1. Introduction

Purpose of the Analysis

The goal of this exploratory data analysis (EDA) is to understand patterns and relationships in the Titanic dataset, with a focus on the factors that influenced passenger survival. Through visual and statistical techniques, we aim to gain insights that could guide predictive modeling or decision-making.

Description of the Dataset

The Titanic dataset contains data about passengers aboard the Titanic, including demographic features (age, sex, class), ticket and fare information, family size, and whether the passenger survived. It is a widely used dataset for classification and survival prediction tasks.

2. Data Overview

.info()

```
<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

Missing Data Heatmap

A heatmap from seaborn helps visually identify columns with missing values, such as Age, Cabin, and Embarked.

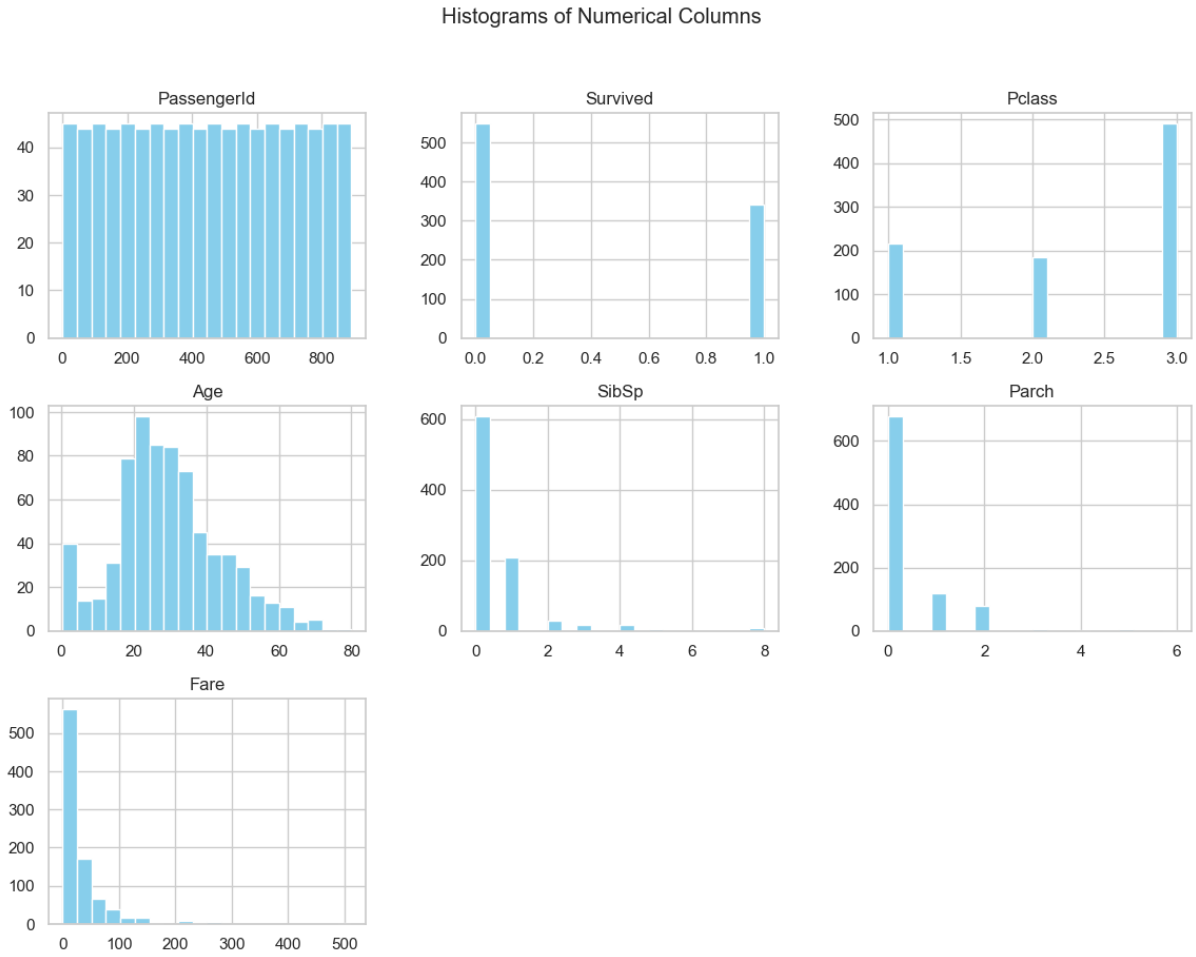
Observations

- Age and Cabin have significant missing values.
- Cabin has the most missing data — it may be dropped or imputed with care.
- Embarked has a few missing values and can be imputed with the most frequent port.

---

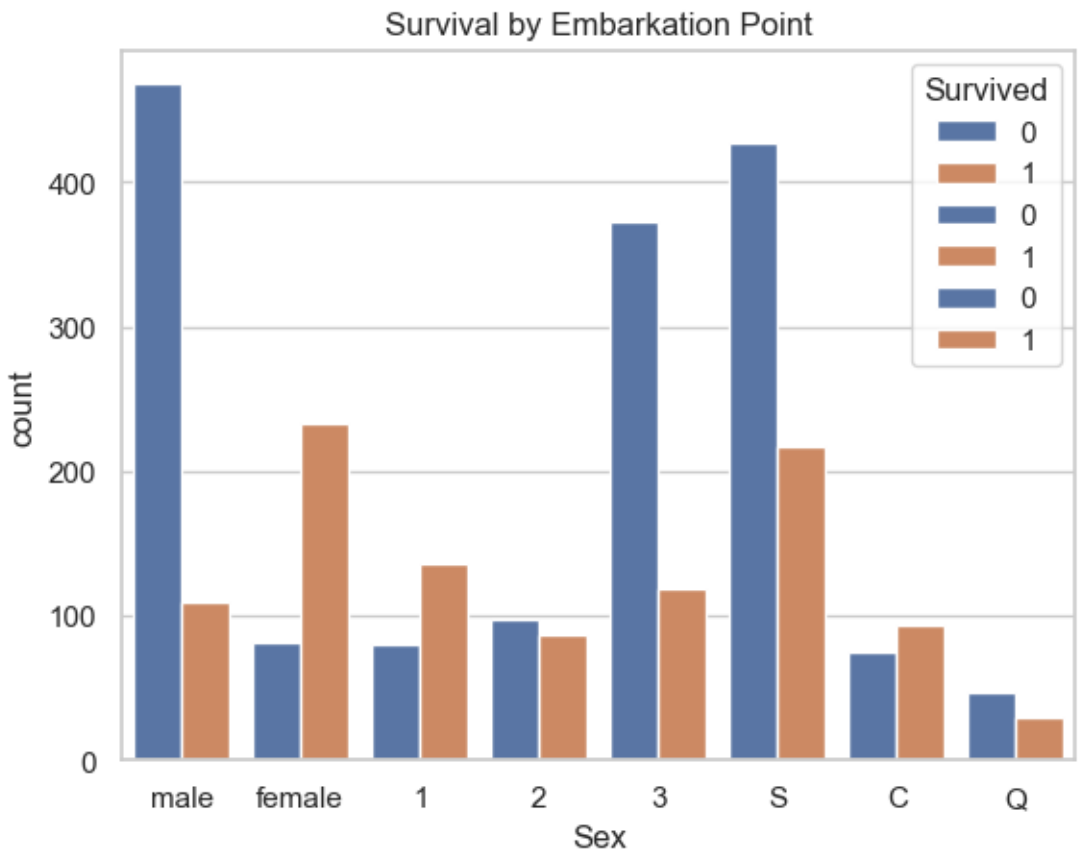
4. Univariate Analysis

Histograms (age, fare, etc.)



- Age is right-skewed; many passengers are in their 20s and 30s.
- Fare is highly skewed with a few very large values.

Countplots (gender, survival, class)



- More males than females on board.
- Higher survival rate for females.
- Most passengers were in third class.

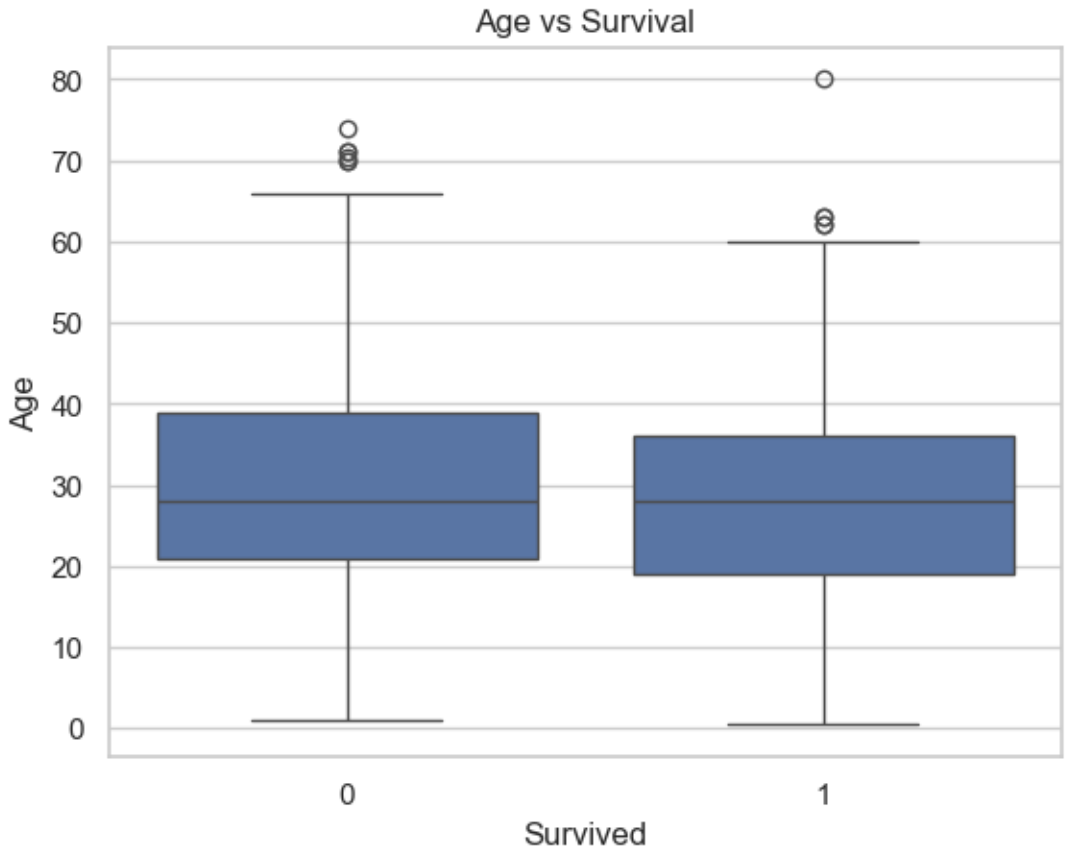
Observations

- Survival is not uniformly distributed across gender and class.
- Fare distribution shows a long tail — some passengers paid significantly more.

---

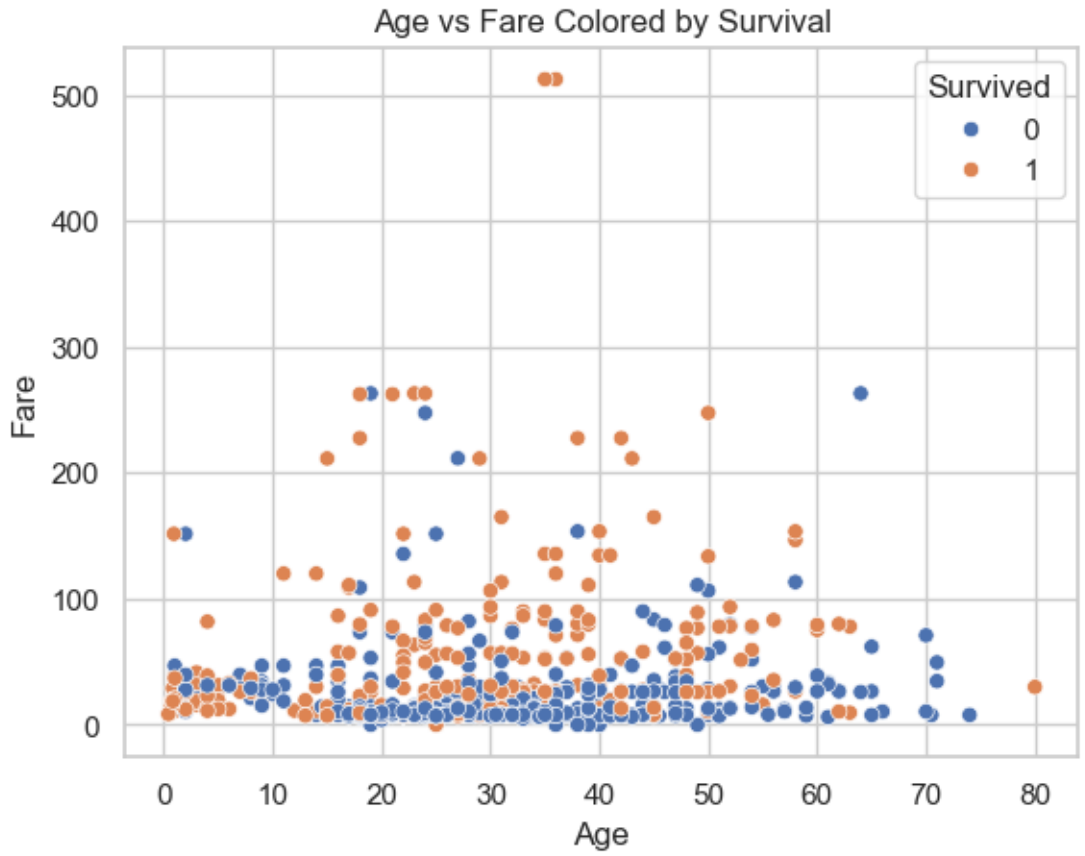
5. Bivariate Analysis

Boxplots (age vs survived, fare vs class)



- Survivors tend to be younger.
- Higher-class passengers paid more on average.

Scatterplots



- Useful to explore relationships between numerical variables (e.g., age vs. fare, age vs. sibsp).

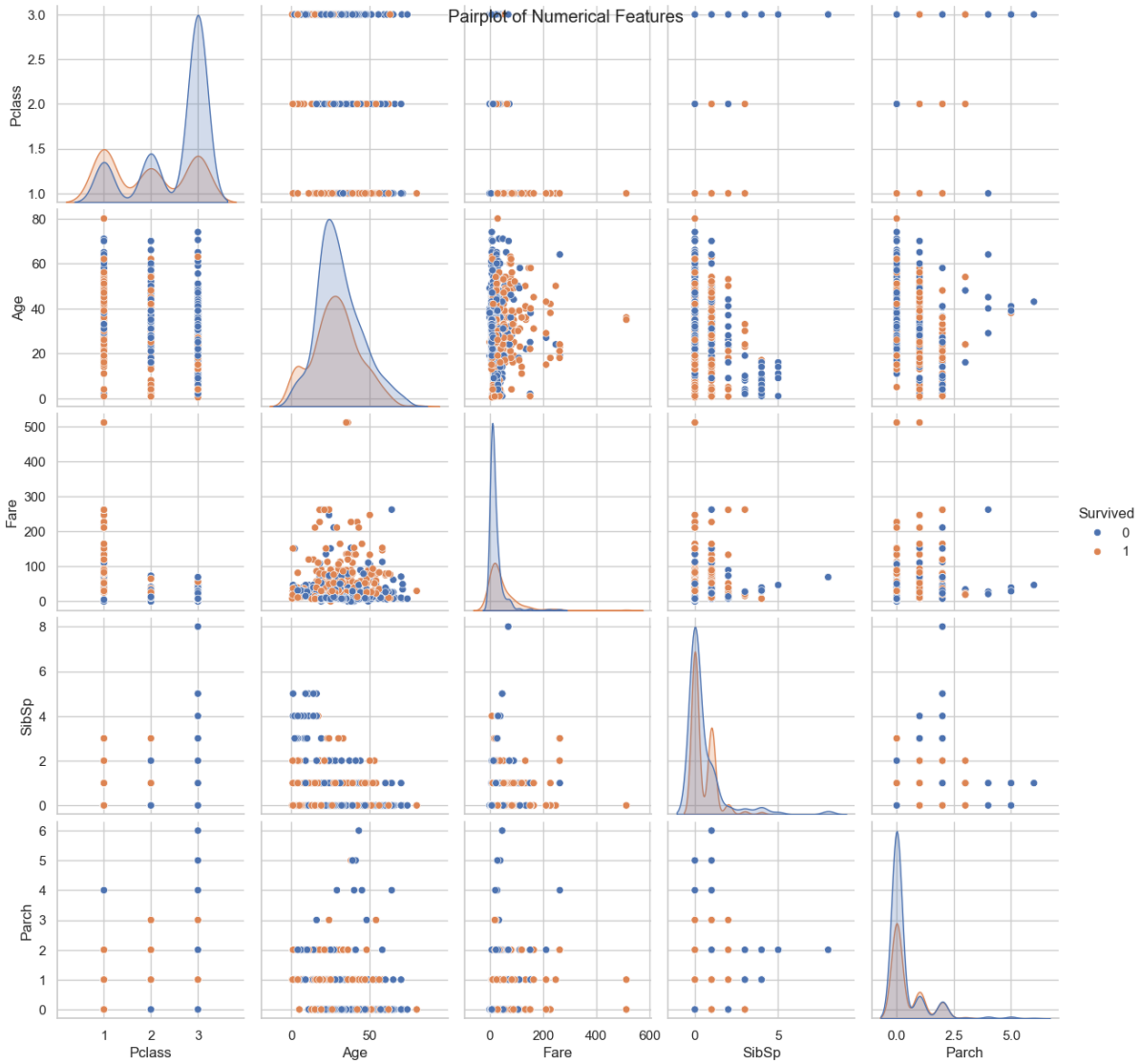
Observations

- First-class passengers generally have higher fares and better survival rates.
- Some outliers exist in both age and fare.

---

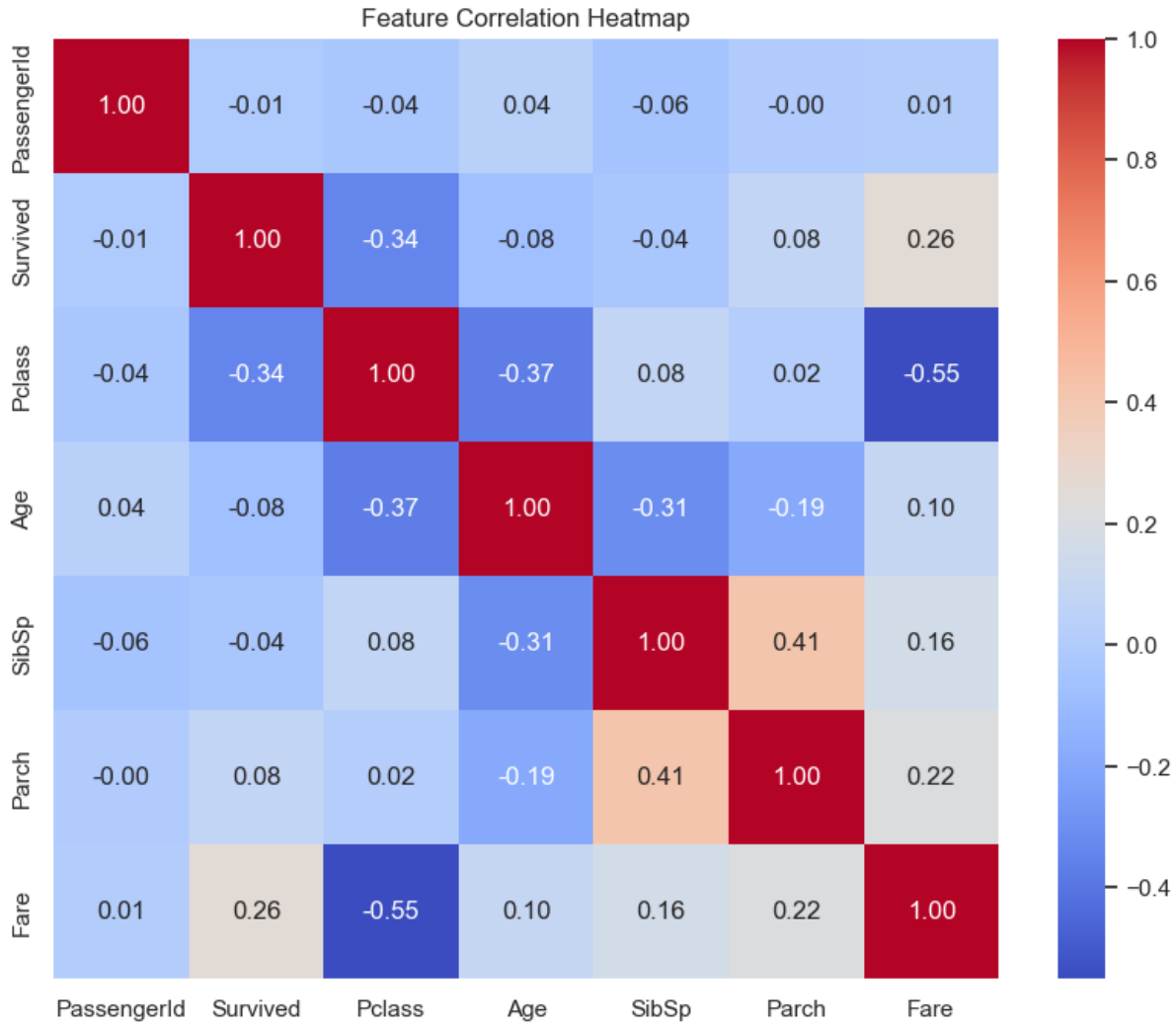
6. Multivariate Analysis

sns.pairplot()



- Shows pairwise relationships between multiple numeric variables.
- Trends between fare, age, and survival start to emerge.

sns.heatmap() (correlation)



- Shows correlation matrix between numerical variables.
- Fare and class have strong negative correlation.
- Survival has positive correlation with being in higher class and being female.

Observations

- Some variables like Pclass, Sex, and Fare show a stronger link to survival.
- Features like SibSp and Parch may have some influence, but are weaker indicators.

7. Summary of Findings

Key Trends

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

Anomalies

- Some passengers paid extremely high fares.
- A few older passengers also survived, contrary to the general age trend.

Interesting Insights

- Family size (combination of SibSp and Parch) may play a role in survival.
- Embarkation point may influence survival slightly, with C (Cherbourg) having a higher rate.