Importing necessary libraries

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:
```
df=pd.read_csv(r"C:\Users\Chandrika\Downloads\titanic\train.csv")  #loading dataset
```

In [4]:
```
df.info() #total information of dataset
```

```
<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.head() #first 5 rows of dataset
```

Out[5]:

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

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
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 [7]:
```
df.describe()
```

Out[7]:

	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

Data Cleaning

In [8]:
```
# Handling missing values
df['Age'].fillna(df['Age'].median(), inplace=True)
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
df.drop(columns=['Cabin'], inplace=True)
```

In [9]:
```
# Verifying no missing values
df.isnull().sum()
```

Out[9]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

Research questions:

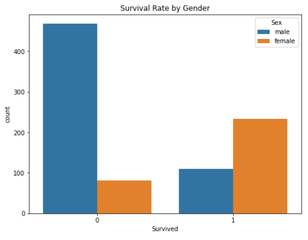
1)What factors influenced the survival rate of passengers on the Titanic?

2)How did age and gender affect survival chances?

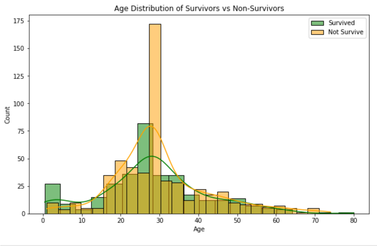
3)What was the survival rate across different passenger classes?

Visualization and Analysis

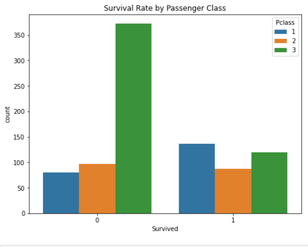
In [10]:
```
# 1. Survival rate based on gender
plt.figure(figsize=(8, 6))
sns.countplot(x='Survived', hue='Sex', data=df)
plt.title('Survival Rate by Gender')
plt.show()
```



In [15]:
```
# 2. Age distribution of survivors vs non-survivors
plt.figure(figsize=(10, 6))
sns.histplot(df[df['Survived'] == 1]['Age'], kde=True, color='green', label='Survived')
sns.histplot(df[df['Survived'] == 0]['Age'], kde=True, color='orange', label='Not Survive')
plt.title('Age Distribution of Survivors vs Non-Survivors')
plt.legend()
plt.show()
```



In [16]:
```
# 3. Survival rate based on passenger class
plt.figure(figsize=(8, 6))
sns.countplot(x='Survived', hue='Pclass', data=df)
plt.title('Survival Rate by Passenger Class')
plt.show()
```



In [17]:
```
# 4. Survival rate based on passenger class and gender
plt.figure(figsize=(10, 6))
sns.catplot(x='Pclass', hue='Sex', col='Survived', data=df, kind='count', height=4, aspect=1.5)
plt.suptitle('Survival Rate by Passenger Class and Gender')
plt.show()
```

```
<Figure size 720x432 with 0 Axes>
```

conclusion

Based on the analysis of the Titanic dataset, we can draw the following conclusions:

1. **Gender Influence:** Females had a significantly higher survival rate compared to males.
2. **Age Influence:** There is a noticeable difference in the age distribution between survivors and non-survivors, with younger passengers having slightly higher survival rates.
3. **Passenger Class Influence:** Passengers in higher classes (First and Second) had a higher survival rate compared to those in Third class.
4. **Combined Factors:** The survival rate varies significantly when considering the combination of passenger class and gender, indicating that first-class females had the highest survival rate.

These insights help us understand the key factors that influenced survival rates during the Titanic disaster.