Titanic Dataset - Exploratory Data Analysis Report

Objective

Perform exploratory data analysis on the Titanic dataset to extract meaningful patterns, trends, and insights using Python libraries (Pandas, Matplotlib, Seaborn).

Dataset Description

- Source: Kaggle Titanic Competition
- Features: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked

Steps Performed

a. Data Loading and Initial Exploration:

- Loaded train.csv using Pandas
- Viewed first few records with .head()
- Checked dataset info and statistical summary with .info() and .describe()

b. Data Cleaning:

- Filled missing Age values with median
- Filled missing Embarked values with mode
- Dropped Cabin column (too many missing values)

Exploratory Data Analysis (EDA)

Univariate Analysis:

- Age Distribution: Most passengers aged between 20-40 years.
- Fare Distribution: Right-skewed distribution.
- Categorical Features: Majority were male, did not survive, belonged to 3rd class, and embarked from Southampton (S).

Bivariate Analysis:

- Correlation Heatmap: Fare and Pclass moderately correlated with survival.
- Pairplot: Clustering based on Sex and Pclass.
- Boxplots: Higher fare linked to better survival.

Titanic Dataset - Exploratory Data Analysis Report

Multivariate Analysis:

- Females in 1st class had the highest survival rate.
- Males in 3rd class had the lowest survival rate.

Key Observations

1. Gender: Females had a higher survival rate than males.
2. Class: 1st class passengers survived more.
3. Fare: Higher fare-paying passengers had better survival chances.
4. Age: Younger passengers had slightly better survival odds.
5. Embarkation Port: Majority embarked from Southampton ('S').

Conclusion

The EDA highlights that gender, passenger class, and fare significantly influenced survival chances on the Titanic. These insights can guide further predictive modeling and feature engineering.