In [8]:
```
# 1. Tools Setup
!pip install pandas matplotlib seaborn
```

```
Requirement already satisfied: pandas in c:\users\win10\anaconda3\lib\site-packages
(2.1.4)
Requirement already satisfied: matplotlib in c:\users\win10\anaconda3\lib\site-packa
ges (3.8.0)
Requirement already satisfied: seaborn in c:\users\win10\anaconda3\lib\site-packages
(0.12.2)
Requirement already satisfied: numpy<2,>=1.23.2 in c:\users\win10\anaconda3\lib\site
-packages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\win10\anaconda3\li
b\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\win10\anaconda3\lib\site-pac
kages (from pandas) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\win10\anaconda3\lib\site-p
ackages (from pandas) (2023.3)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\win10\anaconda3\lib\site
-packages (from matplotlib) (1.2.0)
Requirement already satisfied: cycler>=0.10 in c:\users\win10\anaconda3\lib\site-pac
kages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\win10\anaconda3\lib\sit
e-packages (from matplotlib) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\win10\anaconda3\lib\sit
e-packages (from matplotlib) (1.4.4)
Requirement already satisfied: packaging>=20.0 in c:\users\win10\anaconda3\lib\site-
packages (from matplotlib) (23.1)
Requirement already satisfied: pillow>=6.2.0 in c:\users\win10\anaconda3\lib\site-pa
ckages (from matplotlib) (10.2.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\win10\anaconda3\lib\site
-packages (from matplotlib) (3.0.9)
Requirement already satisfied: six>=1.5 in c:\users\win10\anaconda3\lib\site-package
s (from python-dateutil>=2.8.2->pandas) (1.16.0)
```

In [9]:
```
# 2. Import Libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

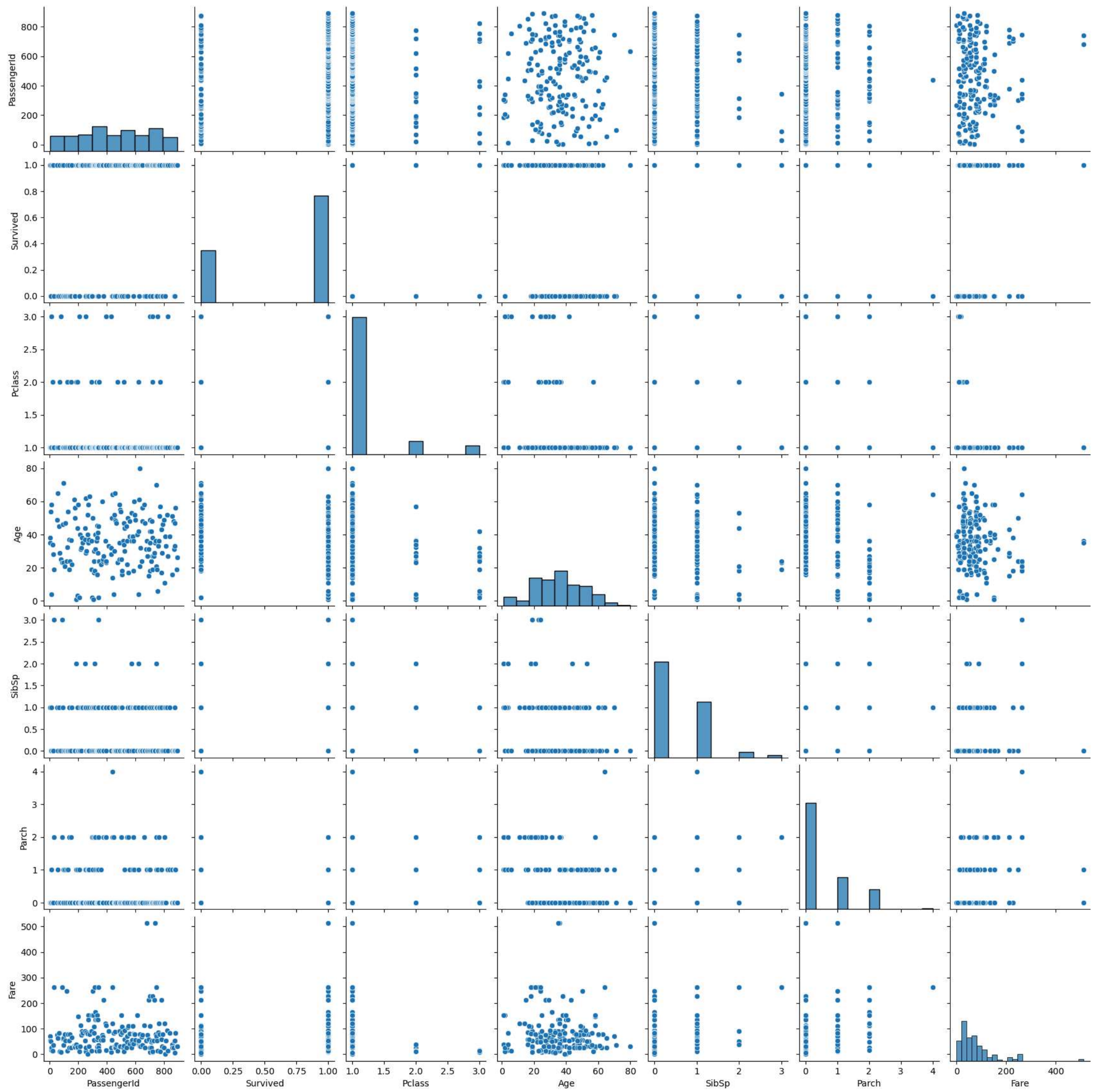
In [10]:
```
#3. Data Loading
df = pd.read_csv('train.csv')
```

In [11]:
```
#4. Basic Data Inspection
#summary statistics
print(df.describe())

#Data types and Non-nu;; counts
print(df.info())

# Value counts for a categorical column(e.g.,'Product_Category')
# print(df['Product_Category'].value_counts())
```

```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000   
mean    446.000000    0.383838    2.308642   29.699118    0.523008   
std     257.353842    0.486592    0.836071   14.526497    1.102743   
min       1.000000    0.000000    1.000000    0.420000    0.000000   
25%     223.500000    0.000000    2.000000   20.125000    0.000000   
50%     446.000000    0.000000    3.000000   28.000000    0.000000   
75%     668.500000    1.000000    3.000000   38.000000    1.000000   
max     891.000000    1.000000    3.000000   80.000000    8.000000   

            Parch        Fare  
count  891.000000  891.000000  
mean     0.381594   32.204208  
std      0.806057   49.693429  
min      0.000000    0.000000  
25%      0.000000    7.910400  
50%      0.000000   14.454200  
75%      0.000000   31.000000  
max      6.000000  512.329200  
<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
None
```

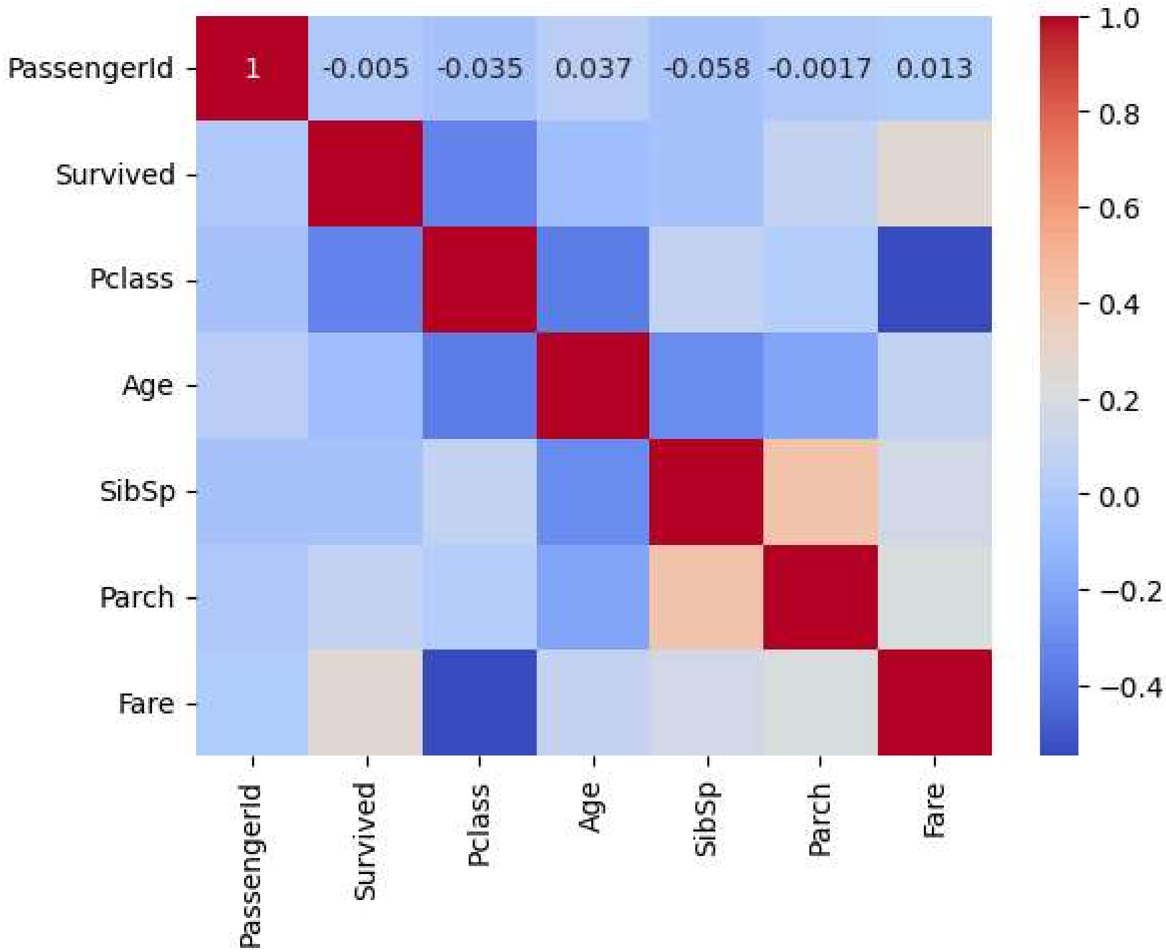
In [30]:

```
#Visualization with Seaborn Pairplot(relationships across dataset)
import numpy as np
df.replace([np.inf,-np.inf],np.nan, inplace=True)
df.dropna(inplace=True)

import warnings
warnings.filterwarnings("ignore")

sns.pairplot(df.select_dtypes(include='number'))
plt.show()
```

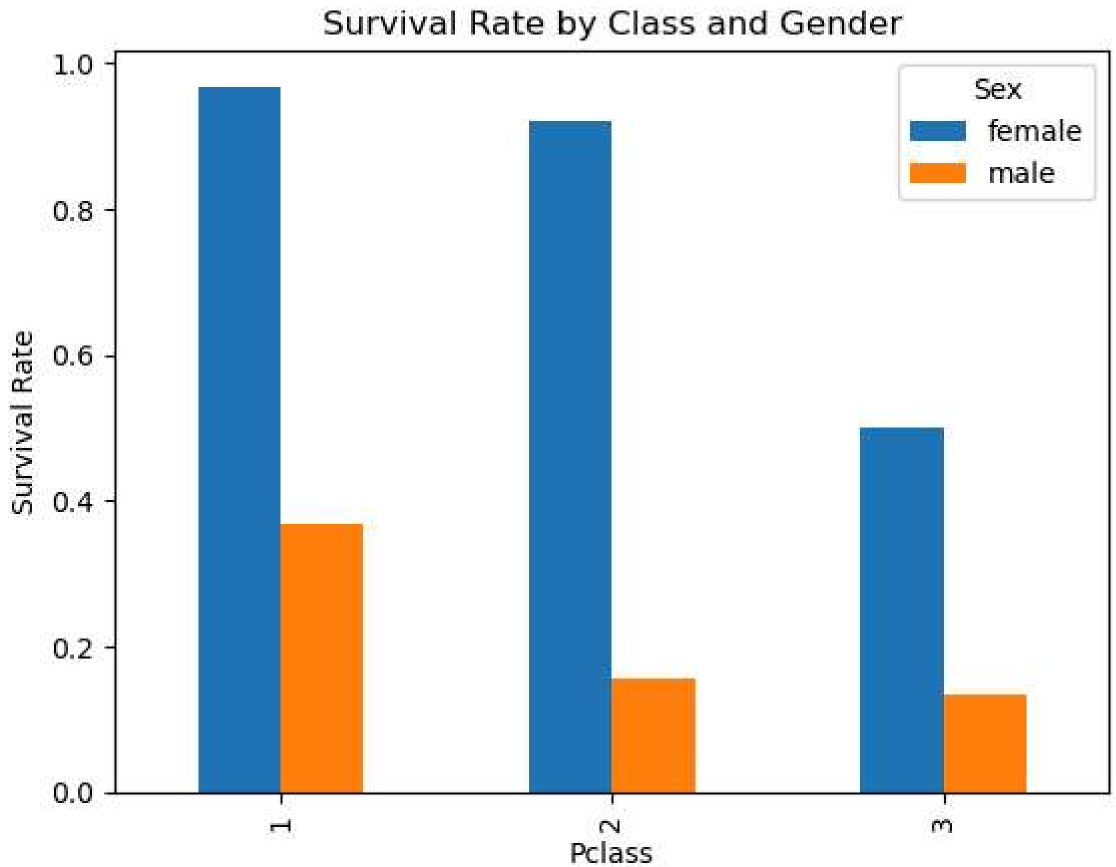


In [14]:

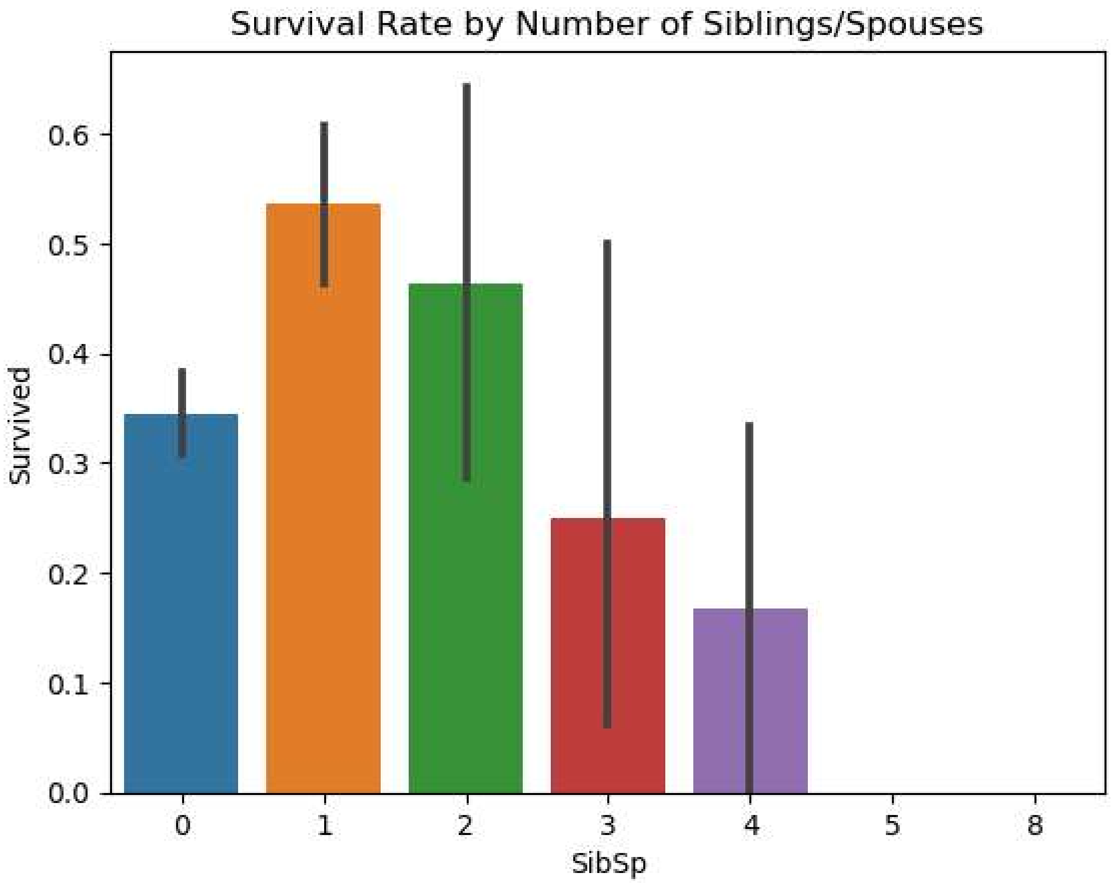
```
#correlation
numeric_df = df.select_dtypes(include = 'number')

#to Calculate the Heatmap
corr = numeric_df.corr()

#To plot the Heatmap
import seaborn as sns
import matplotlib.pyplot as plt

sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.show()
```

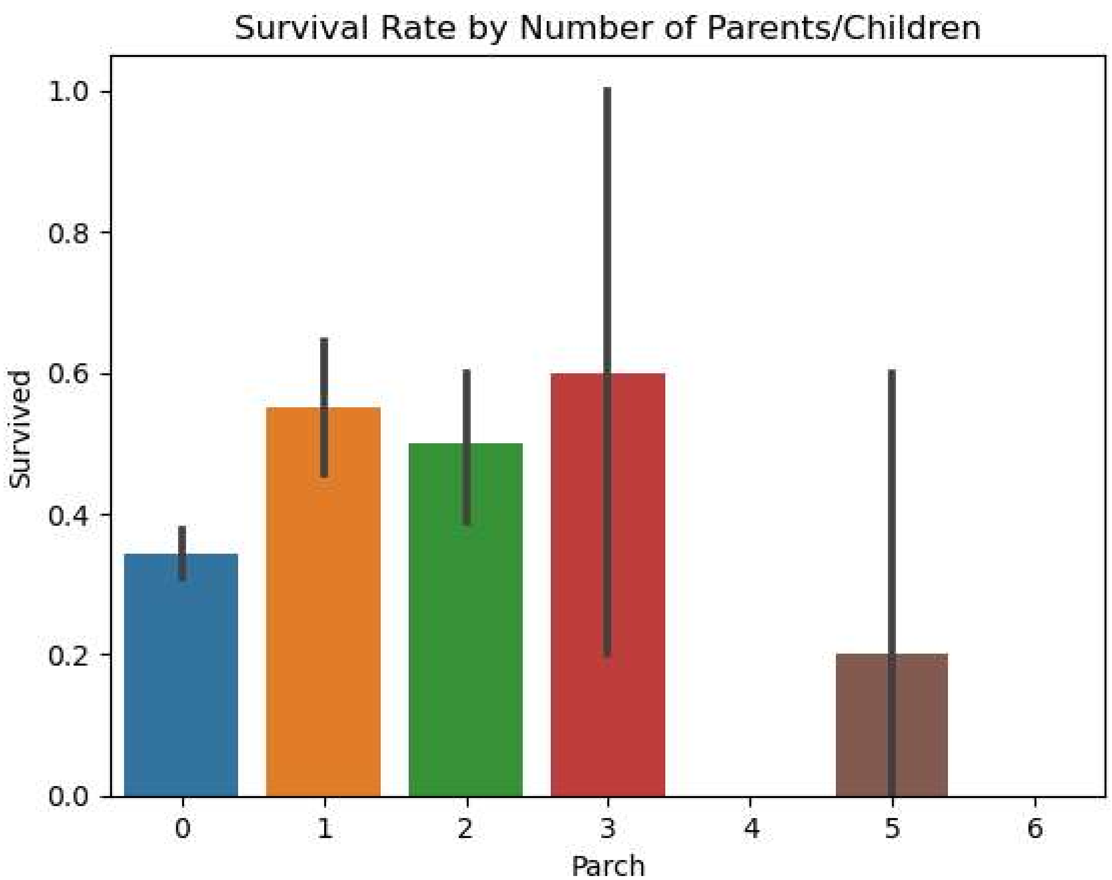


In [19]:
```
#5. Identifying relationship and trends

#Survaivability by class and gender
survival_rate = df.groupby(['Pclass','Sex'])['Survived'].mean().unstack()
survival_rate.plot(kind='bar',stacked= False)
plt.ylabel('Survival Rate')
plt.title('Survival Rate by Class and Gender')
plt.show()
```

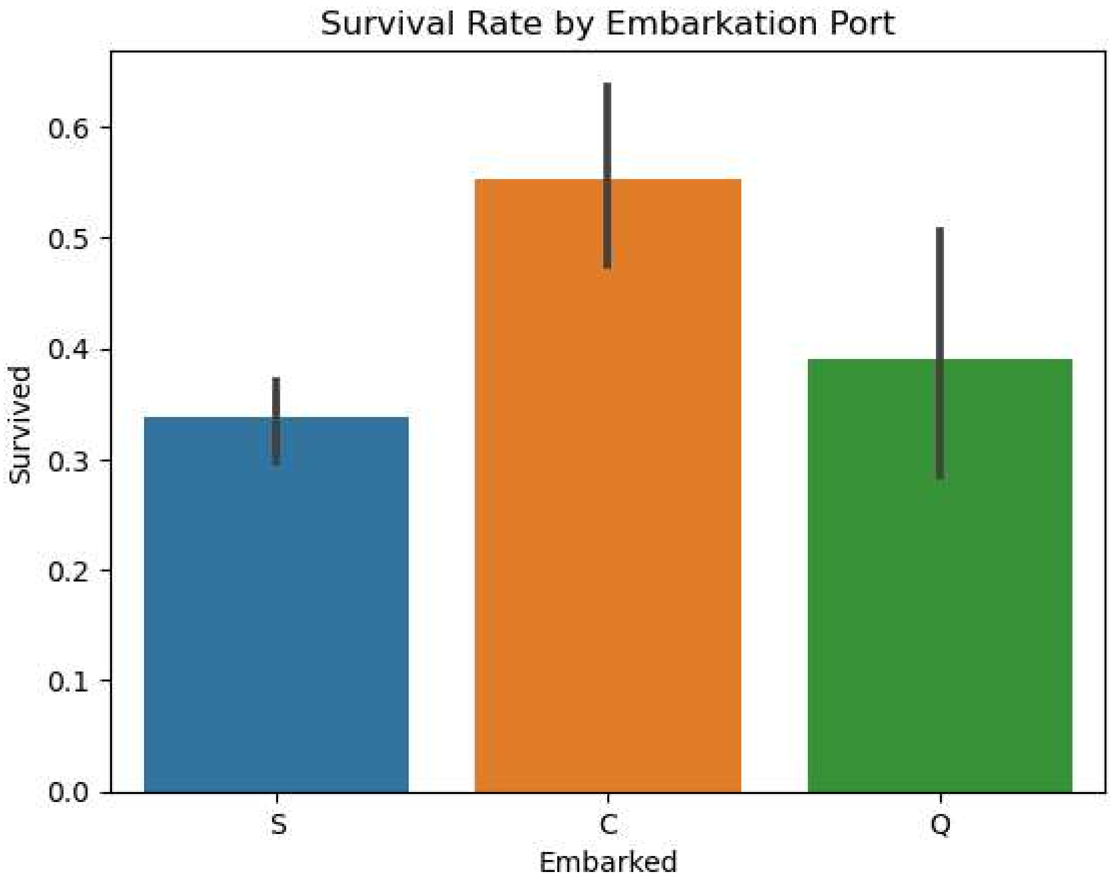


In [24]:
```
#Family Relationships
sns.barplot(x='SibSp', y='Survived', data=df)
plt.title('Survival Rate by Number of Siblings/Spouses')
plt.show()

sns.barplot(x='Parch', y='Survived', data=df)
plt.title('Survival Rate by Number of Parents/Children')
plt.show()
```

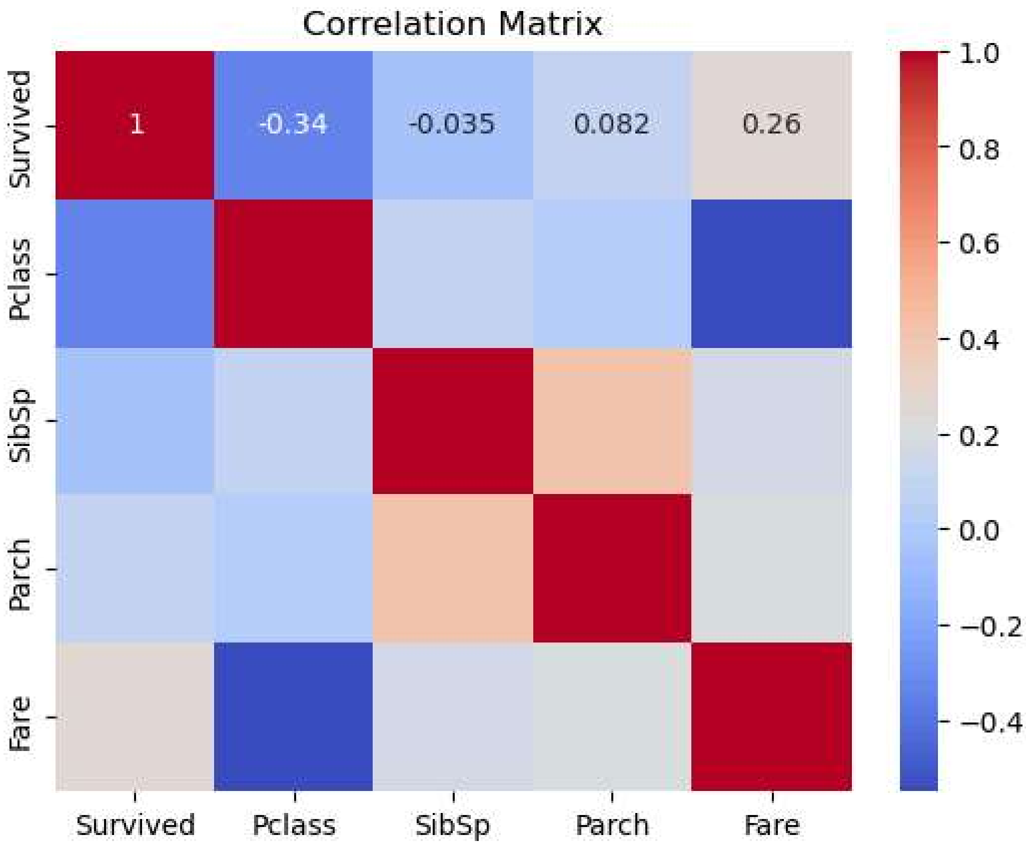

Survival Rate by Number of Siblings/Spouses
Survived
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0 1 2 3 4 5 8
SibSp


Survival Rate by Number of Parents/Children
Survived
0.0
0.2
0.4
0.6
0.8
1.0
0 1 2 3 4 5 6
Parch

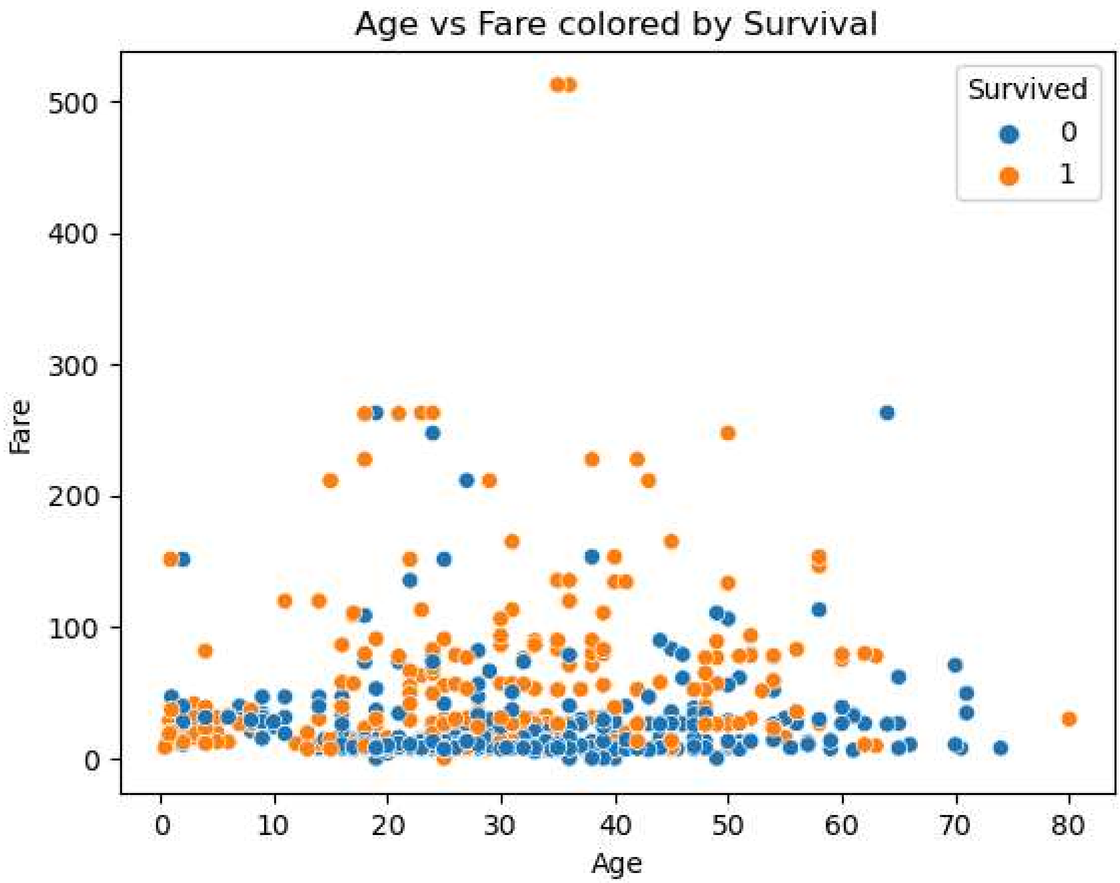
In [25]:
```
#Embarked and Survival
sns.barplot(x='Embarked', y='Survived', data=df)
plt.title('Survival Rate by Embarkation Port')
plt.show()
```



In [26]:
```
#5. Correlation Matrix
numeric_df = df[['Survived', 'Pclass', 'SibSp', 'Parch', 'Fare']]
corr = numeric_df.corr()
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



In [27]:
```
#Visualization with Histograms, Boxplots, Scatterplots.
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df)
plt.title('Age vs Fare colored by Survival')
plt.show()
```



Observation for Each Visual

1. Pairplot.

Data Distribution: The histograms on the diagonal show that numerical features (Age, SibSp, Parch, Fare) are generally right-skewed, indicating a high concentration of passengers who are younger, travel alone, or paid lower fares. Survived vs. Pclass: Visually, the scatter plots and histograms indicate that a significantly higher proportion of survivors (Survived=1) are in Pclass 1, while the majority of non-survivors (Survived=0) are concentrated in Pclass 3. Fare vs. Survived: There is a strong visual trend showing that passengers who paid higher fares were more likely to survive.

2. Survival Rate by Class and Gender.

Gender Dominance: This chart clearly demonstrates that gender was the single most dominant factor for survival. Females had a dramatically higher survival rate than males across all passenger classes. Class Effect: Females in Class 1 had nearly a 100% survival rate. Even females in Class 3 had a survival rate of approximately 50%, which is much higher than males in any class. Males in Class 1 had the highest male survival rate (around 38%), but this rate dropped significantly for males in Class 2 and Class 3 (around 15%).

3. Survival Rate by Number of Siblings/Spouses (SibSp).

Optimal Group Size: Passengers traveling with 1 or 2 siblings/spouses had the highest survival rates (over 50% for both). Traveling Alone: Passengers with 0 SibSp (traveling alone) had a lower survival rate (around 35%) compared to those with small families. Large Groups: Survival rates drop off significantly for passengers in large groups (SibSp of 4, 5, or 8), approaching zero for the largest groups.

4. Survival Rate by Number of Parents/Children (Parch).

Optimal Group Size: Similar to SibSp, passengers with 1, 2, or 3 parents/children had the best chance of survival (rates between 50% and 60%). Traveling Alone: Passengers with 0 Parch (no parents or children) had a survival rate of about 35%.

5. Survival Rate by Embarkation Port.

Port C Advantage: Passengers who embarked at Port C (Cherbourg) had the highest survival rate (approximately 55%). Port S Disadvantage: Passengers from Port S (Southampton) had the lowest survival rate (around 34%). This suggests that passengers boarding at Cherbourg might have been disproportionately from higher passenger classes.

6. Correlation Matrix.

Strongest Predictors of Survival: Pclass has the strongest negative correlation with Survived (−0.34). This indicates that the lower the class number (First Class), the higher the chance of survival. Fare has a moderate positive correlation with Survived (0.26), confirming that higher fares were linked to survival. Other Correlations: Age and SibSp have negligible correlation with survival (close to zero).

7. Age vs Fare colored by Survival.

High Fare is Key: The most striking observation is the clustering of survivors (orange dots, Survived=1) in the high Fare range (above 100). Most points above a fare of 200 are survivors. High Density of Non-Survivors: The greatest density of non-survivors (blue dots, Survived=0) is found in the low Fare range (below 50), regardless of age. Children Survival: Visually, the youngest ages (0-10) appear to have a higher proportion of survivors relative to the overall density of that age cluster.

📝 Summary of Findings

The analysis consistently shows that survival was not random but was highly influenced by socio-economic and demographic factors, adhering to a "women and children first" policy that prioritized certain groups. Gender and Class are Dominant: Female gender and higher passenger class (Pclass 1) were the two most critical factors for survival. Females in Class 1 had a near-perfect survival rate. Wealth was a Major Advantage: The strong positive correlation between Fare and survival (0.26) confirms that passengers who paid more were far more likely to be saved. This is intertwined with Pclass, as higher classes had higher fares. Family Structure Matters: Passengers traveling alone or in very large groups had the lowest survival rates. The highest survival rates were observed for individuals traveling with small families (1 to 3 immediate relatives: SibSp or Parch). Embarkation Bias: Passengers who boarded at Cherbourg (C) had a higher survival rate than those from Southampton (S), suggesting a pre-existing bias in the passenger demographics at that port (likely more high-fare, first-class passengers).