In [35]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('seaborn-v0_8')
sns.set_palette('pastel')
```

In [36]:
```
df = pd.read_csv(r"C:\Users\tedla\Downloads\train_data.csv")
```

In [37]:
```
print("Shape of dataset:", df.shape)
print("\nColumn Names:", df.columns.tolist())
print("\nFirst 5 rows:\n", df.head())
```

```
Shape of dataset: (891, 12)

Column Names: ['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibS
p', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked']

First 5 rows:
    PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

In [38]:
```
print("\nData Info:\n")
print(df.info())
print("\nSummary Statistics:\n")
print(df.describe(include='all'))
```

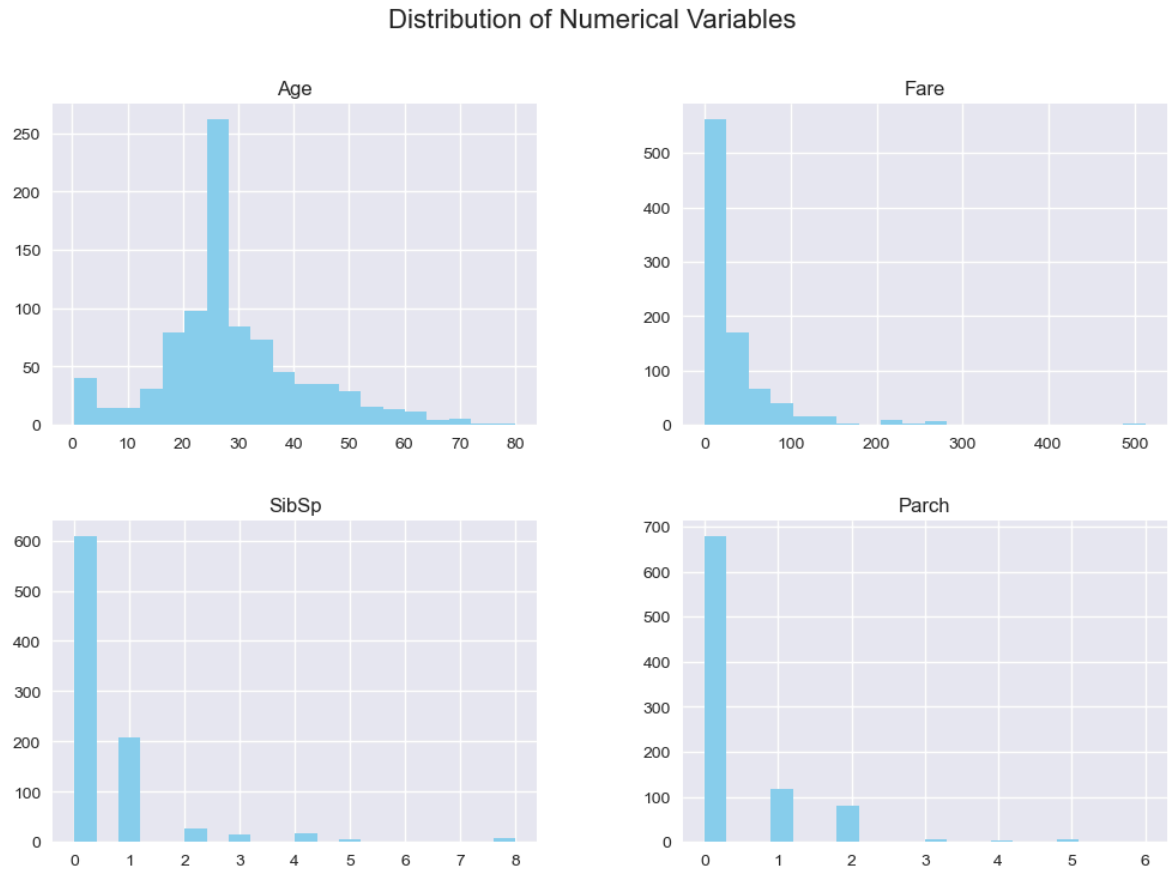
```
Data Info:

<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
None

Summary Statistics:

        PassengerId    Survived      Pclass                    Name   Sex  \
count    891.000000  891.000000  891.000000                     891   891   
unique          NaN         NaN         NaN                     891     2   
top             NaN         NaN         NaN  Braund, Mr. Owen Harris  male   
freq            NaN         NaN         NaN                       1   577   
mean     446.000000    0.383838    2.308642                     NaN   NaN   
std      257.353842    0.486592    0.836071                     NaN   NaN   
min        1.000000    0.000000    1.000000                     NaN   NaN   
25%      223.500000    0.000000    2.000000                     NaN   NaN   
50%      446.000000    0.000000    3.000000                     NaN   NaN   
75%      668.500000    1.000000    3.000000                     NaN   NaN   
max      891.000000    1.000000    3.000000                     NaN   NaN   

                Age       SibSp       Parch  Ticket        Fare    Cabin  \
count   714.000000  891.000000  891.000000     891  891.000000      204   
unique         NaN         NaN         NaN     681         NaN      147   
top            NaN         NaN         NaN  347082         NaN  B96 B98   
freq           NaN         NaN         NaN       7         NaN        4   
mean     29.699118    0.523008    0.381594     NaN   32.204208      NaN   
std      14.526497    1.102743    0.806057     NaN   49.693429      NaN   
min       0.420000    0.000000    0.000000     NaN    0.000000      NaN   
25%      20.125000    0.000000    0.000000     NaN    7.910400      NaN   
50%      28.000000    0.000000    0.000000     NaN   14.454200      NaN   
75%      38.000000    1.000000    0.000000     NaN   31.000000      NaN   
max      80.000000    8.000000    6.000000     NaN  512.329200      NaN   

       Embarked  
count       889  
unique        3  
top           S  
freq        644  
mean        NaN  
std         NaN  
min         NaN  
25%         NaN
```

```
50%          NaN
75%          NaN
max          NaN
```

In [39]:
```
print("\nMissing Values:\n", df.isnull().sum())
df['Age'].fillna(df['Age'].median(), inplace=True)
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
df.drop(columns=['Cabin'], inplace=True)
```

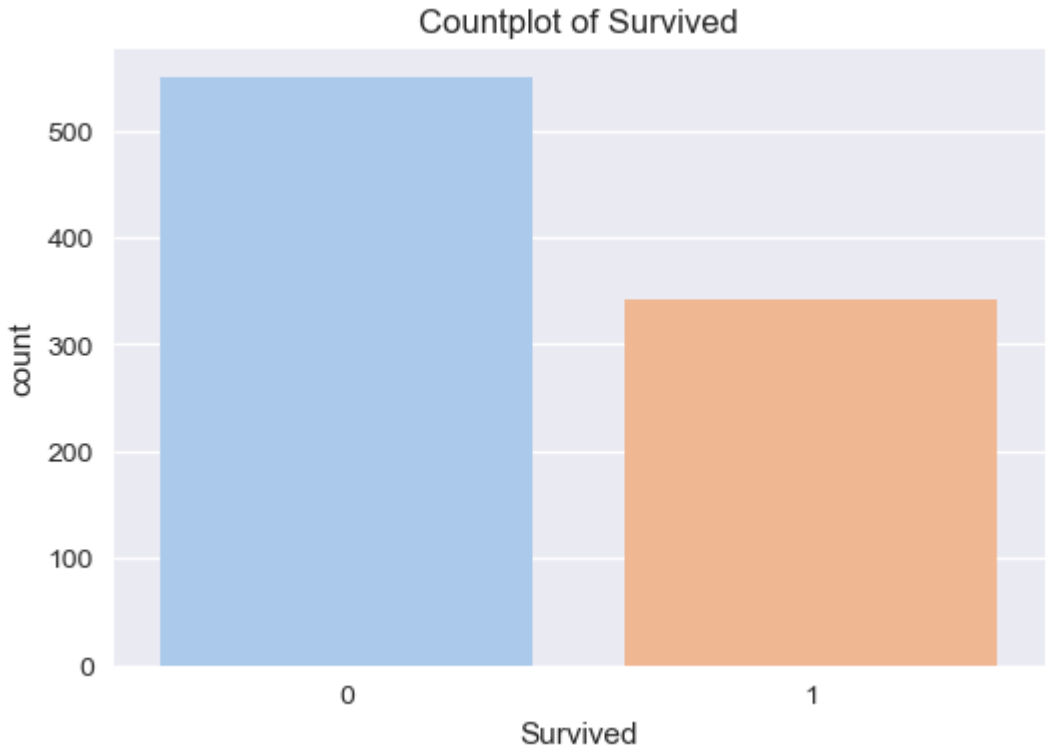
```
Missing Values:
 PassengerId      0
Survived          0
Pclass            0
Name              0
Sex               0
Age             177
SibSp             0
Parch             0
Ticket            0
Fare              0
Cabin           687
Embarked          2
dtype: int64
```

In [40]:
```
num_cols = ['Age', 'Fare', 'SibSp', 'Parch']
cat_cols = ['Survived', 'Pclass', 'Sex', 'Embarked']
df[num_cols].hist(bins=20, figsize=(12,8), color='skyblue')
plt.suptitle("Distribution of Numerical Variables", fontsize=16)
plt.show()
```

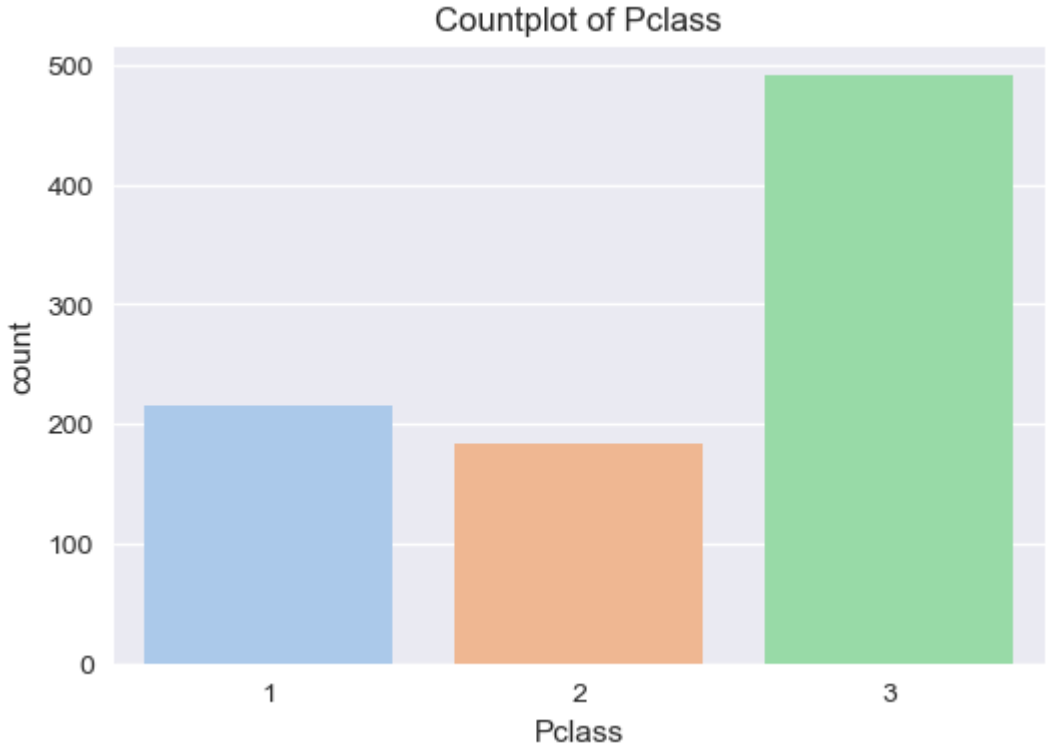


In [48]:
```
for col in cat_cols:
    plt.figure(figsize=(6,4))
    sns.countplot(x=col, data=df)
    plt.title(f"Countplot of {col}")
```

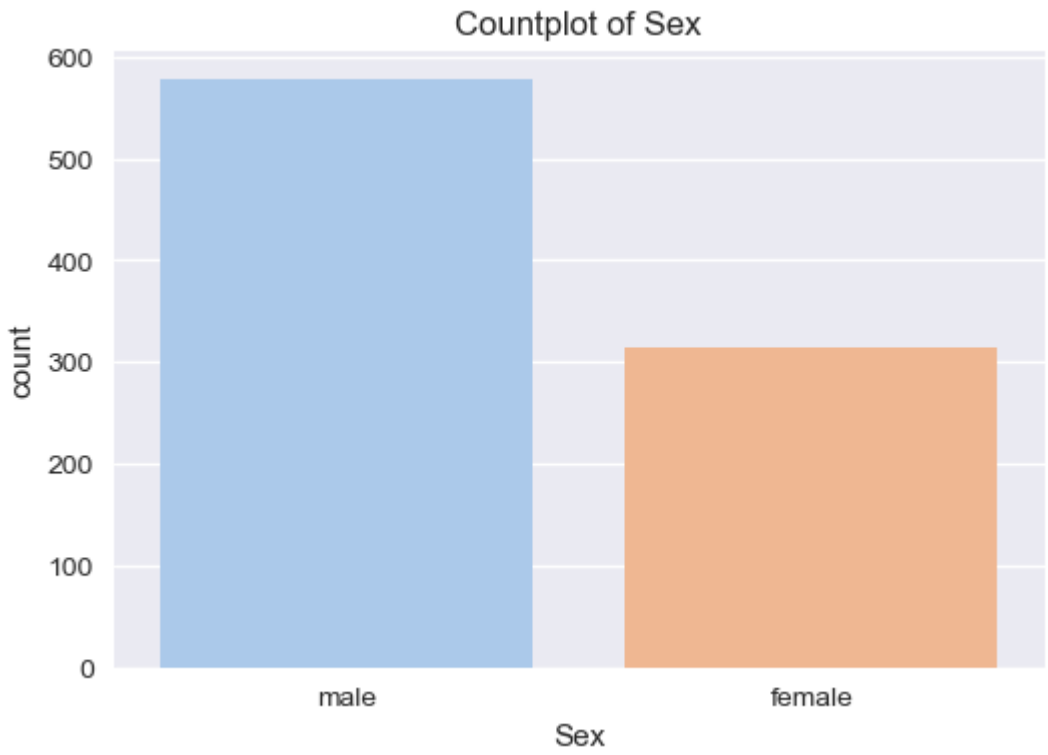
```
    plt.show()
    if col == 'Survived':
        print("Observation: Majority of passengers did not survive (~62%).")
    elif col == 'Pclass':
        print("Observation: Most passengers traveled in 3rd class.")
    elif col == 'Sex':
        print("Observation: More male passengers than female passengers.")
    elif col == 'Embarked':
        print("Observation: Most passengers embarked from port S.")
```



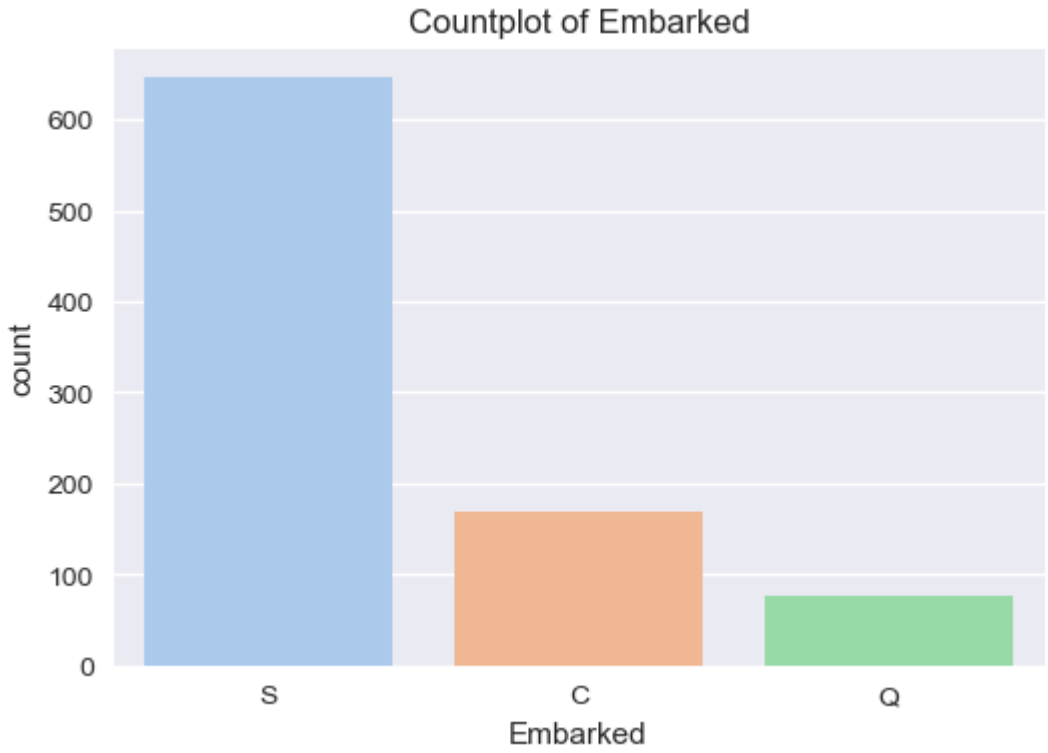

```
Observation: Majority of passengers did not survive (~62%).
```




```
Observation: Most passengers traveled in 3rd class.
```

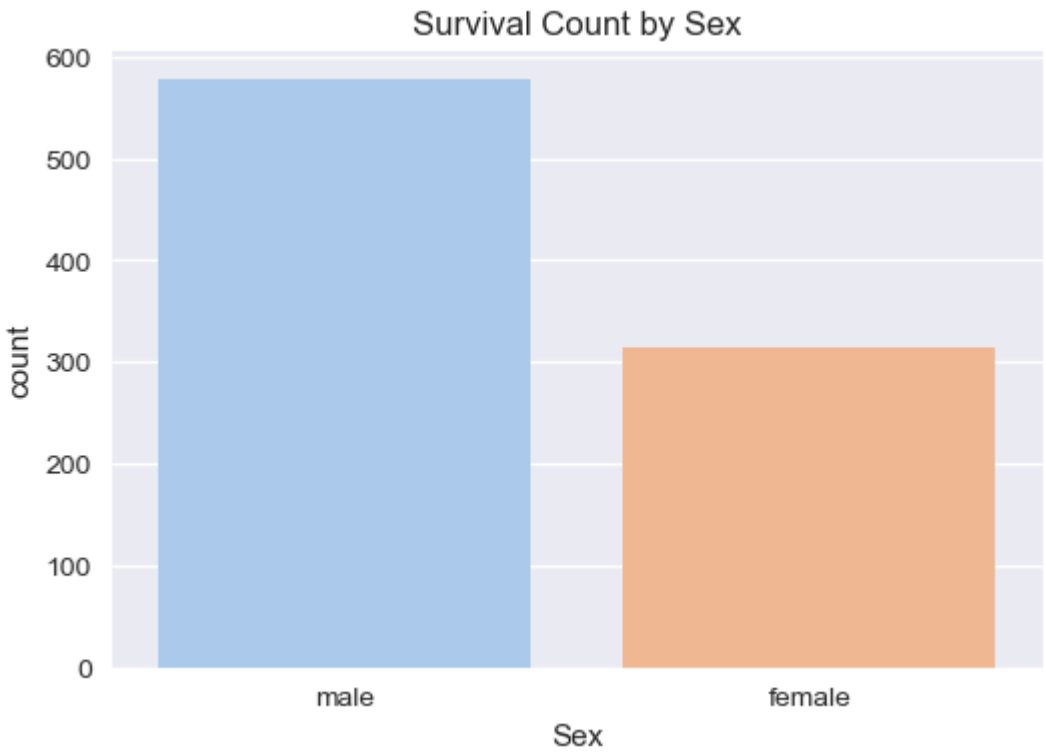


Observation: More male passengers than female passengers.



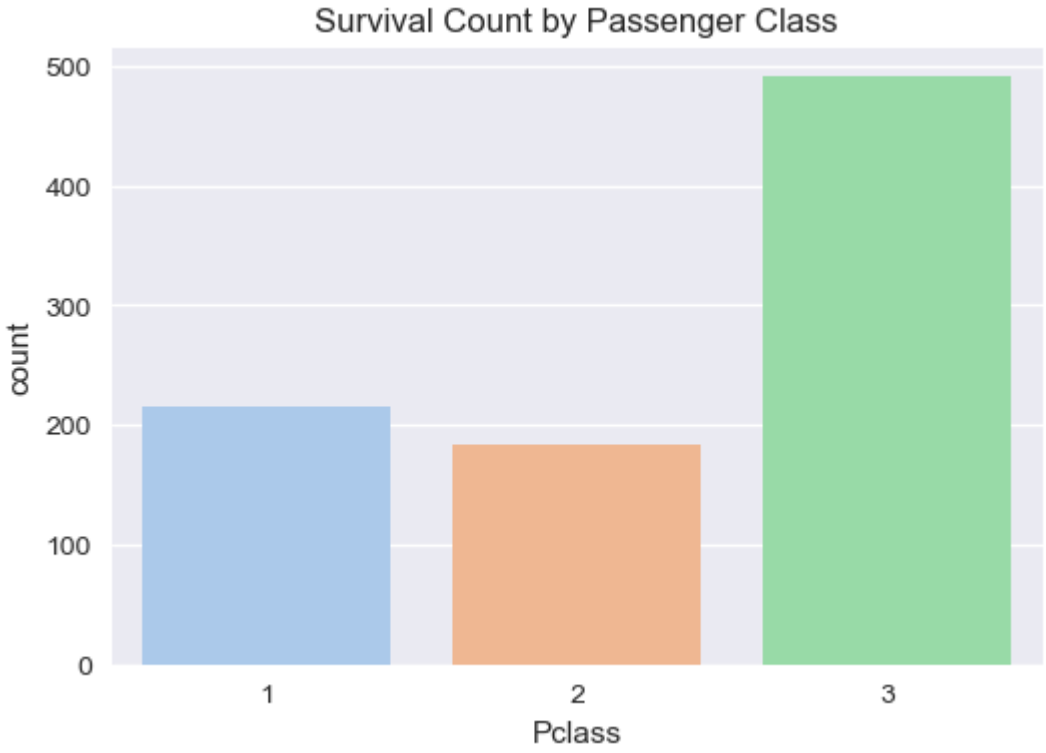
Observation: Most passengers embarked from port S.

In [49]:
```
plt.figure(figsize=(6,4))
sns.countplot(x='Sex', data=df)
plt.title("Survival Count by Sex")
plt.show()
print("Observation: Females had a significantly higher survival rate compared to
```



Observation: Females had a significantly higher survival rate compared to males.

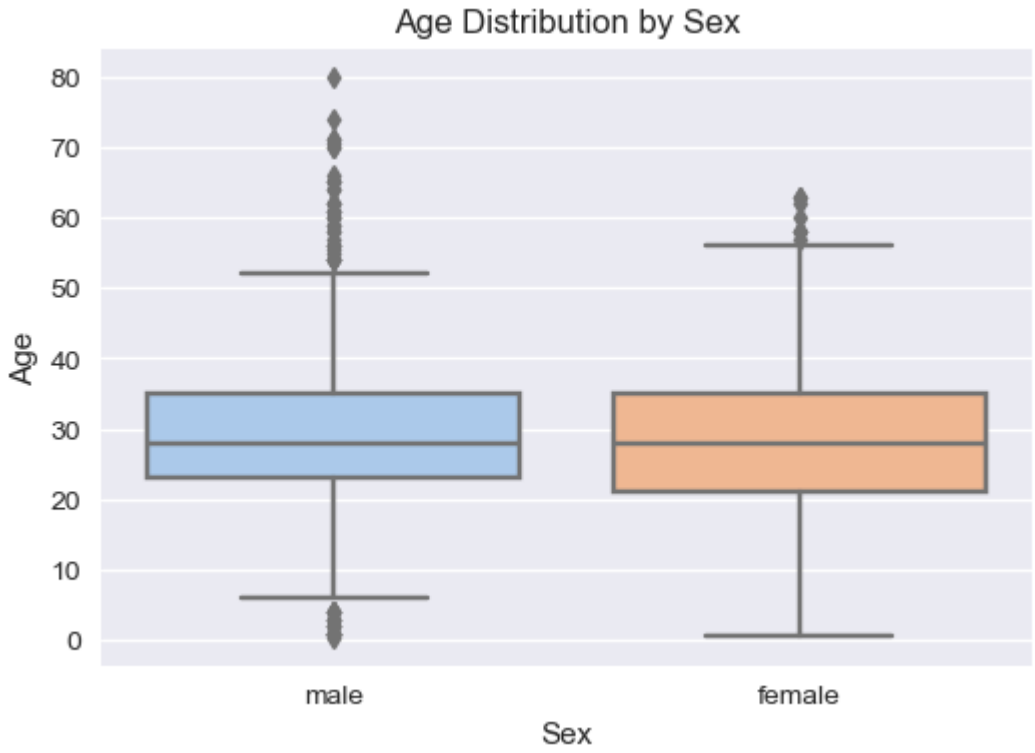
In [50]:
```
plt.figure(figsize=(6,4))
sns.countplot(x='Pclass', data=df)
plt.title("Survival Count by Passenger Class")
plt.show()
print("Observation: First-class passengers had the highest survival rates, while
```



Observation: First-class passengers had the highest survival rates, while third-class had the lowest.

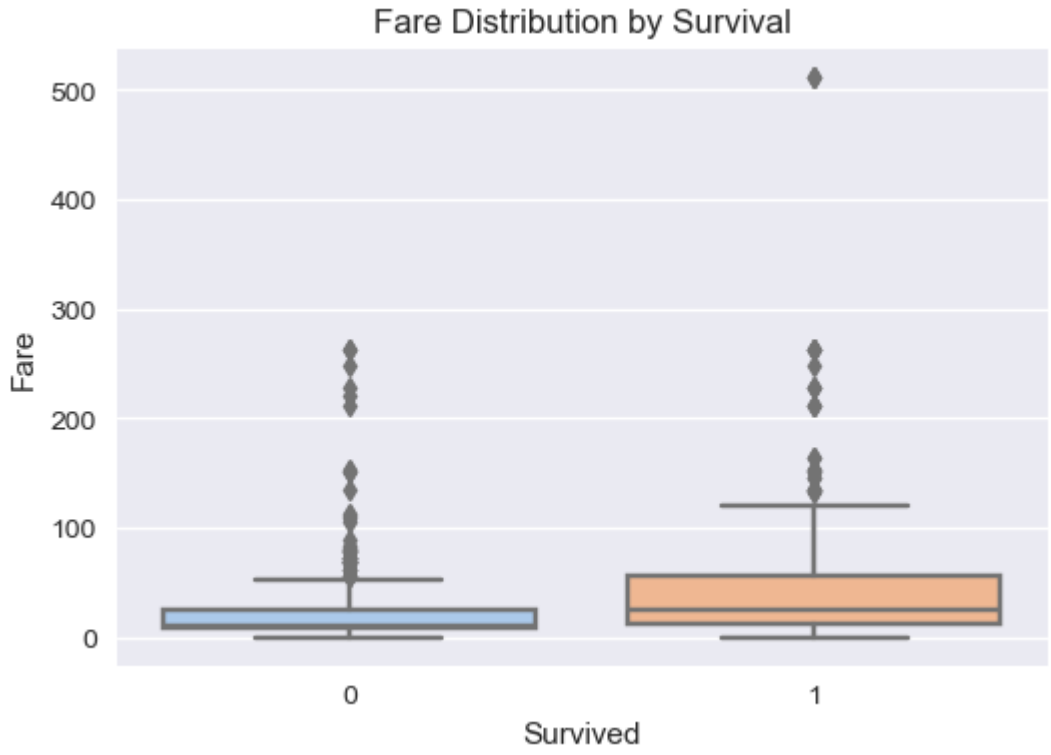
In [51]:
```
plt.figure(figsize=(6,4))
sns.boxplot(x='Sex', y='Age', data=df)
plt.title("Age Distribution by Sex")
```

```
plt.show()
print("Observation: Younger passengers tended to have higher survival rates, but
```



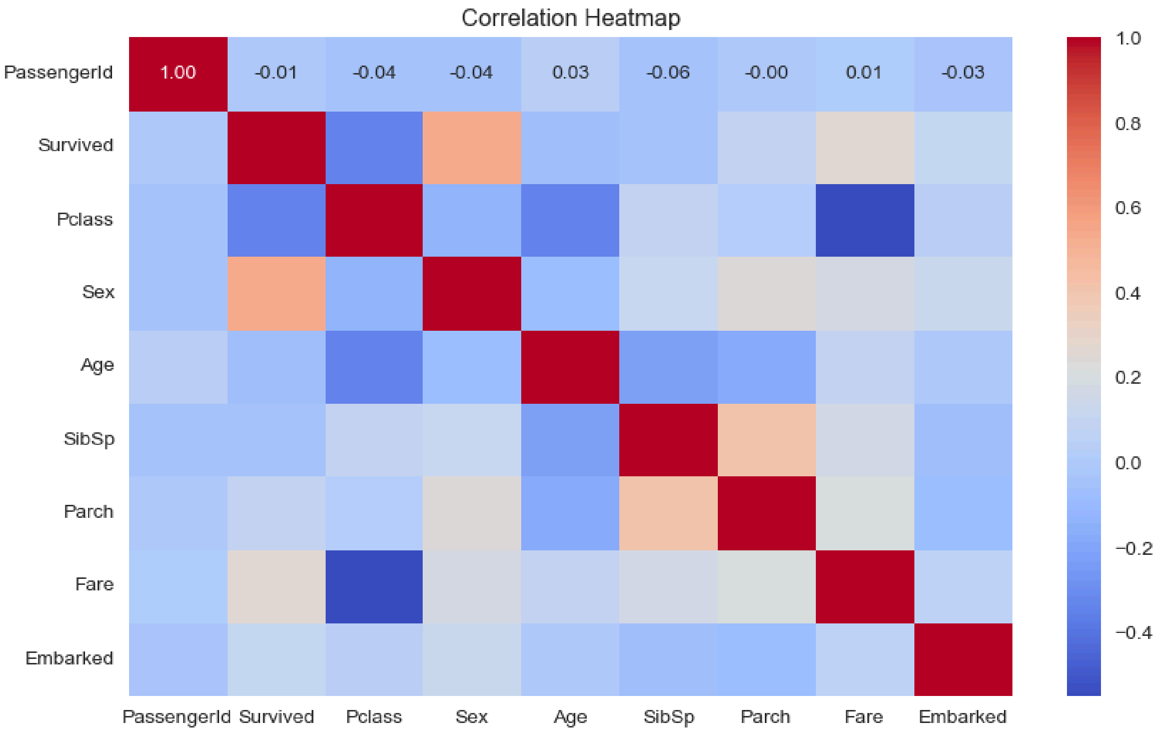

```
Observation: Younger passengers tended to have higher survival rates, but there w
ere survivors in all age groups.
```

In [52]:
```
plt.figure(figsize=(6,4))
sns.boxplot(x='Survived', y='Fare', data=df)
plt.title("Fare Distribution by Survival")
plt.show()
print("Observation: Higher ticket fares are associated with higher survival chan
```



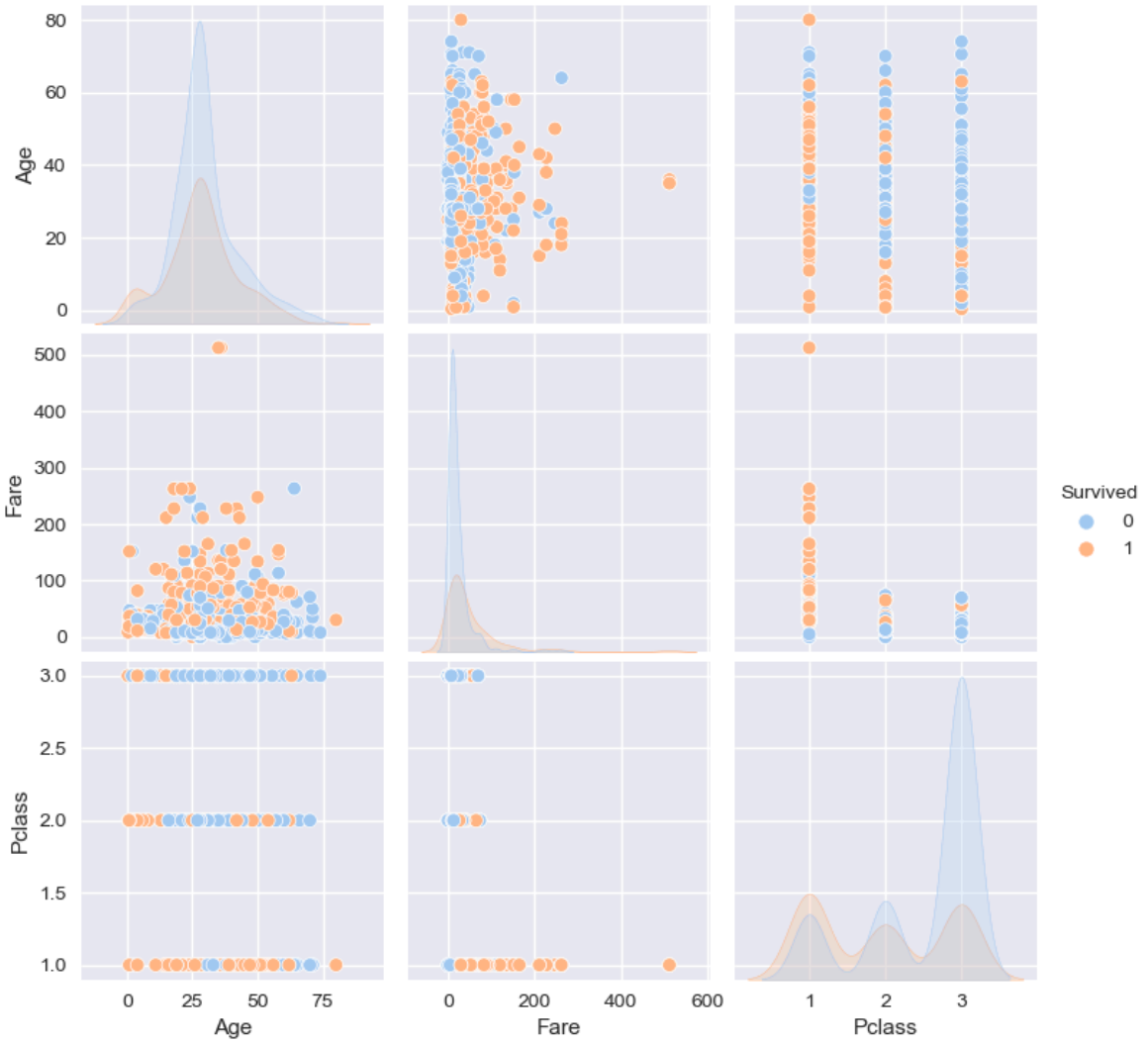

```
Observation: Higher ticket fares are associated with higher survival chances.
```

In [53]:
```
df_corr = df.copy()
df_corr['Sex'] = df_corr['Sex'].map({'male': 0, 'female': 1})
df_corr['Embarked'] = df_corr['Embarked'].map({'S': 0, 'C': 1, 'Q': 2})
plt.figure(figsize=(10,6))
sns.heatmap(df_corr.corr(numeric_only=True), annot=True, cmap='coolwarm', fmt=".
plt.title("Correlation Heatmap")
plt.show()
print("Observation: Strong negative correlation between Pclass and Fare. Surviva
sns.pairplot(df_corr[['Survived', 'Age', 'Fare', 'Pclass']], hue='Survived')
plt.show()
print("Observation: Clear separation in Fare and Pclass between survivors and no
```




```
Observation: Strong negative correlation between Pclass and Fare. Survival is pos
itively correlated with Sex (female) and Fare.
```

```
C:\Users\tedla\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\tedla\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\Users\tedla\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



Observation: Clear separation in Fare and Pclass between survivors and non-survivors. Age shows overlap between both groups.

In [47]:
```
print("""
Summary of Key Insights:
1. Females had a significantly higher survival rate than males.
2. First-class passengers survived more often than those in lower classes.
3. Younger passengers tended to have better survival chances.
4. Higher fares correlated with higher survival probability.
5. Pclass and Fare are negatively correlated.
""")
```

```
Summary of Key Insights:
1. Females had a significantly higher survival rate than males.
2. First-class passengers survived more often than those in lower classes.
3. Younger passengers tended to have better survival chances.
4. Higher fares correlated with higher survival probability.
5. Pclass and Fare are negatively correlated.
```

In [ ]: