**Task-2**

**What is EDA?**

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

EDA helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

**Types of EDA:**

There are four primary types of EDA:

* **Univariate non-graphical.** This is simplest form of data analysis, where the data being analyzed consists of just one variable. Since it’s a single variable, it doesn’t deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
* **Univariate graphical.** Non-graphical methods don’t provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include:
  + Stem-and-leaf plots, which show all data values and the shape of the distribution.
  + Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.
  + Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.
* **Multivariate nongraphical:**Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.
* **Multivariate graphical:** Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

**Advantages of EDA:**

Exploratory Data Analysis (EDA) offers numerous advantages in data analytics, making it an essential step in the data analysis process. Here are some of the key benefits:

**1. Better Understanding of Data**

* **Insight into Data Structure:** EDA helps in comprehending the dataset's structure, including the types of variables, the distribution of values, and the presence of any underlying patterns.
* **Contextual Understanding:** It provides context about the data, which can inform subsequent analysis steps and decision-making processes.

**2. Identifying Data Quality Issues**

* **Detecting Missing Values:** EDA helps identify missing data points and provides insights into the patterns of missingness.
* **Spotting Outliers and Anomalies:** It reveals unusual data points that might indicate errors, data entry issues, or significant phenomena.
* **Assessing Data Consistency:** EDA helps check for inconsistencies, such as unexpected duplicates or mismatches in data.

**3. Informing Data Cleaning and Preparation**

* **Guiding Data Transformation:** It helps decide on necessary transformations, such as normalization, scaling, or encoding categorical variables.
* **Handling Missing Data:** EDA informs strategies for dealing with missing values, whether through imputation, removal, or other methods.
* **Improving Data Quality:** By identifying and addressing data quality issues, EDA ensures that the dataset is reliable and suitable for analysis.

**4. Uncovering Patterns and Relationships**

* **Detecting Trends and Patterns:** EDA can reveal trends over time, seasonal patterns, or other systematic relationships in the data.
* **Understanding Variable Relationships:** It helps identify correlations and interactions between variables, which can be crucial for building predictive models.

**5. Hypothesis Generation**

* **Formulating Hypotheses:** EDA enables the generation of new hypotheses about the data and its underlying processes, guiding further investigation and analysis.
* **Testing Assumptions:** It allows for the validation of assumptions required for various statistical techniques and models.

**6. Model Building and Selection**

* **Feature Selection:** EDA helps identify relevant features and variables that should be included in predictive models.
* **Model Specification:** It informs the choice of appropriate models and techniques based on the data's characteristics and relationships.

**7. Enhanced Visualization and Communication**

* **Data Visualization:** EDA leverages visual tools like histograms, scatter plots, and heatmaps to convey data insights effectively.
* **Storytelling with Data:** By visualizing data, EDA helps communicate findings to stakeholders, making it easier to convey complex information in an understandable format.

**8. Improved Decision Making**

* **Informed Decisions:** With a thorough understanding of the data, decision-makers can make more informed and accurate decisions.
* **Risk Reduction:** EDA helps mitigate risks by identifying potential issues early in the analysis process, ensuring robust and reliable results.

**Disadvantages of EDA:**

While Exploratory Data Analysis (EDA) is a fundamental step in the data analysis process, it does have some disadvantages and limitations. Here are some of the key drawbacks:

**1. Time-Consuming Process**

* **Manual Effort:** EDA often involves a lot of manual effort, especially when dealing with large and complex datasets.
* **Iterative Nature:** The iterative process of exploring, cleaning, and visualizing data can be very time-consuming, delaying further analysis steps.

**2. Subjectivity and Bias**

* **Analyst Bias:** EDA relies heavily on the analyst's judgment, which can introduce biases in interpreting the data.
* **Overfitting:** There is a risk of overfitting the data to the analyst’s expectations, potentially leading to misleading conclusions.

**3. Complexity with Large Datasets**

* **Scalability Issues:** Handling and visualizing very large datasets can be challenging and may require significant computational resources.
* **Information Overload:** Large datasets can produce overwhelming amounts of information, making it difficult to discern useful insights.

**4. Limited by Data Quality**

* **Garbage In, Garbage Out:** The effectiveness of EDA is limited by the quality of the data. Poor-quality data can lead to incorrect conclusions.
* **Incomplete Data:** Incomplete or missing data can skew results and hinder the exploration process.

**5. Lack of Standardization**

* **Ad Hoc Nature:** EDA is often performed in an ad hoc manner without standardized procedures, which can lead to inconsistencies and difficulty in reproducing results.
* **Variable Practices:** Different analysts may use different tools, techniques, and approaches, resulting in varied interpretations of the same dataset.

**6. Potential for Misinterpretation**

* **Misleading Visualizations:** Poorly designed visualizations can mislead the audience, resulting in incorrect conclusions.
* **Correlation vs. Causation:** EDA can identify correlations but cannot establish causation, which might be misinterpreted without further analysis.

**7. Resource Intensive**

* **High Computational Demand:** EDA can be resource-intensive, requiring significant computational power and memory, particularly for large datasets.
* **Specialized Tools:** Effective EDA often requires specialized tools and software, which might not be readily available to all analysts.

**8. Overemphasis on Exploration**

* **Paralysis by Analysis:** Spending too much time on EDA can delay the implementation of actual analysis and decision-making processes.
* **Risk of Overanalyzing:** There is a risk of overanalyzing the data during EDA, leading to wasted effort on insignificant details.

**Features of EDA:**

Exploratory Data Analysis (EDA) is a critical step in the data analytics process, aimed at understanding the dataset's underlying structure, extracting important variables, detecting outliers and anomalies, and testing hypotheses. Here are some key features of EDA:

**1. Data Summarization**

* **Descriptive Statistics:** Compute basic statistics such as mean, median, mode, variance, standard deviation, skewness, and kurtosis.
* **Frequency Distributions:** Understand the distribution of categorical variables by calculating the frequency and relative frequency of each category.

**2. Data Visualization**

* **Univariate Analysis:**
  + **Histograms:** Visualize the distribution of a single numeric variable.
  + **Box Plots:** Identify the spread and outliers in a single numeric variable.
  + **Bar Charts:** Display the frequency of different categories in a categorical variable.
* **Bivariate Analysis:**
  + **Scatter Plots:** Examine relationships between two continuous variables.
  + **Box Plots:** Compare the distribution of a numeric variable across different categories.
  + **Heatmaps:** Show correlations between numeric variables.
* **Multivariate Analysis:**
  + **Pair Plots:** Visualize relationships between multiple pairs of variables.
  + **3D Scatter Plots:** Examine relationships between three continuous variables.

**3. Data Cleaning and Preprocessing**

* **Handling Missing Values:** Identify and address missing values through imputation or removal.
* **Data Transformation:** Apply transformations such as normalization, scaling, or encoding categorical variables.
* **Outlier Detection:** Identify and potentially remove or investigate outliers.

**4. Pattern and Trend Detection**

* **Time Series Analysis:** Identify trends, seasonality, and cyclical patterns in time-series data.
* **Correlation Analysis:** Measure the strength and direction of relationships between numeric variables using correlation coefficients.

**5. Hypothesis Testing**

* **Assumption Checks:** Validate assumptions required for statistical tests and models, such as normality and homoscedasticity.
* **Initial Hypothesis Generation:** Develop hypotheses about relationships and patterns within the data for further analysis.

**6. Feature Engineering**

* **Variable Transformation:** Create new features by transforming existing variables (e.g., log transformation, polynomial features).
* **Feature Extraction:** Identify and extract important features that contribute to the predictive power of models.

**7. Dimensionality Reduction**

* **Principal Component Analysis (PCA):** Reduce the dimensionality of the data while retaining most of the variance.
* **Factor Analysis:** Identify underlying relationships between variables and group them into factors.

**8. Data Distribution Analysis**

* **Normality Tests:** Assess whether a variable follows a normal distribution using tests like the Shapiro-Wilk test.
* **Probability Plots:** Use Q-Q plots to compare the distribution of a variable against a theoretical distribution.

**9. Comparative Analysis**

* **Group Comparisons:** Compare summary statistics and distributions across different groups or categories within the data.

**10. Interactive and Dynamic Analysis**

* **Interactive Visualization Tools:** Use tools like Tableau, Power BI, or Python libraries (e.g., Plotly, Bokeh) to create interactive visualizations for deeper insights.
* **Dynamic Data Exploration:** Use Jupyter Notebooks or other interactive environments to iteratively explore and visualize data.