Titanic - Exploratory Data Analysis

TitanicEDA Report

By Balaji Thakur

Dataset Overview

Fileused: /mnt/data/train[1].csv

Shape (rows, columns): (891, 12)

Columns and dtypes:

- PassengerId: int64
- Survived: int64
- Pclass: int64
- Name: object
- Sex: object
- Age: float64
- SibSp: int64
- Parch: int64
- Ticket: object
- Fare: float64
- Cabin: object
- Embarked: object

Missing values:

- PassengerId: 0
- Survived: 0
- Pclass: 0
- Name: 0
- Sex: 0
- Age: 177
- SibSp: 0
- Parch: 0
- Ticket: 0
- Fare: 0
- Cabin: 687
- Embarked: 2

Numerical Summary (describe())

Numerical summary (selected):

PassengerId: {'count': 891.0, 'mean': 446.0, 'std': 257.3538420152301, 'min': 1.0, '25%': 223.5, '50%':

Survived: {'count': 891.0, 'mean': 0.3838383838383838, 'std': 0.4865924542648575, 'min': 0.0, '25%'

Pclass: {'count': 891.0, 'mean': 2.308641975308642, 'std': 0.836071240977049, 'min': 1.0, '25%': 2.0,

Age: {'count': 714.0, 'mean': 29.69911764705882, 'std': 14.526497332334042, 'min': 0.42, '25%': 20.

SibSp: {'count': 891.0, 'mean': 0.5230078563411896, 'std': 1.1027434322934317, 'min': 0.0, '25%': 0.

Parch: {'count': 891.0, 'mean': 0.38159371492704824, 'std': 0.8060572211299483, 'min': 0.0, '25%': 0

Fare: {'count': 891.0, 'mean': 32.204207968574636, 'std': 49.6934285971809, 'min': 0.0, '25%': 7.910

Categorical Summary (describe(include='object'))

Categoricalsummary (selected):

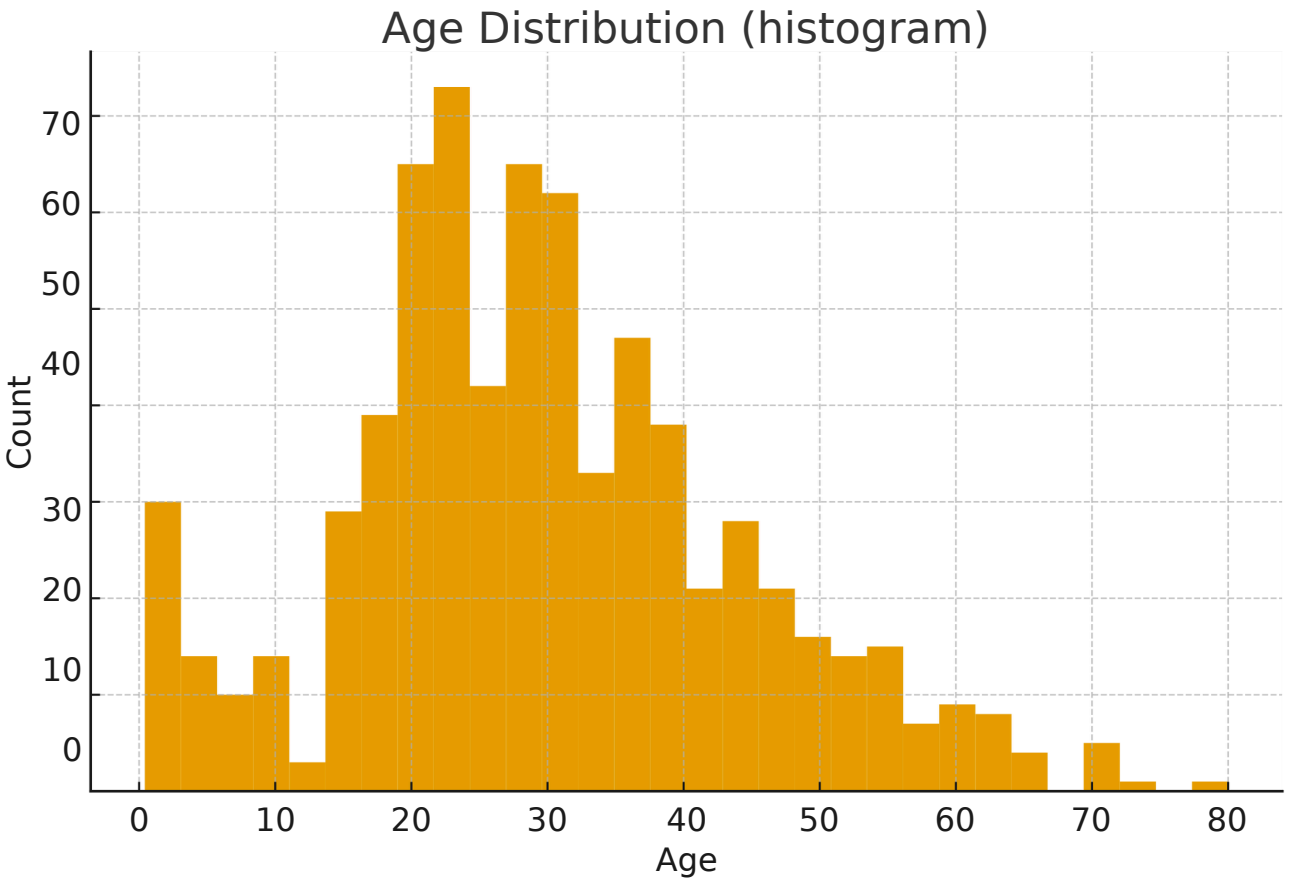
Name: {'count': 891, 'unique': 891, 'top': 'Braund, Mr. Owen Harris', 'freq': 1}

Sex: {'count': 891, 'unique': 2, 'top': 'male', 'freq': 577}

Ticket: {'count': 891, 'unique': 681, 'top': '347082', 'freq': 7}

Cabin: {'count': 204, 'unique': 147, 'top': 'B96 B98', 'freq': 4}

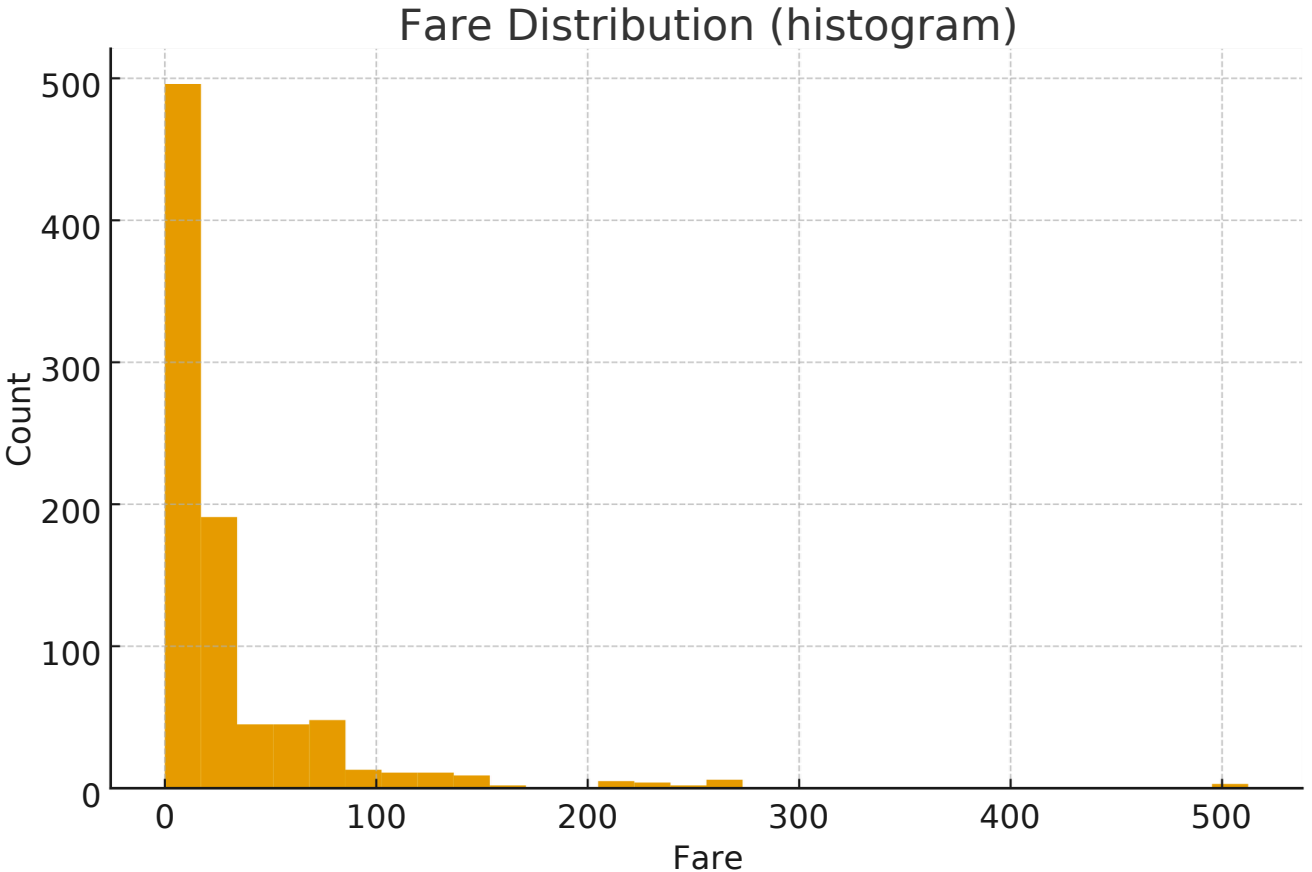
Embarked: {'count': 889, 'unique': 3, 'top': 'S', 'freq': 644}


Age Distribution (histogram)
Count
0
10
20
30
40
50
60
70
Age
0
10
20
30
40
50
60
70
80

Observation - Age Distribution

Observation:Age distribution shows most passengers are young adults.

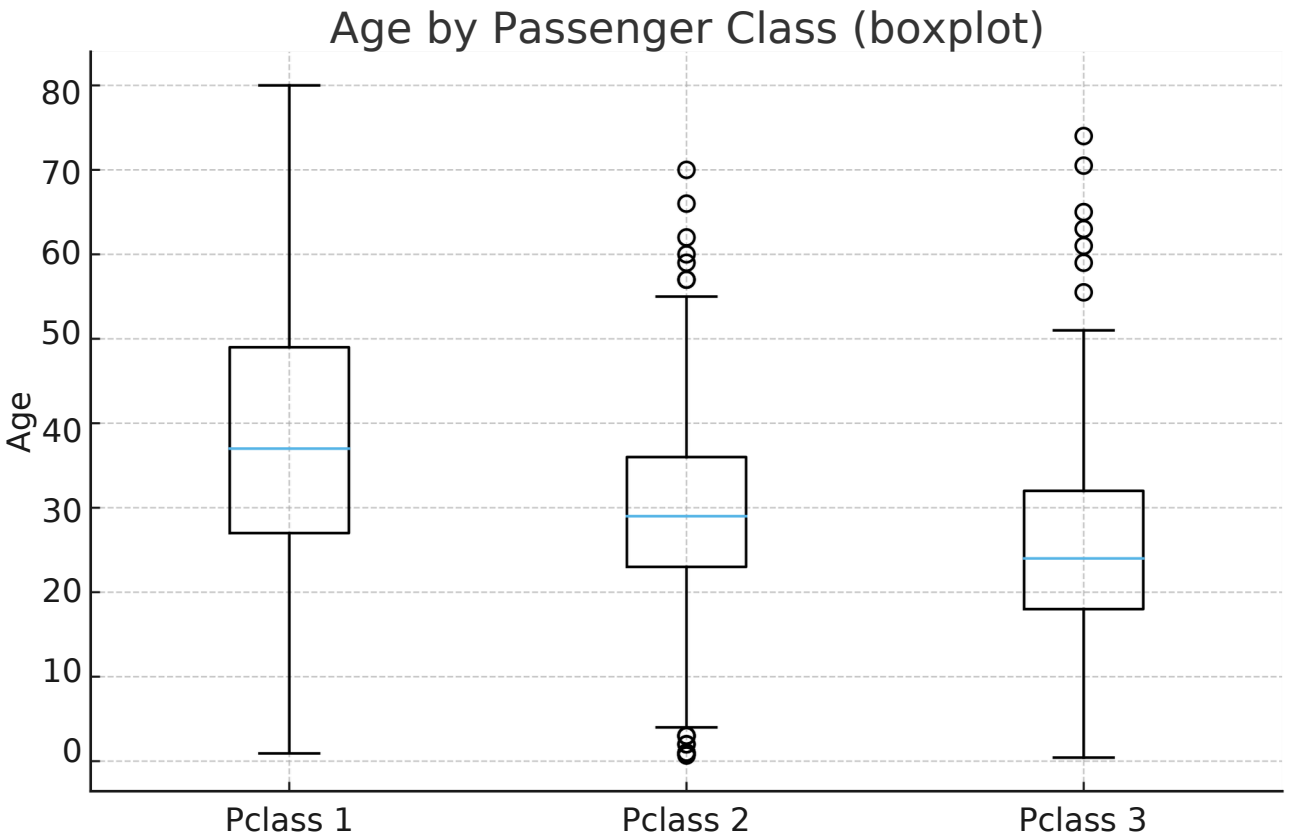
There are missing Age values which may require imputation.


Fare Distribution (histogram)
Count
Fare
0
100
200
300
400
500
0
100
200
300
400
500

Observation - Fare Distribution

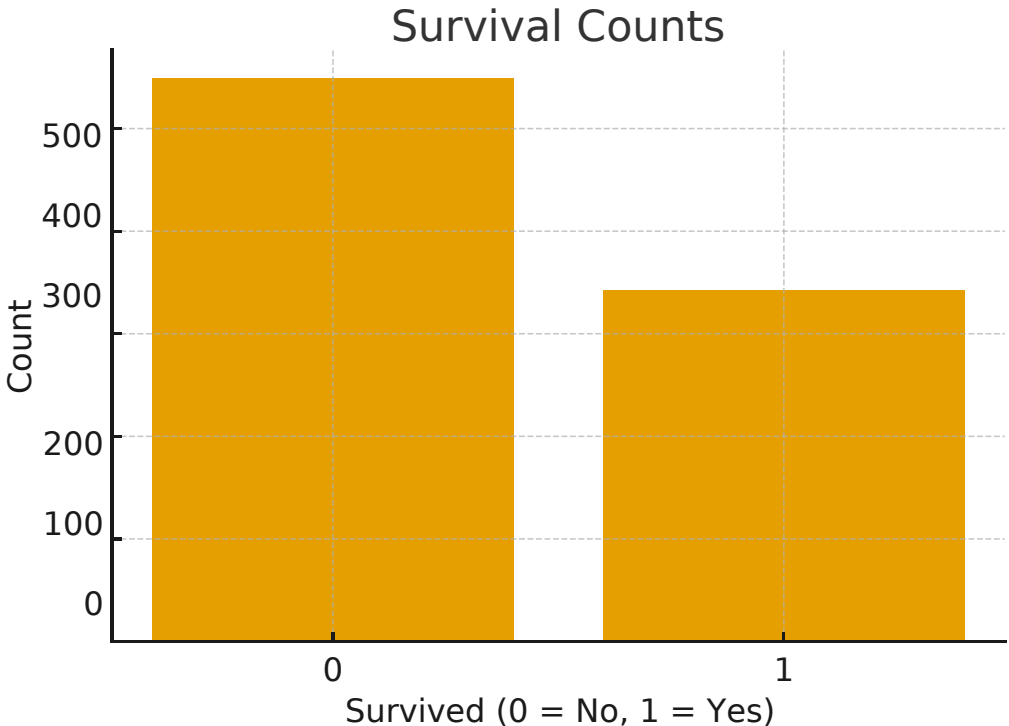
Observation:Faredistribution is right-skewed(some passengers paid very high fares).

Consider log-transforming Fare for models/visualization.


Age by Passenger Class (boxplot)
Age
80
70
60
50
40
30
20
10
0
Pclass 1
Pclass 2
Pclass 3

Observation - Age by Pclass

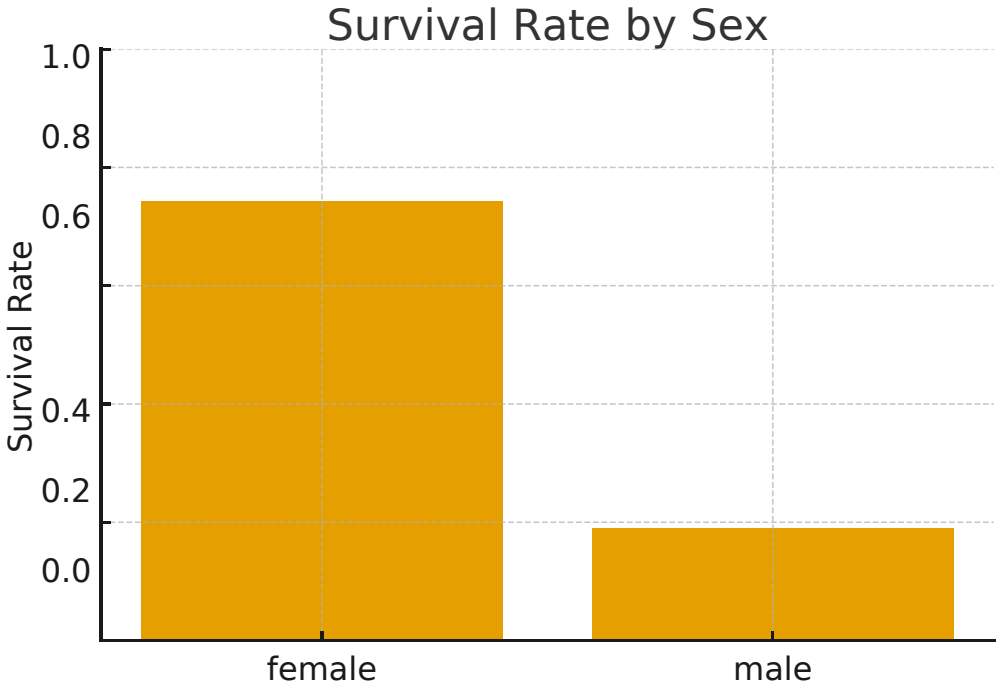
Observation:Higherclasspassengers(Pclass=1) tend to have higher median ages than lower classes.


Survival Counts
Count
500
400
300
200
100
0
0
1
Survived (0 = No, 1 = Yes)

Observation - Survival Counts

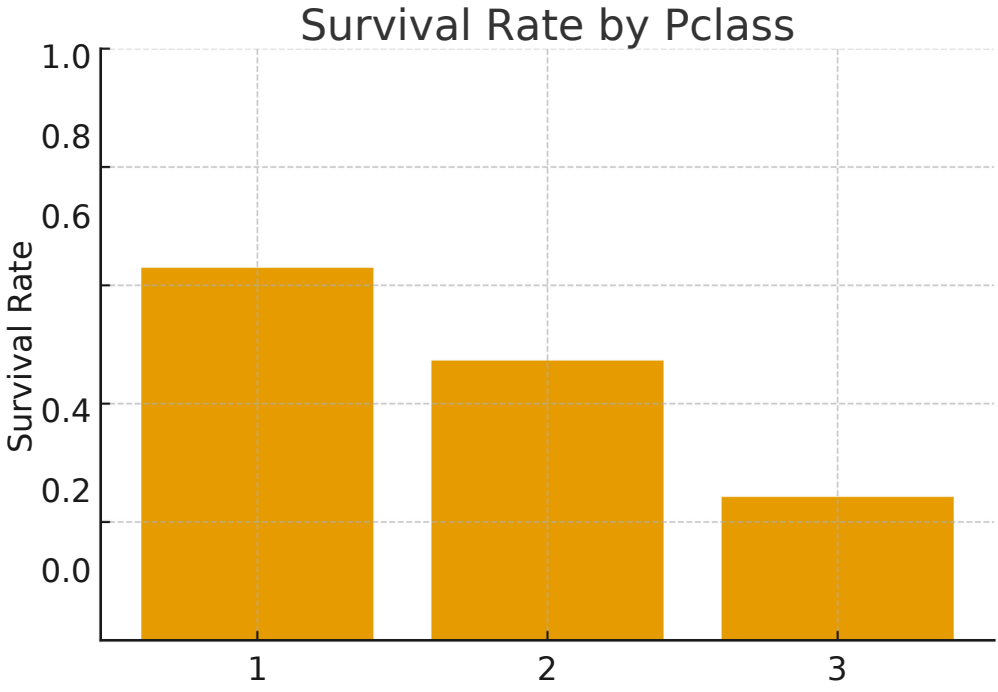
Observation: Survivors= 342, Non-survivors = 549.

Overall more passengers did not survive than survived.


Survival Rate by Sex
Survival Rate
1.0
0.8
0.6
0.4
0.2
0.0
female
male

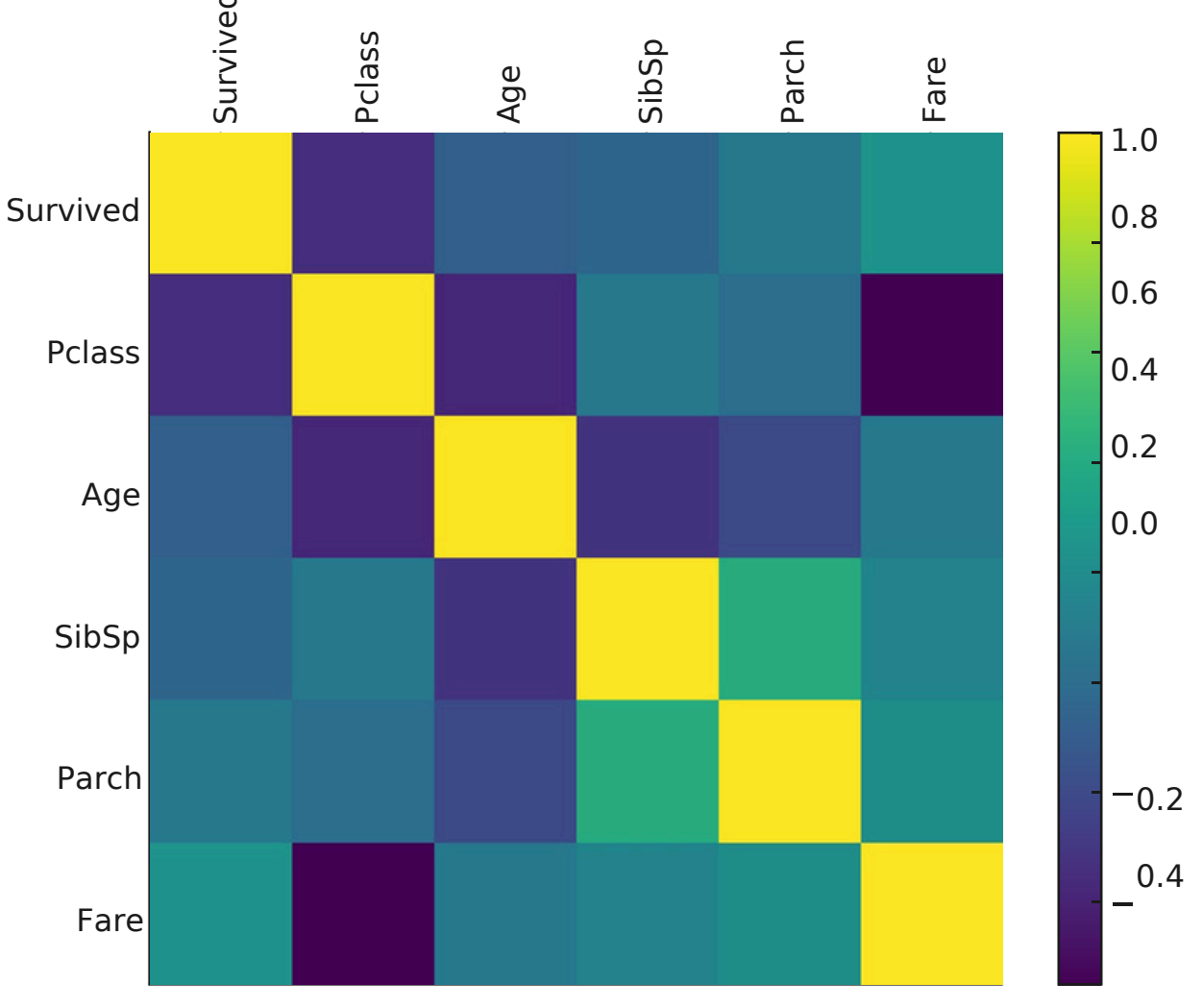
Observation - Survival Rate by Sex

Observation:Femaleshaveamuchhighersurvivalratethan males.


Survival Rate by Pclass
Survival Rate
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3

Observation - Survival Rate by Pclass

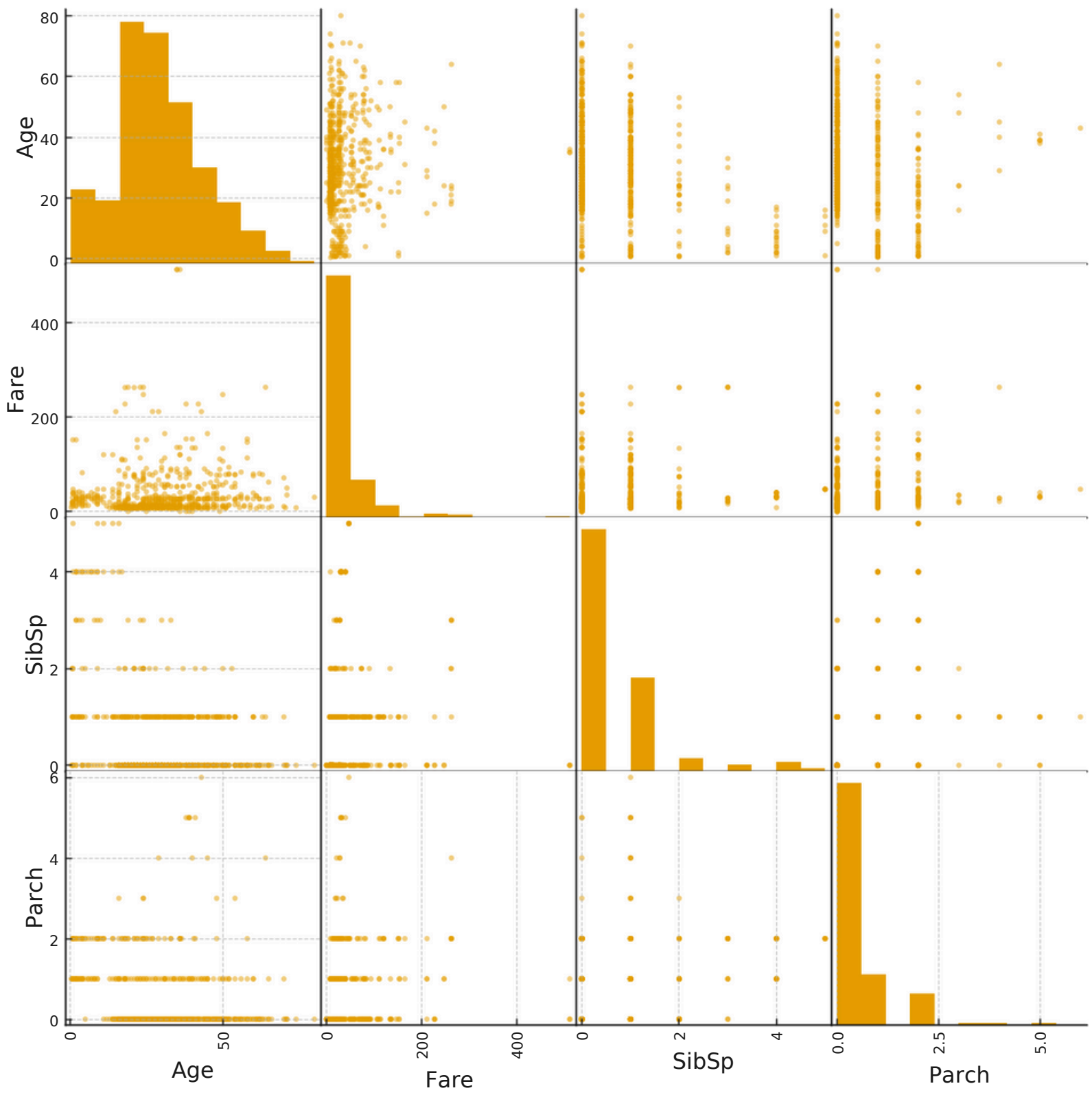
Observation:Firstclasspassengershadhighersurvivalrates than second and third class.


Survived
Pclass
Age
SibSp
Parch
Fare
Survived
Pclass
Age
SibSp
Parch
Fare
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Observation - Correlation Matrix

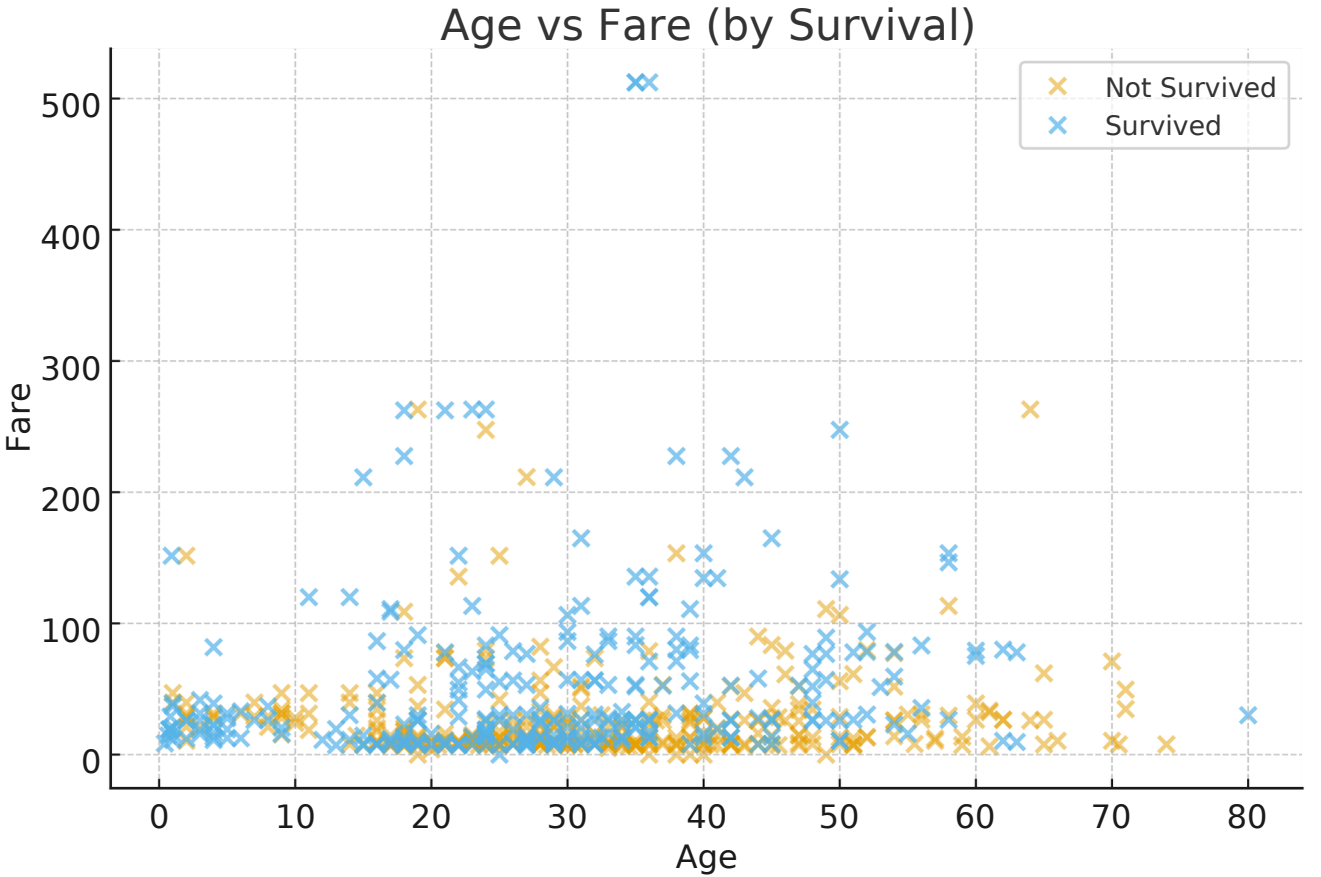
Observation:'Fare'hasapositive correlation with 'Survived'.

Age shows weak correlation with survival in raw form.


Scatter Matrix: Age, Fare, SibSp, Parch
Age
Fare
SibSp
Parch
0
20
40
60
80
200
400
2
4
6
50
2.5
5.0

Observation - Scatter Matrix

Observation:Pairwiserelationshipsareweak;Fare shows some spread compared to Age.


Age vs Fare (by Survival)
Not Survived
Survived
Fare
Age
0
100
200
300
400
500
0
10
20
30
40
50
60
70
80

Observation - Age vs Fare

Observation:Passengerswhopaidhigherfares show higher survival fraction; however, Age alone does

Summary & Next Steps

Key Findings:

- Females had a substantially higher survival rate than males.
- First-class passengers had higher survival rates and higher fares.
- Fare correlates positively with survival; Age has weak direct correlation.
- There are missing values in 'Age' and 'Cabin' which should be handled for modeling.

Suggested next steps:

- Impute missing Age values (median or model-based imputation).
- Extract title from 'Name' and analyze survival by title.
- Feature engineer family size from SibSp + Parch.
- Consider log-transforming Fare for models.