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

Load and Explore Data

```
df = pd.read_csv(r"D:\for internship\train.csv")
df.info()
df.isnull().sum()
for col in df.select_dtypes(include='object').columns:
    print(df[col].value_counts())
```

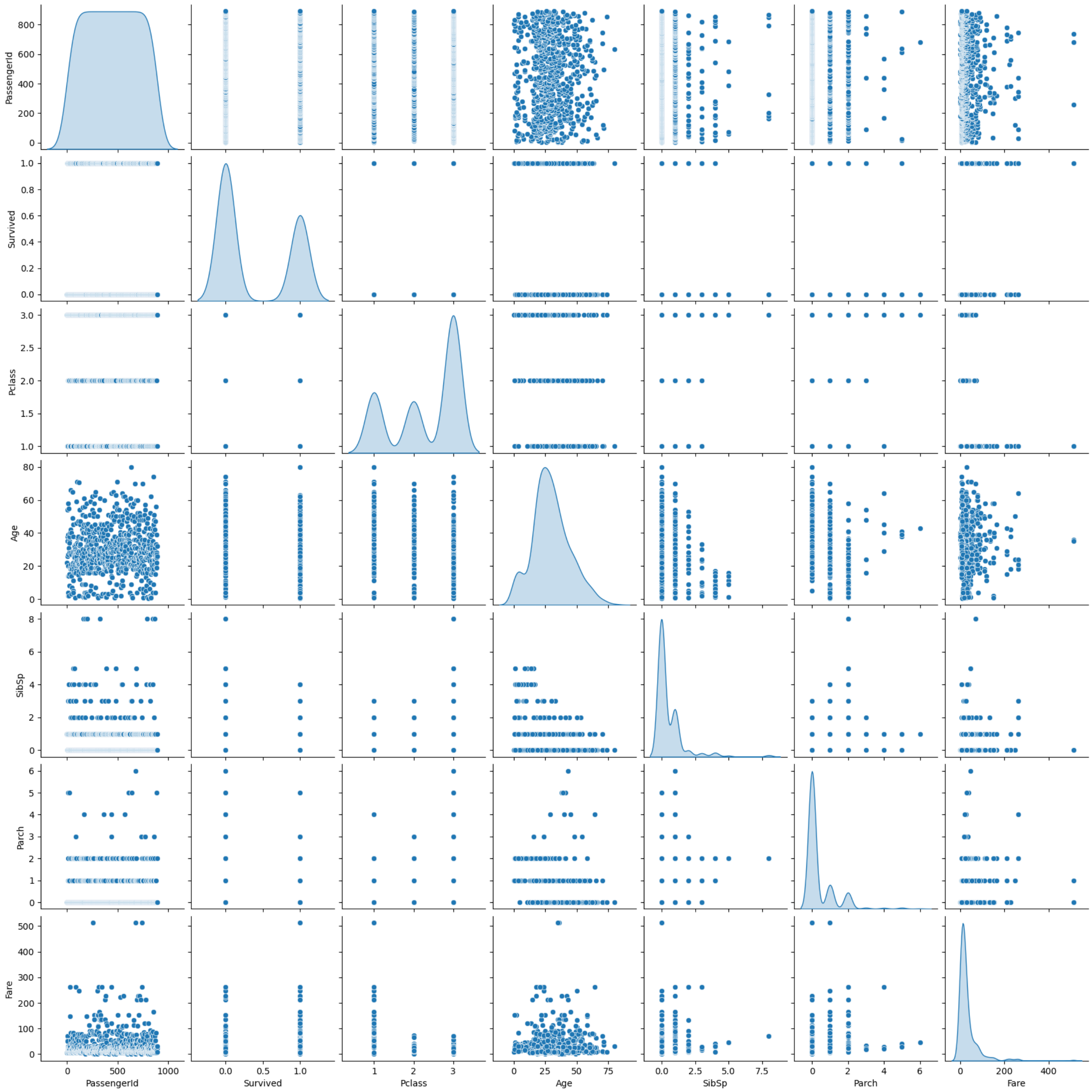
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    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
Name
Dooley, Mr. Patrick                                    1
Braund, Mr. Owen Harris                                1
Cumings, Mrs. John Bradley (Florence Briggs Thayer)    1
Heikkinen, Miss. Laina                                 1
Futrelle, Mrs. Jacques Heath (Lily May Peel)           1
                                                      ..
Hewlett, Mrs. (Mary D Kingcome)                        1
Vestrom, Miss. Hulda Amanda Adolfina                   1
Andersson, Mr. Anders Johan                            1
Saundercock, Mr. William Henry                         1
Bonnell, Miss. Elizabeth                               1
Name: count, Length: 891, dtype: int64
Sex
male      577
female    314
Name: count, dtype: int64
```

```
Ticket
347082              7
1601                7
CA. 2343            7
3101295             6
CA 2144             6
                   ..
PC 17590            1
17463               1
330877              1
373450              1
STON/O2. 3101282    1
Name: count, Length: 681, dtype: int64
Cabin
G6             4
C23 C25 C27    4
B96 B98        4
F2             3
D              3
              ..
E17            1
A24            1
C50            1
B42            1
C148           1
Name: count, Length: 147, dtype: int64
Embarked
S    644
C    168
Q     77
Name: count, dtype: int64
```

Visual Exploration

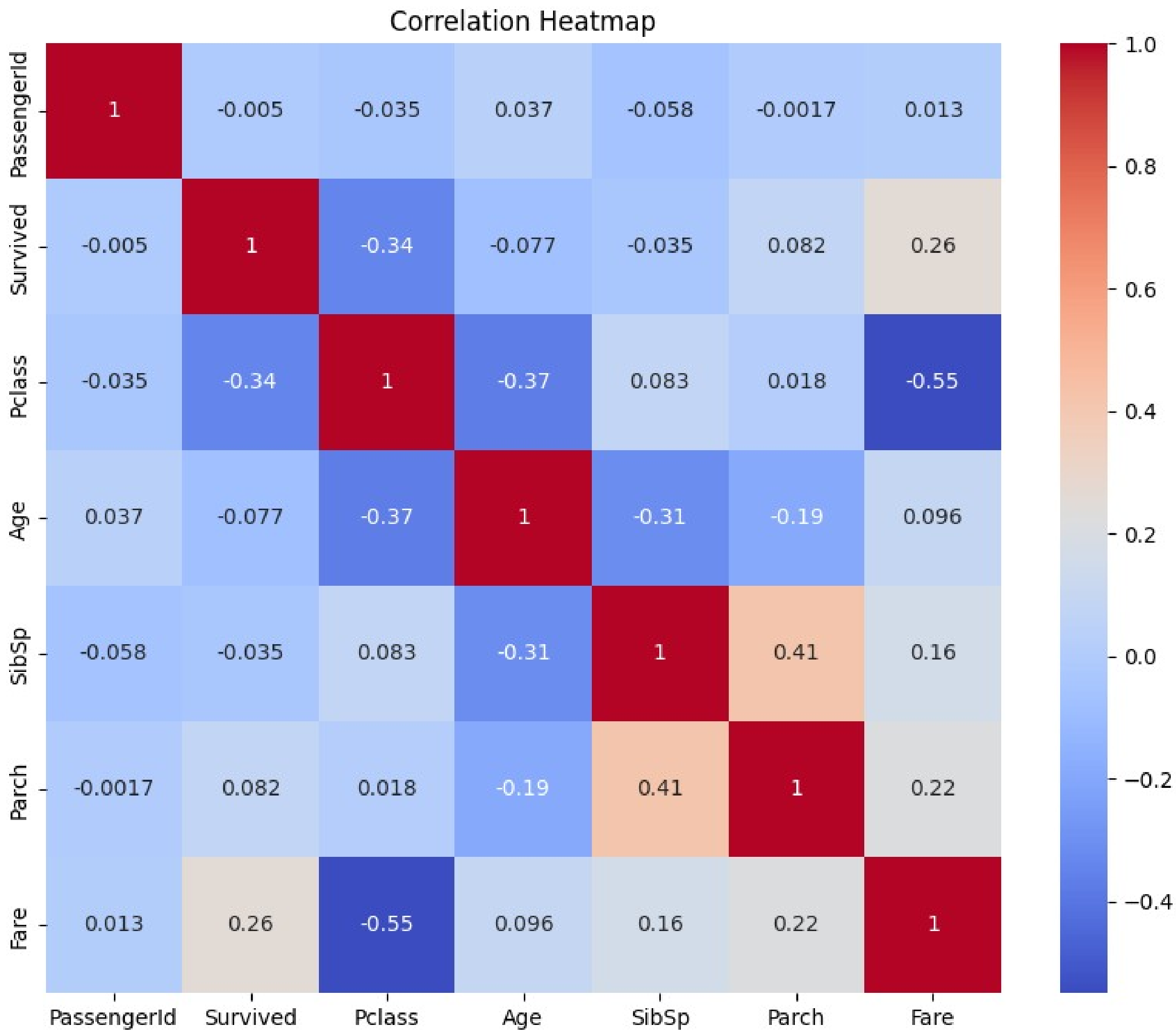
Pairplot for relationships

```
sns.pairplot(df, diag_kind='kde')
plt.show()
```



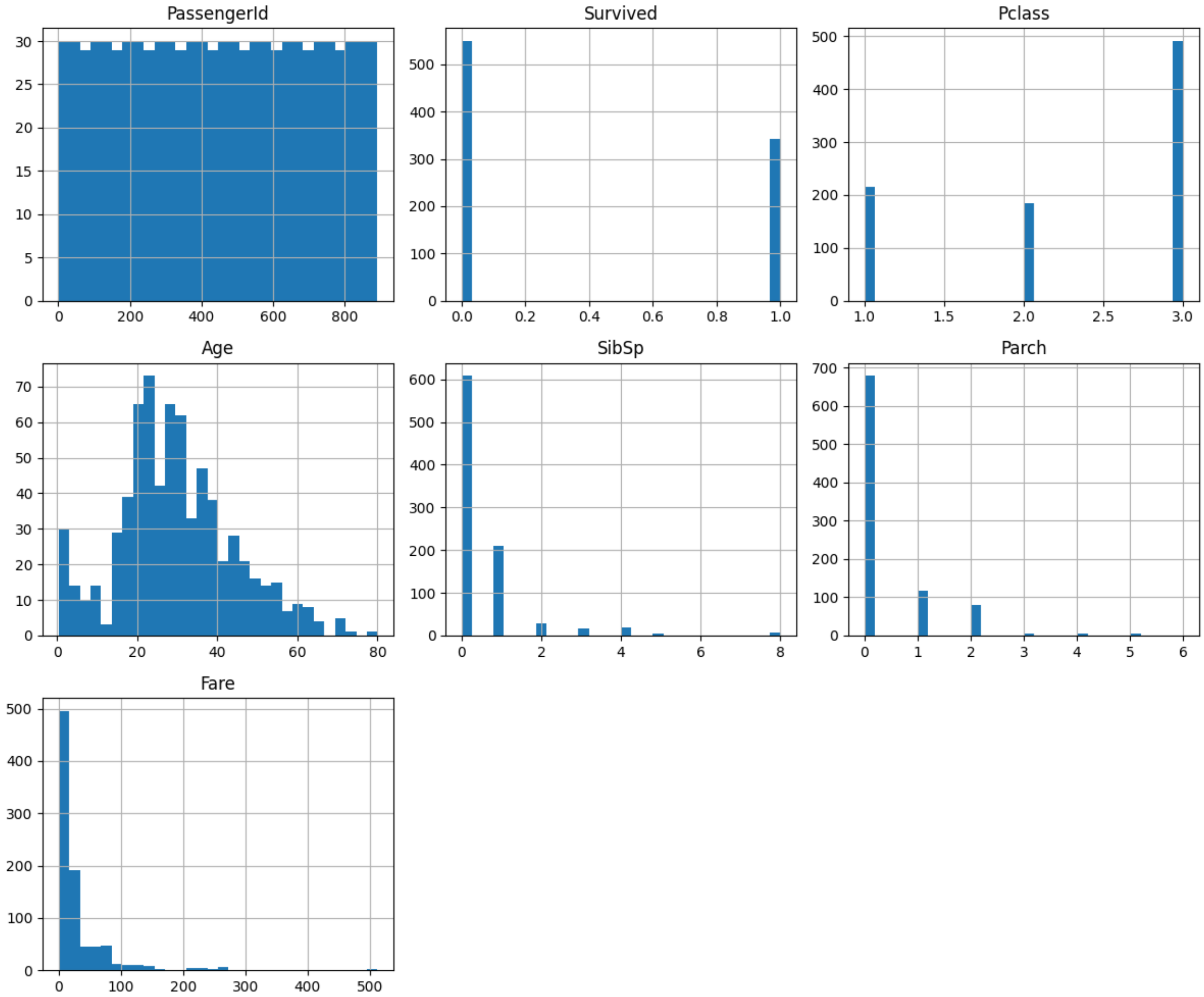
Correlation Heatmap

```
numeric_df = df.select_dtypes(include='number')
plt.figure(figsize=(10,8))
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



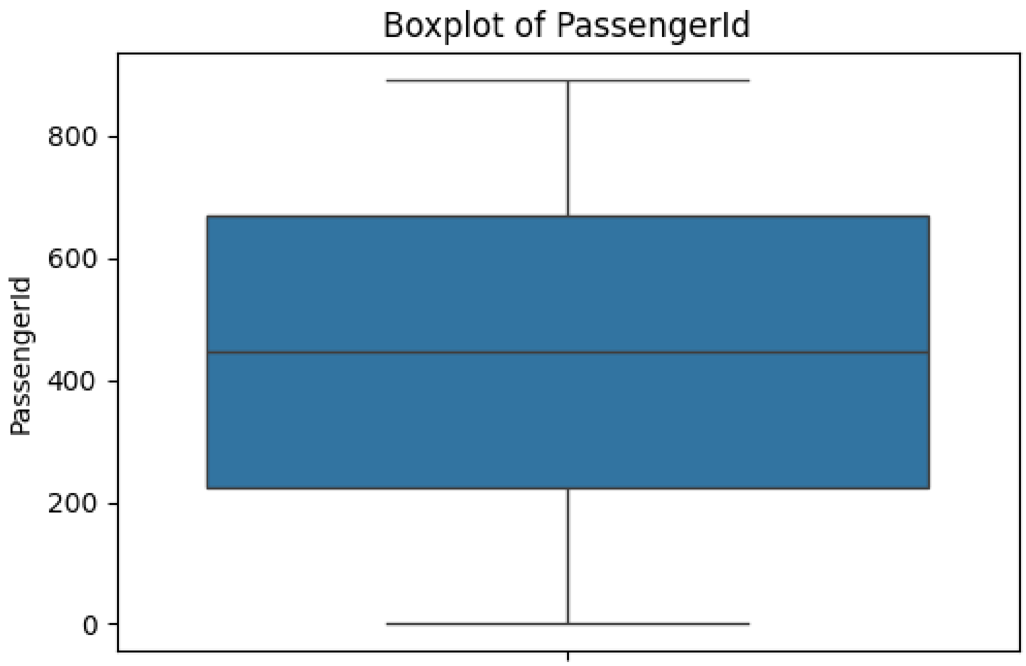
Histograms

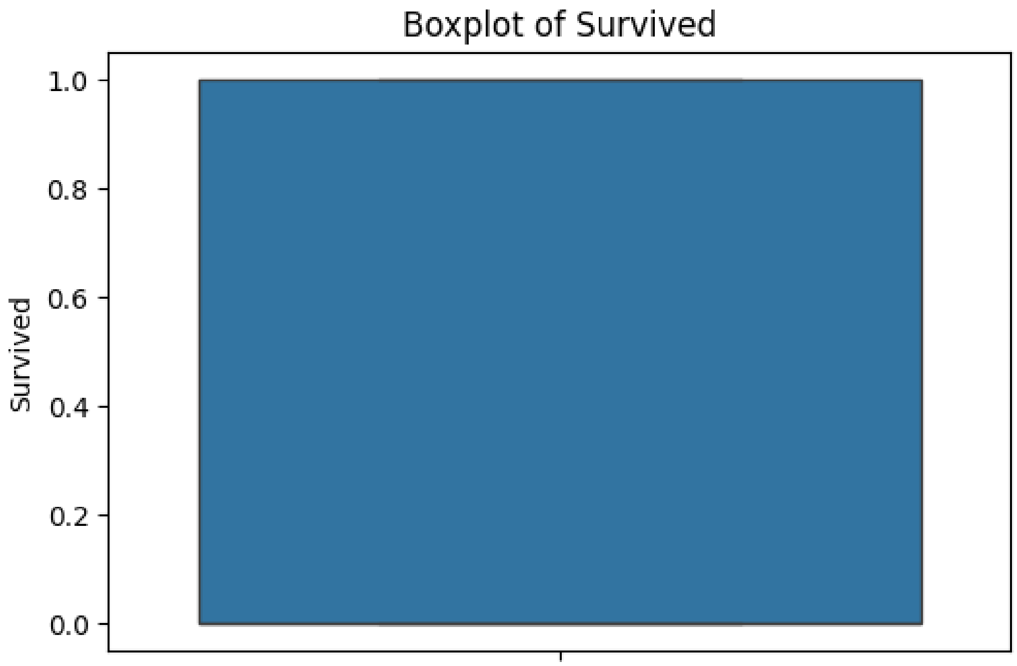
```
df.hist(figsize=(12,10), bins=30)
plt.tight_layout()
plt.show()
```

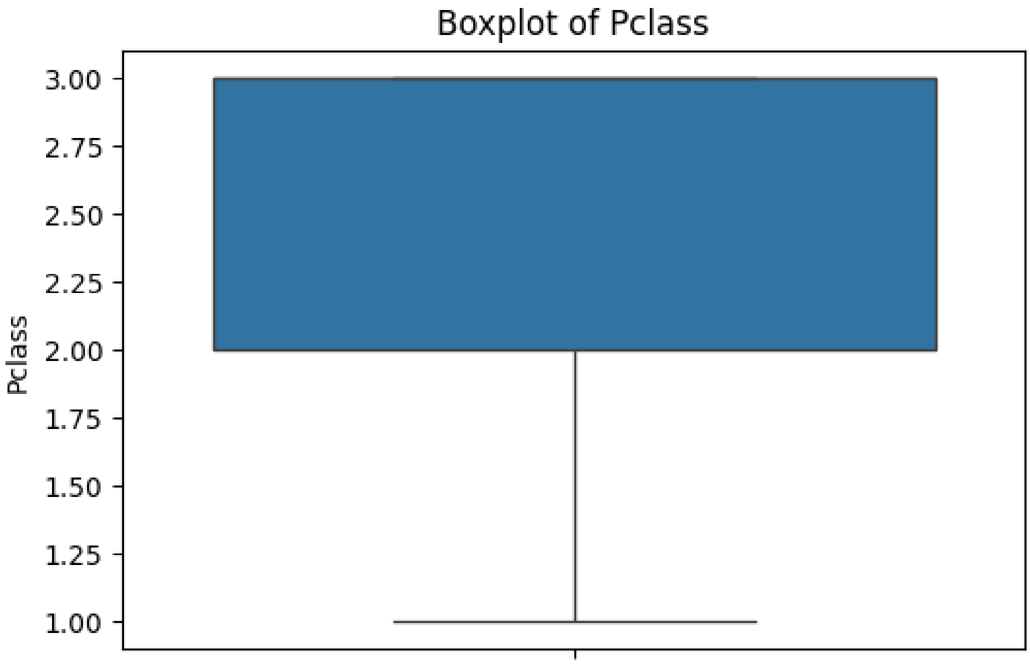


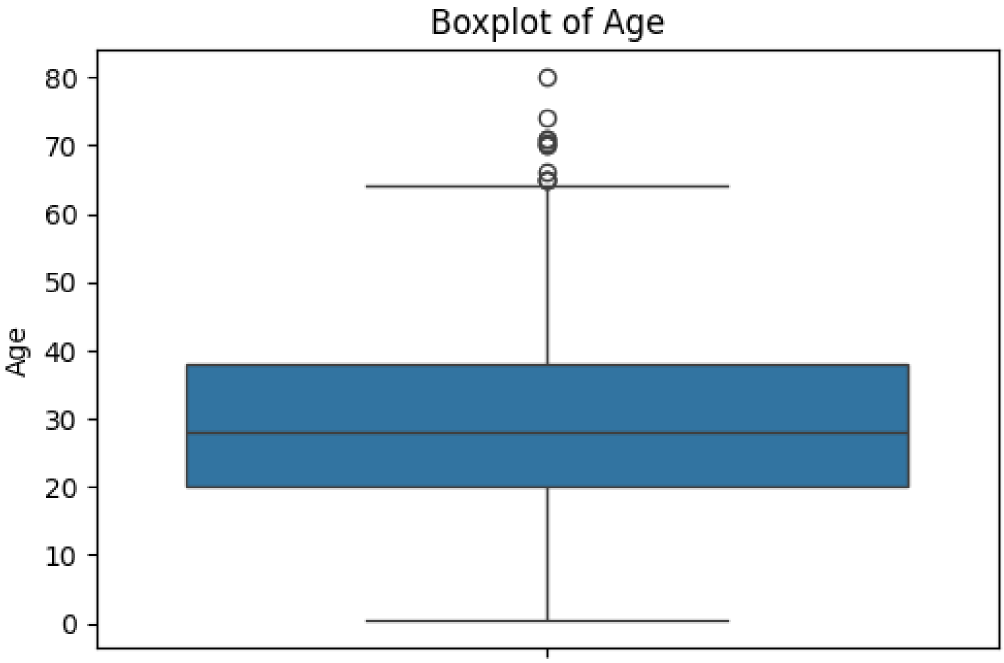
Boxplots

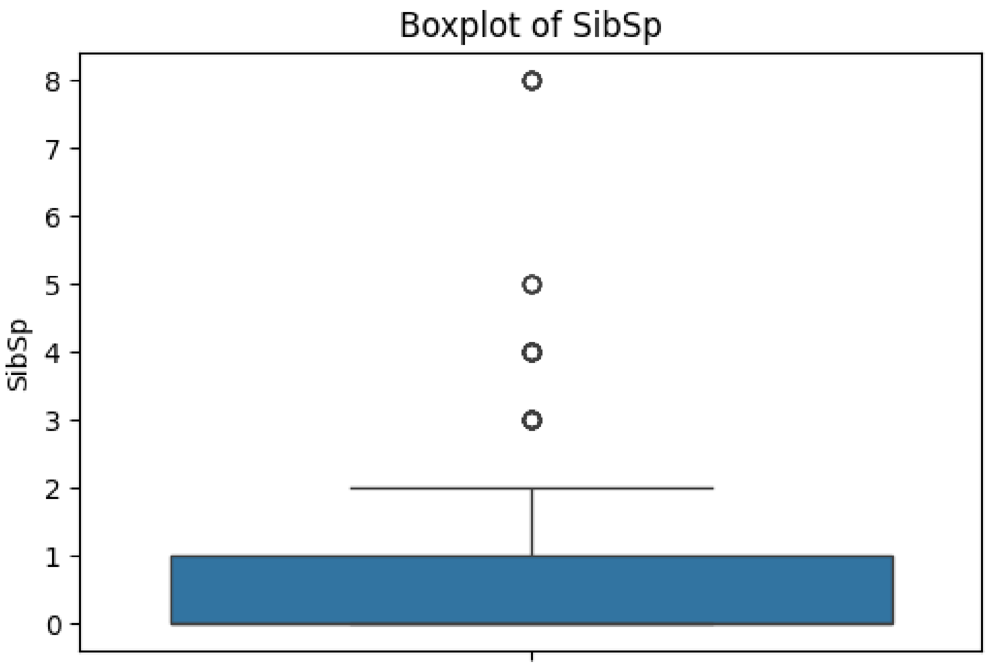
```
for col in df.select_dtypes(include='number').columns:
    plt.figure(figsize=(6,4))
    sns.boxplot(y=df[col])
    plt.title(f'Boxplot of {col}')
    plt.show()
```

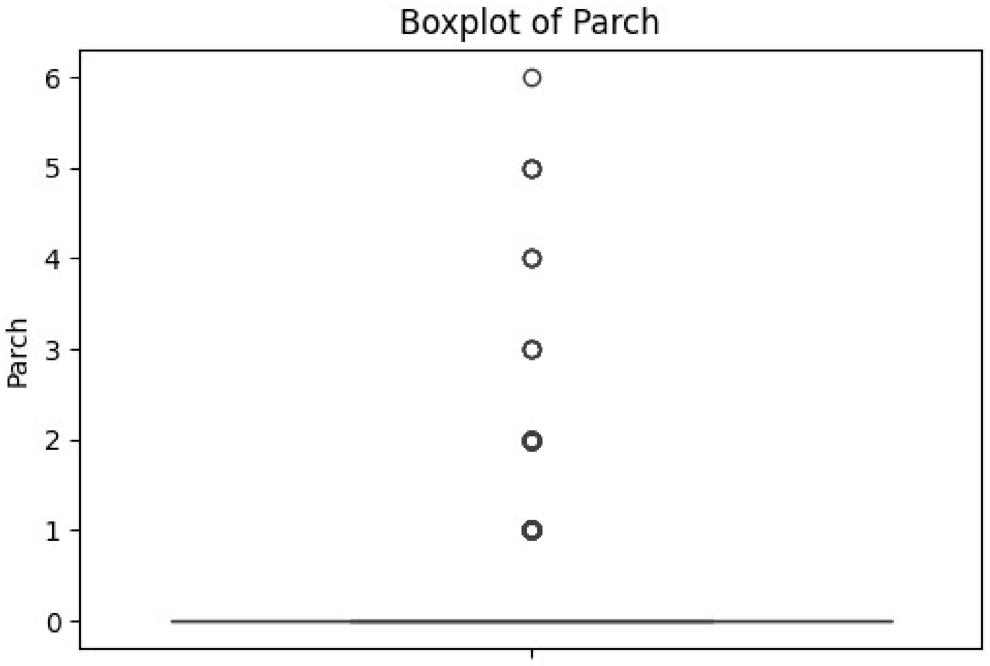

Boxplot of PassengerId
PassengerId
800
600
400
200
0

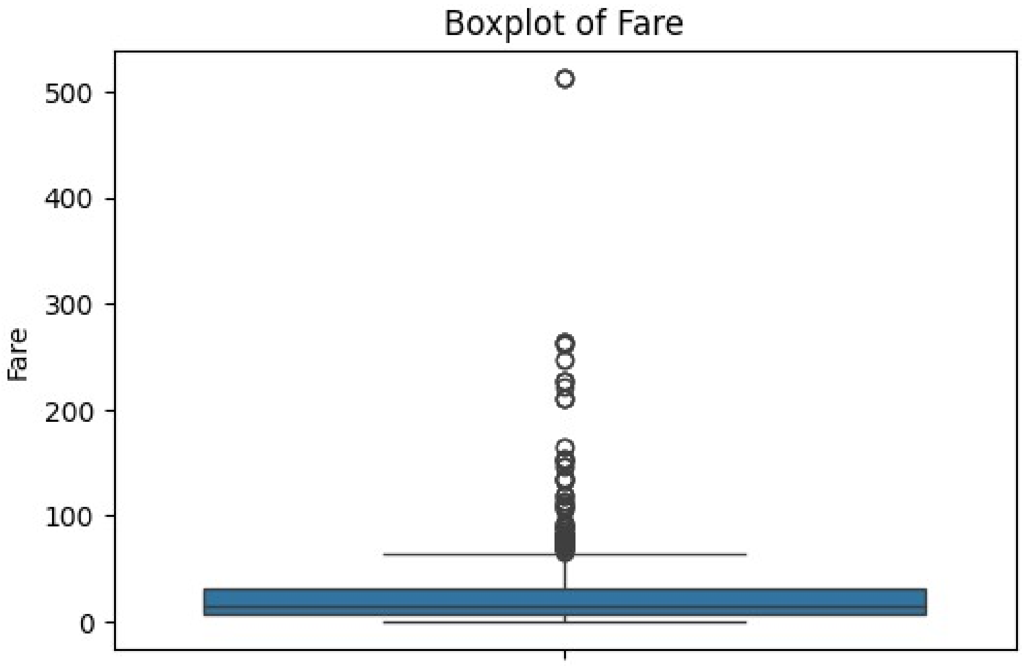

Boxplot of Survived
Survived
1.0
0.8
0.6
0.4
0.2
0.0


Boxplot of Pclass
Pclass
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00


Boxplot of Age
Age
80
70
60
50
40
30
20
10
0

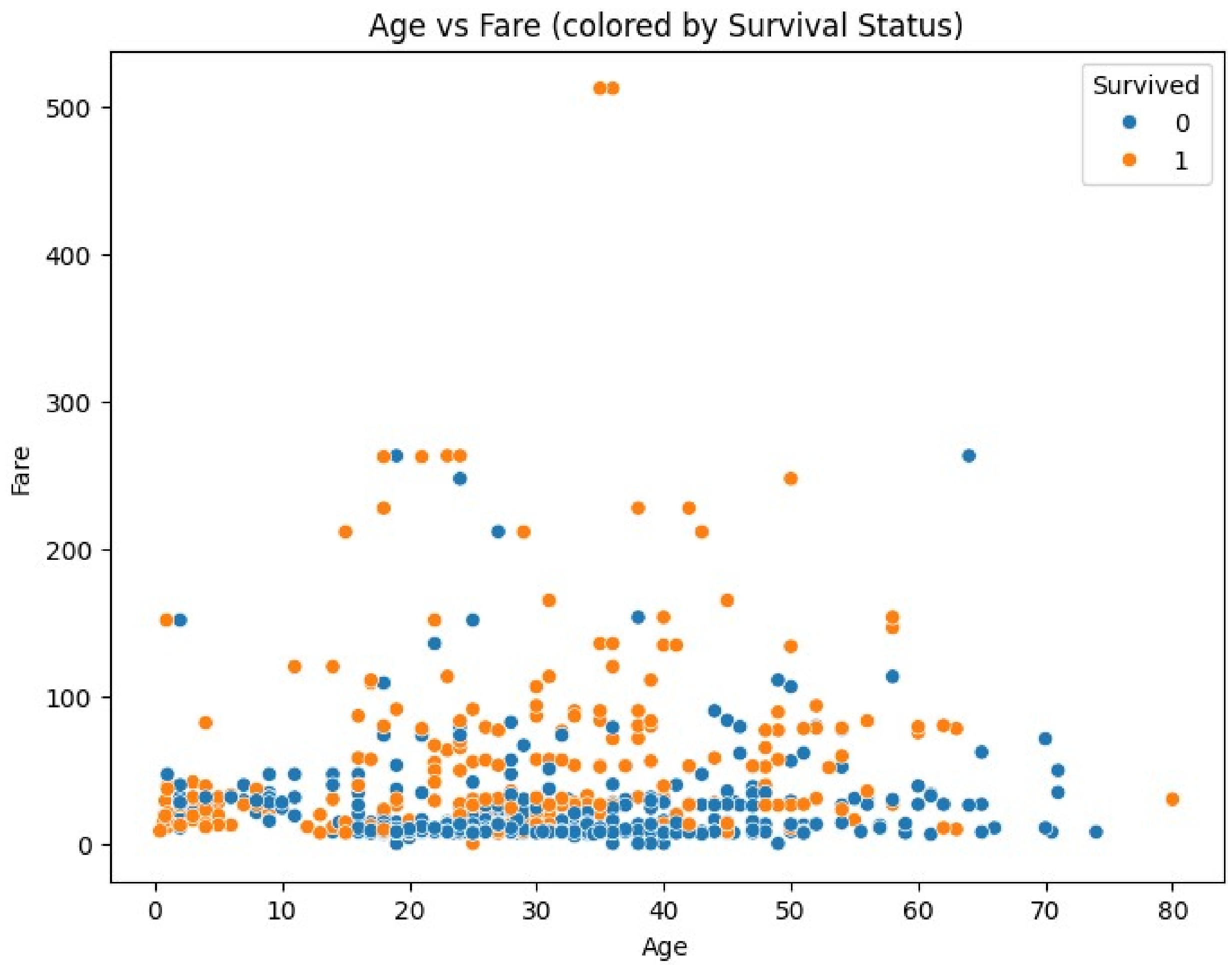

Boxplot of SibSp
SibSp
0
1
2
3
4
5
6
7
8


Boxplot of Parch
Parch
0
1
2
3
4
5
6



Scatterplots for significant variable pairs

```
plt.figure(figsize=(8,6))
sns.scatterplot(x='Age', y='Fare', data=df, hue='Survived')
plt.title('Age vs Fare (colored by Survival Status)')
plt.show()
```



□ Observation for Correlation Heatmap

- Strong positive correlation between `SibSp` and `Parch`, indicating family travel groups.
- Negative correlation between `Pclass` and `Fare`, as 1st class passengers paid higher fares.
- No significant correlation between `Age` and other numeric features.

□ Observation for Age Histogram

- Age distribution is right-skewed.
- Most passengers are aged 20–40, with fewer older passengers.

Observation for Fare Boxplot

- Fares in 1st class are significantly higher and more variable.

- Outliers in fare values are more common in 1st class.

□ Observation for Scatterplot (Age vs Fare, colored by Survival)

- Older and younger passengers had slightly better survival rates.
- Higher fares (1st class) are associated with higher survival.

□ Observation for Scatterplot (Age vs Fare, colored by Survival Status)

- Most passengers in the 20–40 age range paid lower fares (likely 3rd class).
- Higher fares are associated with older passengers (1st class).
- Surviving passengers (colored differently) are more frequent at higher fares, indicating higher survival rates for 1st class.
- There is no strong linear relationship between Age and Fare, but outliers suggest some older, wealthy passengers.