Titanic Dataset – Exploratory Data Analysis (EDA)

In [6]:

```
!pip install seaborn pandas matplotlib
```

```
Requirement already satisfied: seaborn in c:\users\chandini chamiyal\appdata\local\p
rograms\python\python312\lib\site-packages (0.13.2)
Requirement already satisfied: pandas in c:\users\chandini chamiyal\appdata\local\pr
ograms\python\python312\lib\site-packages (2.3.0)
Requirement already satisfied: matplotlib in c:\users\chandini chamiyal\appdata\loca
l\programs\python\python312\lib\site-packages (3.10.3)
Requirement already satisfied: numpy!=1.24.0,>=1.20 in c:\users\chandini chamiyal\ap
pdata\local\programs\python\python312\lib\site-packages (from seaborn) (2.3.1)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\chandini chamiyal
\appdata\local\programs\python\python312\lib\site-packages (from pandas) (2.9.0.post
0)
Requirement already satisfied: pytz>=2020.1 in c:\users\chandini chamiyal\appdata\lo
cal\programs\python\python312\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in c:\users\chandini chamiyal\appdata
\local\programs\python\python312\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\chandini chamiyal\appdat
a\local\programs\python\python312\lib\site-packages (from matplotlib) (1.3.2)
Requirement already satisfied: cycler>=0.10 in c:\users\chandini chamiyal\appdata\lo
cal\programs\python\python312\lib\site-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\chandini chamiyal\appda
ta\local\programs\python\python312\lib\site-packages (from matplotlib) (4.58.4)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\chandini chamiyal\appda
ta\local\programs\python\python312\lib\site-packages (from matplotlib) (1.4.8)
Requirement already satisfied: packaging>=20.0 in c:\users\chandini chamiyal\appdata
\local\programs\python\python312\lib\site-packages (from matplotlib) (25.0)
Requirement already satisfied: pillow>=8 in c:\users\chandini chamiyal\appdata\local
\programs\python\python312\lib\site-packages (from matplotlib) (11.0.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\chandini chamiyal\appdat
a\local\programs\python\python312\lib\site-packages (from matplotlib) (3.2.3)
Requirement already satisfied: six>=1.5 in c:\users\chandini chamiyal\appdata\local
\programs\python\python312\lib\site-packages (from python-dateutil>=2.8.2->pandas)
(1.17.0)
```

```
[notice] A new release of pip is available: 25.0.1 -> 25.1.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

In [7]:

```
import pandas as pd
```

In [8]:

```
df=pd.read_csv("train.csv")
df.head()
```

Out[8]:

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

In [12]:

```
print("\n Dataset Info:")
df.info()
```

```
 Dataset 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 [13]:

```
df.describe()
```

Out[13]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.00000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.20420
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.69342
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.00000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.91040
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.45420
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.00000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.32920

In [15]:

```
df.value_counts()
```

Out[15]:

```
PassengerId  Survived  Pclass  Name
Sex     Age   SibSp  Parch  Ticket    Fare     Cabin        Embarked
2            1         1      Cumings, Mrs. John Bradley (Florence Briggs Thayer)
female  38.0  1      0      PC 17599  71.2833  C85          C            1
4            1         1      Futrelle, Mrs. Jacques Heath (Lily May Peel)
female  35.0  1      0      113803    53.1000  C123         S            1
7            0         1      McCarthy, Mr. Timothy J
male    54.0  0      0      17463     51.8625  E46          S            1
11           1         3      Sandstrom, Miss. Marguerite Rut
female  4.0   1      1      PP 9549   16.7000  G6           S            1
12           1         1      Bonnell, Miss. Elizabeth
female  58.0  0      0      113783    26.5500  C103         S            1

..
872          1         1      Beckwith, Mrs. Richard Leonard (Sallie Monypeny)
female  47.0  1      1      11751     52.5542  D35          S            1
873          0         1      Carlsson, Mr. Frans Olof
male    33.0  0      0      695       5.0000   B51 B53 B55  S            1
880          1         1      Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)
female  56.0  0      1      11767     83.1583  C50          C            1
888          1         1      Graham, Miss. Margaret Edith
female  19.0  0      0      112053    30.0000  B42          S            1
890          1         1      Behr, Mr. Karl Howell
male    26.0  0      0      111369    30.0000  C148         C            1
Name: count, Length: 183, dtype: int64
```

In [24]:

```
df.isnull().sum()
```

Out[24]:

```
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 [25]:

```python
# Count how many survived (0 = No, 1 = Yes)
df['Survived'].value_counts()

# Gender count
df['Sex'].value_counts()

# Passenger Class count
df['Pclass'].value_counts()

# Embarked Port count
df['Embarked'].value_counts()
```

Out[25]:

```
Embarked
S    644
C    168
Q     77
Name: count, dtype: int64
```

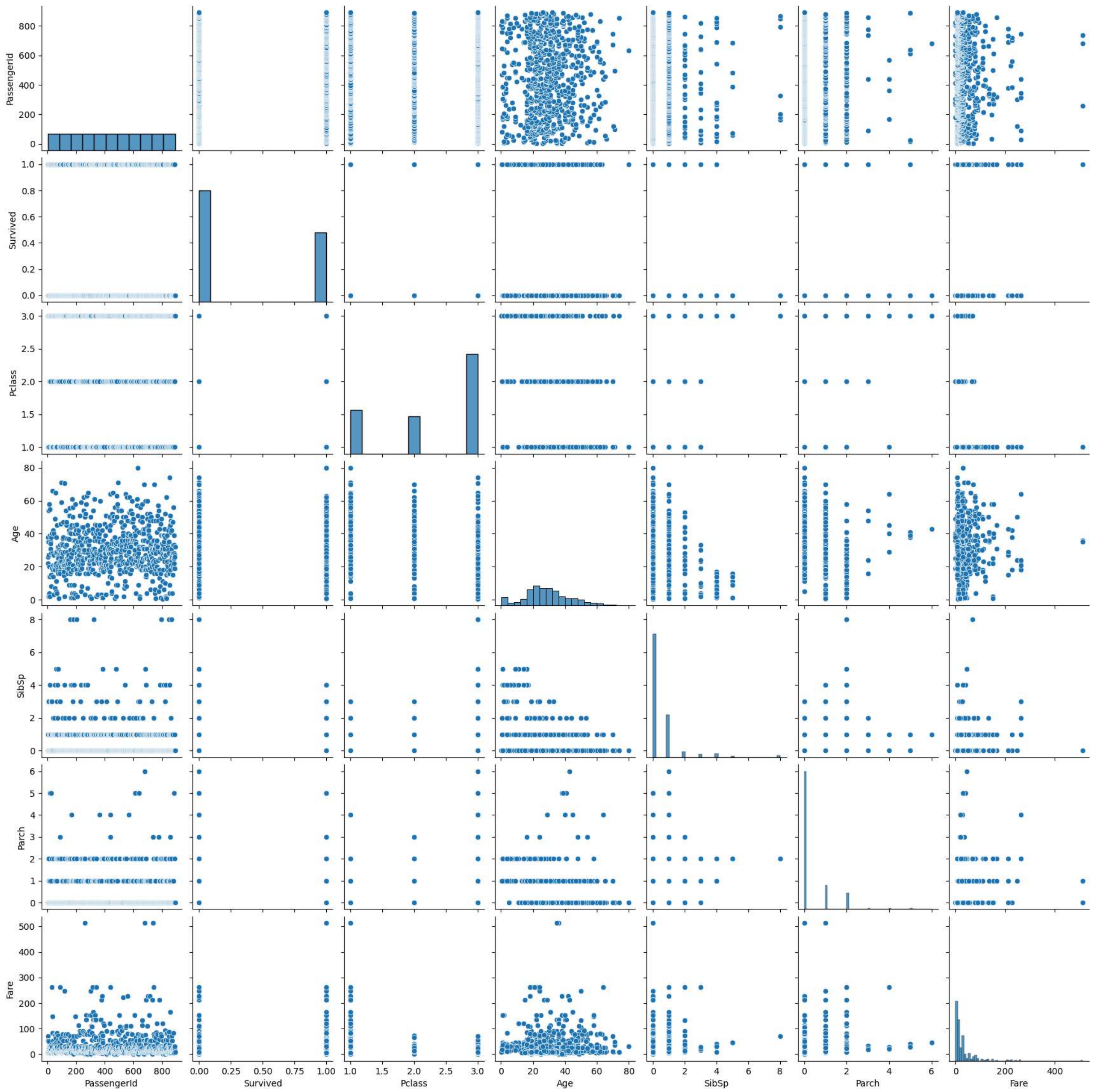
In [16]:

```python
#Visualization
import matplotlib.pyplot as plt
import seaborn as sns
```

```
Matplotlib is building the font cache; this may take a moment.
```

In [17]:

```python
sns.pairplot(df)
plt.show()
```



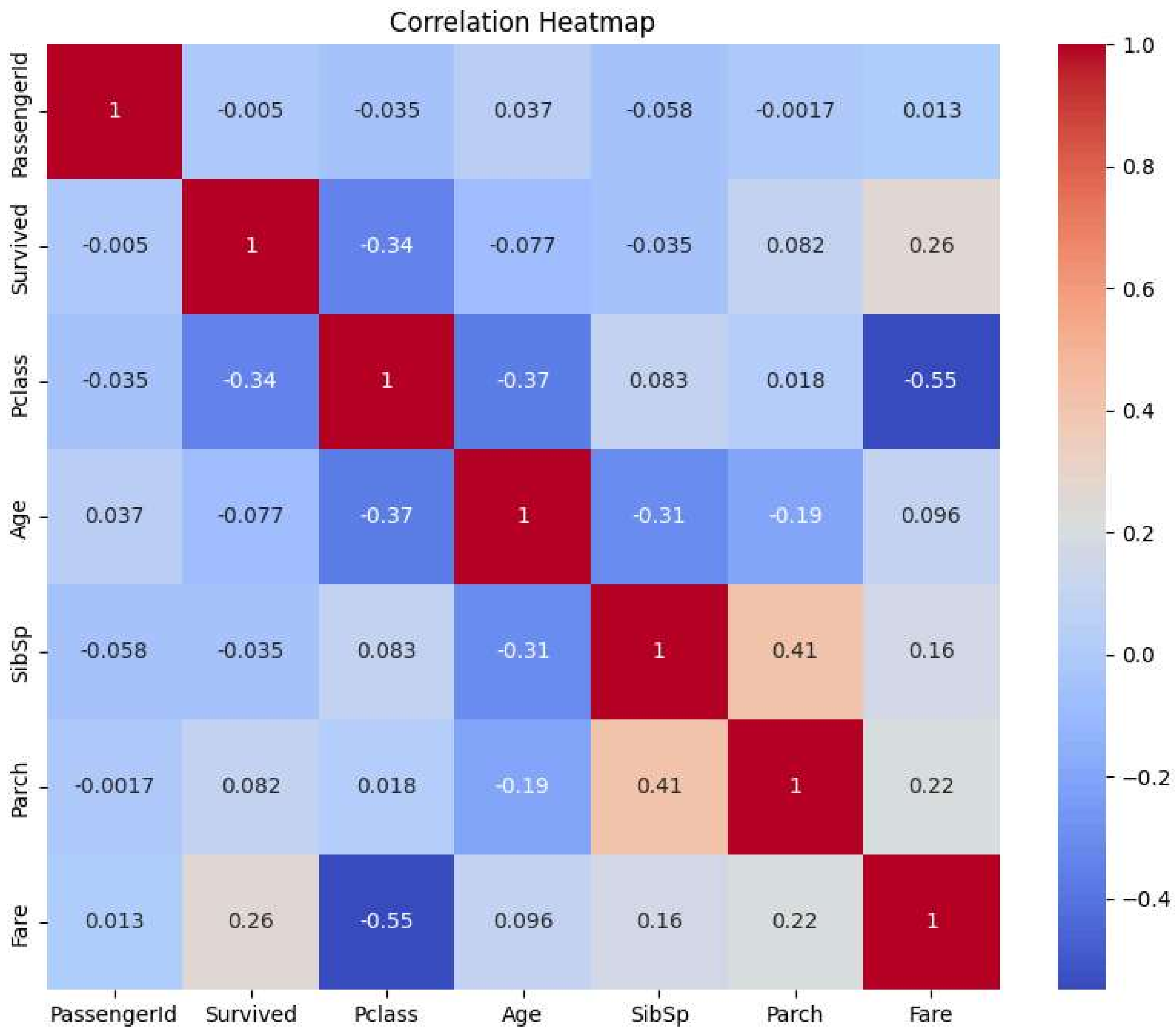
In [23]:

```
numeric_df = df.select_dtypes(include='number')



import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(10, 8))
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```

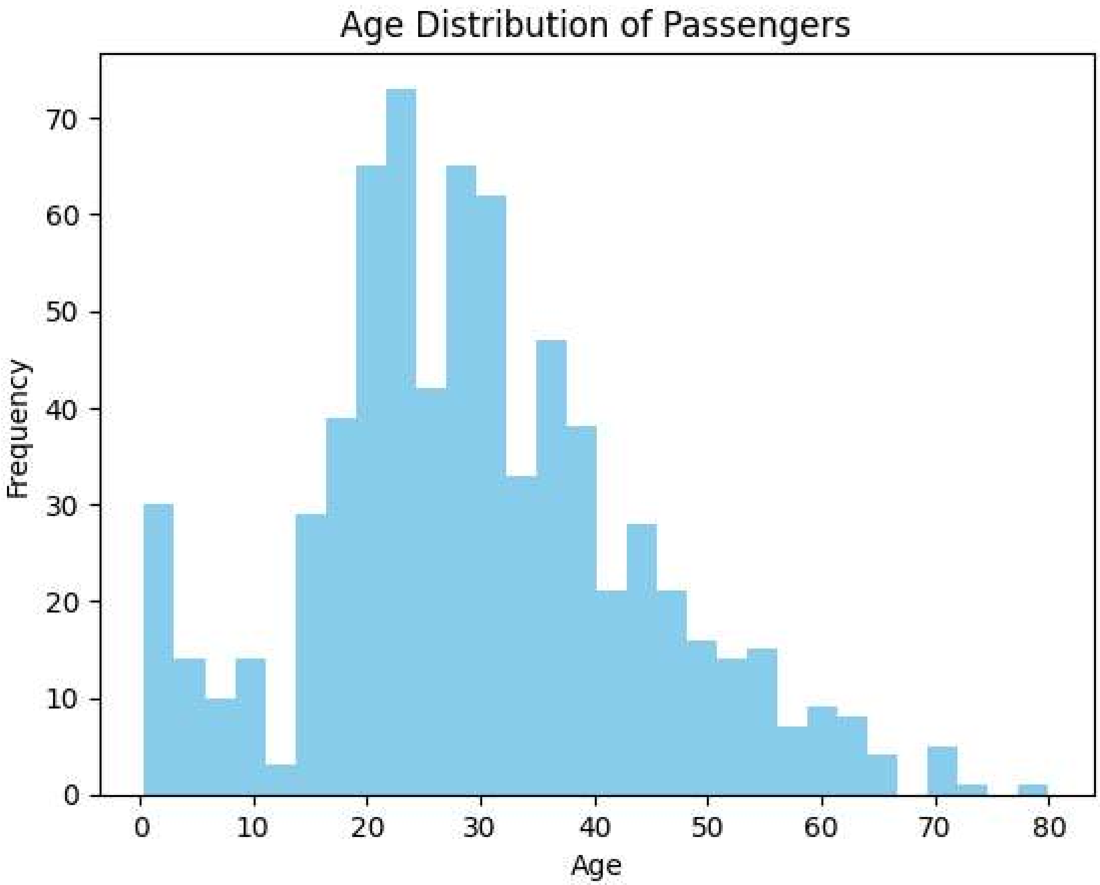


In [ ]:

```
-heatmap shows strong positive correlation between 'income' and 'spending_score'.
-Pairplot shows clusters between 'age' and 'purchases'.
```

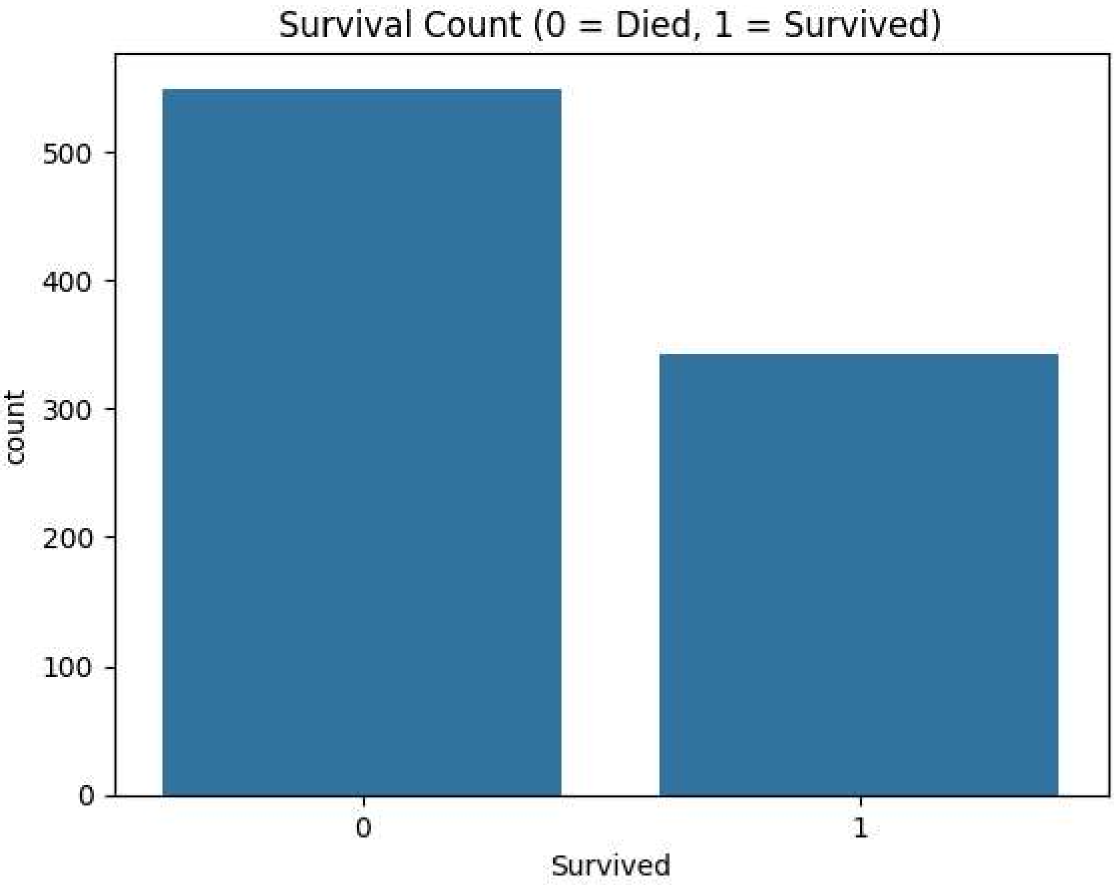
In [26]:

```
#histogram
df['Age'].plot(kind='hist', bins=30, color='skyblue')
plt.title("Age Distribution of Passengers")
plt.xlabel("Age")
plt.show()
print("Observation: Most passengers were aged between 20 and 40 years.")
```



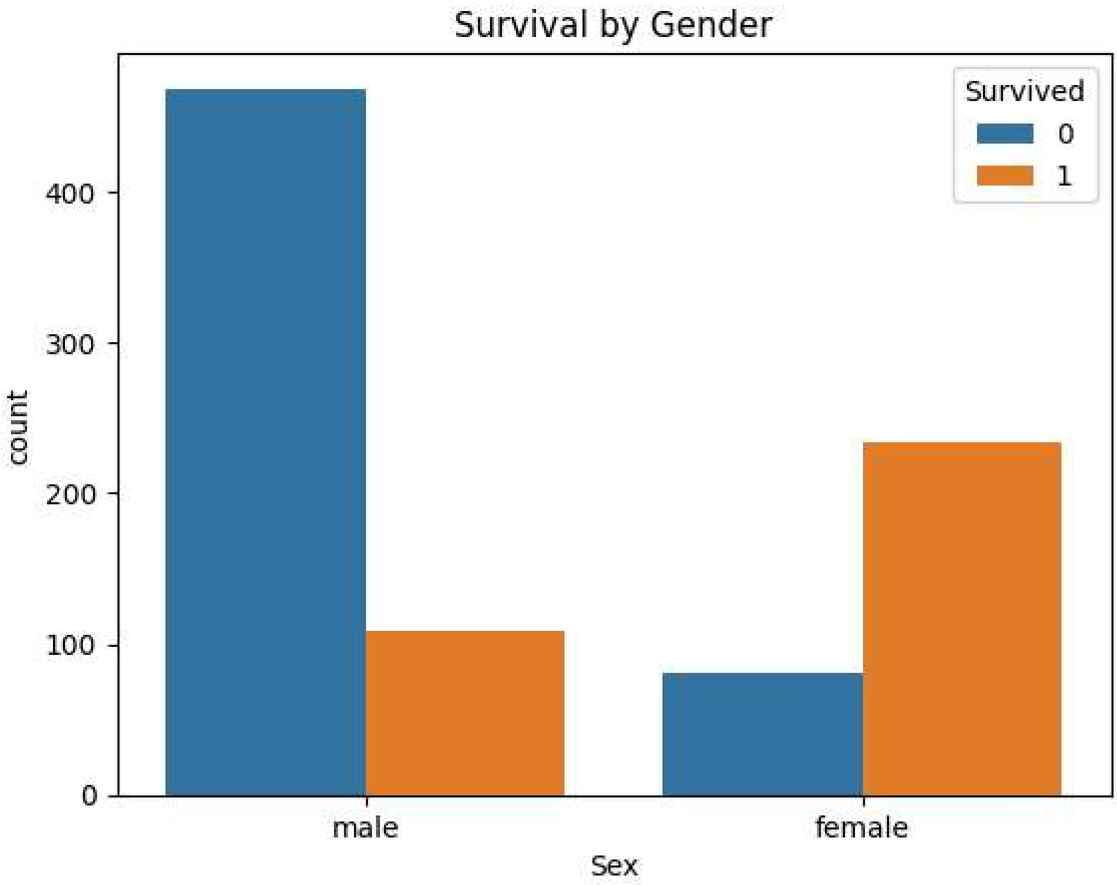
In [29]:

```
# Survival Count
sns.countplot(x='Survived', data=df)
plt.title("Survival Count (0 = Died, 1 = Survived)")
plt.show()
print("Observation: About 38% of passengers survived, while the majority (around 62
```



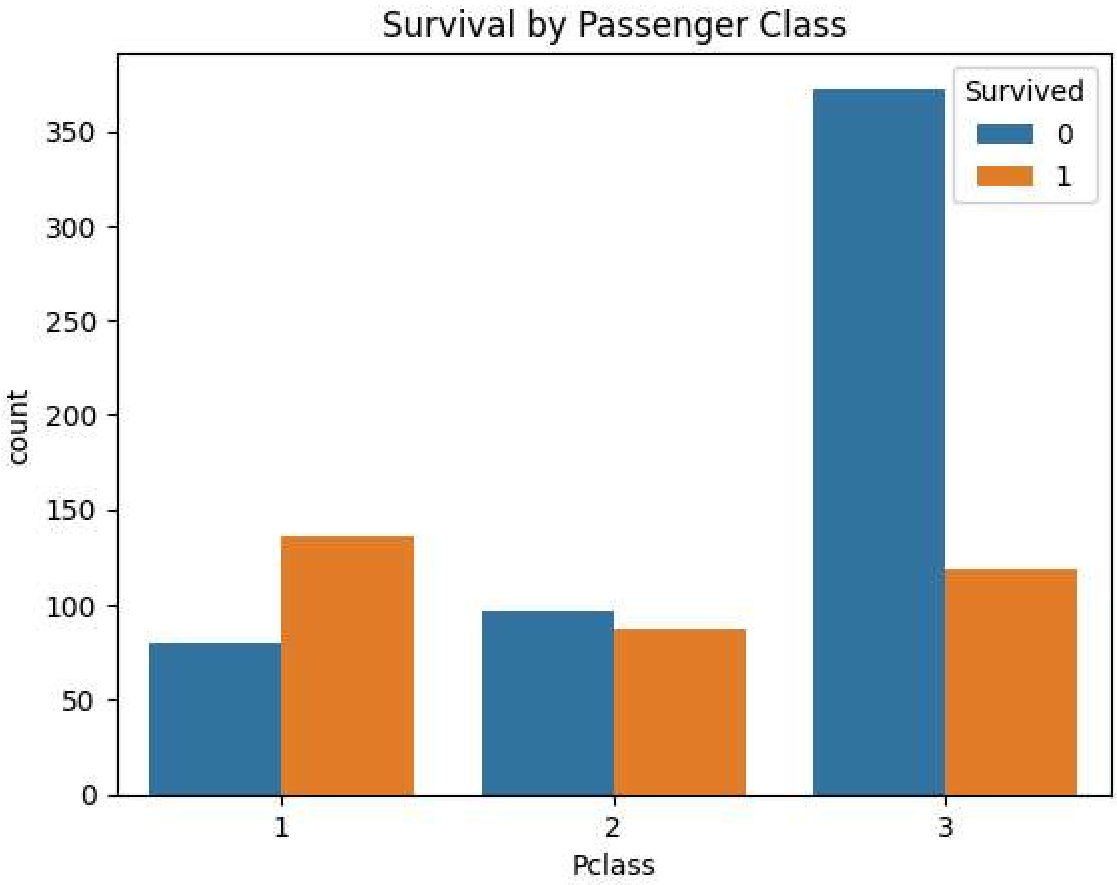
In [30]:

```
# Gender vs Survival
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title("Survival by Gender")
plt.show()
print("Observation: Females had a much higher survival rate compared to males.")
```



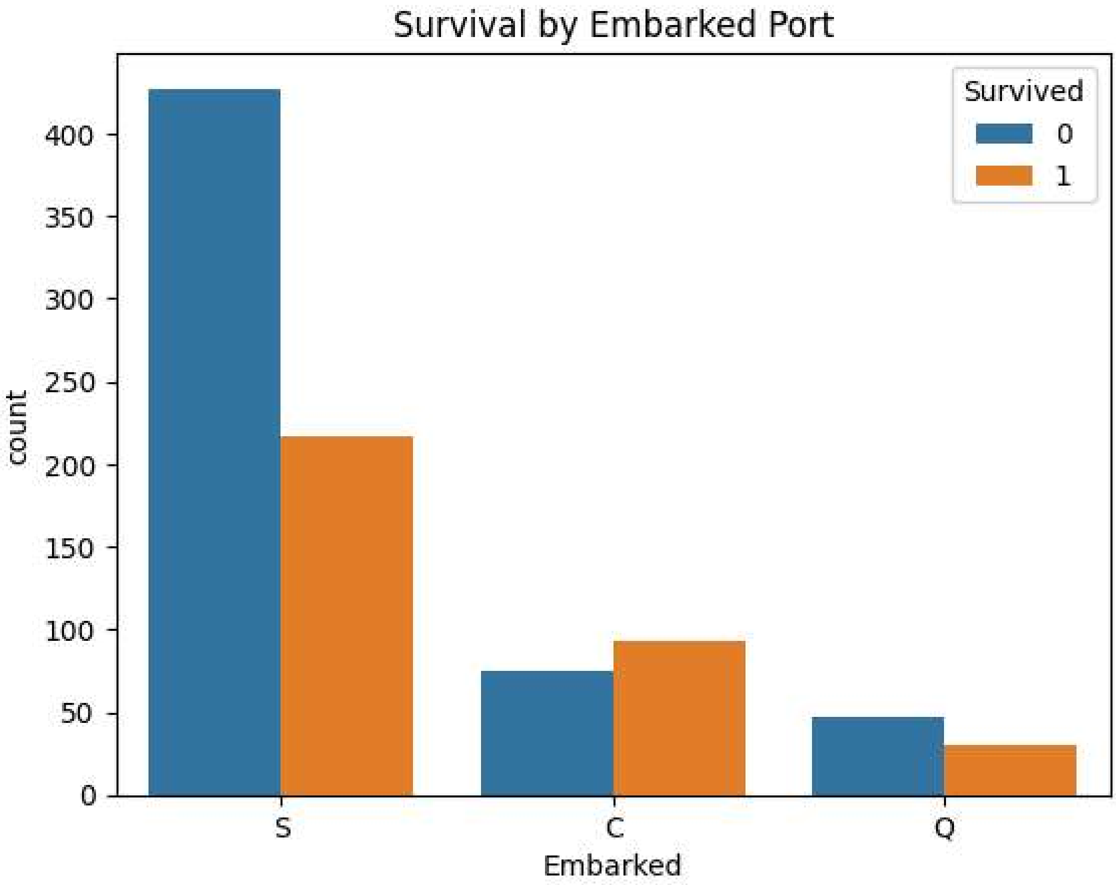
In [31]:

```
#Passenger Class vs Survival
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title("Survival by Passenger Class")
plt.show()
print("Observation: Passengers in 1st class had a higher survival rate than those i
```



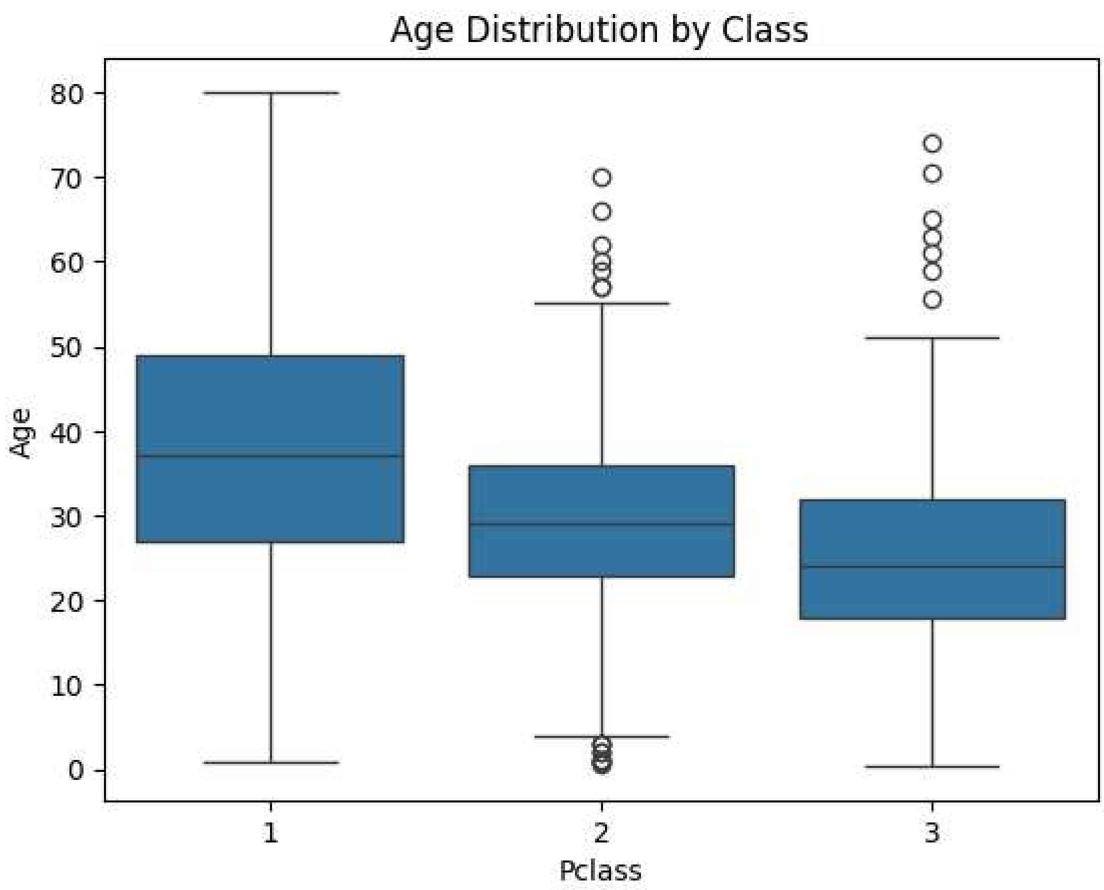
In [32]:

```
#Embarked Port vs Survival
sns.countplot(x='Embarked', hue='Survived', data=df)
plt.title("Survival by Embarked Port")
plt.show()
print("Observation: Passengers who boarded from port 'C' had a better survival rate
```



In [27]:

```
#Age Distribution by Class
sns.boxplot(x='Pclass', y='Age', data=df)
plt.title("Age Distribution by Class")
plt.show()
print("Observation: 1st class passengers were generally older; 3rd class passengers
```



In [34]:

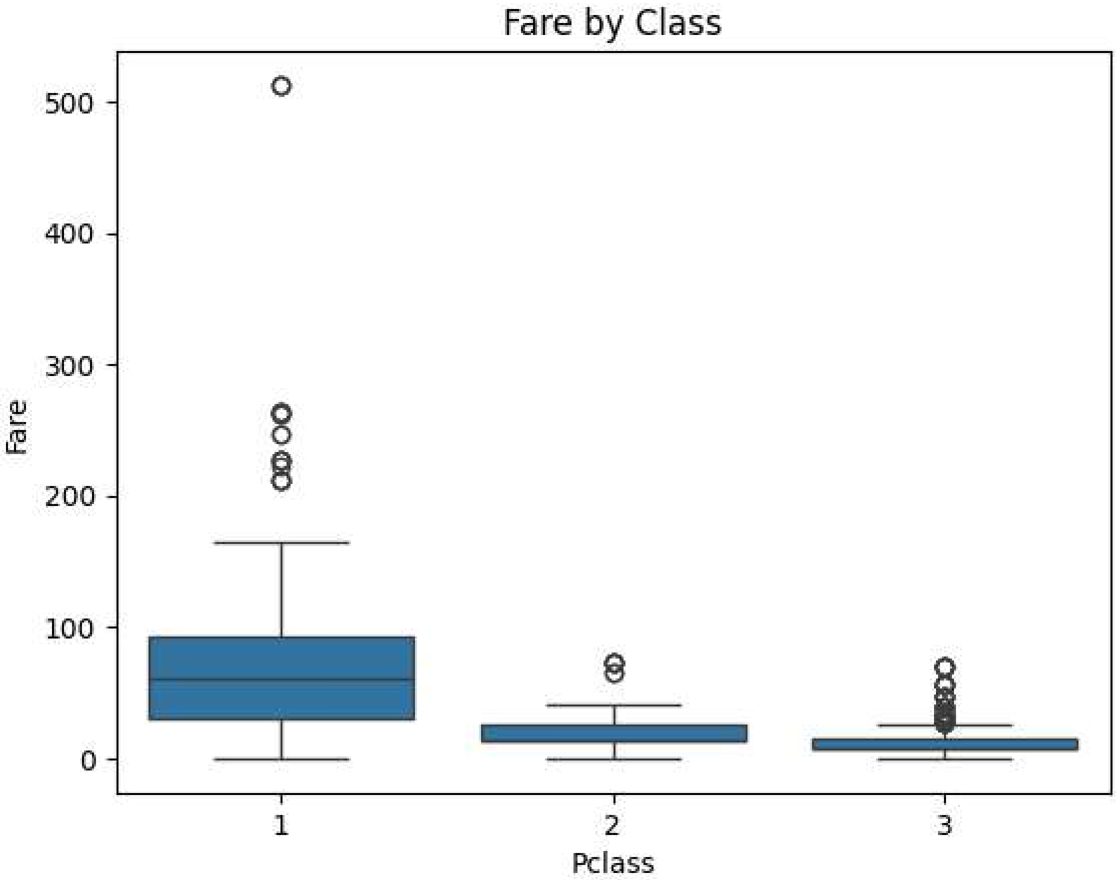
```
##Summary of Findings

-print( Around 38% of passengers survived the Titanic disaster.)
- Females had a significantly higher survival rate than males.
- Passengers in 1st class had much higher survival rates compared to 2nd and 3rd cl
- Most passengers boarded from port 'S', but survival rate was highest from port 'C
- Younger passengers and children had slightly higher chances of survival.
- 1st class passengers paid higher fares and were generally older.
```

```
  Cell In[34], line 5
    - Passengers in 1st class had much higher survival rates compared to 2nd and 3rd
class.
                    ^
SyntaxError: invalid decimal literal
```

In [35]:

```
#Fare by Class
sns.boxplot(x='Pclass', y='Fare', data=df)
plt.title("Fare by Class")
plt.show()
print("Observation: Passengers in 1st class paid higher fares, showing a wide range
```



```
Observation: Passengers in 1st class paid higher fares, showing a wide range of tick
et prices.
```

Summary of Findings

- Around 38% of passengers survived the Titanic disaster.
- Females had a significantly higher survival rate than males.
- Passengers in 1st class had much higher survival rates compared to 2nd and 3rd class.
- Most passengers boarded from port 'S', but survival rate was highest from port 'C'.
- Younger passengers and children had slightly higher chances of survival.
- 1st class passengers paid higher fares and were generally older.