Task-5-EXPLORATORY DATA ANALYSIS (EDA)

Name - Bishal Dey

Tool used - Jupyter Notebook

Title - Extract insights using visual and statistical exploration.

Dataset - Titanic Dataset (train.csv)

DATASET OVERVIEW

Metric	Value
Total Rows	891
Total Columns	12
Target Variable	Survived (0 = No, 1 = Yes)
Missing Columns	Age, Cabin, Embarked

Data Types Summary

- **Numerical Features**: Age, Fare, SibSp, Parch
- **Categorical Features**: Sex, Pclass, Embarked
- **Textual Features**: Name, Ticket, Cabin

REPORT OF FINDINGS

Key Univariate Insights

- **Age**:
  - Distribution is right-skewed with most passengers between 20-40 years.

    - Some missing values.
- **Fare**:
    - Highly right-skewed with outliers.
    - Most passengers paid under 100.
- **Survived**:
    - 61.6% did not survive, 38.4% survived.
- **Pclass**:
    - Majority in class 3 (low class), indicating economic disparity.
- **Sex**:
    - ~65% male, ~35% female.

Visuals Used

- Histograms (Age, Fare, SibSp, Parch)
- Boxplots (Fare vs Survival)
- Countplots (Survived vs Sex, Pclass)
- Heatmap (Correlation Matrix)
- Pairplot (Age, Fare, Pclass, Survived)
-

Observations

- The survival rate of females is significantly higher than males.

- Younger passengers had a slightly higher survival chance.

- Passengers who paid higher fares had better survival rates.

Bivariate & Multivariate Insights

➤ Survival by Sex

- **Females** had a **much higher survival rate** than males.
- Clear gender-based survival disparity.

➤ Age vs Survival

- Children (age <10) had better survival rates.
- Young adults (20–40) had lower survival probability.

➤ Pclass vs Survival

- Higher class passengers had better survival odds:
    - Class 1: High survival.
    - Class 3: Lowest survival.

➤ Fare vs Survival

- Passengers who paid **higher fares** had a **better chance** of survival.

FINAL SUMMARY

Sex and Pclass are the most important predictors for survival.

Females, children, and high-class passengers had higher survival rates.

Fare and Age impact survival moderately.

There are missing values in Age, Cabin, and Embarked.

Data skewness and outliers are present (especially in Fare).

Correlation heatmap reveals meaningful trends without severe multicollinearity.