Task 5

Titanic Dataset Exploratory Data Analysis Report

1. Introduction

This project performs an Exploratory Data Analysis (EDA) on the Titanic dataset using Python libraries such as pandas, numpy, matplotlib, and seaborn. The goal is to identify patterns and relationships in the data related to passenger survival rates based on demographic and ticket-related variables.

2. Initial Data Inspection

- Displayed the first few rows using `df.head()`
- Inspected data types and missing values using `df.info()`
- Described statistical distribution with `df.describe()`
- Counted missing values and examined class, sex, embarked ports, and survival distributions.

3. Data Cleaning Steps

- Filled missing 'Age' values with the median.
- Dropped the 'Cabin' column due to too many missing values.
- Filled missing 'Embarked' values with the mode.

4. Visualizations

4.1 Distribution Plots

- Histogram of `Age` and `Fare` to observe distribution.

4.2 Boxplots

- Age vs Survival
- Fare vs Passenger Class

4.3 Countplots

- Survival rate by Sex
- Survival rate by Passenger Class

4.4 Scatterplot

- Age vs Fare colored by Survival

4.5 Correlation Heatmap

- Heatmap showing correlation between numeric variables like `Age`, `Fare`, `Pclass`, `SibSp`, `Parch`, and `Survived`.

4.6 Pairplot

- Pairwise relationships among selected numerical columns, segmented by survival.

5. Conclusion

Summary of Key Findings from EDA

1. **Gender was a paramount factor in survival:** A significantly higher proportion of females survived compared to males across all passenger classes. This strongly suggests the 'women and children first' protocol was largely followed.
2. **Passenger Class greatly influenced survival:** First-class passengers had the highest survival rates, while third-class passengers experienced the lowest survival rates and the highest number of fatalities. This indicates a clear disparity in survival chances based on socioeconomic status or access to lifeboats.
3. **Age Distribution and Survival:** The majority of passengers were young adults. While the median age for survivors and non-survivors was similar, a notable observation is that children (younger ages) generally had a better chance of survival, aligning with the 'women and children first' protocol.
4. **Fare and Survival:** Passengers who paid higher fares, which are strongly correlated with first-class tickets, had a higher probability of survival. This reinforces the advantage held by higher-class passengers.
5. **Correlations:** There is a strong negative correlation between 'Fare' and 'Pclass' (higher class, lower Pclass number, higher fare). 'Survived' shows some correlation with 'Sex', 'Pclass', and 'Fare'.
6. **Missing Data Handled:** Missing 'Age' values were imputed with the median, and the 'Cabin' column was dropped due to extensive missing data.