Exploratory Data Analysis on the Titanic Dataset

Objective: Understand survival patterns based on gender, class, age, and other features.

Tools Used: Python, Pandas, Seaborn, Matplotlib

Load the Data

```
import pandas as pd

df = pd.read_csv("sample_titanic_data.csv")

df.head()  # View the first 5 rows
```

Understand the Structure

```
df.info()      # Data types and missing values

df.describe()  # Summary stats for numeric columns

df.columns     # List of column names
```

Clean the Data

```
df.isnull().sum()  # Count missing values

# Fill missing 'Age' with median

df['Age'].fillna(df['Age'].median(), inplace=True)

# Fill missing 'Embarked' with mode

df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
```

Analyze the Data

A. Survival Rate

```
df['Survived'].value_counts(normalize=True)
```

B. Survival by Gender

```
pd.crosstab(df['Sex'], df['Survived'], normalize='index')
```

C. Average Age by Class

```
df.groupby('Pclass')['Age'].mean()
```

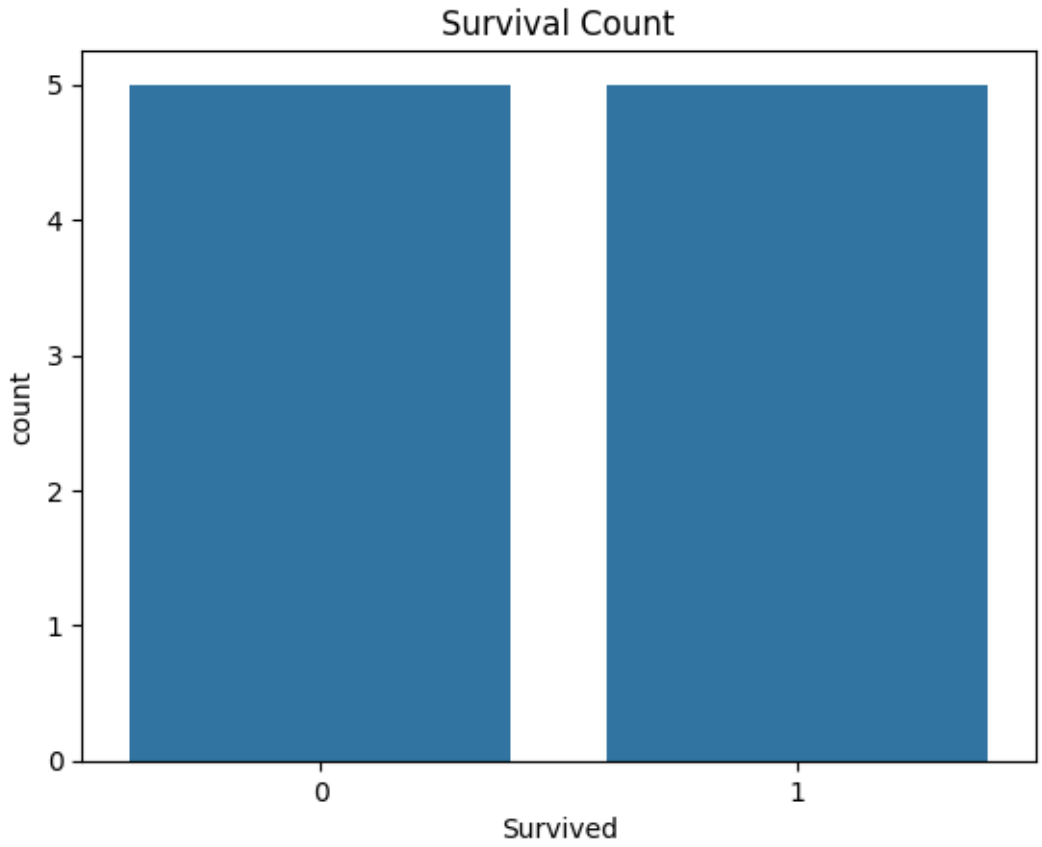
D. Survival by Embarkation Port

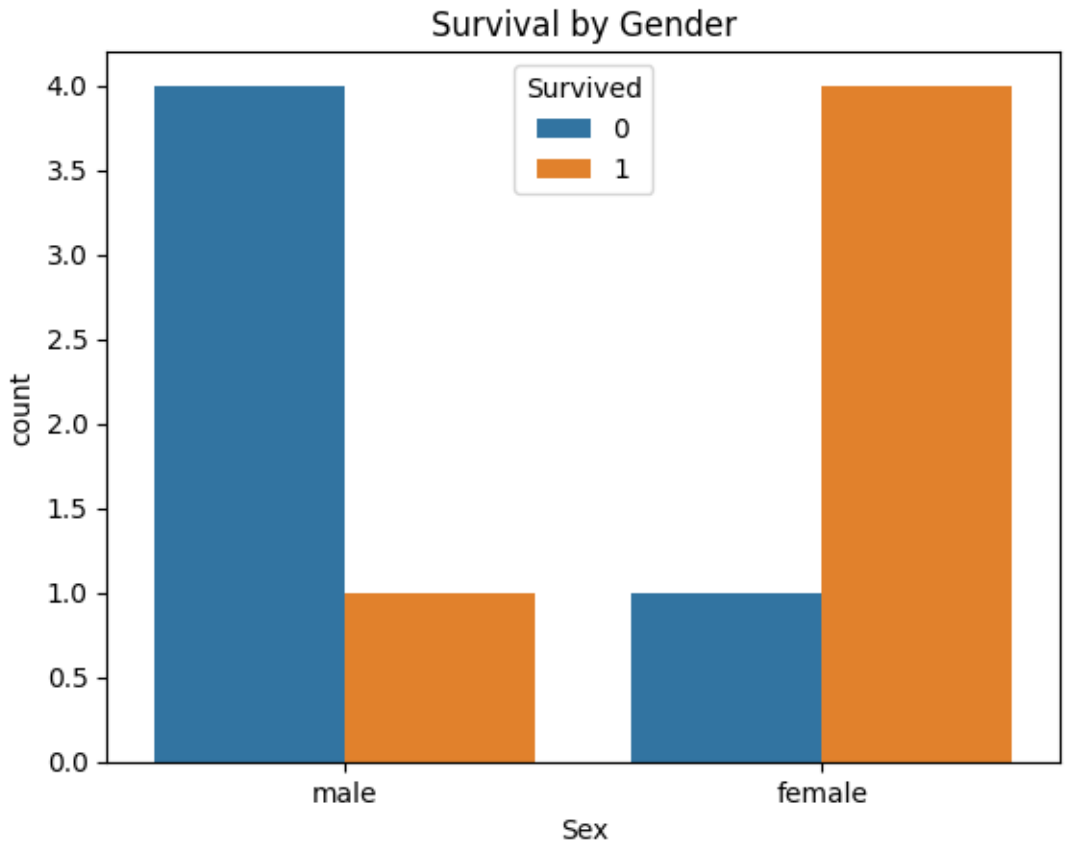
```
pd.crosstab(df['Embarked'], df['Survived'], normalize='index')
```

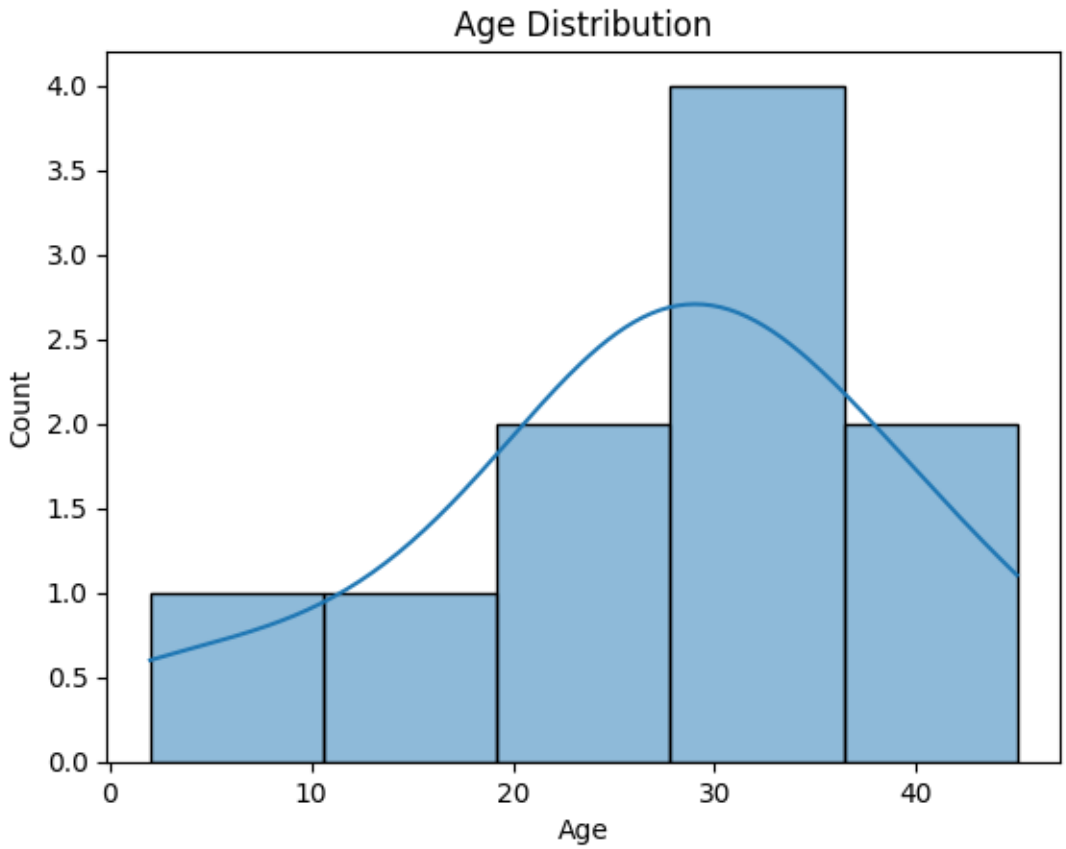
Visualizations

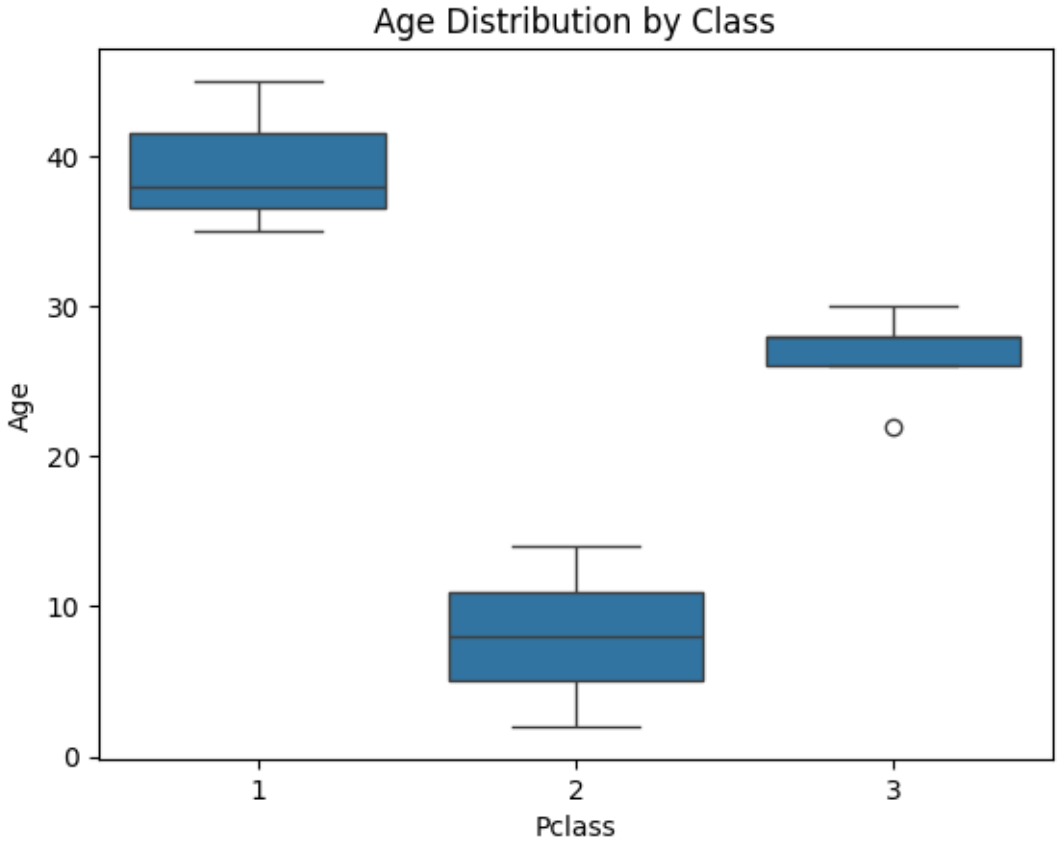
```
import seaborn as sns
import matplotlib.pyplot as plt

# Survival count
sns.countplot(x='Survived', data=df)
plt.title("Survival Count")
plt.show()

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

# Age distribution
sns.histplot(df['Age'], kde=True)
plt.title("Age Distribution")
plt.show()

# Boxplot: Age by Class
sns.boxplot(x='Pclass', y='Age', data=df)
plt.title("Age Distribution by Class")
```

plt.show()


Survival Count
count
Survived
0
1
0
1
2
3
4
5


Survival by Gender
Survived
0
1
count
Sex
male
female
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Age Distribution
Count
Age
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
10
20
30
40


Age Distribution by Class
Age
Pclass
0
10
20
30
40
1
2
3