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


df = pd.read_csv("/Titanic-Dataset.csv")


from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
df.head(10)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
df.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

```
df.info()
```

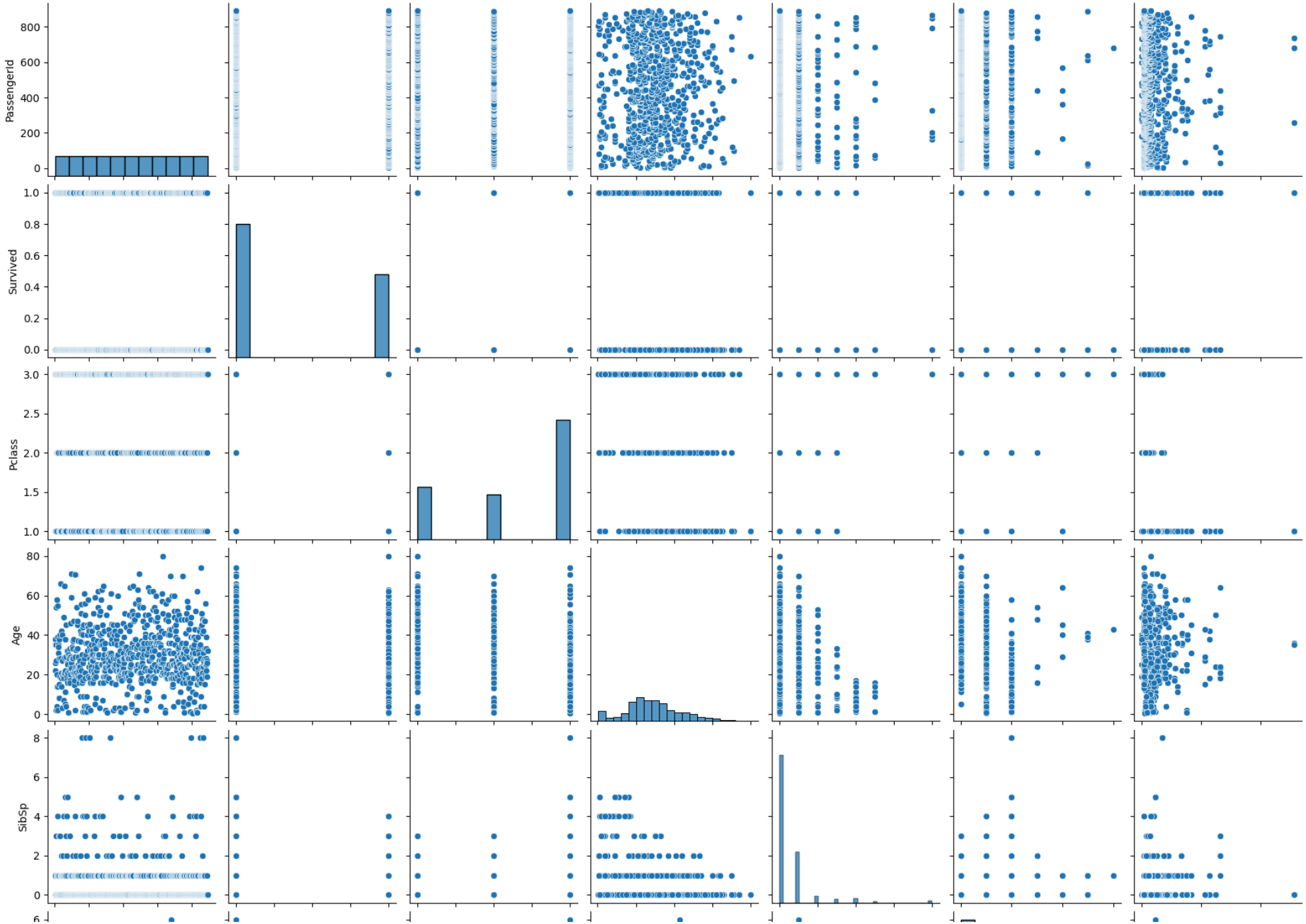
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null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

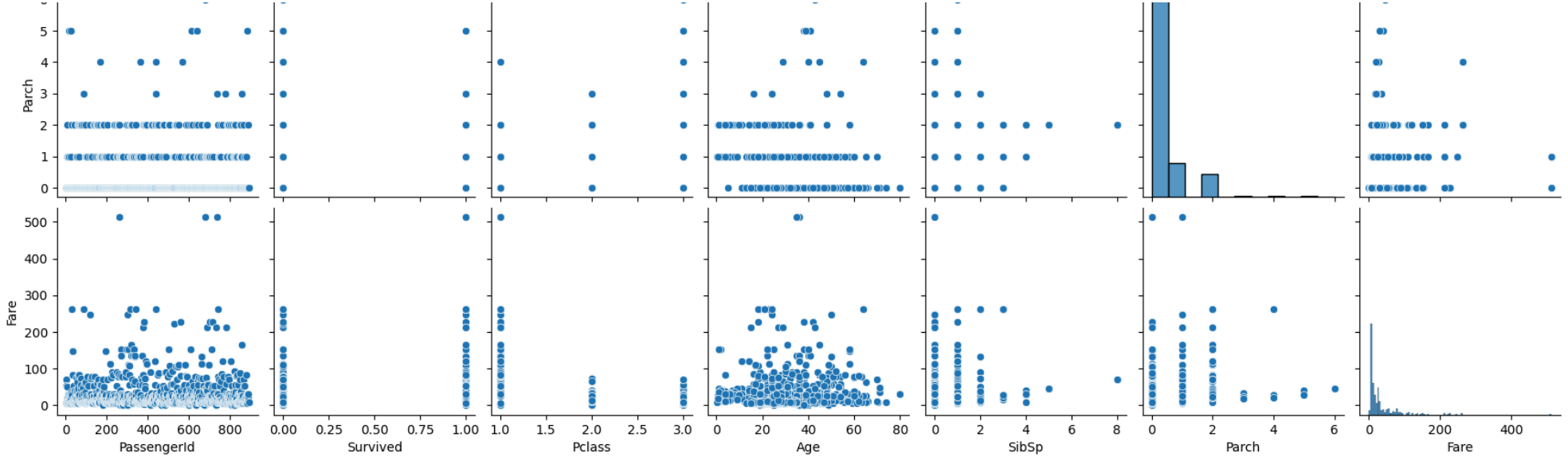
```
df.value_counts()
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	count
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	PC 17599	71.2833	C85	C	1
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	1
7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S	1
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S	1
12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S	1
...	...	...	...	...	...	...	...	...	...	...	...	...
872	1	1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	female	47.0	1	1	11751	52.5542	D35	S	1
873	0	1	Carlsson, Mr. Frans Olof	male	33.0	0	0	695	5.0000	B51 B53 B55	S	1
880	1	1	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.0	0	1	11767	83.1583	C50	C	1
888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S	1
890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C	1

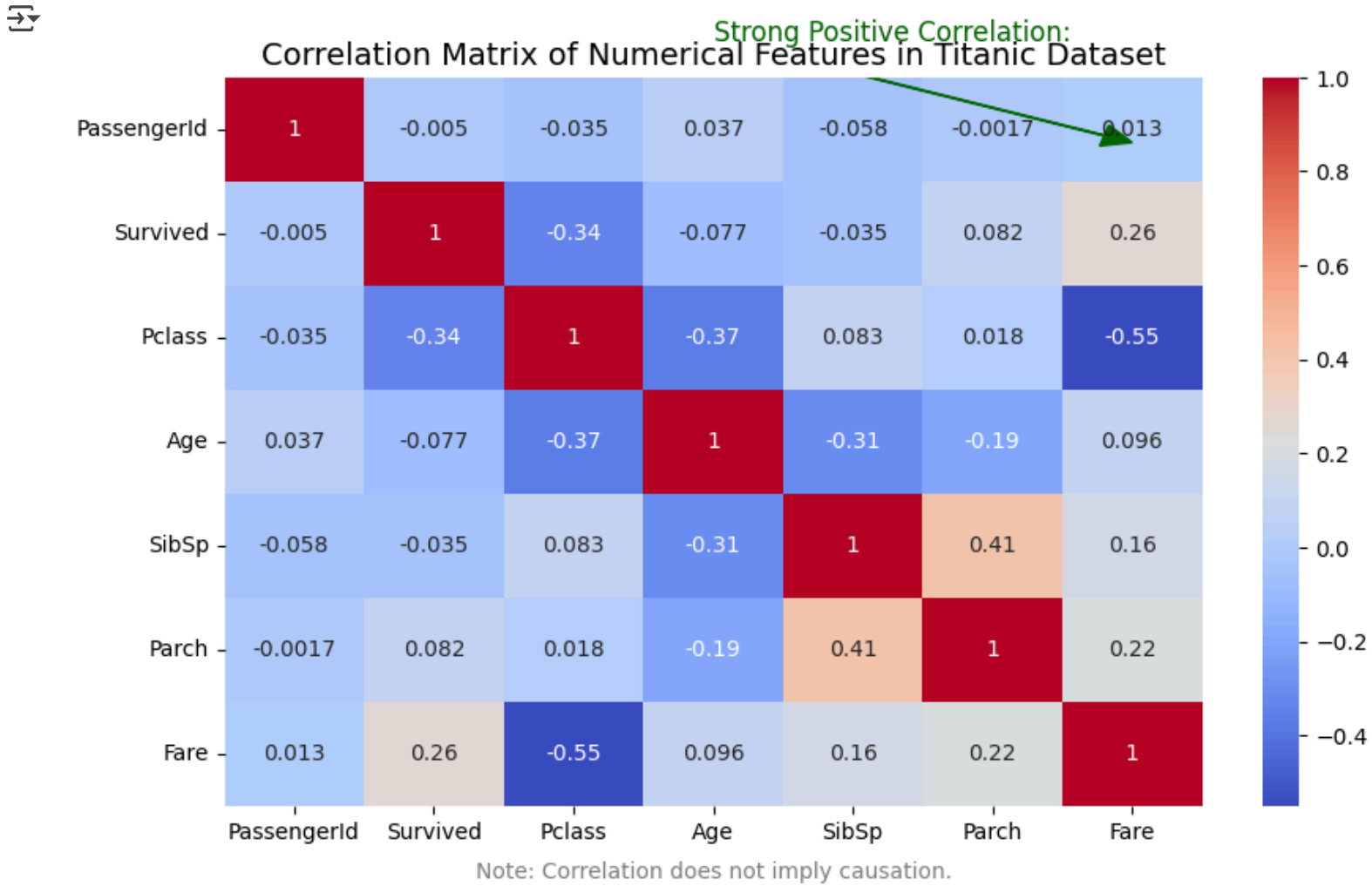
```
sns.pairplot(df)
```

<seaborn.axisgrid.PairGrid at 0x7d29ad095610>

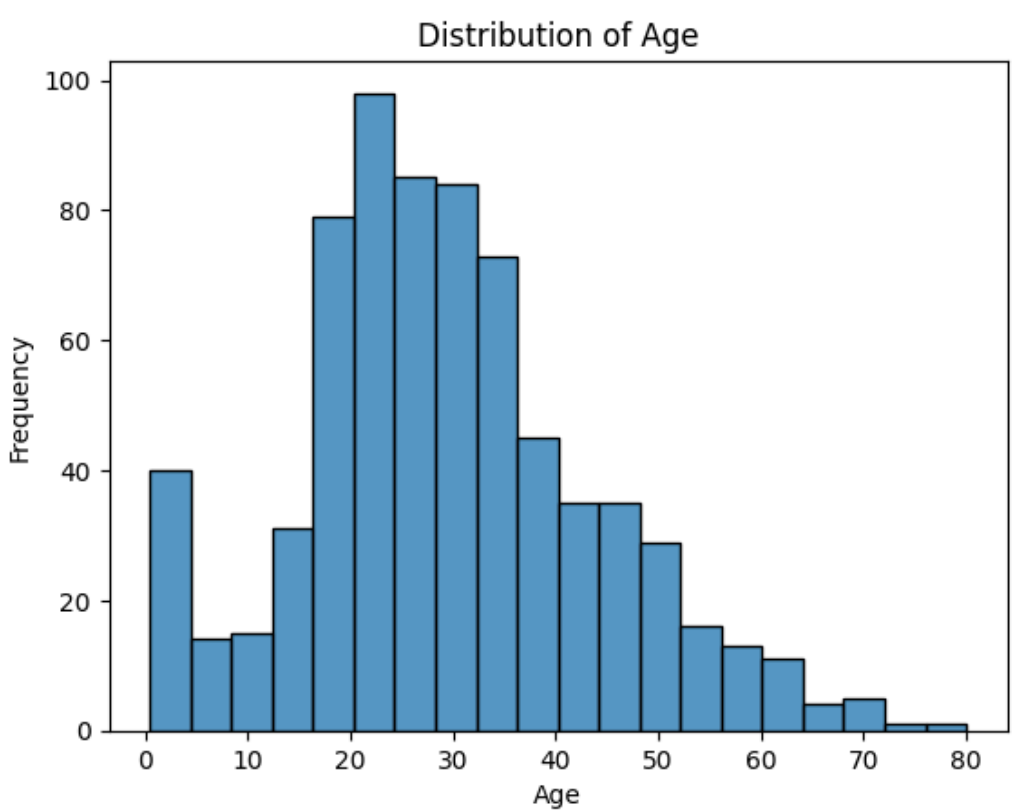


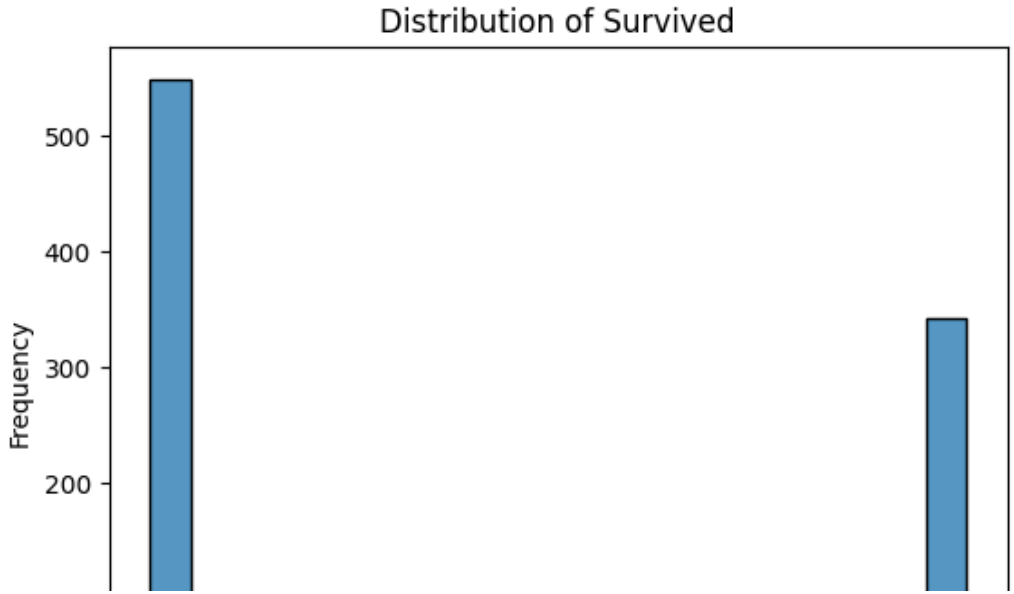


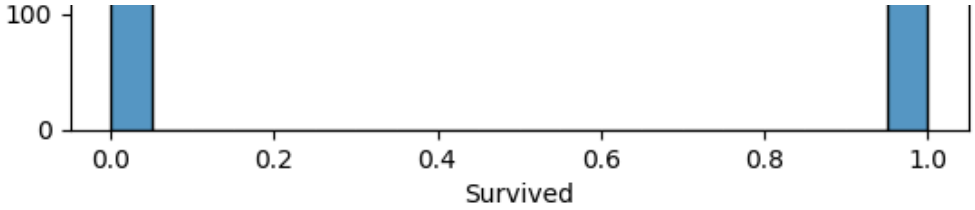

```
numerical_df = df.select_dtypes(include=np.number)
corr_matrix = numerical_df.corr()
plt.figure(figsize=(10, 6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix of Numerical Features in Titanic Dataset', fontsize=14)
plt.text(0.5, 1.05, "Strong Positive Correlation:", color='darkgreen', fontsize=12, transform=plt.gca().transAxes)
plt.arrow(0.6, 1.02, 0.3, -0.1, head_width=0.03, head_length=0.03, fc='darkgreen', ec='darkgreen', transform=plt.gca().transAxes)
plt.text(0.5, -0.1, "Note: Correlation does not imply causation.", color='gray', fontsize=10, transform=plt.gca().transAxes, ha='center')
plt.show()
```




```
# visualizing the distribution of age vs survived rate.
for col in ['Age', 'Survived']:
    sns.histplot(df[col], bins=20)
    plt.title(f'Distribution of {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
    plt.show()
```

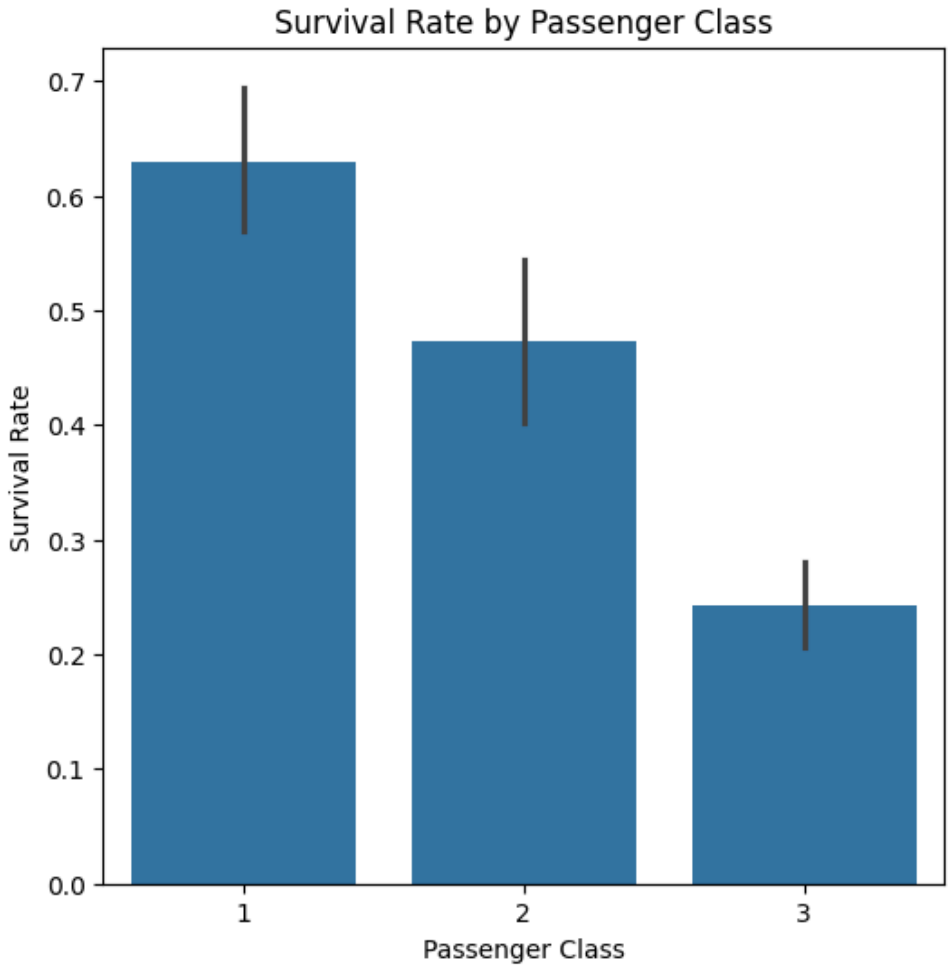

Distribution of Age
Frequency
Age
0
10
20
30
40
50
60
70
80
0
20
40
60
80
100


Distribution of Survived
Frequency
200
300
400
500

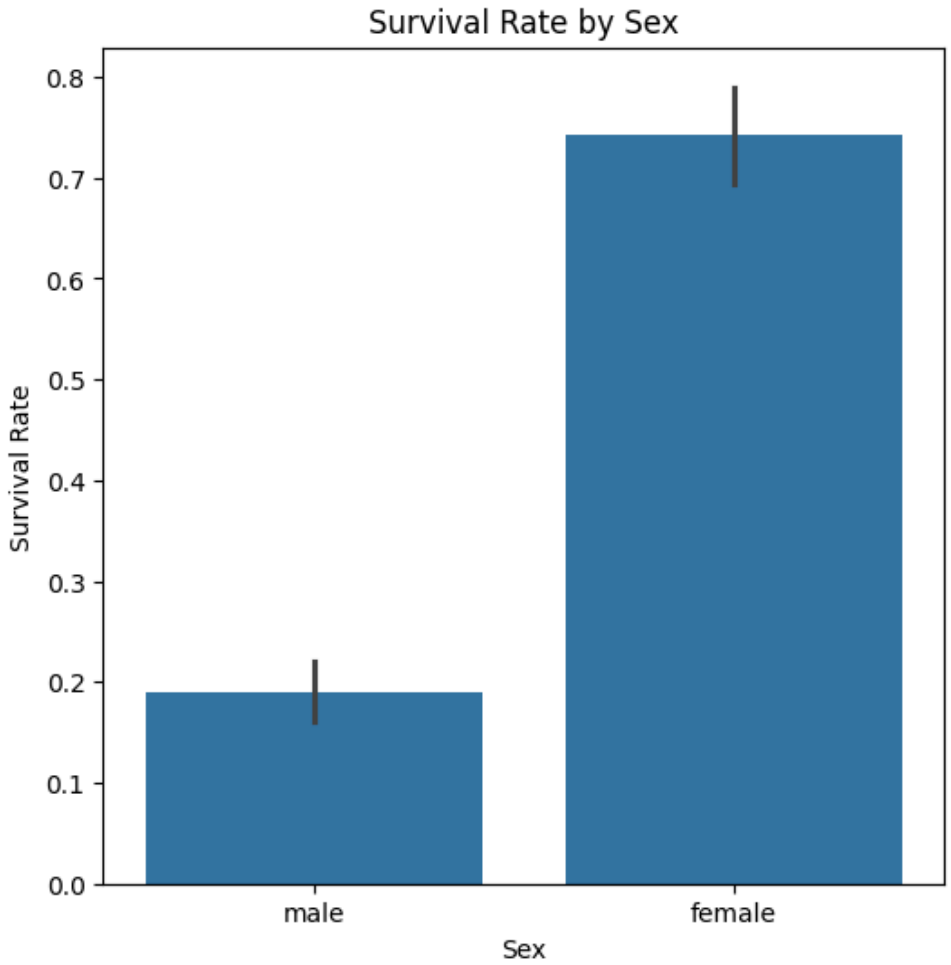


Identifying relationships and trends with the help of visualization.

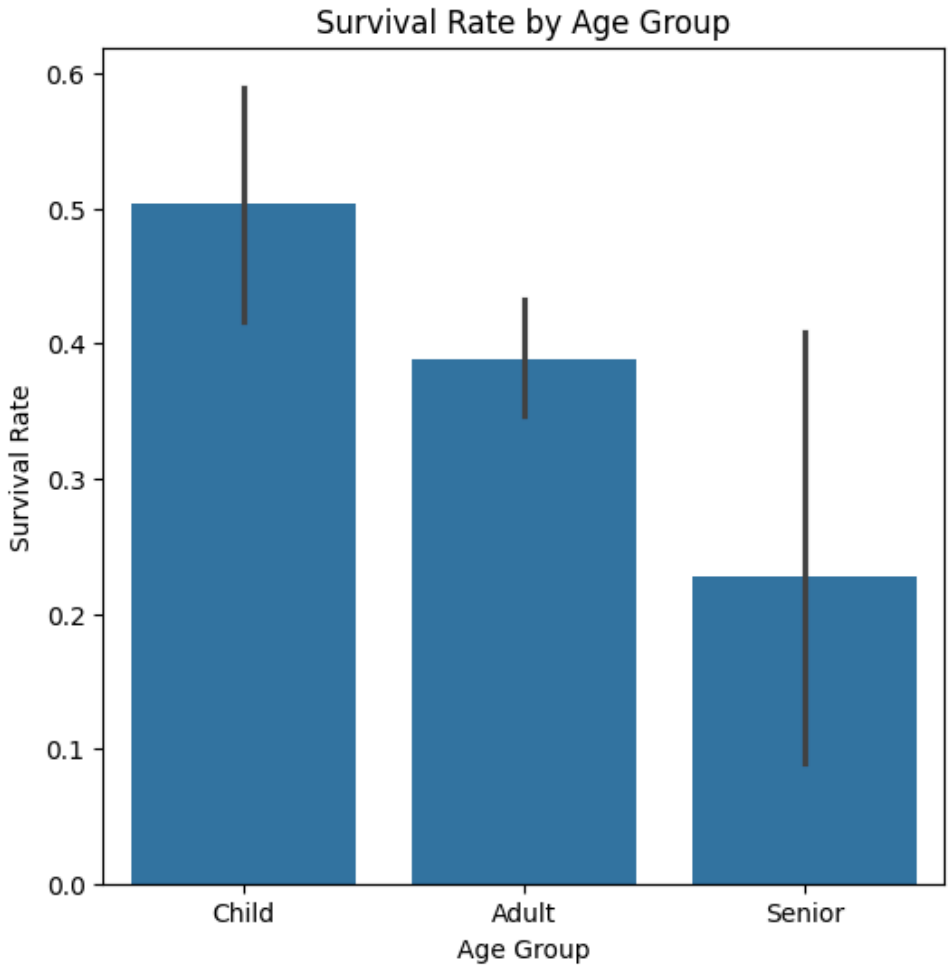
```
# 1. Survival Rate by Passenger Class (Pclass)
plt.figure(figsize=(6, 6))
sns.barplot(x='Pclass', y='Survived', data=df)
plt.title('Survival Rate by Passenger Class')
plt.xlabel('Passenger Class')
plt.ylabel('Survival Rate')
plt.show()
```



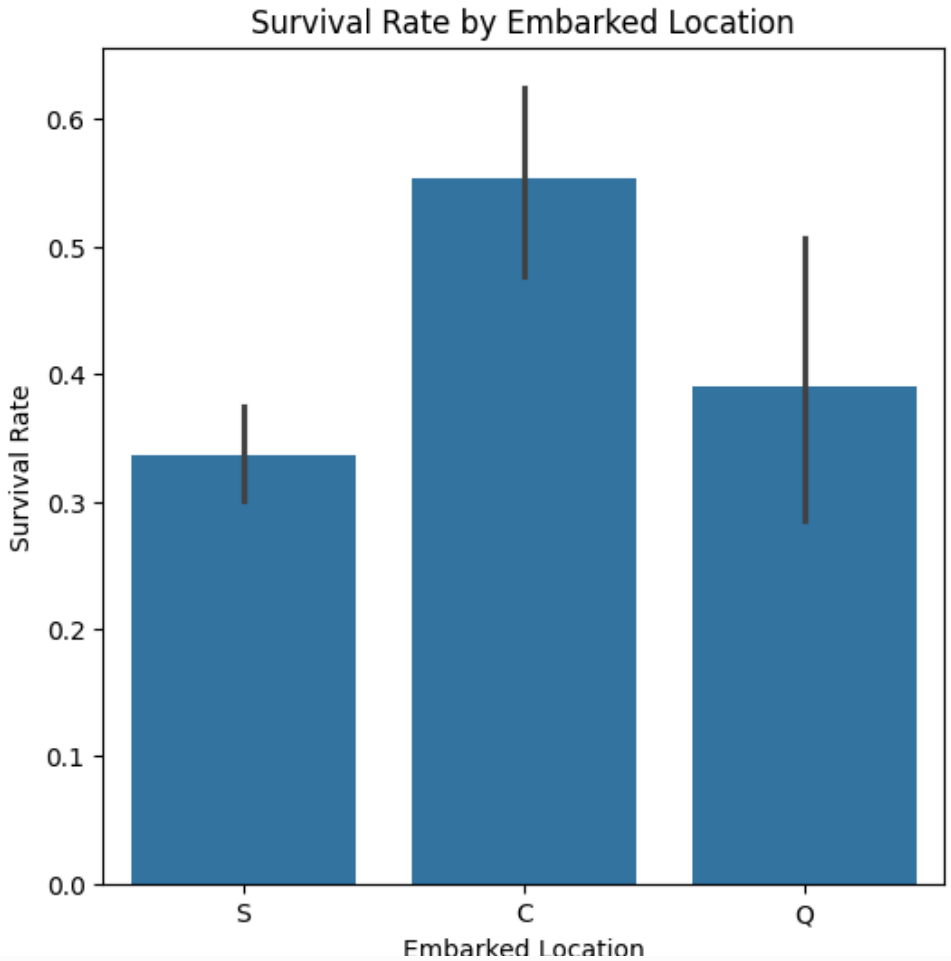

```
# 2. Survival Rate by Sex
plt.figure(figsize=(6, 6))
sns.barplot(x='Sex', y='Survived', data=df)
plt.title('Survival Rate by Sex')
plt.xlabel('Sex')
plt.ylabel('Survival Rate')
plt.show()
```



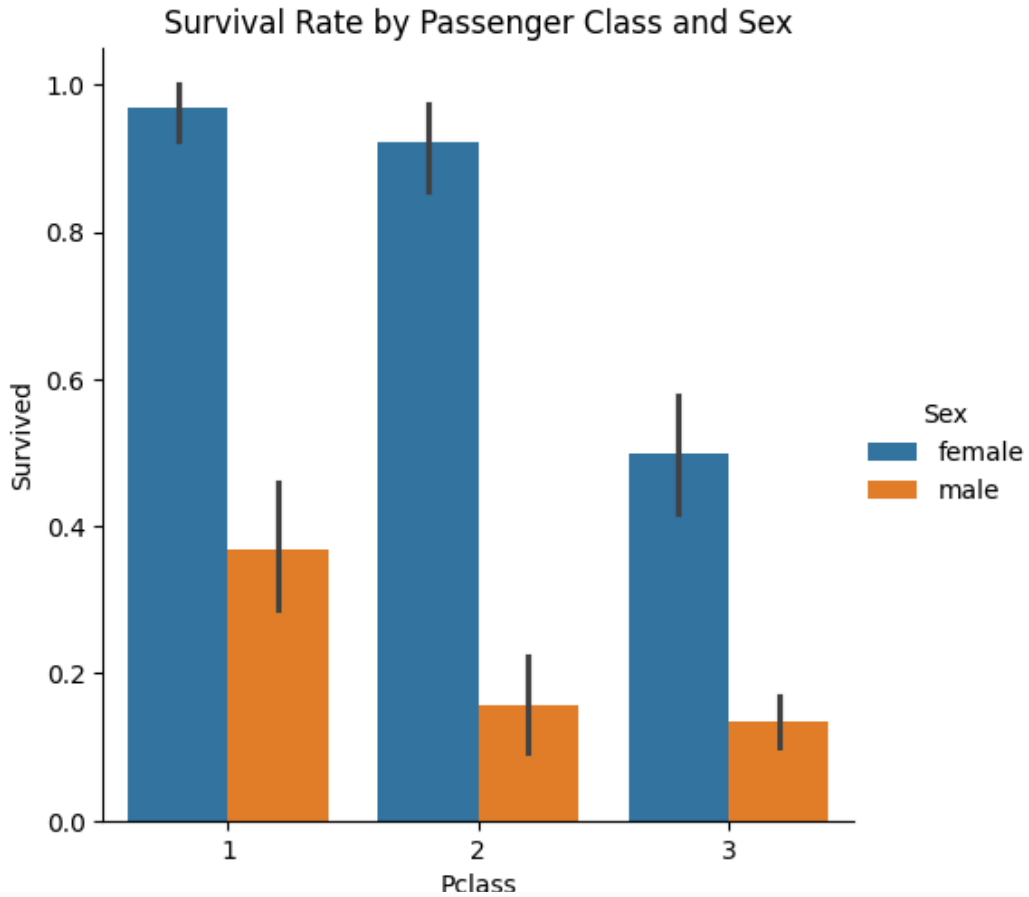

```
# 3. Survival Rate by Age Group (Categorical)
df['AgeGroup'] = pd.cut(df['Age'], bins=[0, 18, 60, np.inf], labels=['Child', 'Adult', 'Senior'])
plt.figure(figsize=(6, 6))
sns.barplot(x='AgeGroup', y='Survived', data=df)
plt.title('Survival Rate by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Survival Rate')
plt.show()
```



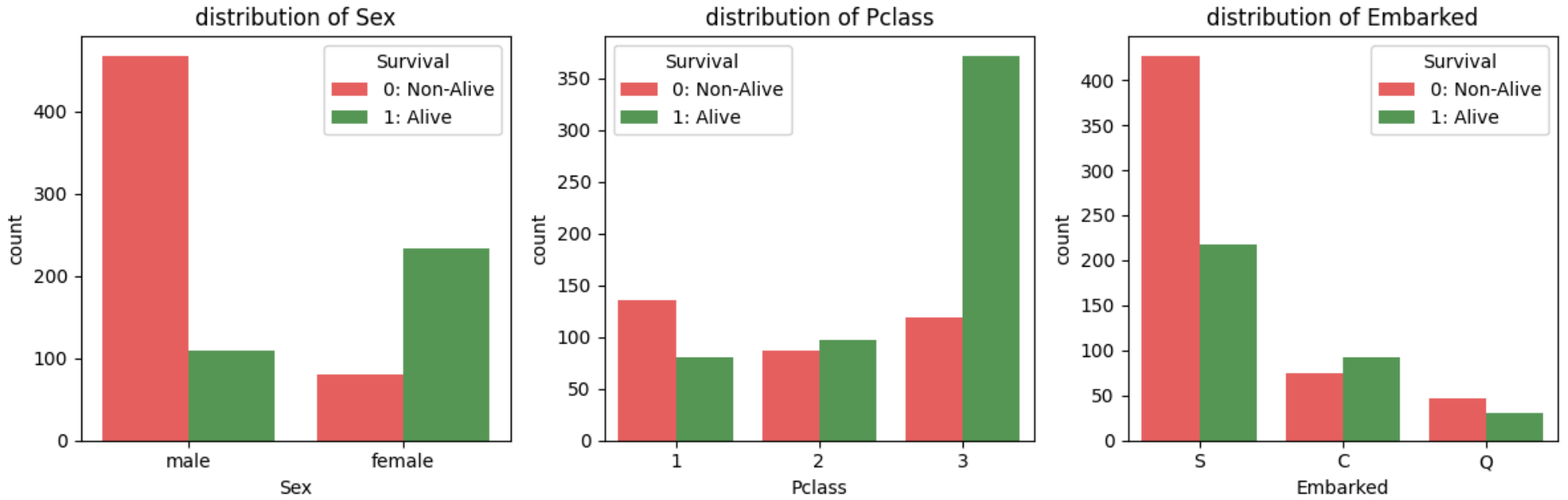

```
# 4. Survival Rate by Embarked Location
plt.figure(figsize=(6, 6))
sns.barplot(x='Embarked', y='Survived', data=df)
plt.title('Survival Rate by Embarked Location')
plt.xlabel('Embarked Location')
plt.ylabel('Survival Rate')
plt.show()
```



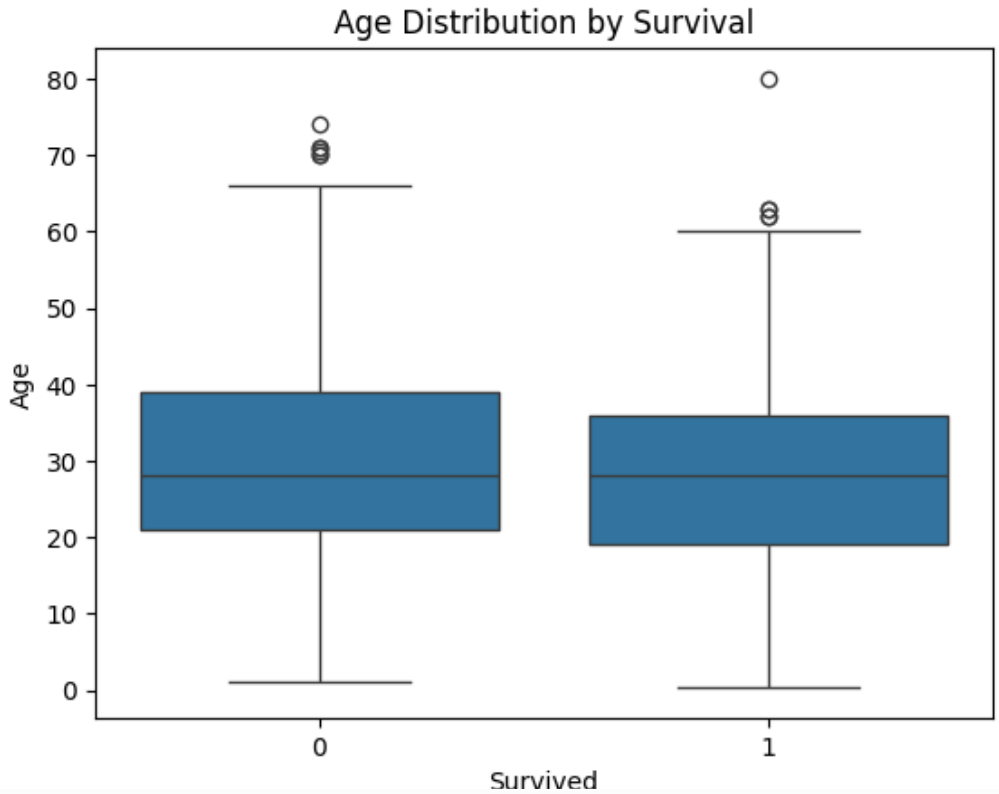

```
# 5.  Combined visualization: Survival rate across multiple factors
sns.catplot(x="Pclass", y="Survived", hue="Sex", kind="bar", data=df)
plt.title('Survival Rate by Passenger Class and Sex')
plt.show()
```



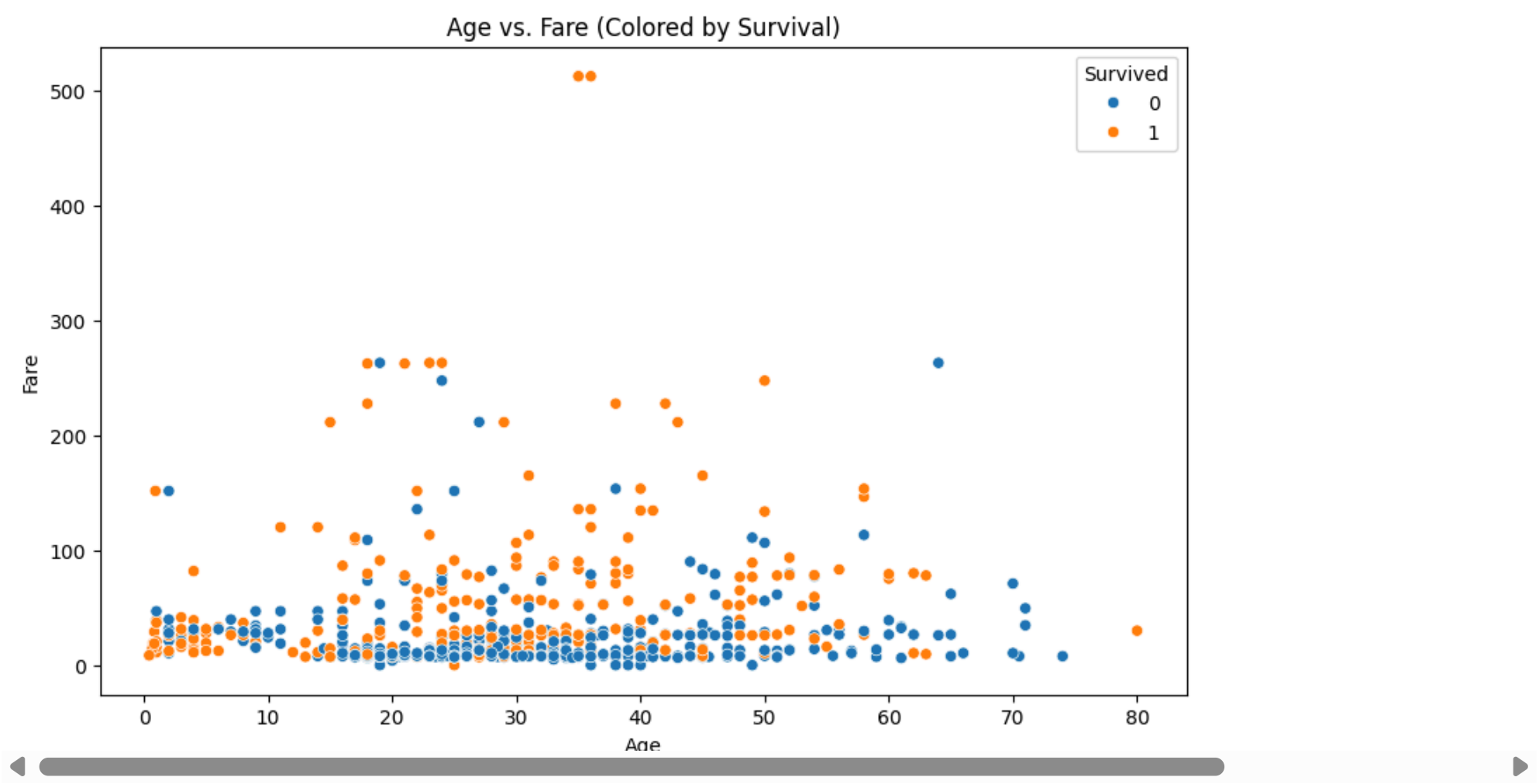

```
# combined visualization of survivalr rate according to distribution on sex, pClass and Embarked.
cat_cols = ["Sex", "Pclass", "Embarked"]
num_cols = ["Age", "SibSp", "Parch", "Fare"]
fig, axes = plt.subplots(1, len(cat_cols), figsize=(4 * len(cat_cols), 4))
for ax, col in zip(axes, cat_cols):
    sns.countplot(data=df, x=col, hue=df["Survived"].map(str), palette=["red","green"], alpha=0.7, ax=ax)
    ax.set_title(f"distribution of {col}")
    ax.legend(title="Survival", labels=["0: Non-Alive", "1: Alive"])
plt.tight_layout()
plt.show()
```



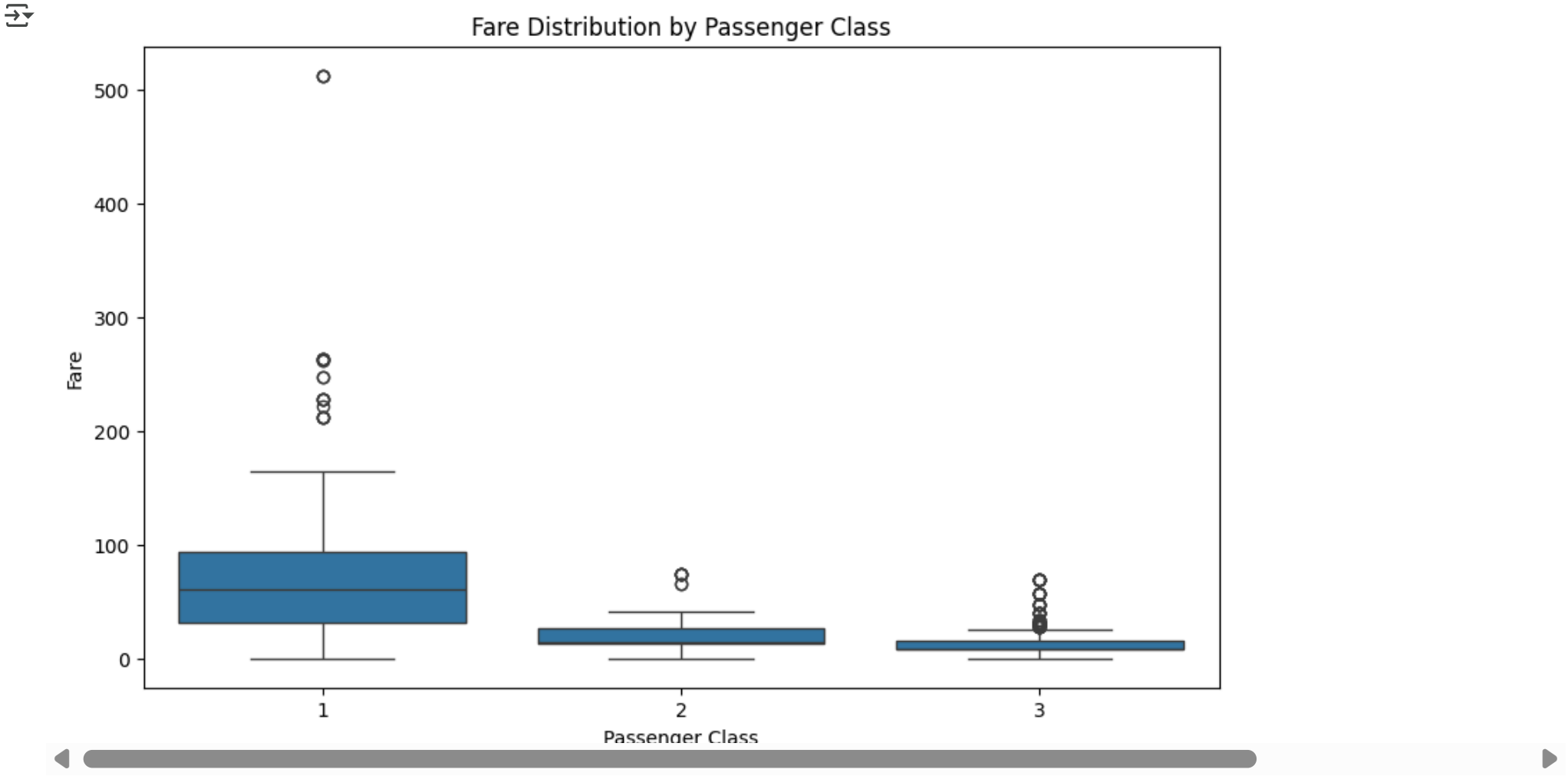

```
# 7. Box plots to compare distributions across groups (Example: Age vs. Survival)
sns.boxplot(x='Survived', y='Age', data=df)
plt.title('Age Distribution by Survival')
plt.show()
```




```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df)
plt.title('Age vs. Fare (Colored by Survival)')
plt.xlabel('Age')
plt.ylabel('Fare')
plt.show()
```

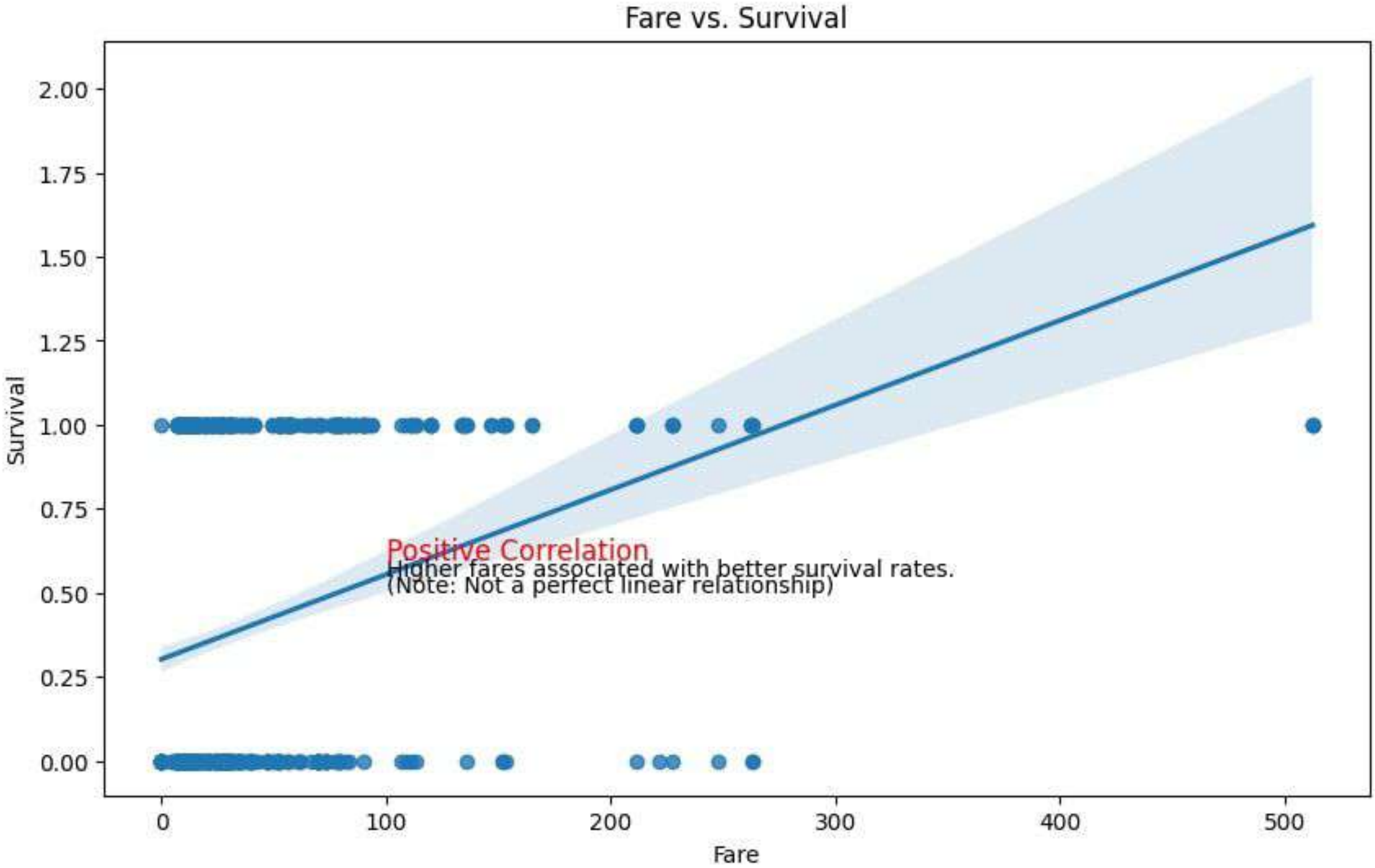



```
# Boxplot of Fare by Pclass
plt.figure(figsize=(10, 6))
sns.boxplot(x='Pclass', y='Fare', data=df)
plt.title('Fare Distribution by Passenger Class')
plt.xlabel('Passenger Class')
plt.ylabel('Fare')
plt.show()
```




```
# Visualizing correlation between Ticket Fare and Survival data.
plt.figure(figsize=(10, 6))
sns.regplot(x='Fare', y='Survived', data=df)
plt.title('Fare vs. Survival')
plt.xlabel('Fare')
plt.ylabel('Survival')

# Add annotation
plt.text(x=100, y=0.6, s="Positive Correlation", fontsize=12, color='red')
plt.text(x=100, y=0.55, s="Higher fares associated with better survival rates.", fontsize=10, color='black')
plt.text(x=100, y=0.5, s="(Note: Not a perfect linear relationship)", fontsize=10, color='black')
```


Fare vs. Survival
Survival
Fare
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
100
200
300
400
500
Positive Correlation
Higher fares associated with better survival rates.
(Note: Not a perfect linear relationship)

summary of findings

The analysis of the Titanic dataset reveals several key factors influencing survival:

Passenger Class (Pclass): First-class passengers

---

had a significantly higher survival rate compared to second and third-class passengers. This suggests class played a crucial role in access to lifeboats or preferential treatment during the evacuation.

Sex:

Women had a substantially higher survival rate than men. This aligns with the historical practice of prioritizing women and children during emergencies.

Age:

While a detailed breakdown by age group is provided, the analysis indicates a potential survival advantage for children and possibly some seniors. Adults appear to have a lower survival rate.

Embarked Location:

There are slight variations in survival rates based on the port of embarkation (C, Q, S). Further investigation is needed to determine the significance of these differences.

Fare:

A positive correlation exists between fare and survival, although it's not a perfect linear relationship. Higher fares likely corresponded to better cabins and potentially easier access to lifeboats.

Combined Factors:

The interaction of passenger class and sex strongly influences survival. The visualizations highlight the combined impact of these factors more clearly than individually.

Other Potential Factors:

The analysis suggests further investigation of sibling/spouse count (SibSp), parent/child count (Parch), and cabin information (Cabin) could uncover additional insights. The ticket information may also be worth exploring with more advanced techniques.