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


sns.set(style="whitegrid")
plt.rcParams['figure.figsize'] = (12, 8)


df = pd.read_csv('train.csv')


print("\n----- Basic Dataset Info -----")
df.info()


print("\n----- Statistical Summary -----")
print(df.describe())

print("\n----- Survival Count -----")
print(df['Survived'].value_counts())
print(df['Survived'].value_counts(normalize=True).round(3) * 100, "%
(Percentage)")

print("\n----- Passenger Class Count -----")
print(df['Pclass'].value_counts())
print(df['Pclass'].value_counts(normalize=True).round(3) * 100, "%
(Percentage)")

print("\n----- Gender Distribution -----")
print(df['Sex'].value_counts())
print(df['Sex'].value_counts(normalize=True).round(3) * 100, "%
(Percentage)")

print("\n----- Embarkation Port -----")
print(df['Embarked'].value_counts())
print(df['Embarked'].value_counts(normalize=True).round(3) * 100, "%
(Percentage)")
```

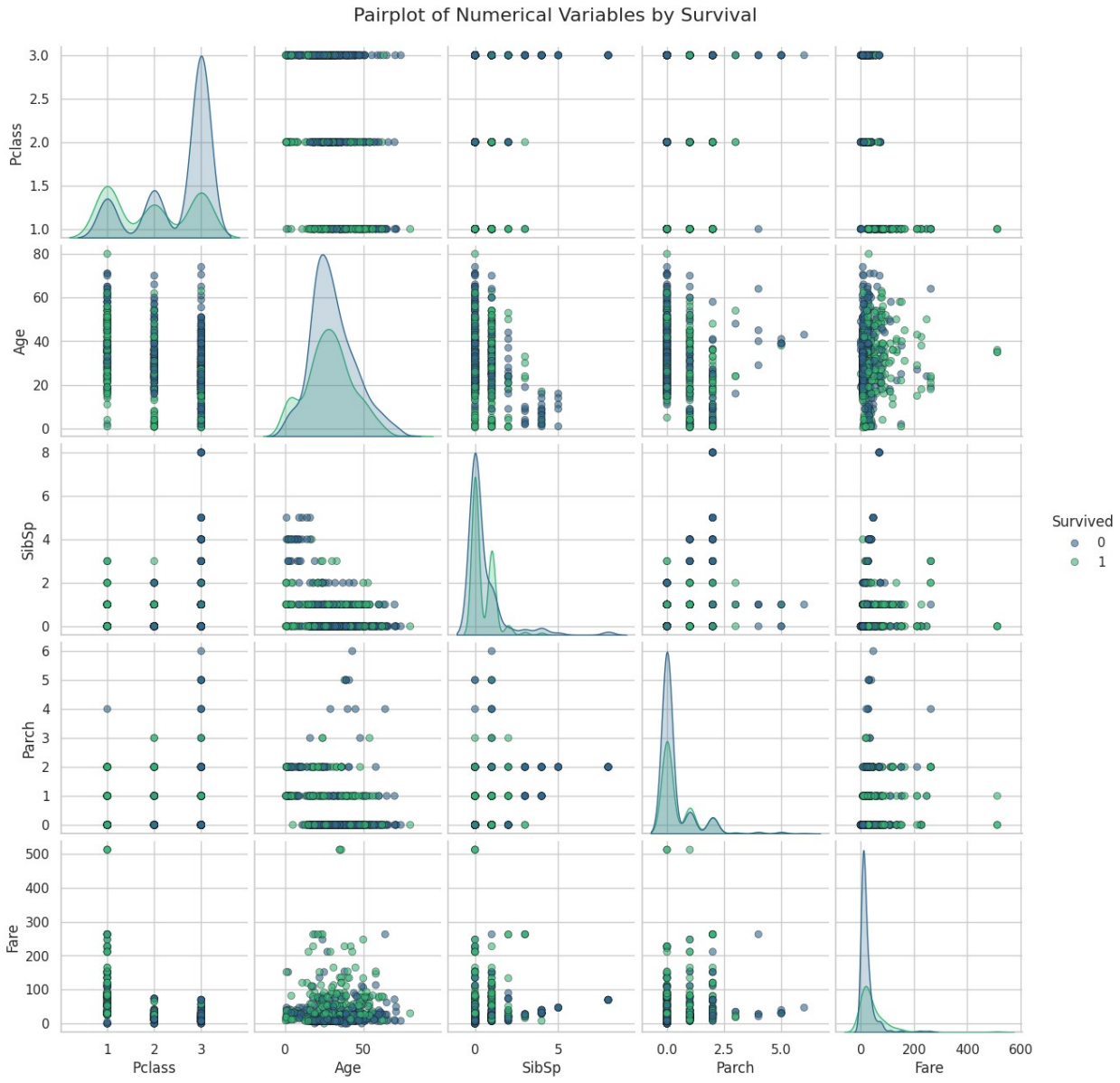
```
----- Basic Dataset Info -----
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
```

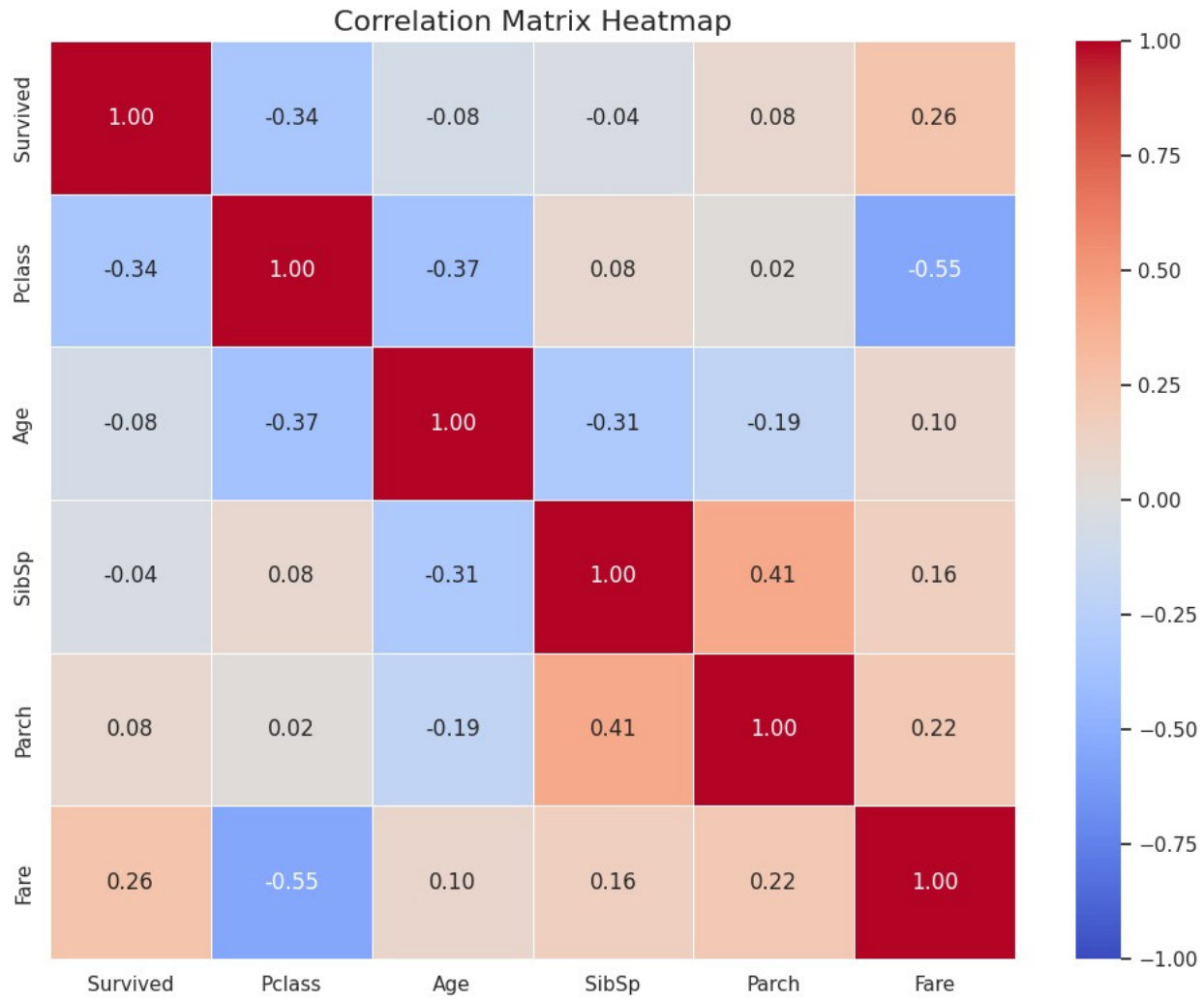
```
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB

----- Statistical Summary -----
       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

----- Survival Count -----
Survived
0    549
1    342
Name: count, dtype: int64
Survived
0    61.6
1    38.4
Name: proportion, dtype: float64 % (Percentage)

----- Passenger Class Count -----
Pclass
3    491
1    216
2    184
Name: count, dtype: int64
Pclass
```

```
3    55.1
1    24.2
2    20.7
Name: proportion, dtype: float64 % (Percentage)

----- Gender Distribution -----
Sex
male      577
female    314
Name: count, dtype: int64
Sex
male      64.8
female    35.2
Name: proportion, dtype: float64 % (Percentage)

----- Embarkation Port -----
Embarked
S    644
C    168
Q     77
Name: count, dtype: int64
Embarked
S    72.4
C    18.9
Q     8.7
Name: proportion, dtype: float64 % (Percentage)


numerical_cols = ['Survived', 'Pclass', 'Age', 'SibSp', 'Parch',
'Fare']

def create_pairplot():

    plt.figure(figsize=(12, 10))
    pairplot = sns.pairplot(df[numerical_cols],
                            hue='Survived',
                            palette='viridis',
                            diag_kind='kde',
                            plot_kws={'alpha': 0.6, 'edgecolor': 'k',
'linewidth': 0.5})
    plt.suptitle('Pairplot of Numerical Variables by Survival',
y=1.02, fontsize=16)
    plt.savefig('my_pairplot.png')
    plt.show()

def create_correlation_heatmap():

    correlation = df[numerical_cols].corr()
```

```
    plt.figure(figsize=(10, 8))
    heatmap = sns.heatmap(correlation,
                          annot=True,
                          cmap='coolwarm',
                          vmin=-1,
                          vmax=1,
                          fmt='.2f',
                          linewidths=0.5)
    plt.title('Correlation Matrix Heatmap', fontsize=16)
    plt.tight_layout()
    plt.savefig('my_correlation_heatmap.png')
    plt.show()

create_pairplot()
create_correlation_heatmap()
```

```
<Figure size 1200x1000 with 0 Axes>
```

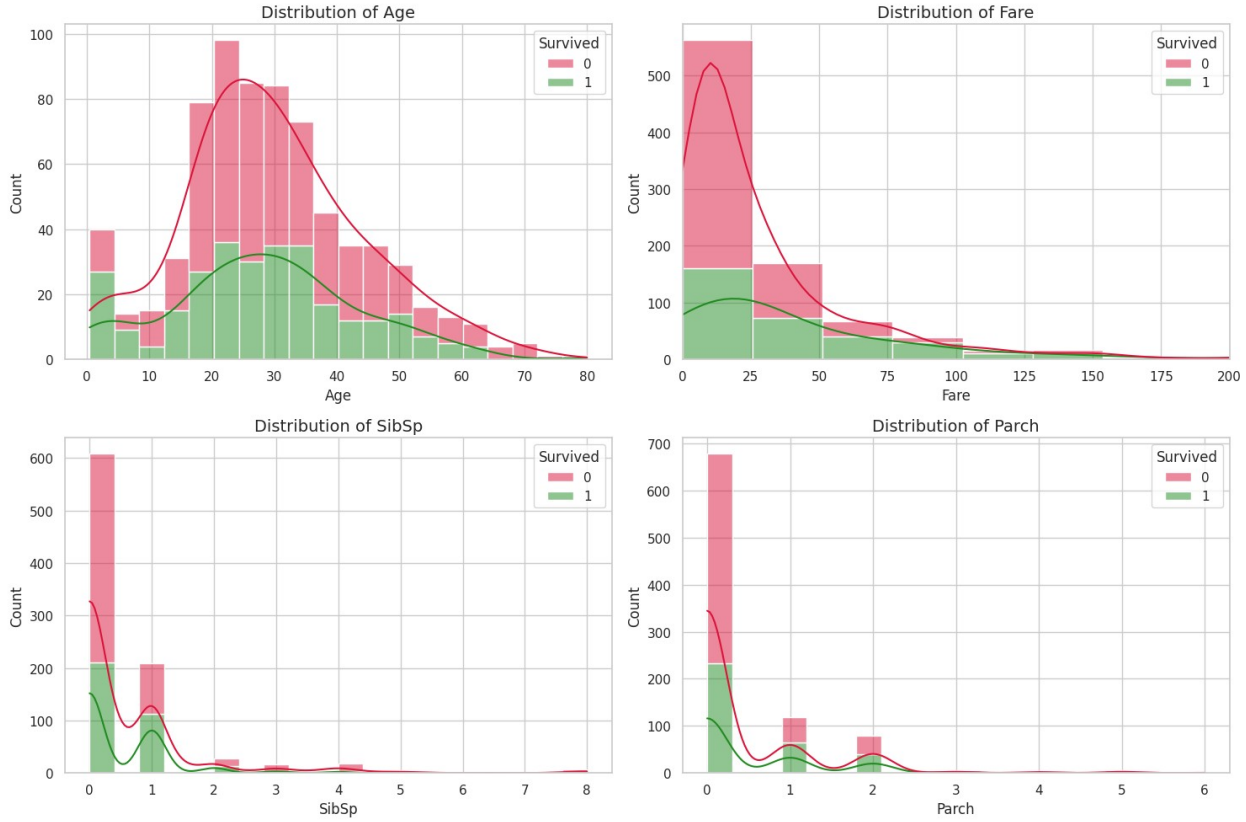

Pairplot of Numerical Variables by Survival
Pclass
Age
SibSp
Parch
Fare
Survived
0
1




```python
def create_histograms():

    plt.figure(figsize=(15, 10))

    features_to_plot = ['Age', 'Fare', 'SibSp', 'Parch']
    for i, feature in enumerate(features_to_plot):
        plt.subplot(2, 2, i+1)

        sns.histplot(data=df, x=feature, hue='Survived',
                     multiple='stack', palette=['crimson',
'forestgreen'],
                     kde=True, bins=20)
        plt.title(f'Distribution of {feature}', fontsize=14)
        if feature == 'Fare':
            plt.xlim(0, 200)

    plt.tight_layout()
    plt.savefig('my_histograms.png')
```

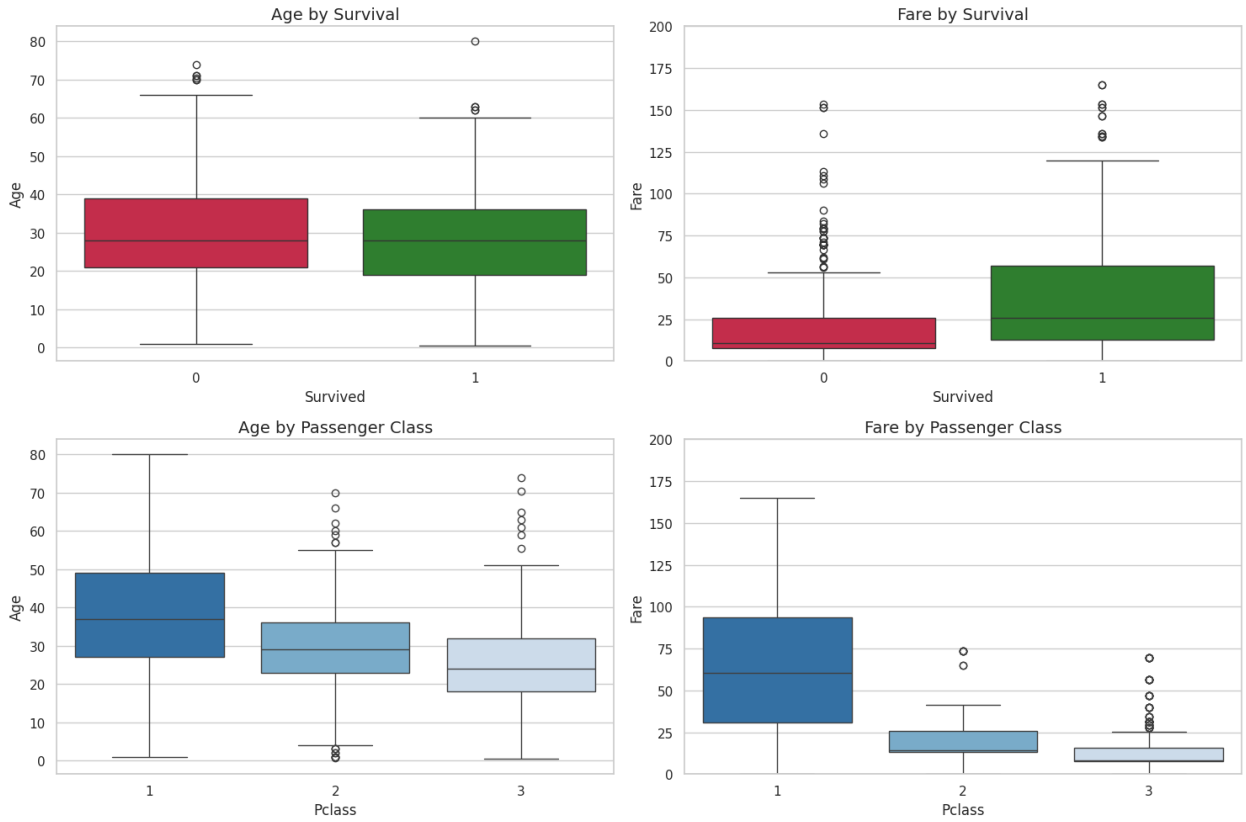
```
    plt.show()

def create_boxplots():

    plt.figure(figsize=(15, 10))


    plt.subplot(2, 2, 1)
    sns.boxplot(x='Survived', y='Age', data=df, palette=['crimson',
'forestgreen'])
    plt.title('Age by Survival', fontsize=14)

    plt.subplot(2, 2, 2)
    sns.boxplot(x='Survived', y='Fare', data=df, palette=['crimson',
'forestgreen'])
    plt.title('Fare by Survival', fontsize=14)
    plt.ylim(0, 200)


    plt.subplot(2, 2, 3)
    sns.boxplot(x='Pclass', y='Age', data=df, palette='Blues_r')
    plt.title('Age by Passenger Class', fontsize=14)

    plt.subplot(2, 2, 4)
    sns.boxplot(x='Pclass', y='Fare', data=df, palette='Blues_r')
    plt.title('Fare by Passenger Class', fontsize=14)
    plt.ylim(0, 200)

    plt.tight_layout()
    plt.savefig('my_boxplots.png')
    plt.show()

def create_scatterplots():

    plt.figure(figsize=(15, 8))


    plt.subplot(1, 2, 1)
    sns.scatterplot(x='Age', y='Fare', data=df, hue='Survived',
                    palette=['crimson', 'forestgreen'], alpha=0.7)
    plt.title('Age vs Fare by Survival Status', fontsize=14)
    plt.ylim(0, 200)


    plt.subplot(1, 2, 2)
    sns.scatterplot(x='SibSp', y='Parch', data=df, hue='Survived',
                    palette=['crimson', 'forestgreen'], size='Fare',
sizes=(20, 200), alpha=0.7)
    plt.title('Family Size Relationship (SibSp vs Parch)',
fontsize=14)
```

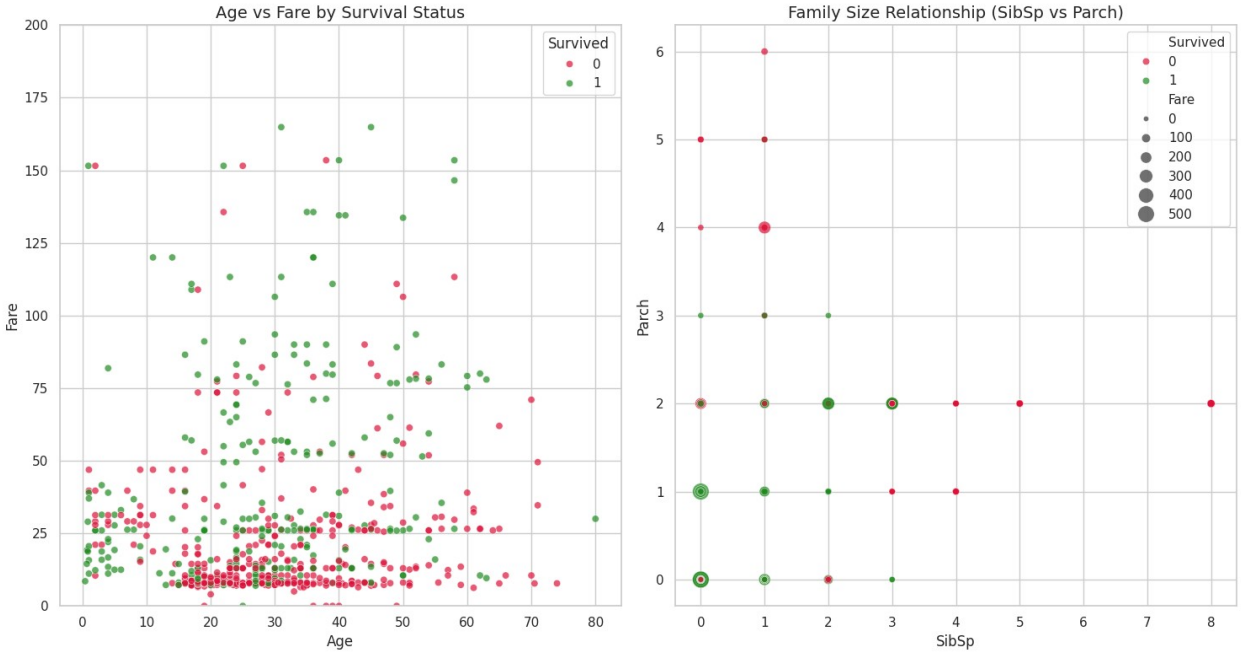
```
    plt.tight_layout()
    plt.savefig('my_scatterplots.png')
    plt.show()

create_histograms()
create_boxplots()
create_scatterplots()
```




```
<ipython-input-11-b2933c9deb54>:26: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  sns.boxplot(x='Survived', y='Age', data=df, palette=['crimson',
'forestgreen'])
<ipython-input-11-b2933c9deb54>:30: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  sns.boxplot(x='Survived', y='Fare', data=df, palette=['crimson',
'forestgreen'])
```

```
<ipython-input-11-b2933c9deb54>:36: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.boxplot(x='Pclass', y='Age', data=df, palette='Blues_r')
<ipython-input-11-b2933c9deb54>:40: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.boxplot(x='Pclass', y='Fare', data=df, palette='Blues_r')
```






```
def analyze_and_observe():
    """
    This function prints my observations from the visualizations
    and identifies relationships and trends in the data.
    """
    print("\n----- My Observations from Basic Exploration -----")
    print("1. I see that the dataset has 891 passengers with varying
degrees of information.")
    print("2. I notice there are missing values in Age (177), Cabin
(687), and Embarked (2).")
    print("3. Only about 38% of passengers survived the disaster.")
    print("4. There are more male passengers (65%) than female
passengers (35%).")
    print("5. The majority of passengers (55%) traveled in third
class.")
    print("6. Most passengers embarked from Southampton (S) at about
72%.")

    print("\n----- My Observations from Pairplot & Correlation Heatmap
-----")
    print("1. I can see that Pclass has a negative correlation with
Survival (-0.34),")
    print("   indicating that passengers in higher classes (lower
Pclass values) were more likely to survive.")
    print("2. Fare shows a positive correlation with Survival (0.26),
suggesting that")
    print("   passengers who paid more had better chances of
survival.")
    print("3. Age has a weak negative correlation with Survival (-
```

```
0.07), which")
    print("   might indicate slightly better survival rates for
younger passengers.")
    print("4. I notice that higher Fare is associated with higher
class (lower Pclass value).")

    print("\n----- My Observations from Histograms -----")
    print("1. The Age distribution appears right-skewed with most
passengers between 20-40 years.")
    print("2. The Fare distribution is heavily right-skewed with most
passengers paying lower fares.")
    print("3. I see that most passengers traveled alone or with very
few family members.")
    print("4. There seems to be better survival rates among passengers
who paid higher fares.")

    print("\n----- My Observations from Boxplots -----")
    print("1. Survivors tend to have slightly lower median age.")
    print("2. There's a significant difference in fares between
survivors and non-survivors,")
    print("   with survivors having paid higher fares on average.")
    print("3. First-class passengers were typically older than third-
class passengers.")
    print("4. There's a clear relationship between class and fare,
with first-class")
    print("   passengers paying much higher fares than others.")

    print("\n----- My Observations from Scatterplots -----")
    print("1. I see a cluster of high-fare passengers with better
survival rates.")
    print("2. Passengers with larger families (high SibSp+Parch)
generally show lower survival rates.")
    print("3. Middle-aged passengers paying higher fares had better
chances of survival.")

analyze_and_observe()
```

```
----- My Observations from Basic Exploration -----
1. I see that the dataset has 891 passengers with varying degrees of
information.
2. I notice there are missing values in Age (177), Cabin (687), and
Embarked (2).
3. Only about 38% of passengers survived the disaster.
4. There are more male passengers (65%) than female passengers (35%).
5. The majority of passengers (55%) traveled in third class.
6. Most passengers embarked from Southampton (S) at about 72%.

----- My Observations from Pairplot & Correlation Heatmap -----
1. I can see that Pclass has a negative correlation with Survival (-
```

```
0.34),
   indicating that passengers in higher classes (lower Pclass values)
were more likely to survive.
2. Fare shows a positive correlation with Survival (0.26), suggesting
that
   passengers who paid more had better chances of survival.
3. Age has a weak negative correlation with Survival (-0.07), which
   might indicate slightly better survival rates for younger
passengers.
4. I notice that higher Fare is associated with higher class (lower
Pclass value).

----- My Observations from Histograms -----
1. The Age distribution appears right-skewed with most passengers
between 20-40 years.
2. The Fare distribution is heavily right-skewed with most passengers
paying lower fares.
3. I see that most passengers traveled alone or with very few family
members.
4. There seems to be better survival rates among passengers who paid
higher fares.

----- My Observations from Boxplots -----
1. Survivors tend to have slightly lower median age.
2. There's a significant difference in fares between survivors and
non-survivors,
   with survivors having paid higher fares on average.
3. First-class passengers were typically older than third-class
passengers.
4. There's a clear relationship between class and fare, with first-
class
   passengers paying much higher fares than others.

----- My Observations from Scatterplots -----
1. I see a cluster of high-fare passengers with better survival rates.
2. Passengers with larger families (high SibSp+Parch) generally show
lower survival rates.
3. Middle-aged passengers paying higher fares had better chances of
survival.
```