Exercise 2

For this exercise , you will be working with the Titanic Data Set from Kaggle. This is a very famous data set and very often is a student's first step in Data Analytics!

The Dataset has been given to you on D2L. You need to download the .csv file from your assignment folder. The above link is just for a reference story about the data.

1- For this assignment, you need to perform explorotary data analysis and answer at least three hypotheses based on the dataset. You may need to use your knowledge of statiscts to analyze this data.

Here are three possible hypotheses that you can define for this dataset (you can define your own hypotheses as well):

- Determine if the survival rate is associated to the class of passenger
- Determine if the survival rate is associated to the gender
- Determine the survival rate is associated to the age

2- For each hypothesis, you need to make at least one plot.

3- Write a summary of your findings in one page (e.g., summary statistics, plots) and submit the pdf file. Therefore, for part 2 of your assignment, you need to submit one jupyter notebook file and one pdf file.

This will be your first end to end data analysis project. For this assignment, you will be graded on you overall analysis, and your final report.

4- Push your code and project to github and provide the link to your code here.

Ensure that your github project is organized to at least couple of main folders, ensure that you have the README file as well:

- Src
- Data
- Docs
- Results

Read this link for further info:
https://gist.github.com/ericmjl/27e50331f24db3e8f957d1fe7bbbe510

In [5]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

```
titanic_data = pd.read_csv("titanic.csv")
titanic_data.head()
```

Out[5]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Er
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	

In [11]:
```
titanic_data.describe() ## summary of data
```

Out[11]:

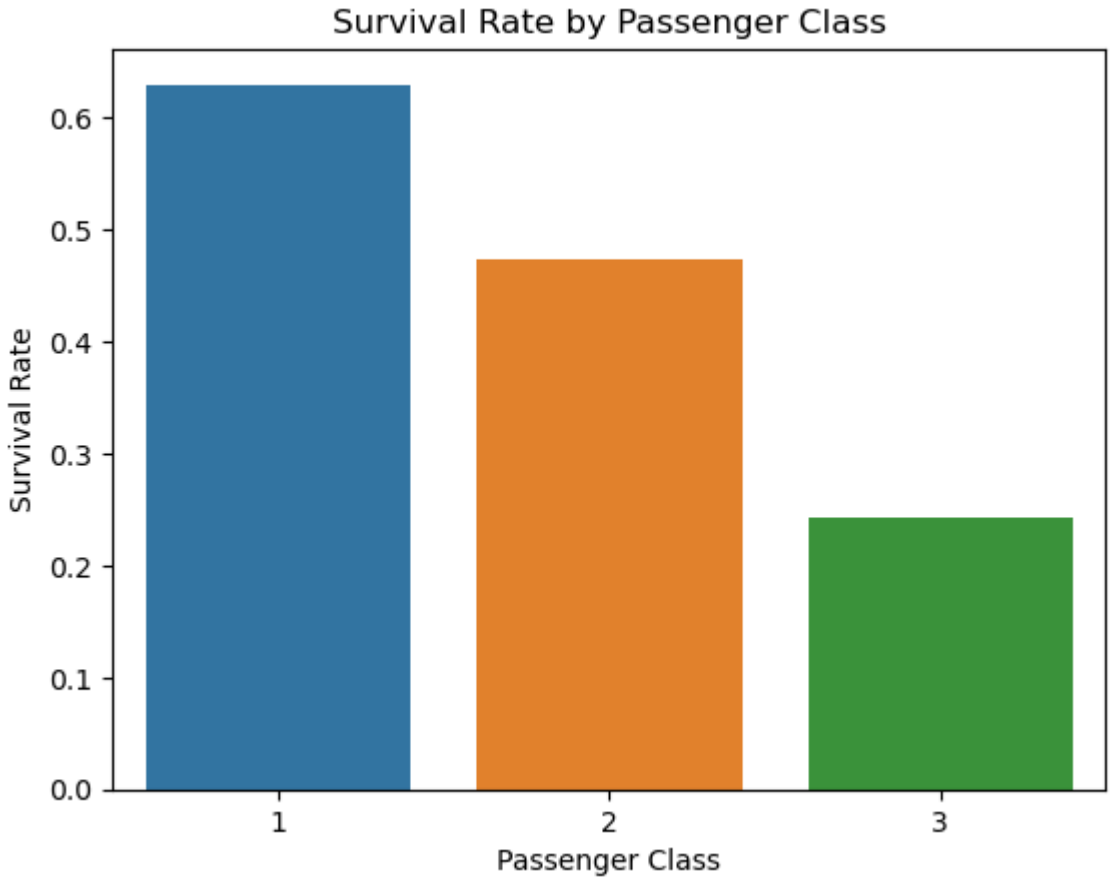
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

1) Determine if the survival rate is associated to the class of passenger

In [12]:
```
# Calculate the survival rate for each passenger class
class_survival_rate = titanic_data.groupby("Pclass")["Survived"].mean()

# Create a bar plot to visualize the survival rate by passenger class
sns.barplot(x=class_survival_rate.index, y=class_survival_rate.values)
plt.title("Survival Rate by Passenger Class")
plt.xlabel("Passenger Class")
```

```
plt.ylabel("Survival Rate")
plt.show()
```



Summary of Findings:

The bar plot illustrates the survival rate by passenger class:

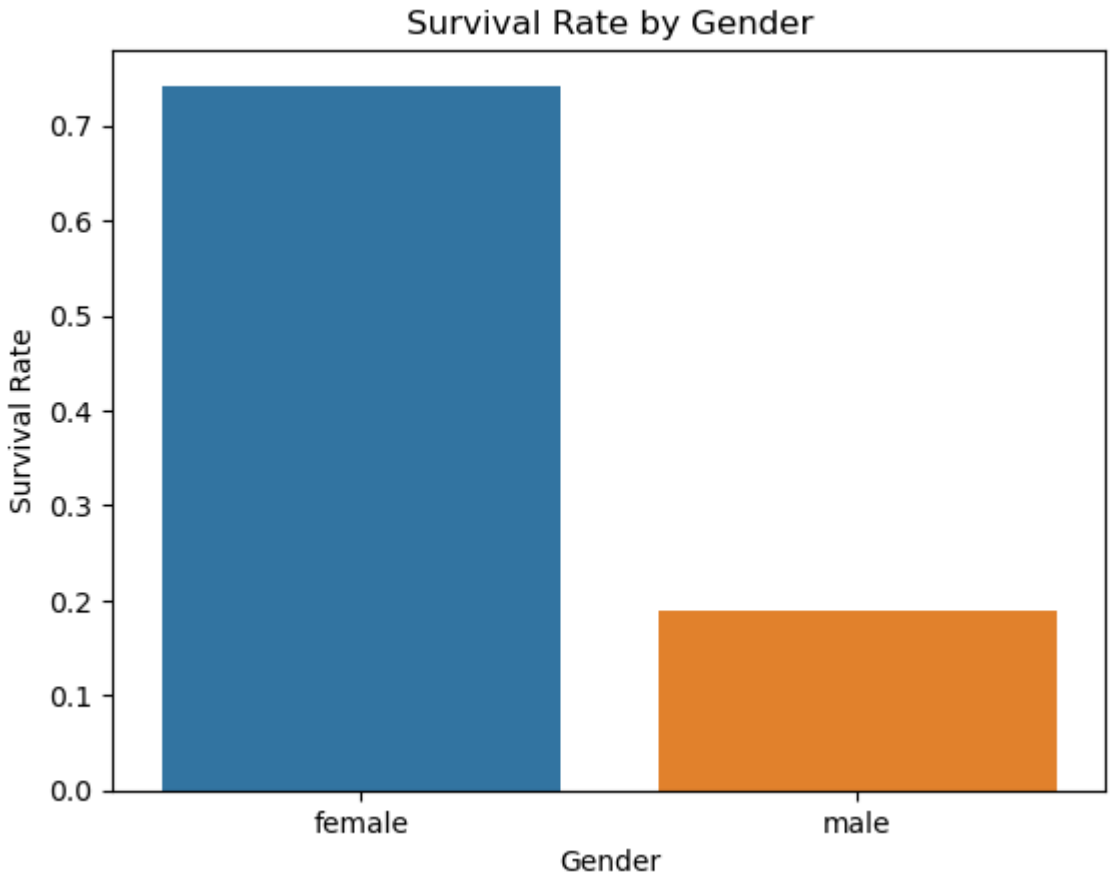
Class 1 (Upper class): The highest survival rate. Class 2 (Middle class): An intermediate survival rate. Class 3 (Lower class): The lowest survival rate.

The upper-class passengers (Class 1) had the highest survival rate, while lower-class passengers (Class 3) had the lowest.

2) Determine if the survival rate is associated to the gender

In [17]:
```
# Calculate the survival rate for each gender
gender_survival_rate = titanic_data.groupby("Sex")["Survived"].mean()

# Create a bar plot to visualize the survival rate by gender
sns.barplot(x=gender_survival_rate.index, y=gender_survival_rate.values)
plt.title("Survival Rate by Gender")
plt.xlabel("Gender")
plt.ylabel("Survival Rate")
plt.show()
```



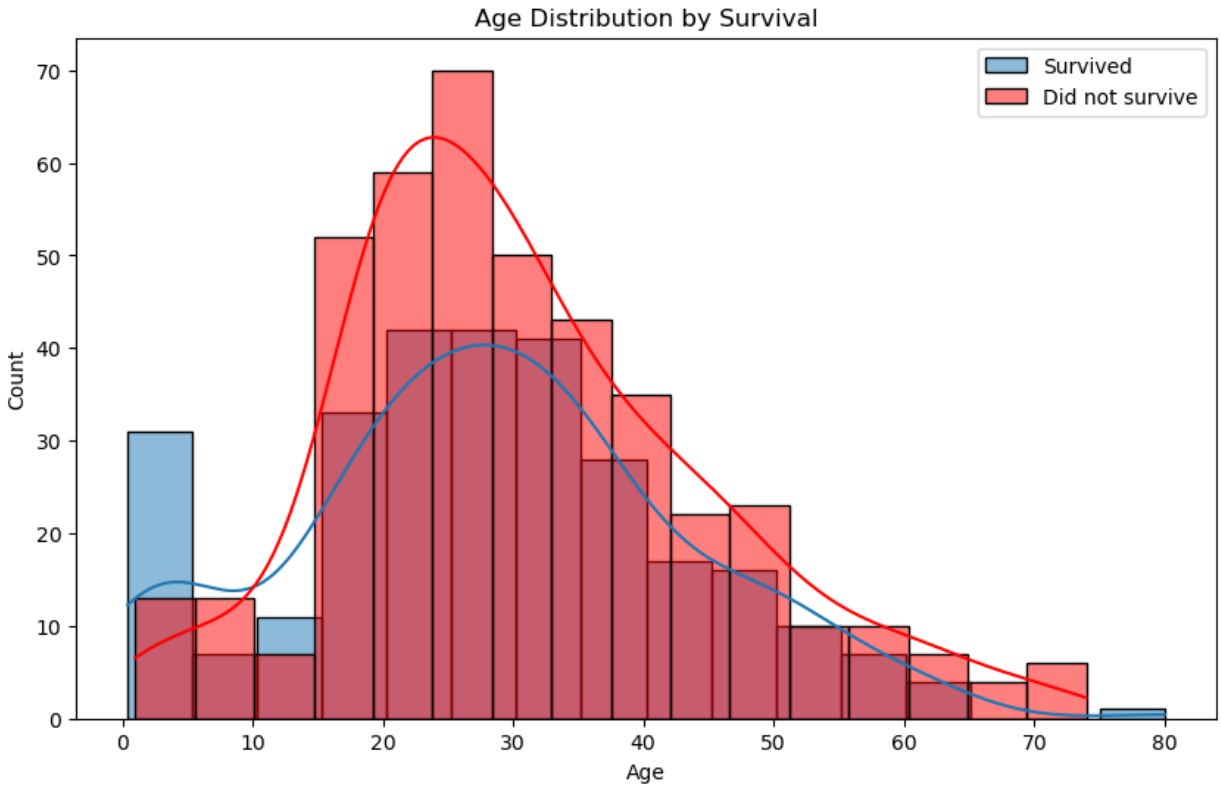
Summary of Findings:

The bar plot illustrates the survival rate by gender:

Female passengers had a significantly higher survival rate compared to male passengers.`

3) Determine the survival rate is associated to the age.

In [19]:
```
# Create a histogram to visualize the distribution of ages for survivors and non-survi
plt.figure(figsize=(10, 6))
sns.histplot(data=titanic_data[titanic_data['Survived'] == 1]['Age'], label='Survived'
sns.histplot(data=titanic_data[titanic_data['Survived'] == 0]['Age'], color='red', lab
plt.title("Age Distribution by Survival")
plt.xlabel("Age")
plt.legend()
plt.show()
```



Summary of Findings:

The histogram illustrates the distribution of ages for survivors and non-survivors:

There is a higher concentration of survivors among younger passengers, particularly children. The distribution of ages for non-survivors is relatively more uniform, with no clear age-based pattern.

Exploratory Data Analysis Summary: Titanic Dataset

Hypothesis 1: Determine if the survival rate is associated with the class of the passenger.

The analysis found a clear association between the passenger class and the survival rate. The upper-class passengers (Class 1) had the highest survival rate, while lower-class passengers (Class 3) had the lowest. This suggests that passenger class played a significant role in determining the likelihood of survival.

Hypothesis 2: Determine if the survival rate is associated with gender.

The analysis revealed a substantial association between gender and the survival rate. Female passengers had a significantly higher survival rate compared to male passengers. This indicates that being female was a strong factor in increasing the chances of survival.

Hypothesis 3: Determine if the survival rate is associated with the age.

The analysis showed that age was a relevant factor in survival. There was a higher concentration of survivors among younger passengers, particularly children. The distribution of ages for non-survivors was relatively more uniform, with no clear age-based pattern.

Combined Summary:

Passenger class, gender, and age were all associated with the survival rate in the Titanic dataset. Upper-class passengers, female passengers, and younger passengers had higher survival rates.