Exploratory Data Analysis of the Titanic Dataset

1. Introduction

1.1 Purpose of the Analysis

This Exploratory Data Analysis (EDA) aims to uncover patterns and relationships within the Titanic dataset, with particular focus on identifying the key factors that influenced passenger survival. Insights obtained through statistical and visual analysis will be useful for guiding future predictive modeling efforts and enhancing data-driven decision-making.

1.2 Dataset Description

The Titanic dataset provides detailed information about the passengers aboard the RMS Titanic. The data includes demographic details (such as age, sex, and passenger class), ticket and fare information, family structure, and survival status. Due to its rich set of features, it is commonly used for classification tasks and survival analysis.

2. Data Overview

An initial inspection using `.info()` gives an overview of the dataset structure, data types, and non-null counts across features.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
             count unique                 top freq       mean         std  \
PassengerId  891.0    NaN                 NaN  NaN      446.0  257.353842
Survived     891.0    NaN                 NaN  NaN   0.383838    0.486592
Pclass       891.0    NaN                 NaN  NaN   2.308642    0.836071
Name           891    891  Dooley, Mr. Patrick    1        NaN         NaN
Sex            891      2                male  577        NaN         NaN
...
S    644
C    168
Q     77
Name: count, dtype: int64
```

3. Missing Value Analysis

3.1 Visual Inspection

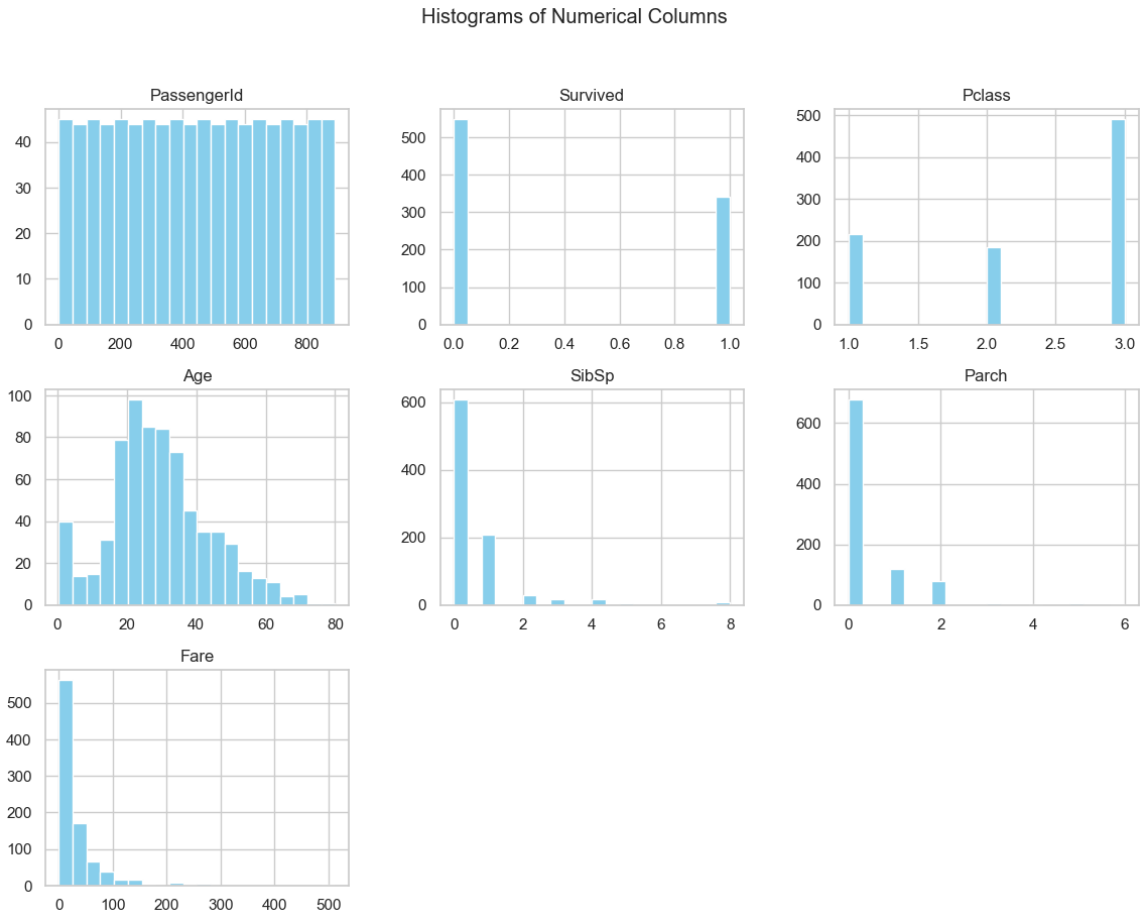
A heatmap generated using Seaborn highlights columns with missing values, most notably:
- Age
- Cabin
- Embarked

3.2 Observations

- Age and Cabin have significant portions of missing data.
- Cabin has the highest number of missing entries and may either be dropped or carefully imputed.
- Embarked has few missing values and can be filled with the most frequent value.

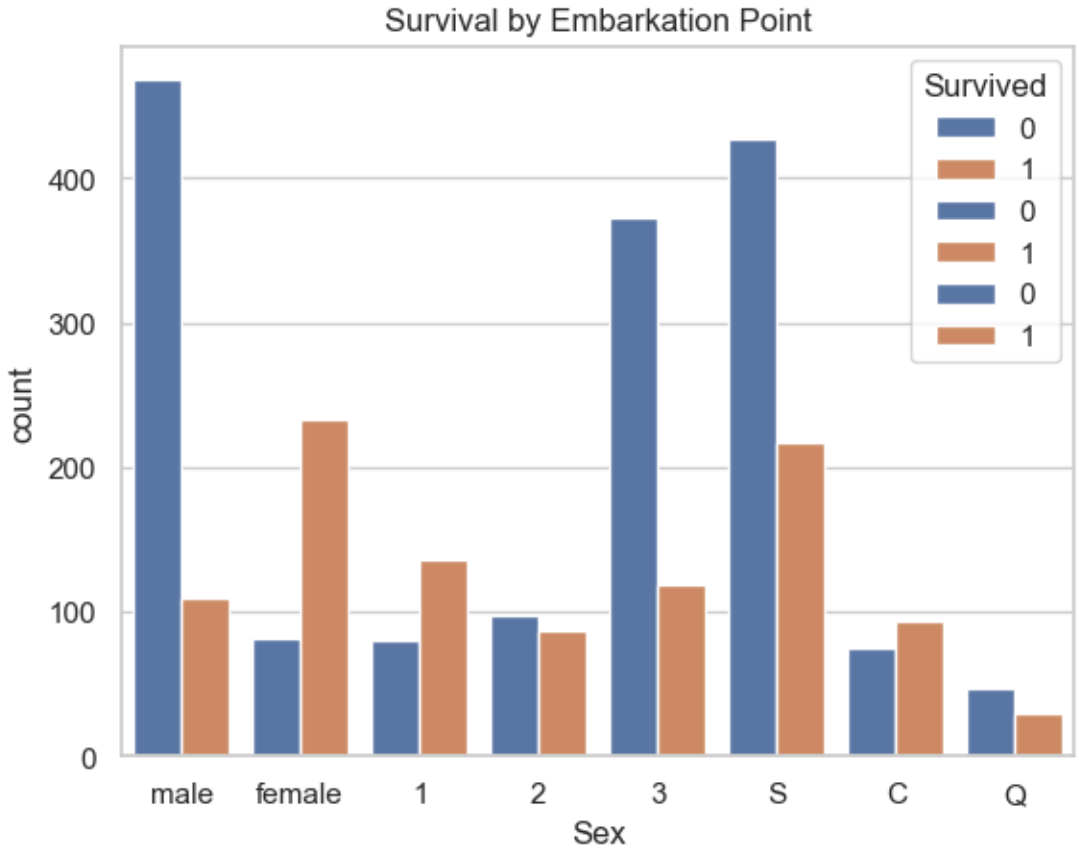
4. Univariate Analysis

4.1 Histograms



- Age: Right-skewed distribution with a concentration of passengers in their 20s and 30s.
- Fare: Highly skewed, with a long tail representing a few extremely high fares.

4.2 Countplots



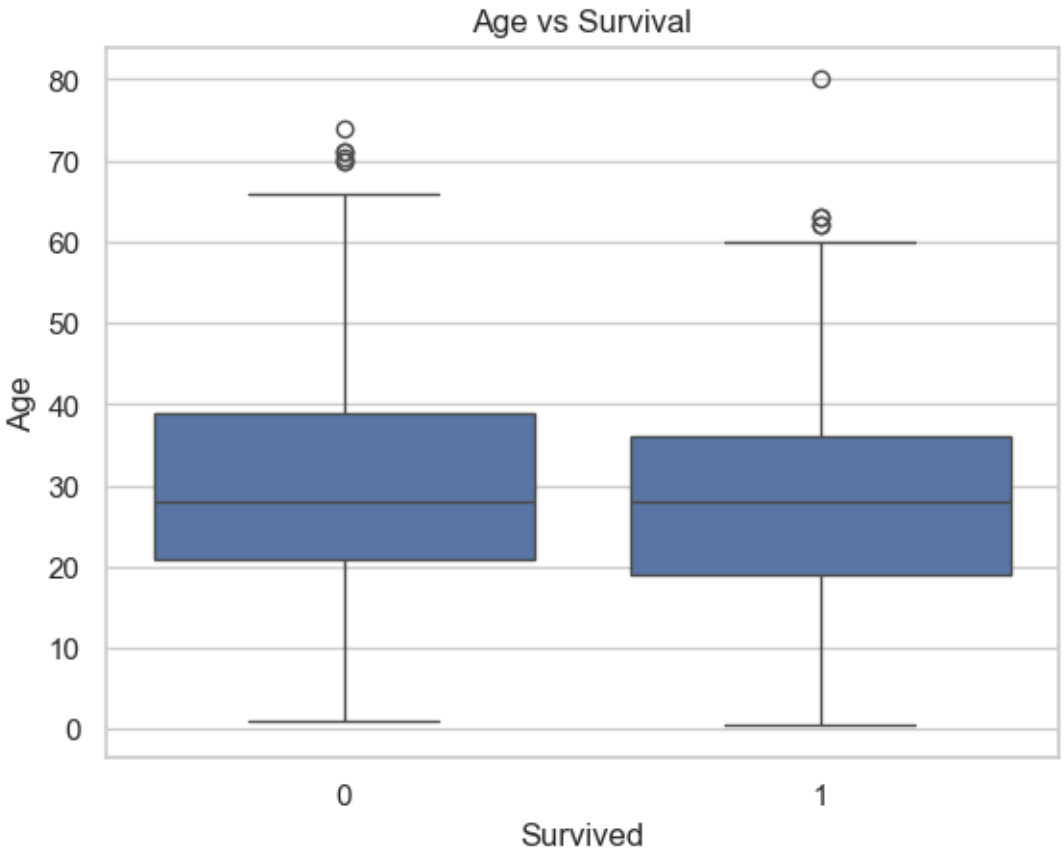
- Gender: More males than females on board.
- Survival: Higher survival rate observed among females.
- Class: Most passengers belonged to third class.

4.3 Observations

- Survival was not uniform across gender and class.
- Fare distribution is uneven, with a small subset of passengers paying very high fares.

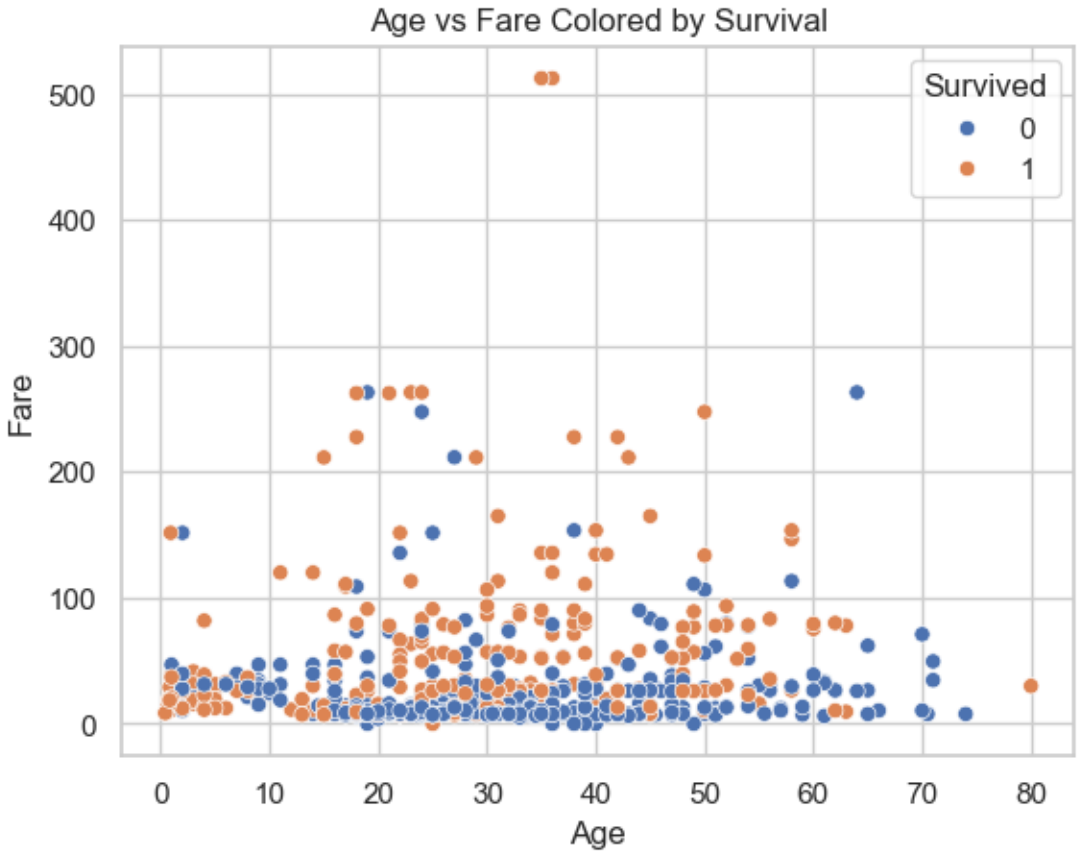
5. Bivariate Analysis

5.1 Boxplots



- Age vs. Survival: Survivors tended to be younger.
- Fare vs. Class: Higher-class passengers paid significantly more.

5.2 Scatterplots



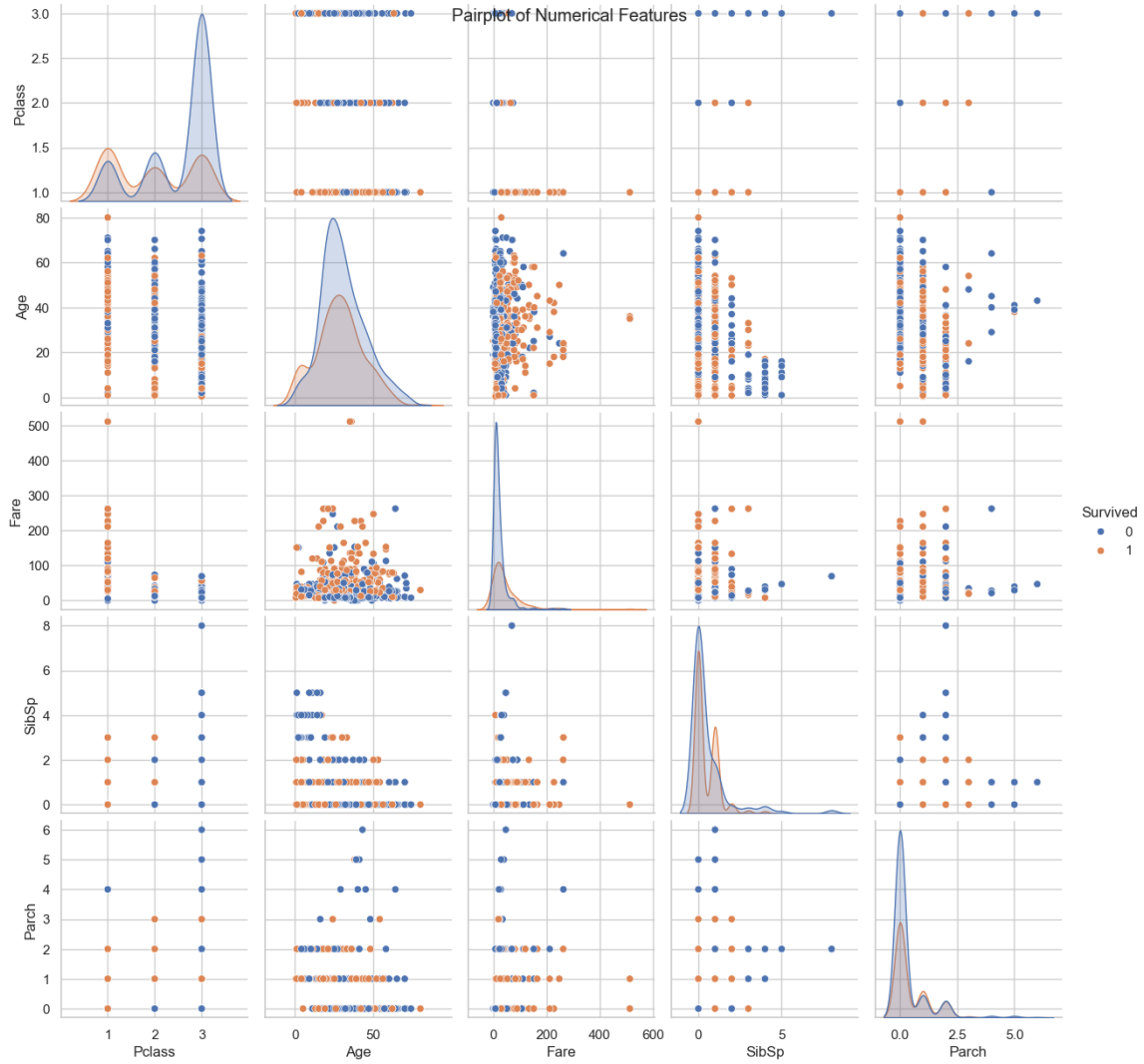
Scatterplots were used to examine relationships between numeric variables, such as:
- Age vs. Fare
- Age vs. SibSp

5.3 Observations

- First-class passengers generally paid higher fares and had better survival rates.
- Several outliers were noted in both the age and fare variables.

6. Multivariate Analysis

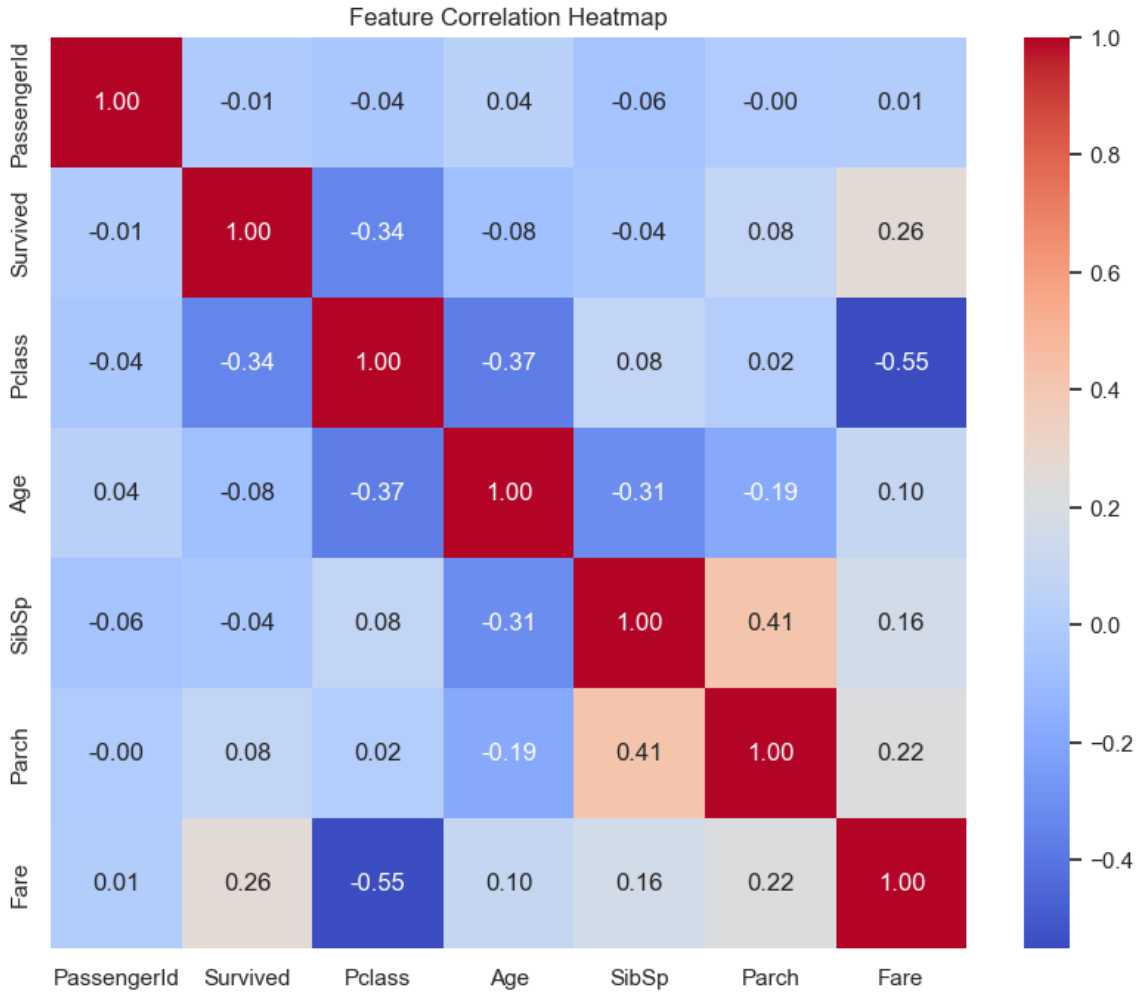
6.1 Pairplot (sns.pairplot)



This visualizes pairwise relationships across several numeric features, revealing trends among:

- Fare
- Age
- Survival

6.2 Correlation Matrix (sns.heatmap)



The correlation heatmap shows:
- A strong negative correlation between Fare and Pclass.
- A positive correlation between Survival and both Pclass (higher class) and Sex (female).

6.3 Observations

- Pclass, Sex, and Fare are among the strongest indicators of survival.
- Features such as SibSp and Parch show weaker, but potentially relevant, influence.

7. Summary of Findings

7.1 Key Trends

- Female passengers had significantly higher survival rates.
- First-class passengers were more likely to survive.
- Younger passengers showed better odds of survival.

7.2 Anomalies

- Some passengers paid unusually high fares.
- A few older passengers survived, defying the general age trend.

7.3 Noteworthy Insights

- Family size (derived from SibSp and Parch) might have influenced survival.
- The Embarked variable showed slight variation in survival rates, with Cherbourg (C) having a higher rate.