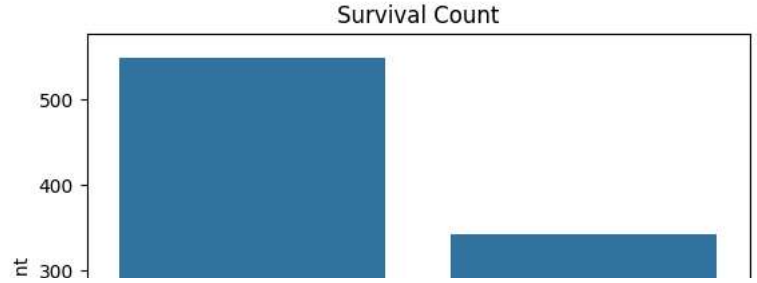
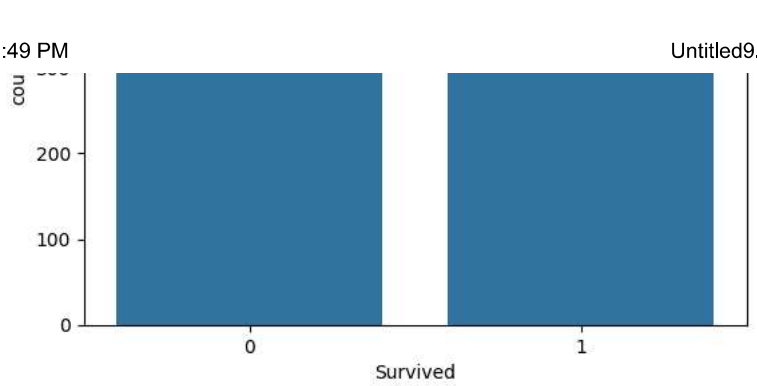
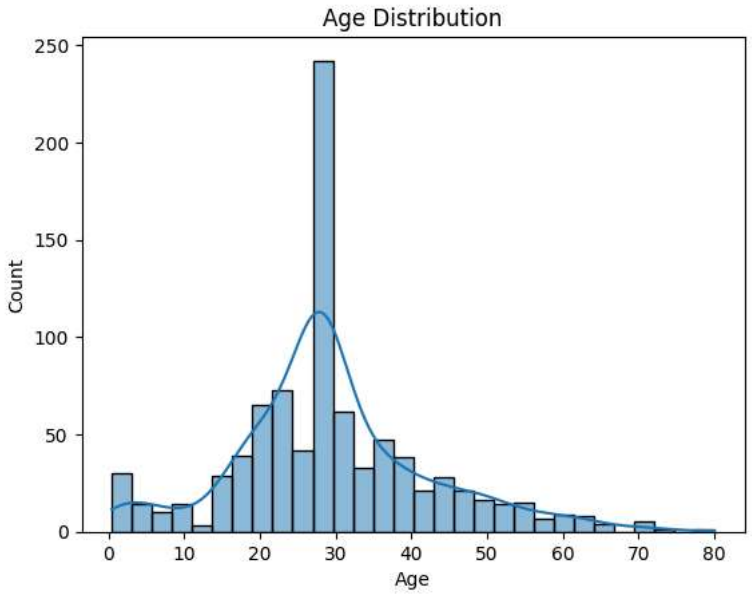
Start coding or generate with AI.

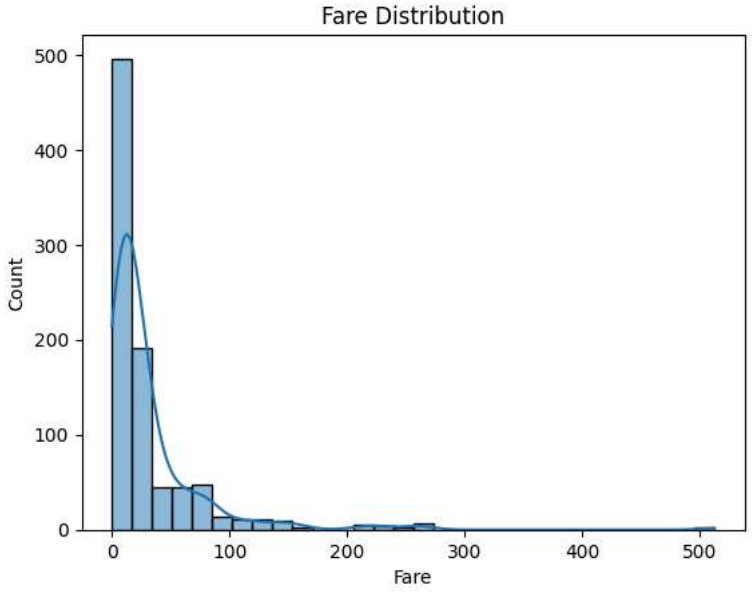
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv('titanic.csv')

# Step 2: Understand the Structure of the Data
print(df.head())
print(df.info())
print(df.describe())

# Step 3: Clean the Data
# Handle missing values
print(df.isnull().sum())
df['Age'].fillna(df['Age'].median(), inplace=True)
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
df.drop(columns='Cabin', inplace=True)  # Dropping 'Cabin' due to many missing values

# Ensure that only numeric columns are selected for the correlation matrix
numeric_df = df.select_dtypes(include=[np.number])

# Step 4: Explore the Data's Characteristics
# Distribution of 'Survived'
sns.countplot(data=df, x='Survived')
plt.title('Survival Count')
plt.show()

# Distribution of 'Age'
sns.histplot(df['Age'], bins=30, kde=True)
plt.title('Age Distribution')
plt.show()

# Step 5: Visualize Distributions
# Histogram of 'Fare'
sns.histplot(df['Fare'], bins=30, kde=True)
plt.title('Fare Distribution')
plt.show()

# Box plot of 'Age' grouped by 'Survived'
sns.boxplot(data=df, x='Survived', y='Age')
plt.title('Age vs. Survival')
plt.show()

# Step 6: Analyze Correlations
# Heatmap of correlations
correlation_matrix = numeric_df.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()

# Step 7: Identify Outliers
# Box plot of 'Fare'
sns.boxplot(data=df, x='Fare')
plt.title('Fare Box Plot')
plt.show()
```

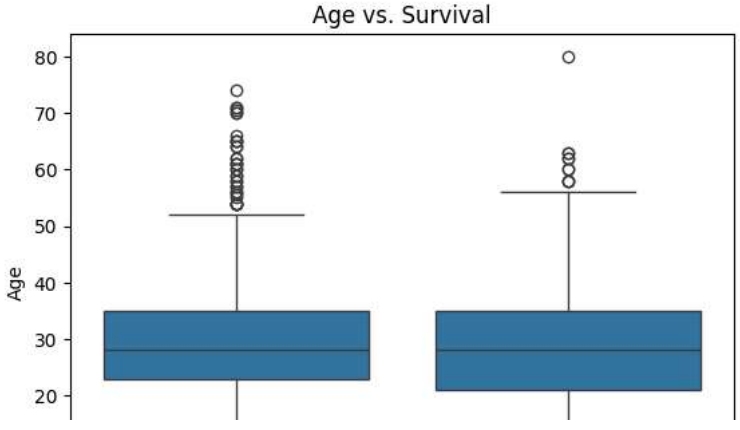
```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
<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
None
       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
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```



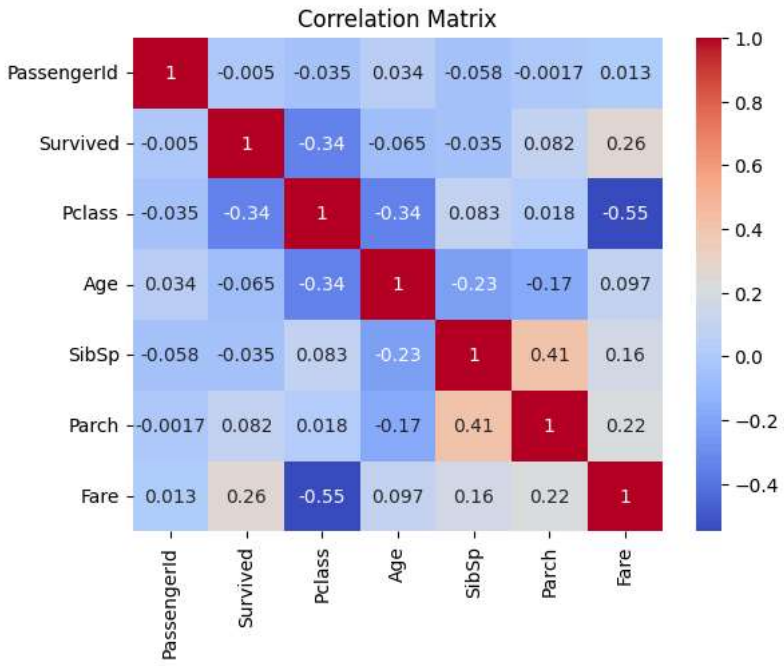

cou
200
100
0
0
1
Survived

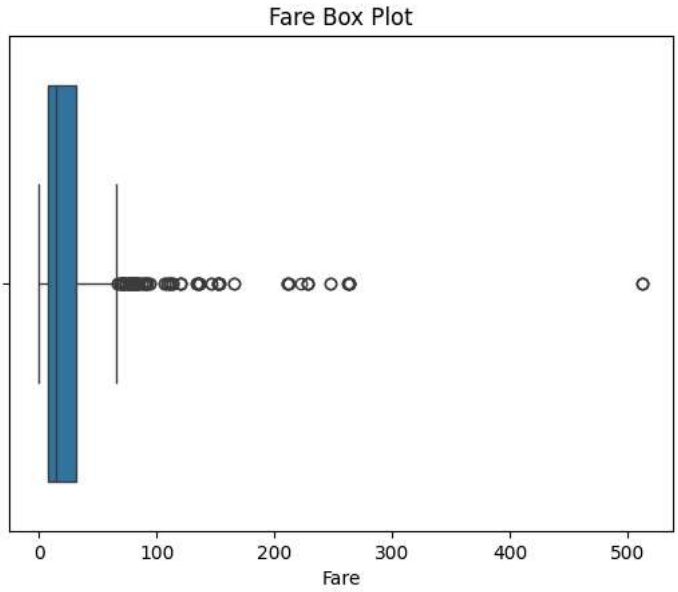

Age Distribution
250
200
150
100
50
0
Count
0
10
20
30
40
50
60
70
80
Age


Fare Distribution
500
400
300
200
100
0
Count
0
100
200
300
400
500
Fare


Age vs. Survival
80
70
60
50
40
30
20
Age


10
0
0
1
Survived


Correlation Matrix
PassengerId 1 -0.005 -0.035 0.034 -0.058 -0.0017 0.013
Survived -0.005 1 -0.34 -0.065 -0.035 0.082 0.26
Pclass -0.035 -0.34 1 -0.34 0.083 0.018 -0.55
Age 0.034 -0.065 -0.34 1 -0.23 -0.17 0.097
SibSp -0.058 -0.035 0.083 -0.23 1 0.41 0.16
Parch -0.0017 0.082 0.018 -0.17 0.41 1 0.22
Fare 0.013 0.26 -0.55 0.097 0.16 0.22 1
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Fare Box Plot
0
100
200
300
400
500
Fare