Prodigy InfoTech : TASK 2

Titanic Dataset - Data Cleaning & EDA

In [8]:

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

Load the dataset

In [9]:

```
train_df = pd.read_csv("train.csv")
```

Data Cleaning

In [10]:

```
# Clean the dataset

df = train_df.copy()
```

Fill missing 'Age' with median

In [11]:

```
df['Age'] = df['Age'].fillna(df['Age'].median())
```

Fill missing 'Embarked' with mode

In [12]:

```
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
```

In [13]:

```
# Drop 'Cabin' (too many nulls), 'Ticket', and 'Name'
df.drop(columns=['Cabin', 'Ticket', 'Name'], inplace=True)

# Encode categorical variables
df['Sex'] = df['Sex'].map({'male': 0, 'female': 1})
df['Embarked'] = df['Embarked'].map({'S': 0, 'C': 1, 'Q': 2})
```

```
# Create FamilySize feature
df['FamilySize'] = df['SibSp'] + df['Parch']
```

Exploratory Data Analysis (EDA)

In [14]:

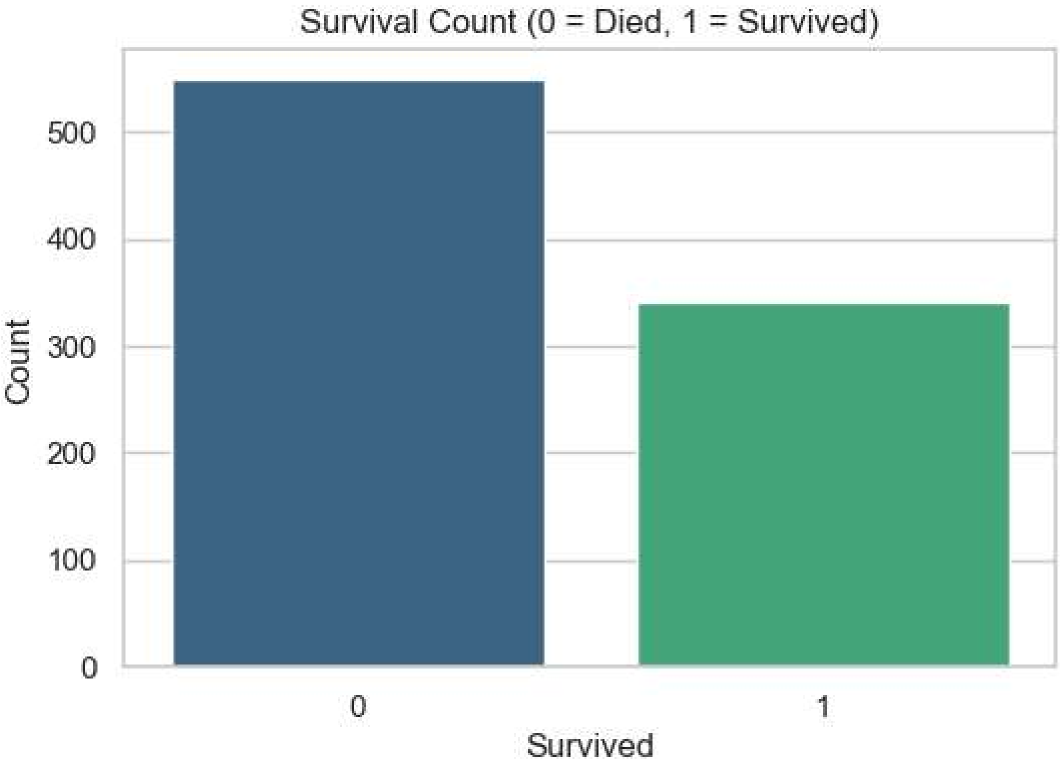
```
sns.set(style="whitegrid")
```

1. Survival Count

In [15]:

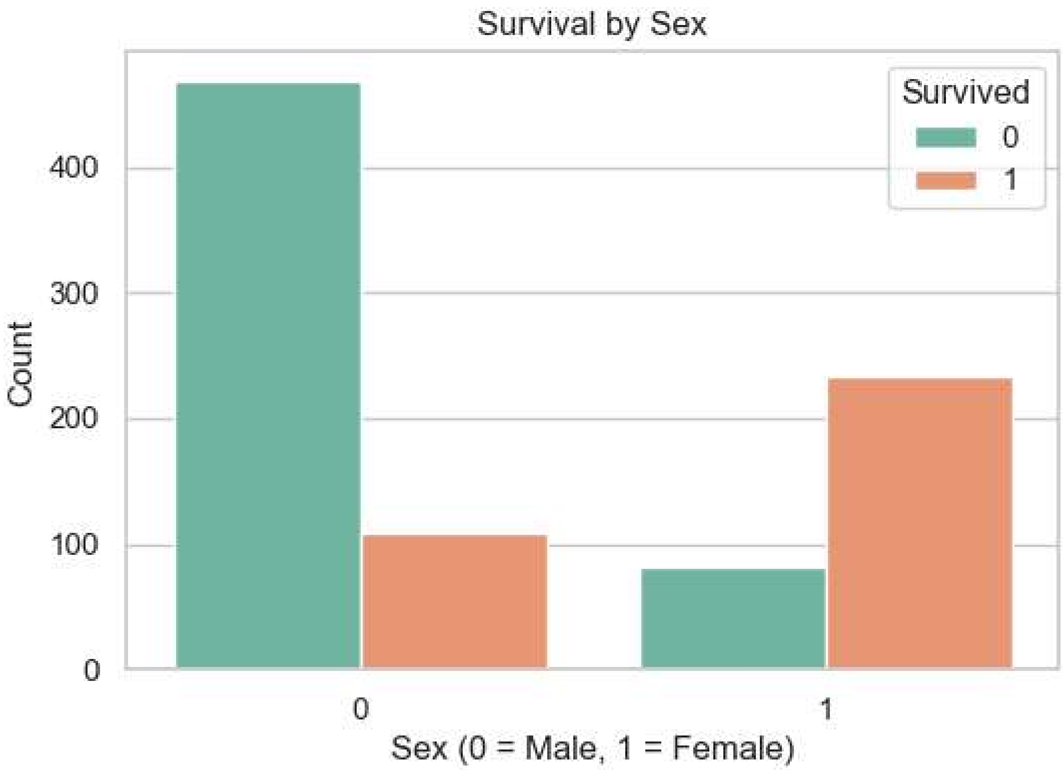
```
plt.figure(figsize=(6,4))
sns.countplot(data=df, x='Survived', palette='viridis')
plt.title('Survival Count (0 = Died, 1 = Survived)')
plt.xlabel('Survived')
plt.ylabel('Count')
plt.show()
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_12060\2306356995.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.1
4.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(data=df, x='Survived', palette='viridis')
```



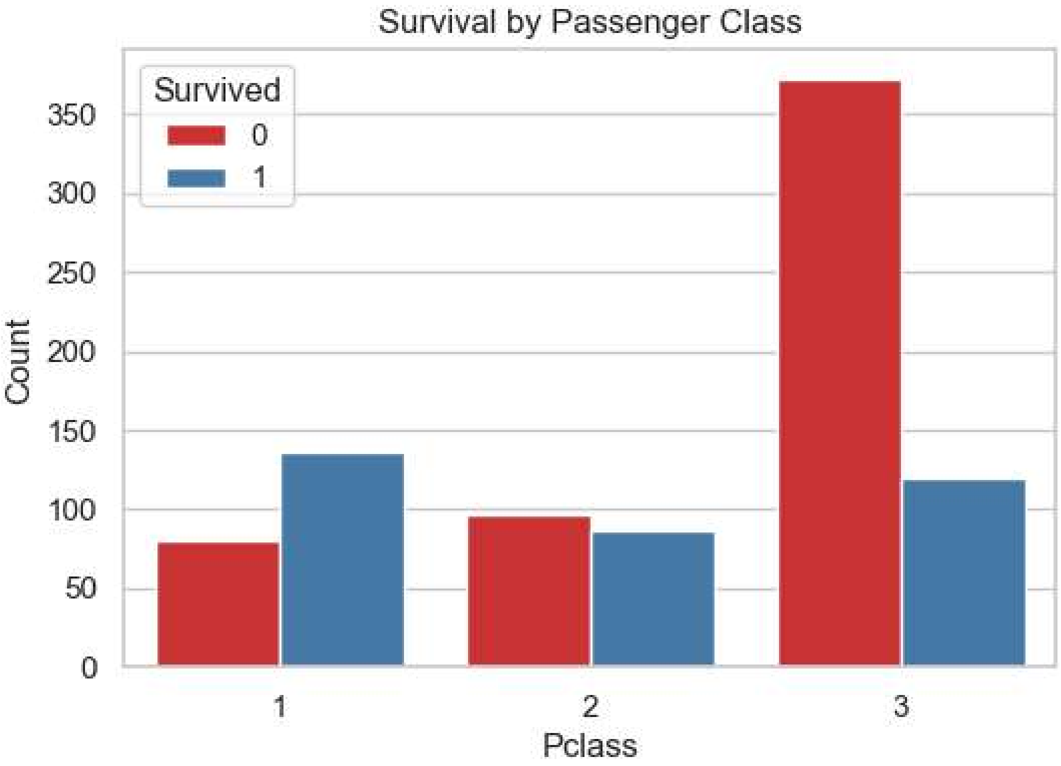
2. Survival by Sex

```
In [16]: plt.figure(figsize=(6,4))
         sns.countplot(data=df, x='Sex', hue='Survived', palette='Set2')
         plt.title('Survival by Sex')
         plt.xlabel('Sex (0 = Male, 1 = Female)')
         plt.ylabel('Count')
         plt.show()
```



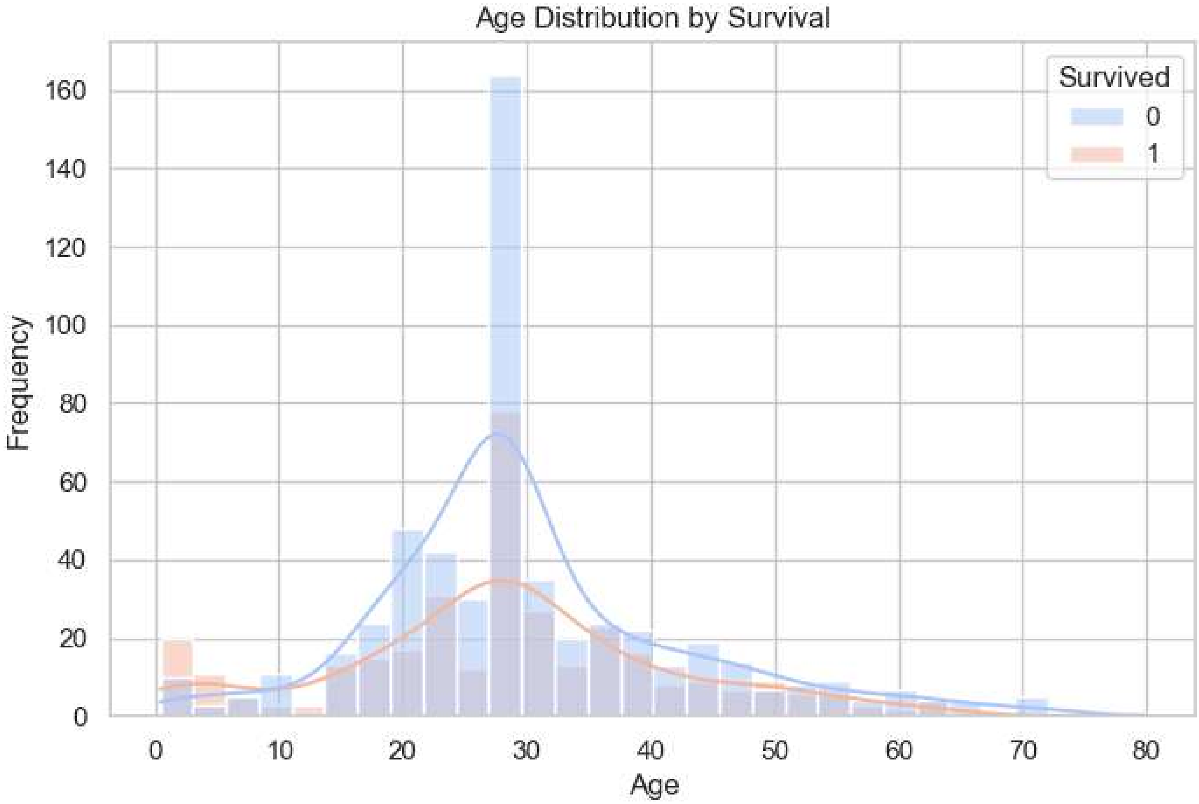
3. Survival by Pclass

In [17]:
```
plt.figure(figsize=(6,4))
sns.countplot(data=df, x='Pclass', hue='Survived', palette='Set1')
plt.title('Survival by Passenger Class')
plt.xlabel('Pclass')
plt.ylabel('Count')
plt.show()
```



4. Age Distribution by Survival

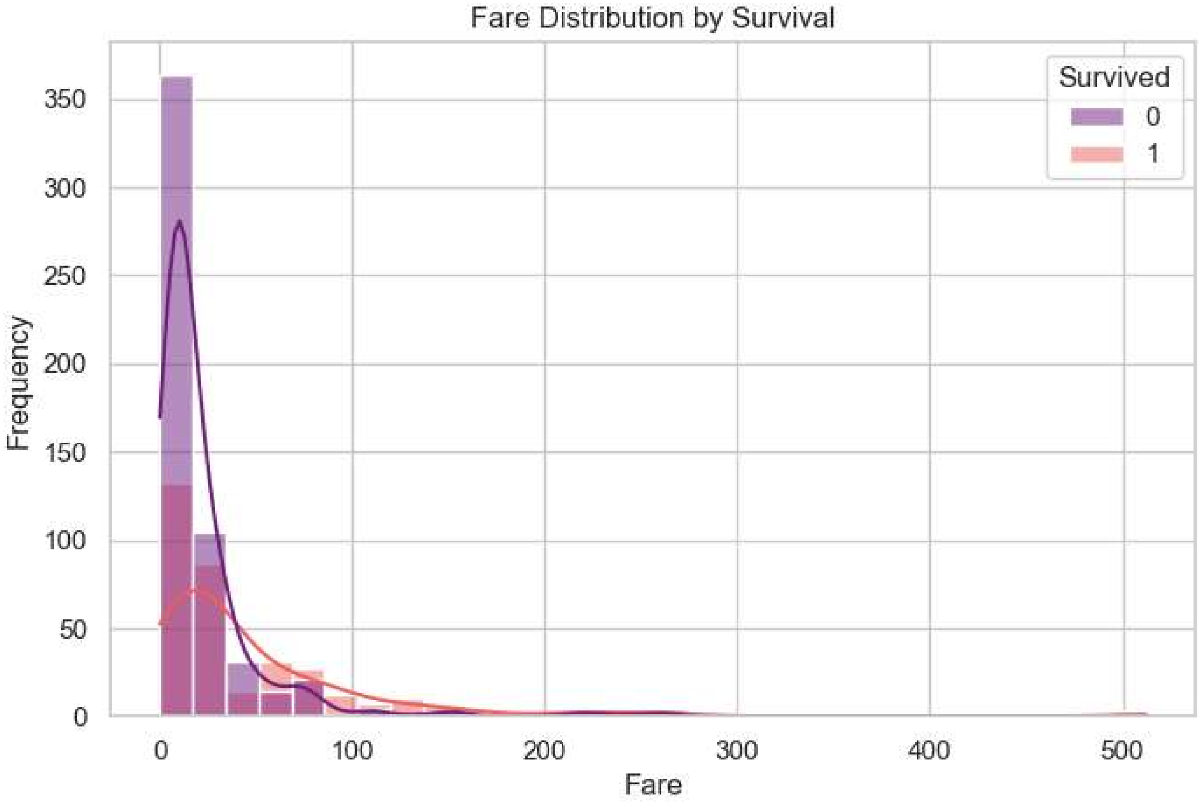
In [18]:
```
plt.figure(figsize=(8,5))
sns.histplot(data=df, x='Age', hue='Survived', kde=True, bins=30, palette='coolwarm
plt.title('Age Distribution by Survival')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



5. Fare Distribution by Survival

In [19]:

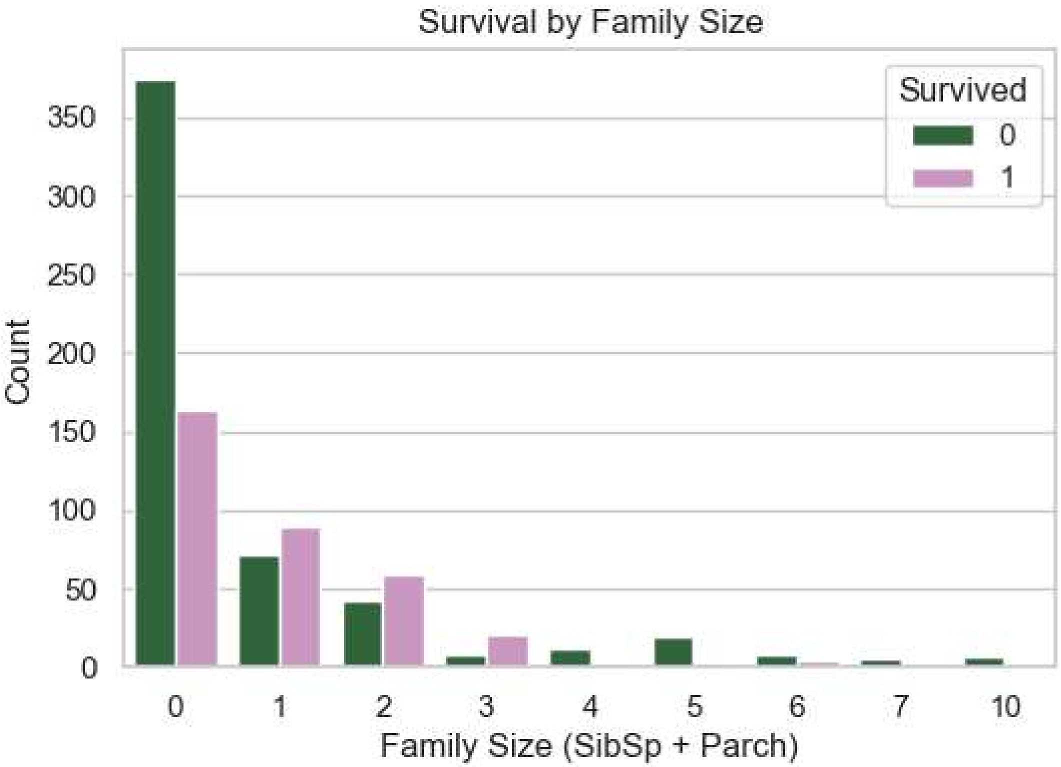
```python
plt.figure(figsize=(8,5))
sns.histplot(data=df, x='Fare', hue='Survived', kde=True, bins=30, palette='magma')
plt.title('Fare Distribution by Survival')
plt.xlabel('Fare')
plt.ylabel('Frequency')
plt.show()
```



6. Family Size vs Survival

In [20]:

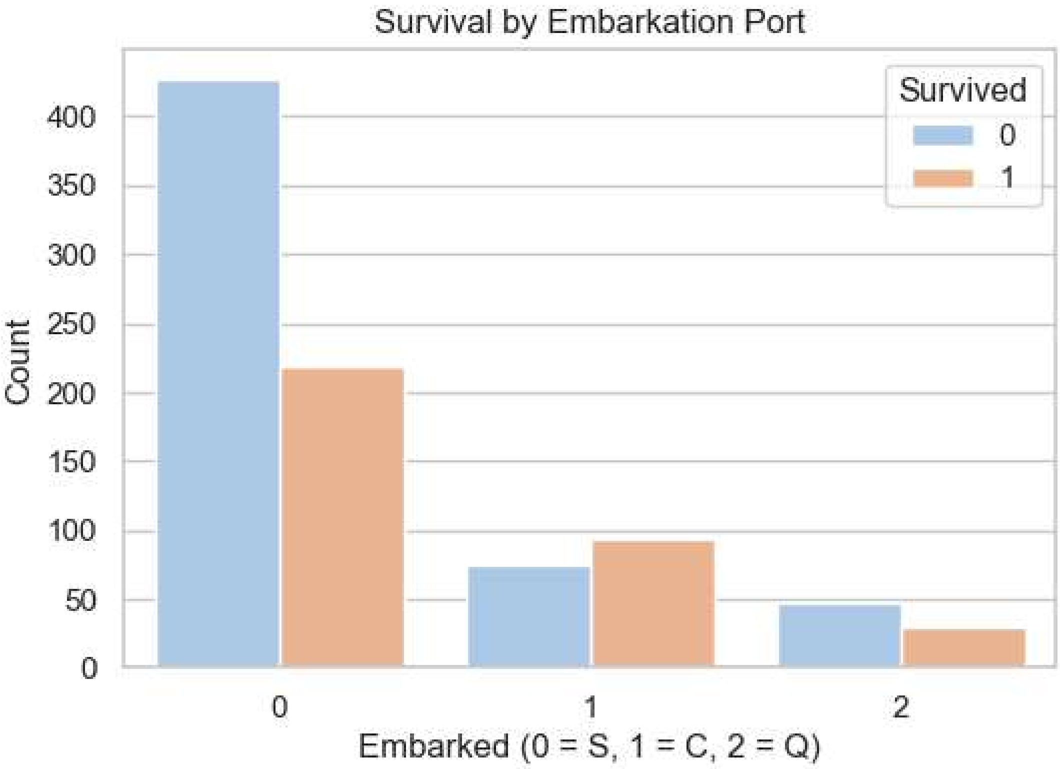
```
plt.figure(figsize=(6,4))
sns.countplot(data=df, x='FamilySize', hue='Survived', palette='cubehelix')
plt.title('Survival by Family Size')
plt.xlabel('Family Size (SibSp + Parch)')
plt.ylabel('Count')
plt.show()
```



7. Survival by Embarkation Port

In [21]:

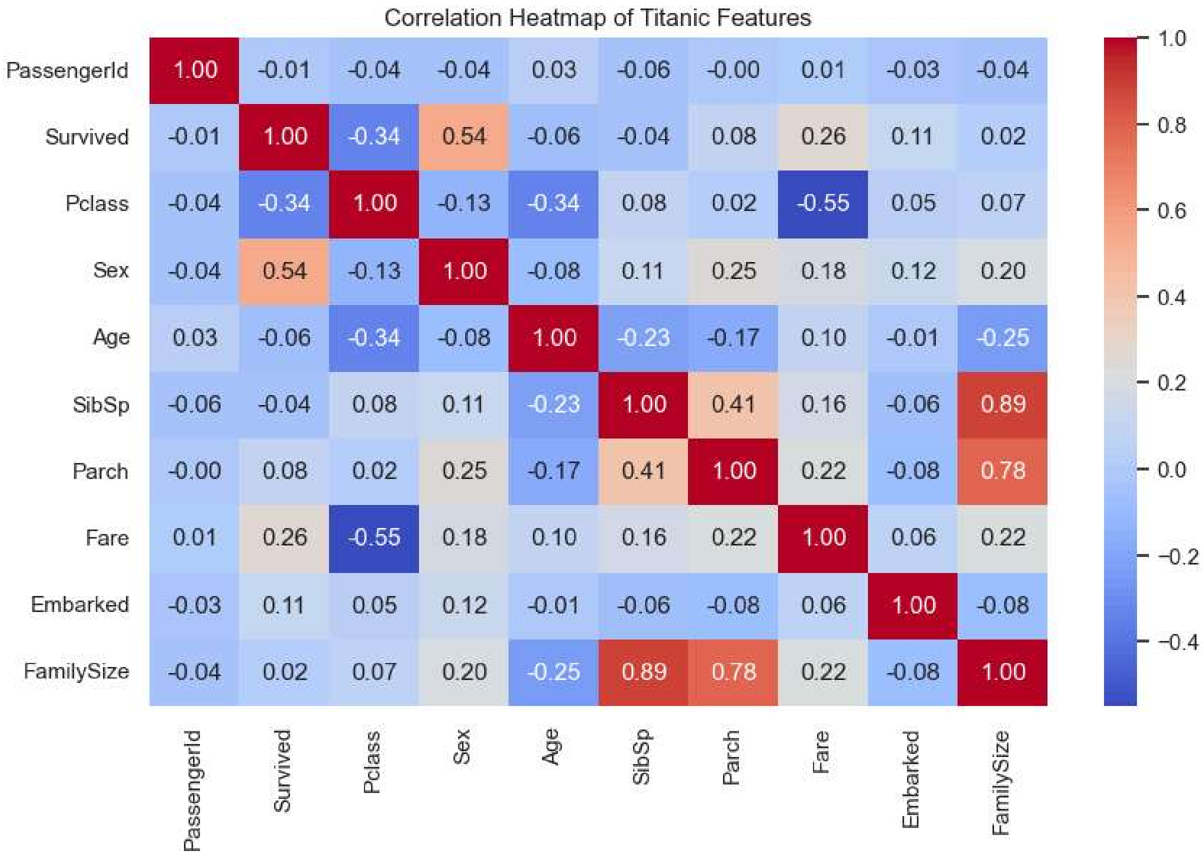
```
plt.figure(figsize=(6,4))
sns.countplot(data=df, x='Embarked', hue='Survived', palette='pastel')
plt.title('Survival by Embarkation Port')
plt.xlabel('Embarked (0 = S, 1 = C, 2 = Q)')
plt.ylabel('Count')
plt.show()
```



8. Correlation Heatmap

In [22]:

```
plt.figure(figsize=(10, 6))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap of Titanic Features')
plt.show()
```


Correlation Heatmap of Titanic Features
PassengerId 1.00 -0.01 -0.04 -0.04 0.03 -0.06 -0.00 0.01 -0.03 -0.04
Survived -0.01 1.00 -0.34 0.54 -0.06 -0.04 0.08 0.26 0.11 0.02
Pclass -0.04 -0.34 1.00 -0.13 -0.34 0.08 0.02 -0.55 0.05 0.07
Sex -0.04 0.54 -0.13 1.00 -0.08 0.11 0.25 0.18 0.12 0.20
Age 0.03 -0.06 -0.34 -0.08 1.00 -0.23 -0.17 0.10 -0.01 -0.25
SibSp -0.06 -0.04 0.08 0.11 -0.23 1.00 0.41 0.16 -0.06 0.89
Parch -0.00 0.08 0.02 0.25 -0.17 0.41 1.00 0.22 -0.08 0.78
Fare 0.01 0.26 -0.55 0.18 0.10 0.16 0.22 1.00 0.06 0.22
Embarked -0.03 0.11 0.05 0.12 -0.01 -0.06 -0.08 0.06 1.00 -0.08
FamilySize -0.04 0.02 0.07 0.20 -0.25 0.89 0.78 0.22 -0.08 1.00
PassengerId
Survived
Pclass
Sex
Age
SibSp
Parch
Fare
Embarked
FamilySize
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4