ASSIGNMENT – WEEK 4

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

# Use Seaborn's clean style
sns.set(style="whitegrid")

# 1. Load Titanic dataset
df = sns.load_dataset("titanic")

# 2. Show basic info
print("🔹 First 5 rows of the dataset:")
print(df.head(), "\n")

print("🔹 Dataset shape (rows, columns):", df.shape, "\n")

print("🔹 Missing values per column:")
print(df.isnull().sum(), "\n")

# 3. Plot survival distribution
plt.figure(figsize=(6, 4))
sns.countplot(x="survived", data=df)
plt.title("Survival Count (0 = No, 1 = Yes)")
plt.xlabel("Survived")
plt.ylabel("Number of Passengers")
plt.show()
```

```
# 4. Age distribution
plt.figure(figsize=(8, 4))
sns.histplot(df["age"].dropna(), kde=True, bins=30)
plt.title("Passenger Age Distribution")
plt.xlabel("Age")
plt.show()

# 5. Describe age stats
print("🔹 Age statistics:\n", df["age"].describe(), "\n")

# 6. Boxplots for outliers (Age & Fare by Survival)
fig, axes = plt.subplots(1, 2, figsize=(12, 5))
sns.boxplot(x='survived', y='age', data=df, ax=axes[0])
axes[0].set_title("Age vs Survival")

sns.boxplot(x='survived', y='fare', data=df, ax=axes[1])
axes[1].set_title("Fare vs Survival")

plt.tight_layout()
plt.show()

# 7. Survival by passenger class
plt.figure(figsize=(8, 4))
sns.countplot(x="pclass", hue="survived", data=df)
plt.title("Survival by Passenger Class")
plt.xlabel("Class")
plt.ylabel("Number of Passengers")
plt.legend(title="Survived", labels=["No", "Yes"])
plt.show()
```

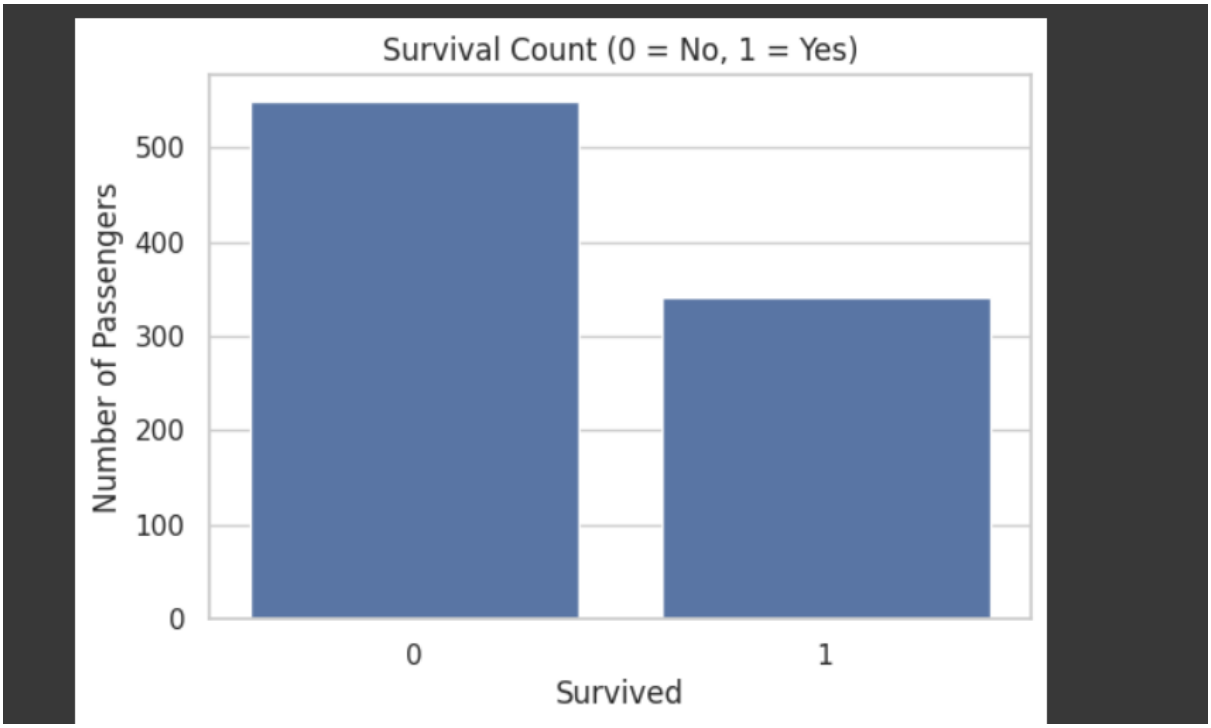
```
# 8. Survival by gender
plt.figure(figsize=(8, 4))
sns.countplot(x="sex", hue="survived", data=df)
plt.title("Survival by Sex")
plt.xlabel("Sex")
plt.ylabel("Number of Passengers")
plt.legend(title="Survived", labels=["No", "Yes"])
plt.show()

# 9. Correlation heatmap
plt.figure(figsize=(6, 5))
numeric_df = df[['survived', 'pclass', 'age', 'sibsp', 'parch', 'fare']]
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title("Feature Correlation Heatmap")
plt.show()


# 10. Survival by Embarked location
plt.figure(figsize=(8, 4))
sns.countplot(x="embarked", hue="survived", data=df)
plt.title("Survival by Embarkation Port")
plt.xlabel("Embarked")
plt.ylabel("Number of Passengers")
plt.legend(title="Survived", labels=["No", "Yes"])
plt.show()

# 11. Pairplot of numeric features
sns.pairplot(numeric_df)
```

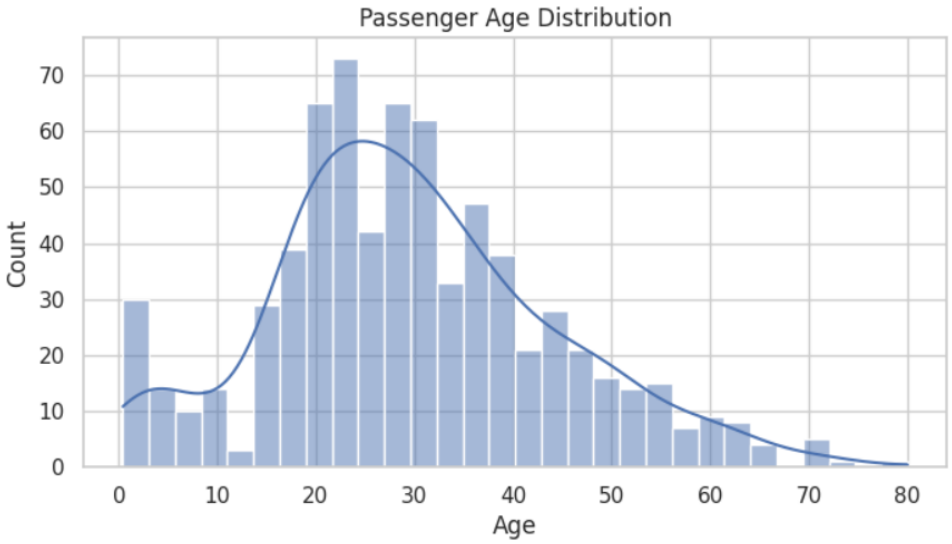
```
plt.suptitle("Pairplot of Numeric Features", y=1.02)
plt.show()


print("\nFINAL EDA SUMMARY")
print("- Dataset has", df.shape[0], "rows and", df.shape[1], "columns.")
print("- Missing values found in: age, embarked, deck, and embarked_town.")
print("- Age and Fare show outliers — visible in box plots.")
print("- More passengers died (0) than survived (1).")
print("- Females had a higher survival rate than males.")
print("- Passengers in Class 1 had better survival chances than those in Class 3.")
print("- Port of embarkation and fare show some influence on survival.")
print("- Correlation heatmap reveals modest relationships: fare vs survival, class vs survival.")
print("- Dataset is useful for classification tasks like predicting survival.")
```

OUTPUT:

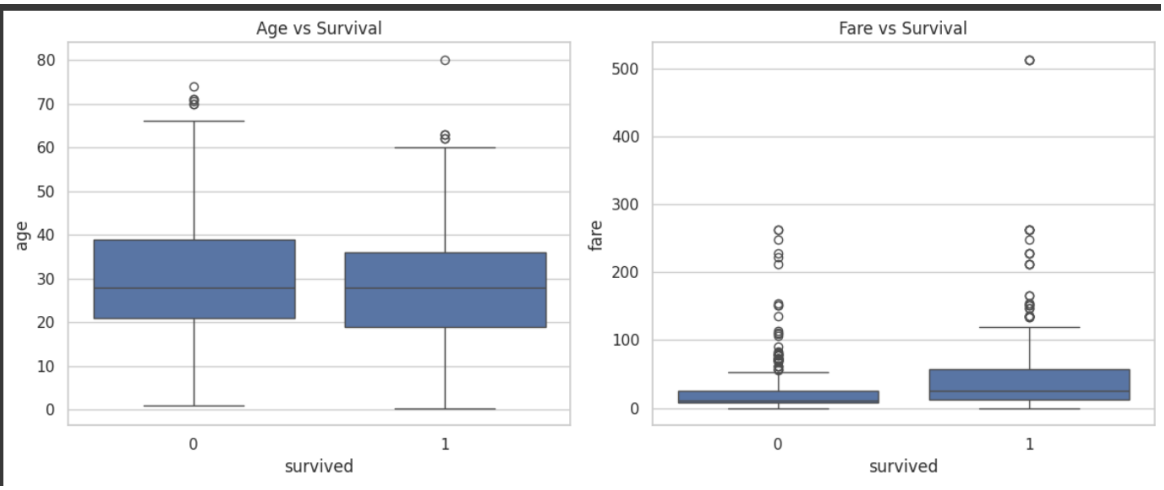
```
🔹 First 5 rows of the dataset:
   survived  pclass     sex   age  sibsp  parch     fare embarked  class  \
0         0       3    male  22.0      1      0   7.2500        S  Third
1         1       1  female  38.0      1      0  71.2833        C  First
2         1       3  female  26.0      0      0   7.9250        S  Third
3         1       1  female  35.0      1      0  53.1000        S  First
4         0       3    male  35.0      0      0   8.0500        S  Third

     who  adult_male deck  embark_town alive  alone
0    man        True  NaN  Southampton    no  False
1  woman       False    C    Cherbourg   yes  False
2  woman       False  NaN  Southampton   yes   True
3  woman       False    C  Southampton   yes  False
4    man        True  NaN  Southampton    no   True

🔹 Dataset shape (rows, columns): (891, 15)
```

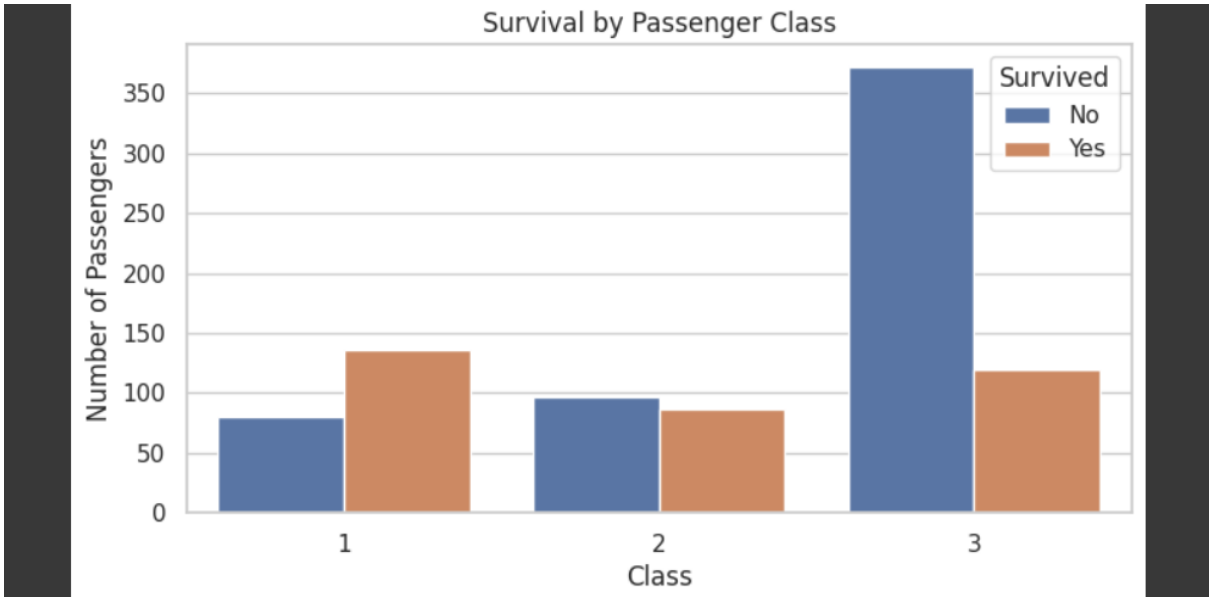
```
🔹 Missing values per column:
survived         0
pclass           0
sex              0
age            177
sibsp            0
parch            0
fare             0
embarked         2
class            0
who              0
adult_male       0
deck           688
embark_town      2
alive            0
alone            0
dtype: int64
```

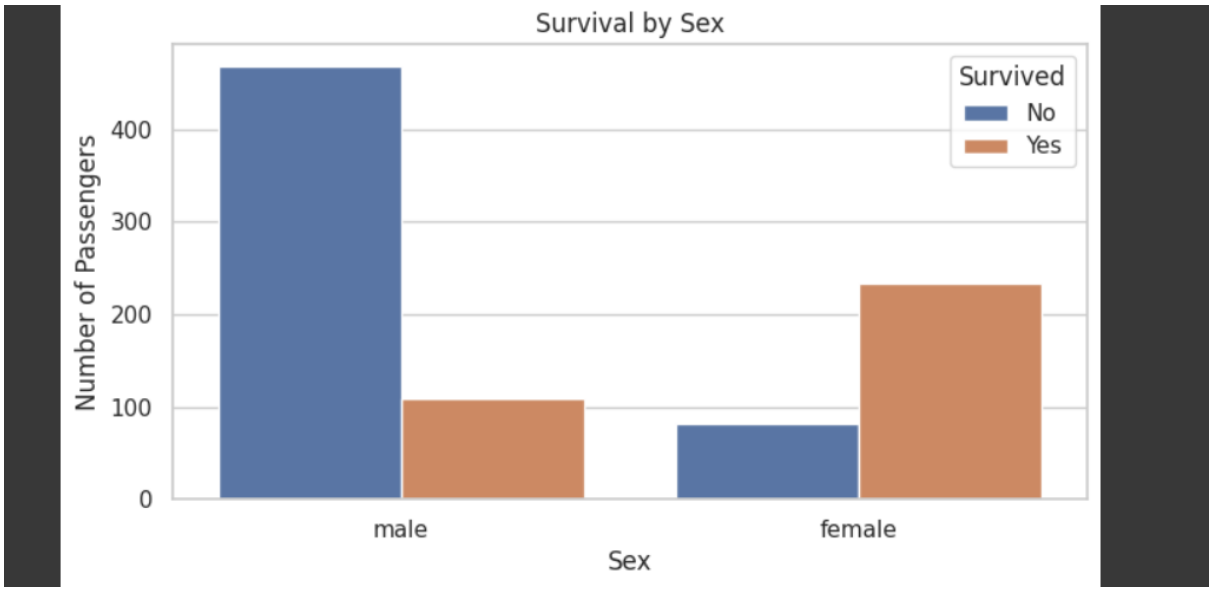


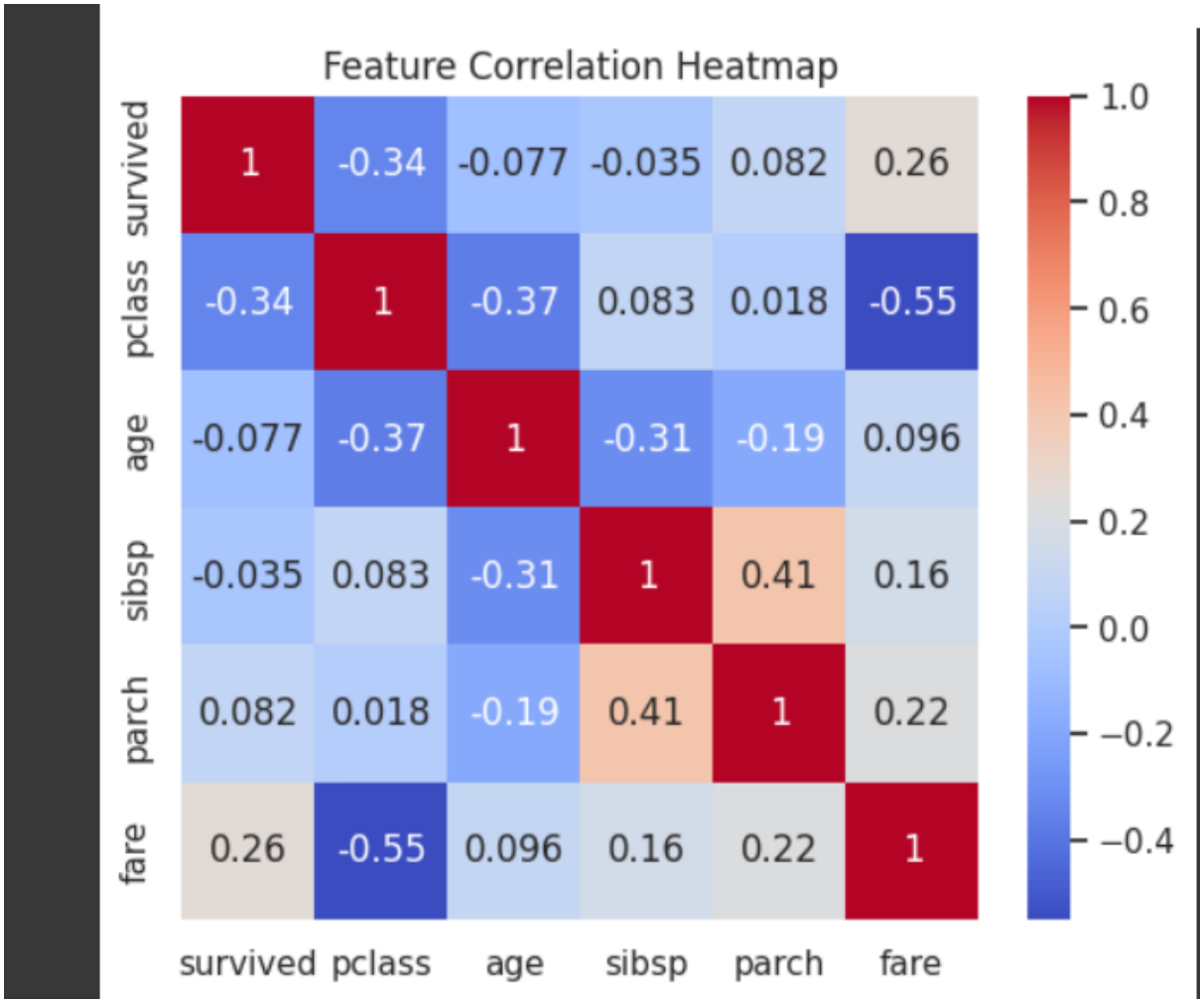


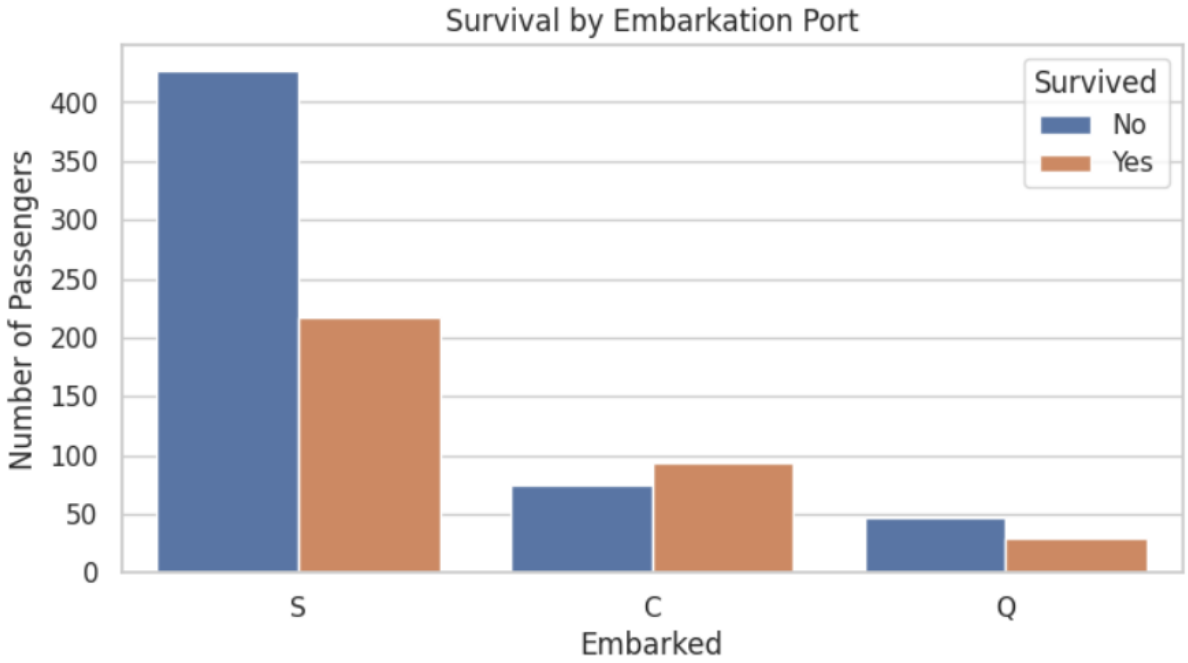

```
🔹 Age statistics:
count    714.000000
mean      29.699118
std       14.526497
min        0.420000
25%       20.125000
50%       28.000000
75%       38.000000
max       80.000000
Name: age, dtype: float64
```

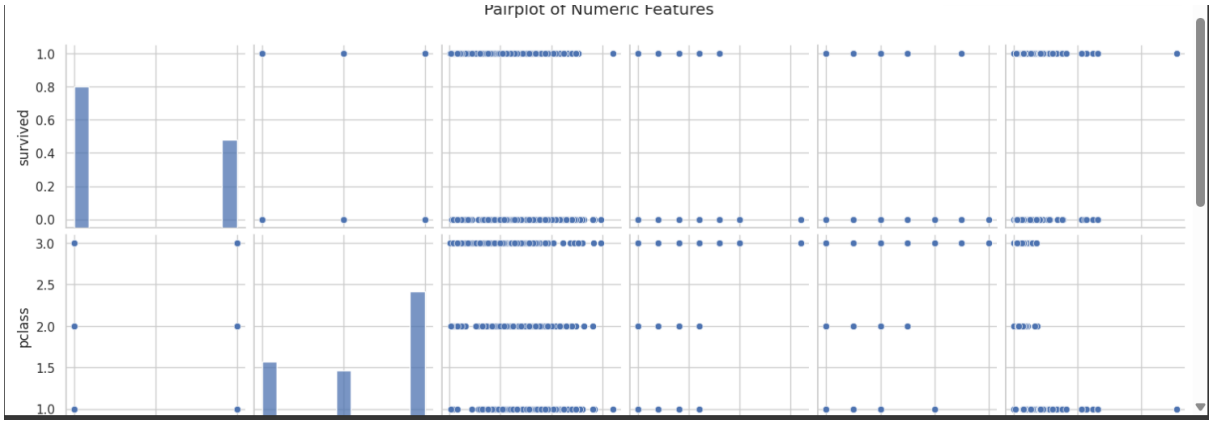


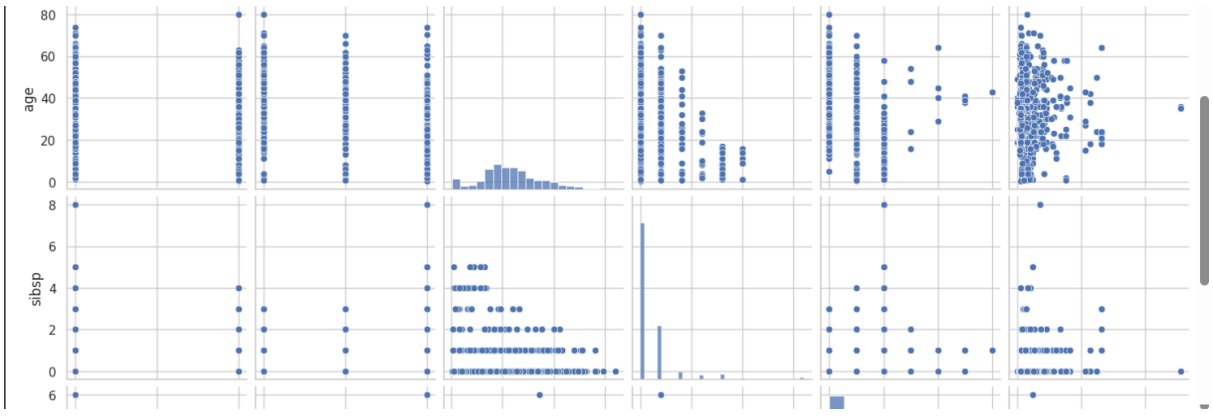

Survival by Passenger Class
Survived
No
Yes
Number of Passengers
0
50
100
150
200
250
300
350
1
2
3
Class

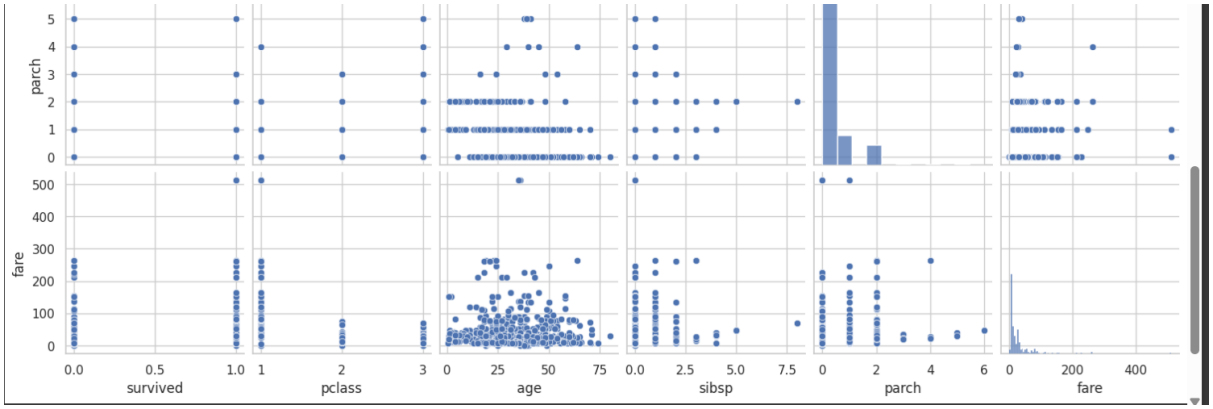

Survival by Sex
Survived
No
Yes
Number of Passengers
0
100
200
300
400
male
female
Sex


Feature Correlation Heatmap
survived
pclass
age
sibsp
parch
fare
1
-0.34
-0.077
-0.035
0.082
0.26
-0.34
1
-0.37
0.083
0.018
-0.55
-0.077
-0.37
1
-0.31
-0.19
0.096
-0.035
0.083
-0.31
1
0.41
0.16
0.082
0.018
-0.19
0.41
1
0.22
0.26
-0.55
0.096
0.16
0.22
1
survived
pclass
age
sibsp
parch
fare
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Survival by Embarkation Port
Survived
No
Yes
Number of Passengers
0
50
100
150
200
250
300
350
400
S
C
Q
Embarked


Pairplot of Numeric Features
survived
pclass
1.0
0.8
0.6
0.4
0.2
0.0
3.0
2.5
2.0
1.5
1.0


age
sibsp
80
60
40
20
0
8
6
4
2
0
6




```
FINAL EDA SUMMARY
- Dataset has 891 rows and 15 columns.
- Missing values found in: age, embarked, deck, and embarked_town.
- Age and Fare show outliers – visible in box plots.
- More passengers died (0) than survived (1).
- Females had a higher survival rate than males.
- Passengers in Class 1 had better survival chances than those in Class 3.
- Port of embarkation and fare show some influence on survival.
- Correlation heatmap reveals modest relationships: fare vs survival, class vs survival.
- Dataset is useful for classification tasks like predicting survival.
```