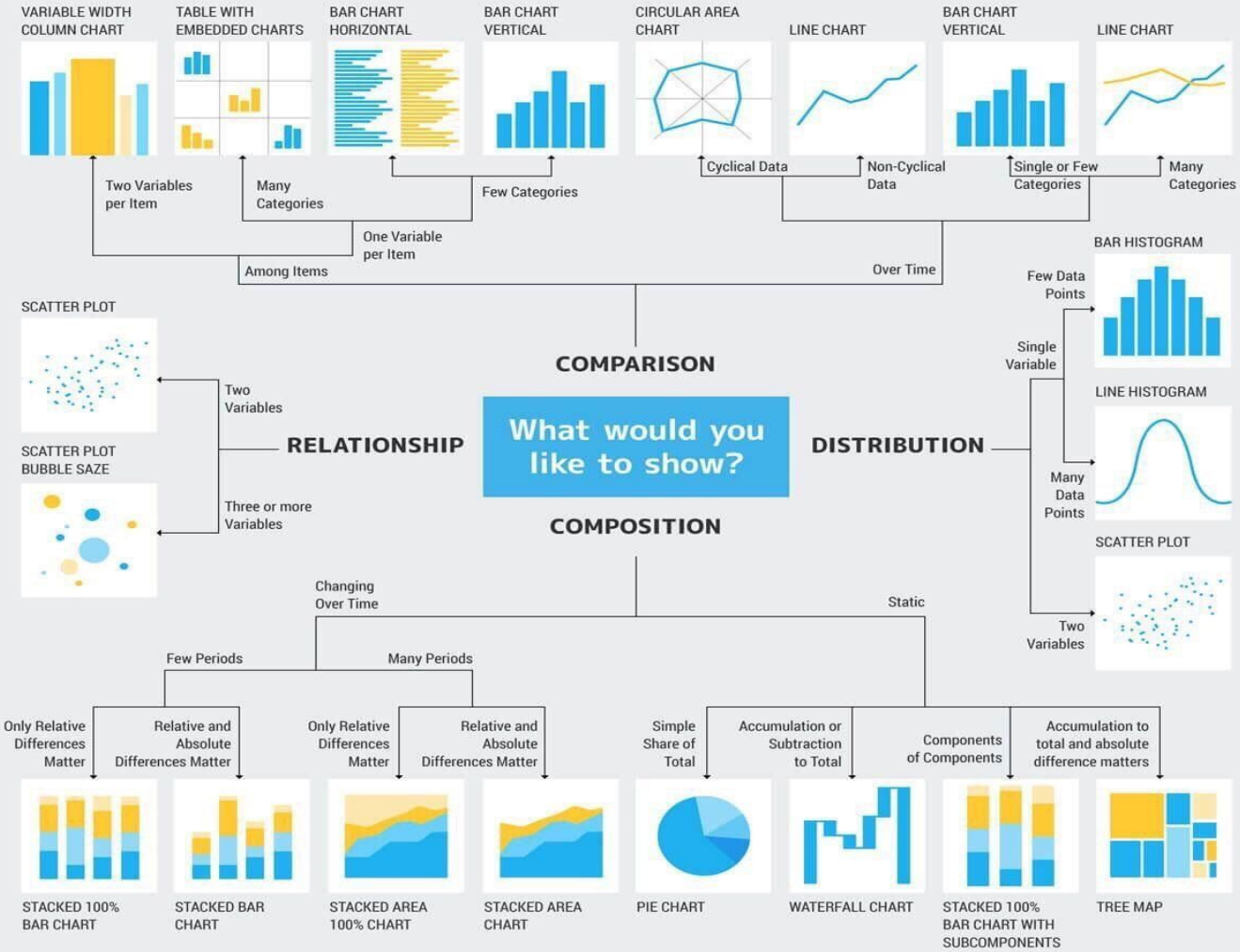
- **Charts for Showing Distribution**
  - Used to understand the spread and shape of data.
  - Histogram: Shows the frequency distribution of numerical data (e.g., age distribution of customers).
  - Box Plot: Displays median, quartiles, and outliers (e.g., salary distribution in a company).
  - Example: Understanding exam score distribution among students.
- **Charts for Showing Relationships (Correlation)**
  - Used to analyze the relationship between two or more variables.
  - Scatter Plot: Best for visualizing relationships between two continuous variables (e.g., study hours vs. exam scores).
  - Heatmap: Uses color intensity to show relationships in matrix-style data (e.g., correlation between different stock prices).
  - Example: Examining the relationship between temperature and ice cream sales.

4.6 Choosing the Right Chart for the Data

- **Charts for Showing Composition (Proportions)**
  - Used to display parts of a whole.
  - Pie Chart: Shows percentage distribution but is not recommended for too many categories.
  - Donut Chart: A variation of a pie chart with a hole in the middle.
  - Stacked Bar/Column Chart: Shows both total and individual contributions of sub-categories.
- **Charts for Time Series Data**
  - Used for data that changes over time.
  - Line Chart: Best for showing trends over time (e.g., stock prices over months).
  - Area Chart: Similar to a line chart but filled to show volume (e.g., website traffic over a year).
  - Candlestick Chart: Used in finance to show stock price movements.
  - Example: Tracking daily COVID-19 cases over several months.

4.6 Choosing the Right Chart for the Data

- **Charts for Showing Composition (Proportions)**
  - Used to display parts of a whole.
  - Pie Chart: Shows percentage distribution but is not recommended for too many categories.
  - Donut Chart: A variation of a pie chart with a hole in the middle.
  - Stacked Bar/Column Chart: Shows both total and individual contributions of sub-categories.
- **Charts for Time Series Data**
  - Used for data that changes over time.
  - Line Chart: Best for showing trends over time (e.g., stock prices over months).
  - Area Chart: Similar to a line chart but filled to show volume (e.g., website traffic over a year).
  - Candlestick Chart: Used in finance to show stock price movements.
  - Example: Tracking daily COVID-19 cases over several months.


What would you like to show?
COMPARISON
RELATIONSHIP
DISTRIBUTION
COMPOSITION
VARIABLE WIDTH COLUMN CHART
TABLE WITH EMBEDDED CHARTS
BAR CHART HORIZONTAL
BAR CHART VERTICAL
CIRCULAR AREA CHART
LINE CHART
BAR CHART VERTICAL
LINE CHART
Two Variables per Item
Many Categories
Few Categories
One Variable per Item
Among Items
Cyclical Data
Non-Cyclical Data
Single or Few Categories
Many Categories
Over Time
SCATTER PLOT
Two Variables
SCATTER PLOT BUBBLE SAZE
Three or more Variables
BAR HISTOGRAM
Few Data Points
Single Variable
LINE HISTOGRAM
Many Data Points
SCATTER PLOT
Two Variables
Changing Over Time
Static
Few Periods
Many Periods
Only Relative Differences Matter
Relative and Absolute Differences Matter
Only Relative Differences Matter
Relative and Absolute Differences Matter
Simple Share of Total
Accumulation or Subtraction to Total
Components of Components
Accumulation to total and absolute difference matters
STACKED 100% BAR CHART
STACKED BAR CHART
STACKED AREA 100% CHART
STACKED AREA CHART
PIE CHART
WATERFALL CHART
STACKED 100% BAR CHART WITH SUBCOMPONENTS
TREE MAP

5.1 Feature Engineering

- Feature engineering is the process of transforming raw data into meaningful features that better represent the problem, improving both modeling performance and analytical outcomes.
- **Importance**
  - Improves Model Performance – Well-crafted features help machine learning models capture patterns in data more effectively.
  - Reduces Overfitting – Removing irrelevant or redundant features prevents the model from memorizing noise.
  - Enhances Interpretability – Meaningful features make it easier to understand how the model makes predictions.
  - Optimizes Computational Efficiency – Reducing the number of features decreases memory usage and speeds up model training.
  - Enables Better Generalization – Good features help the model perform well on unseen data.

5.2 Feature Engineering Process

- Feature engineering consists of three main processes:
  - Extraction
  - Selection
  - Reduction
- **Feature Extraction**
  - Involves deriving useful features from raw data.
  - Also involves process of creating new features from existing data to better capture relevant information.
  - Examples:
    - Extracting the number of words in a text document.
    - Generating statistical summaries (mean, variance) from sensor data.

5.2 Feature Engineering Process

- **Feature Selection**
  - Identifies the most important features while eliminating irrelevant or redundant ones.
  - The process of identifying the most relevant features and choose n most relevant subset of features.
  - Helps in improving model performance and reducing computational complexity.
  - Example:
    - Removing highly correlated features in a dataset to avoid multicollinearity in regression models.
- **Feature Reduction**
  - Reduces the dimensionality of data while preserving essential information.
  - Helps when dealing with high-dimensional datasets (e.g., images, text).
  - Techniques:
    - Principal Component Analysis (PCA): Projects features into a lower-dimensional space.

5.3 Feature Selection Methods

- Feature selection is the process of identifying and selecting the most relevant features (variables) from a dataset to improve the performance of a machine learning model.
- The goal is to remove irrelevant, redundant, or noisy data while retaining the most important features for better model efficiency and accuracy.
- These methods can be categorized into three types:
    - Filter Methods
    - Wrapper Methods
    - Embedded Methods

5.3.1 Feature Selection Methods: Filter Methods

- Filter methods assess individual features using statistical tests before applying any machine learning model.
- Select features based on statistical properties like correlation, variance, and mutual information.
- Independent of the machine learning model.
- Steps:
  - Evaluate each feature importance using statistical techniques (e.g., correlation, Chi-square).
  - Rank the features based on evaluation scores.
  - Select top-ranked features and discard irrelevant ones.
- Examples:
  - Variance Threshold: Removes features with low variance.
  - Correlation Analysis: Eliminates features that are highly correlated with each other.

5.3.2 Feature Selection Methods: Wrapper Methods

- Wrapper methods evaluate feature subsets based on model performance, iterating through different subsets to identify the optimal combination.
- Use machine learning models to evaluate feature subsets iteratively.
- More computationally expensive than filter methods but often yield better results.
- Forward Selection:
  - Start with no features, and iteratively add the best-performing feature.
  - Evaluate the model after each feature addition.
  - Stop when no significant improvement is observed.
- Backward Elimination:
  - Start with all features, and iteratively remove the least important ones based on model performance.
  - Evaluate the model after each feature removal.
  - Stop when removing features no longer improves the model.

5.3.3 Feature Selection Methods: Embedded Methods

- Perform feature selection as part of the model training process.
- Efficient and commonly used in decision trees and regularized regression models.
- Examples:
  - Lasso Regression (L1 Regularization): Shrinks irrelevant features' coefficients to zero.
  - Tree-based Models (e.g., Random Forest, XGBoost): Rank feature importance based on decision splits.

5.4 Correlation and Causation

- **Correlation**
  - Measures the relationship between two variables.
  - Does not imply causation.
  - Example: Ice cream sales and drowning incidents are correlated (both increase in summer), but one does not cause the other.
- **Causation**
  - Indicates a direct cause-and-effect relationship between variables.
  - Requires experiments or domain knowledge to confirm.
  - Example: Smoking causes lung cancer.

5.5.1 Feature Representation in Text

- Bag of Words (BoW)
  - Converts text into a frequency-based numerical vector.
  - “Ram and Hari like Math and Science” is represented as [“Ram”:1, “and”:2, “Hari”:1, “like”:1, “Math”:1, “Science”:1]
- TF-IDF (Term Frequency - Inverse Document Frequency)
  - Weighs words based on importance across multiple documents.
  - Example:
    - Document 1: "Cats are great pets." Document 2: "Dogs are good pets." Document 3: "Birds are great pets.“
    - “good” has a higher TF-IDF score than "great" because it's rarer across documents.
- Word Embeddings:
  - Word2Vec, GloVe, FastText: Represent words in dense vector space based on semantic relationships.
  - BERT, GPT Embeddings: Contextual embeddings for capturing word meanings in different contexts.
  - Example: Cat = [0.8 0.2], Dog = [0.8 0.9], House=[0.1 0.2]

5.5.2 Feature Representation in Images

- Pixel Values
  - Represents images as matrices of pixel intensity values (RGB).
- Edge Features
  - Using techniques like Sobel filters to detect edges.
- Convolutional Neural Networks (CNNs)
  - Automatically extracts hierarchical features from images for tasks like classification.
  - Histograms of Oriented Gradients (HOG): Captures edge and shape information useful for object detection.