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

sns.set(style="whitegrid")
plt.rcParams['figure.figsize'] = (10, 6)
```

Load Titanic dataset (ensure titanic.csv is in the same directory)

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

```
# Load Titanic dataset (ensure titanic.csv is in the same directory)
df = pd.read_csv("/content/Titanic-Dataset.csv")
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 2117
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 1759
2	3	1	3	Heikkinen, Miss	female	26.0	0	0	STON/O2

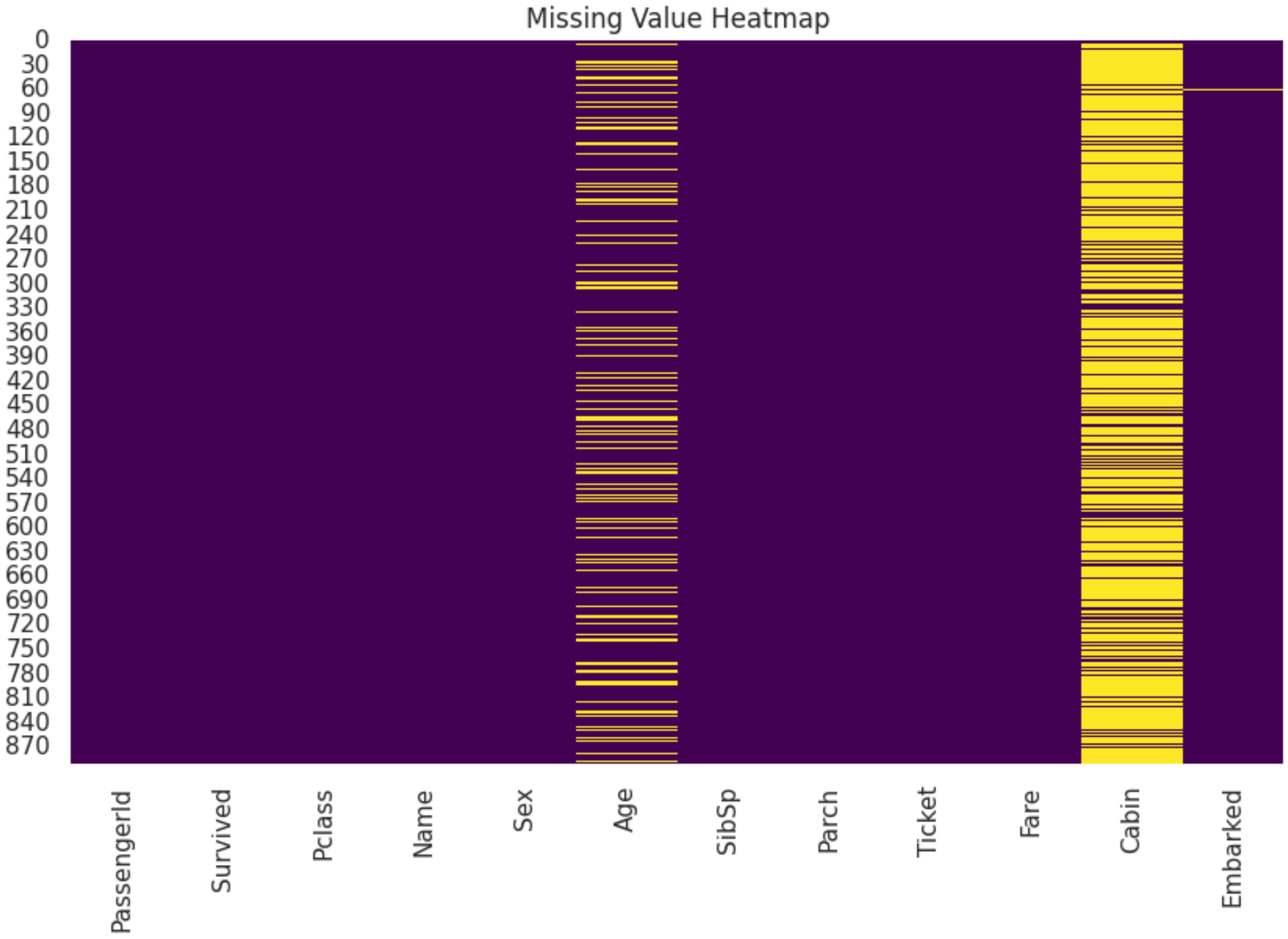
Next steps: Generate code with df | View recommended plots | New interactive sheet

```
# Shape and basic info
print("Shape of dataset:", df.shape)
df.info()

# Summary statistics
df.describe(include='all')
```

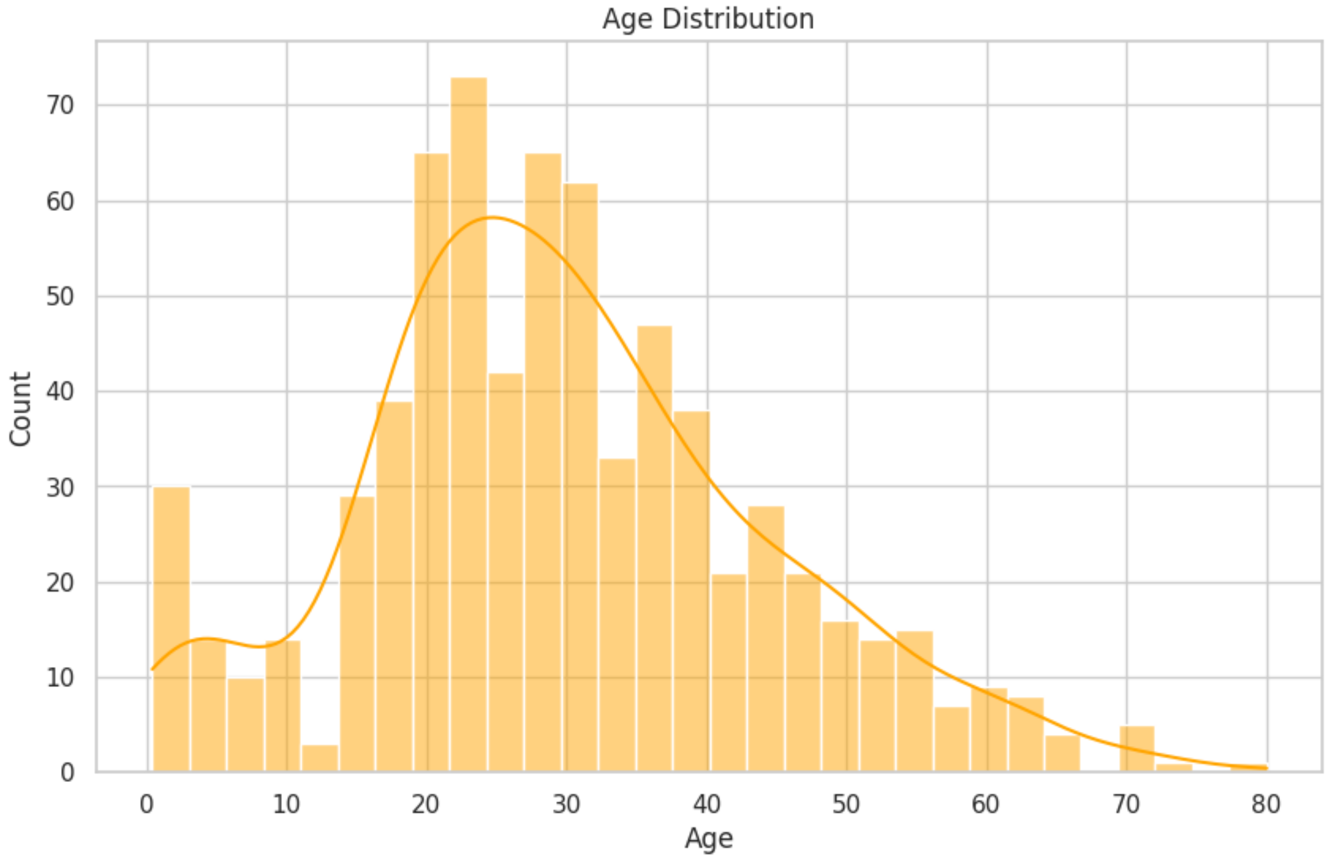
```
Shape of dataset: (891, 12)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64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

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp
count	891.000000	891.000000	891.000000	891	891	714.000000	891.000000
unique	NaN	NaN	NaN	891	2	NaN	NaN
top	NaN	NaN	NaN	Dooley, Mr. Patrick	male	NaN	NaN
freq	NaN	NaN	NaN	1	577	NaN	NaN
mean	446.000000	0.383838	2.308642	NaN	NaN	29.699118	0.523008
std	257.353842	0.486592	0.836071	NaN	NaN	14.526497	1.102743
min	1.000000	0.000000	1.000000	NaN	NaN	0.420000	0.000000
25%	223.500000	0.000000	2.000000	NaN	NaN	20.125000	0.000000
50%	446.000000	0.000000	3.000000	NaN	NaN	28.000000	0.000000
75%	668.500000	1.000000	3.000000	NaN	NaN	38.000000	1.000000
max	891.000000	1.000000	3.000000	NaN	NaN	80.000000	8.000000

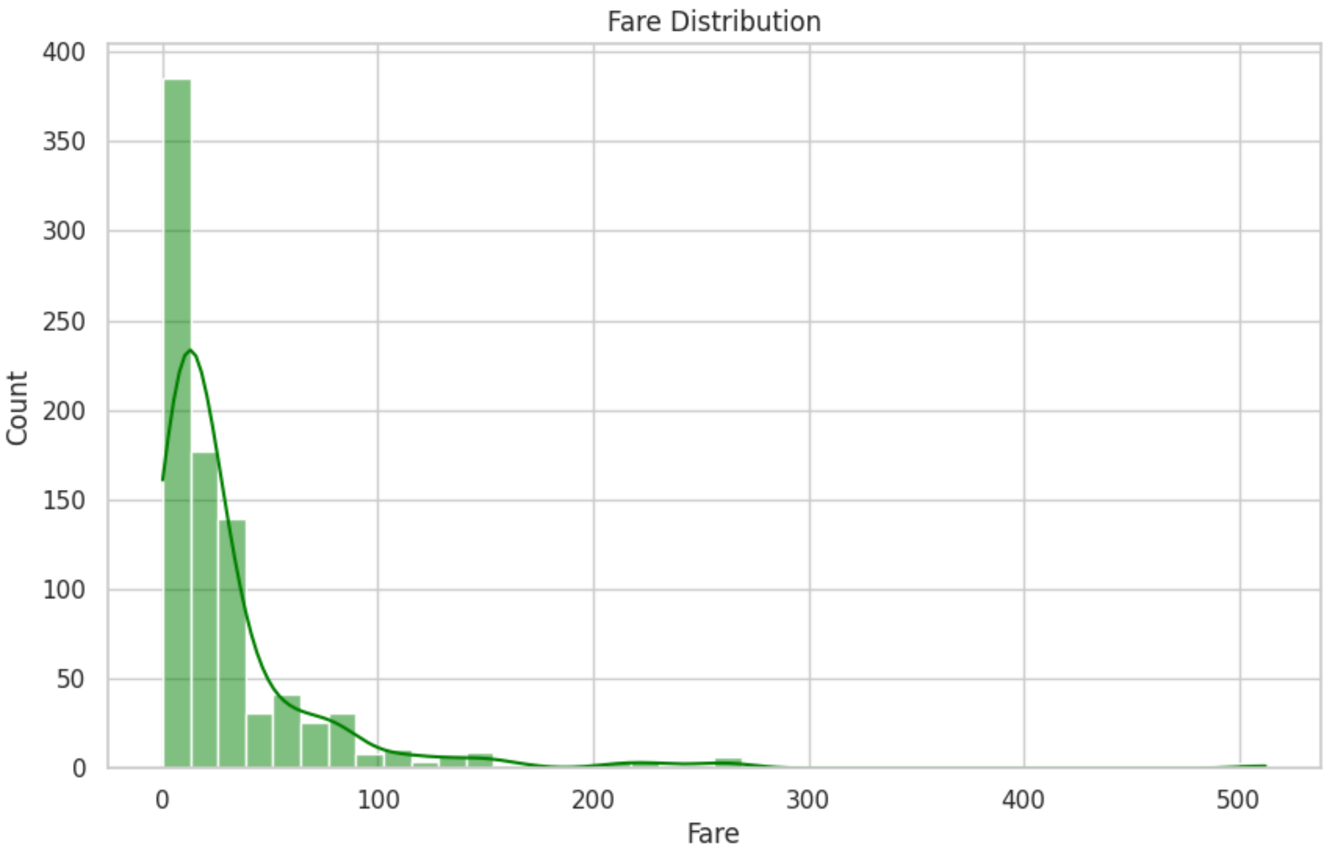
```
# Count of missing values
df.isnull().sum()

# Visualize missing values
sns.heatmap(df.isnull(), cbar=False, cmap='viridis')
plt.title("Missing Value Heatmap")
plt.show()
```



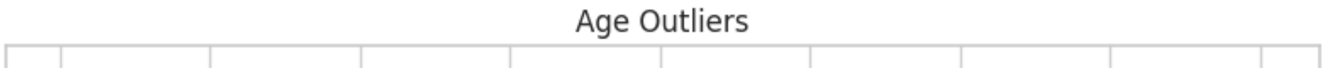
```
sns.histplot(df['Age'].dropna(), kde=True, bins=30, color='orange')
plt.title("Age Distribution")
plt.xlabel("Age")
plt.show()
```

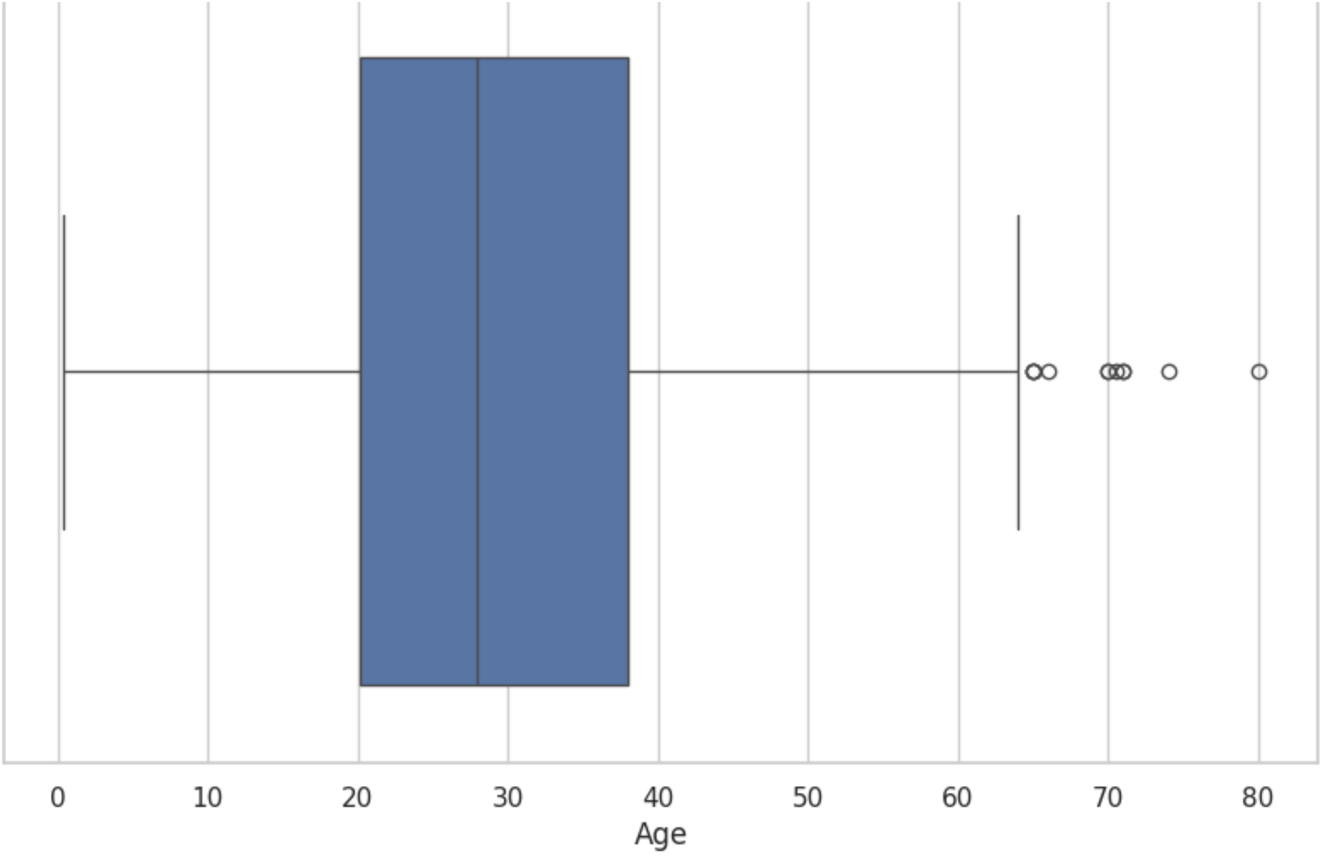


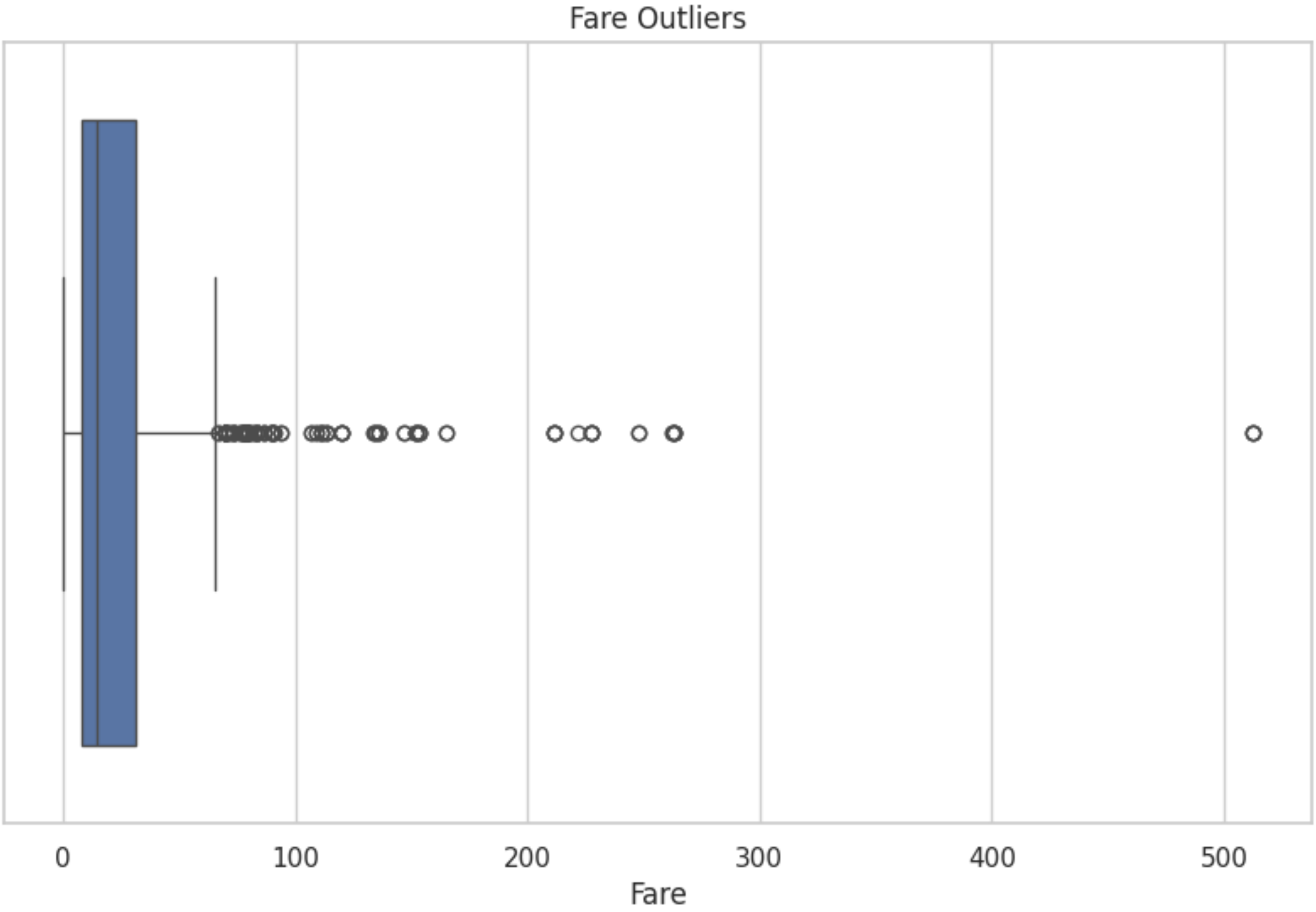
```
sns.histplot(df['Fare'], kde=True, bins=40, color='green')
plt.title("Fare Distribution")
plt.xlabel("Fare")
plt.show()
```



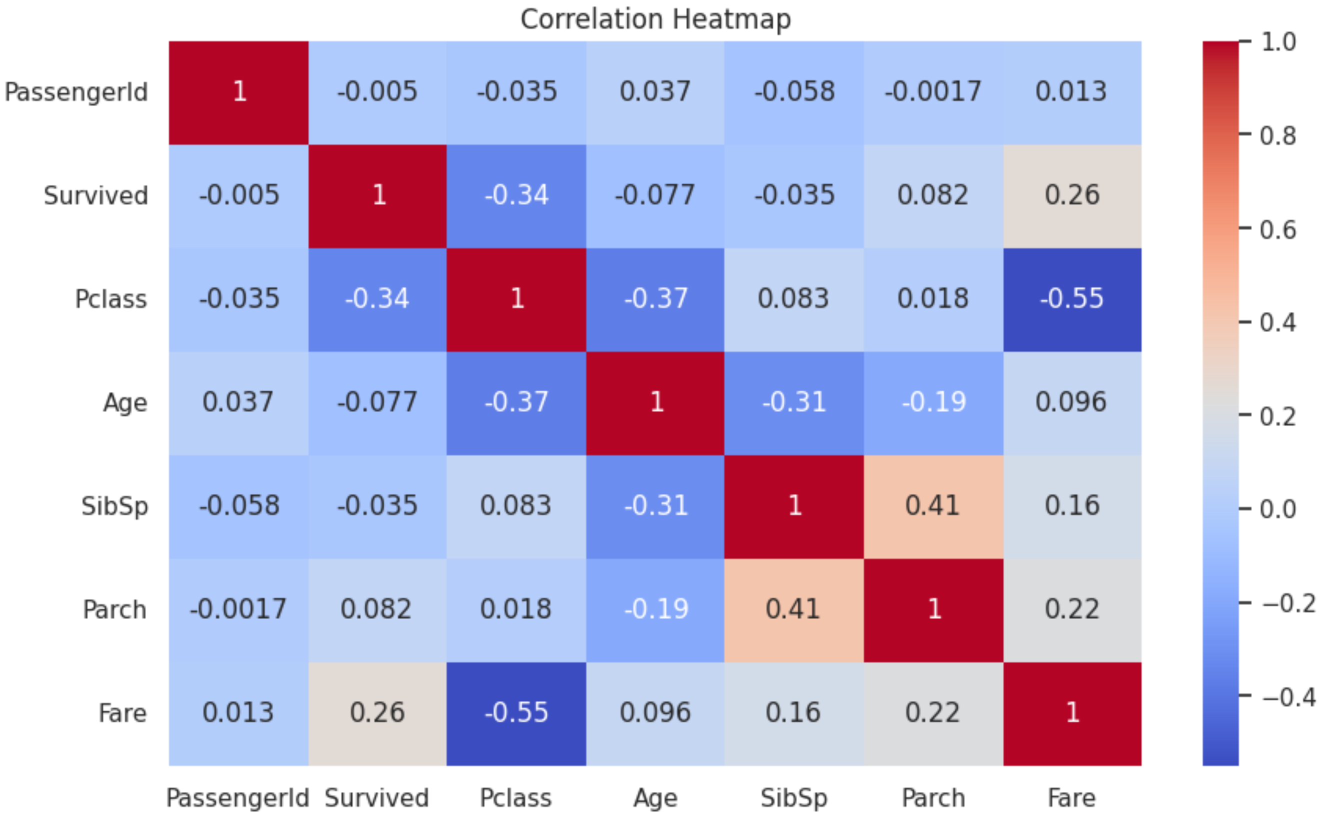
```
sns.boxplot(x='Age', data=df)
plt.title("Age Outliers")
plt.show()

sns.boxplot(x='Fare', data=df)
plt.title("Fare Outliers")
plt.show()
```



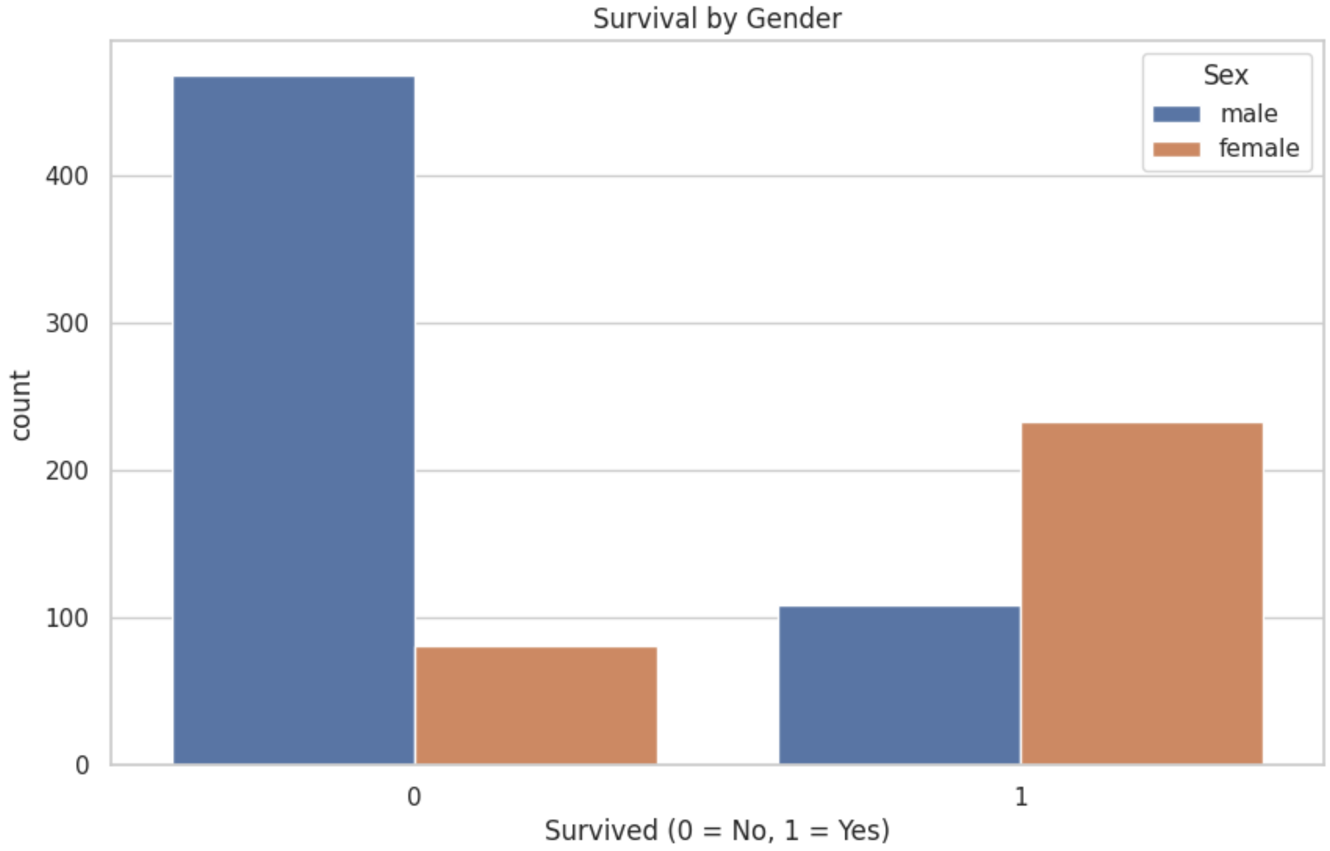

0
10
20
30
40
50
60
70
80
Age


Fare Outliers
0
100
200
300
400
500
Fare

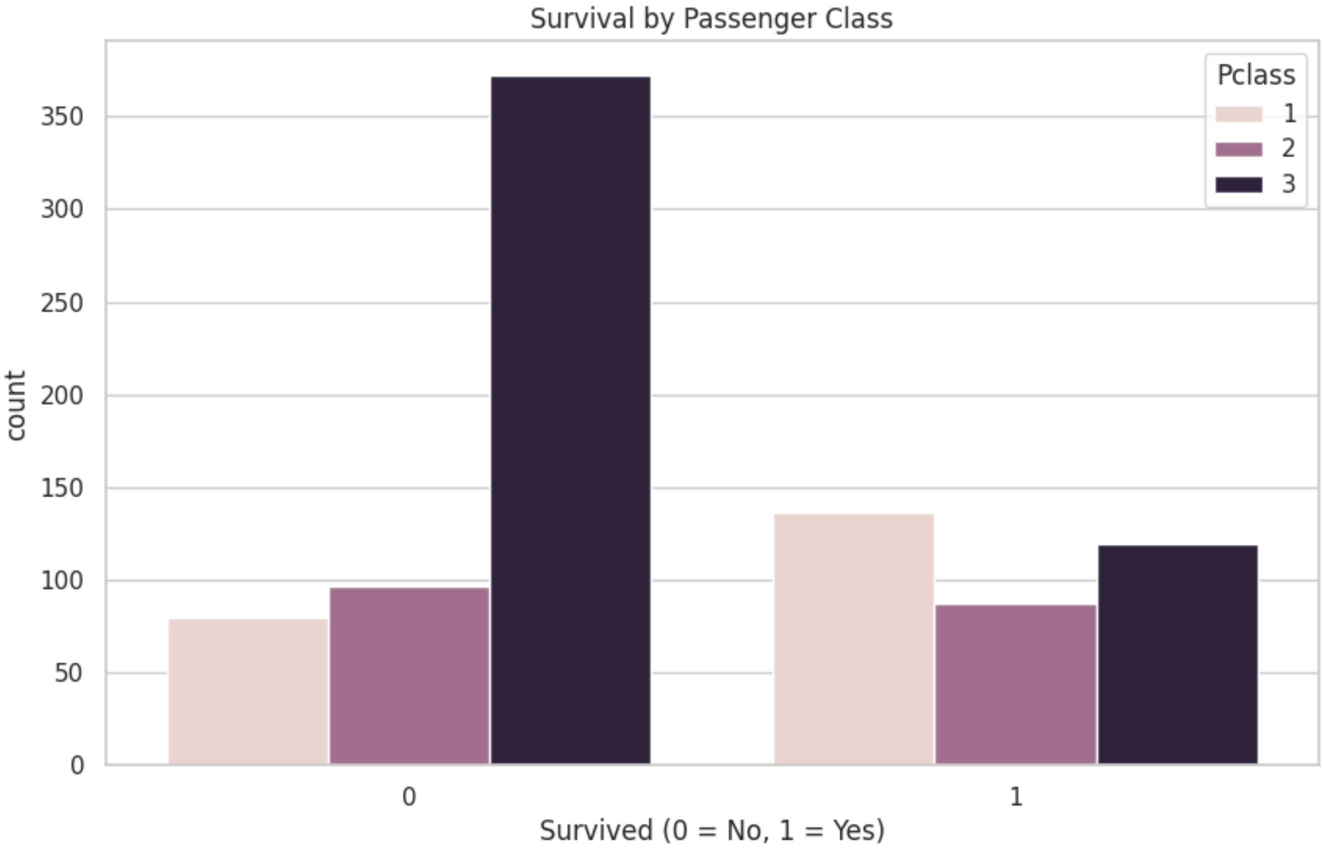
```
numeric_data = df.select_dtypes(include=['int64', 'float64'])
sns.heatmap(numeric_data.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```



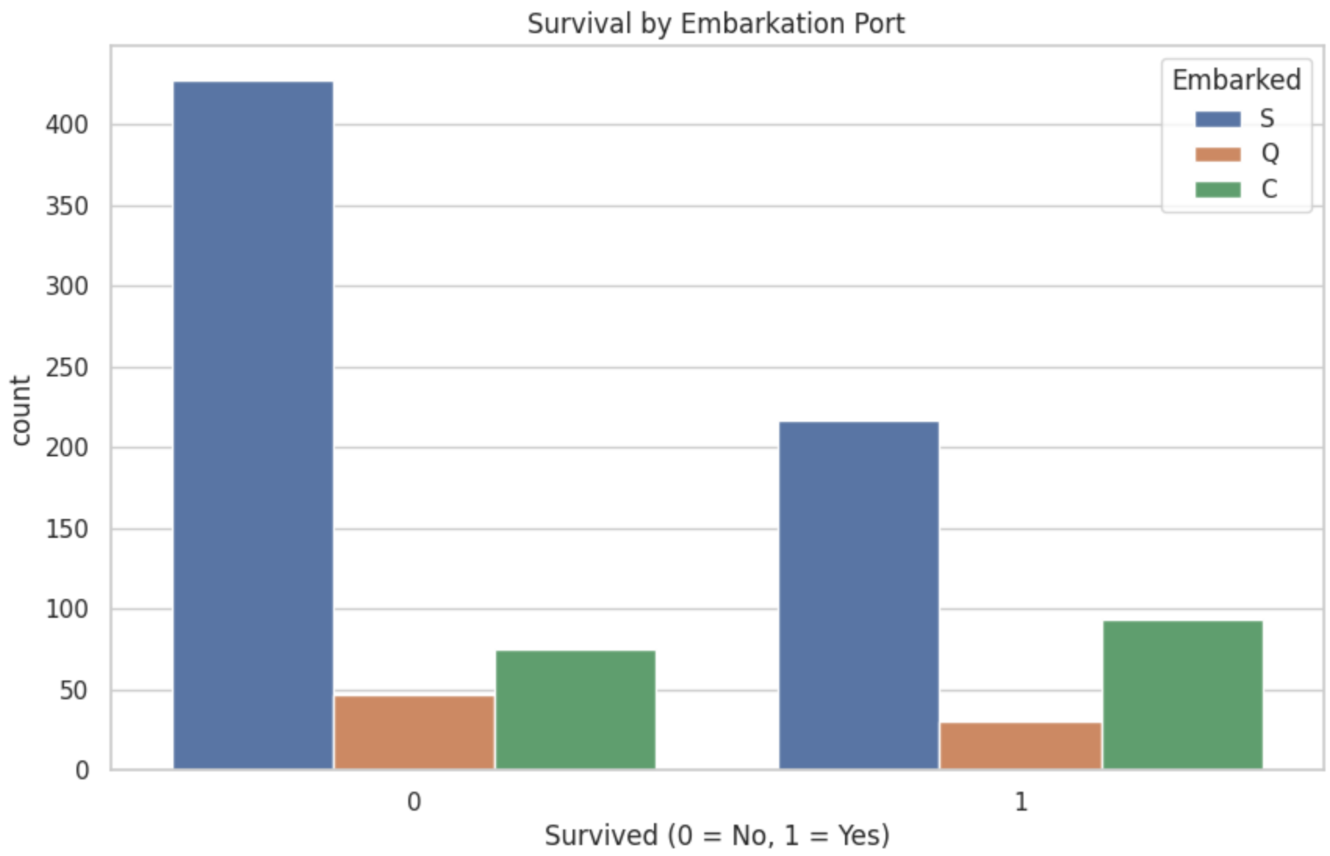
```
sns.countplot(x='Survived', hue='Sex', data=df)
plt.title("Survival by Gender")
plt.xlabel("Survived (0 = No, 1 = Yes)")
plt.show()
```



```
sns.countplot(x='Survived', hue='Pclass', data=df)
plt.title("Survival by Passenger Class")
plt.xlabel("Survived (0 = No, 1 = Yes)")
plt.show()
```



```
sns.countplot(x='Survived', hue='Embarked', data=df)
plt.title("Survival by Embarkation Port")
plt.xlabel("Survived (0 = No, 1 = Yes)")
plt.show()
```



```
# Step 9 — Key Insights from EDA

insights = """
🧠 Key Insights from Titanic EDA

🔍 Missing Values:
- 'Cabin' column has a significant amount of missing values.
- 'Age' has missing values that might require imputation.
- 'Embarked' has 2 missing entries.

📊 Distributions:
```

```
- Most passengers are between 20 and 40 years old.
- 'Fare' column has outliers with values exceeding $500.

📦 Outliers:
- Fare distribution is right-skewed with several high-end outliers.
- Minor outliers also present in Age.

🔗 Relationships:
- Females had a much higher survival rate compared to males.
- 1st Class passengers had better survival outcomes than 2nd and 3rd class.
- Embarked = 'C' port showed better survival rates.
- Positive correlation between Fare and Survival.

✅ Conclusion:
- Gender, passenger class, and port of embarkation significantly affected survi
- Handling missing values and scaling features may be necessary before applying
"""


print(insights)
```

```
🧠 Key Insights from Titanic EDA

🔍 Missing Values:
- 'Cabin' column has a significant amount of missing values.
- 'Age' has missing values that might require imputation.
- 'Embarked' has 2 missing entries.

📊 Distributions:
- Most passengers are between 20 and 40 years old.
- 'Fare' column has outliers with values exceeding $500.

📦 Outliers:
- Fare distribution is right-skewed with several high-end outliers.
- Minor outliers also present in Age.

🔗 Relationships:
- Females had a much higher survival rate compared to males.
- 1st Class passengers had better survival outcomes than 2nd and 3rd class.
- Embarked = 'C' port showed better survival rates.
- Positive correlation between Fare and Survival.

✅ Conclusion:
- Gender, passenger class, and port of embarkation significantly affected s
- Handling missing values and scaling features may be necessary before appl
```