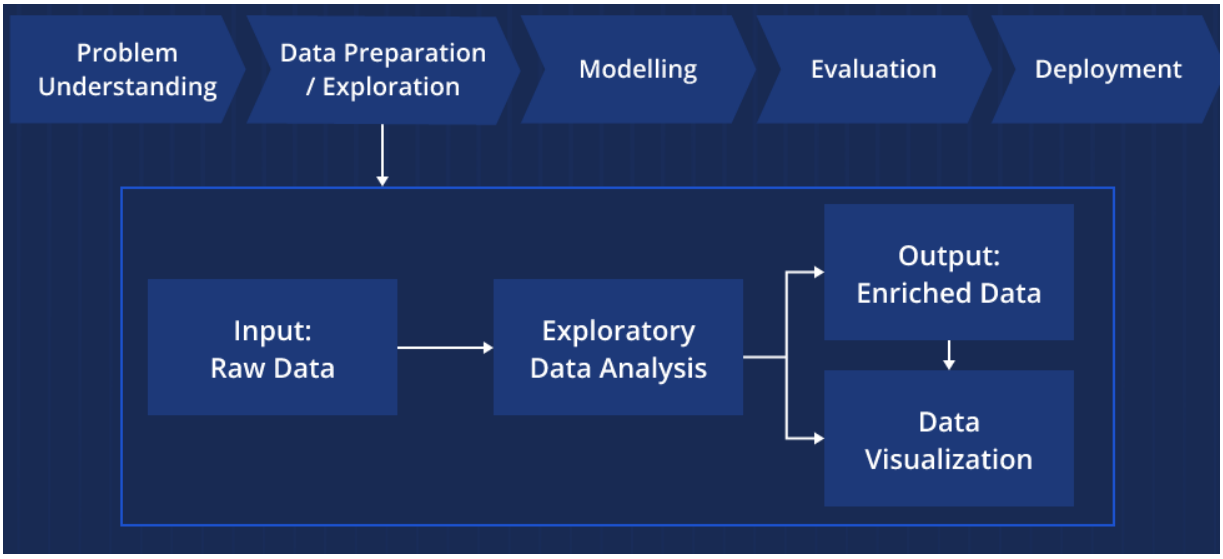
Exploratory Data Analysis (EDA)

What is Exploratory Data Analysis (EDA)



- EDA is a crucial aspect of data science that helps in understanding the underlying structure of the data. It involves analyzing and summarizing data visually and statistically to uncover patterns or relationships.

- Refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

- It is the preliminary method used for interpreting data before undertaking any formal modeling or hypothesis testing.

- **Data exploration** is a critical initial step in the data analysis process, where analysts examine large datasets to uncover patterns, outliers, and relationships before formal modeling and analysis occur.

- EDA utilizes various statistical techniques and powerful data visualization tools to understand the data's key characteristics, quality, and structure.
- Open-source tools like Python, R, Tableau enable robust data visualization during exploration of data through methods like histograms, scatter plots, box plots, and more.
- EDA can assist in identifying missing or incomplete data, outliers and inconsistencies that can impact the statistical analysis of data.
- EDA is done to get a sense of data and discover any potential problems or issues which need to be addressed before formal modeling or hypothesis testing.

The main goals of EDA:

- Gaining a deeper understanding of the data
- Identifying data quality issues
- Developing initial insights and hypotheses
- Selecting features for modeling or further analysis

EDA methods and techniques

Common techniques and methods used in Exploratory Data Analysis include the following:

1. **Data visualization**

Data visualization involves generating visual representations of the data using graphs, charts, and other graphical techniques. Data visualization enables a quick and easy understanding of patterns and relationships within data. Visualization techniques include scatter plots, histograms, heatmaps and box plots.

2. **Correlation analysis**

Using correlation analysis, one can analyze the relationships between pairs of variables to identify any correlations or dependencies between them. Correlation analysis helps in feature selection and in building predictive models. Common

correlation techniques include Pearson's correlation coefficient, Spearman's rank correlation coefficient and Kendall's tau correlation coefficient.

3. **Descriptive statistics**

It involves calculating summary statistics such as mean, median, mode, standard deviation and variance to gain insights into the distribution of data. The mean is the average value of the data set and provides an idea of the central tendency of the data. The median is the mid-value in a sorted list of values and provides another measure of central tendency. The mode is the most common value in the data set.

4. **Clustering**

Clustering techniques such as K-means clustering, hierarchical clustering, and DBSCAN clustering help identify patterns and relationships within a dataset by grouping similar data points together based on their characteristics.

5. **Outlier detection**

Outliers are data points that vary or deviate significantly from the rest of the data and can have a crucial impact on the accuracy of models. Identifying and removing outliers from data using methods like Z-score, interquartile range (IQR) and box plots method can help improve the data quality and the models' accuracy.

How to Perform Exploratory Data Analysis (EDA):

I. Data Collection:

The first step in EDA.

Data can come from a variety of sources, including databases, websites, APIs, and file storage services. It's essential to ensure that the data is collected in a format that can be easily analyzed and manipulated.

Process of Collecting Data

To collect data, consider the following steps:

1. Define the research question you want to answer.

2. Identify the data needed to answer the question.

3. Identify potential data sources, such as databases, websites, APIs, or file. storage services.

4. Collect the data in a format that can be easily analyzed and manipulated.

II. Importing Libraries:

When performing EDA, you'll need to import several libraries to facilitate data manipulation, visualization, and analysis.

III. Data Cleaning:

Before exploring and analyzing the data through EDA, data cleaning is necessary. It involves identifying and resolving data quality issues such as missing values, outliers, and duplicates. These issues can affect the accuracy of the insights derived during EDA.

Steps in Data Cleaning

To clean your data effectively, follow these steps:

1. Identify and remove duplicates.

Identifying and removing duplicates from a dataset is an essential data cleaning step to ensure data integrity and accuracy in any data analysis or machine learning project. Duplicates can arise due to data entry errors, system glitches, or other reasons, and they can skew analyses and lead to misleading results.

2. Identify and deal with missing values in the data.

Dealing with missing values is a crucial aspect of data cleaning and preprocessing, as missing data can lead to biased analysis and inaccurate results. There are various strategies to handle missing values, depending on the nature of the data and the amount of missingness.

Based on the evaluation, choose an appropriate strategy to handle missing values. Some common techniques include:

a. Deletion:

If missing values are limited and randomly distributed, you may choose to remove rows or columns with missing data.

b. Imputation:

If you have a significant amount of missing data or cannot afford to lose entire rows or columns, impute missing values with reasonable estimates.

Common imputation methods include mean, median, mode, forward-fill, backward-fill, and regression-based imputation.

3. Address outliers by either removing or transforming them.

Outliers are data points that deviate significantly from the rest of the data and can be caused by errors in data collection or represent extreme cases that need special consideration. Dealing with outliers can be done in two primary ways:

a. Removing Outliers:

Removing outliers involves simply eliminating data points that are considered outliers from the dataset. This method is suitable when you are confident that the outliers are due to data entry errors or are irrelevant to the analysis

b. Transforming Outliers:

Instead of removing outliers, another approach is to transform the data to reduce the impact of outliers on the analysis. Common transformation techniques include log transformation, square root transformation, or Box-Cox transformation.

4. Remove or correct irrelevant data.

Irrelevant data includes information that is not useful or applicable to the analysis, either due to data entry errors or because it does not align with the research question or analysis objectives.

Step 1: Identify Irrelevant Data Start by examining your dataset and identifying any data points or columns that do not contribute to the analysis or have no meaningful relevance. This could include:

a. Columns with constant values.

b. Columns with mostly missing data.

c. Data points or rows with missing values in critical fields.

Step 2: Evaluate Data Quality For data that seems potentially irrelevant, check for data quality issues, such as data entry errors, typos, or inconsistencies. Evaluate whether these issues can be corrected or if the data should be removed.

Step 3: Remove Irrelevant data once you have identified irrelevant data, you can remove it from the dataset.

5. Detect and deal with any invalid or inconsistent data.

Invalid data refers to information that does not conform to the expected format or valid range, while inconsistent data refers to contradictory or conflicting entries within the dataset.

Step 1: Identify Invalid or Inconsistent Data Start by thoroughly examining your dataset and identifying potential issues, such as:

a. Data outside valid ranges.

b. Contradictory information.

c. Format discrepancies.

d. Data type inconsistencies.

Step 2: Data Validation Perform data validation checks to verify the integrity of the data. Common validation checks include:

a. Ensure that numeric values fall within valid ranges.

b. Validate that data entries match the expected format.

c. Cross-check related fields to ensure their consistency.

Step 3: Handle Invalid or Inconsistent Data The appropriate method for handling invalid or inconsistent data depends on the specific context and the nature of the errors. Options include:

a. Removal.

b. Imputation.

c. Manual Correction.

d. Data Transformation.

IV. Data Exploration:

Exploring the data is the next step in the EDA process. It helps to identify patterns and trends and forms the basis of the data analysis phase. Common EDA techniques include the use of summary statistics, visualizations, and correlation matrices.

Process of Exploring Data

To explore the data effectively, follow these four steps:

1. Calculate summary statistics such as mean, median, and mode to understand the central tendencies and distributions of the data.

Mean: The mean, also known as the average, is calculated by summing all the values in the dataset and dividing by the number of data points.

Median: The median is the middle value in a dataset when it is sorted in ascending order. If the dataset has an odd number of data points, the median is the middle value. If the dataset has an even number of data points, the median is the average of the two middle values.

2. Visualize the data using graphs, charts, and histograms. This will help to identify any patterns or anomalies in the data.

- Scatter Plot
- Line Chart
- Histogram
- Box Plot

3. Examine the correlations between variables using correlation matrices or scatter plots.

Correlation Matrix:

A correlation matrix is a table that displays the correlation coefficients between all pairs of numeric variables in the dataset. Correlation coefficients quantify the strength and direction of the linear relationship between two variables. The values range from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no linear correlation.

Scatter Plot:

A scatter plot is a graphical representation of the relationship between two continuous variables. Each data point is plotted as a point on the chart, with

one variable represented on the x-axis and the other on the y-axis. Scatter plots help identify patterns and trends in the data, such as linear, non-linear, or no relationship.

4. Use hypothesis tests to form initial insights and generate new hypotheses.

Hypothesis testing is a statistical method used to make inferences about a population based on sample data. It allows you to test assumptions or claims about the population and draw conclusions from the observed data. By using hypothesis tests, you can form initial insights and generate new hypotheses, which can guide further data analysis and investigation.

a. Formulate the Null and Alternative Hypotheses.

b. Choose a Statistical Test.

c. Set the Significance Level (Alpha).

d. Calculate the Test Statistic and P-value.

e. Make a decision, draw conclusions and form New Hypotheses.

Example: Chi-square Test for Independence:

Null Hypothesis (H_0): There is no association between smoking status (smoker/non-smoker) and the incidence of lung cancer.

Alternative Hypothesis (H_1): There is an association between smoking status and the incidence of lung cancer.

Perform a chi-square test for independence to assess whether smoking status and the incidence of lung cancer are related. If the p-value is less than the

significance level, you may reject the null hypothesis and conclude that smoking status and lung cancer incidence are associated.

V. Data Analysis:

The data analysis phase involves using various techniques to gain insights into the underlying structure and relationships in the data. Common techniques used include clustering, regression, and classification.

Analyzing Data Effectively

To analyze your data effectively, consider the following:

1. Decide what type of analysis is appropriate for the research question.
2. Implement the chosen analysis technique.
3. Interpret the results and identify patterns and trends in the data.
4. Update further research questions and explore in more detail.

Types of EDA techniques

Several types of exploratory data analysis techniques can be used to gain insights into data. Some common types of EDA include:

Univariate non-graphical

Univariate non-graphical exploratory data analysis is a simple yet fundamental method for examining information that includes utilizing only one variable to analyze the data. Univariate non-graphical EDA focuses on figuring out the underlying distribution or pattern in the data and mentions objective facts about the population. This procedure includes the examination of the attributes of the population distribution, including spread, central tendency, skewness and kurtosis.

- An average or middle value of a distribution is called the central tendency. A common measure of central tendency is the mean, followed by the median and mode. As a measure of central tendency, the median may be preferred if the distribution is skewed or concerns are raised about outliers.
- Spread shows how far off the information values are from the central tendency. The standard deviation and variance are two valuable proportions of the spread. The variance is the mean of the square of the individual deviations, and the standard deviation is the foundation of the variance.
- Skewness and kurtosis are two more helpful univariate descriptors of the distribution. Skewness is a metric of the asymmetry of the distribution, while kurtosis is a proportion of the peakedness of the distribution contrasted with an ordinary dispersion.

Outlier detection is also important in univariate non-graphical EDA, as outliers can significantly impact the distribution and distort statistical analysis results.

Multivariate non-graphical

Multivariate non-graphical EDA is a technique used to explore the relationship between two or more variables through cross-tabulation or statistics. It is useful

for identifying patterns and relationships between variables. This analysis is particularly useful when multiple variables exist in a dataset, and you want to see how they relate.

Cross-tabulation is a helpful extension of tabulation for categorical data. Cross-tabulation is preferable when there are two variables involved. To do this, create a two-way table with column headings corresponding to the number of one variable and row headings corresponding to the number of the other two variables. Next, fill the counts with all subjects with the same pair of levels.

We produce statistics for quantitative variables individually for each level of each categorical variable and one quantitative variable, and then we compare the statistics across all categorical variables. The purpose of multivariate non-graphical EDA is to identify relationships between variables and understand how they are related. Examining the relationship between variables makes it possible to discover patterns and trends that may not be immediately obvious from examining individual variables in isolation.

Univariate graphical

A univariate graphical EDA technique employs a variety of graphs to gain insight into a single variable's distribution. These graphical techniques enable us to gain a quick understanding of shapes, central tendencies, spreads, modalities, skewnesses, and outliers of the data we are studying. The following are some of the most commonly used univariate graphical EDA techniques:

1. **Histogram:** This is one of the most basic graphs used in EDA. A histogram is a bar plot that displays the frequency or proportion of cases in each of several intervals (bins) of a variable's values. The height of each bar represents the count or proportion of observations that fall within each interval. Histograms provide an intuitive sense of the shape and spread of the distribution, as well as any outliers.
2. **Stem-and-leaf plots:** A stem-and-leaf plot is an alternative to a histogram that displays each data value along with its magnitude. In a stem-and-leaf plot, each data value is split into a stem and leaf, with the stem representing the leading digits and the leaf representing the trailing digits.

This type of plot provides a visual representation of the data's distribution and can highlight features such as symmetry and skewness.

3. **Boxplots:** Boxplots, also known as box-and-whisker plots, provide a visual summary of the distribution's central tendency, spread and outliers. The box in a boxplot represents the data's interquartile range (IQR), with the median line inside the box. The whiskers extend from the box to the smallest and largest observations within 1.5 times the IQR from the box. Data points outside of the whiskers are considered outliers.
4. **Quantile-normal plots:** A quantile-normal plot, also known as a Q-Q plot, assesses the data distribution by comparing the observed values to the expected values from a normal distribution. In a Q-Q plot, the observed data is plotted against the quantiles of a normal distribution. The points should lie along a straight line if the data is normally distributed. If the data deviates from normality, the plot will reveal any skewness, kurtosis, or outliers.

Multivariate graphical

A multivariate graphical EDA displays relationships between two or more data sets using graphics. When examining relationships between variables beyond two, this technique is used to gain a more comprehensive understanding of the data. A grouped barplot is one of the most commonly used multivariate graphical techniques, with each group representing one level of one variable and each bar representing its amount.

Multivariate graphics can also be represented in scatterplots, run charts, heat maps, multivariate charts, and bubble charts.

- **Scatterplots** are graphical representations displaying the relationship between two quantitative/numerical variables. It consists of plotting one variable on the x-axis and another on the y-axis. On the plot, each point represents an observation. Scatterplots make it possible to identify outliers or patterns in the data and the direction and strength of the relationship between any two variables.
- **A run chart** is a line graph that shows how data changes over time. It is a simple but powerful tool for tracking changes and monitoring trends in

data. Run charts can be used to detect trends, cycles, or shifts in a process over time.

- **A multivariate chart** illustrates the relationship between factors and responses. It is a type of scatterplot that depicts relationships between several variables simultaneously. A multivariate chart depicts the relationship between variables and identifies patterns or clusters in the data.
- **Bubble chart** is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot. The size of each circle represents a value of a third variable. Bubble charts are often used to compare data sets with three variables, as they provide an easy way to visualize the relationships between these variables.

Visualization techniques in EDA

Visualization techniques play an essential role in EDA, enabling us to explore and understand complex data structures and relationships visually. Some common visualization techniques used in EDA are:

1. **Histograms:** Histograms are graphical representations that show the distribution of numerical variables. They help understand the central tendency and spread of the data by visualizing the frequency distribution.
2. **Boxplots:** A boxplot is a graph showing the distribution of a numerical variable. This visualization technique helps identify any outliers and understand the spread of the data by visualizing its quartiles.
3. **Heatmaps:** They are graphical representations of data in which colors represent values. They are often used to display complex data sets, providing a quick and easy way to visualize patterns and trends in large amounts of data.

4. **Bar charts:** A bar chart is a graph that shows the distribution of a categorical variable. It is used to visualize the frequency distribution of the data, which helps to understand the relative frequency of each category.
5. **Line charts:** A line chart is a graph that shows the trend of a numerical variable over time. It is used to visualize the changes in the data over time and to identify any patterns or trends.
6. **Pie charts:** Pie charts are a graph that showcases the proportion of a categorical variable. It is used to visualize each category's relative proportion and understand the data distribution.

Exploratory data analysis tools

Spreadsheet software

Due to its simplicity, familiar interface and basic statistical analysis capabilities, spreadsheet software such as Microsoft Excel, Google Sheets, or LibreOffice Calc is commonly used for EDA. Using them, users can sort, filter, manipulate data and perform basic statistical analysis, like calculating the mean, median and standard deviation.

Statistical software

Specialized statistical software such as R or Python and their various libraries and packages offer more advanced statistical analysis tools, including regression analysis, hypothesis testing, and time series analysis. This software allows users to write customized functions and perform complex statistical analyses on large datasets.

Data visualization software

Visualization software like Tableau, Power BI, or QlikView enables users to create interactive and dynamic data visualizations. These tools help users to identify patterns and relationships in the data, allowing for more informed decision-making. They also offer various types of charts and graphs, as well as the ability to create dashboards and reports. The software allows data to be easily shared and published, making it useful for collaborative projects or presentations.

Programming languages

Programming languages such as R, Python, Julia and MATLAB offer powerful numerical computing capabilities and provide access to various statistical analysis tools. These languages can be used to write customized functions for specific analysis needs and are particularly useful when working with large datasets. They also enable the automation of repetitive tasks, besides bringing flexibility in data handling and manipulation.

Business Intelligence (BI) tools

BI tools like SAP BusinessObjects, IBM Cognos or Oracle BI offer a range of functionalities, including data exploration, dashboards and reports. They allow users to visualize and analyze data from various sources, including databases and spreadsheets. They provide data preparation tools and quality management tools that can be used in business settings to help organizations make data-driven decisions.

Data mining tools

Data mining tools such as KNIME, RapidMiner or Weka provide a range of functionalities, including data preprocessing, clustering, classification and association rule mining. These tools are particularly useful for identifying patterns and relationships in large datasets and building predictive models. Data mining tools are used in various industries, including finance, healthcare and retail.

Cloud-based tools

Cloud-based platforms such as Google Cloud, Amazon Web Services (AWS) and Microsoft Azure offer a range of tools and services for data analysis. They provide a scalable and flexible infrastructure for storing and processing data and offer a range of data analysis and visualization tools. Cloud-based tools are particularly useful for working with large and complex datasets, as they offer high-performance computing resources and the ability to scale up or down depending on the project's needs.

Text analytics tools

Text analytics tools like RapidMiner and SAS Text Analytics are used to analyze unstructured data, such as text documents or social media posts. They use natural

language processing (NLP) techniques to extract insights from text data, such as sentiment analysis, entity recognition and topic modeling. Text analytics tools are used in a range of industries, including marketing, customer service and political analysis.

Geographic Information System (GIS) tools

GIS tools such as ArcGIS and QGIS are used to analyze and visualize geospatial data. They allow users to map data and perform spatial analysis, such as identifying patterns and trends in geographical data or performing spatial queries. GIS tools are used in a range of industries, including urban planning, environmental management and transportation.