Exploratory Data Analysis (EDA) Report

Submitted By: Ainsley Sveta A

Dataset: Titanic Survival Dataset

Tools Used: Python, Pandas, Matplotlib, Seaborn

1. Objective

To perform exploratory data analysis on the Titanic dataset in order to identify key variables influencing passenger survival. This involves uncovering patterns, correlations, and trends through descriptive statistics and visualizations.

2. Dataset Overview

- Total Rows: 891
- Total Columns: 12
- Target Variable: Survived (0 = No, 1 = Yes)
- Missing Data:
    - Age: 177 missing values
    - Cabin: 687 missing values
    - Embarked: 2 missing values

3. Key Variables

- Survived: Whether the passenger survived (binary)
- Pclass: Ticket class (1st, 2nd, 3rd)
- Sex: Gender of the passenger
- Age: Age in years
- SibSp/Parch: Family aboard
- Fare: Fare paid
- Embarked: Port of embarkation

4. Univariate Analysis

4.1 Categorical Features:

- Sex: ~65% male, ~35% female
- Pclass: Majority of passengers belonged to 3rd class
- Embarked: Most passengers embarked from 'S' (Southampton)

4.2 Numerical Features:

- Age: Wide distribution, average age ~29.7
- Fare: Highly skewed with a few outliers; median ~14.45
- SibSp/Parch: Most passengers traveled alone or with one relative

5. Bivariate Analysis

5.1 Survival by Gender

- Survival rate for females was significantly higher than for males.

5.2 Survival by Class

- 1st class passengers had the highest survival rate.
- Survival rates decreased from 1st to 3rd class.

5.3 Age vs Survival

- Young children had slightly better survival chances.
- Survivors' median age was lower than non-survivors.

6. Correlation Analysis

- Positive Correlations with Survival:
    - Fare: Higher fare indicates higher chance of survival
    - Pclass: Negative correlation (lower class, lower survival)
- Weak Correlation with Age, SibSp, and Parch

7. Visualizations Used

- Histograms for Age, Fare, SibSp, Parch

- Boxplots for Age vs Survived
- Countplots for Survived by Sex and Class
- Heatmap for correlation matrix
- Pairplot for multivariate relationships

8. Key Insights

- Female passengers had a much higher survival probability than males.
- Higher-class passengers (1st class) were more likely to survive.
- Passengers who paid higher fares tended to survive more.
- Age and family size (SibSp + Parch) had moderate effects.

Conclusion

The EDA reveals that survival on the Titanic was strongly influenced by gender, class, and fare. These findings provide a foundation for building machine learning models and deeper statistical analysis.