REPORT- TASK 5

Exploratory Data Analysis (EDA) Report

Dataset: Titanic Survival Data
Objective: Extract insights using visual and statistical exploration.

Step 1: Tools Used

- **Python Libraries:**
  - Pandas – for data manipulation
  - Matplotlib & Seaborn – for visualizations

Step 2: Dataset Overview

- Total entries: **891**
- Features: **12**
- Target variable: **Survived** (1 = survived, 0 = did not survive)

Step 3: Statistical Exploration

df.info()

- Found 3 columns with **missing data**:
  - Age: 177 missing
  - Cabin: 687 missing
  - Embarked: 2 missing
- Majority of columns are numerical, with a few categorical (Sex, Embarked, Cabin)

df.describe()

- **Mean age:** ~29.7 years
- **Fare** ranges widely from 0 to 512 (potential outliers)
- **Pclass** median is 3 → many passengers were in third class

df.isnull().sum()

- High missing rate in Cabin (~77%) → may be dropped or filled with a placeholder
- Moderate missing rate in Age (~20%) → could use median or model-based imputation

df['Survived'].value_counts()

- **549 (61.6%)** passengers did not survive
- **342 (38.4%)** survived
- Class imbalance noted (important for ML model training)

Step 4: Visual Exploration

Univariate Analysis

- **Age Distribution**: Normal distribution with a concentration between 20–40 years
- **Survival Distribution**: More non-survivors than survivors

Bivariate Analysis

- **Survival vs Sex**:
  - **Females had significantly higher survival rates** than males
- **Survival vs Pclass**:
  - **First-class passengers** had better survival rates
- **Survival vs Age**:
  - **Younger passengers** had slightly better survival rates
- **Survival vs Fare**:
  - Passengers who paid **higher fares** were more likely to survive

Multivariate Analysis

- **Correlation Heatmap**:
  - Fare and Pclass are moderately correlated
  - Survived is correlated with Pclass, Sex, and Fare
- **Pairplot**:
  - Visual separation seen in Fare and Pclass for survivors vs non-survivors

---

Step 5: Summary of Findings

1. **Sex** is a strong predictor of survival — females were prioritized during evacuation.
2. **Pclass** affects survival — higher-class passengers had better survival rates.
3. **Fare** is positively associated with survival — reflects socio-economic bias.
4. **Age** shows a slight trend — younger passengers had a better chance.
5. **Missing data** needs to be handled carefully, especially in Age and Cabin.
6. **Imbalanced classes** (more deaths than survivors) should be addressed in predictive modelling.
