Exploratory Data Analysis (EDA)

Objective:

Extract insights using visual and statistical exploration.

Tools:

Python (Pandas, Matplotlib, Seaborn)

Observation:

In this analysis, we performed an exploratory study of the Titanic dataset using Python libraries such as pandas, matplotlib, and seaborn. Key findings from the exploration are:

Data Structure: The dataset contains both numerical and categorical variables, with some missing values in important columns like Age, Cabin, and Embarked.

Missing Data: Significant missing data was found in the Cabin column, suggesting it might be useful to drop or impute this column in future modeling.

Passenger Demographics:

Most passengers were young adults.

There were more male passengers than female passengers.

Survival Analysis:

Females had a higher survival rate compared to males.

Passengers in higher classes (Pclass 1) had better survival chances.

Correlation Analysis:

Features like Fare and Pclass showed notable correlations with survival, indicating their potential importance for predictive modeling.

Some features are weakly correlated, implying that further feature engineering might be needed for better model performance.

Summary:

This project involves performing Exploratory Data Analysis (EDA) on the Titanic dataset with the goal of uncovering insights about the passengers and factors related to their survival. The analysis follows a structured approach:

Data Loading: The dataset is imported using pandas and the first few records are examined.

Basic Exploration:

Information about data types, non-null values, and statistical measures (mean, median, standard deviation) was obtained.

Missing values were identified, particularly in the Age, Cabin, and Embarked columns.

Categorical Analysis:

Value counts for categorical columns (e.g., Sex, Embarked, Pclass) were studied to understand distributions.

Visualization:

Histograms showed the distribution of numerical features such as Age, Fare, and SibSp.

Boxplots were used to detect potential outliers in numeric columns.

Correlation Study:

A heatmap was generated to observe relationships between numerical features.

Features like Fare and Pclass showed notable correlations with survival status.

The notebook provides a solid preliminary understanding of the dataset, highlights important variables, and sets the stage for future predictive modeling or deeper statistical analysis.