Task : Exploratory Data Analysis (EDA)

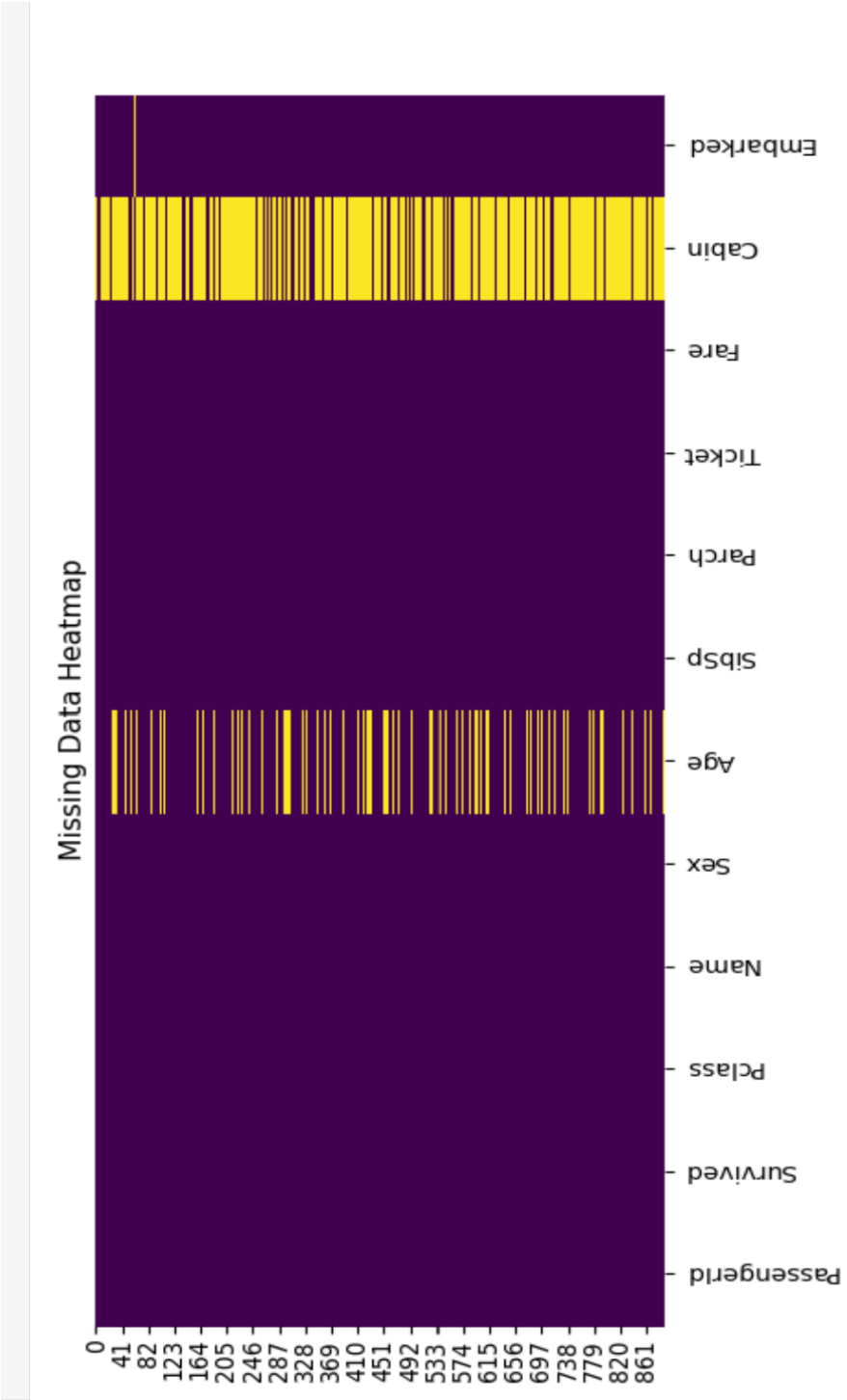
Objective: Extract insights using visual and statistical exploration.
Tools: Python (Pandas, Matplotlib, Seaborn)

Observation 1:

Number of rows and columns
Missing values (e.g., Age, Cabin)
Categorical distributions (male/female, embarked locations)
Balance of target variable (`Survived`)

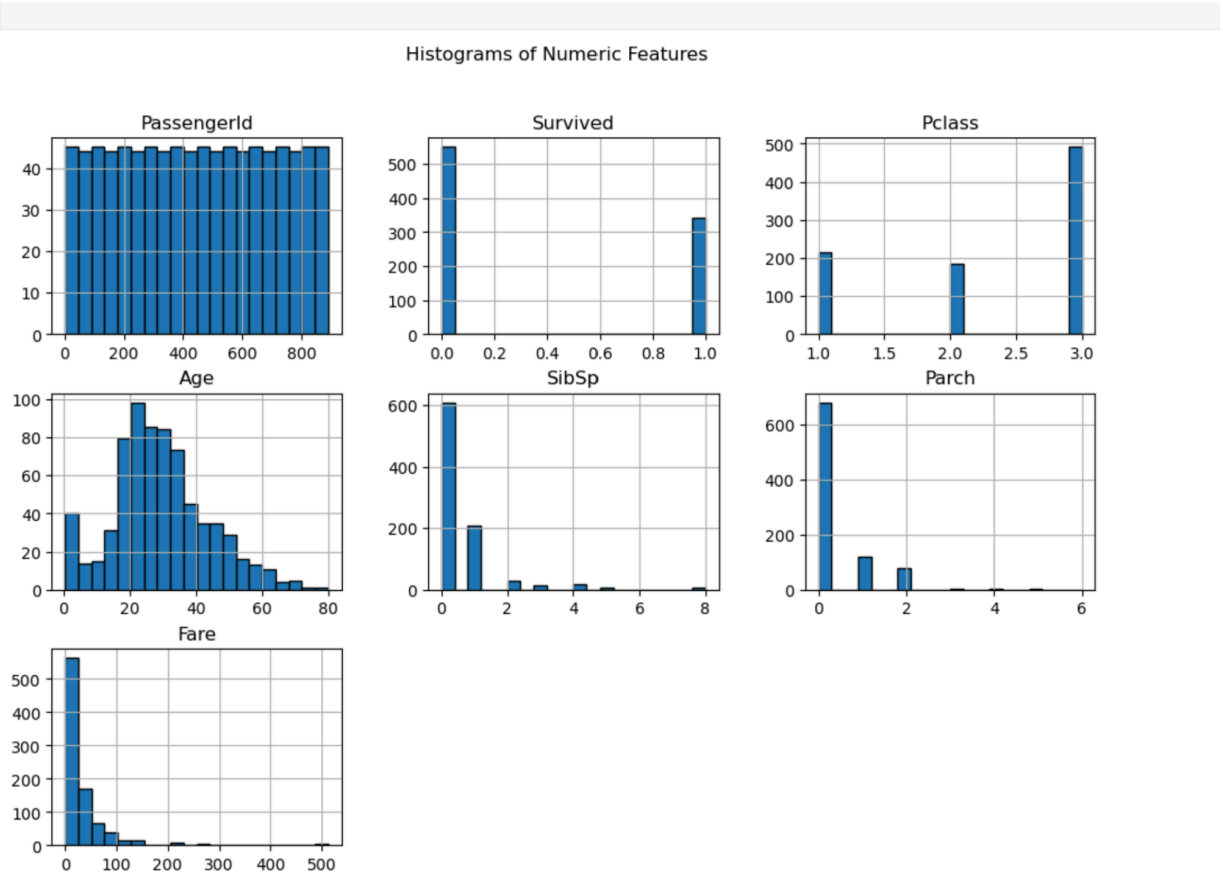
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    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

Value counts for Survived:
Survived
0    549
1    342
Name: count, dtype: int64

Value counts for Pclass:
Pclass
3    491
1    216
2    184
Name: count, dtype: int64

Value counts for Sex:
Sex
male      577
female    314
Name: count, dtype: int64

Value counts for Embarked:
Embarked
S      644
C      168
Q       77
NaN      2
Name: count, dtype: int64
```

Observation 2 : Identifying which columns need imputation or can be ignored, Cabin has way many missing values as compared to Age.



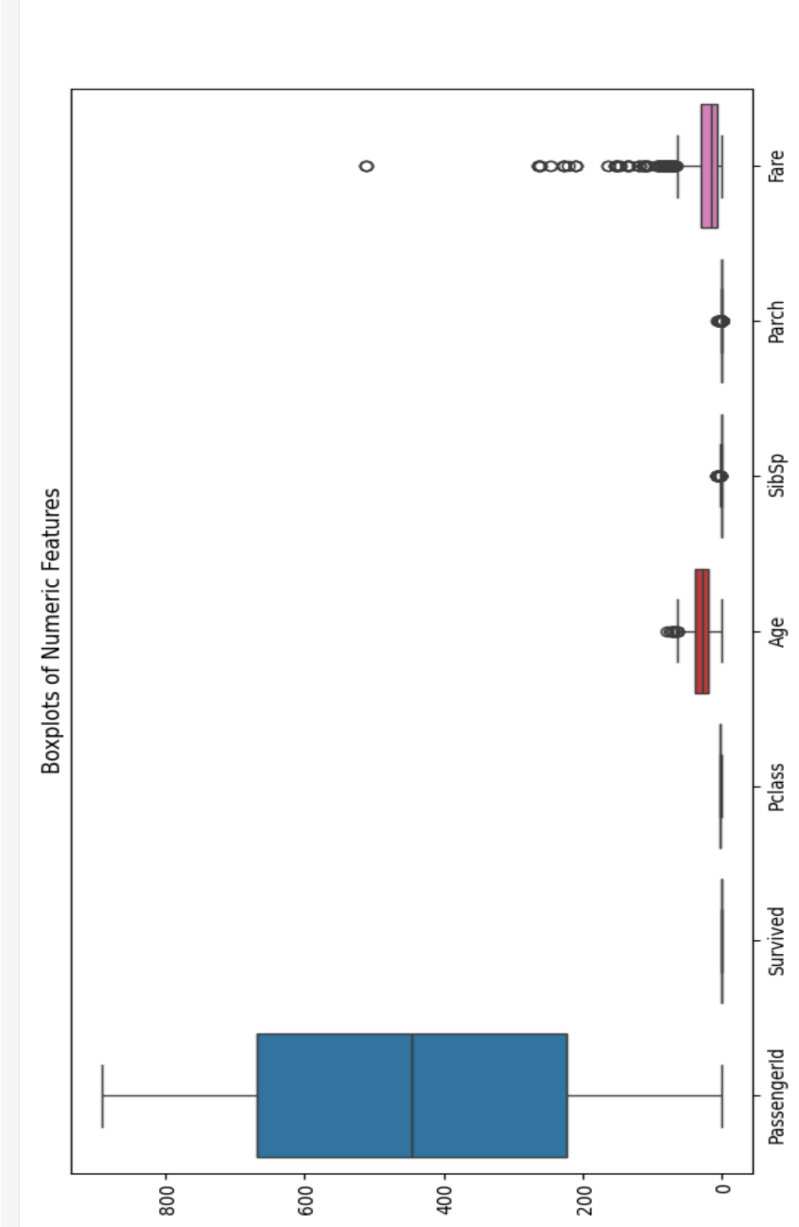
Observation 3 :
PassengerId: Distributed fairly evenly, just a unique identifier with no meaningful pattern.
Survived: Two clear groups: more passengers did not survive than did. Shows imbalance in the target variable.
Pclass (Passenger Class): Most passengers were in 3rd class, fewer in 1st and 2nd class, indicating more lower-class travelers.
Age: Slightly right-skewed with more young adults and fewer older passengers. Some missing values likely exist. Children and teens are visible in lower age bins.
SibSp (Siblings/Spouses aboard):
Most passengers traveled alone. A few had between 1 and 8 siblings or spouses aboard.
Parch (Parents/Children aboard): Most passengers had no parents or children aboard. Only a few had between 1 and 6.
Fare: Mostly low fares with a few very high ones, showing inequality and reflecting class differences.

Key Takeaways:
Most passengers were 3rd class adults traveling alone and paying low fares. More passengers did not survive, suggesting factors like class influenced survival. Age and fare are skewed, and some outliers exist in fare and family-related features.



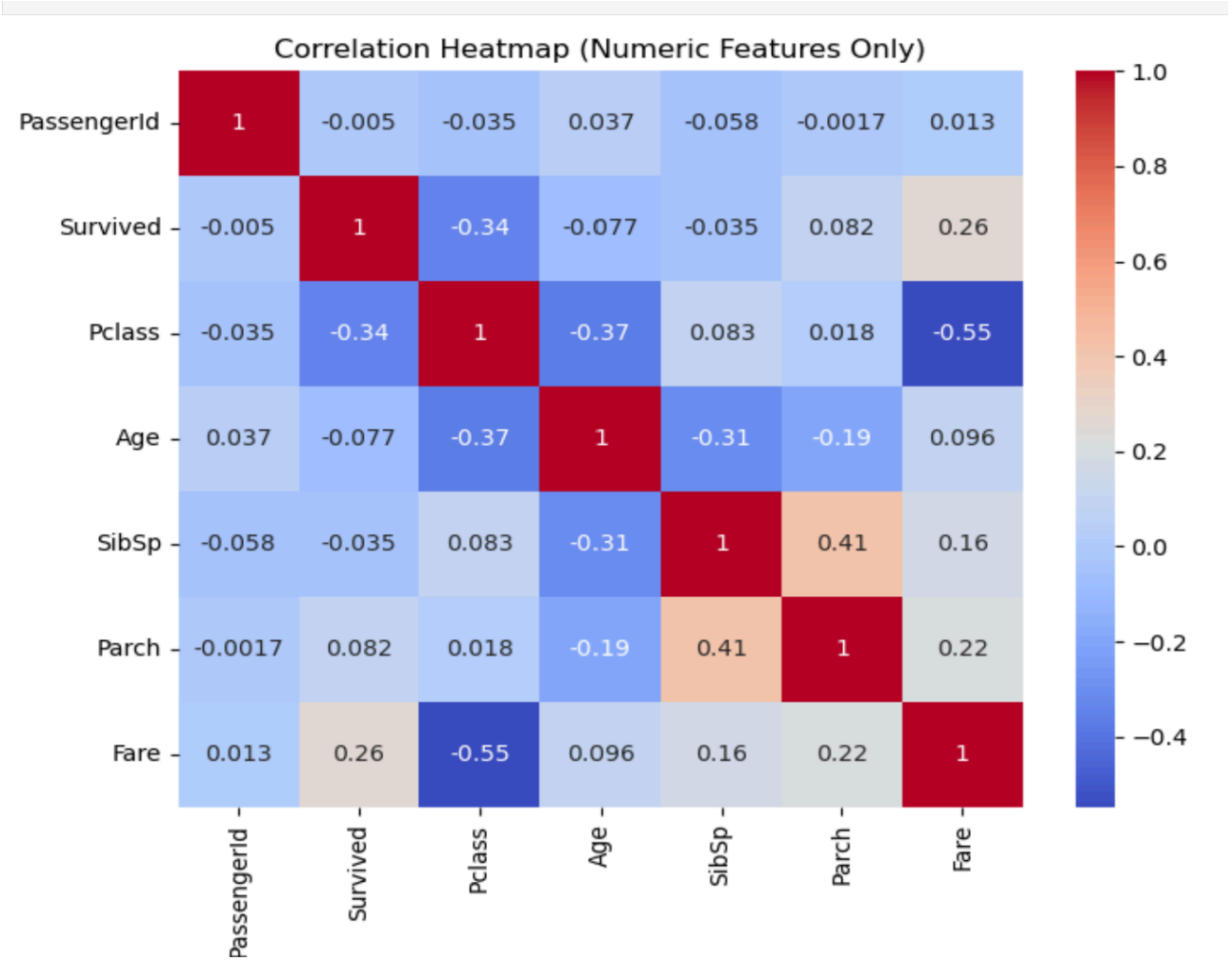
Observation 4 :

PassengerId: follows a uniform spread since it is just an identifier
Survived: has only two categories, 0 and 1
Pclass: is limited to three values, 1 to 3
Age: is mostly between 20 and 40 with a few older outliers
SibSp: is usually 0, meaning most passengers traveled alone, with a few larger counts
Parch: is also mostly 0, with very few cases of parents or children aboard
Fare: is right-skewed with many low fares and a few very high outliers



Observation 5 :

Survived: linked more with class and fare, higher survival for first-class and high fares
Pclass: negatively related with survival, lower class meant lower chance of survival
Age: weakly related overall, younger passengers more often with family
SibSp: moderately related with Parch, families often traveled together
Parch: shows connection with SibSp, both rise when families traveled
Fare: positively related with survival, higher fares tied to better chances
PassengerId: no real relation with any feature

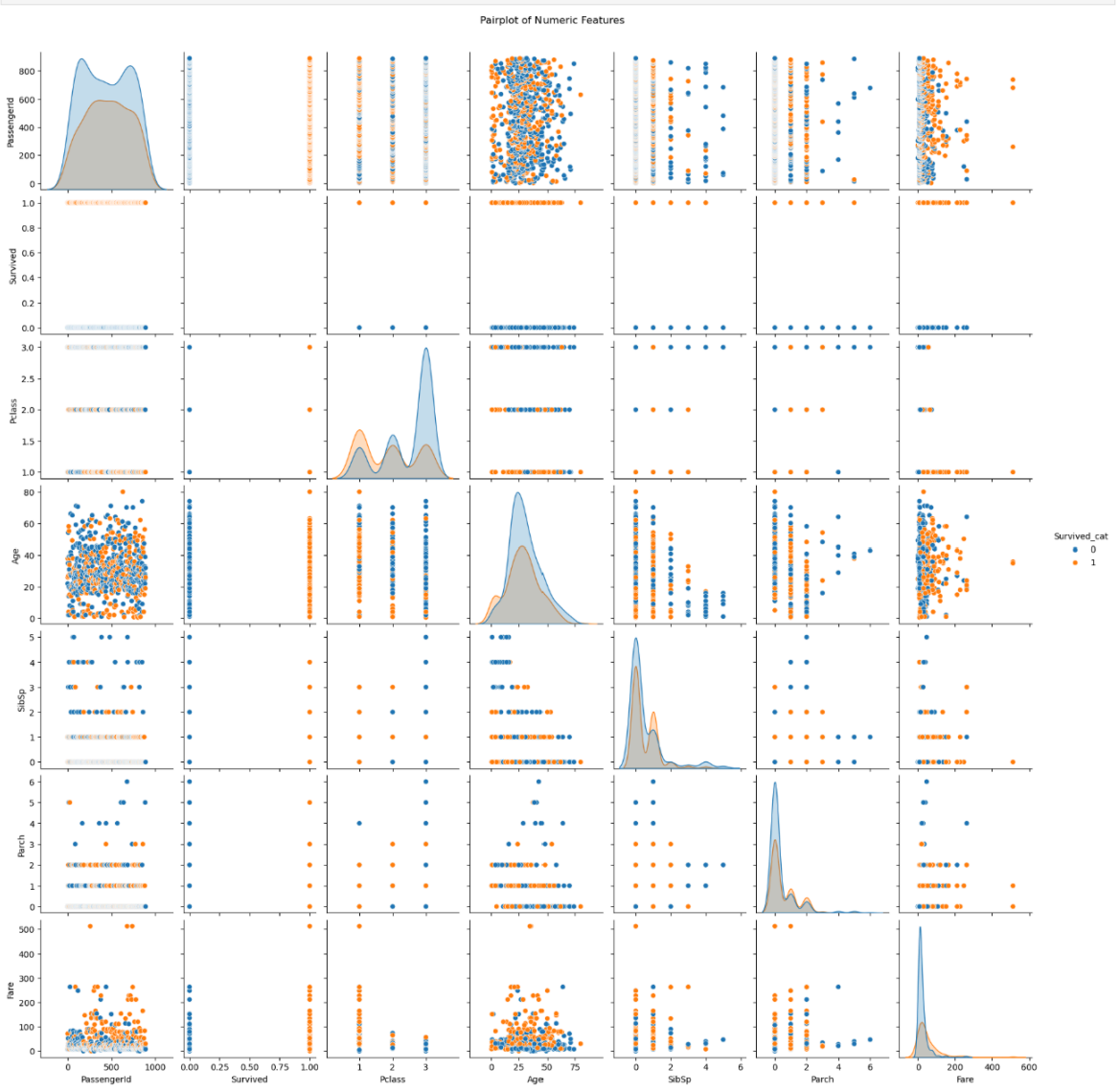


Observation 6:

Survived: higher survival is visible in first class and higher fare ranges
Pclass: lower classes group around low fares and lower survival
Age: most passengers fall between 20–40, children show relatively higher survival
SibSp: mostly 0, few large families with mixed outcomes
Parch: mostly 0, small family groups sometimes have better survival
Fare: strongly right-skewed, higher fares connect with first class and more survival
PassengerId: evenly spread, no meaningful relation

Key Takeaways

• Class and fare were strong indicators of survival chances
• Children and smaller families tended to survive more often
• Most passengers were traveling alone with low fares and lower survival rates
• Identifiers like PassengerId add no value to analysis

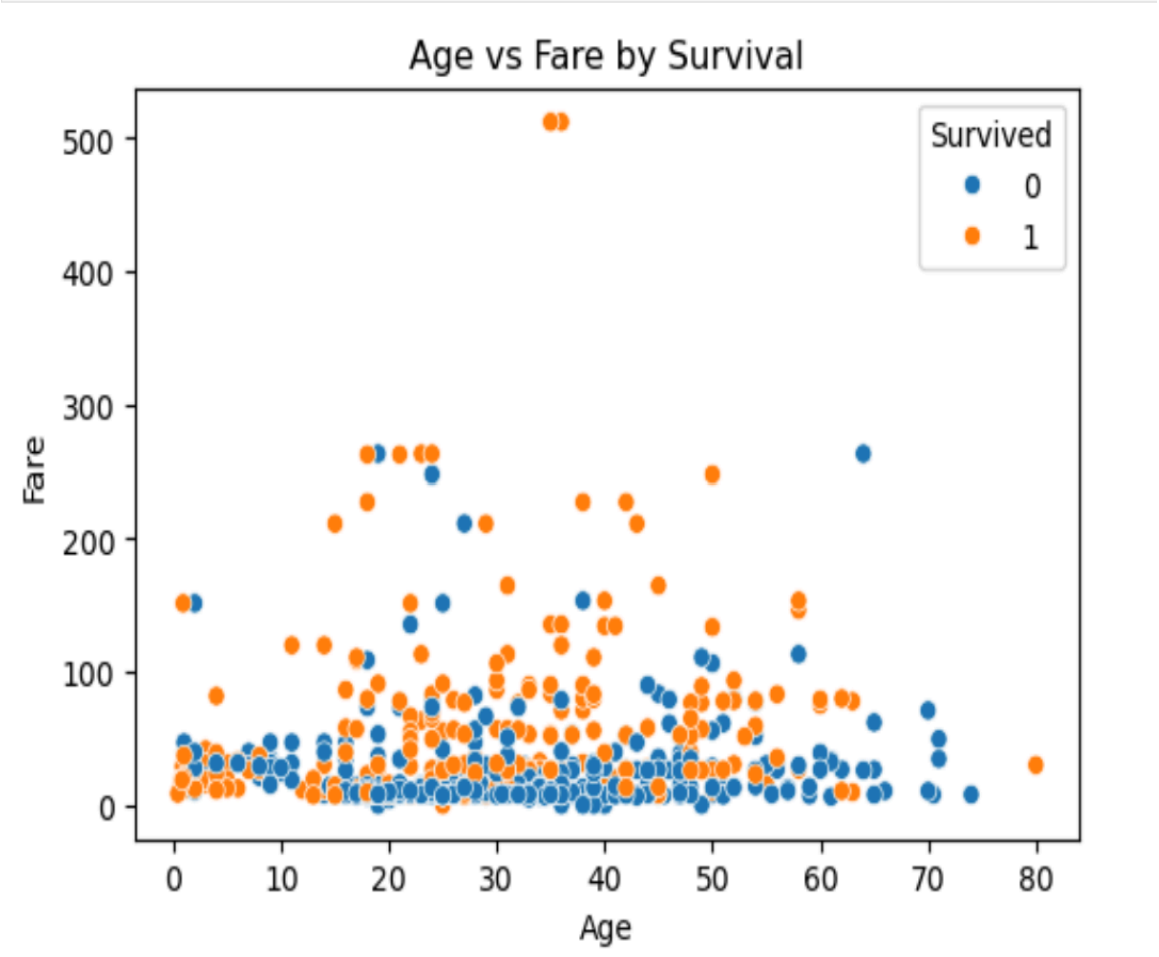


Observation 7: Are survivors clustered by age or fare?

Survivors are not strongly clustered by age — survival is spread across all age groups.
Higher fares show more survivors, especially above 100, where most points are orange.
At lower fares, both survivors and non-survivors appear in large numbers, so fare plays a clearer role than age.
Children and younger passengers also show some higher survival compared to older ones.

Key Takeaways

Fare is a stronger indicator of survival than age.
Passengers paying higher fares (likely first class) had better survival chances.
Survival was not limited to any specific age, but children benefited slightly more.



Summary of Findings

• The dataset has 891 rows and 13 columns
• Missing data is most significant in Cabin (687 values) and Age (177 values), with a few in Embarked (2 values)
• Survival shows positive correlation with Fare and Parch, and negative correlation with Pclass, Age, and SibSp
• Women had a much higher chance of survival than men
• First-class passengers survived more often than those in lower classes
• Larger families sometimes faced lower survival chances compared to individuals or small families
• Age and Fare are both skewed, so transformations may be useful for predictive modeling

Overall Observation

The Titanic dataset, with 891 rows and 13 columns, provides interesting insights into who survived and why. Missing data is mainly in the Cabin column, which is too sparse to be useful, and Age, which can be filled in with imputation. Most other fields are complete.

Looking at the data, it is clear that class and fare had a big impact on survival. Passengers in higher classes and those who paid higher fares were much more likely to survive, highlighting the role of socio-economic privilege. Gender was also crucial, with women surviving at much higher rates than men. Family size had a mixed effect. Smaller families sometimes had a better chance, but larger families often reduced survival.

Age and Fare distributions are skewed. Children and passengers who paid very high fares tended to survive more, while most passengers were in 3rd class, which had the lowest survival rate. PassengerId does not provide any meaningful insight.

Correlation analysis shows that survival is most closely linked to Fare, Pclass, and Sex, while Age, number of siblings and spouses, and number of parents and children have weaker connections. Visualizations support this, showing Fare is a stronger predictor of survival than Age, although children had a slight advantage.

Key Takeaway

Survival on the Titanic was strongly influenced by socio-economic status and gender, moderately by family composition, and only slightly by age. Any predictive model should focus on Pclass, Sex, and Fare, handle missing Age values carefully, and ignore Cabin or PassengerId since they do not contribute meaningfully.