In [12]:

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

##Load the dataset:
df = pd.read_excel(r"C:\Users\prachi athalye\Desktop\Titanic.xlsx")
##Explore the dataset:
# Display the first few rows of the dataset
print(df.head())
# Check the shape of the dataset
print(df.shape)
# Check the data types of each column
print(df.dtypes)
# Check for missing values
print(df.isnull().sum())
# Check basic statistics of numerical columns
print(df.describe())
```

```
   PassengerId  Survived  Pclass  \
0          892         0       3   
1          893         1       3   
2          894         0       2   
3          895         0       3   
4          896         1       3   

                                           Name     Sex   Age  SibSp  Par
ch  \
0                              Kelly, Mr. James    male  34.5      0      
0   
1              Wilkes, Mrs. James (Ellen Needs)  female  47.0      1      
0   
2                     Myles, Mr. Thomas Francis    male  62.0      0      
0   
3                              Wirz, Mr. Albert    male  27.0      0      
0   
4  Hirvonen, Mrs. Alexander (Helga E Lindqvist)  female  22.0      1      
1   

    Ticket     Fare Cabin Embarked  
0   330911   7.8292   NaN        Q  
1   363272   7.0000   NaN        S  
2   240276   9.6875   NaN        Q  
3   315154   8.6625   NaN        S  
4  3101298  12.2875   NaN        S  
(418, 12)
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
       PassengerId    Survived      Pclass         Age       SibSp  \
count   418.000000  418.000000  418.000000  332.000000  418.000000   
mean   1100.500000    0.363636    2.265550   30.272590    0.447368   
std     120.810458    0.481622    0.841838   14.181209    0.896760   
min     892.000000    0.000000    1.000000    0.170000    0.000000   
25%     996.250000    0.000000    1.000000   21.000000    0.000000   
50%    1100.500000    0.000000    3.000000   27.000000    0.000000   
75%    1204.750000    1.000000    3.000000   39.000000    1.000000   
```

```
max    1309.000000    1.000000    3.000000   76.000000    8.000000

            Parch        Fare
count  418.000000  417.000000
mean     0.392344   35.627188
std      0.981429   55.907576
min      0.000000    0.000000
25%      0.000000    7.895800
50%      0.000000   14.454200
75%      0.000000   31.500000
max      9.000000  512.329200
```

In [10]:

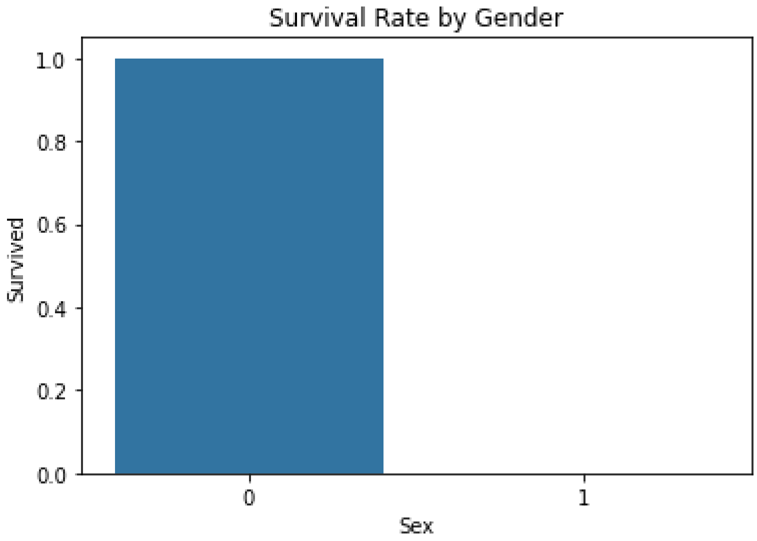
```
##Data Cleaning:
# Drop unnecessary columns
df = df.drop(['PassengerId', 'Name', 'Ticket', 'Cabin'], axis=1)
# Fill missing values
df['Age'] = df['Age'].fillna(df['Age'].median())
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
# Convert categorical variables to numeric
df['Sex'] = df['Sex'].map({'female': 0, 'male': 1})
df['Embarked'] = df['Embarked'].map({'S': 0, 'C': 1, 'Q': 2})
# Check if missing values have been filled
print(df.isnull().sum())
```

```
Survived    0
Pclass      0
Sex         0
Age         0
SibSp       0
Parch       0
Fare        1
Embarked    0
dtype: int64
```

In [11]:

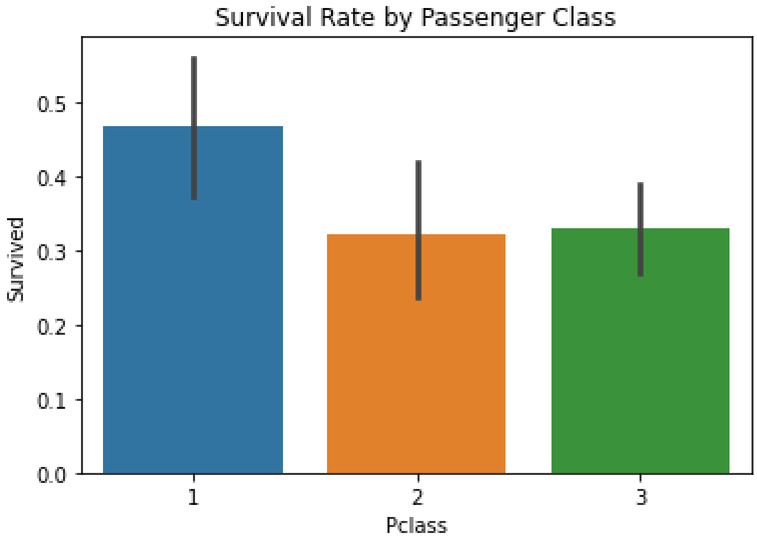
```
##Exploratory Data Analysis:
# Calculate the survival rate
survival_rate = df['Survived'].mean()
print("Survival Rate:", survival_rate)
# Visualize the survival rate by gender
sns.barplot(x='Sex', y='Survived', data=df)
plt.title("Survival Rate by Gender")
plt.show()
```

```
Survival Rate: 0.36363636363636365
```

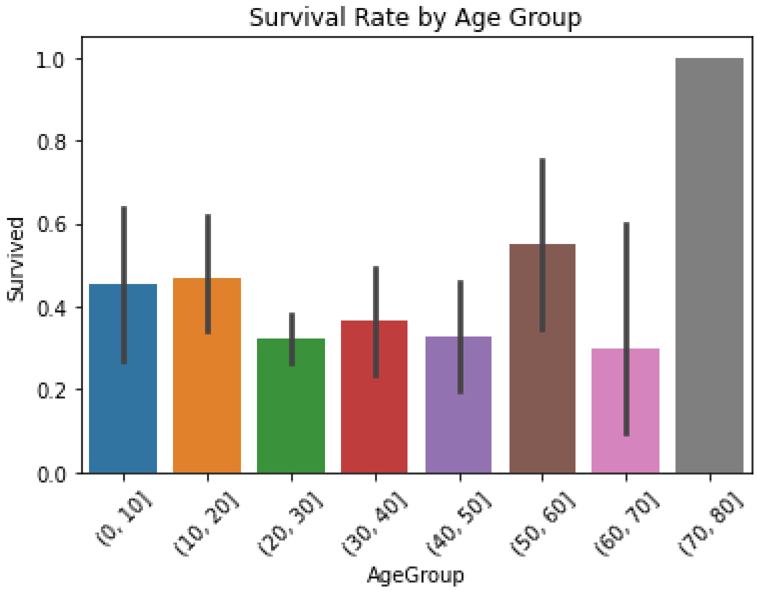


In [6]:

```
# Visualize the survival rate by passenger class
sns.barplot(x='Pclass', y='Survived', data=df)
plt.title("Survival Rate by Passenger Class")
plt.show()
```



In [7]:
```
# Visualize the survival rate by age group
df['AgeGroup'] = pd.cut(df['Age'], bins=[0, 10, 20, 30, 40, 50, 60, 70, 80]
sns.barplot(x='AgeGroup', y='Survived', data=df)
plt.title("Survival Rate by Age Group")
plt.xticks(rotation=45)
plt.show()
```


Survival Rate by Age Group
Survived
(0, 10]
(10, 20]
(20, 30]
(30, 40]
(40, 50]
(50, 60]
(60, 70]
(70, 80]
AgeGroup