task5-1

April 19, 2025

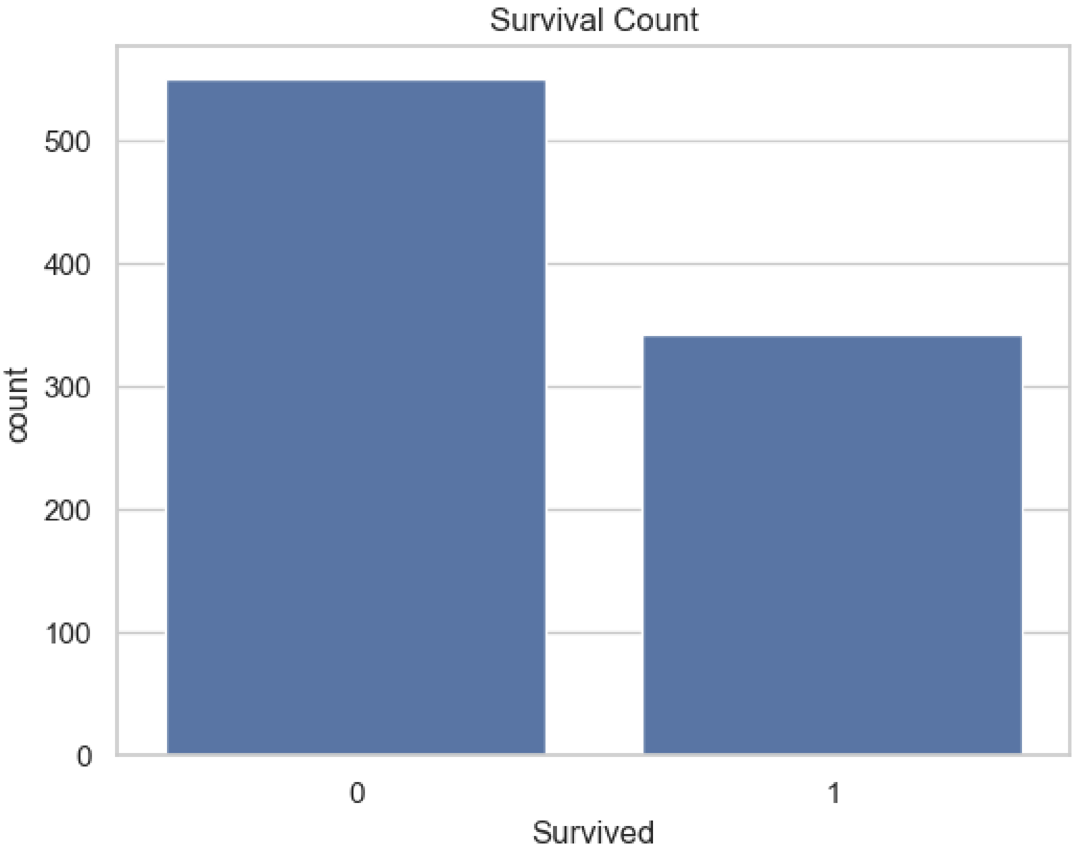
[9]:

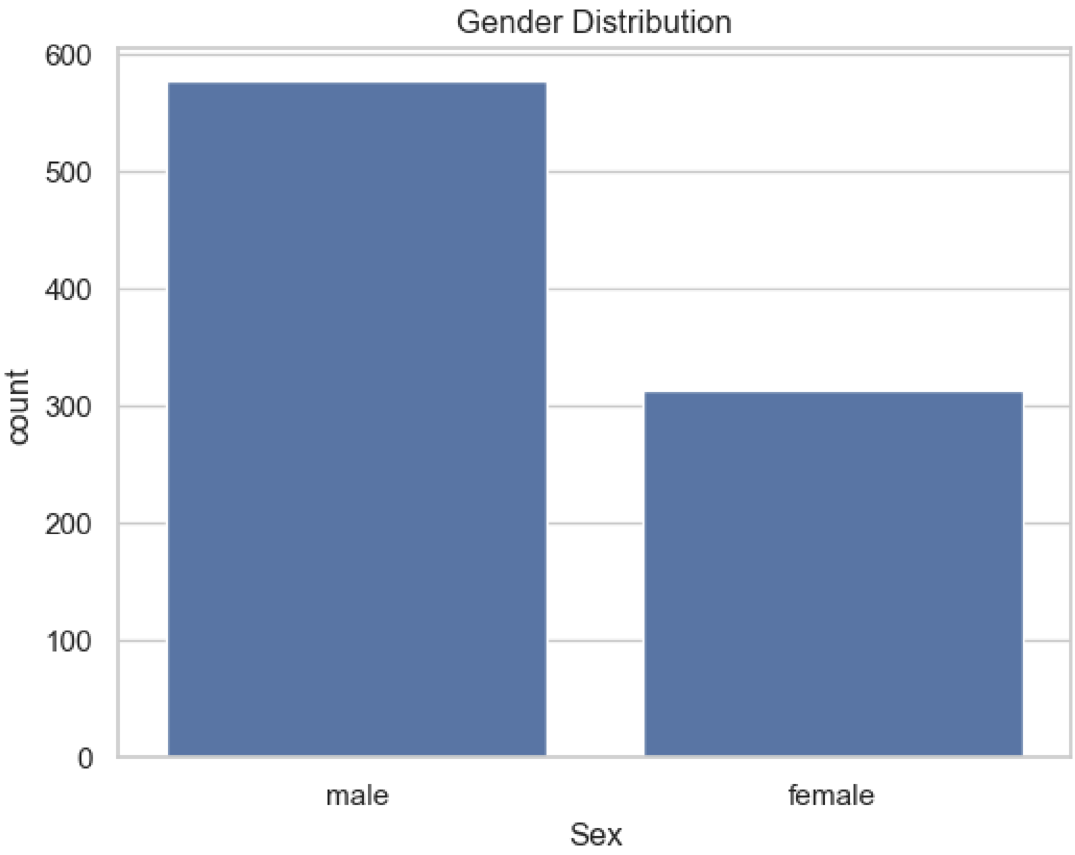
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Set seaborn style
sns.set(style="whitegrid")

# Load the dataset
df = pd.read_csv(r"C:\Users\NIHAR\OneDrive\Documents\train (1).csv")

# Basic structure
df.info()

# Statistical summary
df.describe()

# Check missing values
df.isnull().sum()

# First 5 rows
df.head()
# Value counts for categorical features
print(df['Survived'].value_counts())
print(df['Sex'].value_counts())
print(df['Pclass'].value_counts())

# Plot barplots
sns.countplot(x='Survived', data=df)
plt.title('Survival Count')
plt.show()
sns.countplot(x='Sex', data=df)
plt.title('Gender Distribution')
plt.show()

sns.countplot(x='Pclass', data=df)
plt.title('Passenger Class Distribution')
plt.show()
```

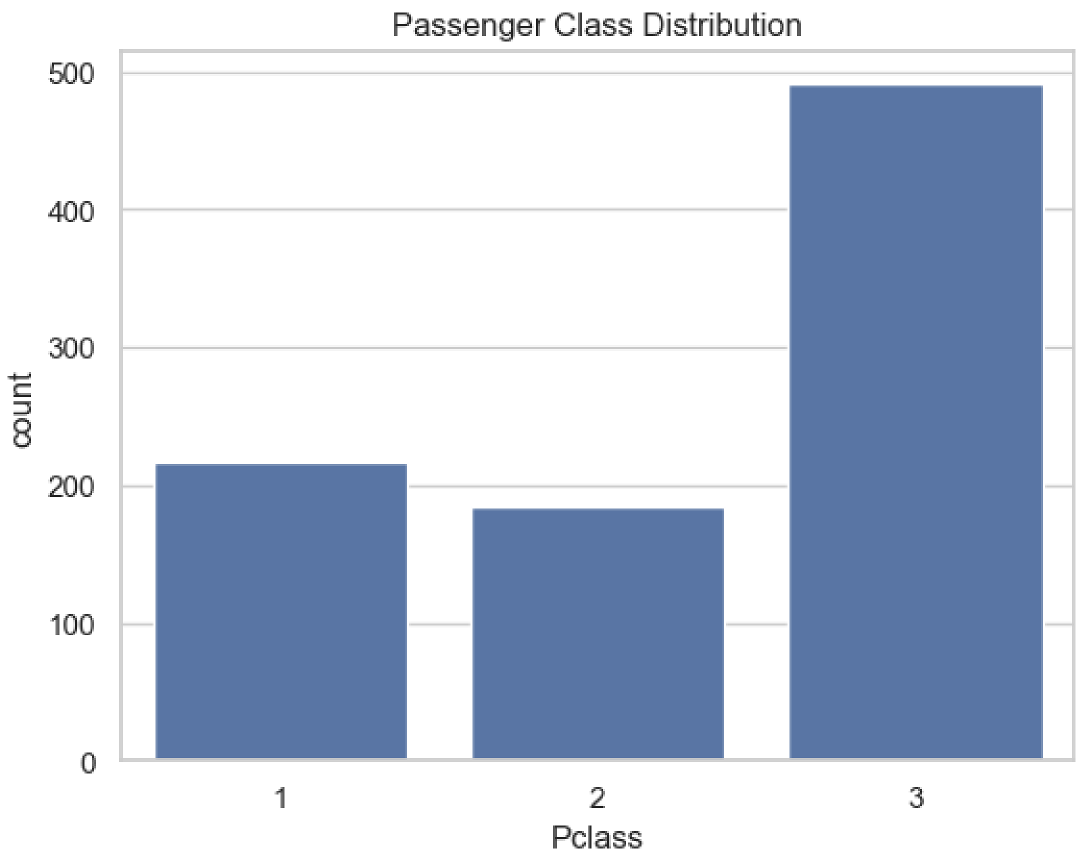
```
# Histograms
df['Age'].hist(bins=20)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()

df['Fare'].hist(bins=20)
plt.title('Fare Distribution')
plt.xlabel('Fare')
plt.ylabel('Count')
plt.show()

# Boxplots
sns.boxplot(y='Age', data=df)
plt.title('Boxplot of Age')
plt.show()
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title('Survival by Passenger Class')
plt.show()

sns.countplot(x='Sex', hue='Survived', data=df)
plt.title('Survival by Gender')
plt.show()

# Only for numerical features
# Select only numeric columns for correlation matrix
numeric_df = df.select_dtypes(include=['number'])

plt.figure(figsize=(10, 8))
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Correlation Matrix of Numeric Features")
plt.show()


# Check again
df.isnull().sum()

# Fill missing Age with median
df['Age'] = df['Age'].fillna(df['Age'].median())


# Drop 'Cabin' or fill with 'Unknown'
df.drop(columns='Cabin', inplace=True)  # or use: df['Cabin'].fillna('Unknown', inplace=True)
```

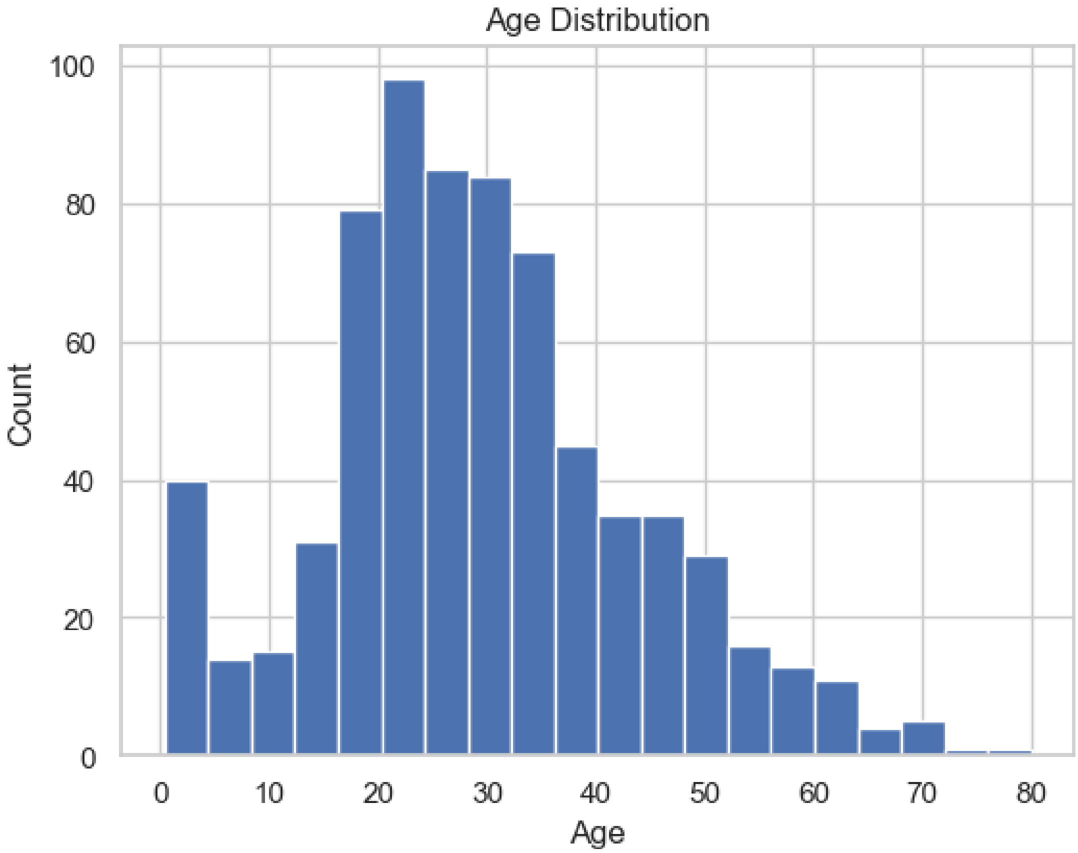
```python
# Drop remaining nulls
df.dropna(inplace=True)
```

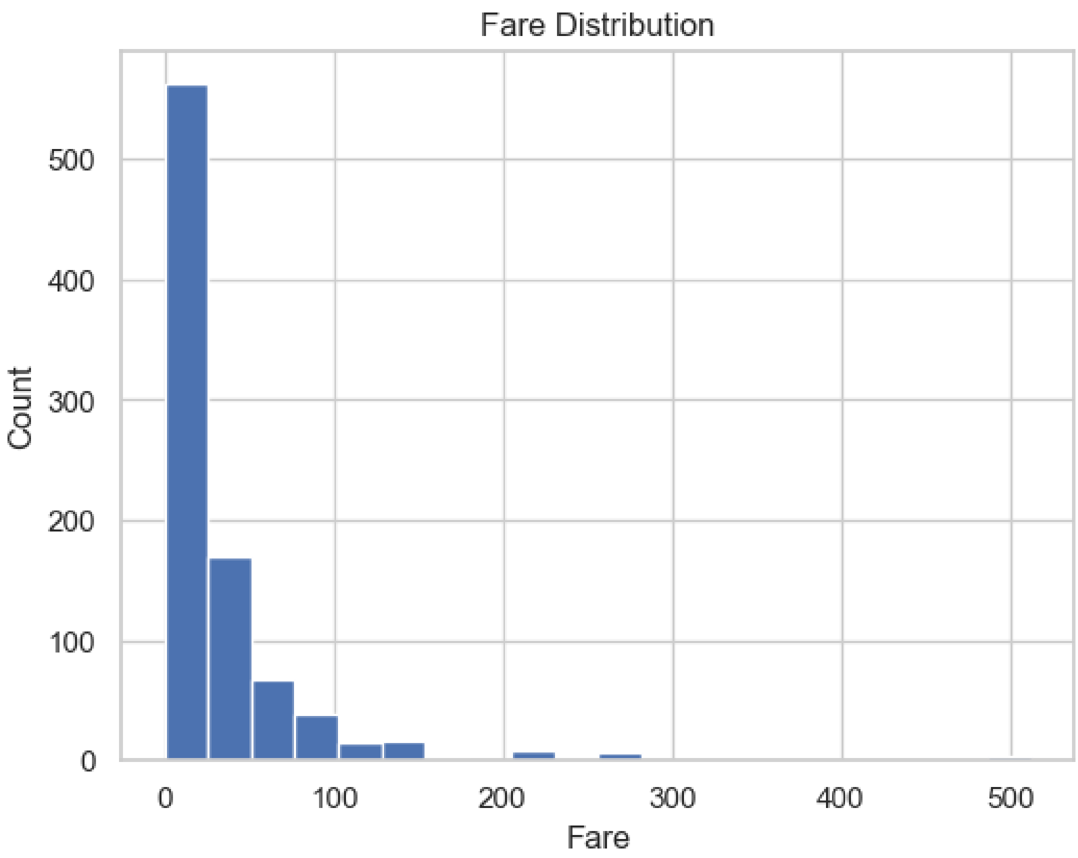
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    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
Survived
0    549
1    342
Name: count, dtype: int64
Sex
male      577
female    314
Name: count, dtype: int64
Pclass
3    491
1    216
2    184
Name: count, dtype: int64
```

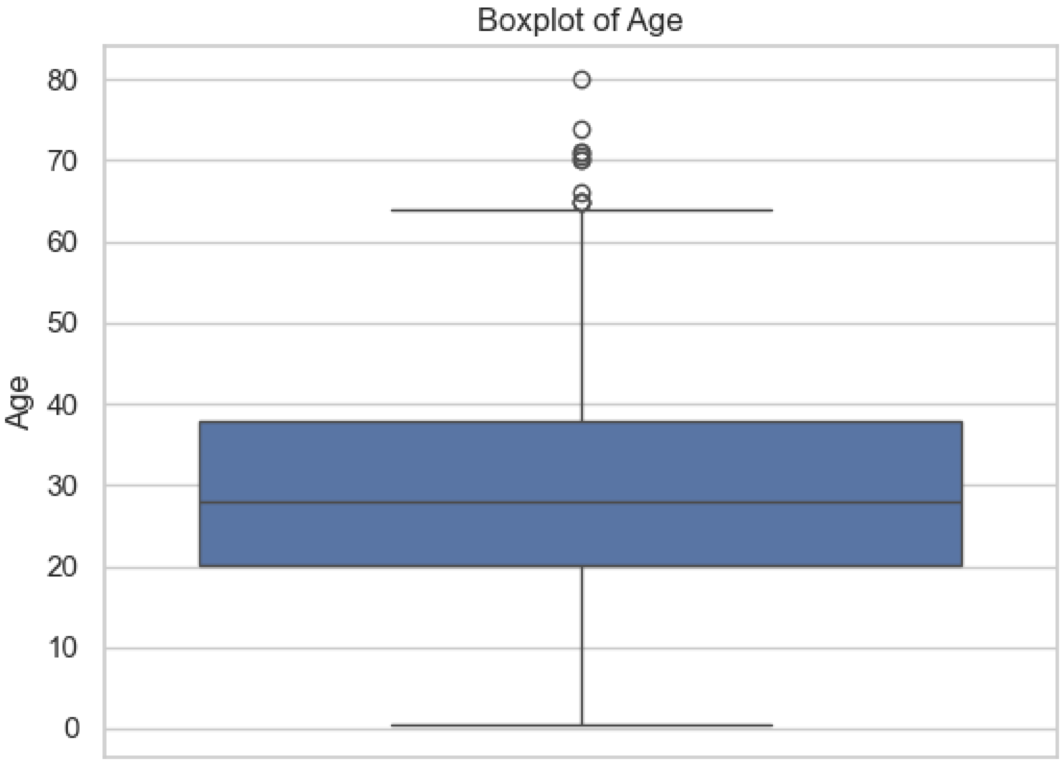

Survival Count
count
500
400
300
200
100
0
0
1
Survived

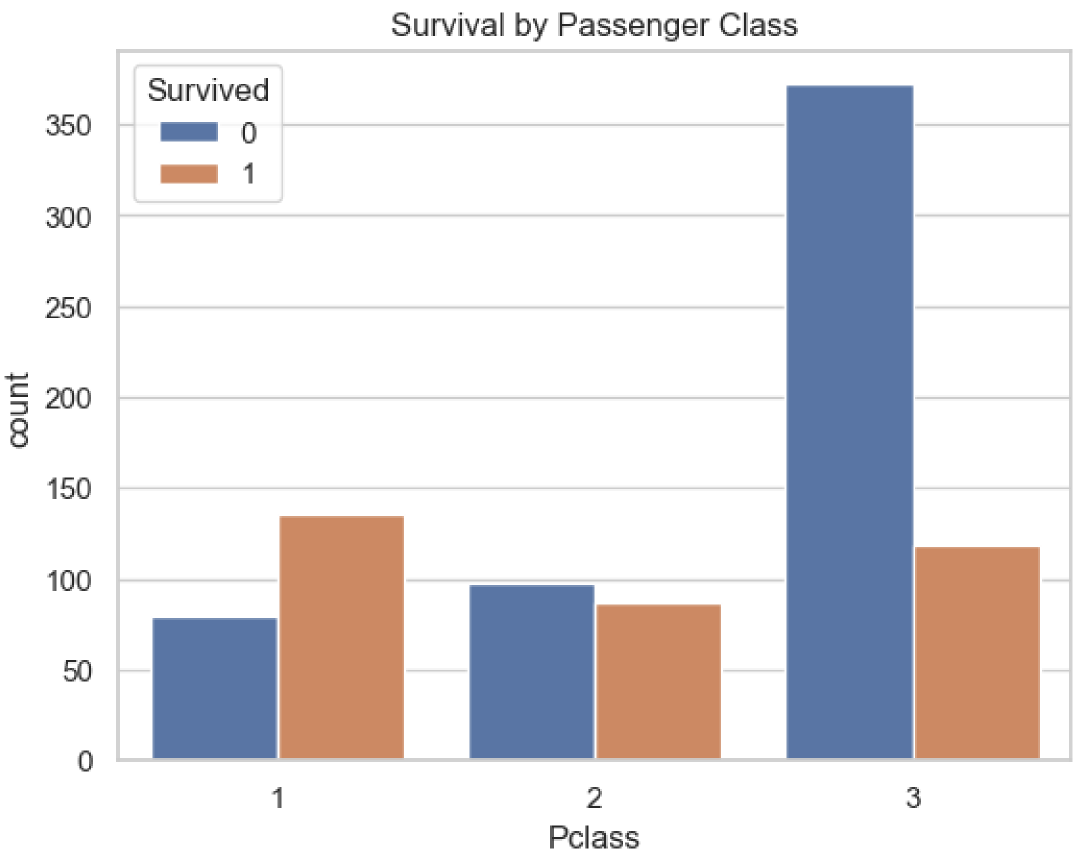

Gender Distribution
600
500
400
300
200
100
0
count
male
female
Sex

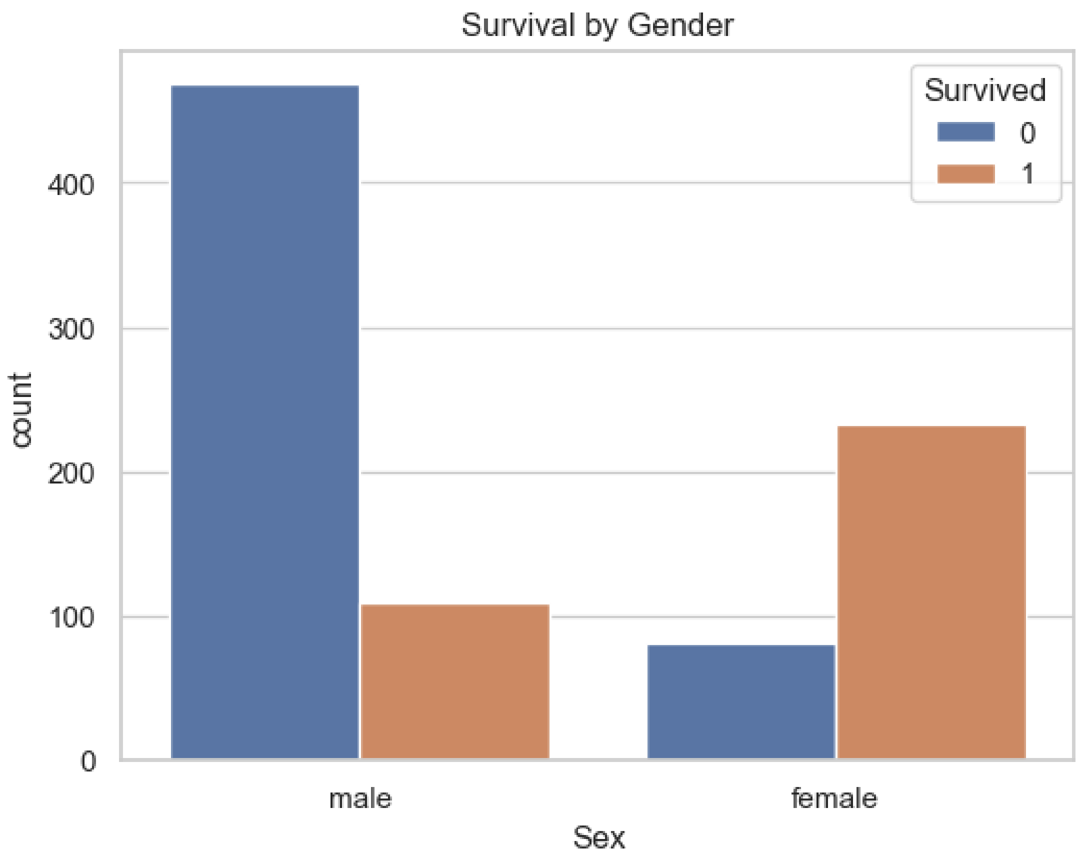

Passenger Class Distribution
500
400
300
200
100
0
count
1
2
3
Pclass

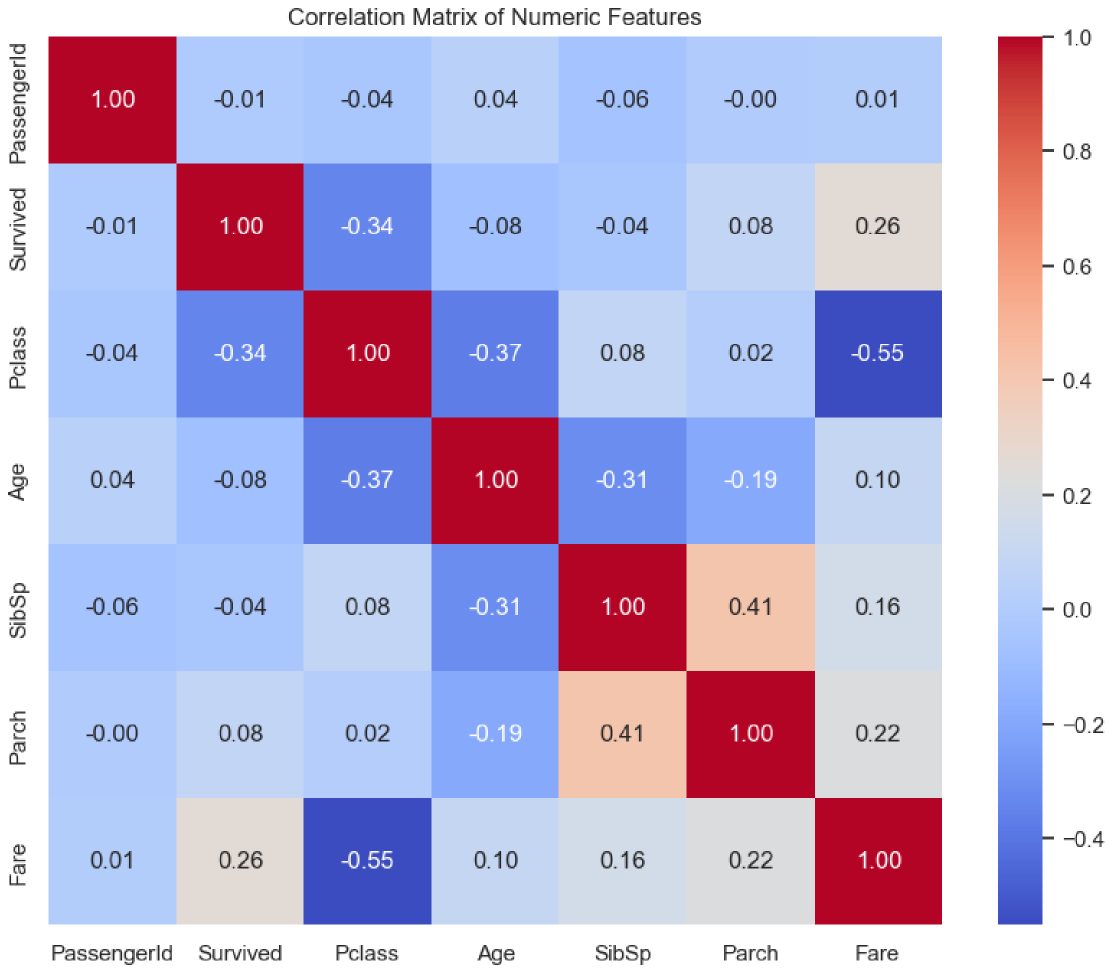

Age Distribution
Count
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
Age


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


Boxplot of Age
Age
80
70
60
50
40
30
20
10
0


Survival by Passenger Class
Survived
0
1
count
0
50
100
150
200
250
300
350
1
2
3
Pclass


Survival by Gender
Survived
0
1
count
400
300
200
100
0
male
female
Sex



[ ]:

```
File > Download as > PDF via LaTeX (.pdf)
```