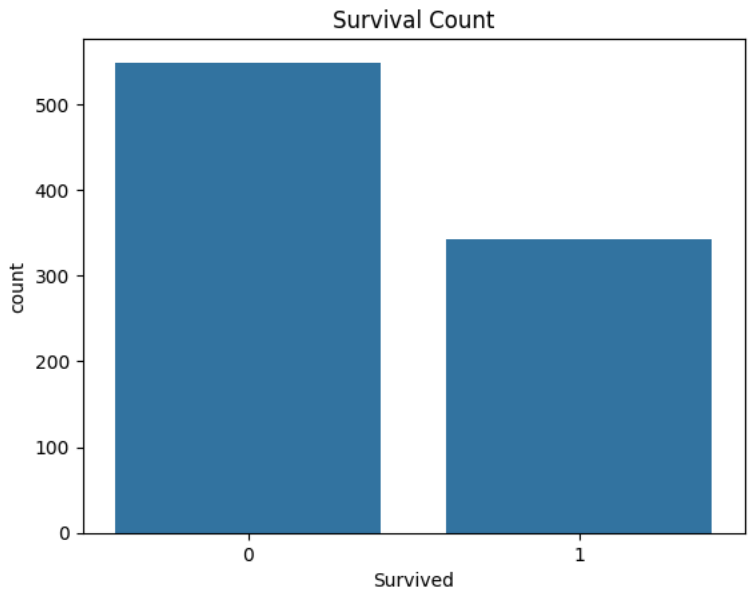
```
# Titanic EDA - Internship Task 5

# Import Libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Load Dataset
df = pd.read_csv("train.csv")
```

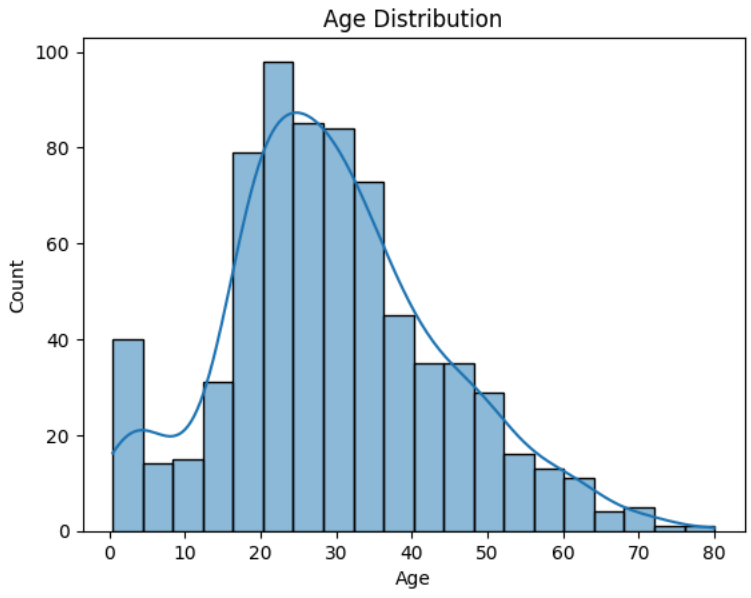
```
# Initial Data Exploration
print("Shape of dataset:", df.shape)
print("\nInfo:")
df.info()
print("\nSummary statistics:")
print(df.describe())
print("\nMissing values:")
print(df.isnull().sum())
print("\nUnique values per column:")
print(df.nunique())
```

```
 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

Summary statistics:
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200

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
dtype: int64

Unique values per column:
PassengerId    891
Survived         2
Pclass           3
Name           891
Sex              2
Age             88
SibSp            7
Parch            7
Ticket         681
Fare           248
Cabin          147
```

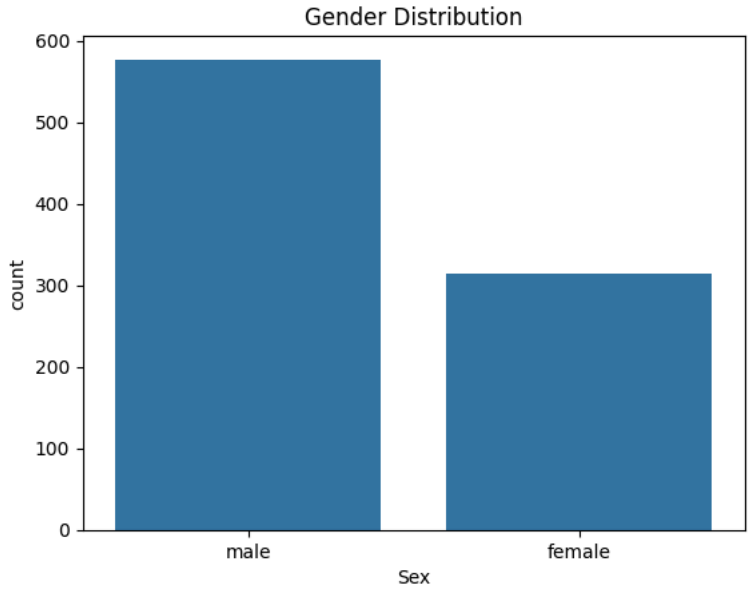
```
# Univariate Analysis
sns.countplot(x='Survived', data=df)
plt.title('Survival Count')
plt.show()
```



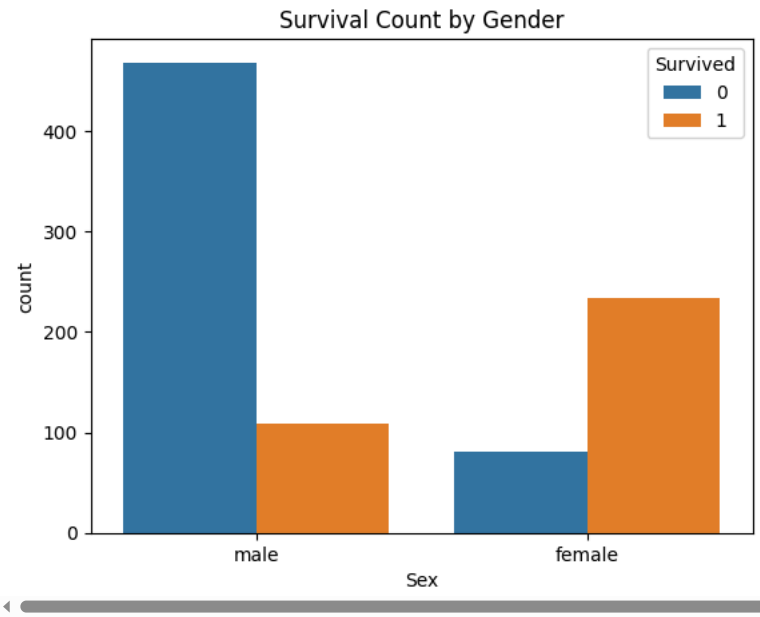

```
# Univariate Analysis
sns.histplot(df['Age'].dropna(), kde=True)
plt.title('Age Distribution')
plt.show()
```



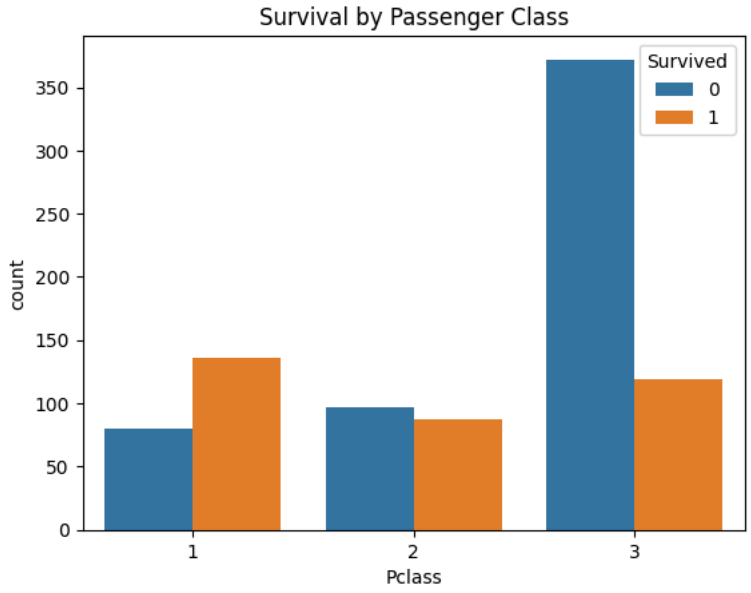

```
# Univariate Analysis
sns.countplot(x='Sex', data=df)
plt.title('Gender Distribution')
plt.show()
```



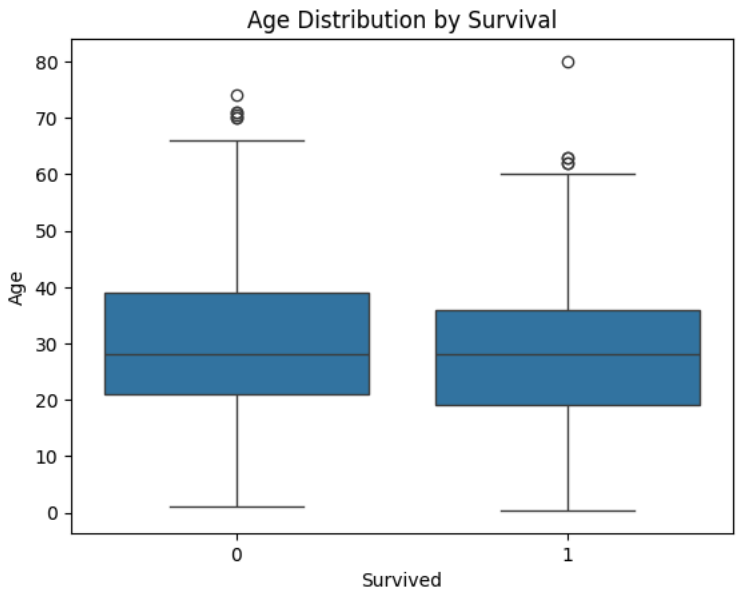

```
# Bivariate Analysis
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title('Survival Count by Gender')
plt.show()
```



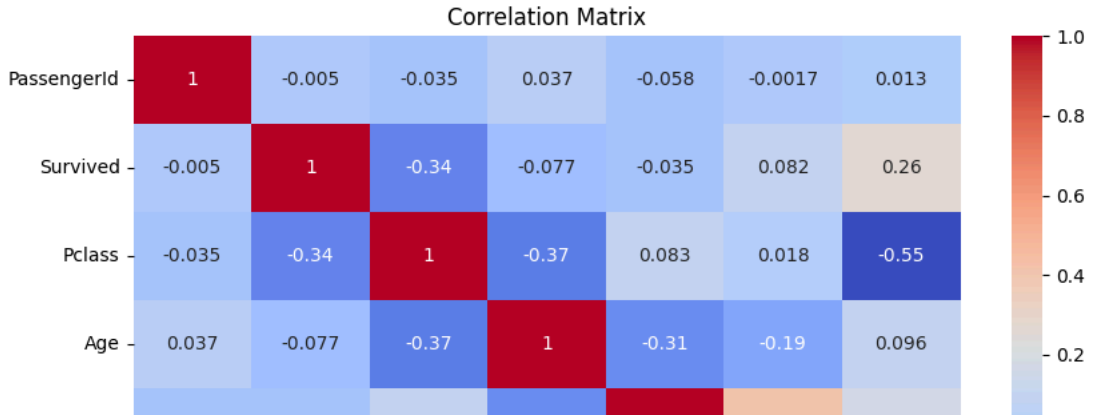

```
# Bivariate Analysis
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title('Survival by Passenger Class')
plt.show()
```




```
# Bivariate Analysis
sns.boxplot(x='Survived', y='Age', data=df)
plt.title('Age Distribution by Survival')
plt.show()
```




```
# Select only numeric columns for correlation
numeric_df = df.select_dtypes(include=['int64', 'float64'])

# Plot correlation heatmap
plt.figure(figsize=(10,6))
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```


Correlation Matrix
PassengerId
Survived
Pclass
Age
1
-0.005
-0.035
0.037
-0.058
-0.0017
0.013
-0.005
1
-0.34
-0.077
-0.035
0.082
0.26
-0.035
-0.34
1
-0.37
0.083
0.018
-0.55
0.037
-0.077
-0.37
1
-0.31
-0.19
0.096
1.0
0.8
0.6
0.4
0.2