Titanic Exploratory Data Analysis (EDA) Report

Objective:

Analyse the Titanic dataset using statistical and visual exploration techniques to uncover insights related to survival, passenger demographics, and relationships among variables.

Step 1: Setup and Load Data

Observation:

- The training dataset contains 891 rows and 12 columns.
- Some columns contain missing values, especially Age, Cabin, and Embarked.

Step 2: Summary Statistics

Observation:

- Cabin has a large number of missing values.
- Age and Embarked also have some missing entries.
- Columns like Name, Ticket, and Cabin are non-numeric and may need preprocessing later.

Step 3: Univariate Analysis

A. Target Variable – Survived

Observation:

- About 38% of passengers survived, while 62% did not.

B. Categorical Features

Observations:

- Pclass: 1st class had the highest survival rate, 3rd class the lowest.
- Sex: Females had much higher survival rates than males.
- Embarked: Passengers who boarded at Cherbourg ('C') had better survival rates.

C. Numerical Features

Observations:

- Age: Right-skewed, most passengers were 20–40 years old.
- Fare: Highly skewed, some fares are extremely high.

- SibSp and Parch: Most passengers were alone or with one family member.

Step 4: Bivariate Relationships

A. Boxplots for Age and Fare vs Survival

Observations:

- Survivors tended to be slightly younger.
- Survivors generally paid higher fares.

B. Correlation Heatmap

Observation:

- Fare and Pclass are moderately correlated with survival.
- SibSp and Parch are also positively correlated.

C. Pairplot (Selected Columns)

Observation:

- A visible separation in Fare and Pclass among survivors.
- Survivors tend to be clustered in higher Fare and lower Pclass.

Summary of Key Insights:

1. Sex: Female passengers were significantly more likely to survive.
2. Pclass: Passengers in 1st class had the highest survival rates.
3. Fare: Higher ticket fare was positively associated with survival.
4. Age: Children had slightly higher survival rates, though not as pronounced as sex or class.
5. Embarked: Passengers from Cherbourg (C) had higher survival rates.
6. Missing Data:
    - Age and Cabin have missing values.
    - Cabin is likely too sparse for useful analysis without imputation or simplification.