In [1]:
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```python
sns.set(style="whitegrid")
plt.style.use("ggplot")
```

In [3]:
```python
train_df = pd.read_csv("train.csv")
test_df = pd.read_csv("test.csv")
gender_df = pd.read_csv("gender_submission.csv")
```

In [5]:
```python
print("Dataset Info:")
print(train_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
None
```

In [6]:
```python
print("Dataset Info:")
print(test_df.info())
```

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  418 non-null    int64
 1   Pclass       418 non-null    int64
 2   Name         418 non-null    object
 3   Sex          418 non-null    object
 4   Age          332 non-null    float64
 5   SibSp        418 non-null    int64
 6   Parch        418 non-null    int64
 7   Ticket       418 non-null    object
 8   Fare         417 non-null    float64
 9   Cabin        91 non-null     object
 10  Embarked     418 non-null    object
dtypes: float64(2), int64(4), object(5)
memory usage: 36.0+ KB
None
```

In [7]:
```python
print("Dataset Info:")
print(gender_df.info())
```

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  418 non-null    int64
 1   Survived     418 non-null    int64
dtypes: int64(2)
memory usage: 6.7 KB
None
```

In [8]:
```python
print("\nDataset Description:")
print(train_df.describe())
```

```
Dataset Description:
       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
```

In [9]:
```python
print("\nDataset Description:")
print(test_df.describe())
```

```
Dataset Description:
       PassengerId      Pclass         Age       SibSp       Parch        Fare
count   418.000000  418.000000  332.000000  418.000000  418.000000  417.000000
mean   1100.500000    2.265550   30.272590    0.447368    0.392344   35.627188
std     120.810458    0.841838   14.181209    0.896760    0.981429   55.907576
min     892.000000    1.000000    0.170000    0.000000    0.000000    0.000000
25%     996.250000    1.000000   21.000000    0.000000    0.000000    7.895800
50%    1100.500000    3.000000   27.000000    0.000000    0.000000   14.454200
75%    1204.750000    3.000000   39.000000    1.000000    0.000000   31.500000
max    1309.000000    3.000000   76.000000    8.000000    9.000000  512.329200
```

In [10]:
```python
print("\nDataset Description:")
print(gender_df.describe())
```

```
Dataset Description:
       PassengerId    Survived
count   418.000000  418.000000
mean   1100.500000    0.363636
std     120.810458    0.481622
min     892.000000    0.000000
25%     996.250000    0.000000
50%    1100.500000    0.000000
75%    1204.750000    1.000000
max    1309.000000    1.000000
```

In [11]:
```python
print("\nFirst 5 Rows:")
print(train_df.head())
```

```
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 [12]:
```python
print("\nFirst 5 Rows:")
print(test_df.head())
```

```
First 5 Rows:
   PassengerId  Pclass                                          Name     Sex  \
0          892       3                              Kelly, Mr. James    male
1          893       3              Wilkes, Mrs. James (Ellen Needs)  female
2          894       2                     Myles, Mr. Thomas Francis    male
3          895       3                              Wirz, Mr. Albert    male
4          896       3  Hirvonen, Mrs. Alexander (Helga E Lindqvist)  female

    Age  SibSp  Parch   Ticket     Fare Cabin Embarked
0  34.5      0      0   330911   7.8292   NaN        Q
1  47.0      1      0   363272   7.0000   NaN        S
2  62.0      0      0   240276   9.6875   NaN        Q
3  27.0      0      0   315154   8.6625   NaN        S
4  22.0      1      1  3101298  12.2875   NaN        S
```

In [13]:
```python
print("\nFirst 5 Rows:")
print(gender_df.head())
```

```
First 5 Rows:
   PassengerId  Survived
0          892         0
1          893         1
2          894         0
3          895         0
4          896         1
```

In [14]:
```python
print("\nMissing Values:")
print(train_df.isnull().sum())
```

```
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 [15]:
```python
print("\nMissing Values:")
print(test_df.isnull().sum())
```

```
Missing Values:
PassengerId      0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

In [16]:
```python
print("\nMissing Values:")
print(gender_df.isnull().sum())
```

```
Missing Values:
PassengerId    0
Survived       0
dtype: int64
```

In [17]:
```python
print(train_df['Sex'].value_counts())
```

```
male      577
female    314
Name: Sex, dtype: int64
```

In [18]:
```python
print(test_df['Sex'].value_counts())
```

```
male      266
female    152
Name: Sex, dtype: int64
```

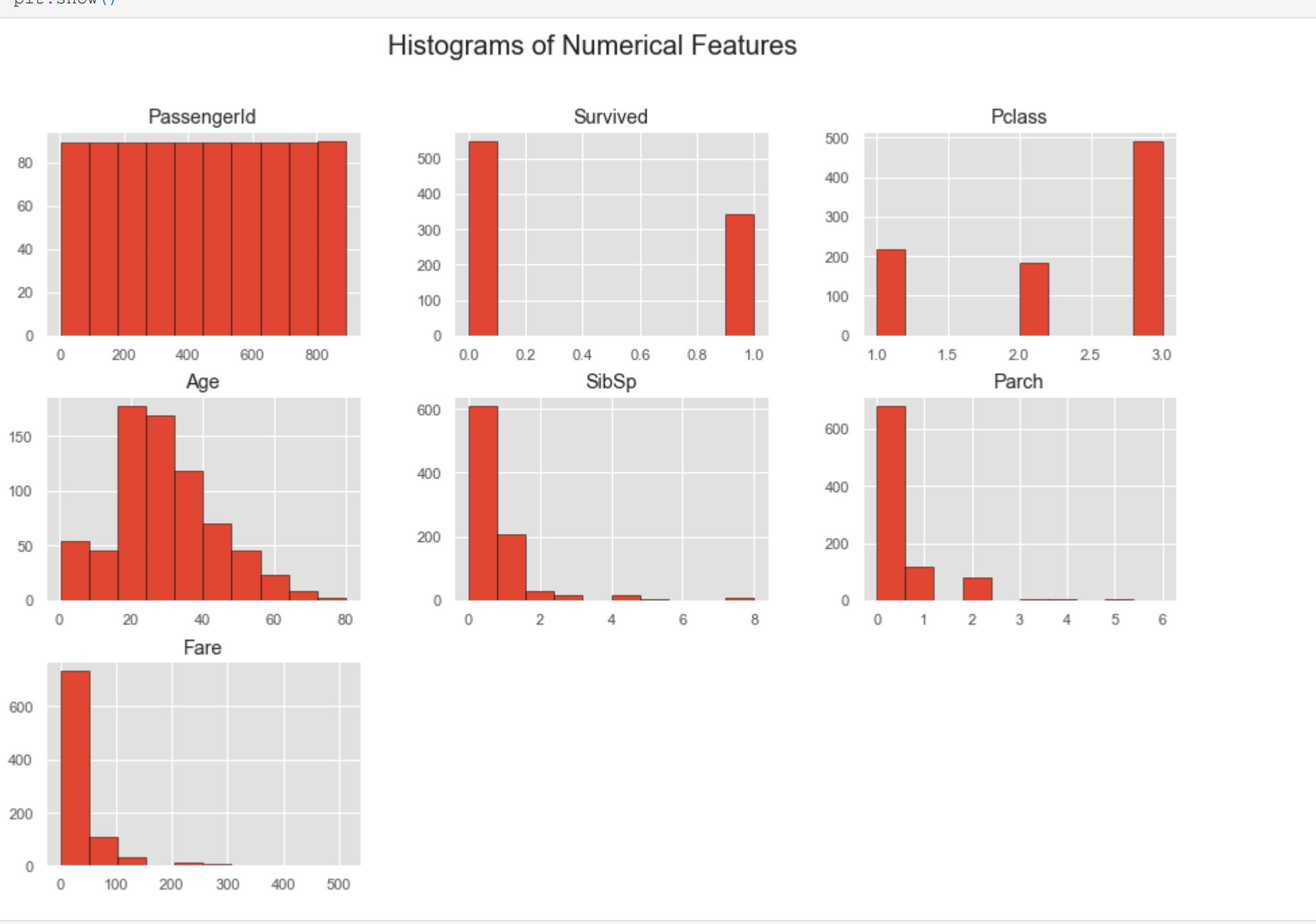
In [19]:
```python
print(train_df['Pclass'].value_counts())
```

```
3    491
1    216
2    184
Name: Pclass, dtype: int64
```

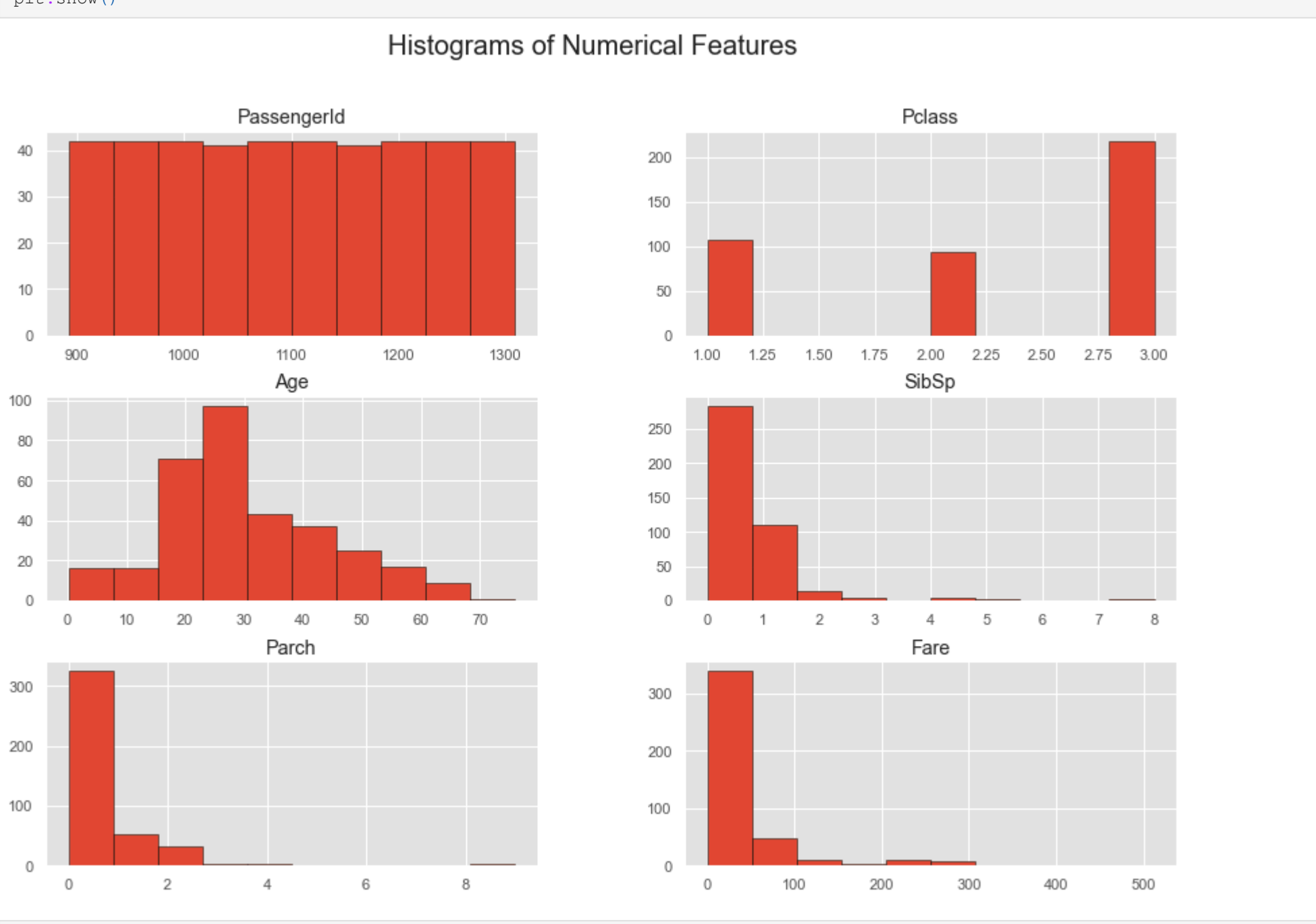
In [20]:
```python
print(test_df['Pclass'].value_counts())
```

```
3    218
1    107
2     93
Name: Pclass, dtype: int64
```

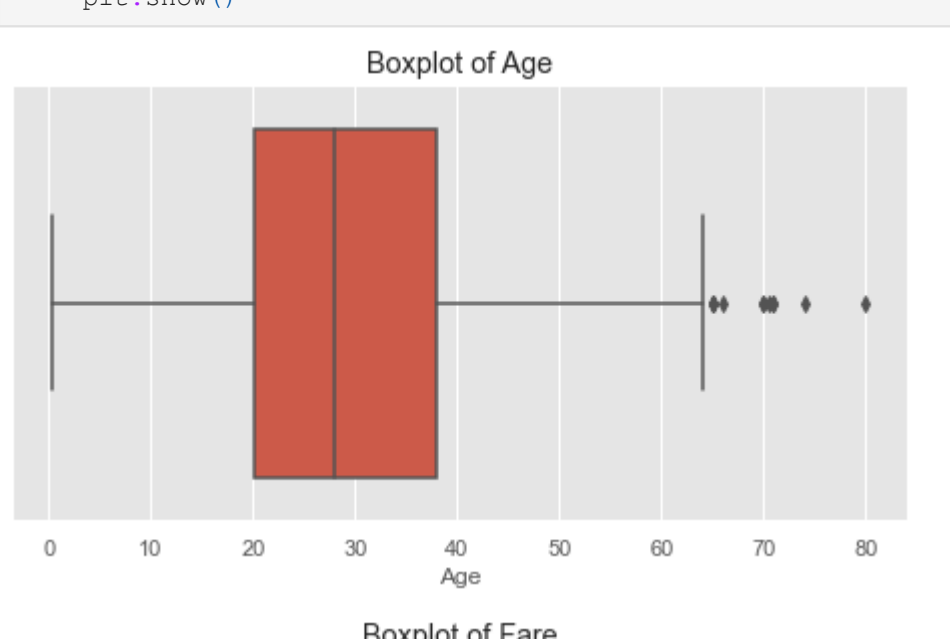
In [22]:
```python
train_df.hist(figsize=(15, 10), edgecolor='black')
plt.suptitle('Histograms of Numerical Features', fontsize=20)
plt.show()
```

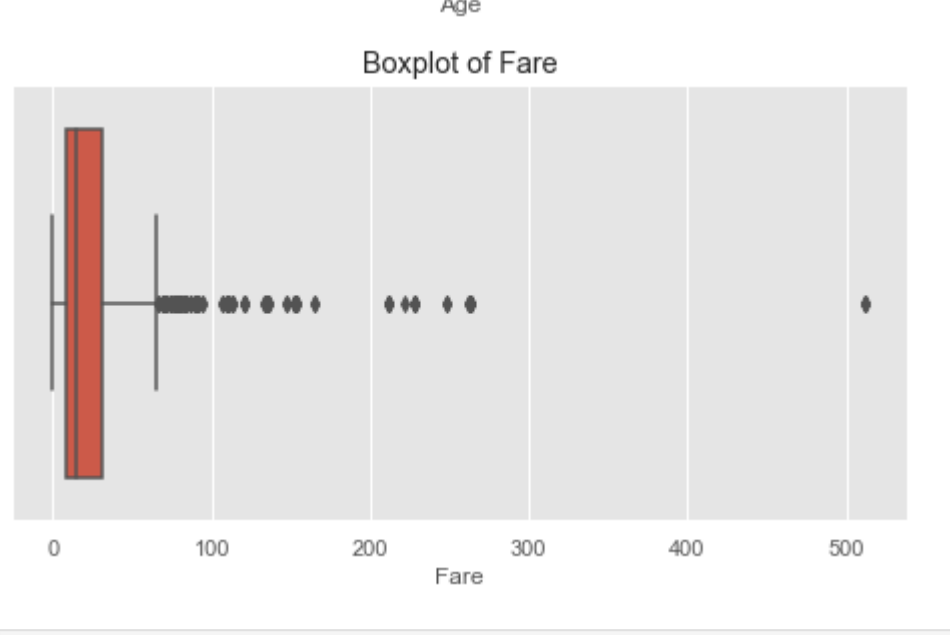


In [23]:
```python
test_df.hist(figsize=(15, 10), edgecolor='black')
plt.suptitle('Histograms of Numerical Features', fontsize=20)
plt.show()
```

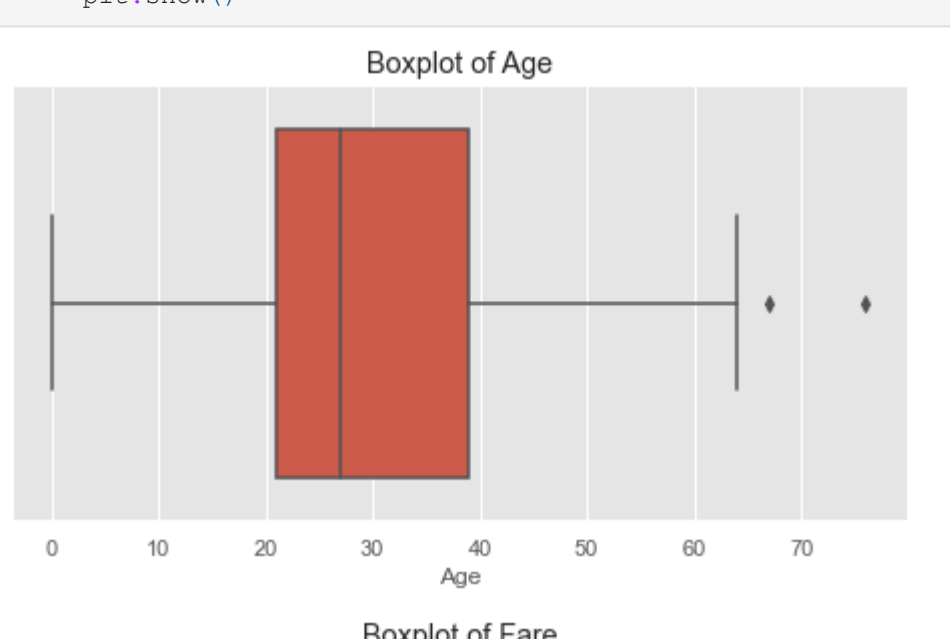


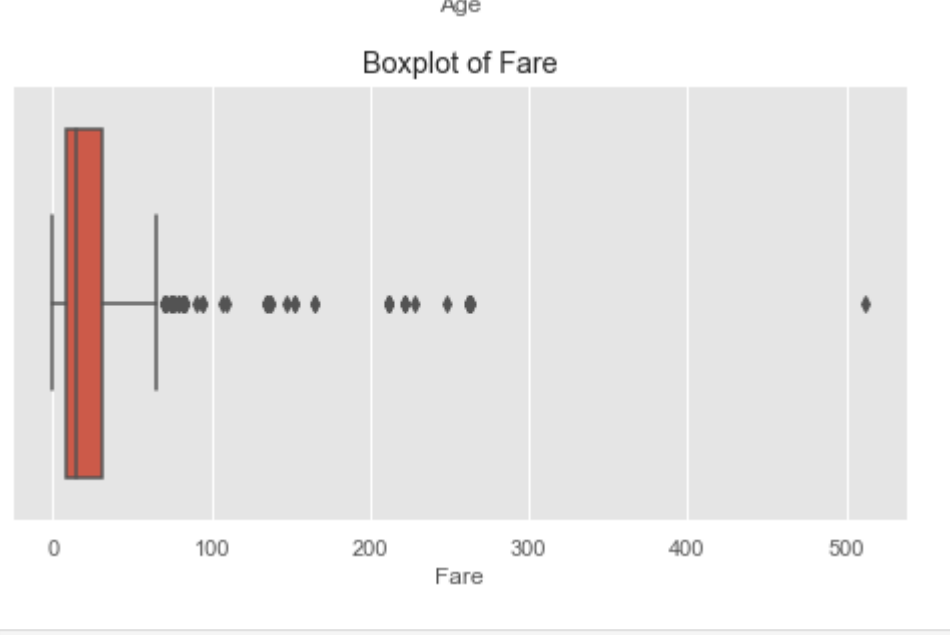
In [24]:
```python
features = ['Age', 'Fare']
for feature in features:
    plt.figure(figsize=(8, 4))
    sns.boxplot(x=train_df[feature])
    plt.title(f'Boxplot of {feature}')
    plt.show()
```





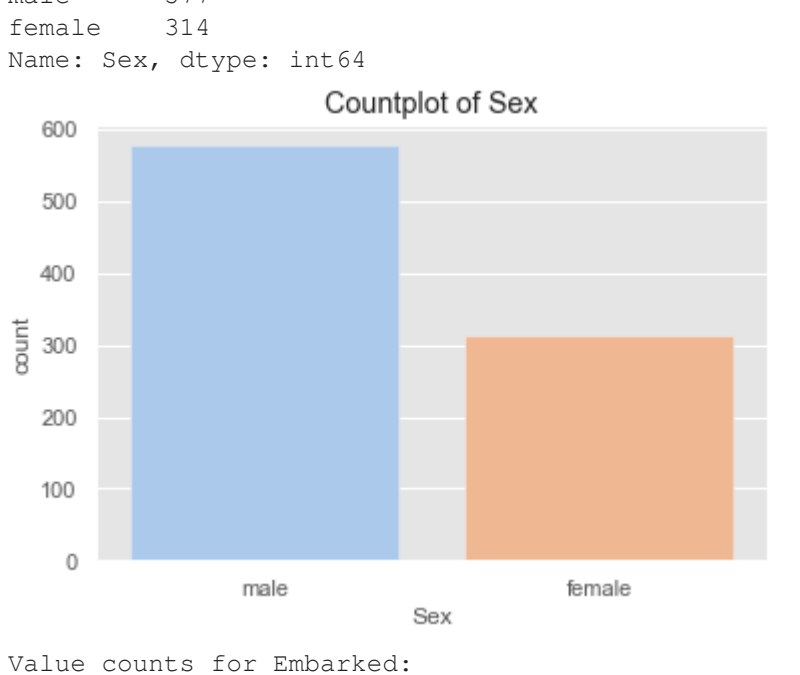
In [25]:
```python
features = ['Age', 'Fare']
for feature in features:
    plt.figure(figsize=(8, 4))
    sns.boxplot(x=test_df[feature])
    plt.title(f'Boxplot of {feature}')
    plt.show()
```



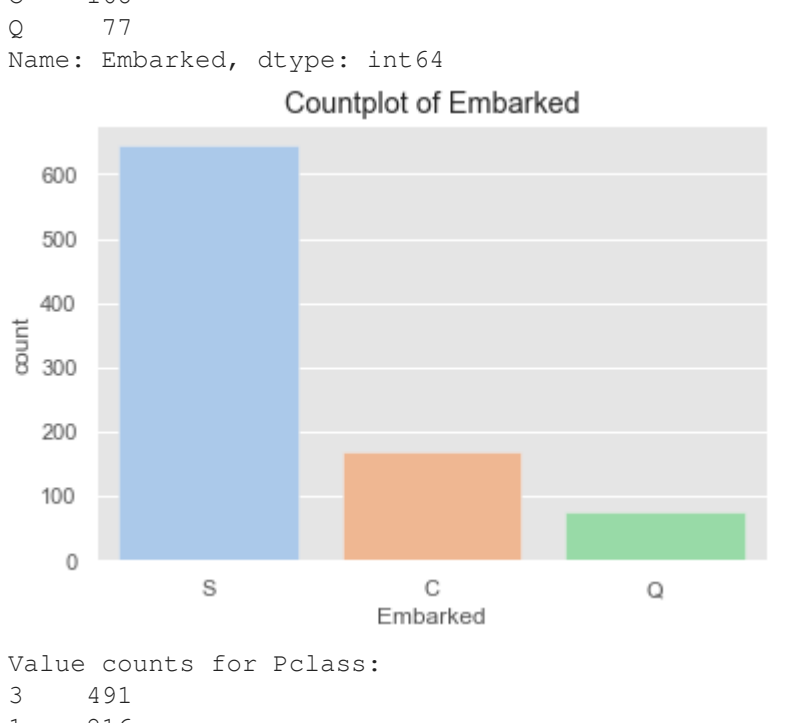


In [28]:
```python
categorical_features = ['Sex', 'Embarked', 'Pclass', 'Survived']

for feature in categorical_features:
    print(f"\nValue counts for {feature}:\n{train_df[feature].value_counts()}")

    sns.countplot(x=feature, data=train_df, palette='pastel')
    plt.title(f'Countplot of {feature}')
    plt.show()
```

```
Value counts for Sex:
male      577
female    314
Name: Sex, dtype: int64
```

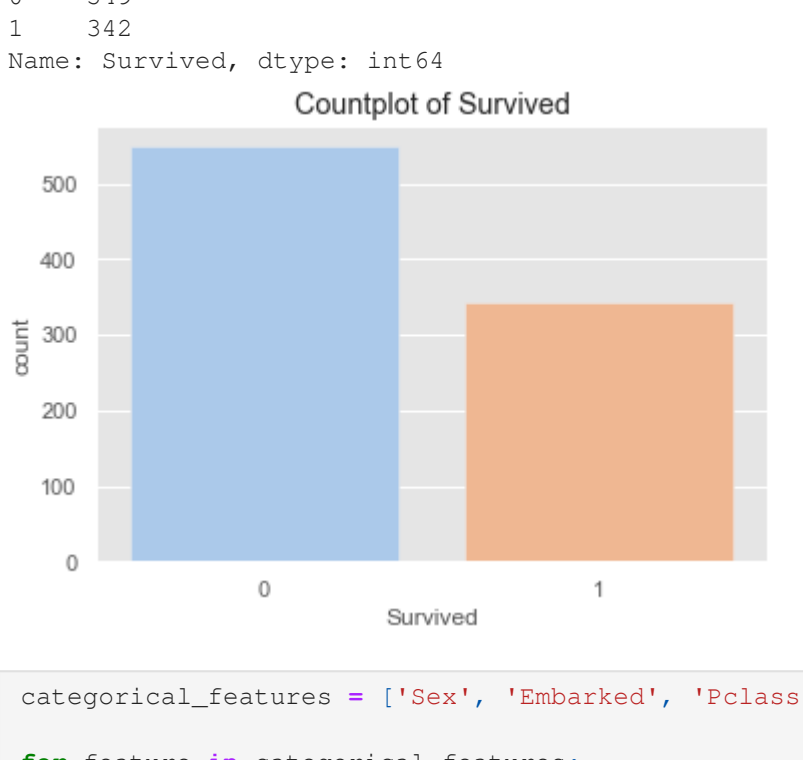



```
Value counts for Embarked:
S    644
C    168
Q     77
Name: Embarked, dtype: int64
```



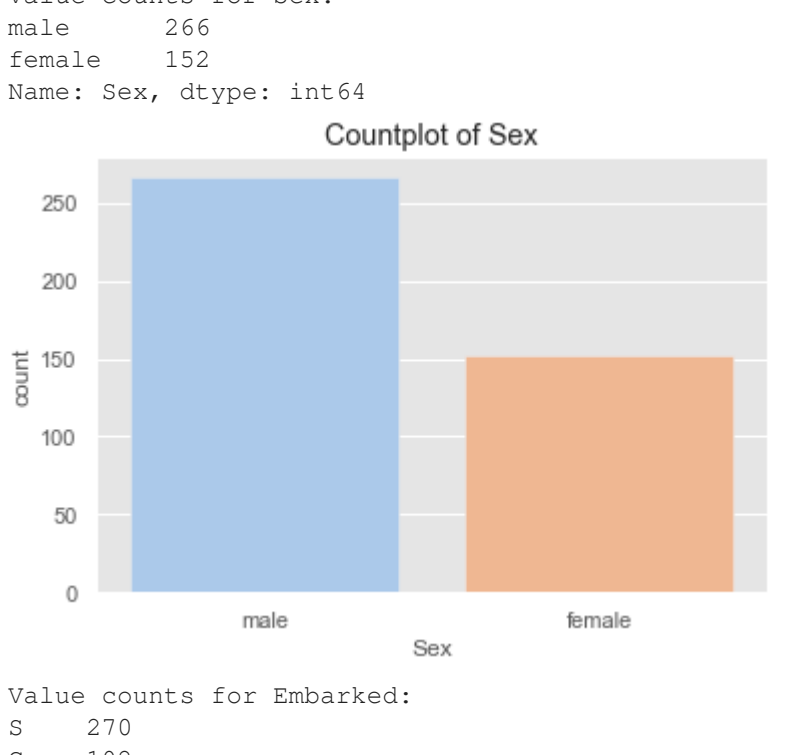

```
Value counts for Pclass:
3    491
1    216
2    184
Name: Pclass, dtype: int64
```


```
Value counts for Survived:
0    549
1    342
Name: Survived, dtype: int64
```

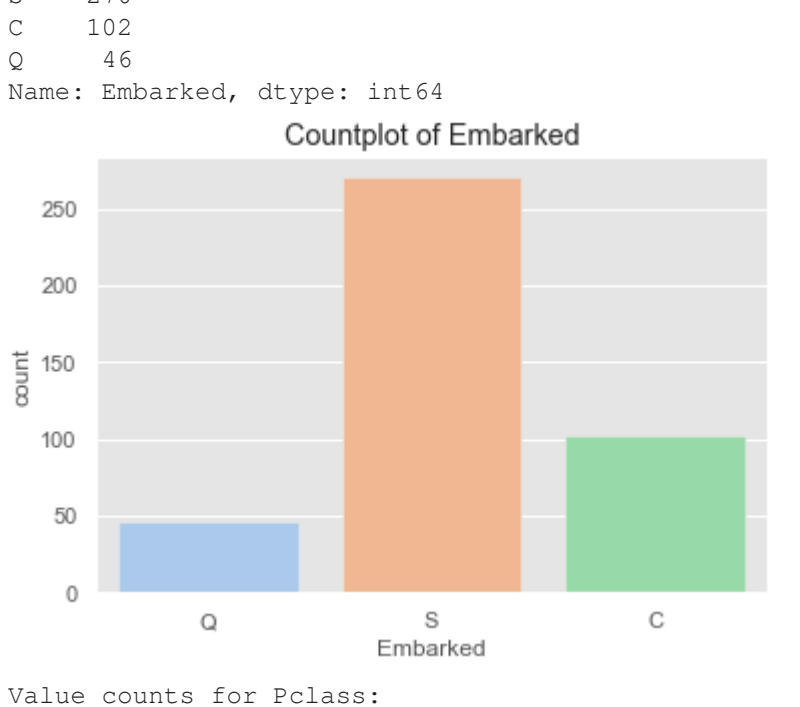


In [30]:
```python
categorical_features = ['Sex', 'Embarked', 'Pclass']

for feature in categorical_features:
    print(f"\nValue counts for {feature}:\n{test_df[feature].value_counts()}")

    sns.countplot(x=feature, data=test_df, palette='pastel')
    plt.title(f'Countplot of {feature}')
    plt.show()
```

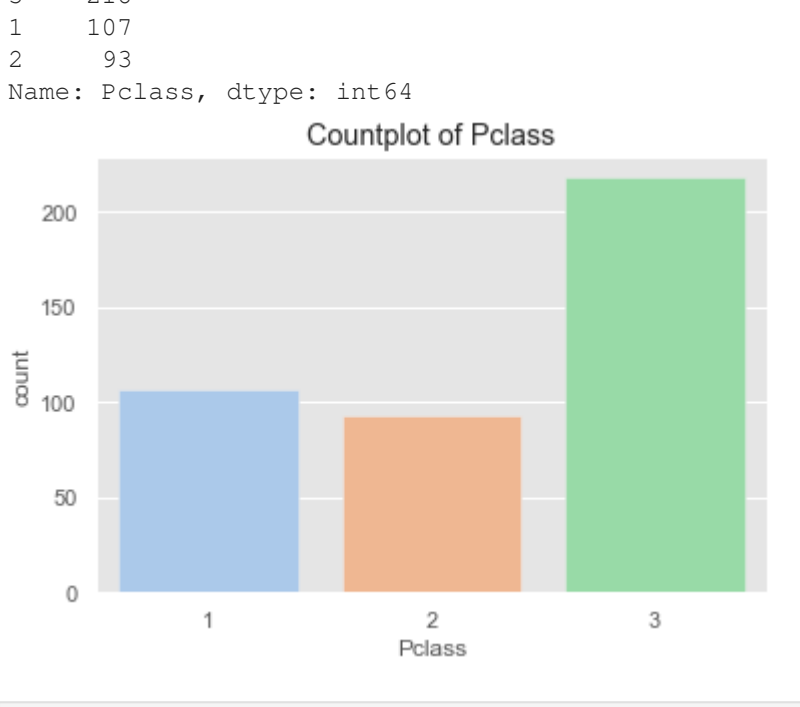
```
Value counts for Sex:
male      266
female    152
Name: Sex, dtype: int64
```



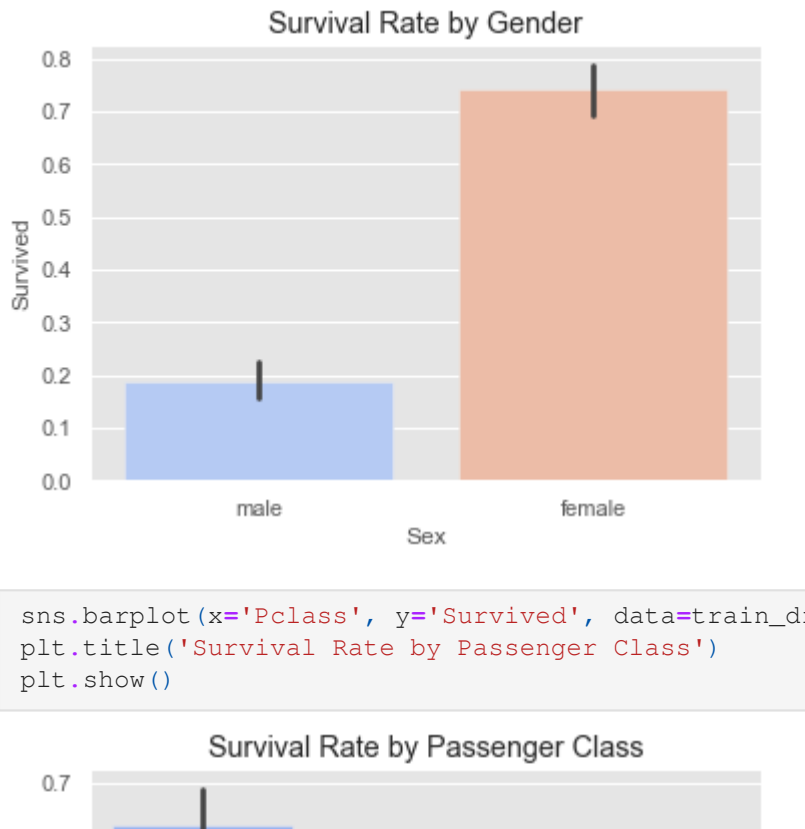

```
Value counts for Embarked:
S    270
C    102
Q     46
Name: Embarked, dtype: int64
```



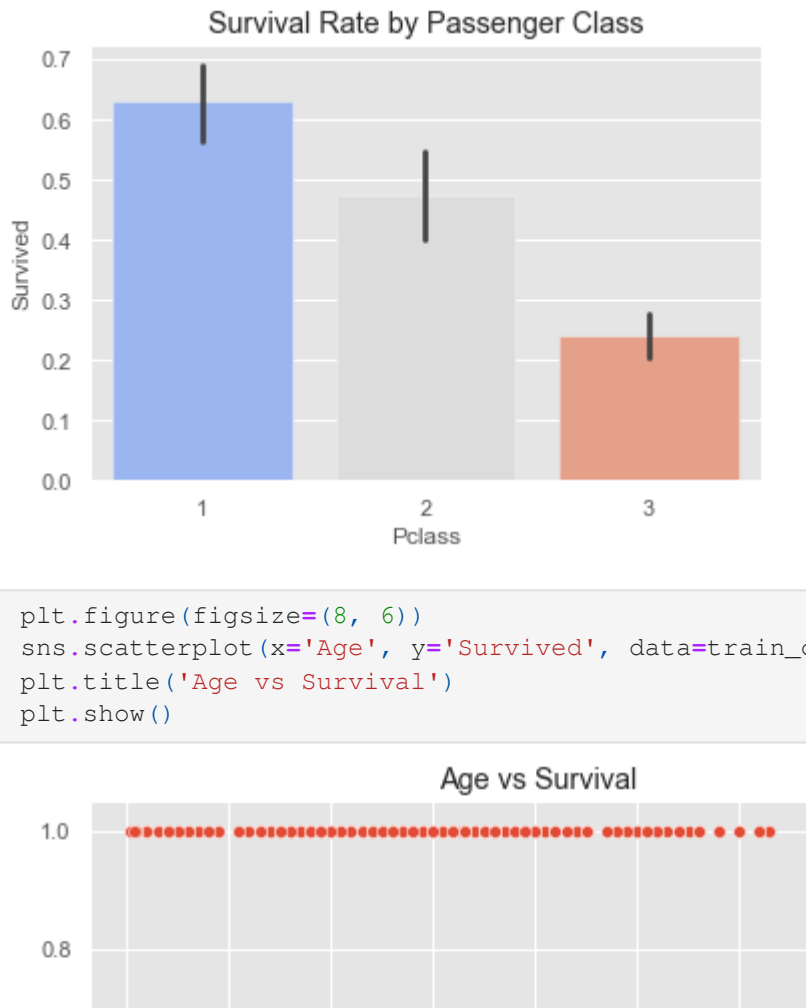

```
Value counts for Pclass:
3    218
1    107
2     93
Name: Pclass, dtype: int64
```



In [31]:
```python
sns.barplot(x='Sex', y='Survived', data=train_df, palette='coolwarm')
plt.title('Survival Rate by Gender')
plt.show()
```

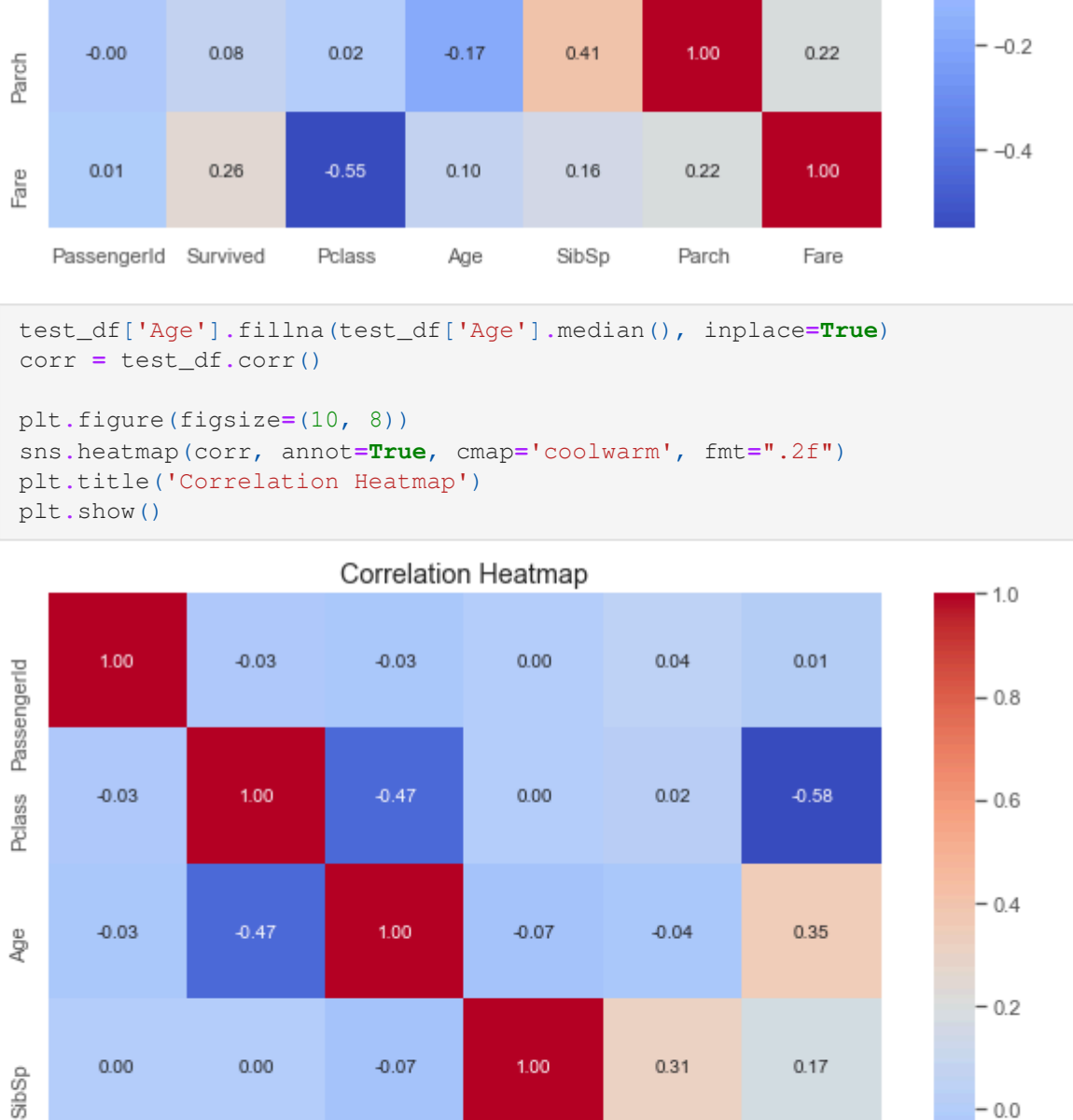


In [33]:
```python
sns.barplot(x='Pclass', y='Survived', data=train_df, palette='coolwarm')
plt.title('Survival Rate by Passenger Class')
plt.show()
```

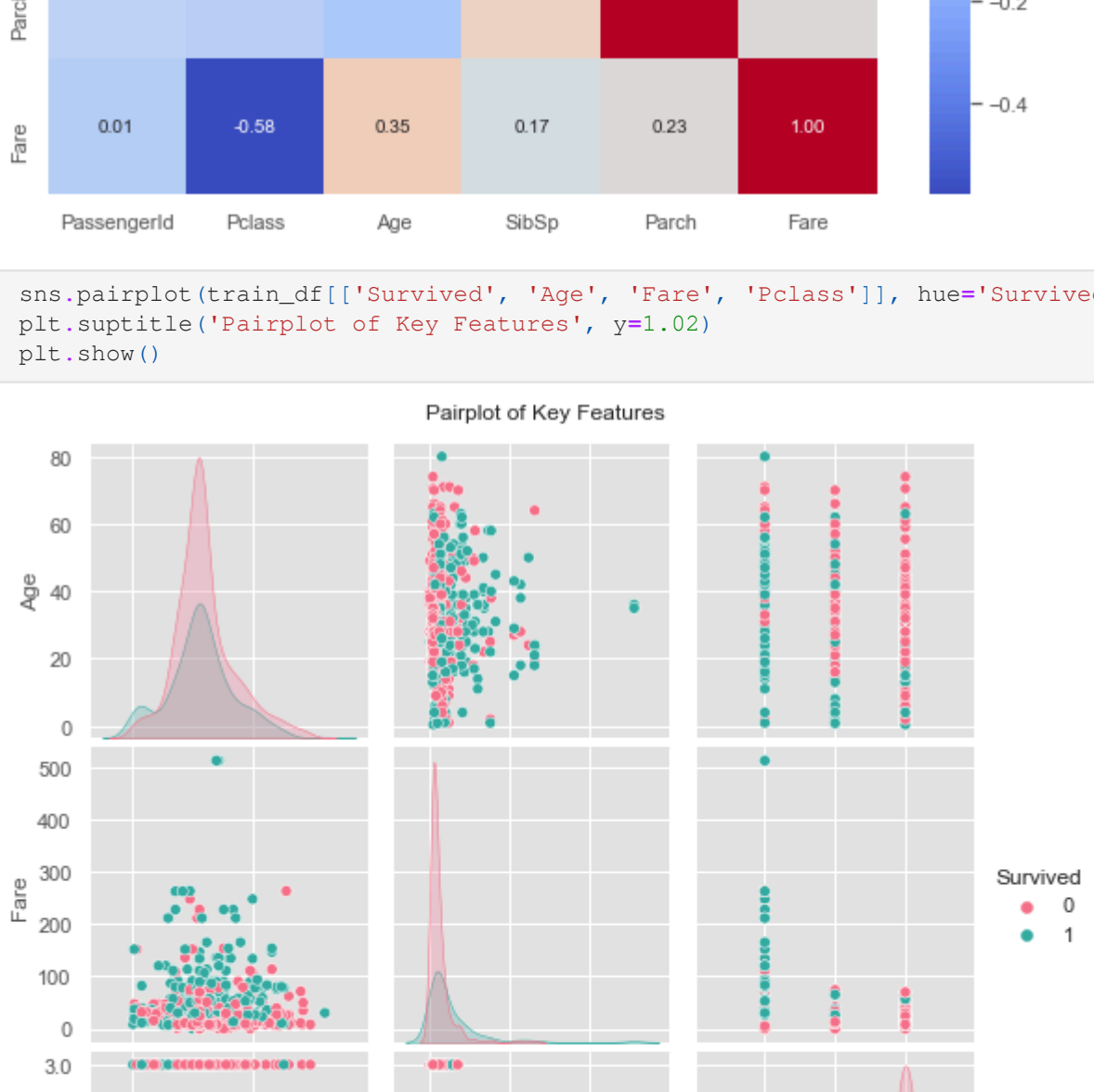


In [35]:
```python
plt.figure(figsize=(8, 6))
sns.scatterplot(x='Age', y='Survived', data=train_df)
plt.title('Age vs Survival')
plt.show()
```

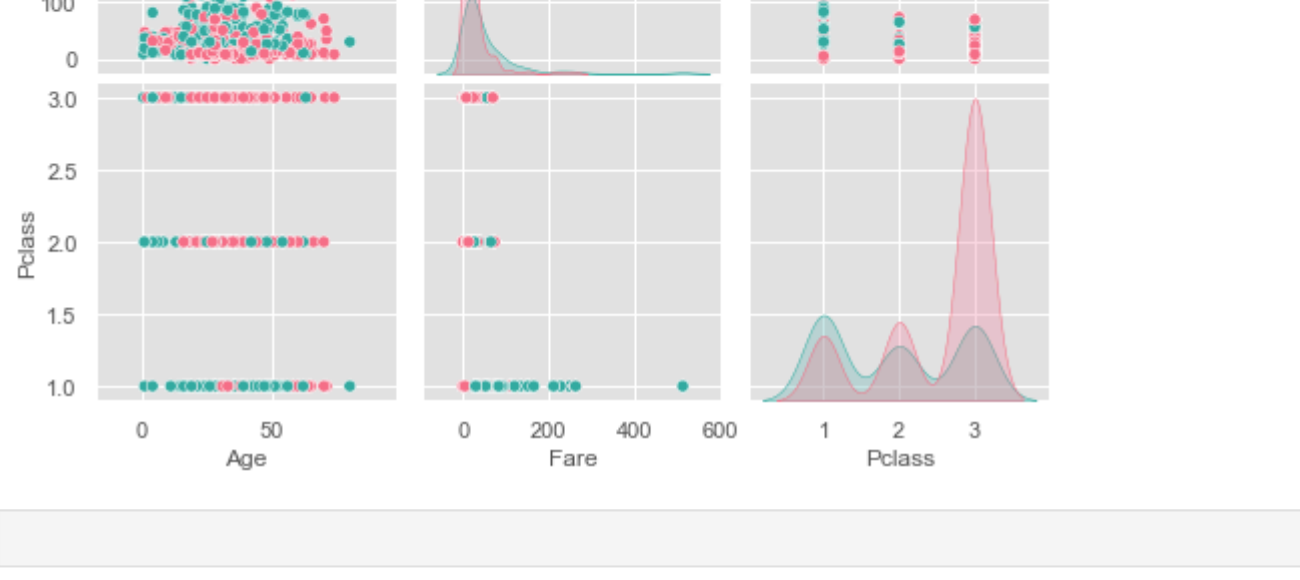
In [37]:
```python
train_df['Age'].fillna(train_df['Age'].median(), inplace=True)
corr = train_df.corr()

plt.figure(figsize=(10, 8))
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```



In [38]:
```python
test_df['Age'].fillna(test_df['Age'].median(), inplace=True)
corr = test_df.corr()

plt.figure(figsize=(10, 8))
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```



In [39]:
```python
sns.pairplot(train_df[['Survived', 'Age', 'Fare', 'Pclass']], hue='Survived', palette='husl')
plt.suptitle('Pairplot of Key Features', y=1.02)
plt.show()
```



In [ ]: