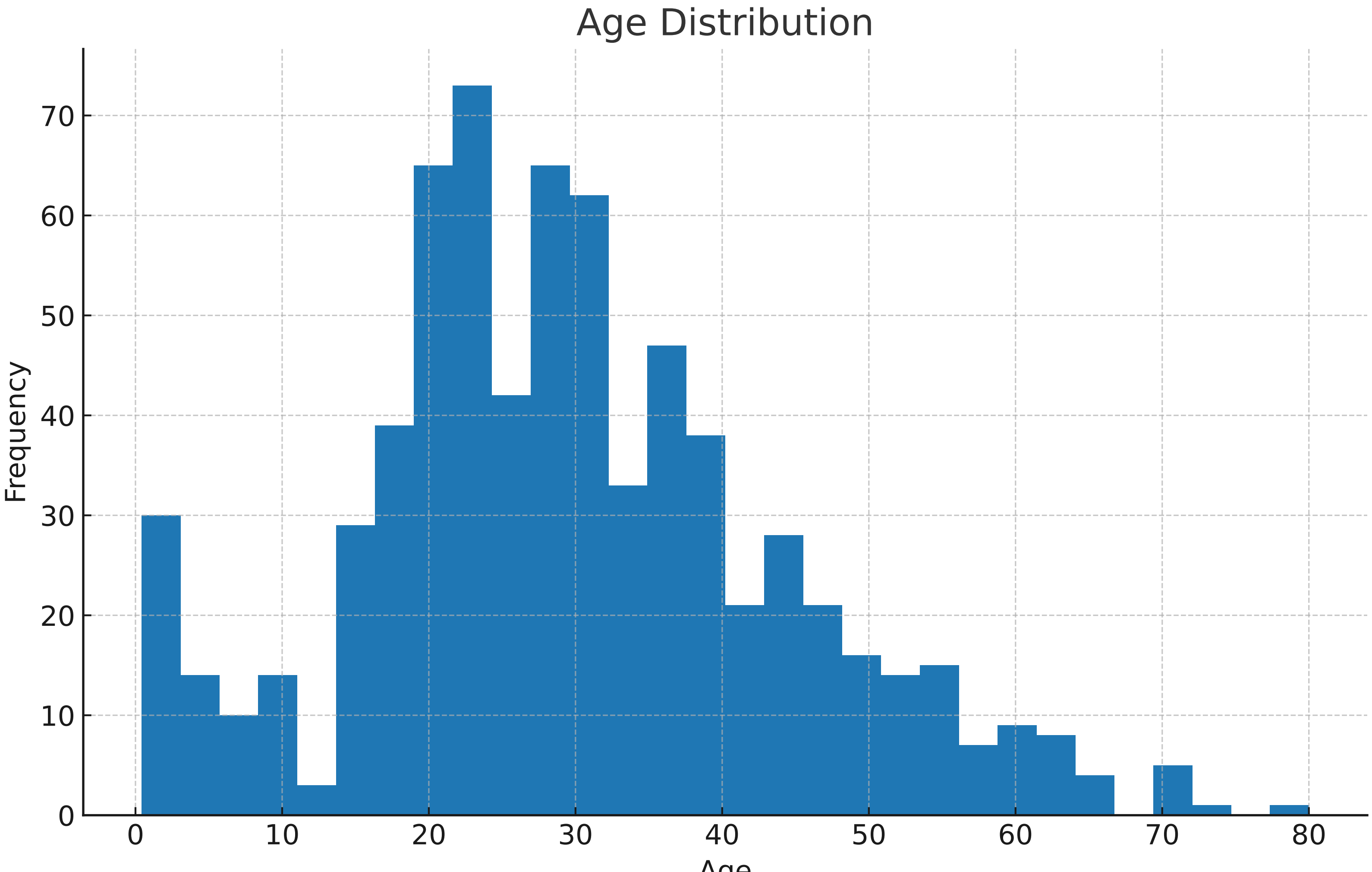
Titanic Dataset: Exploratory Data Analysis

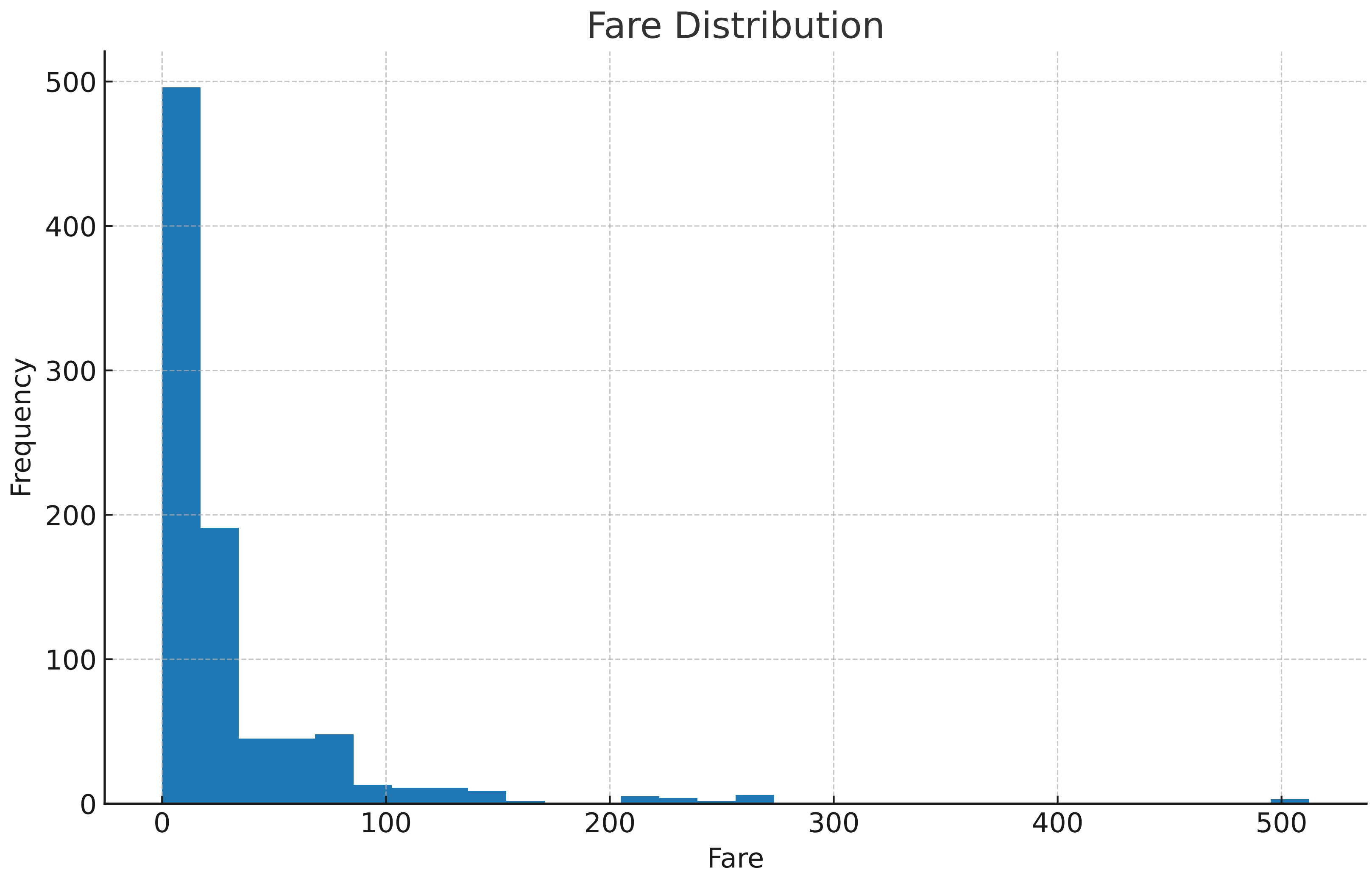
Generated on 2025-08-19 14:41:58

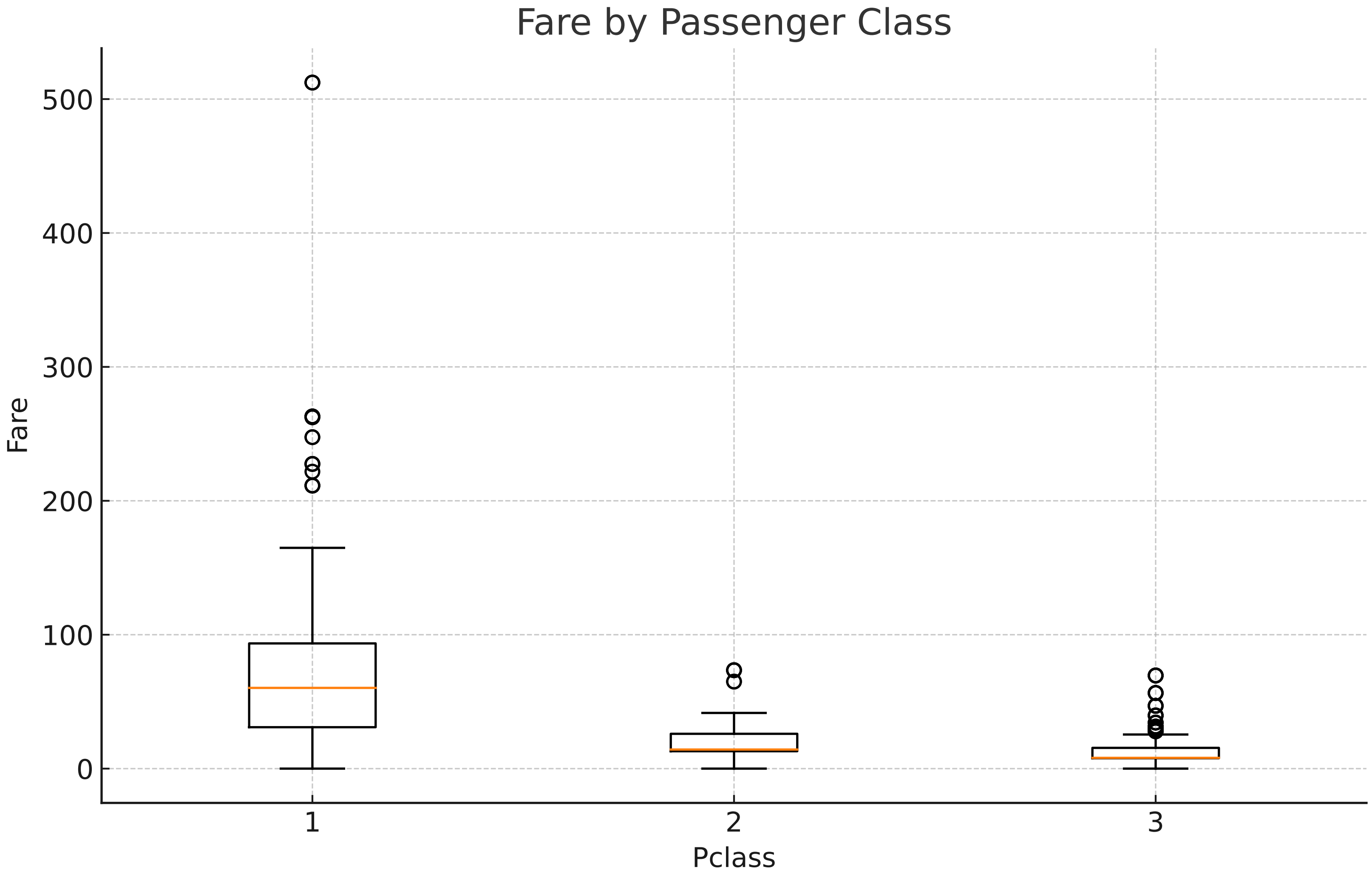
Using uploaded titanic.csv (Kaggle Titanic train.csv)

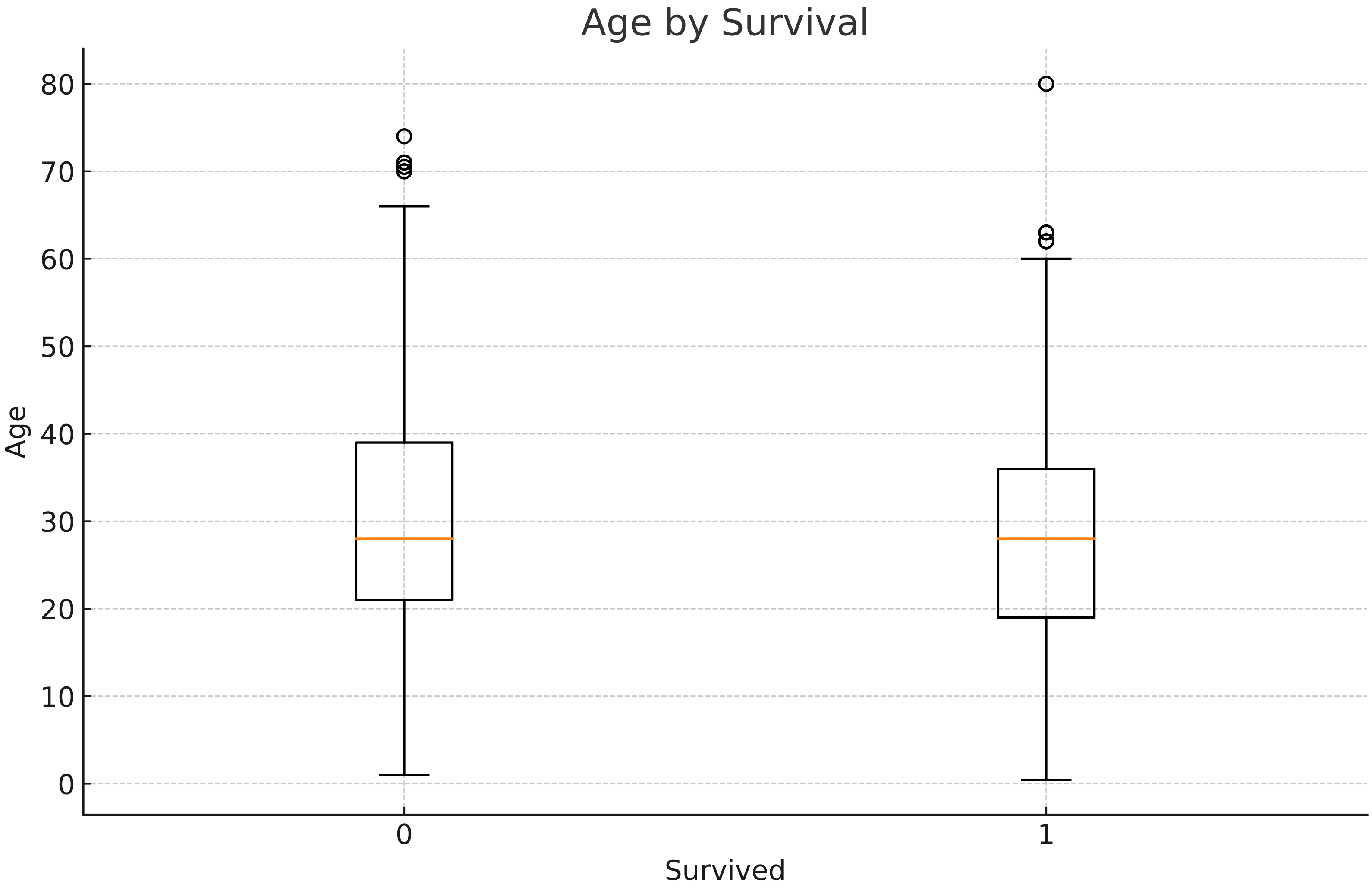
```
Rows: 891, Columns: 12

=== .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


=== Missing Values ===
Cabin          687
Age            177
Embarked         2
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
SibSp            0
Parch            0
Ticket           0
Fare             0
```

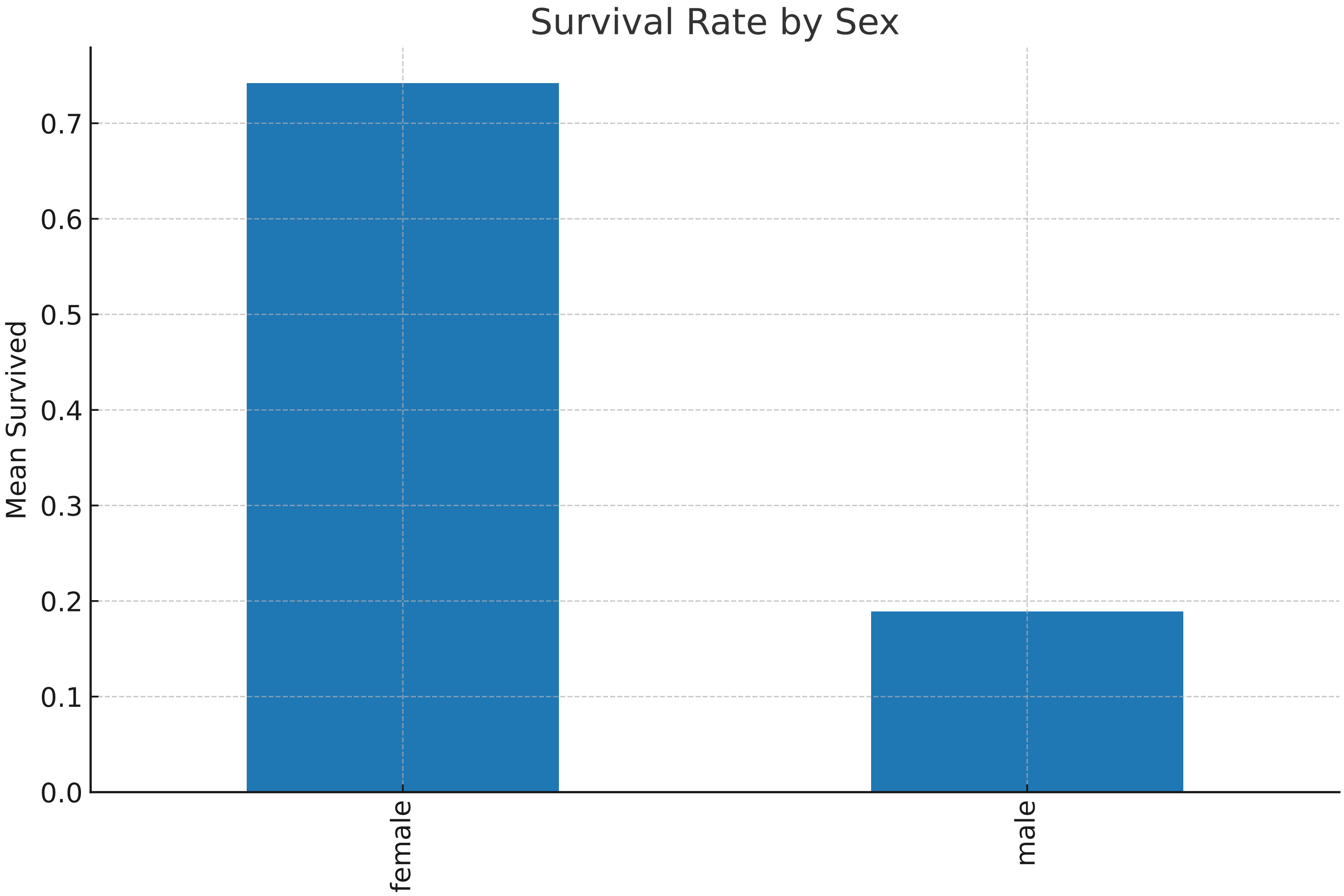
```
        PassengerId    Survived      Pclass                     Name   Sex         Age       SibSp       Parch  Ticket        Fare    Cabin Embarke
count    891.000000  891.000000  891.000000                      891   891  714.000000  891.000000  891.000000     891  891.000000      204      88
unique          NaN         NaN         NaN                      891     2         NaN         NaN         NaN     681         NaN      147
top             NaN         NaN         NaN  Braund, Mr. Owen Harris  male         NaN         NaN         NaN  347082         NaN  B96 B98
freq            NaN         NaN         NaN                        1   577         NaN         NaN         NaN       7         NaN        4      64
mean     446.000000    0.383838    2.308642                      NaN   NaN   29.699118    0.523008    0.381594     NaN   32.204208      NaN      Na
std      257.353842    0.486592    0.836071                      NaN   NaN   14.526497    1.102743    0.806057     NaN   49.693429      NaN      Na
min        1.000000    0.000000    1.000000                      NaN   NaN    0.420000    0.000000    0.000000     NaN    0.000000      NaN      Na
25%      223.500000    0.000000    2.000000                      NaN   NaN   20.125000    0.000000    0.000000     NaN    7.910400      NaN      Na
50%      446.000000    0.000000    3.000000                      NaN   NaN   28.000000    0.000000    0.000000     NaN   14.454200      NaN      Na
75%      668.500000    1.000000    3.000000                      NaN   NaN   38.000000    1.000000    0.000000     NaN   31.000000      NaN      Na
max      891.000000    1.000000    3.000000                      NaN   NaN   80.000000    8.000000    6.000000     NaN  512.329200      NaN      Na
```

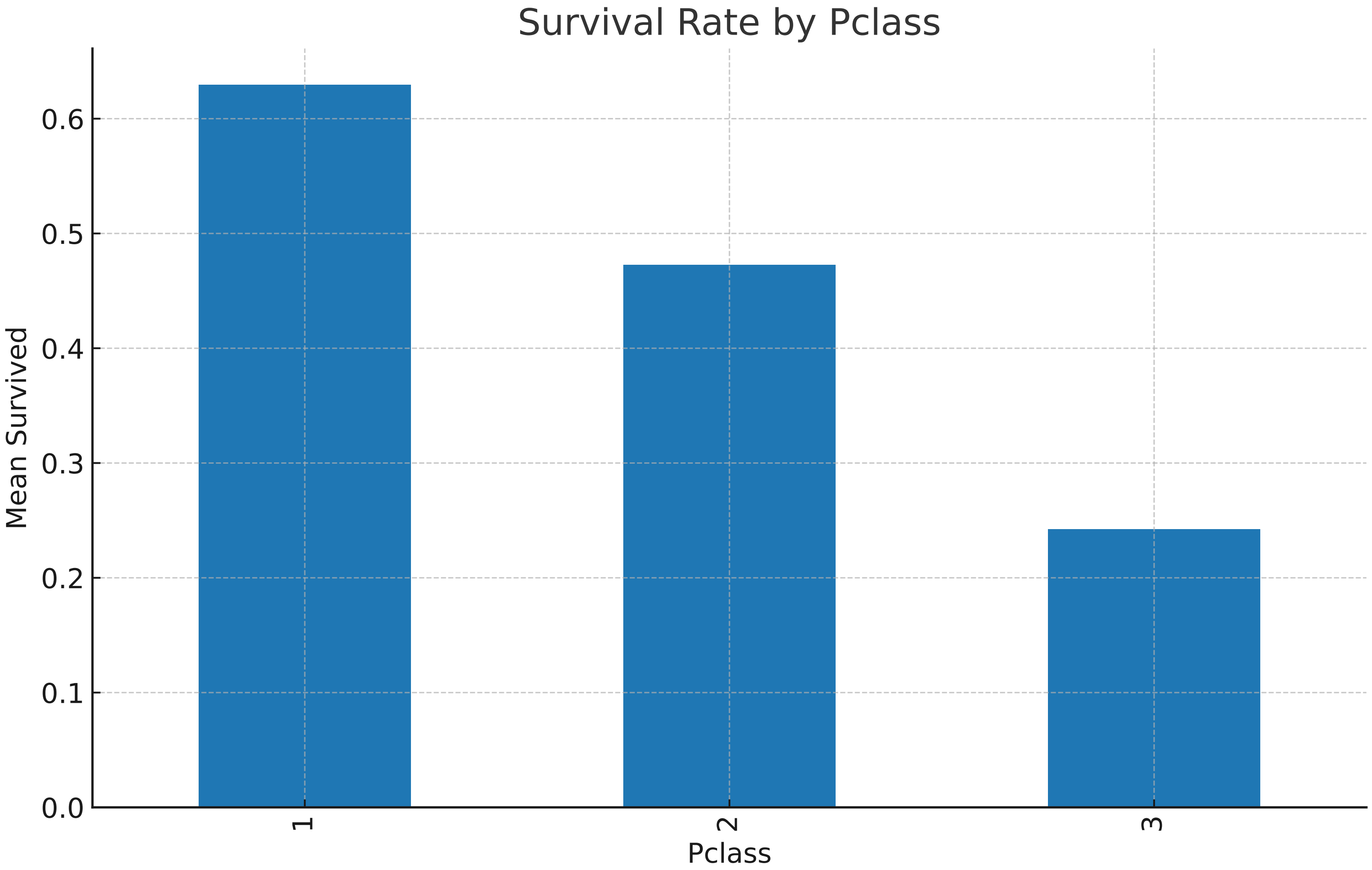

Age Distribution
Frequency
Age
0
10
20
30
40
50
60
70
0
10
20
30
40
50
60
70
80

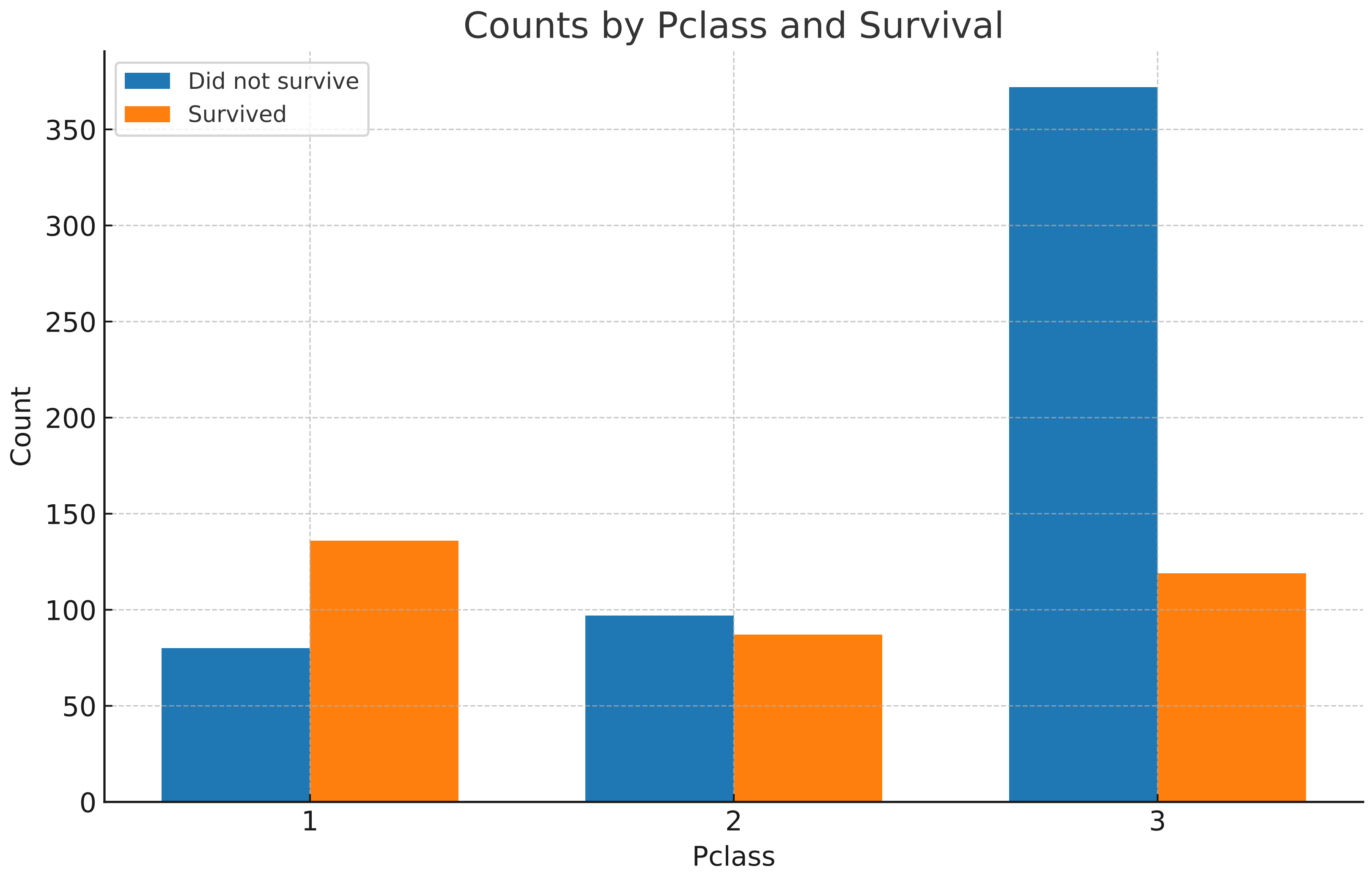

Fare Distribution
Frequency
500
400
300
200
100
0
0
100
200
300
400
500
Fare

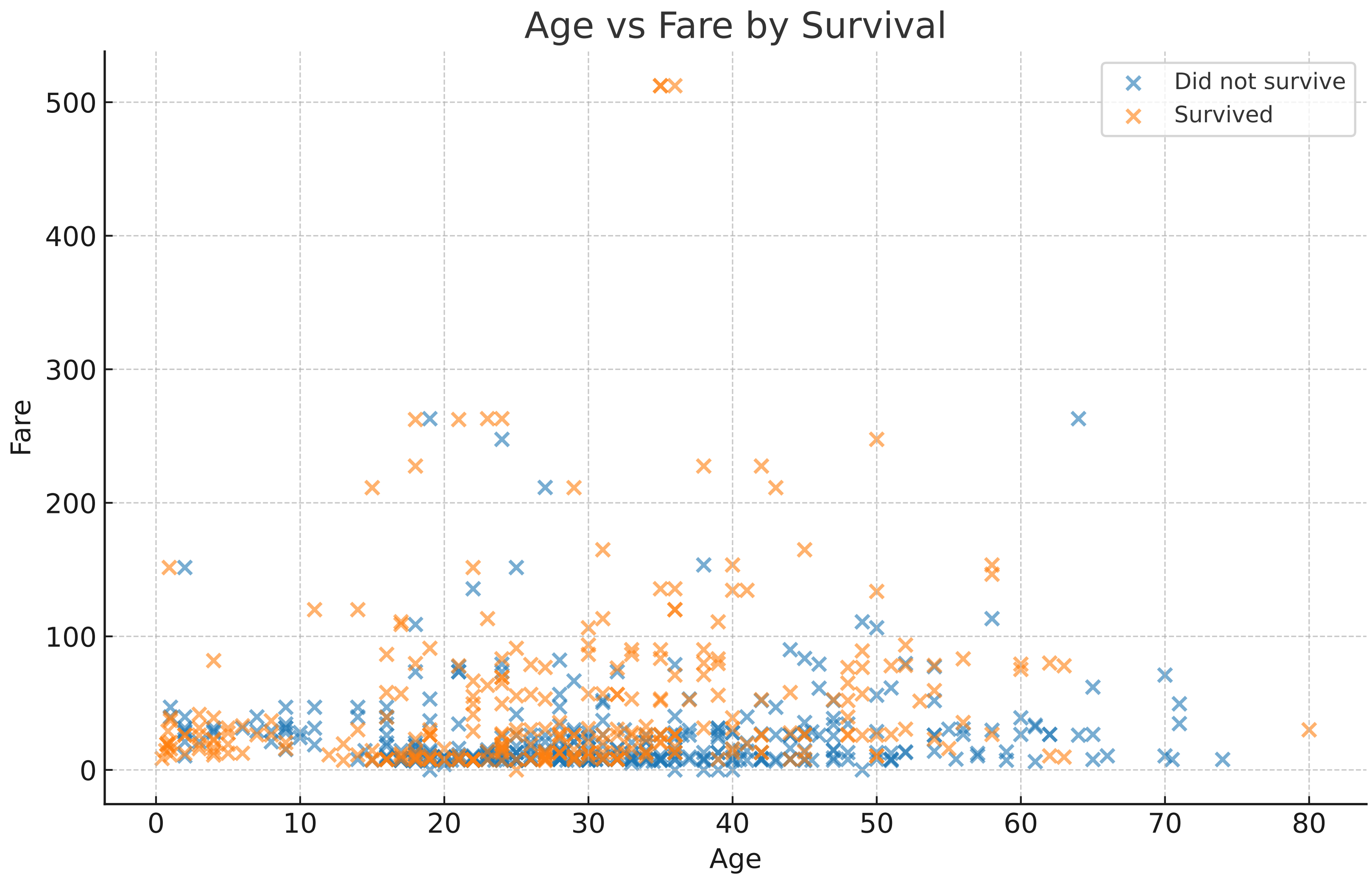

Fare by Passenger Class
Fare
500
400
300
200
100
0
1
2
3
Pclass

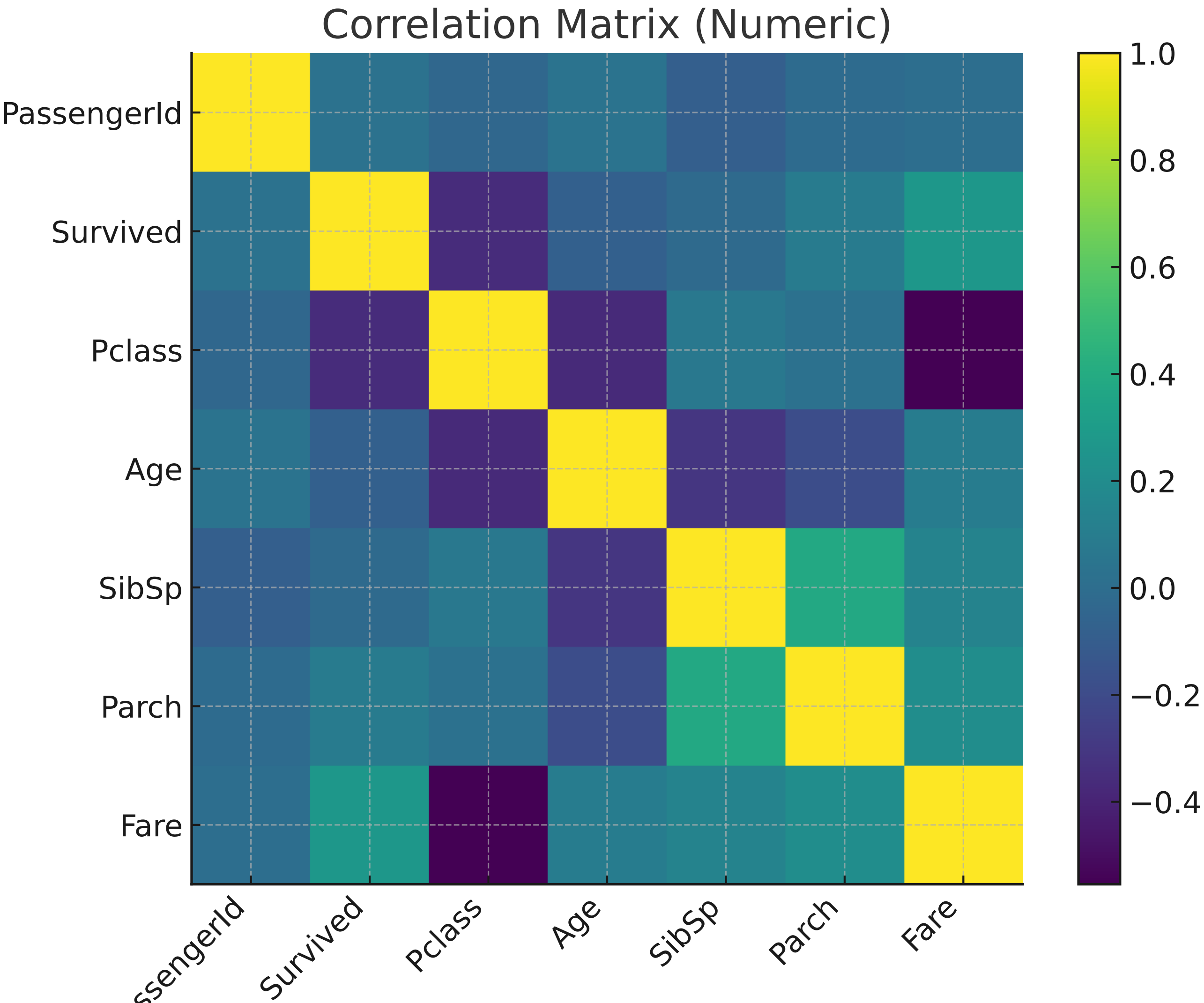

Age by Survival
Age
80
70
60
50
40
30
20
10
0
0
1
Survived


Survival Rate by Sex
Mean Survived
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
female
male


Survival Rate by Pclass
Mean Survived
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
Pclass


Counts by Pclass and Survival
Did not survive
Survived
Count
0
50
100
150
200
250
300
350
1
2
3
Pclass


Age vs Fare by Survival
Did not survive
Survived
Fare
Age
0
100
200
300
400
500
0
10
20
30
40
50
60
70
80


Correlation Matrix (Numeric)
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
ssengerId
Survived
Pclass
Age
SibSp
Parch
Fare
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

- Age: skewed toward young adults; missing values exist.
- Fare: right-skewed; wide spread especially in 1st class.
- Fare vs Pclass: strong separation, higher class => higher fare.
- Age vs Survival: children had higher survival than some adults.
- Survival Rate by Sex: females survived at higher rates.
- Survival Rate by Pclass: higher classes show higher survival.
- Counts: 3rd class had many non-survivors.
- Age vs Fare scatter: survivors more common at higher fares.
- Correlations: Survived positively linked to Fare, negatively to Pclass.

Summary of Findings:

- Survival higher among females and higher passenger classes.
- Fare (wealth) is associated with better survival.
- Children had better chances of survival.
- Missing Age/Embarked values need imputation for modeling.
- Next steps: feature engineering (family size, titles, cabin decks), preprocessing for ML models.