EXPLORATORY DATA ANALYSIS (EDA)

INTRODUCTION

- Exploratory Data Analysis (EDA) is a crucial and initial step in data science projects.
- It involves analyzing and visualizing data to understand its key characteristics, structures,uncover patterns, and identify relationships between variables.
- By understanding the data, we can identify potential issues, clean it up.
- By this we can prepare it for further analysis like statistical modeling or machine learning.
- This leads to more reliable and accurate results.

OBJECTIVES

Data Understanding:

- For the data understanding some steps are include like:-
    - Determine the number of data points (rows) and variables (columns) we are working with.
    - Check for missing values and understand their potential impact on our analysis.
    - Check are the numbers numbers (like age) or words (like color)
    - Calculate basic statistics like mean, median, standard deviation to get a feel for the data's central tendency and spread.
    - Explore how data points are distributed (normal, skewed, etc.) using techniques like histograms.

Identifying Patterns:

- Create charts and graphs like histograms, scatter plots, box plots, and heatmaps to visualize trends and relationships between variables.
- See if data points tend to group together based on certain characteristics.
- If our data has a time component, exploring how patterns change over time.
- Look for trends and connections using charts and graphs.
- Identify data points that fall significantly outside the expected range and investigate further (might be errors or interesting exceptions).

Data Cleaning:

- After identifying patterns and potential issues, you might need to clean the data. This could involve:-
    - Correct any errors if we find in the data.
    - Decide how to handle missing information.
    - Ensure all the data uses the same format (like dates).
    - Decide what to do with outliers (remove, fix, investigate further).

Feature Selection:

- The feature selection is used when it is needed only.
- Choose the data points most relevant to what you're trying to learn.This can improve the accuracy of your models and reduce training time.
- Don't pick data that tells you the same thing twice(like avoiding repetations).
- Use statistical techniques to identify features that have the most significant impact on your target variable.

Hypothesis Generation:

- Based on our EDA findings, formulate hypotheses to be tested through statistical modeling or machine learning.
- These hypotheses should be clear, concise, and testable with our data.
- Based on your findings, guess what might be true about the data.
- Like "People buy more ice cream when it's hot" based on a connection we saw in the data.

Techniques and Methods

- There are many techniques and methods used in EDA. Here are some of the most common ones:

Summary Statistics:

- There are basic calculations that describe a dataset's central tendency (mean, median) and spread (standard deviation, variance).
- Identify the minimum and maximum values in our data set.
- Like finding the average height of people in a dataset.
- Tells us basic things about the data.
- Divide our data into four equal parts (quartiles) to understand data distribution.

Data Visualization:

- Charts and graphs to see the data in a clear way.
- The Line charts shows trends and changes over time.
- The Pie charts represent proportions of categorical data.
- The Heatmaps visualize relationships between many variables at once.
- Like bar charts to see how many people have each hair color.

Correlation Analysis:

- This helps measure the strength and direction of the linear relationship between two variables.
- Measure the strength and direction of the linear relationship between two variables (values range from -1 to 1).
- See if two things are related (like ice cream sales and temperature).
- Doesn't mean one causes the other, just that they might be connected.

Dimensionality Reduction:

- The Dimensionality reduction is used when it is needed only.
- If you have a lot of features (variables), this process can help reduce them to a smaller set while retaining most of the information.
- The Dimensionality reduction have some techniques like Principal Component Analysis(PCA),Feature Selection.
- If we have a lot of data points, group similar ones together.
- Makes it easier to analyze the data without losing important information.

Clustering:

- This is unsupervised learning that groups data points into similar clusters based on their features.
- Group similar data points together automatically.
- Like putting all the red cars in one pile and all the blue cars in another.
- There are different type of clustering like K-means clustering,Hierarchical clustering.
- The K-means clustering means groups data points into a predefined number of clusters based on similarities.
- The Hierarchical clustering means builds a hierarchy of clusters, allowing you to explore the data at different granularities.

Outlier Detection:

- This involves identifying data points that fall outside the expected range.
- Find data points that seem very different from the rest.
- Might be mistakes or interesting exceptions.
- There are some different methods of outlier detection they are Interquartile Range (IQR) and Z-scores.
- The Interquartile Range (IQR) identifies the data points that fall outside a specific range based on the quartiles.
- The Z-scores Standardize our data and identify points that deviate significantly from the mean (standard deviation).

Missing Value Imputation:

- This deals with handling missing data points in your dataset.
- There are different techniques for imputation, like replacing them with mean/median/mode or using more sophisticated methods.
- The Mean/Median/Mode imputation fill missing values with the average, middle value, or most frequent value in the dataset (be cautious with this approach).
- How to handle data points that are missing information.
- Like filling the blank on a missing age with an average age.

Conclusion:

- The Exploratory Data Analysis is more crucial part for the data analysis or data science.
- It is like the first step of cleaning and organizing the data before we can find the hidden treasures (valuable insights) inside.
- The hidden tresures are like visualizing it and finding patterns.
- By understanding that we can fix any issues having and seeing how things are connect and we can ready to use the data much more powerful ways.