Titanic Dataset - EDA and Feature Engineering

This report contains Exploratory Data Analysis (EDA) and Feature Engineering steps performed on the Titanic Dataset.

It includes Python code, output summaries, and visualizations to understand the factors affecting survival and to prepare

the dataset for machine learning model building.

Python Code Used

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder


data = pd.read_csv("Titanic-Dataset.csv")


# Data Cleaning
data['Age'].fillna(data['Age'].median(), inplace=True)
data['Embarked'].fillna(data['Embarked'].mode()[0], inplace=True)
data.drop('Cabin', axis=1, inplace=True)


# Feature Engineering
data['FamilySize'] = data['SibSp'] + data['Parch'] + 1
data['IsAlone'] = 1
data.loc[data['FamilySize'] > 1, 'IsAlone'] = 0
data['Title'] = data['Name'].str.extract(' ([A-Za-z]+)\.', expand=False)
data['Title']                                                     =
data['Title'].replace(['Lady','Countess','Capt','Col','Don','Dr','Major','Rev','Sir','Jonkheer','D
ona'],'Rare')
data['Title'] = data['Title'].replace({'Mlle':'Miss','Ms':'Miss','Mme':'Mrs'})


# Encoding
label = LabelEncoder()
data['Sex'] = label.fit_transform(data['Sex'])
data['Embarked'] = label.fit_transform(data['Embarked'])
data['Title'] = label.fit_transform(data['Title'])
data.drop(['Name','Ticket'], axis=1, inplace=True)
```
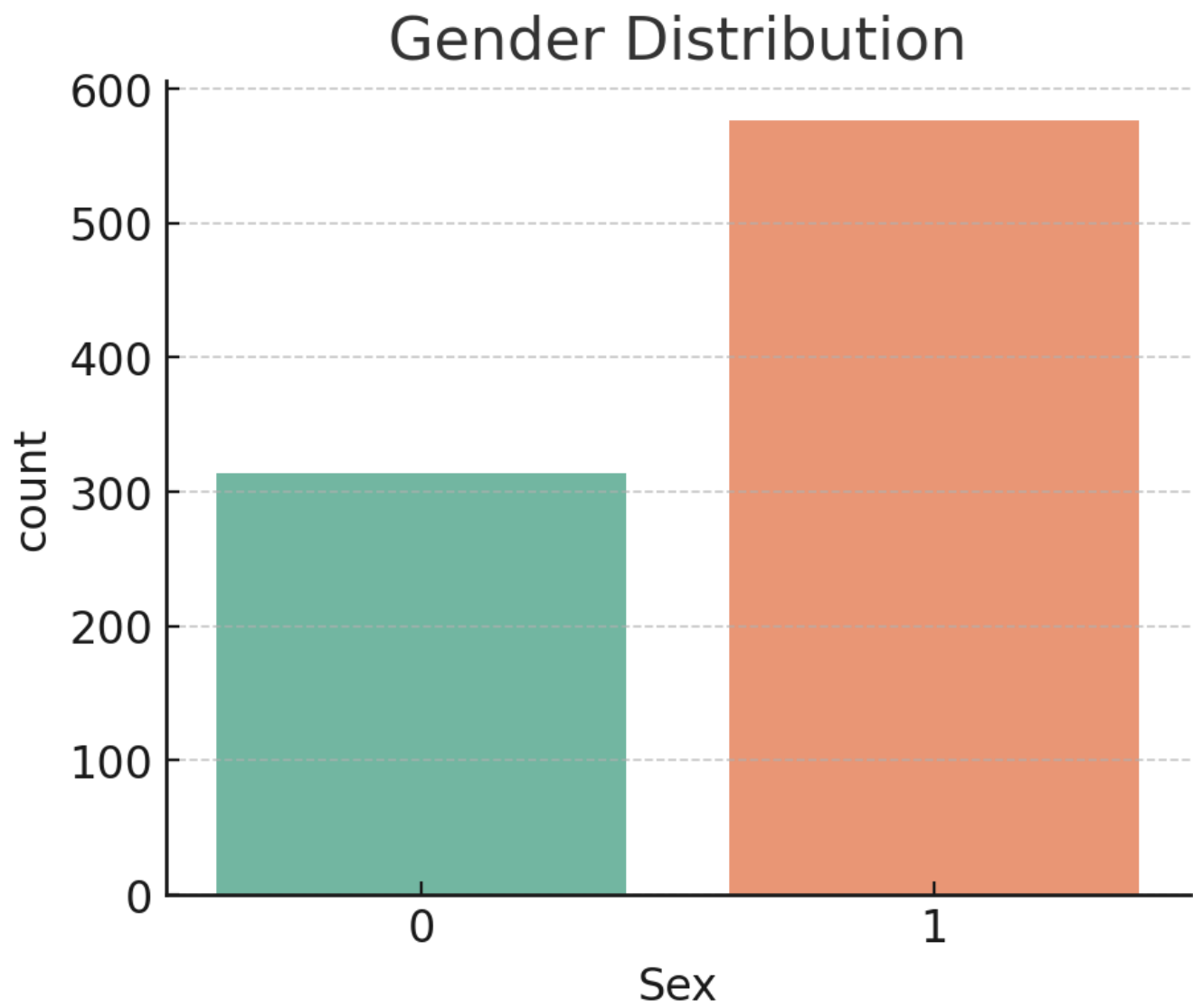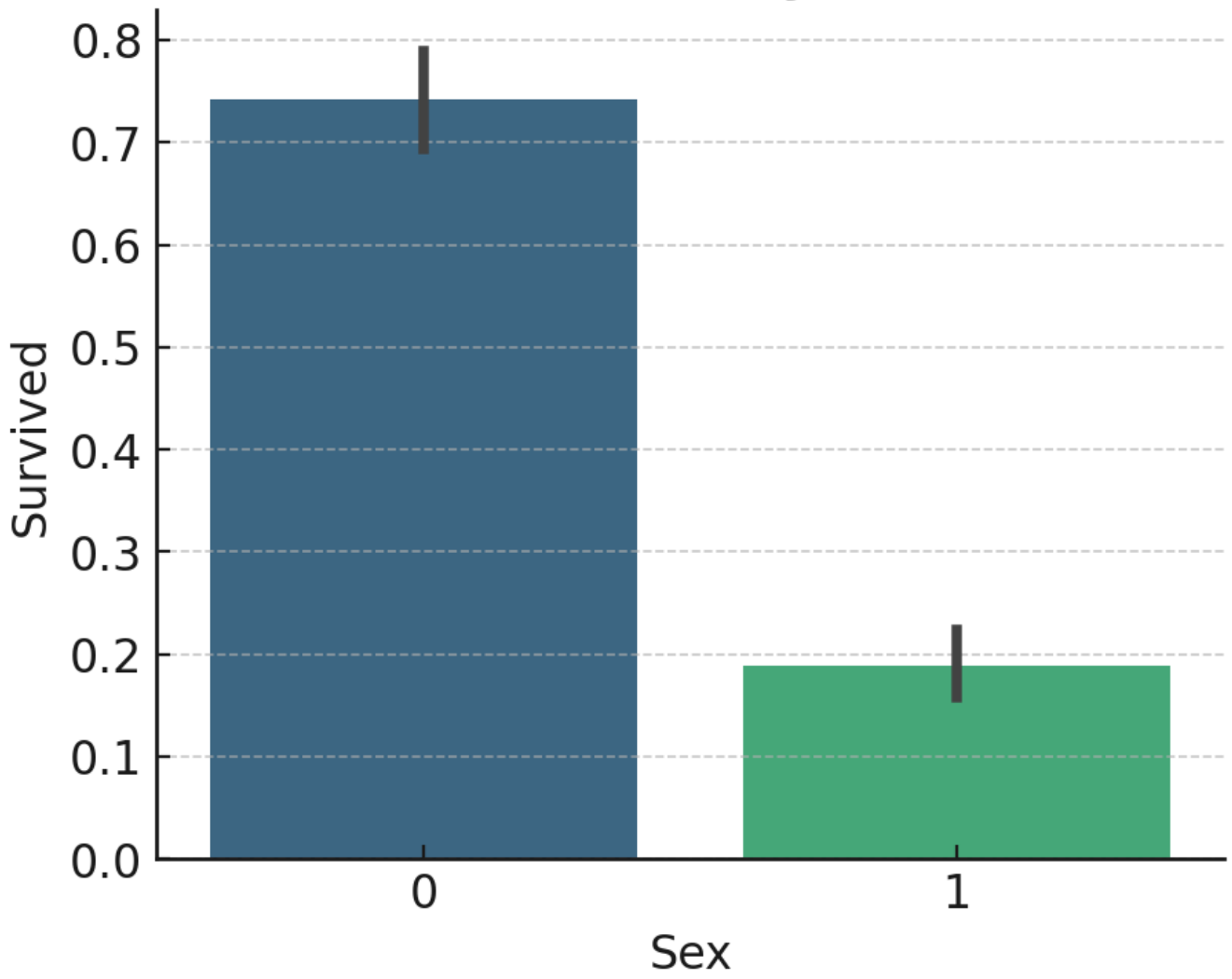
Output Summary

- Dataset Shape: (891, 12)

- Missing Columns: Age, Cabin, Embarked

- Cleaned: Filled Age and Embarked, Dropped Cabin

- New Columns: FamilySize, IsAlone, Title

- Encoded: Sex, Embarked, Title

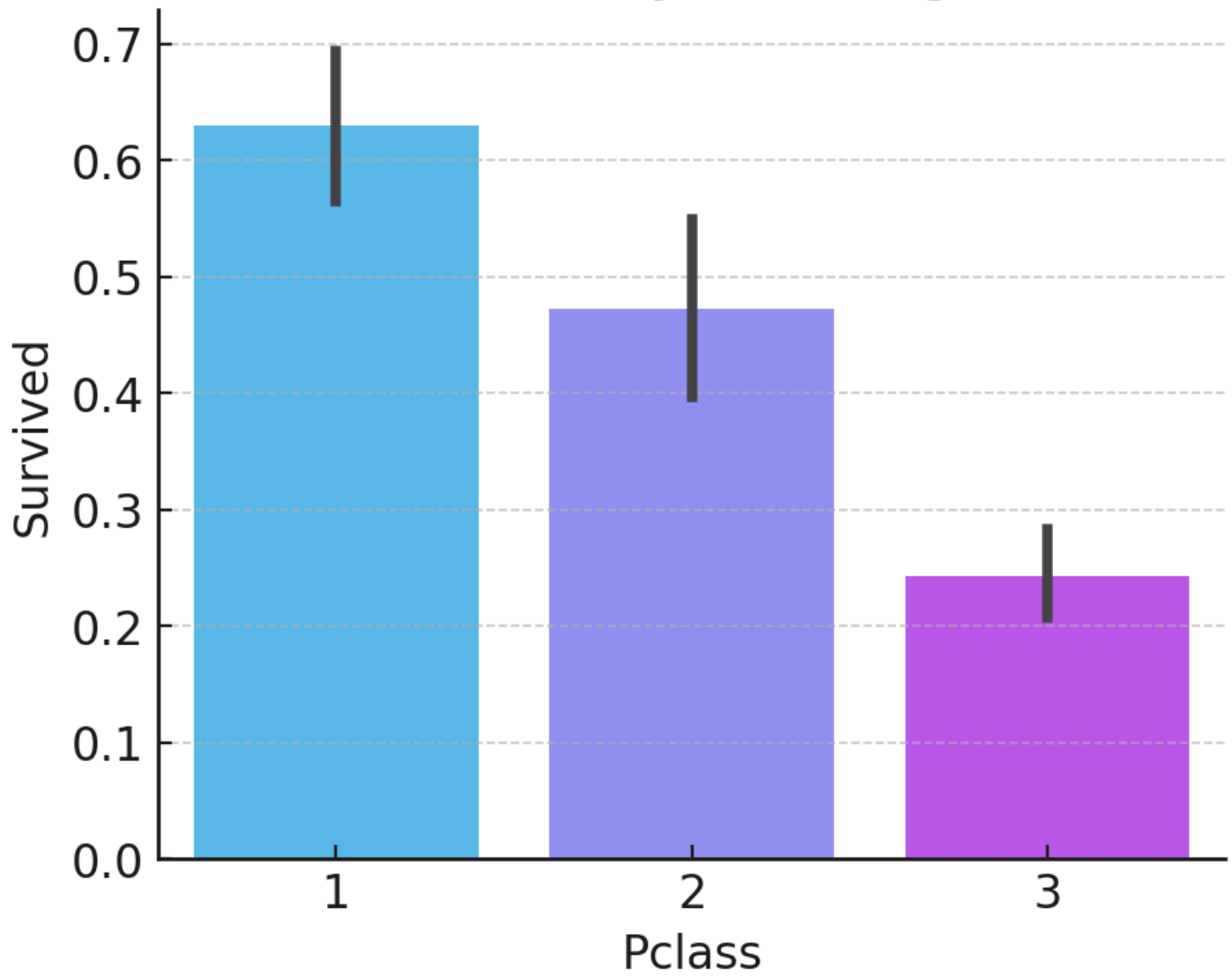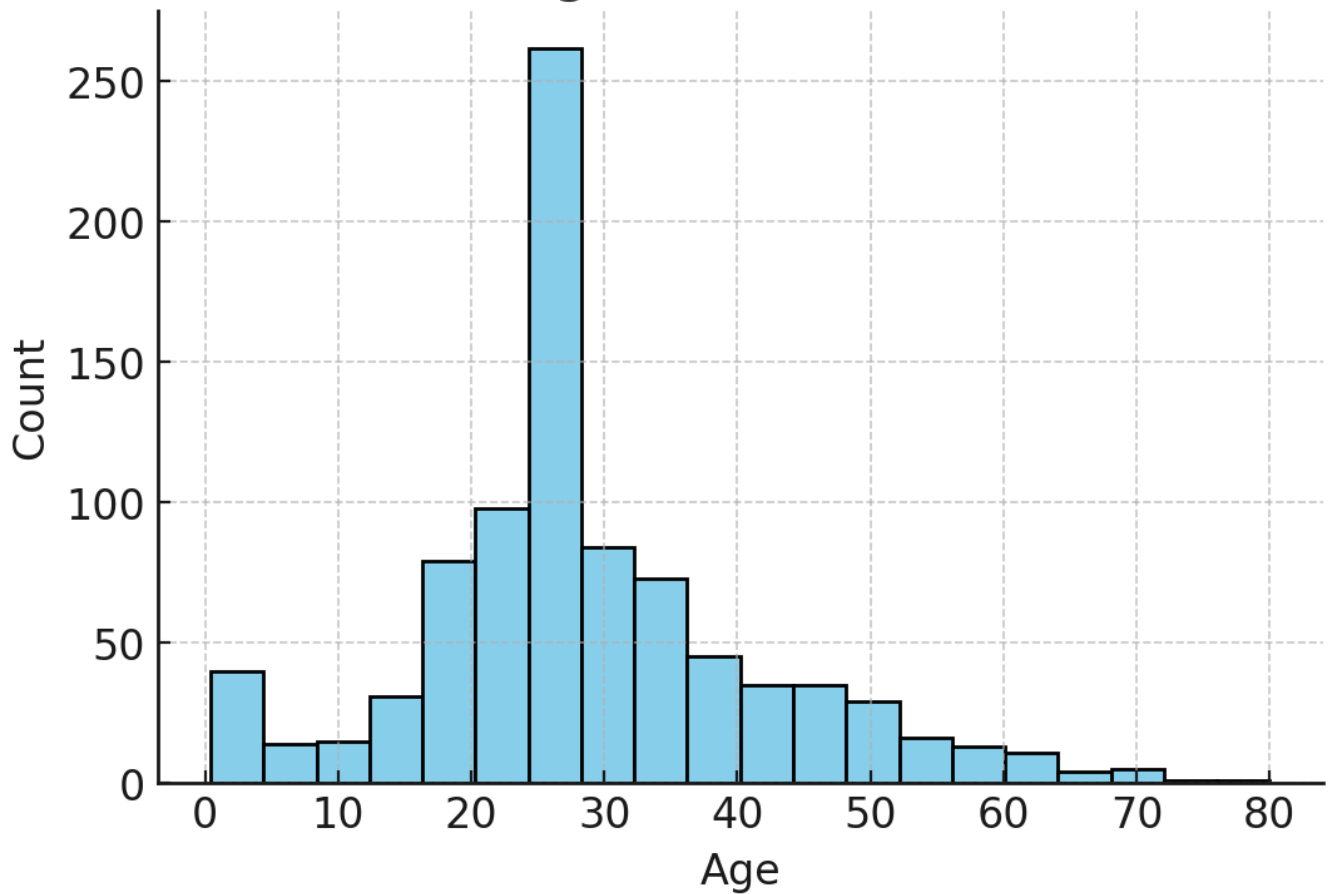- Final Dataset Ready for Model Building

Visualizations

Survival Rate by Gender

Survival Rate by Passenger Class

Age Distribution

Correlation Heatmap