Exploratory Data Analysis on Titanic Dataset

Rahul Garg

Objective

The objective of this project is to perform Exploratory Data Analysis (EDA) on the Titanic dataset using Python (Pandas, Matplotlib, Seaborn) to discover patterns, trends, and anomalies related to passenger survival.

Dataset Description

The dataset used is the classic Titanic dataset, consisting of three files:

- **train.csv**: Training data used for EDA (891 records).
- **test.csv**: Test data for prediction (not used here).
- **gender_submission.csv**: Sample output format.

The **train.csv** file includes the following key columns:

- `Survived`: Target variable (0 = No, 1 = Yes)
- `Pclass, Sex, Age, SibSp, Parch, Fare, Embarked`: Features

Tools Used

- Python 3.10
- Pandas, Matplotlib, Seaborn
- Jupyter Notebook

Initial Exploration

Missing Values

- **Age**: 177 missing values
- **Cabin**: 687 missing values (mostly empty)
- **Embarked**: 2 missing

Survival Counts

- **Not Survived (0)**: 549 passengers
- **Survived (1)**: 342 passengers

Survival Percentages

- **Survived**: 38.4%
- **Not Survived**: 61.6%

Univariate Analysis

- **Sex**: More males than females; survival higher among females.
- **Pclass**: Class 1 passengers had a higher survival rate.
- **Age**: Distribution shows most passengers are young adults.
- **Fare**: Positively skewed; higher fares correlate with higher survival.

Visualizations

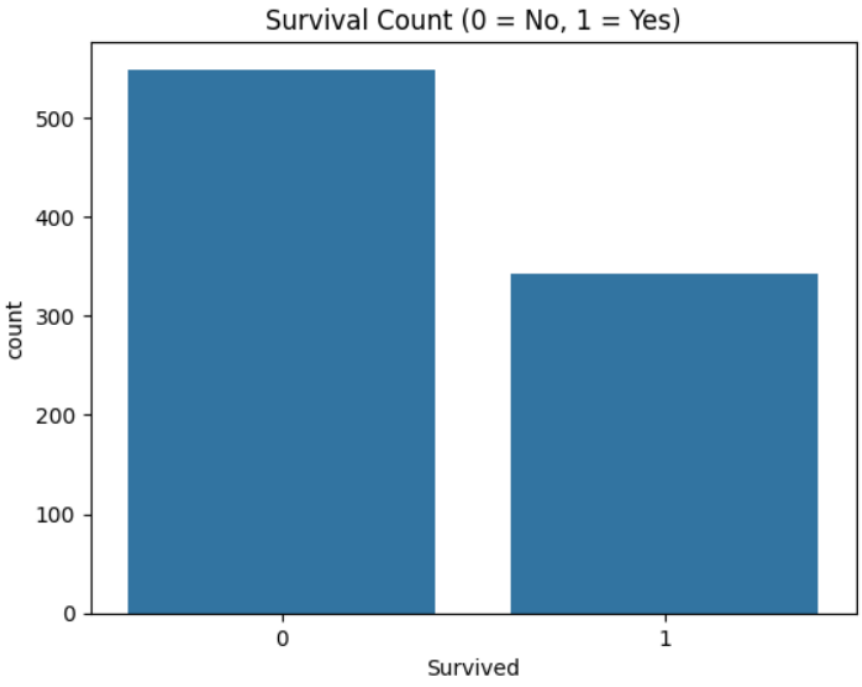

Figure 1: Count of Survivors vs Non-Survivors

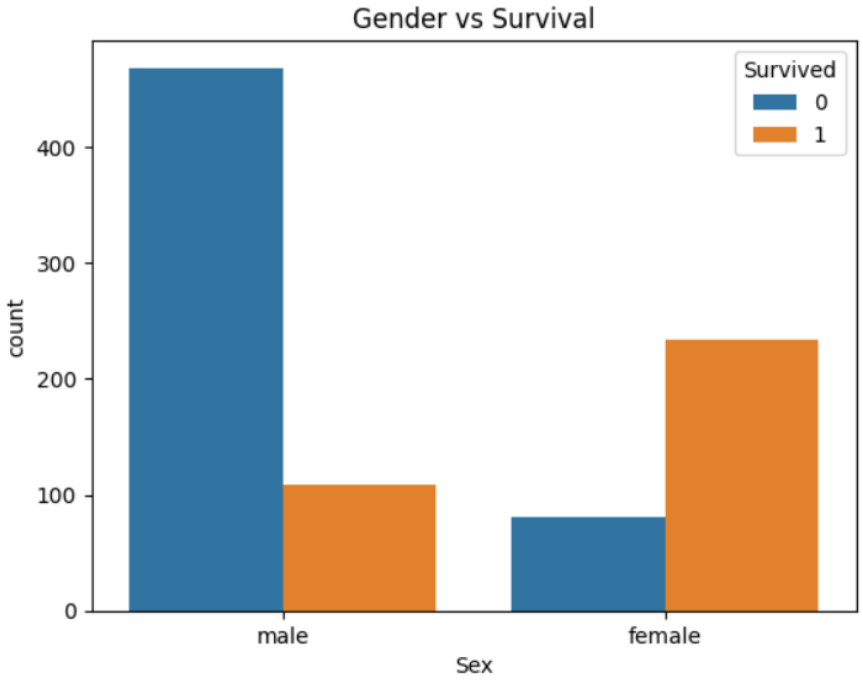

Figure 2: Survival by Gender

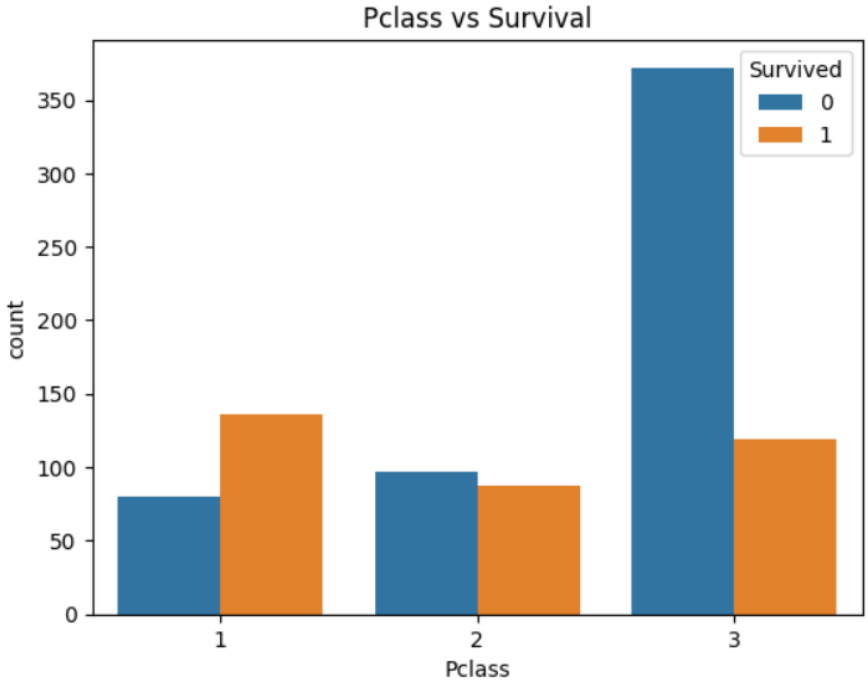

Figure 3: Survival by Passenger Class

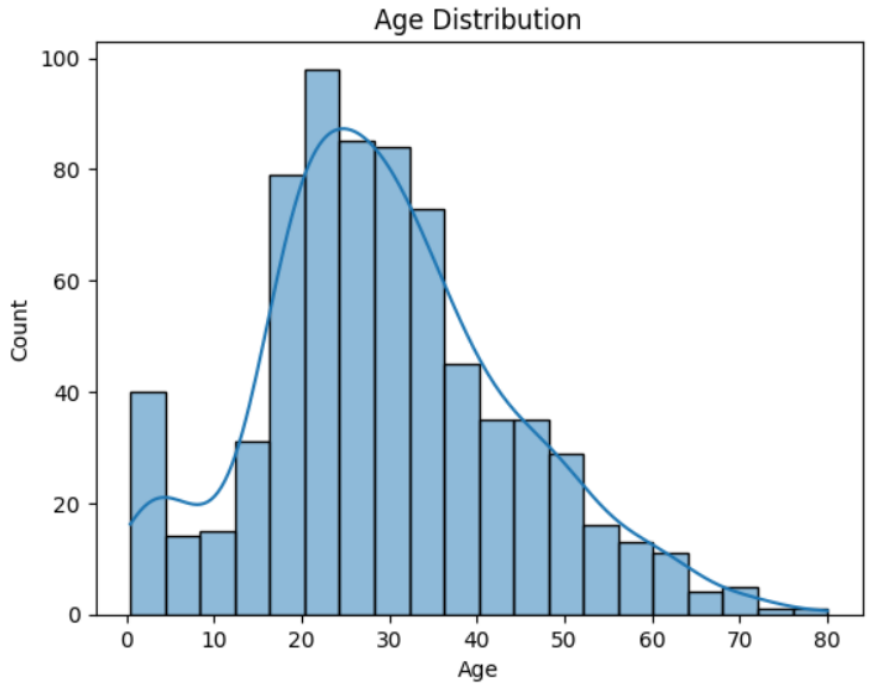

Figure 4: Age Distribution

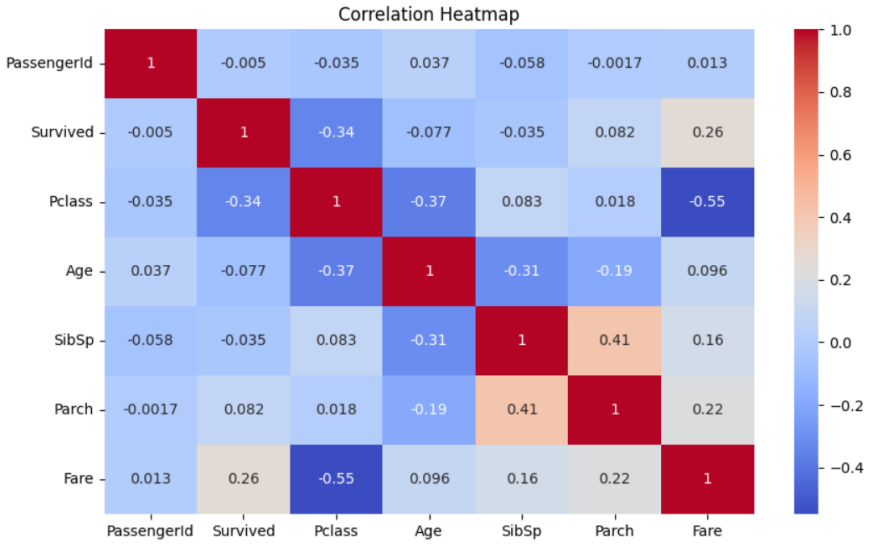

Figure 5: Correlation Heatmap

Multivariate Insights

- Women in higher classes had the highest survival.
- Pclass 3 male passengers had the lowest survival.
- Young children had higher survival rates.

Handling Missing Values

- **Age**: Filled with median age.
- **Embarked**: Filled with mode.
- **Cabin**: Dropped (too many missing values).

Conclusion

Through this EDA:

- We identified clear trends in survival related to gender, class, and age.
- Visualizations helped uncover correlations and potential features for future modeling.
- The dataset is now prepared for machine learning analysis.