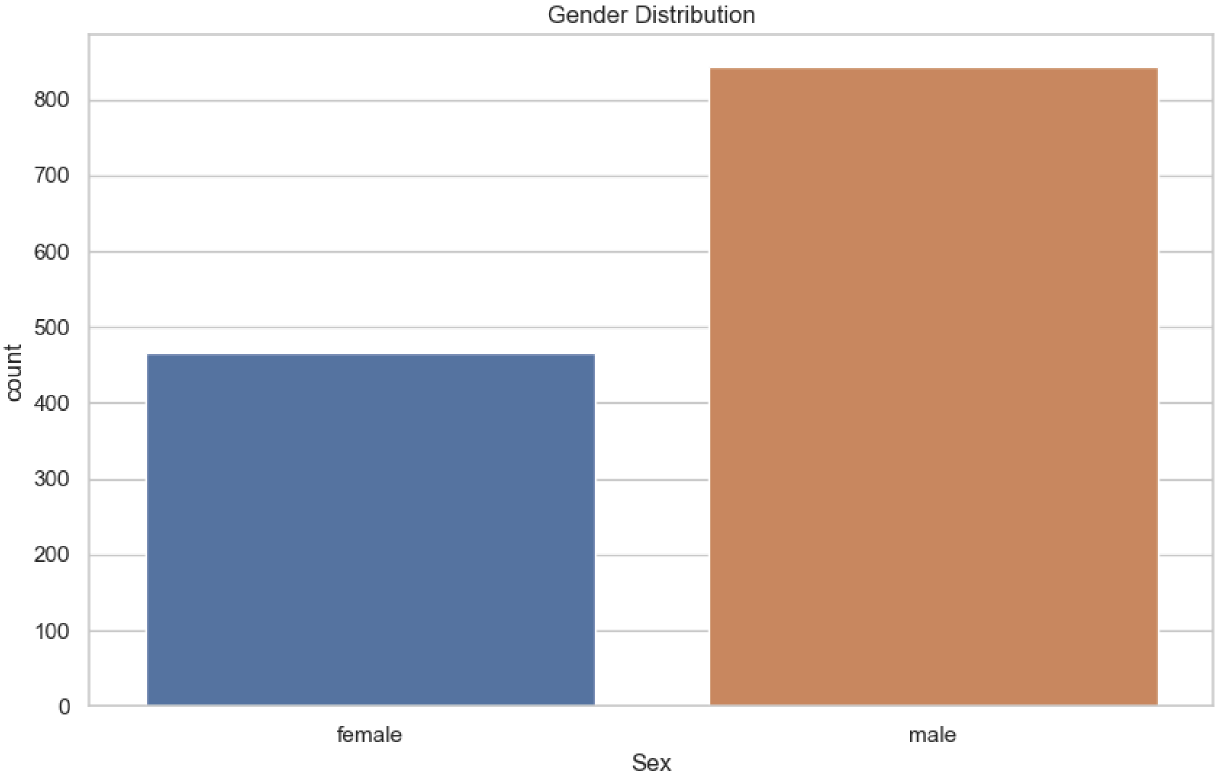
In [4]:

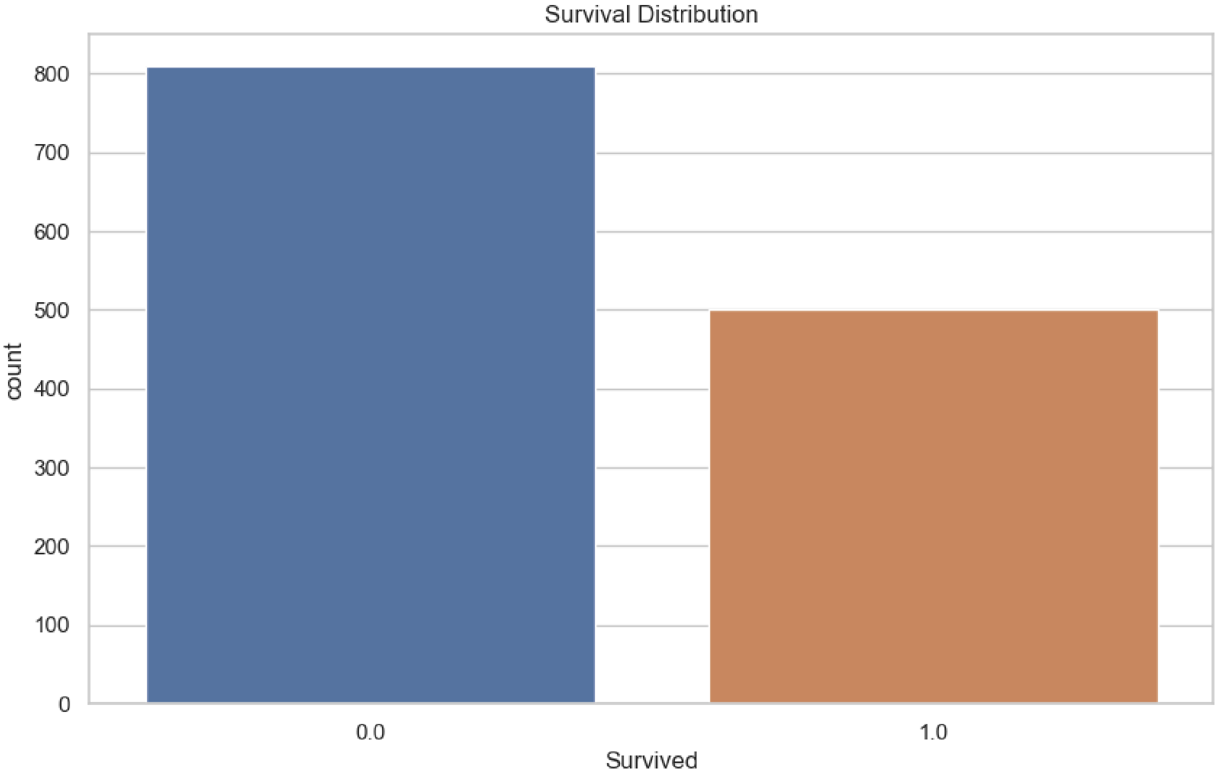
```python
# TASK 5: Exploratory Data Analysis - Titanic Dataset

# Step 1: Import Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Setup display settings
pd.set_option('display.max_columns', 100)
sns.set(style='whitegrid')
plt.rcParams['figure.figsize'] = (10,6)

# Step 2: Load the Dataset
# Make sure the Titanic dataset (train.csv) is in the same folder
df = pd.read_csv('train.csv')
print("✅ Dataset loaded successfully!")
print("Shape of dataset:", df.shape)
print(df.head())

# Step 3: Basic Information
print("\n--- Dataset Info ---")
print(df.info())

print("\n--- Missing Values ---")
print(df.isnull().sum())

print("\n--- Summary Statistics ---")
print(df.describe(include='all').T)

# Step 4: Data Cleaning (Safe Version)
# Clean column names (remove spaces and make consistent capitalization)
df.columns = df.columns.str.strip().str.capitalize()

# Fill missing Age
if 'Age' in df.columns:
    df['Age'] = df['Age'].fillna(df['Age'].median())

# Fill missing Embarked
if 'Embarked' in df.columns:
    df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])

# Drop unwanted columns if they exist
for col in ['Cabin', 'Ticket']:
    if col in df.columns:
        df.drop(columns=[col], inplace=True)

print("\n✅ Missing values after cleaning:")
print(df.isnull().sum())

# Step 5: Feature Engineering
if all(x in df.columns for x in ['Sibsp', 'Parch']):
    df['Familysize'] = df['Sibsp'] + df['Parch'] + 1
    df['Isalone'] = (df['Familysize'] == 1).astype(int)

if 'Name' in df.columns:
    df['Title'] = df['Name'].str.extract('([A-Za-z]+)\.')
    df['Title'] = df['Title'].replace(['Mlle', 'Ms'], 'Miss')
    df['Title'] = df['Title'].replace(['Mme'], 'Mrs')

# Step 6: Univariate Analysis
if 'Sex' in df.columns:
    sns.countplot(x='Sex', data=df)
    plt.title('Gender Distribution')
    plt.show()

if 'Survived' in df.columns:
    sns.countplot(x='Survived', data=df)
    plt.title('Survival Distribution')
    plt.show()

if 'Age' in df.columns:
    sns.histplot(df['Age'], bins=30, kde=True)
    plt.title('Age Distribution')
    plt.show()

if 'Fare' in df.columns:
    sns.histplot(df['Fare'], bins=30, kde=True, color='green')
    plt.title('Fare Distribution')
    plt.show()

# Step 7: Bivariate Analysis
if all(x in df.columns for x in ['Sex', 'Survived']):
    sns.countplot(x='Sex', hue='Survived', data=df)
    plt.title('Survival by Gender')
    plt.show()

if all(x in df.columns for x in ['Pclass', 'Survived']):
    sns.countplot(x='Pclass', hue='Survived', data=df)
    plt.title('Survival by Passenger Class')
    plt.show()

if all(x in df.columns for x in ['Survived', 'Age']):
    sns.boxplot(x='Survived', y='Age', data=df)
    plt.title('Age vs Survival')
    plt.show()

if all(x in df.columns for x in ['Survived', 'Fare']):
    sns.boxplot(x='Survived', y='Fare', data=df)
    plt.title('Fare vs Survival')
    plt.show()

# Step 8: Multivariate Analysis
num_df = df.select_dtypes(include=[np.number])
if not num_df.empty:
    corr = num_df.corr()
    sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
    plt.title('Correlation Matrix')
    plt.show()

# Step 9: Summary of Observations
print("\n✅ SUMMARY OF FINDINGS:")
print("""
1. More females survived than males.
2. Passengers in 1st class had higher survival rates.
3. Younger passengers had slightly better chances of survival.
4. Higher fares were linked to higher survival probability.
5. Passengers traveling alone had lower survival rates.
""")

# Step 10: Export Cleaned Data
df.to_csv('titanic_cleaned.csv', index=False)
print("\n✅ Cleaned dataset saved as 'titanic_cleaned.csv'")
```

```
✅ Dataset loaded successfully!
Shape of dataset: (1310, 14)
   pclass  survived                                             name     sex  \
0     1.0       1.0                    Allen, Miss. Elisabeth Walton  female   
1     1.0       1.0                   Allison, Master. Hudson Trevor    male   
2     1.0       0.0                     Allison, Miss. Helen Loraine  female   
3     1.0       0.0             Allison, Mr. Hudson Joshua Creighton    male   
4     1.0       0.0  Allison, Mrs. Hudson J C (Bessie Waldo Daniels)  female   

       age  sibsp  parch  ticket      fare    cabin embarked boat   body  \
0  29.0000    0.0    0.0   24160  211.3375       B5        S    2    NaN   
1   0.9167    1.0    2.0  113781  151.5500  C22 C26        S   11    NaN   
2   2.0000    1.0    2.0  113781  151.5500  C22 C26        S  NaN    NaN   
3  30.0000    1.0    2.0  113781  151.5500  C22 C26        S  NaN  135.0   
4  25.0000    1.0    2.0  113781  151.5500  C22 C26        S  NaN    NaN   

                         home.dest  
0                     St Louis, MO  
1  Montreal, PQ / Chesterville, ON  
2  Montreal, PQ / Chesterville, ON  
3  Montreal, PQ / Chesterville, ON  
4  Montreal, PQ / Chesterville, ON  

--- Dataset Info ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1310 entries, 0 to 1309
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   pclass     1309 non-null   float64
 1   survived   1309 non-null   float64
 2   name       1309 non-null   object 
 3   sex        1309 non-null   object 
 4   age        1046 non-null   float64
 5   sibsp      1309 non-null   float64
 6   parch      1309 non-null   float64
 7   ticket     1309 non-null   object 
 8   fare       1308 non-null   float64
 9   cabin      295 non-null    object 
 10  embarked   1307 non-null   object 
 11  boat       486 non-null    object 
 12  body       121 non-null    float64
 13  home.dest  745 non-null    object 
dtypes: float64(7), object(7)
memory usage: 143.4+ KB
None

--- Missing Values ---
pclass          1
survived        1
name            1
sex             1
age           264
sibsp           1
parch           1
ticket          1
fare            2
cabin        1015
embarked        3
boat          824
```

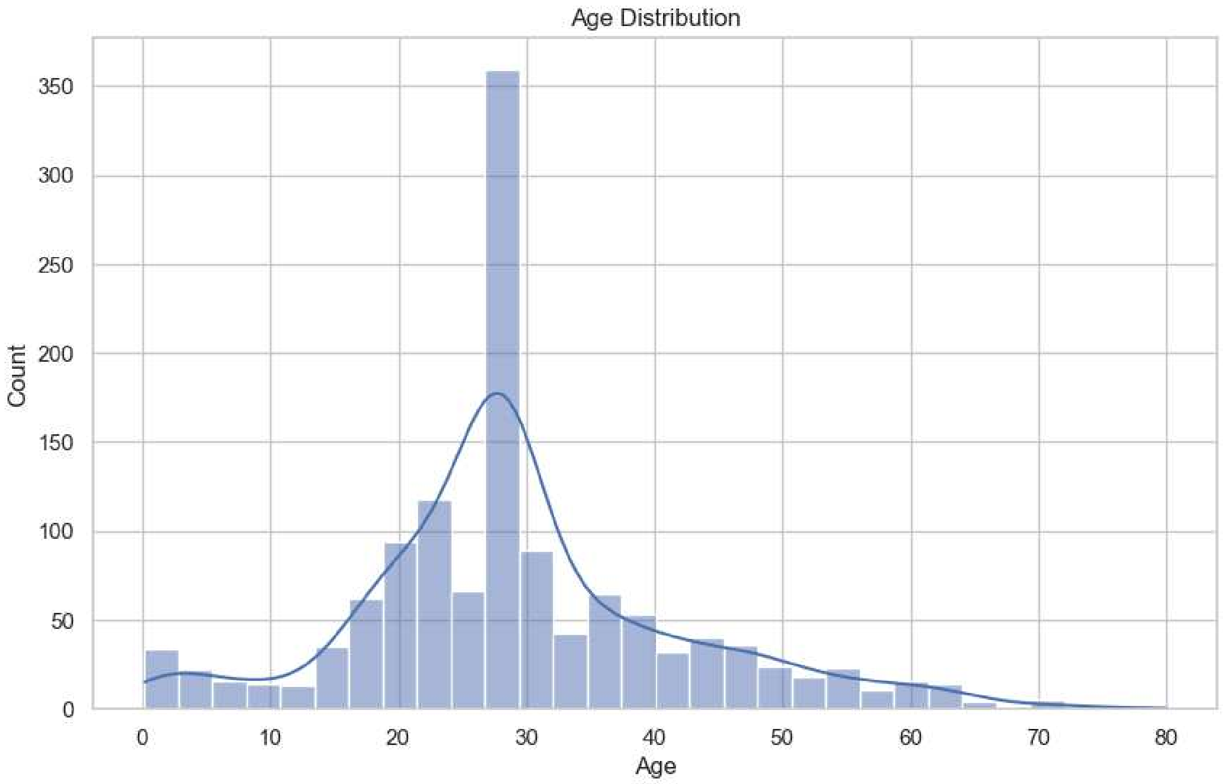
```
body         1189
home.dest     565
dtype: int64

--- Summary Statistics ---
            count unique                   top freq        mean        std  \
pclass     1309.0    NaN                   NaN  NaN    2.294882   0.837836
survived   1309.0    NaN                   NaN  NaN    0.381971   0.486055
name         1309   1307  Connolly, Miss. Kate    2         NaN        NaN
sex          1309      2                  male  843         NaN        NaN
age        1046.0    NaN                   NaN  NaN   29.881135    14.4135
sibsp      1309.0    NaN                   NaN  NaN    0.498854   1.041658
parch      1309.0    NaN                   NaN  NaN    0.385027    0.86556
ticket       1309    929              CA. 2343   11         NaN        NaN
fare       1308.0    NaN                   NaN  NaN   33.295479  51.758668
cabin         295    186           C23 C25 C27    6         NaN        NaN
embarked     1307      3                     S  914         NaN        NaN
boat          486     27                    13   39         NaN        NaN
body        121.0    NaN                   NaN  NaN  160.809917  97.696922
home.dest     745    369          New York, NY   64         NaN        NaN

              min     25%      50%     75%       max
pclass        1.0     2.0      3.0     3.0       3.0
survived      0.0     0.0      0.0     1.0       1.0
name          NaN     NaN      NaN     NaN       NaN
sex           NaN     NaN      NaN     NaN       NaN
age        0.1667    21.0     28.0    39.0      80.0
sibsp         0.0     0.0      0.0     1.0       8.0
parch         0.0     0.0      0.0     0.0       9.0
ticket        NaN     NaN      NaN     NaN       NaN
fare          0.0  7.8958  14.4542  31.275  512.3292
cabin         NaN     NaN      NaN     NaN       NaN
embarked      NaN     NaN      NaN     NaN       NaN
boat          NaN     NaN      NaN     NaN       NaN
body          1.0    72.0    155.0   256.0     328.0
home.dest     NaN     NaN      NaN     NaN       NaN

✅ Missing values after cleaning:
Pclass          1
Survived        1
Name            1
Sex             1
Age             0
Sibsp           1
Parch           1
Fare            2
Embarked        0
Boat          824
Body         1189
Home.dest     565
dtype: int64
```



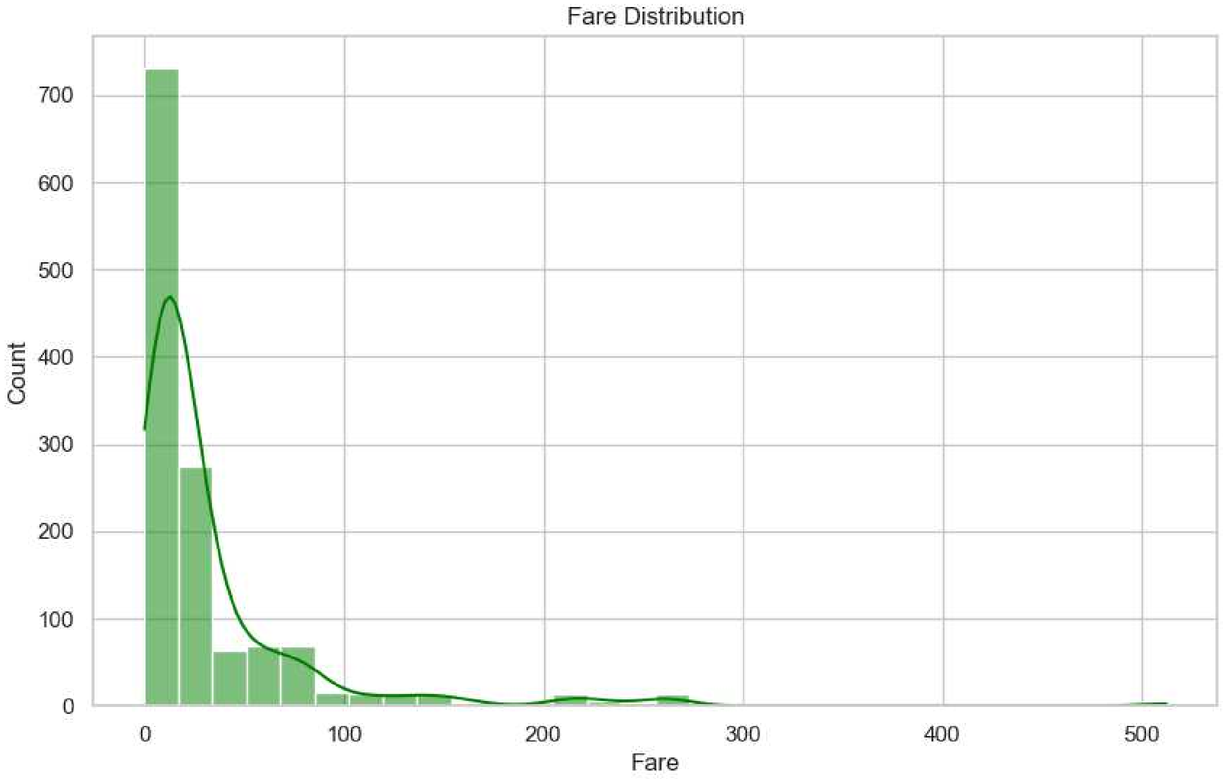
Gender Distribution



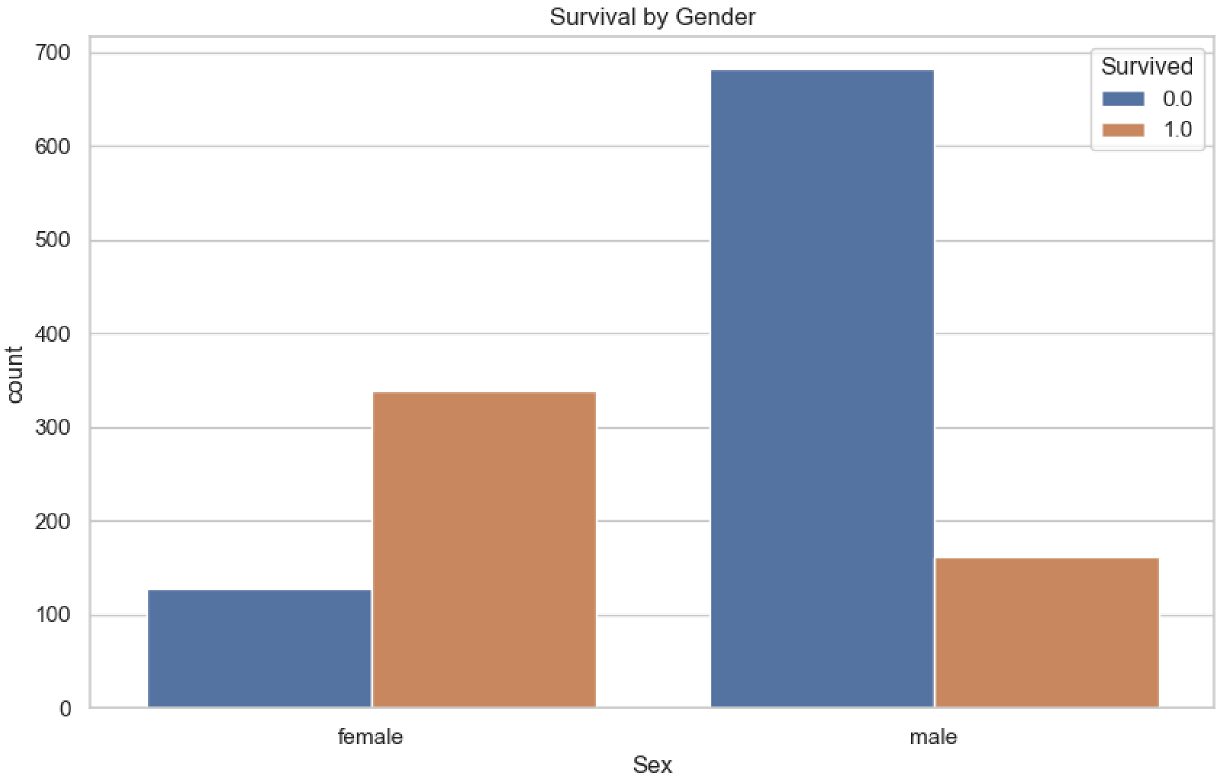
Survival Distribution



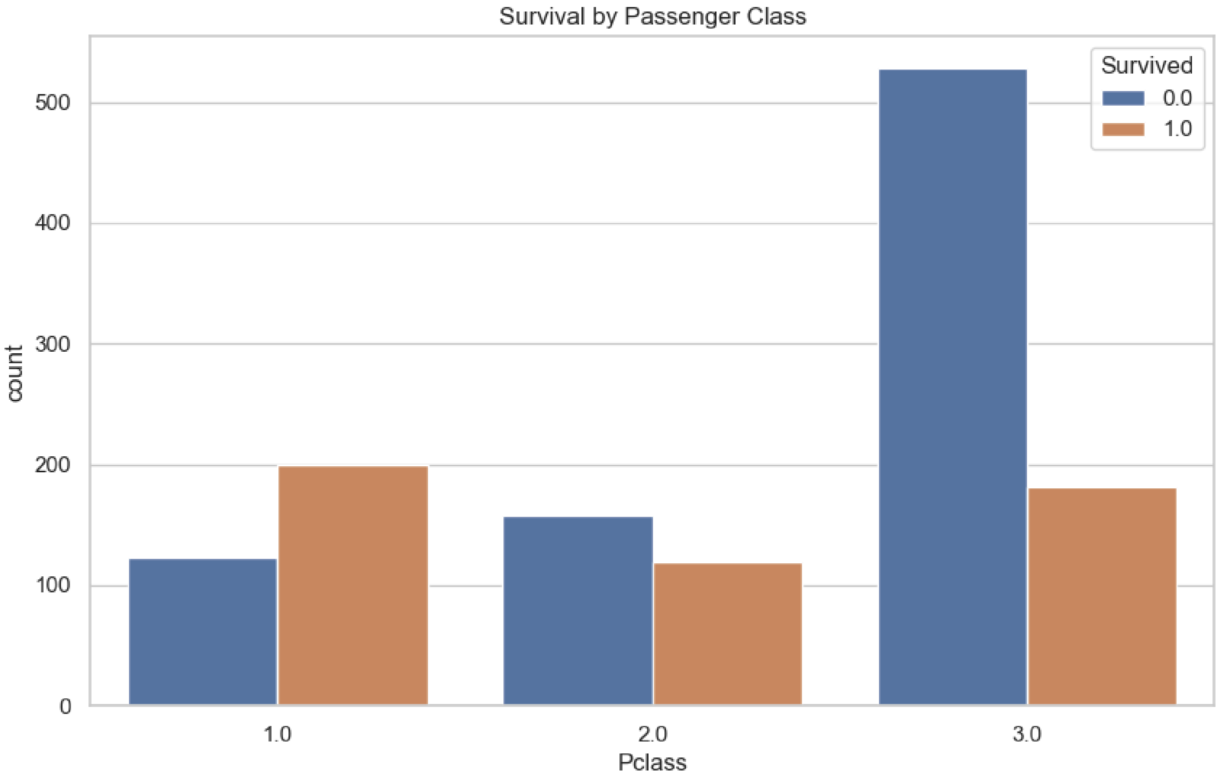
Age Distribution



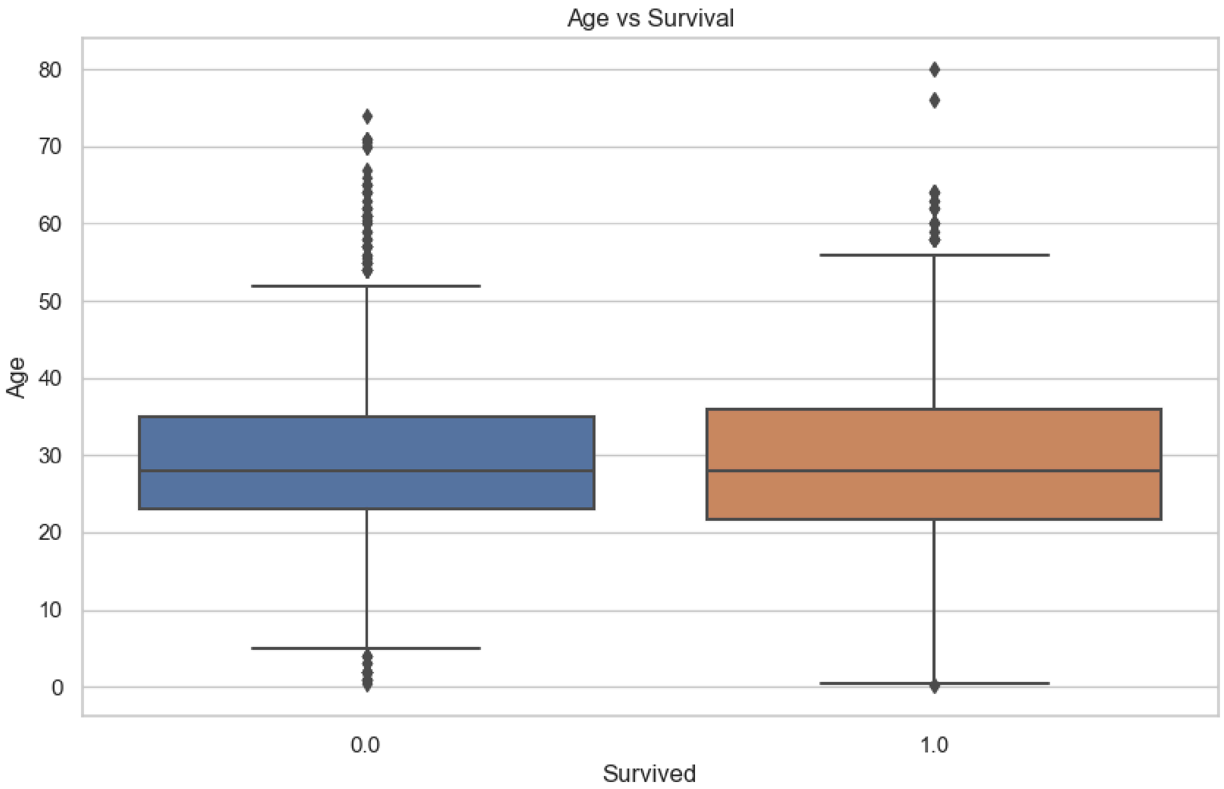
Fare Distribution



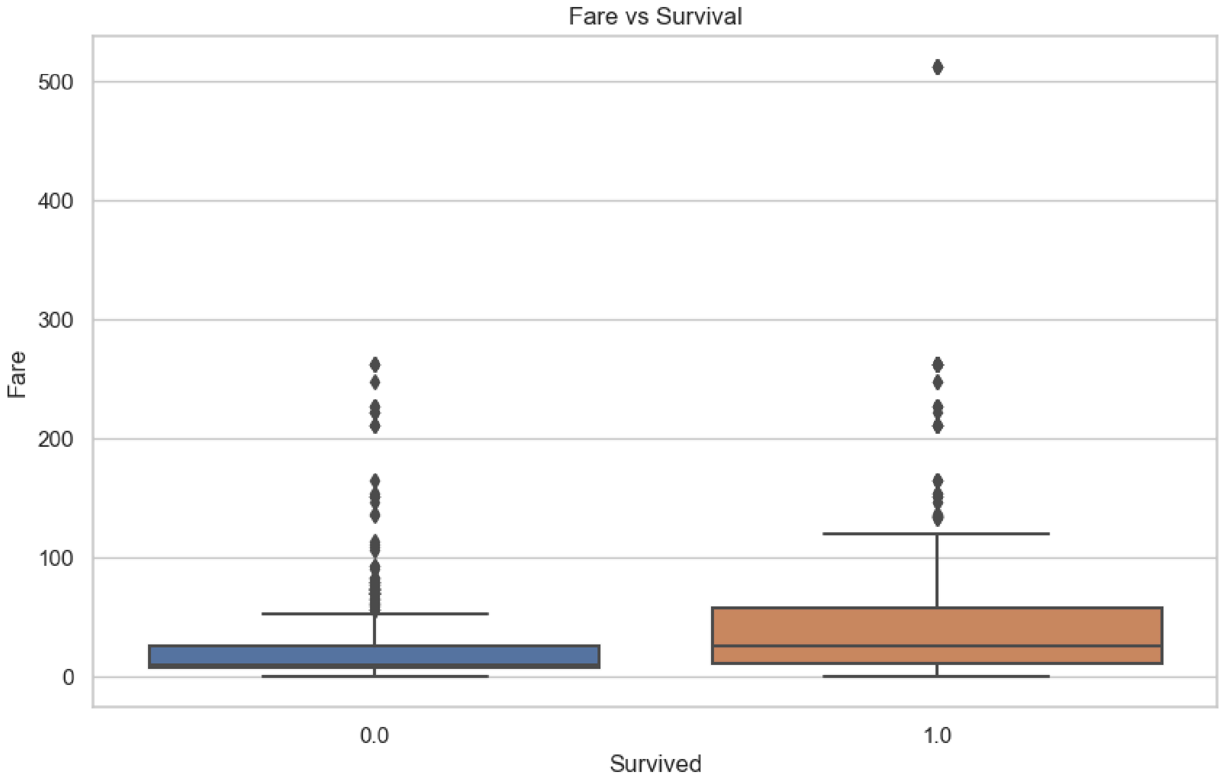
Survival by Gender



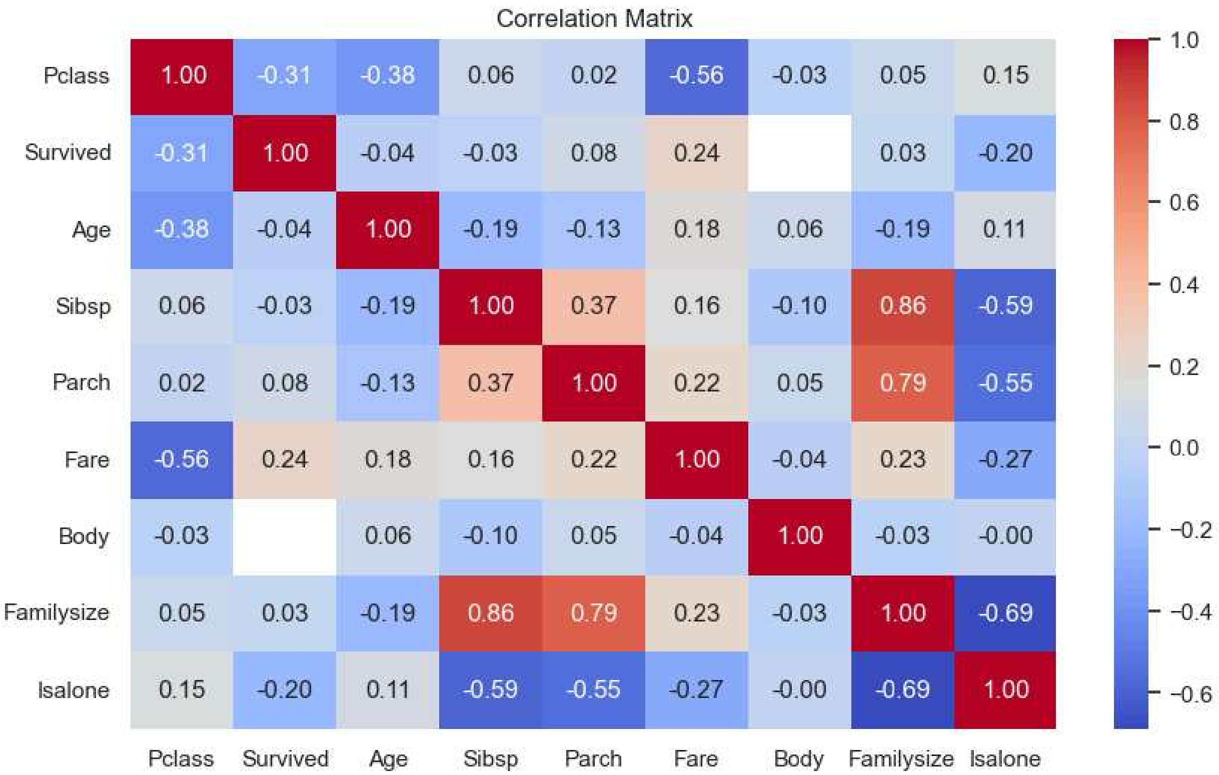
Survival by Passenger Class



Age vs Survival



Fare vs Survival



Correlation Matrix

	Pclass	Survived	Age	Sibsp	Parch	Fare	Body	Familysize	Isalone
Pclass	1.00	-0.31	-0.38	0.06	0.02	-0.56	-0.03	0.05	0.15
Survived	-0.31	1.00	-0.04	-0.03	0.08	0.24		0.03	-0.20
Age	-0.38	-0.04	1.00	-0.19	-0.13	0.18	0.06	-0.19	0.11
Sibsp	0.06	-0.03	-0.19	1.00	0.37	0.16	-0.10	0.86	-0.59
Parch	0.02	0.08	-0.13	0.37	1.00	0.22	0.05	0.79	-0.55
Fare	-0.56	0.24	0.18	0.16	0.22	1.00	-0.04	0.23	-0.27
Body	-0.03		0.06	-0.10	0.05	-0.04	1.00	-0.03	-0.00
Familysize	0.05	0.03	-0.19	0.86	0.79	0.23	-0.03	1.00	-0.69
Isalone	0.15	-0.20	0.11	-0.59	-0.55	-0.27	-0.00	-0.69	1.00

```
✅ SUMMARY OF FINDINGS:

1. More females survived than males.
2. Passengers in 1st class had higher survival rates.
3. Younger passengers had slightly better chances of survival.
4. Higher fares were linked to higher survival probability.
5. Passengers traveling alone had lower survival rates.


✅ Cleaned dataset saved as 'titanic_cleaned.csv'
```

In [ ]: