# EXPLORATORY DATA ANALYSIS – TITANIC

## COMPREHENSIVE STATISTICAL & VISUAL EXPLORATION

**Data Sources:**

train.csv (n=891), test.csv (n=418), gender_submission.csv

**Objective:**

- Extract insights via descriptive statistics and visualization.
- Identify relationships and trends related to survival.

**Tools:**

Python (Pandas, Matplotlib, Seaborn)

# Data Overview

## .info() excerpt:

```
<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
```

## Numeric .describe() (top rows):

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| PassengerId | 891.0 | 446.00 | 257.35 | 1.00 | 223.50 | 446.00 | 668.5 | 891.00 |
| Survived | 891.0 | 0.38 | 0.49 | 0.00 | 0.00 | 0.00 | 1.0 | 1.00 |
| Pclass | 891.0 | 2.31 | 0.84 | 1.00 | 2.00 | 3.00 | 3.0 | 3.00 |
| Age | 714.0 | 29.70 | 14.53 | 0.42 | 20.12 | 28.00 | 38.0 | 80.00 |
| SibSp | 891.0 | 0.52 | 1.10 | 0.00 | 0.00 | 0.00 | 1.0 | 8.00 |
| Parch | 891.0 | 0.38 | 0.81 | 0.00 | 0.00 | 0.00 | 0.0 | 6.00 |
| Fare | 891.0 | 32.20 | 49.69 | 0.00 | 7.91 | 14.45 | 31.0 | 512.33 |

**Observations:**

- Training set: 891 rows, 13 columns.
- Missingness in Age and Cabin; minor in Embarked.
- Fare is highly right-skewed; scales vary across numeric features.
- 'Survived' is class-imbalanced (more non-survivors).

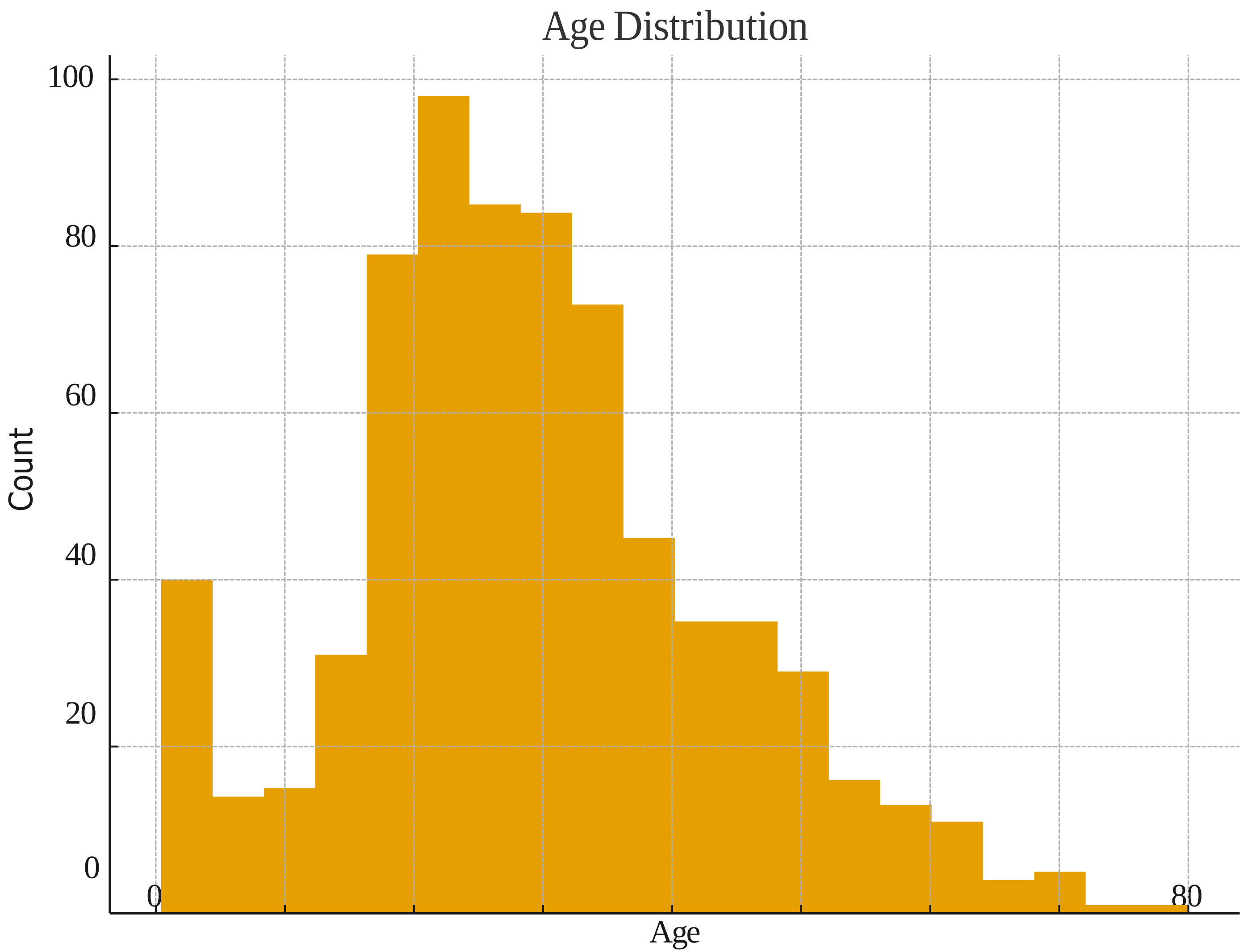
# Missing Values

## Missing values (top 15):

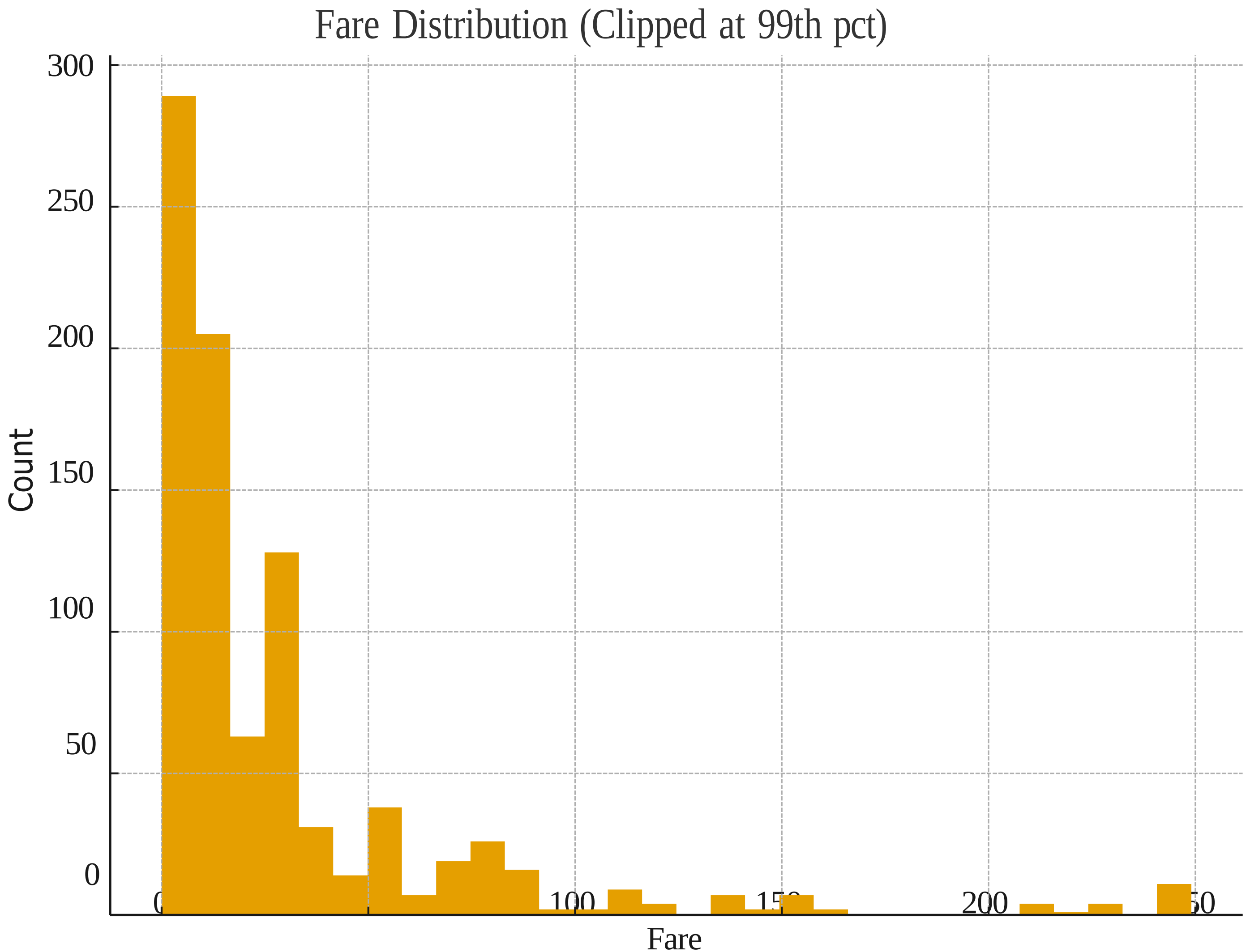
| | missing | percent |
|---|---|---|
| Cabin | 687 | 77.10 |
| Age | 177 | 19.87 |
| Embarked | 2 | 0.22 |
| PassengerId | 0 | 0.00 |
| Survived | 0 | 0.00 |
| Pclass | 0 | 0.00 |
| Name | 0 | 0.00 |
| Sex | 0 | 0.00 |
| SibSp | 0 | 0.00 |
| Parch | 0 | 0.00 |
| Ticket | 0 | 0.00 |
| Fare | 0 | 0.00 |

**Observations:**

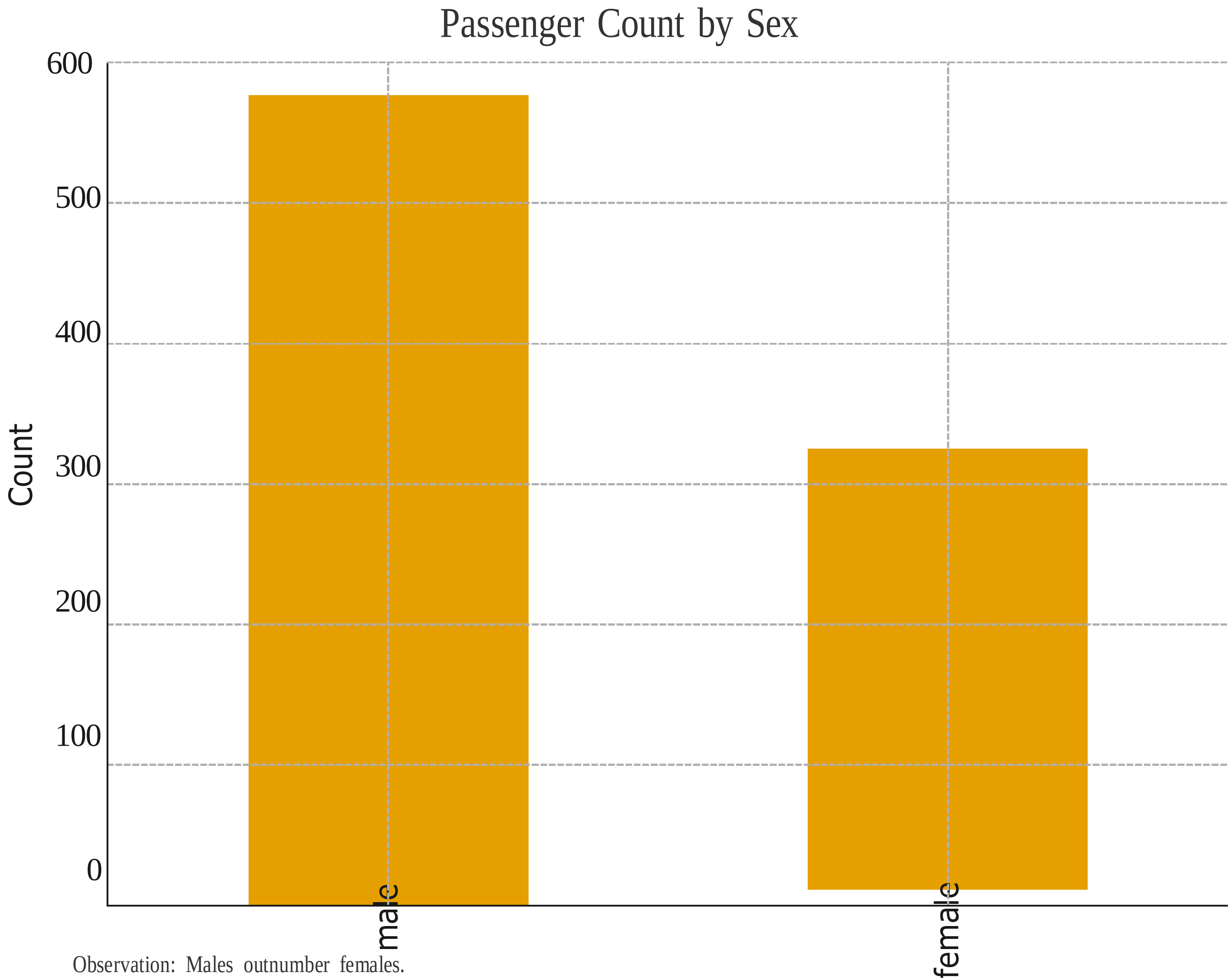
- 'Cabin' is mostly missing - use a 'has_cabin' flag or drop.
- 'Age' missingness needs imputation (e.g., by Title+Pclass medians).
- 'Embarked' missing can be filled with mode.



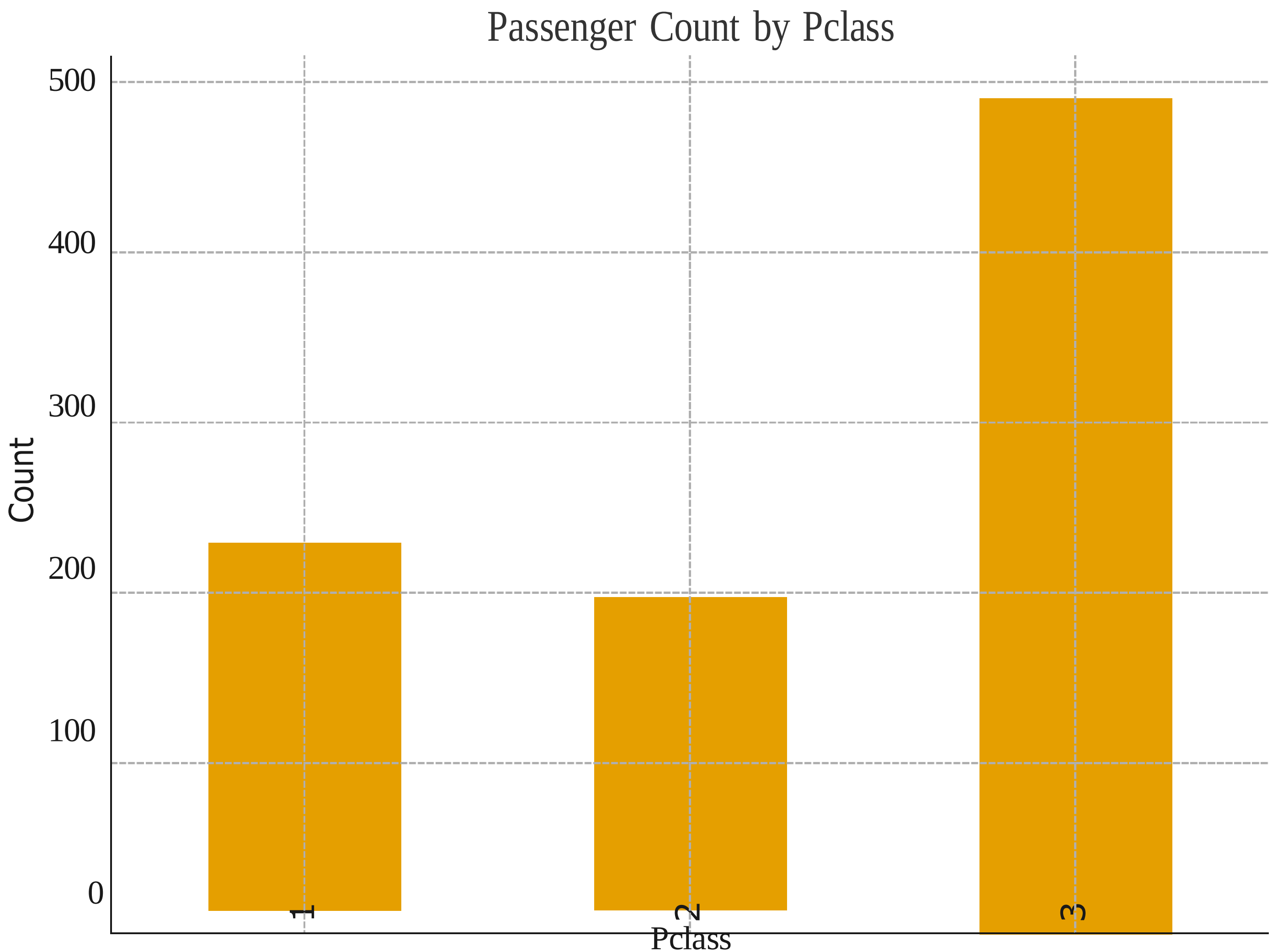

Observation: Age concentrates around young adults (20 40).



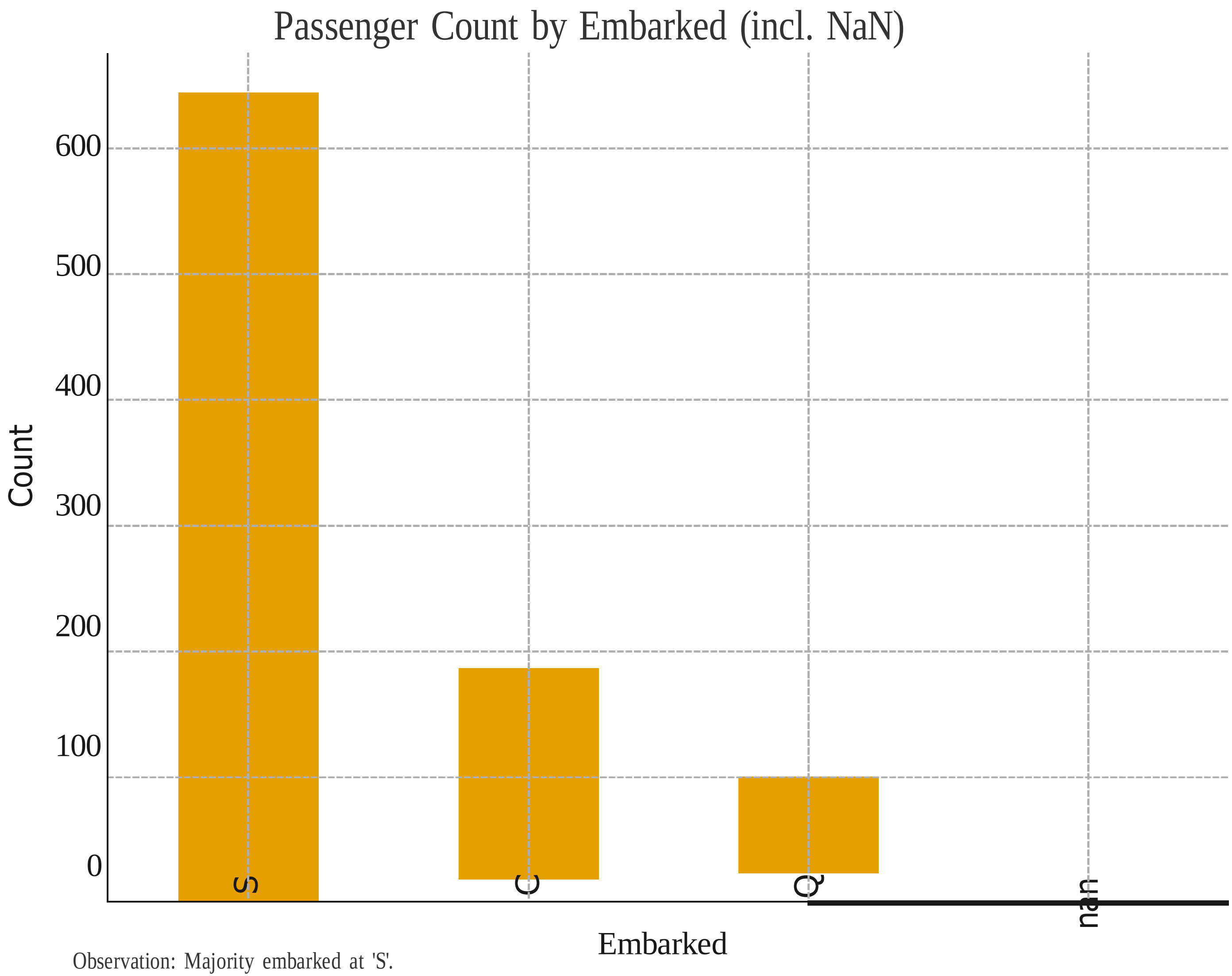
Observation: Strong right skew; consider log transform.

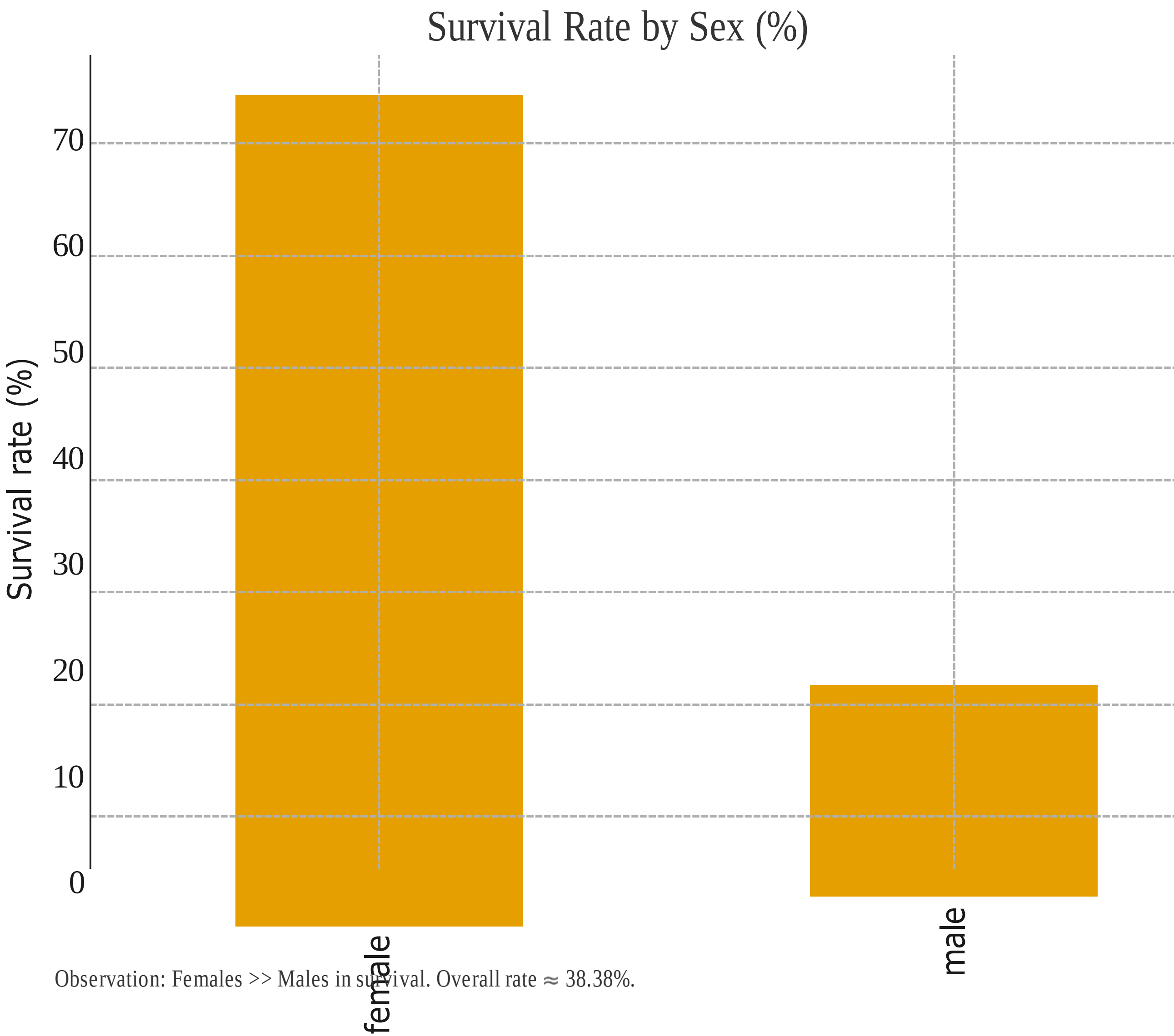



Observation: Males outnumber females.

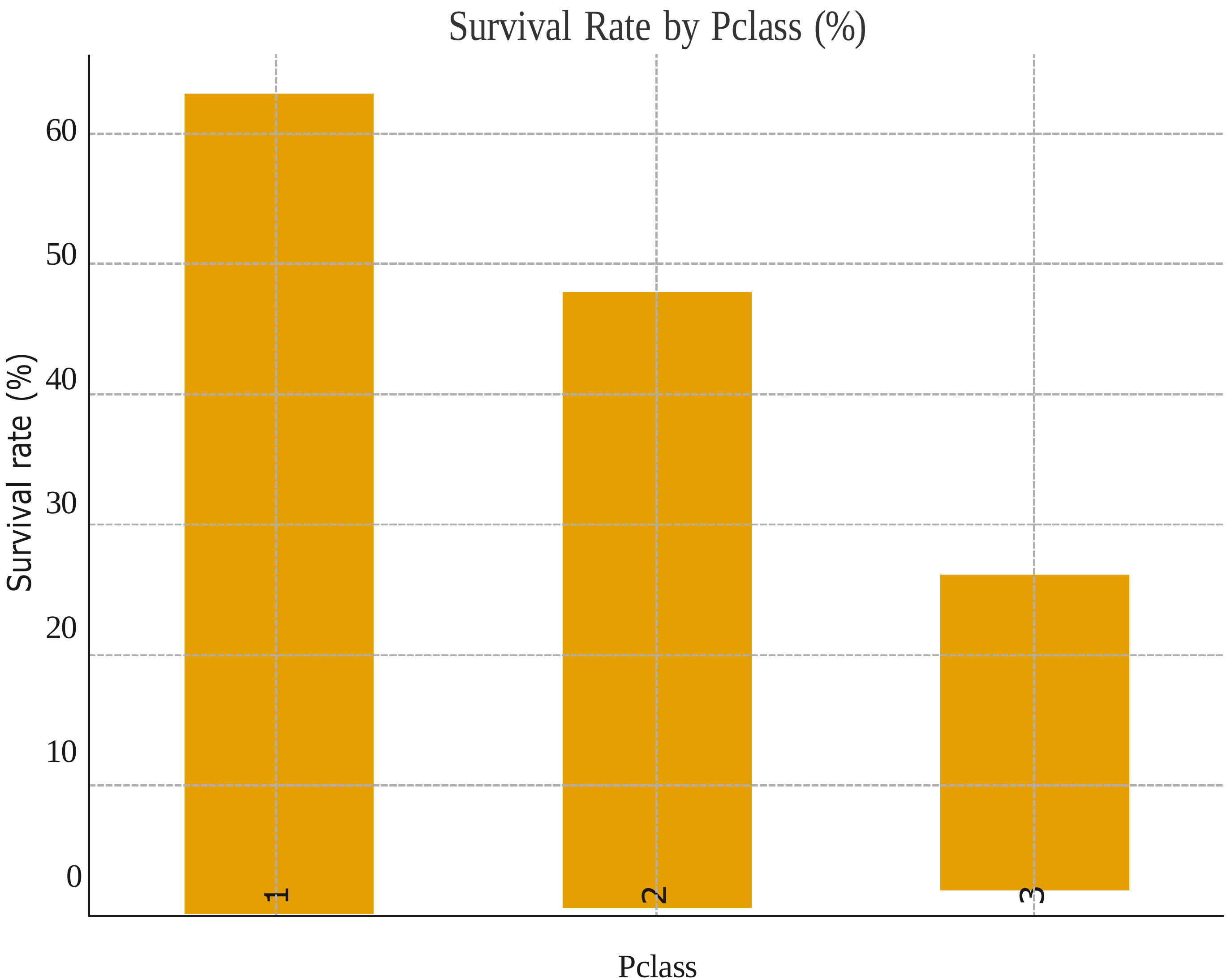



Observation: 3rd class is the largest cohort.

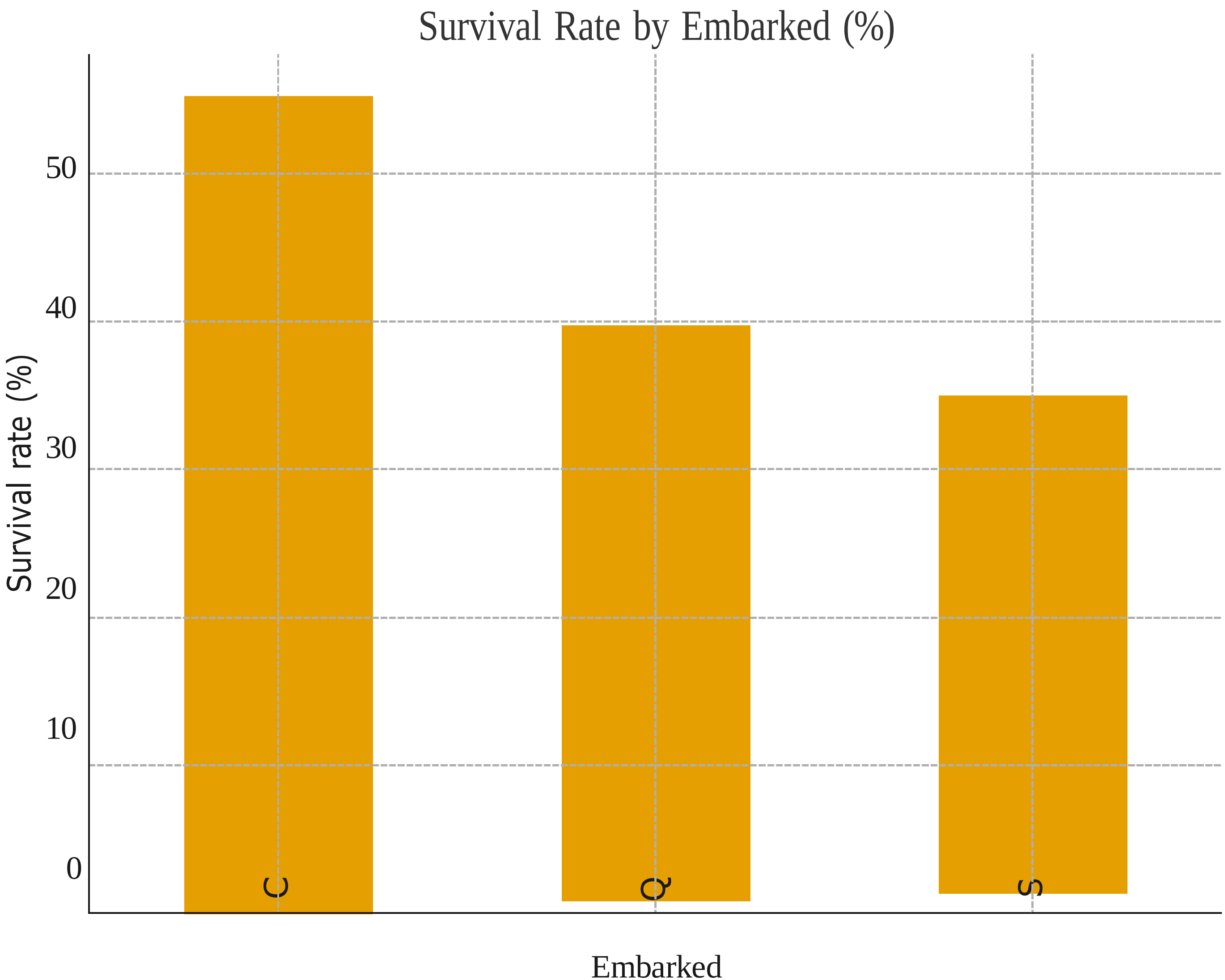

Passenger Count by Embarked (incl. NaN)
Count
0
100
200
300
400
500
600
S
C
Q
nan
Embarked
Observation: Majority embarked at 'S'.



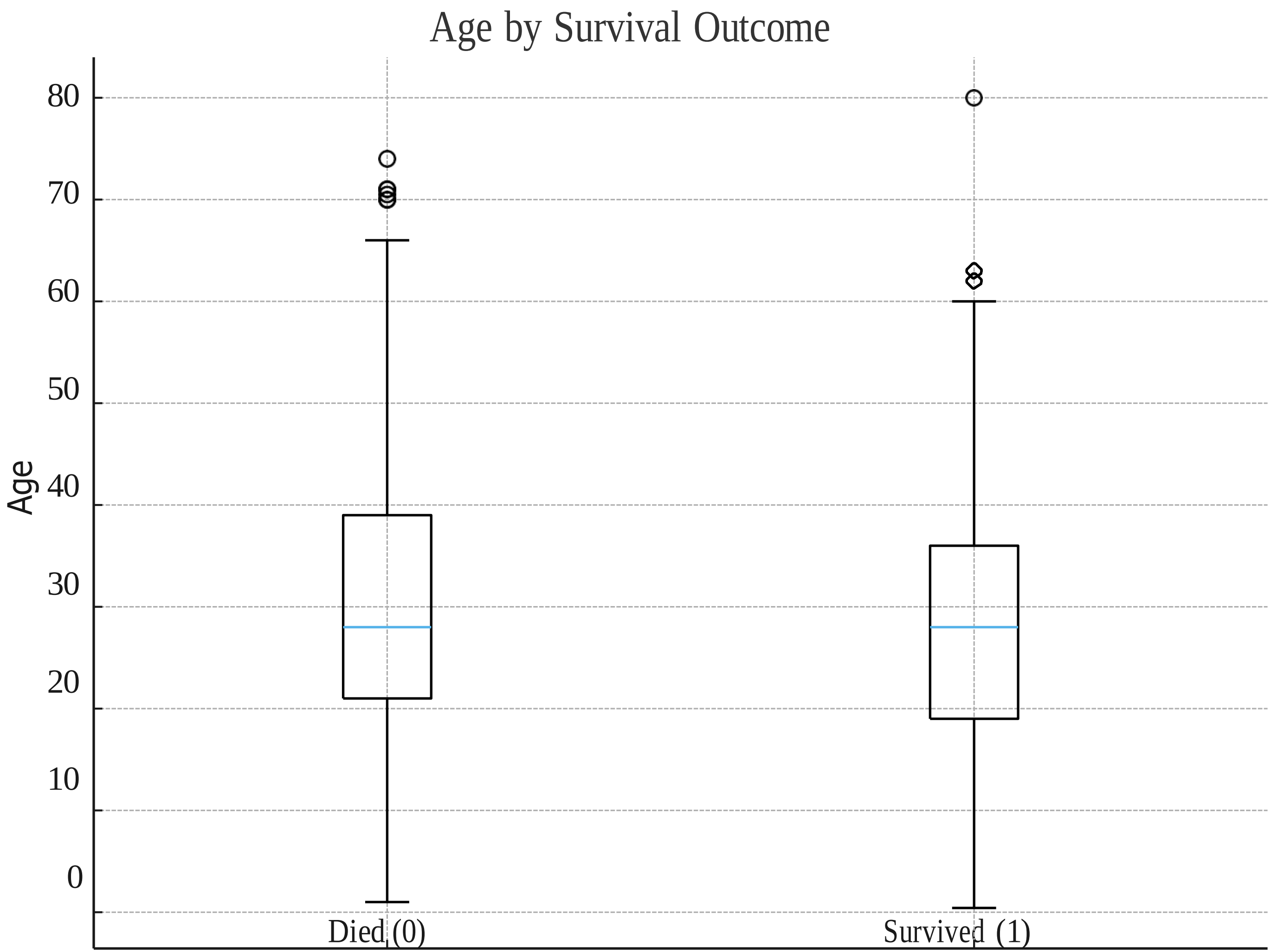
Survival Rate by Sex (%)
Survival rate (%)
70
60
50
40
30
20
10
0
female
male
Observation: Females >> Males in survival. Overall rate ≈ 38.38%.



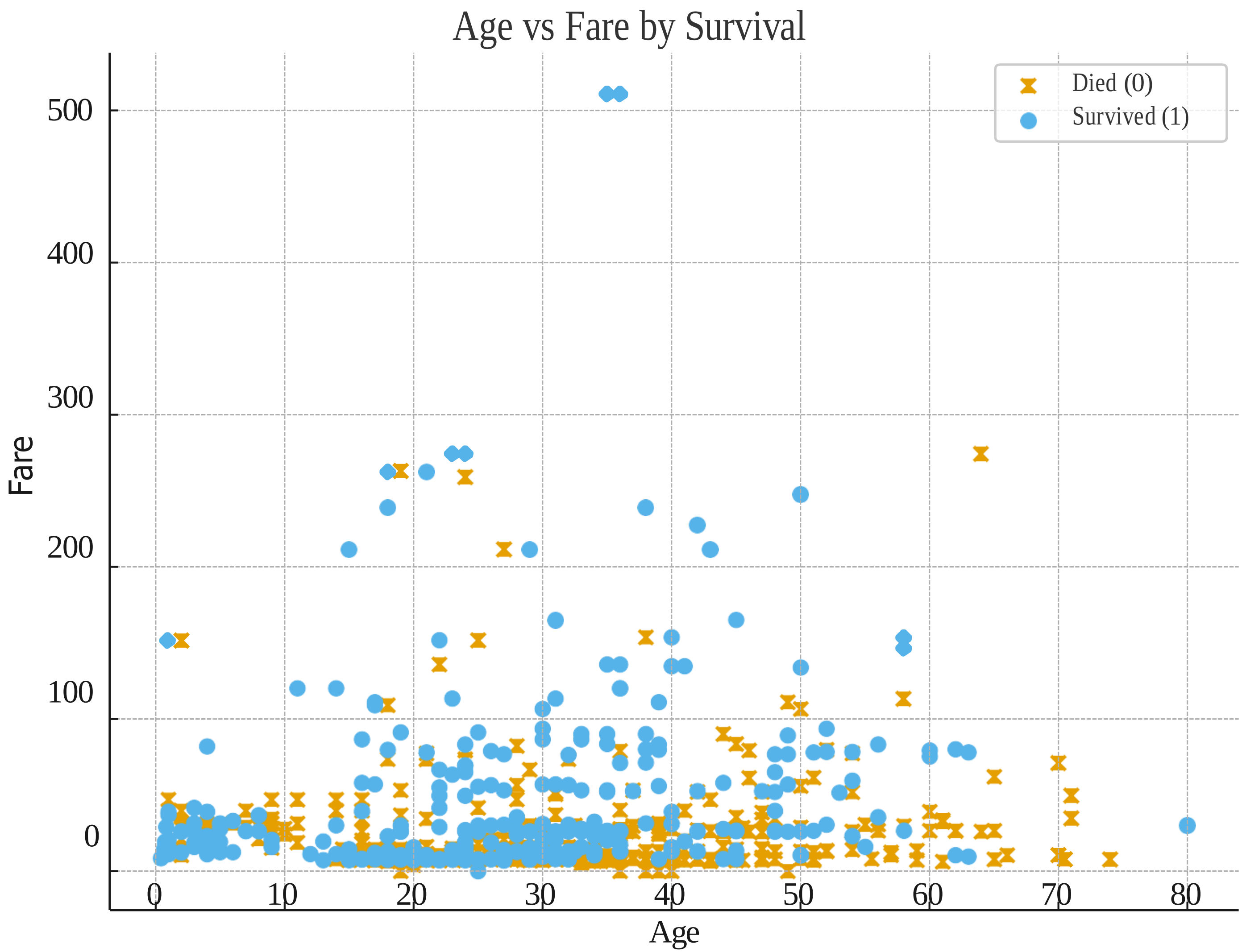

Observation: 1st class survival highest; class is a strong driver.



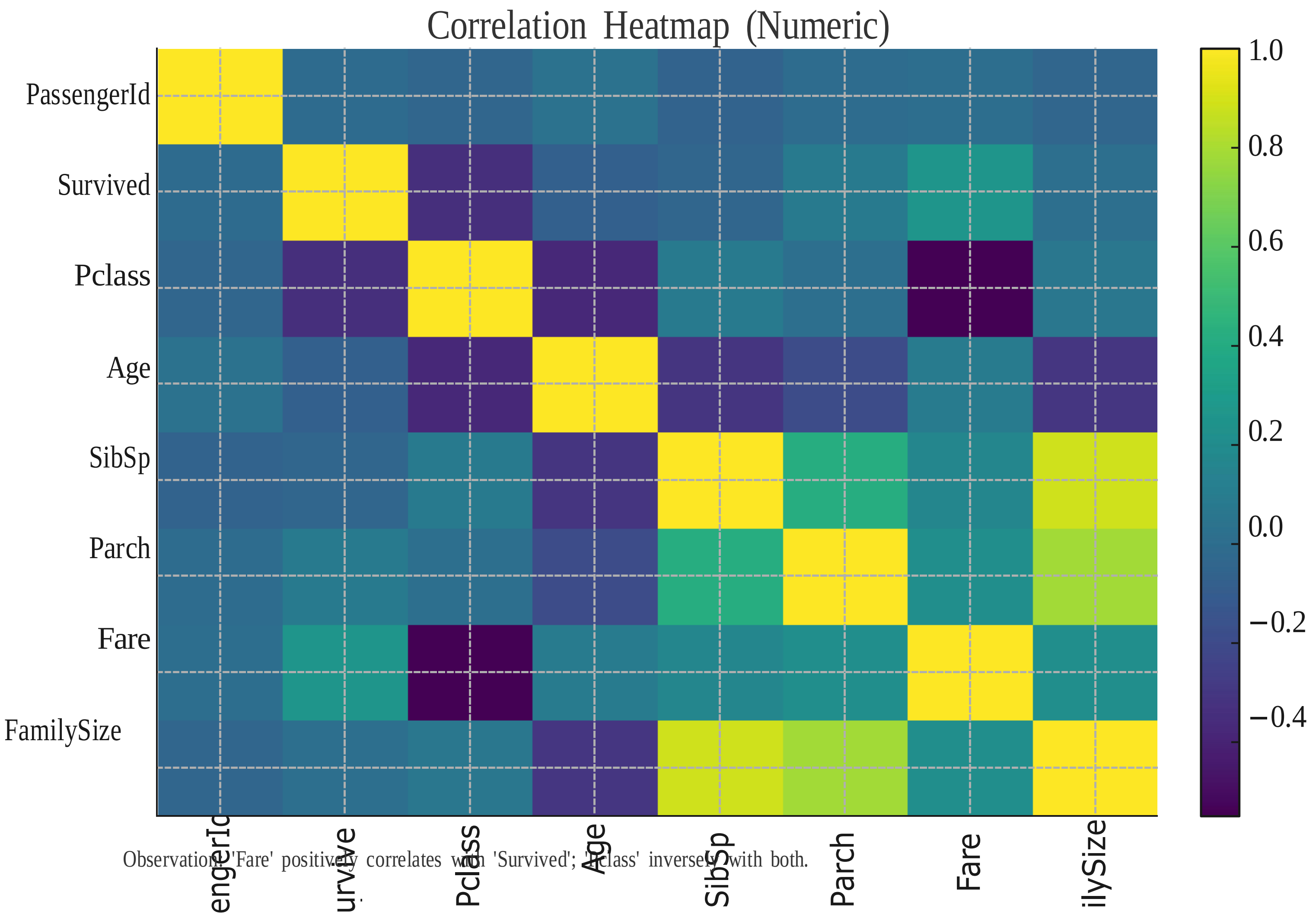

Observation: 'C' often exceeds 'S'/'Q' in survival.




Observation: Survivors skew slightly younger; children benefited from evacuation priority.




Observation: Survivors concentrate at higher fares (proxying Pclass).



Correlation Heatmap (Numeric)
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
FamilySize
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Observation: 'Fare' positively correlates with 'Survived'; 'Pclass' inversely with both.

## Summary & Next Steps

**Key Findings:**

- Overall survival rate: 38.38%.
- Sex and Class are the strongest differentiators.
- Higher fares (proxying socioeconomics/class) associate with higher survival.
- Age effects are nuanced; children tend to survive more.
- Missingness: Cabin (heavy), Age (moderate), Embarked (minor).

**Recommendations:**

- Impute Age (Title+Pclass medians); add has_cabin flag.
- Feature engineering: FamilySize, IsAlone, Title from Name; consider Fare log transform.
- Baselines: Logistic Regression, Decision Tree, Random Forest; stratified CV.