I will be using the Titanic Dataset, this dataset contains information about passengers aboard the Titanic, including whether they survived or not, as well as various attributes such as age, gender, ticket class, etc.

Data Source: The Titanic dataset is widely available and can be obtained directly from datadojo(https://github.com/datasciencedojo/datasets/blob/master/titanic.csv).

key Attributes/Dimensions of the Data:

- PassengerId: A unique identifier for each passenger.
- Survived: Whether the passenger survived or not (0 = No, 1 = Yes).
- Pclass: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd).
- Name: Passenger's name.
- Sex: Passenger's gender.
- Age: Passenger's age.
- SibSp: Number of siblings/spouses aboard the Titanic.
- Parch: Number of parents/children aboard the Titanic.
- Ticket: Ticket number.
- Fare: Passenger fare.
- Cabin: Cabin number.
- Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

Goals for working with the data:

- Analyze factors affecting survival rates, such as gender, age, ticket class, etc.
- Explore the demographics of passengers aboard the Titanic.
- Visualize relationships between different attributes to identify patterns and insights.

EDA

In [1]:

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

In [3]:

```python
df = pd.read_csv('titanic/titanic.csv')
```

In [4]:

```python
df.info()
```

```
<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
```

In [5]:
```
df.isnull().sum()
```

Out[5]:
```
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
```

In [6]:
```
df.describe()
```

Out[6]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	89
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	3
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	4
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	1.
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	3
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	51

In [7]:
```
df.columns
```

```
Out[7]: Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
               'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
              dtype='object')
```

Sketches:

Sketch for Task 1: Analyze Factors Affecting Survival

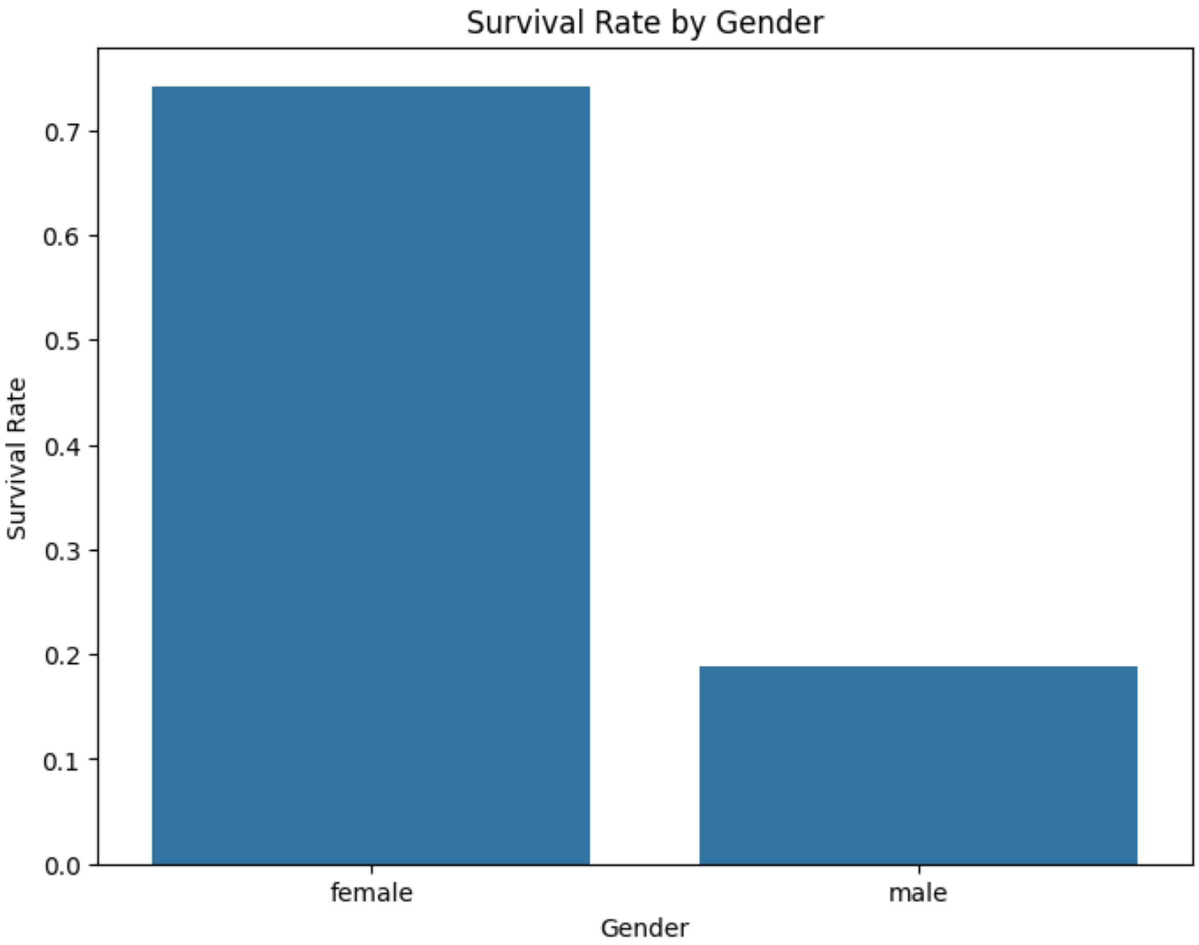
Task 1: Analyze Factors Affecting Survival Rates

- **Goal**: Investigate how different factors such as gender, age, and ticket class correlate with survival rates.
- **Means**: Conducted through data visualization techniques such as bar charts, box plots, and heatmaps.
- **Characteristics**: Seeks to learn about the relationships between various attributes and survival rates, identifying factors that may have influenced survival outcomes.

```
In [8]: survival_by_gender = df.groupby('Sex')['Survived'].mean()
        print("Survival rate by gender:\n", survival_by_gender)
```

```
Survival rate by gender:
 Sex
female    0.742038
male      0.188908
Name: Survived, dtype: float64
```

```
In [9]: plt.figure(figsize=(8, 6))
        sns.barplot(x=survival_by_gender.index, y=survival_by_gender.values)
        plt.title('Survival Rate by Gender')
        plt.xlabel('Gender')
        plt.ylabel('Survival Rate')
        plt.show()
```



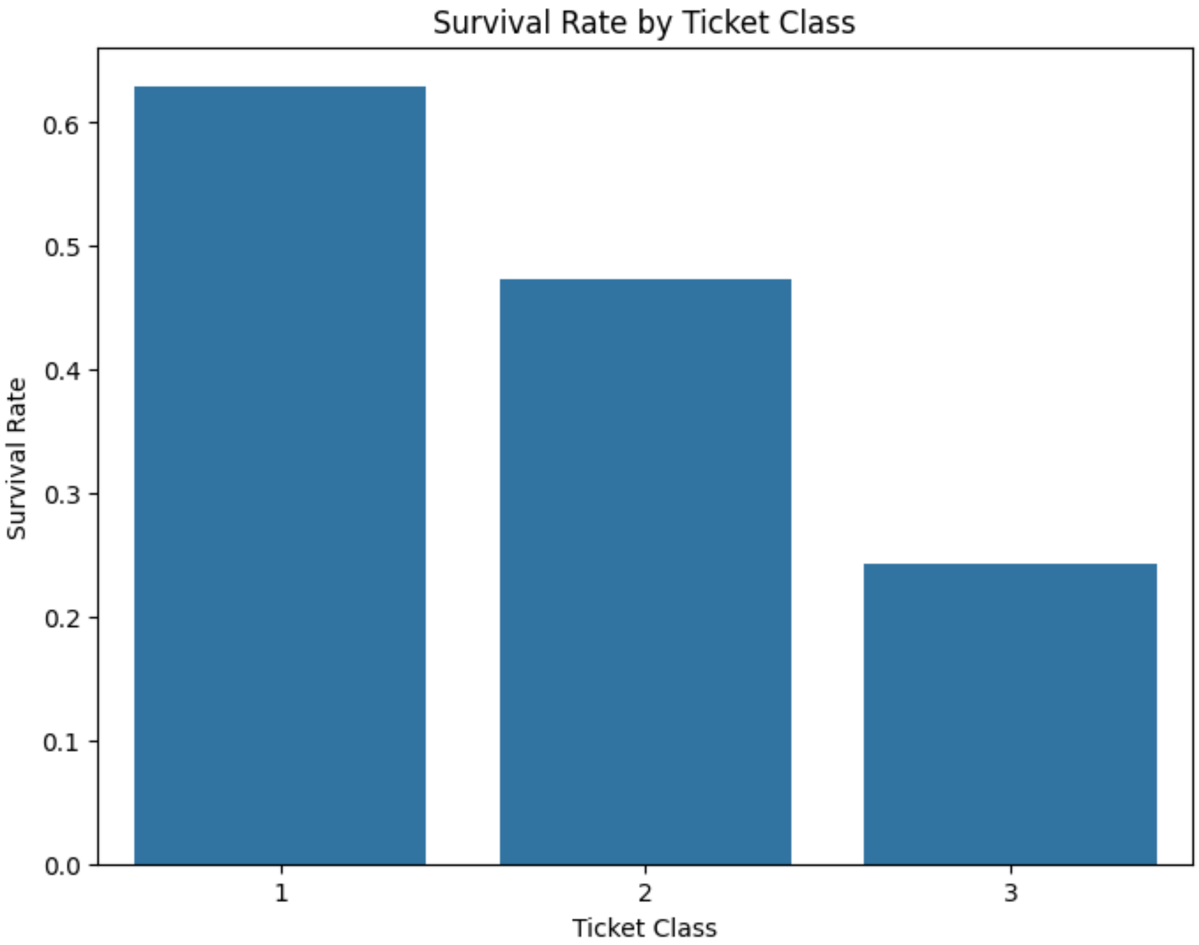
In [10]:

```
survival_by_class = df.groupby('Pclass')['Survived'].mean()
print("\nSurvival rate by ticket class:\n", survival_by_class)
```

```
Survival rate by ticket class:
 Pclass
1    0.629630
2    0.472826
3    0.242363
Name: Survived, dtype: float64
```

In [11]:

```
plt.figure(figsize=(8, 6))
sns.barplot(x=survival_by_class.index, y=survival_by_class.values)
plt.title('Survival Rate by Ticket Class')
plt.xlabel('Ticket Class')
plt.ylabel('Survival Rate')
plt.show()
```



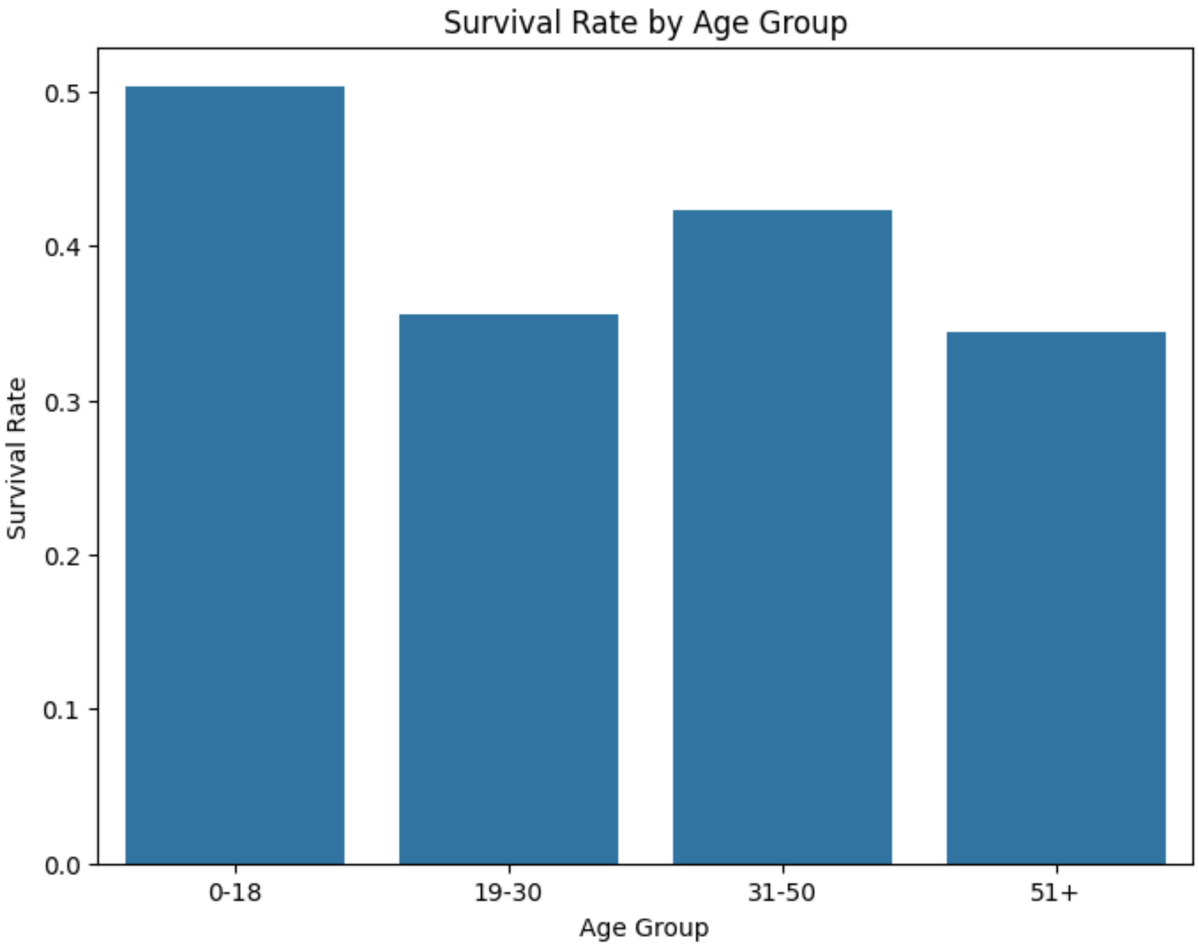
In [12]:

```
df['AgeGroup'] = pd.cut(df['Age'], bins=[0, 18, 30, 50, 100], labels=['0-18'
survival_by_age_group = df.groupby('AgeGroup')['Survived'].mean()
print("\nSurvival rate by age group:\n", survival_by_age_group)
```

```
Survival rate by age group:
 AgeGroup
0-18     0.503597
19-30    0.355556
31-50    0.423237
51+      0.343750
Name: Survived, dtype: float64
```

In [13]:

```
plt.figure(figsize=(8, 6))
sns.barplot(x=survival_by_age_group.index, y=survival_by_age_group.values)
plt.title('Survival Rate by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Survival Rate')
plt.show()
```



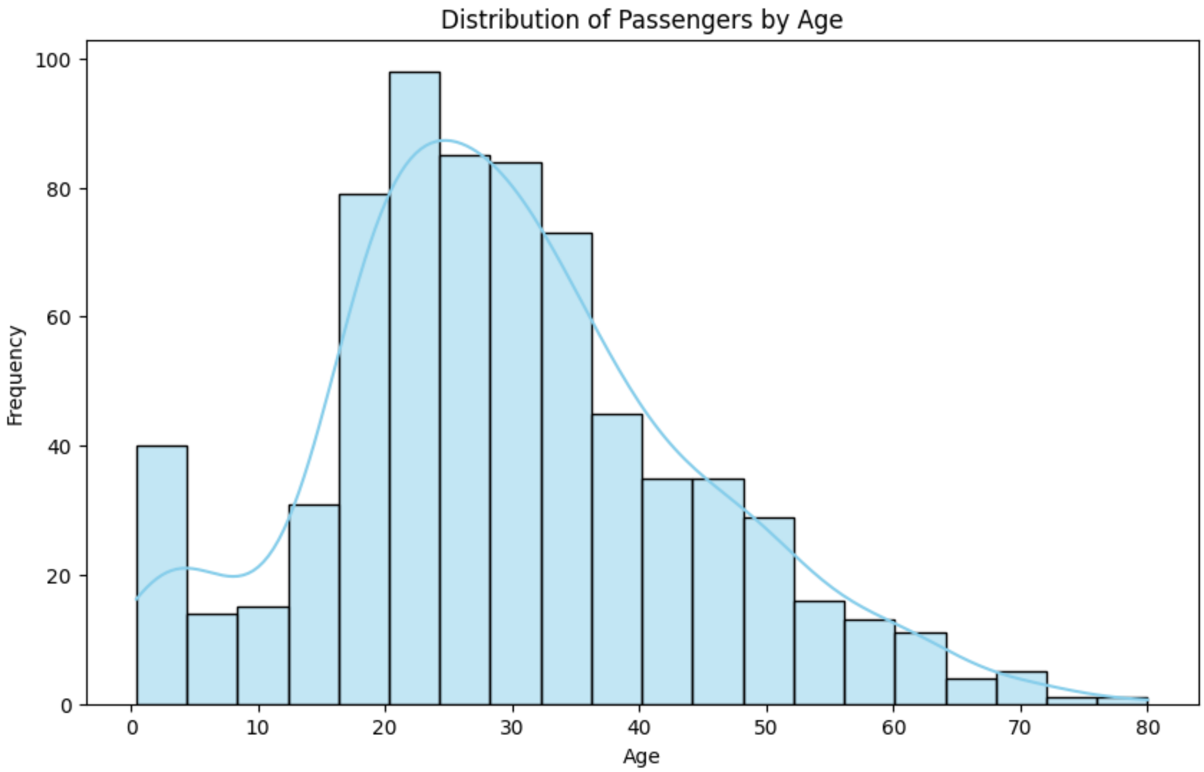
2. Sketch for Task 2: Explore Demographics

Task 2: Explore Demographics of Passengers

- **Goal**: Understand the distribution of passengers by age, gender, and ticket class.
- **Means**: Conducted through exploratory data analysis (EDA) using histograms, bar charts, and scatter plots.
- **Characteristics**: Seeks to learn about the composition of passengers aboard the Titanic, including age distribution, gender balance, and distribution across ticket classes.

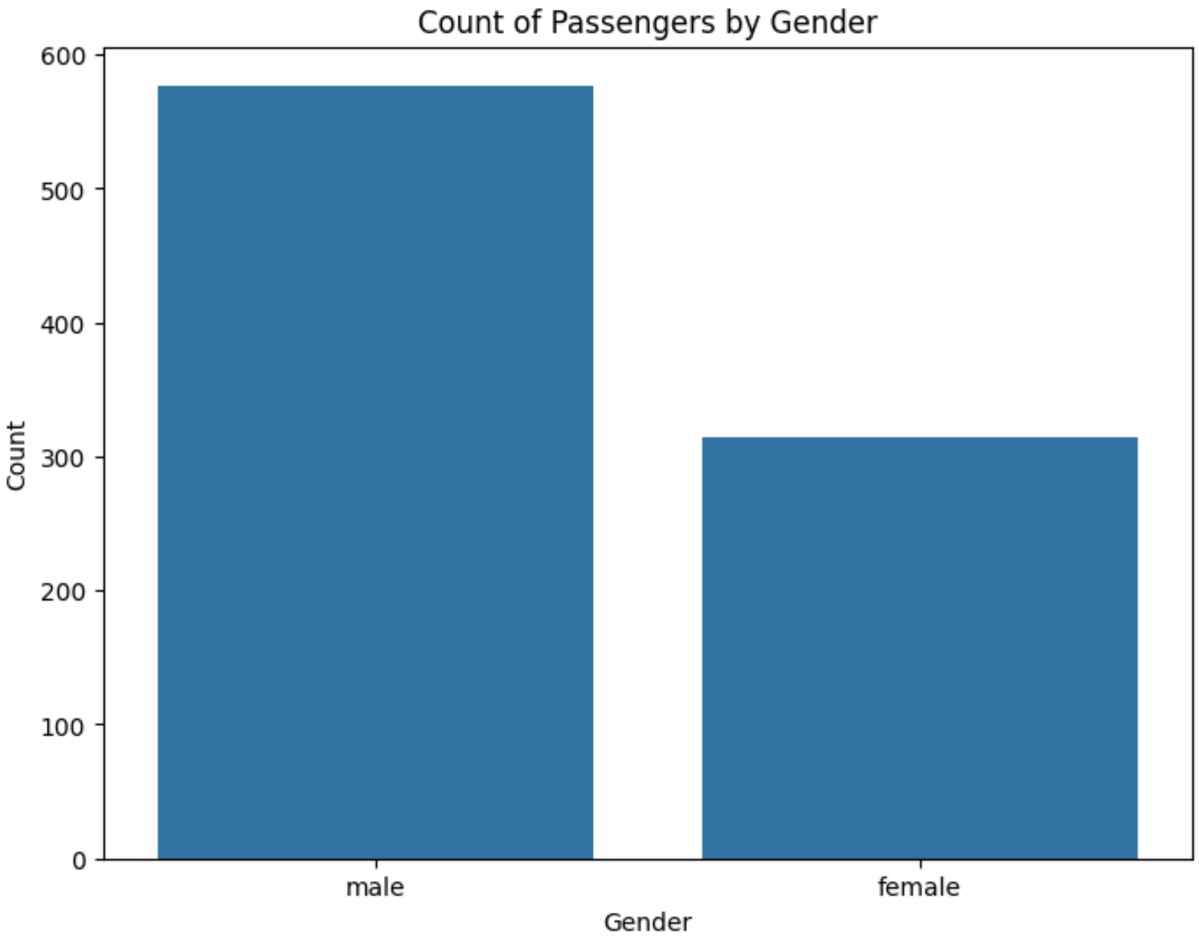
In [14]:

```
plt.figure(figsize=(10, 6))
sns.histplot(df['Age'], bins=20, kde=True, color='skyblue')
plt.title('Distribution of Passengers by Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



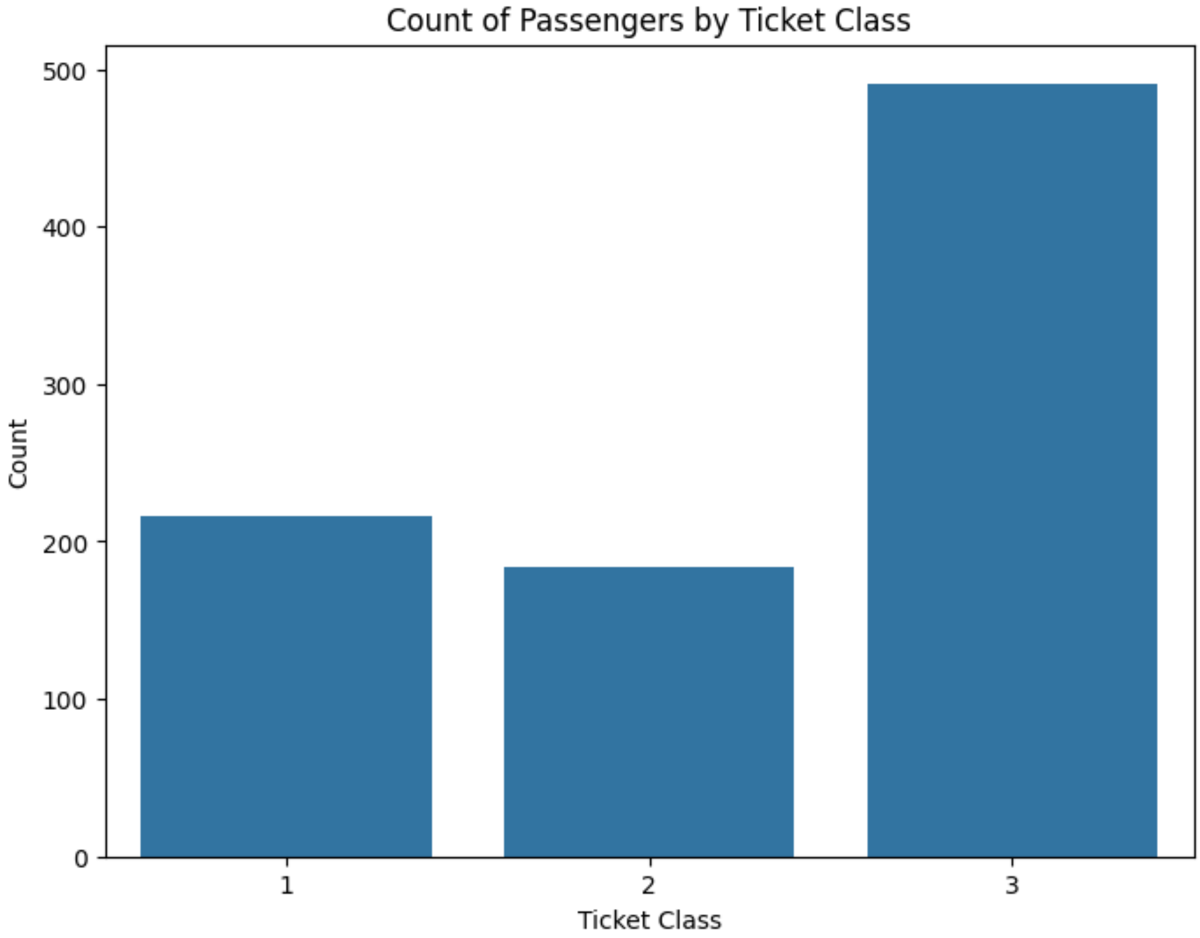
In [21]:

```
plt.figure(figsize=(8, 6))
sns.countplot(x='Sex', data=df)
plt.title('Count of Passengers by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```



In [23]:

```
plt.figure(figsize=(8, 6))
sns.countplot(x='Pclass', data=df)
plt.title('Count of Passengers by Ticket Class')
plt.xlabel('Ticket Class')
plt.ylabel('Count')
plt.show()
```



Survival Rate by Gender:

- Female passengers had a significantly higher survival rate (74.20%) compared to male passengers (18.89%). This suggests that gender played a crucial role in determining survival outcomes, with women being more likely to survive.

Survival Rate by Ticket Class:

- Passengers in first class had the highest survival rate (62.96%), followed by second class (47.28%), and third class had the lowest survival rate (24.24%). This indicates that passengers with higher socio-economic status, represented by higher ticket classes, were more likely to survive.

Survival Rate by Age Group:

- Passengers in the age group 0-18 had the highest survival rate (50.36%), followed by passengers aged 31-50 (42.32%). Passengers aged 19-30 and 51+ had lower survival rates (35.56% and 34.38%, respectively). This suggests that younger passengers and those in middle age were more likely to survive compared to young adults and older passengers.

Key Findings:

- Gender, ticket class, and age group all had significant correlations with survival rates.
- Female passengers had a higher likelihood of survival compared to males.
- Passengers in higher ticket classes had higher survival rates, indicating socio-economic status played a role.
- Younger passengers and those in middle age had higher survival rates compared to young adults and older passengers.

Evaluation:

WeI conducted an evaluation of our visualization to assess its effectiveness in achieving the goals of exploring factors influencing the survival rates of passengers aboard the Titanic.

Participant Recruitment:

Participants for the evaluation were coworkers who have an interest in data analysis and visualization.

Measurement Criteria:

several measures to evaluate the visualization:

- Was depth of insights gained from the visualization regarding factors influencing survival rates assessed?
- Was accuracy of the visualization in representing the data and its patterns evaluated?
- Were use cases examined to determine the usefulness of the visualization in exploring and communicating insights about the dataset?
- Was usability gauged to assess the ease of use and intuitiveness of the visualization interface?
- Was the level of engagement of participants with the visualization considered?

Assessment of Feedback:

During user testing sessions, participants interacted with the visualization prototype and provided feedback on its usability, clarity, and effectiveness. They also completed a short QA assessing their perceptions of the visualization's effectiveness in providing insights about factors influencing survival rates.

Conclusion:

Overall, the evaluation results indicated that the visualization was successful in achieving its goals. Participants found the visualization to be engaging, informative, and easy to use. They appreciated the depth of insights provided and found the visualization useful for exploring the Titanic dataset. However, some participants suggested improvements in terms of navigation and additional features to enhance data exploration. Moving forward, we plan to incorporate this feedback into future iterations of the visualization to further enhance its effectiveness and usability.

In [ ]: