In [79]:
```
!pip install pandas
import pandas as pd
```

```
Requirement already satisfied: pandas in c:\users\larona\appdata\local\programs\pyth
on\python313\lib\site-packages (2.3.1)
Requirement already satisfied: numpy>=1.26.0 in c:\users\larona\appdata\local\progra
ms\python\python313\lib\site-packages (from pandas) (2.3.2)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\larona\appdata\loc
al\programs\python\python313\lib\site-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in c:\users\larona\appdata\local\program
s\python\python313\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in c:\users\larona\appdata\local\progr
ams\python\python313\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: six>=1.5 in c:\users\larona\appdata\local\programs\py
thon\python313\lib\site-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
```

In [80]:
```
train = pd.read_excel("train.xlsx", engine="openpyxl")
train
```

Out[80]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7

891 rows × 12 columns

In [81]:
```
test = pd.read_excel("test.xlsx", engine="openpyxl")
test
```

Out[81]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	N
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	N
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	N
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	N
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	N
...	...	...	...	...	...	...	...	...	...	
413	1305	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	N
414	1306	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C
415	1307	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	N
416	1308	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	N
417	1309	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	N

418 rows × 11 columns

In [82]:
```
gender = pd.read_excel("gender.xlsx", engine="openpyxl")
gender
```

Out[82]:

	PassengerId	Survived
0	892	0
1	893	1
2	894	0
3	895	0
4	896	1
...	...	...
413	1305	0
414	1306	1
415	1307	0
416	1308	0
417	1309	0

418 rows × 2 columns

In [83]:

```
# Merge test and gender_submission on PassengerId
updatedTest = test.merge(gender, on="PassengerId")

# Reorder columns: Survived right after PassengerId
cols = updatedTest.columns.tolist()
# Remove 'Survived' from the list and insert it at index 1
cols.insert(1, cols.pop(cols.index('Survived')))
updatedTest = updatedTest[cols]

# Verify
print(updatedTest.head())
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	0	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

In [84]:
```
updatedTest
```

Out[84]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	
2	894	0	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	
3	895	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	
4	896	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	1
...	...	...	...	...	...	...	...	...	...	
413	1305	0	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	
414	1306	1	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	10
415	1307	0	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	
416	1308	0	3	Ware, Mr. Frederick	male	NaN	0	0	359309	
417	1309	0	3	Peter, Master. Michael J	male	NaN	1	1	2668	2

418 rows × 12 columns

In [85]:
```
# Concatenate train and test_with_survival vertically
full_data = pd.concat([train, updatedTest], ignore_index=True)

# Verify
print(full_data.shape)  # Should be (1309, number_of_columns)
print(full_data.head())
print(full_data.tail())
```

```
(1309, 12)
   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
      PassengerId  Survived  Pclass                          Name     Sex  \
1304         1305         0       3            Spector, Mr. Woolf    male
1305         1306         1       1  Oliva y Ocana, Dona. Fermina  female
1306         1307         0       3  Saether, Mr. Simon Sivertsen    male
1307         1308         0       3           Ware, Mr. Frederick    male
1308         1309         0       3      Peter, Master. Michael J    male

       Age  SibSp  Parch              Ticket      Fare Cabin Embarked
1304   NaN      0      0           A.5. 3236    8.0500   NaN        S
1305  39.0      0      0            PC 17758  108.9000  C105        C
1306  38.5      0      0  SOTON/O.Q. 3101262    7.2500   NaN        S
1307   NaN      0      0              359309    8.0500   NaN        S
1308   NaN      1      1                2668   22.3583   NaN        C
```

In [86]:
```
full_data
```

Out[86]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450
...	...	...	...	...	...	...	...	...	...
1304	1305	0	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236
1305	1306	1	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758
1306	1307	0	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262
1307	1308	0	3	Ware, Mr. Frederick	male	NaN	0	0	359309
1308	1309	0	3	Peter, Master. Michael J	male	NaN	1	1	2668

1309 rows × 12 columns

In [87]:
```
print("Missing Age values before:", full_data['Age'].isnull().sum())
```

```
Missing Age values before: 263
```

In [88]:
```
median_age = full_data['Age'].median()
print("Median Age:", median_age)
```

```
Median Age: 28.0
```

In [89]:
```
full_data.fillna({'Age': median_age}, inplace=True)
```

In [90]:
```
print("Missing Age values after:", full_data['Age'].isnull().sum())
```

```
Missing Age values after: 0
```

In [91]:
```
print("Missing Age values before:", full_data['Survived'].isnull().sum())
print("Missing Age values before:", full_data['Pclass'].isnull().sum())
print("Missing Age values before:", full_data['Embarked'].isnull().sum())
```

```
Missing Age values before: 0
Missing Age values before: 0
Missing Age values before: 2
```

In [92]:
```
full_data[full_data['Embarked'].isna()]
```

Out[92]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cab
61	62	1	1	Icard, Miss. Amelie	female	38.0	0	0	113572	80.0	B
829	830	1	1	Stone, Mrs. George Nelson (Martha Evelyn)	female	62.0	0	0	113572	80.0	B

In [93]:
```
full_data[(full_data['Pclass'] == 1) & (full_data['Fare'] > 70) & (full_data['Fare'
```

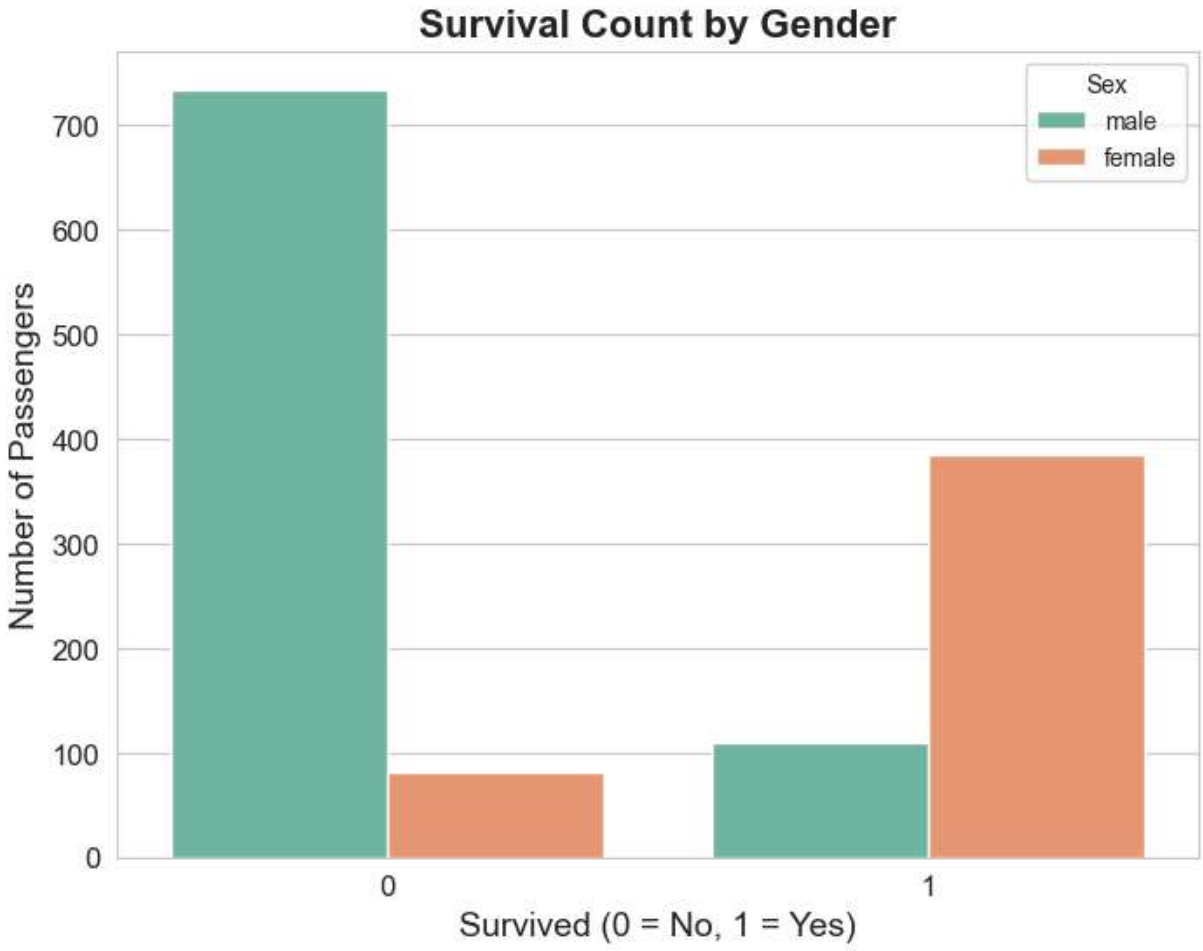
Out[93]:
```
Embarked  Fare
C         79.2000    6
          83.1583    6
S         86.5000    3
C         76.7292    3
S         81.8583    3
          79.6500    3
          78.8500    3
          77.9583    3
          77.2875    2
          83.4750    2
          82.2667    2
C         71.2833    2
S         71.0000    2
C         75.2417    2
          82.1708    2
          78.2667    2
          76.2917    2
          75.2500    2
          89.1042    2
Name: count, dtype: int64
```

In [94]:
```
full_data.loc[full_data['Embarked'].isna(), 'Embarked'] = 'C'
print("Missing Age values after:", full_data['Embarked'].isnull().sum())
```

```
Missing Age values after: 0
```

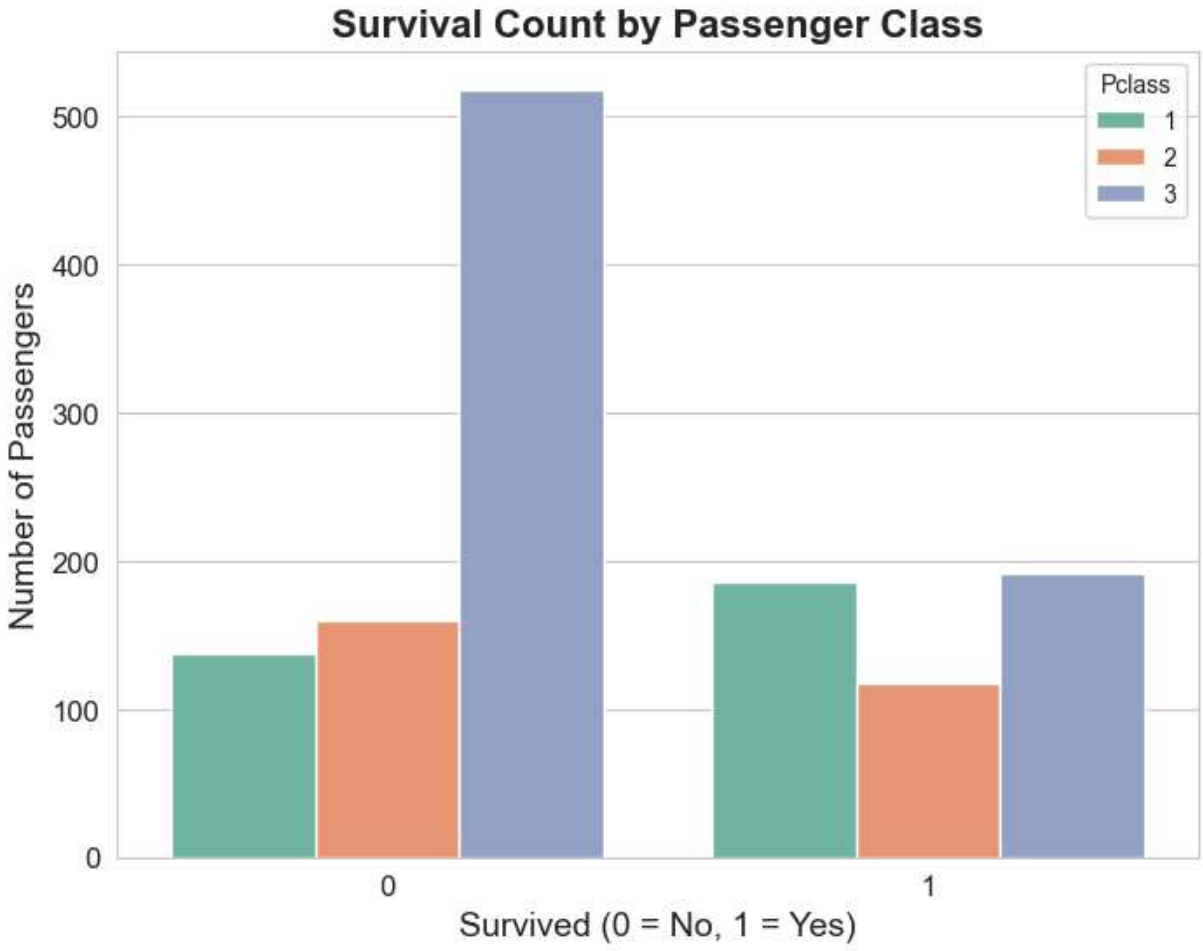
To see which gender between the males and females had more survivors

In [95]:
```
import matplotlib.pyplot as plt
import seaborn as sns

# Set the style for better visuals
sns.set_style("whitegrid")

# Create the countplot
plt.figure(figsize=(8, 6))
sns.countplot(x='Survived', hue='Sex', data=full_data, palette='Set2')

# Add labels and title
plt.title('Survival Count by Gender', fontsize=16, fontweight='bold')
plt.xlabel('Survived (0 = No, 1 = Yes)', fontsize=14)
plt.ylabel('Number of Passengers', fontsize=14)

# Increase font size for x and y ticks
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)

# Show the chart
plt.savefig('Survival_Count_by_Gender.png', dpi=300, bbox_inches='tight')
plt.show()
```



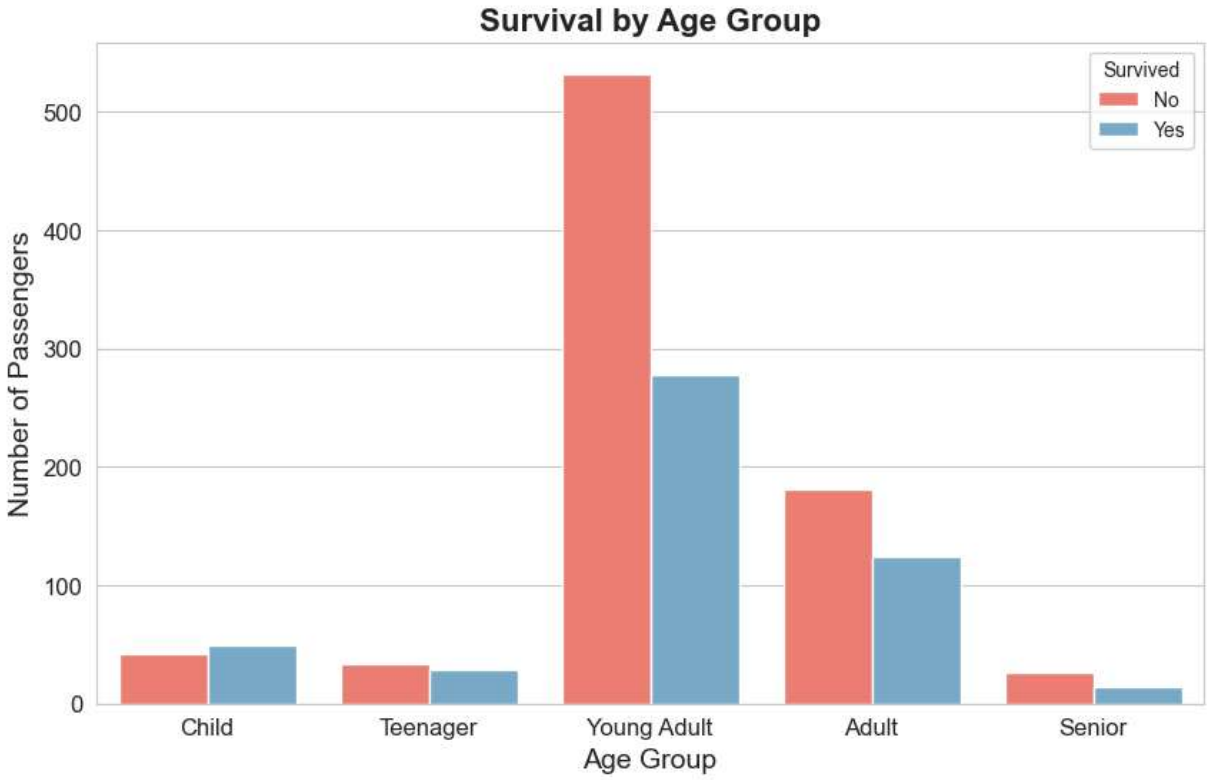
To see which passengers who registered for which class had the most survivors

In [96]:
```
# Set the style for better visuals
sns.set_style("whitegrid")

# Create the countplot
plt.figure(figsize=(8, 6))
sns.countplot(x='Survived', hue='Pclass', data=full_data, palette='Set2')

# Add labels and title
plt.title('Survival Count by Passenger Class', fontsize=16, fontweight='bold')
plt.xlabel('Survived (0 = No, 1 = Yes)', fontsize=14)
plt.ylabel('Number of Passengers', fontsize=14)

# Increase font size for x and y ticks
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)

# Show the chart
plt.savefig('Survival_Count_by_Class.png', dpi=300, bbox_inches='tight')
plt.show()
```



Then showing which age range had the most survivor

In [97]:
```
# Define age bins and labels
bins = [0, 12, 18, 35, 60, 100]  #child: 0-11 teenager: 12-17  young adult: 18-35 A
labels = ['Child', 'Teenager', 'Young Adult', 'Adult', 'Senior']

# Plot countplot with age bins directly
plt.figure(figsize=(10, 6))
sns.countplot(x=pd.cut(full_data['Age'], bins=bins, labels=labels, right=False), hu

# Add labels and title
plt.title('Survival by Age Group', fontsize=16, fontweight='bold')
plt.xlabel('Age Group', fontsize=14)
plt.ylabel('Number of Passengers', fontsize=14)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.legend(title='Survived', labels=['No', 'Yes'])

plt.savefig('Survival_Count_by_Age_Range', dpi=300, bbox_inches='tight')
plt.show()
```



In [98]:
```
print("Missing Age values before:", full_data['Fare'].isnull().sum())
```

```
Missing Age values before: 1
```

In [99]:
```
# Find the row with missing Fare
missing_idx = full_data['Fare'].isna()

# Fill with median fare for same Pclass and Embarked
full_data.loc[missing_idx, 'Fare'] = full_data.groupby(['Pclass', 'Embarked'])['Far
```

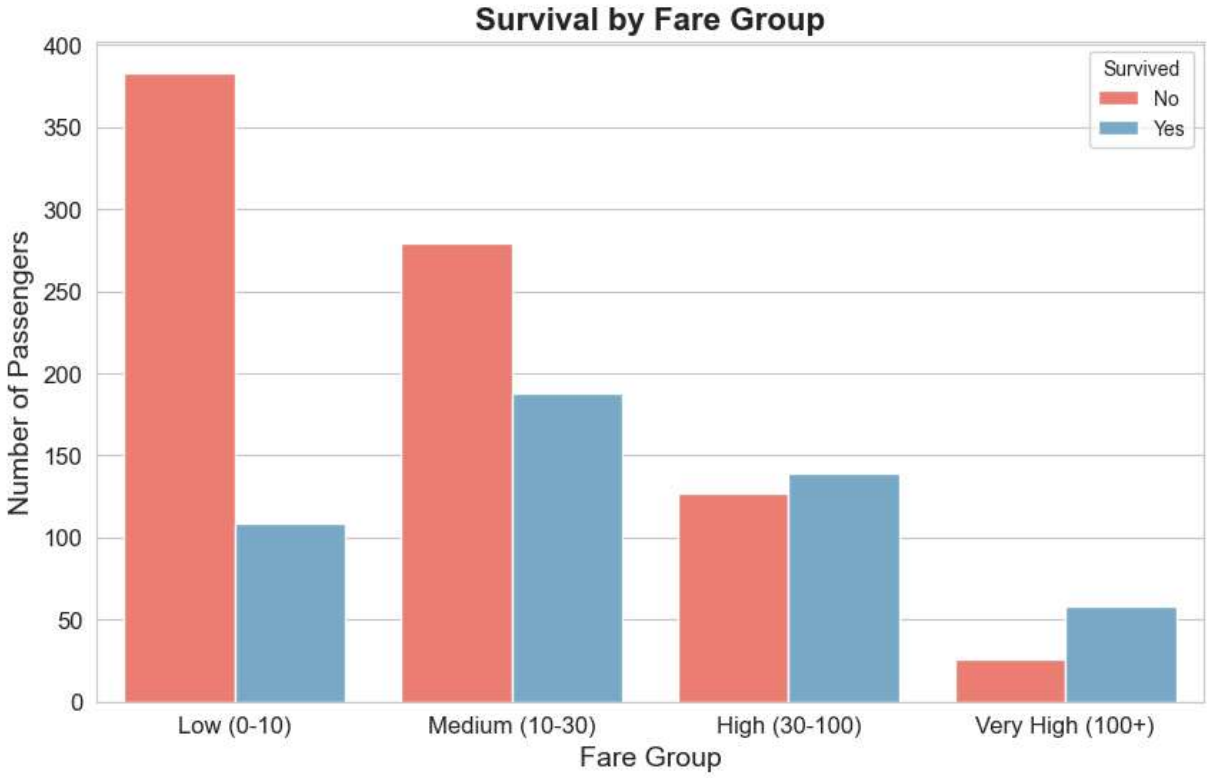
In [100…
```
full_data['Fare'].isna().sum()  # Should return 0
```

Out[100…
```
np.int64(0)
```

Survival by Fare Group in the train dataset

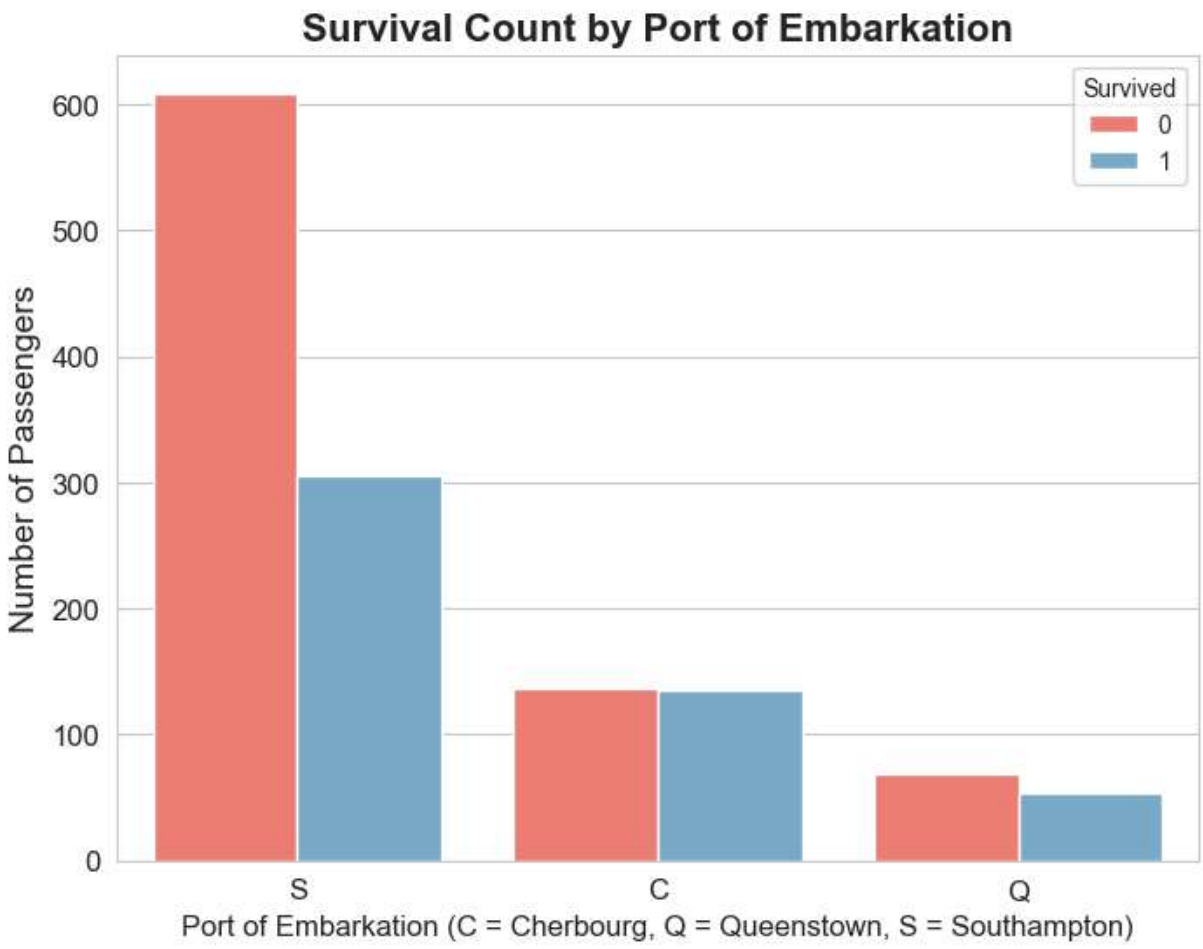
In [101…
```
# Define fare bins and labels
fare_bins = [0, 10, 30, 100, 600]
fare_labels = ['Low (0-10)', 'Medium (10-30)', 'High (30-100)', 'Very High (100+)']

# Plot countplot with fare bins
plt.figure(figsize=(10, 6))
sns.countplot(x=pd.cut(full_data['Fare'], bins=fare_bins, labels=fare_labels, right

# Add labels and title
plt.title('Survival by Fare Group', fontsize=16, fontweight='bold')
plt.xlabel('Fare Group', fontsize=14)
plt.ylabel('Number of Passengers', fontsize=14)
plt.xticks(fontsize=12)
```

```
plt.yticks(fontsize=12)
plt.legend(title='Survived', labels=['No', 'Yes'])

plt.savefig('Survival_Count_by_Fare_Group.png', dpi=300, bbox_inches='tight')
plt.show()
```



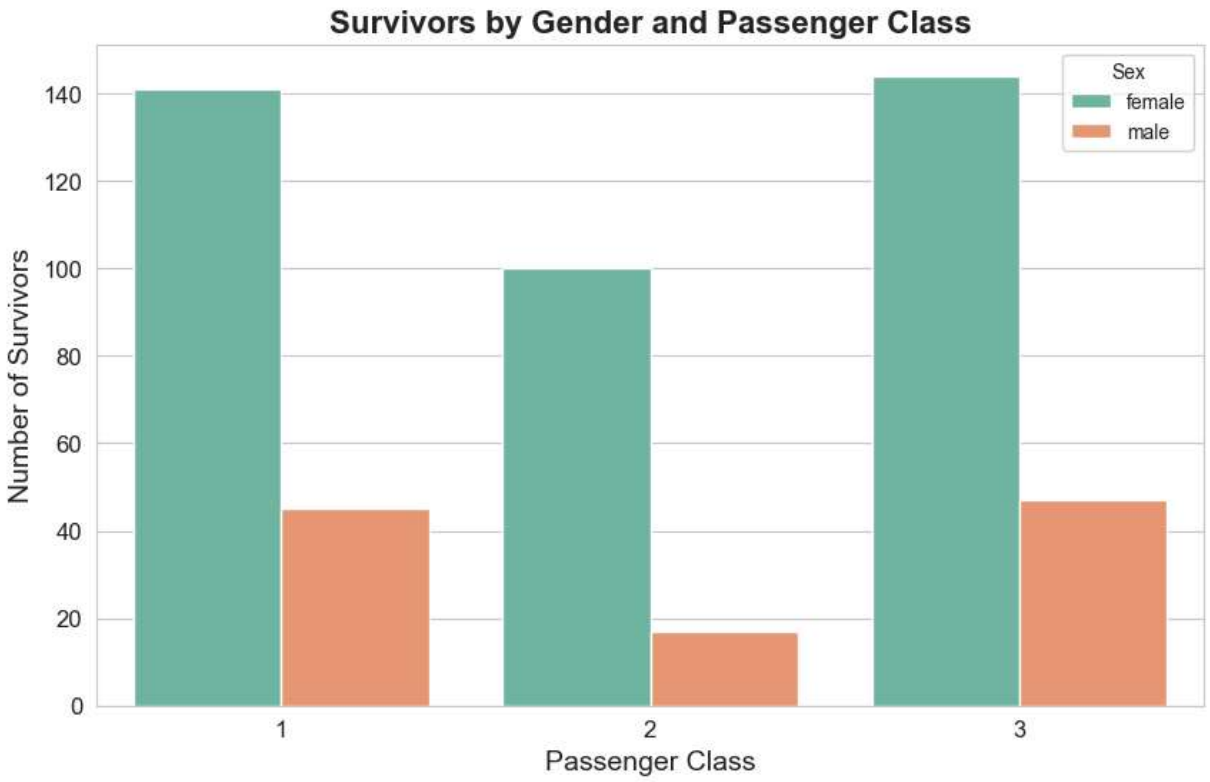
In [102…

```
# Set the style for better visuals
sns.set_style("whitegrid")

# Create the countplot (Embarked vs Survival)
plt.figure(figsize=(8, 6))
sns.countplot(x='Embarked', hue='Survived', data=full_data, palette=['#FF6F61','#6B

# Add labels and title
plt.title('Survival Count by Port of Embarkation', fontsize=16, fontweight='bold')
plt.xlabel('Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)',
plt.ylabel('Number of Passengers', fontsize=14)

# Increase font size for x and y ticks
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)

# Show the chart
plt.savefig('Survival_Count_by_Port_of_Embarkation.png', dpi=300, bbox_inches='tigh
plt.show()
```



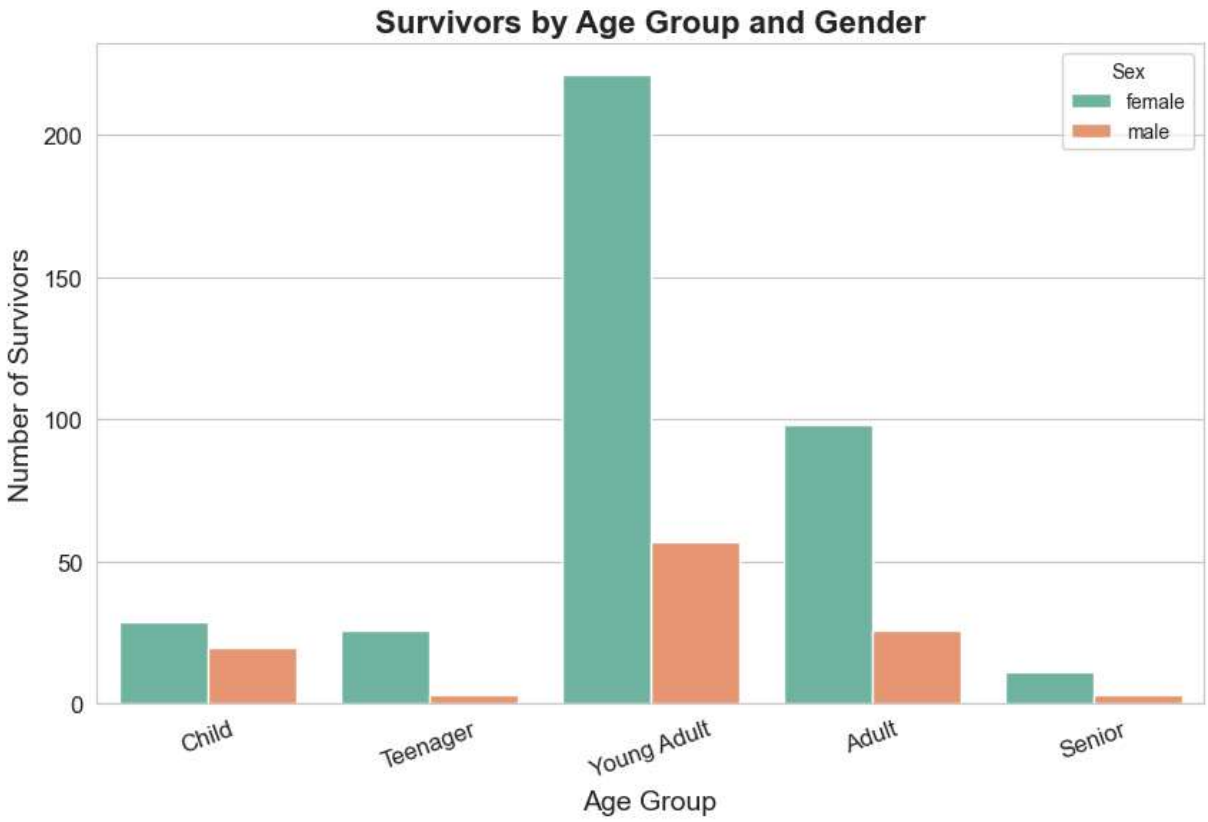
In [103…

```
# Set the style for better visuals
sns.set_style("whitegrid")

# Create the countplot (Sex vs Pclass, split by Survival)
plt.figure(figsize=(10, 6))
sns.countplot(x='Pclass', hue='Sex', data=full_data[full_data['Survived']==1], pale

# Add labels and title
plt.title('Survivors by Gender and Passenger Class', fontsize=16, fontweight='bold'
plt.xlabel('Passenger Class', fontsize=14)
plt.ylabel('Number of Survivors', fontsize=14)

# Increase font size for x and y ticks
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)

plt.savefig('Survival_Count_by_Age_and_Passenger_Class.png', dpi=300, bbox_inches='
plt.show()
```



In [104…

```
# Define age bins and labels
bins = [0, 12, 18, 35, 60, 100]
labels = ['Child', 'Teenager', 'Young Adult', 'Adult', 'Senior']

# Add AgeGroup column
full_data['AgeGroup'] = pd.cut(full_data['Age'], bins=bins, labels=labels, right=Fa

# Plot survivors by AgeGroup and Sex
plt.figure(figsize=(10, 6))
sns.countplot(x='AgeGroup', hue='Sex', data=full_data[full_data['Survived']==1], pa

# Add labels and title
plt.title('Survivors by Age Group and Gender', fontsize=16, fontweight='bold')
plt.xlabel('Age Group', fontsize=14)
plt.ylabel('Number of Survivors', fontsize=14)

# Rotate x labels for clarity
plt.xticks(rotation=20, fontsize=12)
plt.yticks(fontsize=12)

plt.savefig('Survival_Count_by_Age_Group_and_Gender.png', dpi=300, bbox_inches='tig
plt.show()
```


Survivors by Age Group and Gender
Sex
female
male
Number of Survivors
0
50
100
150
200
Child
Teenager
Young Adult
Adult
Senior
Age Group

In [ ]: