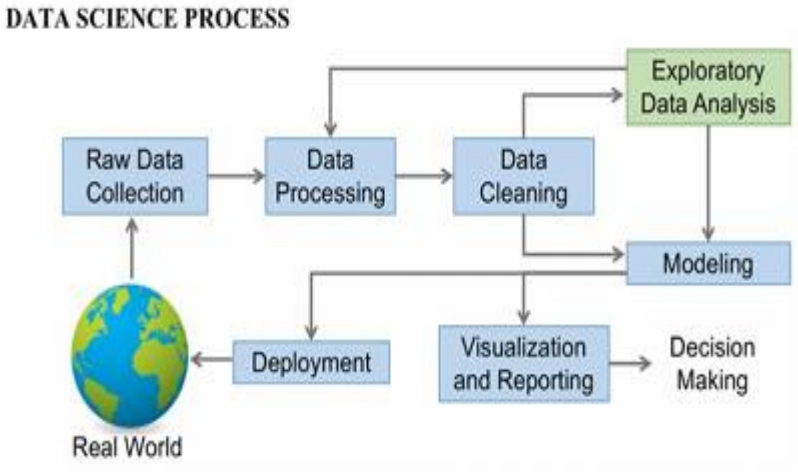
UNIT III

Data Science Process

The Data Science Process is a structured framework used to extract meaningful insights and knowledge from data. It involves a series of systematic steps starting from problem understanding to decision-making.

The key steps involved in Data Science Modelling are:



- **Step 1: Understanding the Problem** The first step involved in Data Science Modelling is understanding the problem. A Data Scientist listens for keywords and phrases when interviewing a line-of-business expert about a business challenge. The Data Scientist breaks down the problem into a procedural flow.
- **Step 2: Data Extraction** The next step in Data Science Modelling is Data Extraction. Not just any Data, but the Unstructured Data pieces you collect, relevant to the business problem you're trying to address. The Data Extraction is done from various sources online, surveys, and existing Databases.
- **Step 3: Data Cleaning** Data cleaning is the process of detecting, correcting and ensuring that your given data set is free from error, consistent and usable by identifying any errors or corruptions in the data, correcting or deleting them,

or manually processing them as needed to prevent the error from corrupting our final analysis.

- **Steps In Data Preprocessing:**
    - Gathering the data
    - Import the dataset & Libraries
    - Dealing with Missing Values
    - Divide the dataset into Dependent & Independent variable
    - dealing with Categorical values
    - Split the dataset into training and test set
    - Feature Scaling
- **Step 4: Exploratory Data Analysis** Exploratory Data Analysis (EDA) is a robust technique for familiarising yourself with Data and extracting useful insights. Data Scientists use Statistics and Visualisation tools to summarise Central Measurements and variability to perform EDA.
- **Step 5: Feature Selection** Feature Selection is the process of identifying and selecting the features that contribute the most to the prediction variable or output that you are interested in, either automatically or manually. if the features are strong enough, the Machine Learning Algorithm will give fantastic outcomes.
- **Step 6: Incorporating Machine Learning Algorithms** This is one of the most crucial processes in Data Science Modelling as the Machine Learning Algorithm aids in creating a usable Data Model. There are three types of Machine Learning methods that are incorporated:
    - **Supervised Learning :** a fundamental type of machine learning where the algorithm is trained on a labeled dataset. The training data provides

the algorithm with a basic understanding of the problem and the expected output for a given input.

    - **Unsupervised Learning :** a type of machine learning that works with unlabeled data. Instead of using predefined labels, the algorithm identifies hidden structures and patterns within the data on its own. Examples include Principal Component Analysis and clustering.

- **Step 7: Testing the Models** After building the machine learning or statistical model, the next step is to evaluate its performance using unseen/test data.
- **Key goals:**
    - Check accuracy, precision, recall, F1-score, etc.
    - Understand if the model is overfitting or underfitting

Model Evaluation Metrics

- **Accuracy:** This is the most common metric for classification problems. It is the ratio of correct predictions to the total number of predictions made. It can be calculated as

(True Positive + True Negative) / Total Sample.

- **Precision:** This measures the fraction of true positives among all positive predictions made by the model. It is calculated as

True Positives / (True Positives + False Positives).

- **Recall:** This measures the fraction of actual positive predictions that were correctly identified by the model. It is calculated as

True Positives / (True Positives + False Negatives).

- **F1-score:** This metric incorporates both precision and recall to provide a single score for a model's performance. It ranges from 0 to 1, where a higher score indicates better model performance.

- **Overfitting:** This occurs when a model learns the training data too well, capturing noise and random fluctuations rather than the underlying pattern. An overfitted model performs very well on the training data but poorly on new, unseen data.
- **Underfitting:** This happens when a model is too simple to capture the underlying patterns in the training data. An underfitted model performs poorly on both the training and test data.

- **Step 8: Deploying the Model** Once a model is tested and finalized, it is deployed into a production environment where it can make predictions on new, real-world data.

Data Visualization

- Data visualization is the graphical representation of data and information using visual elements like charts, graphs, and maps to help people understand complex datasets and identify trends, outliers, and patterns.
- It is a powerful way to tell a story with data by transforming spreadsheets of numbers into visuals that are easy to interpret.
- The process of data visualization is part art and part science, with the goal of communicating data clearly and effectively to an audience.

The Process of Data Visualization

- **Data Collection and Preparation:** This initial stage involves gathering relevant data from various sources like databases, surveys, or APIs. It is crucial to ensure the data is accurate, complete, and aligned with the visualization goals.
- **Data Cleaning and Transformation:** This step focuses on handling missing values, outliers, and preparing the data into a format suitable for visualization tools.
- **Choosing Visualization Types:** Selecting the appropriate charts, graphs, or maps based on the nature of the data is a critical step. For example, a line chart is good for displaying trends over time, while a bar chart is effective for comparing quantities across categories.
- **Designing and Creating Visuals:** This involves using data visualization tools to design the visuals. It includes choosing color schemes, labels, and other visual elements to ensure clarity and easy interpretation.

- **Interpretation and Analysis:** Once the visualizations are created, they are analyzed to extract insights, recognize patterns, and make informed decisions.

Why Data Visualization is Important :

- **Simplifies complex data:** It makes large datasets easier to understand by converting them into visual formats like charts and graphs.
- **Reveals patterns and trends:** It helps quickly spot patterns and relationships in data that are hard to see in raw numbers.
- **Saves time:** Visuals allow for faster data interpretation.
- **Improves communication:** It makes it easier to share insights with others, regardless of their technical background.
- **Supports decision-making:** Clear visualizations lead to more informed, data-driven decisions.
- **Facilitates exploration:** Accessible visuals encourage more opportunities for collaboration and analysis.

Data Visualization Types :

Charts and Graphs

- **Bar Chart:** Compares quantities across different categories. The length of each bar represents the value of a variable.
- **Line Chart:** Displays trends or changes over a continuous interval, such as stock prices over a period of time.
- **Pie Chart:** Shows the proportion of parts to a whole, with each "slice" representing a percentage of the total.
- **Scatter Plot:** Reveals the relationship between two variables, often used to identify correlations.

- **Histogram:** A type of bar chart that shows the distribution of a continuous measure by splitting it into different bins. It helps in identifying where values are concentrated and where there are gaps.
- **Heat Map:** A graphical representation of data where values are depicted by color. It is often used to show the intensity of a variable in a matrix.

Other Visualizations

- **Table:** Displays data in rows and columns.
- **Infographic**: A combination of visuals and words to represent data, often using charts and diagrams.
- **Dashboard:** A collection of multiple visualizations displayed in a single interface to provide real-time insights and interactive features**.**

Data Visualization Tools

- **Tableau**: One of the most popular and powerful tools, known for its ease of use and ability to connect to hundreds of data sources.

  It offers a variety of products, including Tableau Desktop, Tableau Server, and a free version called Tableau Public.

  Tableau Desktop allows you to create interactive reports and dashboards, while Tableau Server is used to share workbooks with licensed users across an organization.
- **Power BI:** A data visualization tool by Microsoft that integrates easily with existing applications and supports a wide range of backend databases. It is a complete tool for creating stunning visualizations and delivering real-time insights.

- **Python Libraries:** Python offers several powerful libraries for data visualization.
    - **Matplotlib:** A fundamental library for 2D plots of arrays, providing low-level control and a wide variety of plots like line, bar, scatter, and histograms.
    - **Seaborn:** Built on top of Matplotlib, this library offers a high-level interface with great default styles and built-in themes for enhanced visualization.
    - **Plotly**: An interactive, open-source library that can create a variety of visualizations, including scientific and 3D graphs. Plotly graphs can be viewed in web browsers, Jupyter notebooks, or as standalone HTML files.
- **Google Charts:** A free and popular option for creating interactive data visualizations for the web. It uses HTML5/SVG technology to generate charts and can pull data from various sources.
- **Dundas BI:** Offers highly customizable data visualizations with interactive charts, maps, and scorecards. It gives users full control over visual elements, simplifying the process of cleansing and modeling large datasets.
- **JupyteR:** A web-based application that allows users to create and share documents containing live code, narrative text, equations, and visualizations. It's ideal for rapid prototyping and statistical modeling**.**

Exploratory Data Analysis (EDA)

- Exploratory Data Analysis (EDA) is a crucial step in the data science process.
- It involves using statistical summaries and graphical representations to analyze and investigate datasets, aiming to uncover patterns, detect anomalies (outliers), and summarize key characteristics.
- The primary goal of EDA is to understand the data before making any formal assumptions or moving on to modeling.
- It is an iterative process, starting with questions, searching for answers through visualization and transformation, and then refining those questions based on new insights.

Purpose of EDA

- **Better Understanding of Data:** EDA helps in understanding the structure, content, and relationships within the dataset.
- **Identification of Patterns:** It allows the discovery of hidden trends and relationships in the data.
- **Detection of Outliers and Anomalies:** EDA helps identify data points that are significantly different from others.
- **Data Quality Assessment**: It helps in detecting missing values and inconsistencies in the data.
- **Informing Model Selection**: Insights from EDA guide the selection of appropriate statistical models for the data.

Key Features of EDA

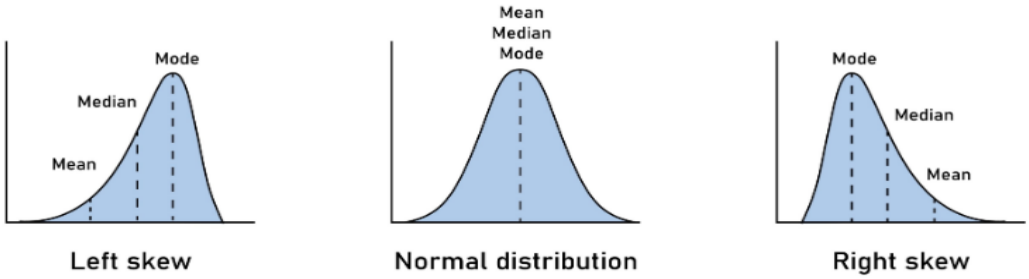
- **Flexibility and Attitude:** It's an approach that is free from rigid assumptions and lets the data "speak for itself".
- **Iterative Cycle:** EDA is not a one-time task but a continuous loop of questioning, exploring, and refining.
- **Use of Visual and Non-Visual Methods:** It employs both statistical summaries (non-graphical methods) and a wide array of plots and charts (graphical methods) to gain insights.

Types of EDA

1. Univariate Non-Graphical EDA

This type of analysis focuses on summarizing a single variable without creating any plots. The goal is to understand the central tendency, spread, and shape of the data's distribution.

- **Measures of Central Tendency:** The mean, median, and mode are used to find the typical or middle value of the dataset.
- **Measures of Spread:** Variance and standard deviation are used to quantify how much the data values deviate from the central value.
- **Skewness and Kurtosis:** These measures describe the asymmetry and "peakedness" of the distribution.
- Example: Analyzing 'Age' from a dataset of customers.



2. Multivariate Non-Graphical EDA

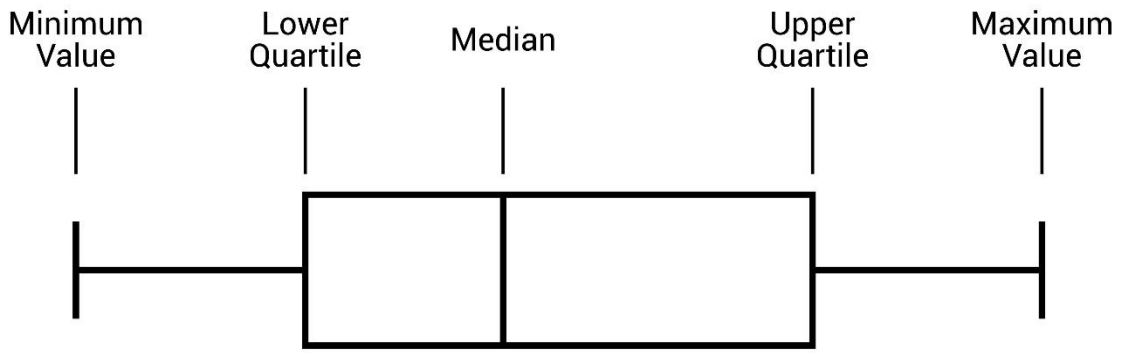
This method explores the relationships between two or more variables using statistical summaries.

- **Cross-Tabulation:** Used for categorical data, this involves creating a two-way table to show the frequency of different combinations of categories between two variables. For example, a cross-tabulation of 'Gender' and 'Political Party' could show how many men and women belong to each party.
- **Summary Statistics by Group:** For a combination of categorical and quantitative variables, statistics (like mean or median) are calculated for the quantitative variable across each level of the categorical variable to allow for comparison.
- Example: Analyzing the relationship between 'Hours Spent on Site' and 'Purchase Amount' from a customer dataset.

3. Univariate Graphical EDA

This type uses a single variable to create visual representations, which can often provide a more complete picture of the data than non-graphical methods alone.

- **Histograms:** Bar plots where each bar represents the frequency of data within a specific range of values. They are excellent for quickly understanding the data's central tendency, spread, modality, and shape.
- **Box Plots:** These plots summarize data based on the five-number summary (minimum, first quartile, median, third quartile, and maximum), and are particularly useful for identifying outliers and understanding distribution symmetry.



- **Stem-and-Leaf Plots:** A textual plot that displays both the shape of the distribution and all the data values.
- Example: Visualizing the 'Age' variable from a customer dataset.

4. Multivariate Graphical EDA

This approach is used to visualize the relationships and interactions between multiple variables, which is crucial for identifying correlations and complex patterns.

- **Scatter Plots:** A fundamental tool for visualizing the relationship between two quantitative variables, where each data point is plotted according to its values on the x and y axes.
- **Heat Maps:** A graphical representation of data where values are represented by colors, often used to show the correlation matrix of multiple variables. It helps in quickly identifying which variable pairs are highly correlated.
- **Pair Plots:** A grid of scatter plots that visualizes the relationship between each numerical variable in a dataset. It is useful for a quick, high-level overview of all potential pairwise correlations.
- Example: Analyzing the relationship between 'Hours Spent on Site' and 'Purchase Amount' from a customer dataset.