Exploratory Data Analysis on the Titanic Dataset

TASK 5 – DATA ANALYST INTERNSHIP
OBJECTIVE: EXTRACT INSIGHTS USING VISUAL AND STATISTICAL EXPLORATION
TOOLS USED: PYTHON (PANDAS, SEABORN, MATPLOTLIB)
SUBMITTED BY: *SIRISHA D*

Table of Contents

Introduction

The Titanic dataset provides detailed information about the passengers on the RMS Titanic. This project aims to uncover patterns that contributed to survival during the disaster by performing Exploratory Data Analysis (EDA).

Goals:

- Clean and prepare the dataset
- Identify patterns, trends, and anomalies
- Visualize insights related to survival

Dataset Overview

```
import pandas as pd
df = pd.read_csv("train.csv")
df.info()
df.describe()
df.isnull().sum()
```

Insights:

- 891 rows and 12 columns
- Missing values in Age, Cabin, and Embarked
- Target variable: Survived (0 = Died, 1 = Survived)

Data Cleaning

```
df['Age'].fillna(df['Age'].median(), inplace=True)
df['Embarked'].fillna(df['Embarked'].mode()[0],
    inplace=True)
df.drop(columns=['Cabin'], inplace=True)
```

Explanation:

- Age filled with median to avoid skew
- Embarked filled with mode (most frequent port)
- Cabin dropped (too many missing values)

Feature Engineering

```
df['FamilySize'] = df['SibSp'] + df['Parch'] + 1
df['Title'] = df['Name'].apply(lambda name:
    name.split(',')[1].split('.')[0].strip())
rare_titles = df['Title'].value_counts()[df['Title'].value_counts()
    < 10].index
df['Title'] = df['Title'].replace(rare_titles, 'Rare')
```

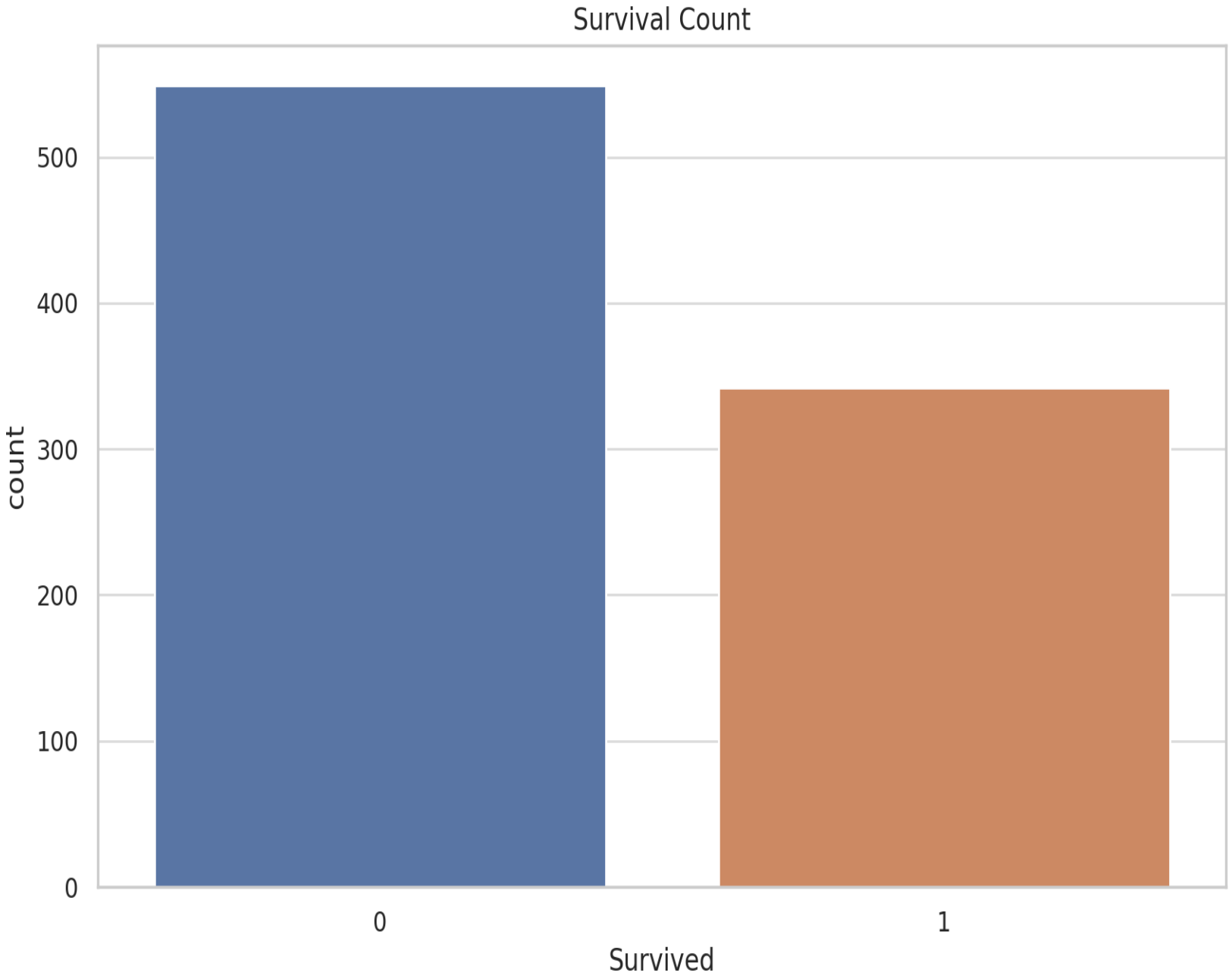
New Features:

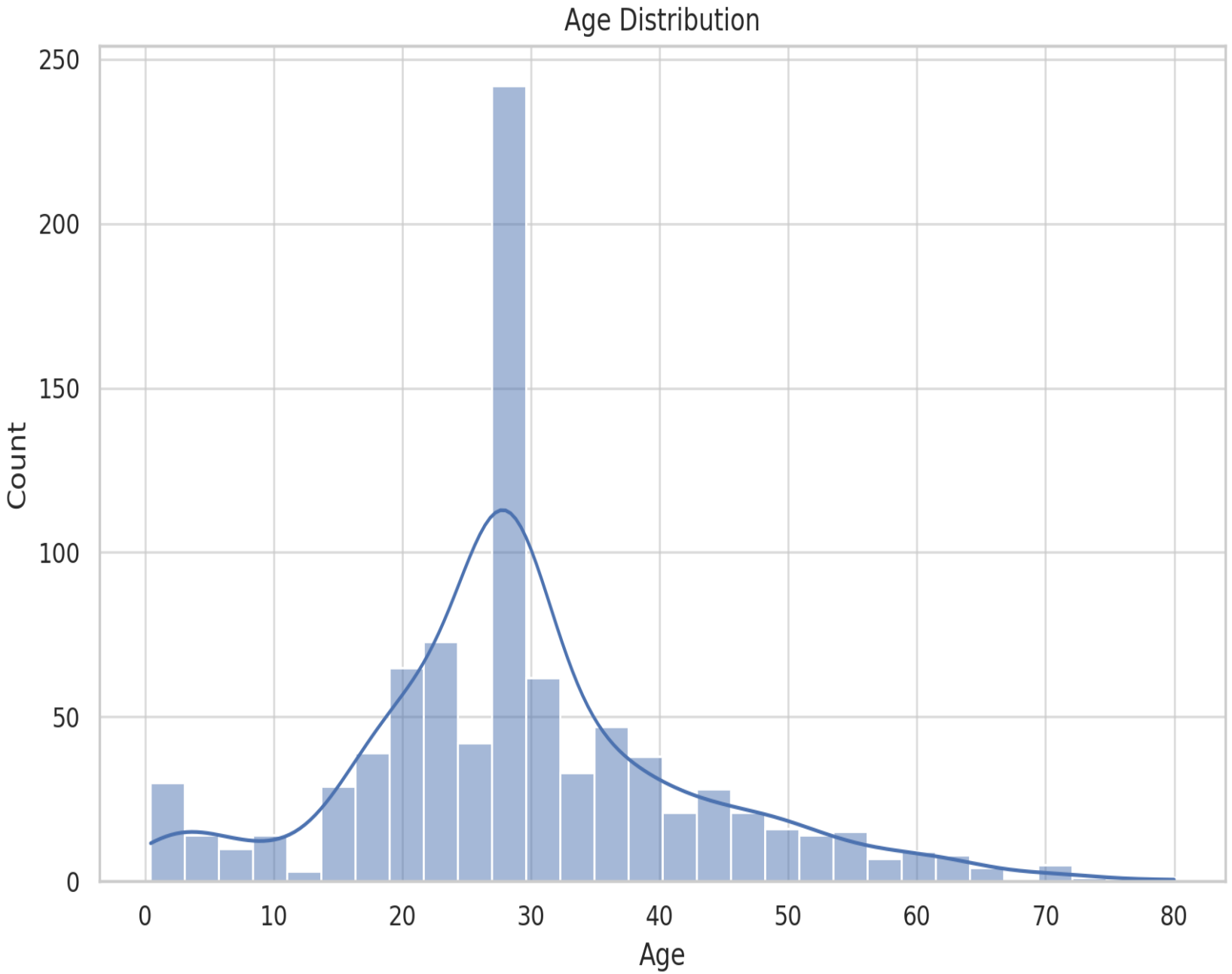
- FamilySize: Total number of family members aboard
- Title: Extracted from names, grouped rare ones

Univariate Analysis

```
import seaborn as sns
import matplotlib.pyplot as plt


sns.countplot(x='Survived', data=df)
sns.histplot(df['Age'], kde=True)
sns.countplot(x='Pclass', data=df)
sns.countplot(x='Sex', data=df)
```


Survival Count
500
400
300
200
100
0
count
0
1
Survived

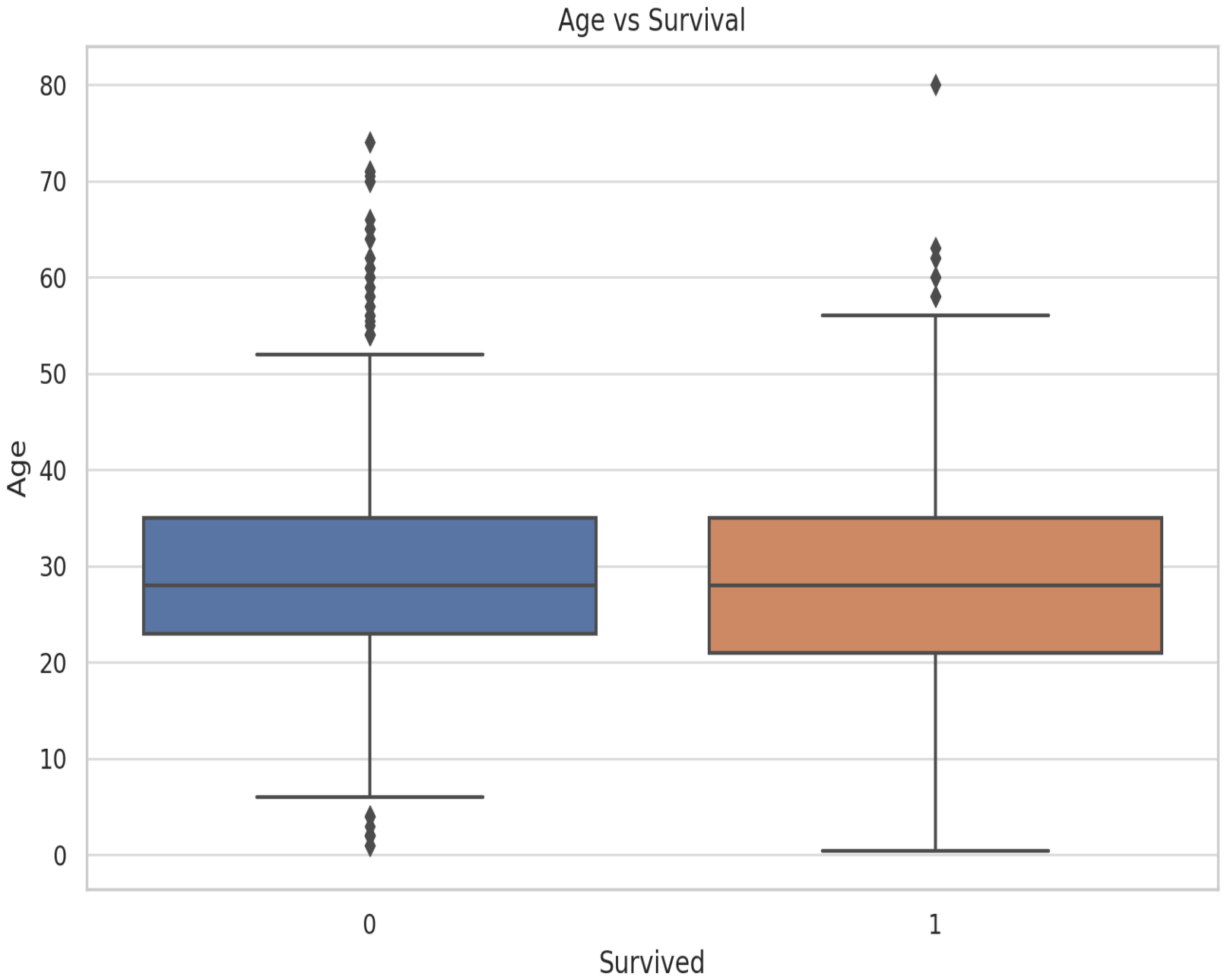

Age Distribution
Count
Age
0
50
100
150
200
250
0
10
20
30
40
50
60
70
80

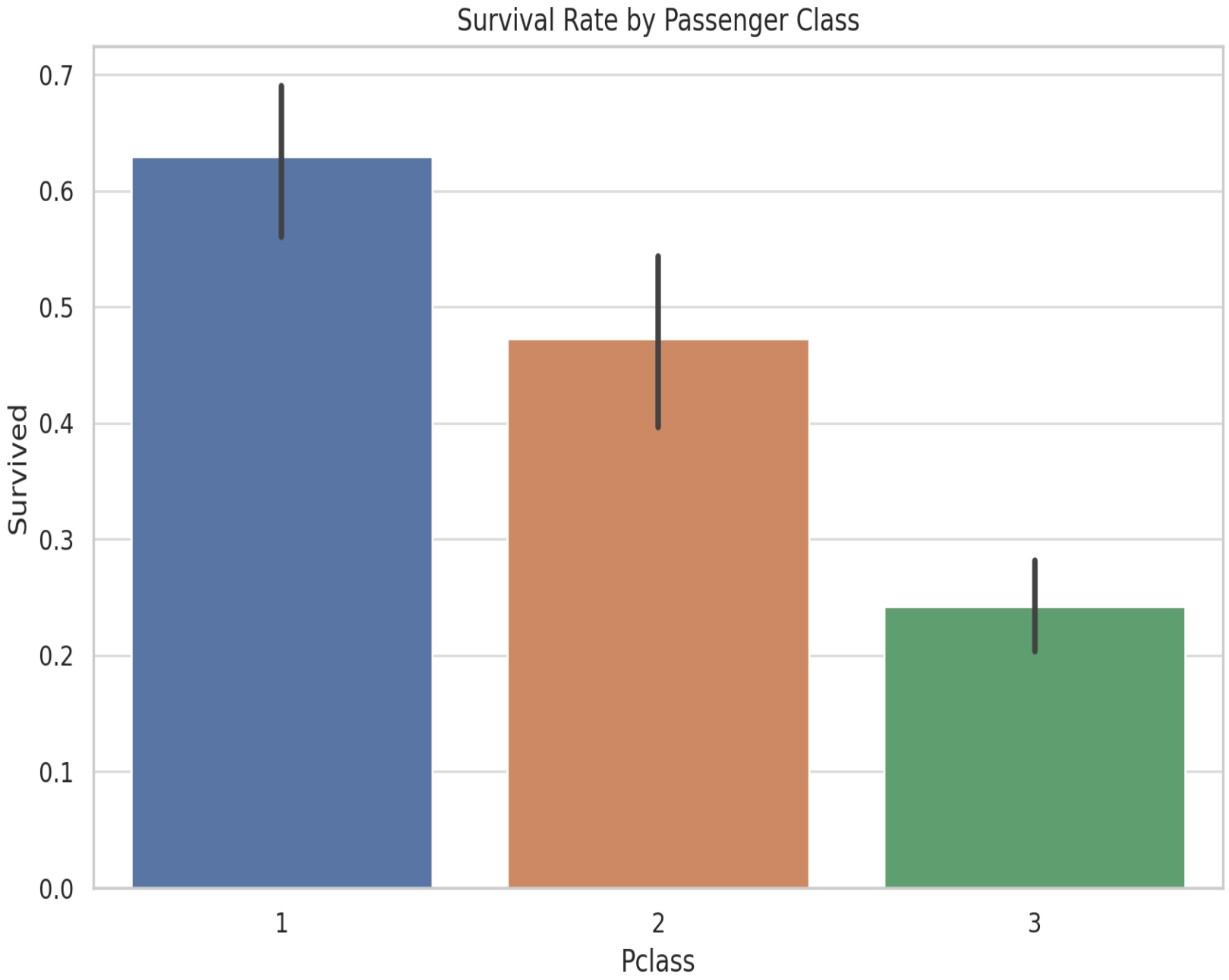
Observations

- Most passengers didn’t survive (62%)
- Age mostly between 20–40
- More 3rd class passengers
- More males than females

Bivariate Analysis

```
sns.boxplot(x='Survived', y='Age', data=df)
sns.barplot(x='Pclass', y='Survived', data=df)
sns.countplot(x='Sex', hue='Survived', data=df)
```


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

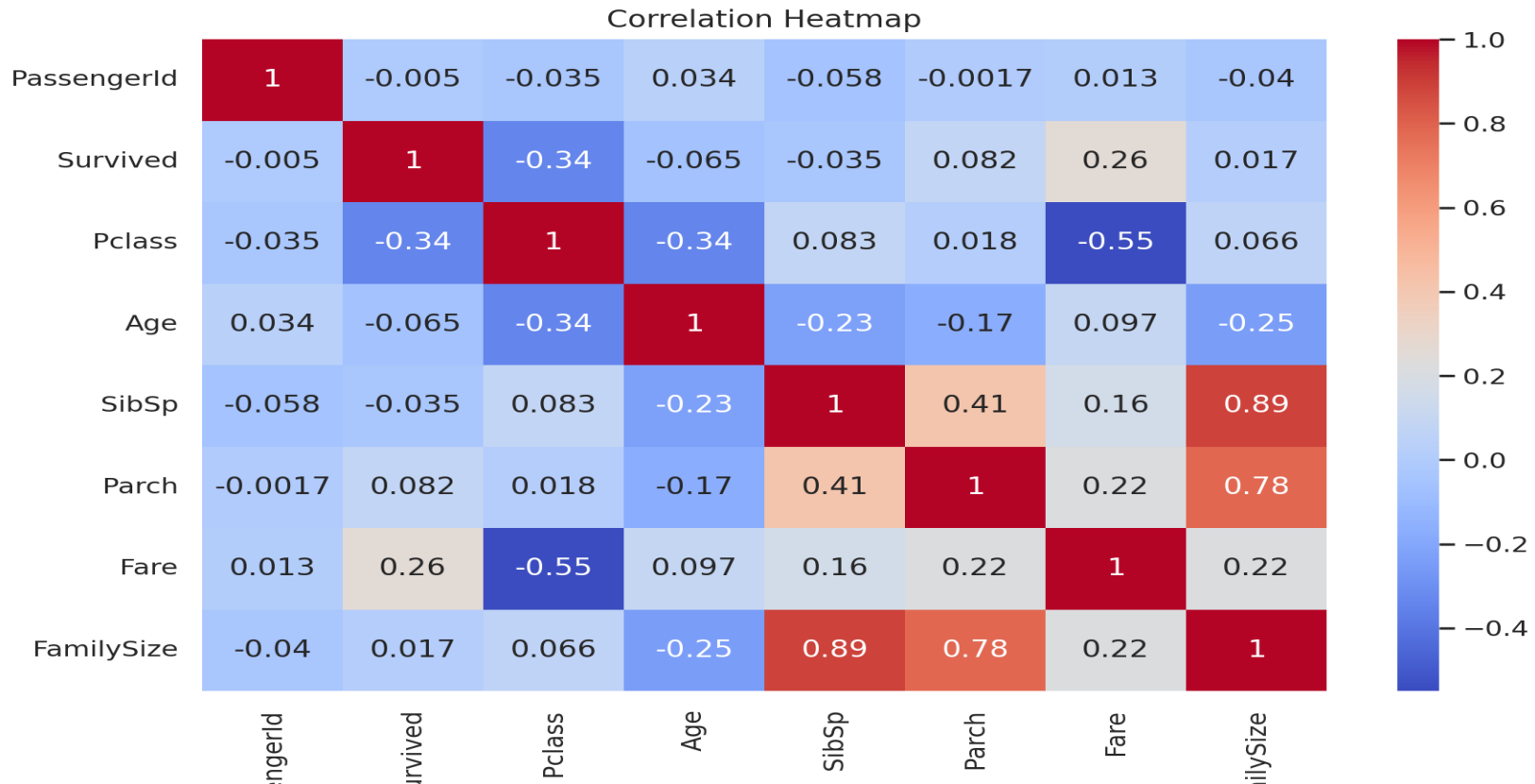

Survival Rate by Passenger Class
Survived
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
3
Pclass

Observations:

- Younger passengers had higher survival rates
- 1st class had better survival
- Female survival rate far higher than male

Multivariate Analysis

```
sns.heatmap(df.corr(numeric_only=True), annot=True,
    cmap='coolwarm')
sns.pairplot(df[['Survived', 'Pclass', 'Age', 'Fare']])
```



Observations:

- ‘Fare’, ‘Sex’, and ‘Pclass’ highly correlated with survival
- Pairplots show clusters in Age/Fare/Pclass vs Survival

Visual Observations Summary

Visual		Observation
Survival Count	-	Only ~38% survived
Age Distribution	-	Most passengers were 20–30 years old
Age vs Survival	-	Younger passengers had better survival
Class vs Survival	-	1st class had highest survival
Gender vs Survival	-	Females survived more
Correlation	-	Fare & Sex positively linked to survival

Summary of Findings

- Women had a much higher survival rate
- First-class passengers were more likely to survive
- Children and small families had better survival odds
- Higher fare = better cabins = higher survival
- Titles like ‘Miss’ and ‘Mrs’ helped identify survival trends

Conclusion

- This analysis highlights how age, gender, class, and social titles significantly impacted survival on the Titanic.
- These insights are valuable in understanding how demographic factors influence crisis outcomes and can guide planning and safety strategies.