Uncovering Insights in the Titanic Dataset using Exploratory Data Analysis

Example Dataset :"Titanic: Machine Learning from Disaster" from Kaggle.

Objective:

- Understand the factors that affect the survival rate on the Titanic.
- Perform data cleaning and preprocessing.
- Visualize data to identify patterns and relationships.
- Derive meaningful insights from the analysis.

Introduction

The Titanic dataset provides information on the passengers aboard the RMS Titanic, which sank in 1912. This analysis aims to explore the factors influencing the survival rates of passengers using various exploratory data analysis (EDA) techniques.

Dataset Description

- The dataset includes the following attributes:
-
- PassengerId: Unique ID for each passenger.
- Survived: Survival status (0 = No, 1 = Yes).
- Pclass: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd).
- Name: Name of the passenger.
- Sex: Gender of the passenger.
- Age: Age of the passenger.
- SibSp: Number of siblings/spouses aboard.
- Parch: Number of parents/children aboard.
- Ticket: Ticket number.
- Fare: Passenger fare.
- Cabin: Cabin number.
- Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

Exploratory Data Analysis:

Load the data

```python
import pandas as pd
import numpy as np

# Load dataset
df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/titanic_sample.csv')

# Display first few rows
print(df.head())
```

```
   PassengerId  Survived     Pclass  \
1            0         3     Braund
2            1         1    Cumings
3            1         3  Heikkinen
4            1         1   Futrelle
5            0         3      Allen

                                          Name     Sex   Age  SibSp  Parch  \
1                              Mr. Owen Harris    male  22.0      1      0
2   Mrs. John Bradley (Florence Briggs Thayer)  female  38.0      1      0
3                                  Miss. Laina  female  26.0      0      0
4          Mrs. Jacques Heath (Lily May Peel)  female  35.0      1      0
5                            Mr. William Henry    male  35.0      0      0

            Ticket     Fare Cabin Embarked
1        A/5 21171   7.2500   NaN        S
2         PC 17599  71.2833   C85        C
3  STON/O2. 3101282   7.9250   NaN        S
4           113803  53.1000  C123        S
5           373450   8.0500   NaN        S
```

Data Cleaning :Handle missing values, outliers, and duplicates.

```python
[2] # Check for missing values
    print(df.isnull().sum())

    # Fill missing 'Age' with median
    df['Age'].fillna(df['Age'].median(), inplace=True)

    # Fill missing 'Embarked' with mode
    df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)

    # Drop 'Cabin' due to too many missing values
    df.drop(columns=['Cabin'], inplace=True)

    # Check for duplicates and drop them
    df.drop_duplicates(inplace=True)
```

```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            1
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          7
Embarked       0
dtype: int64
```

Summary Statistics

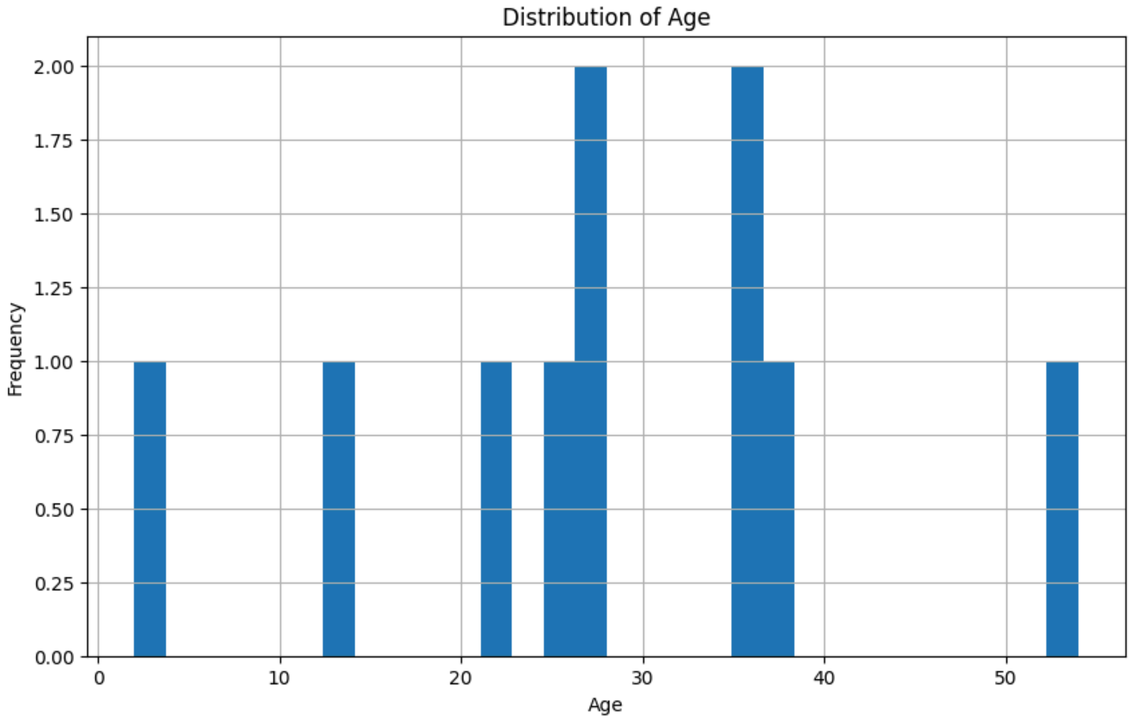
Summarize the dataset to understand its distribution

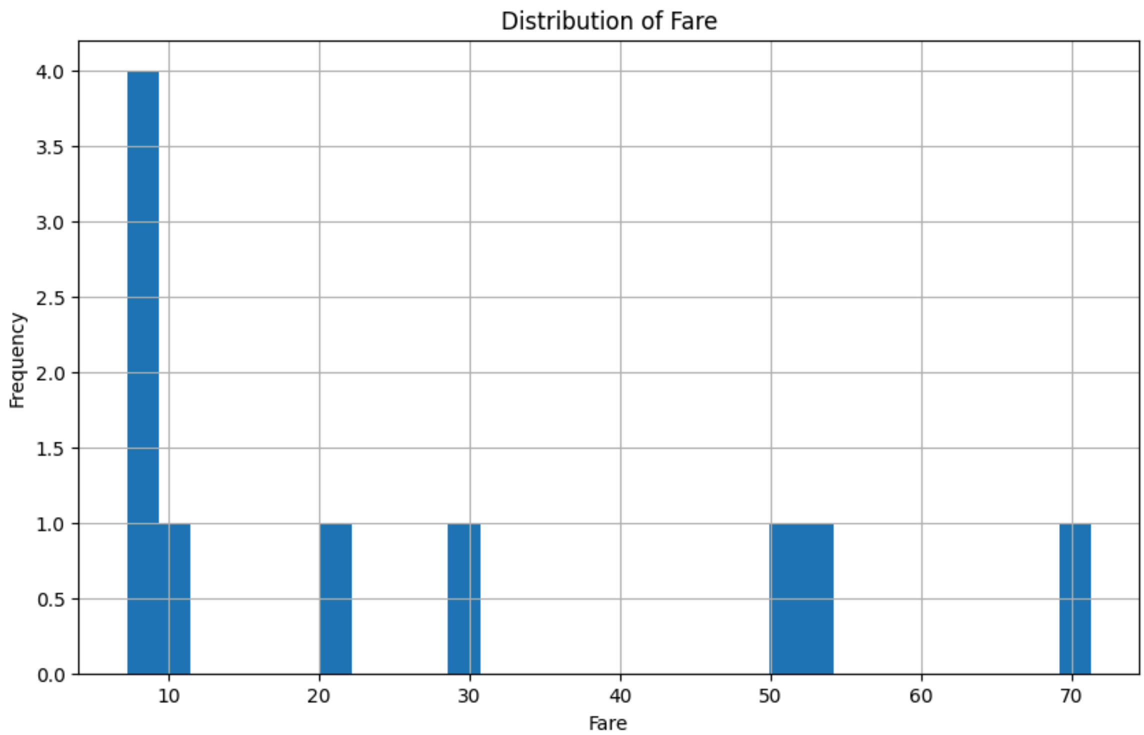
```
[3] print(df.describe())
```

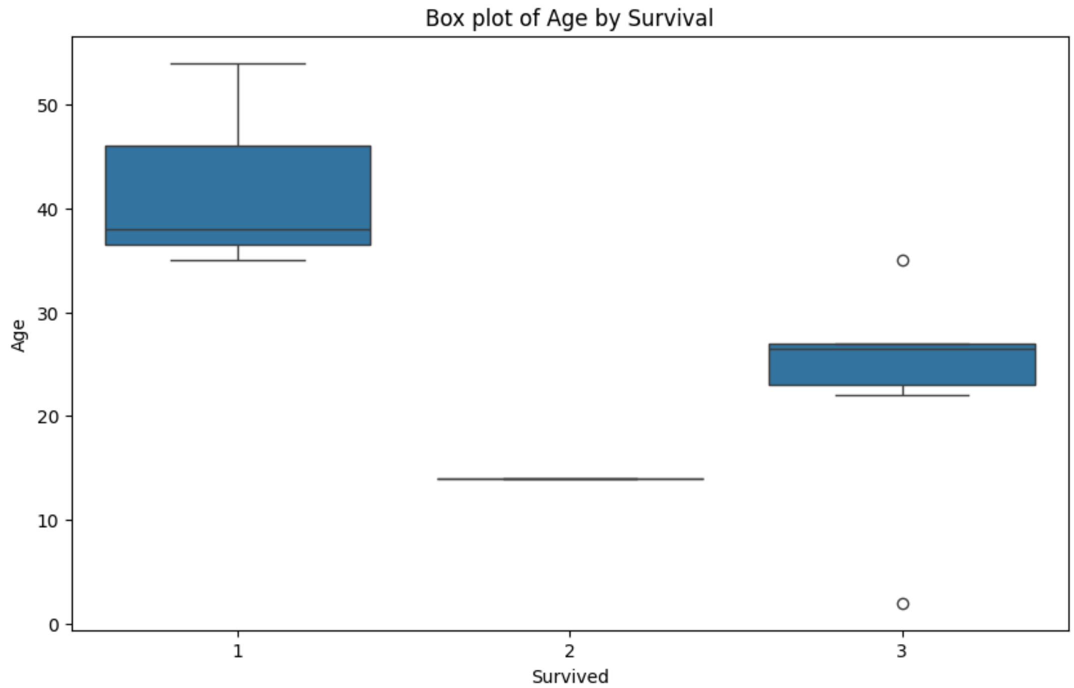
```
       PassengerId   Survived        Age      SibSp      Parch       Fare
count    10.000000  10.000000  10.000000  10.000000  10.000000  10.000000
mean      0.500000   2.300000  28.000000   0.700000   0.300000  27.020820
std       0.527046   0.948683  14.094916   0.948683   0.674949  23.601938
min       0.000000   1.000000   2.000000   0.000000   0.000000   7.250000
25%       0.000000   1.250000  23.000000   0.000000   0.000000   8.152075
50%       0.500000   3.000000  27.000000   0.500000   0.000000  16.104150
75%       1.000000   3.000000  35.000000   1.000000   0.000000  46.414575
max       1.000000   3.000000  54.000000   3.000000   2.000000  71.283300
```

Data Visualization and Discussion

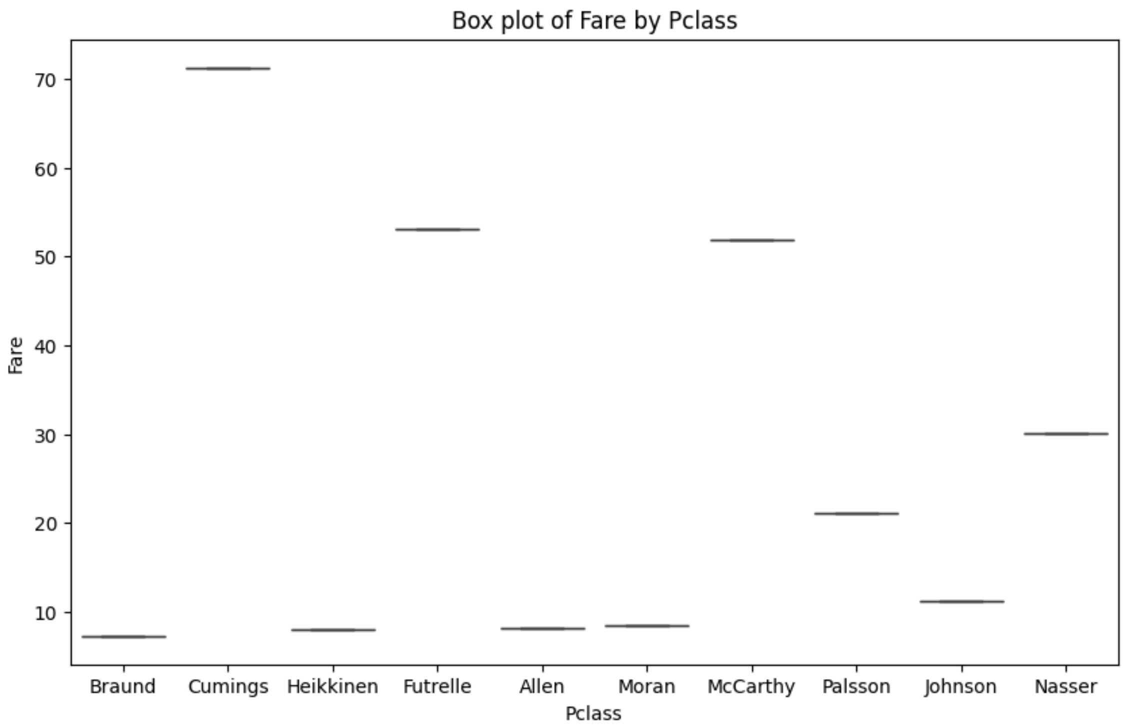
```
# Histograms
plt.figure(figsize=(10, 6))
df['Age'].hist(bins=30)
plt.title('Distribution of Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()

plt.figure(figsize=(10, 6))
df['Fare'].hist(bins=30)
plt.title('Distribution of Fare')
plt.xlabel('Fare')
plt.ylabel('Frequency')
plt.show()

# Box plots
plt.figure(figsize=(10, 6))
sns.boxplot(x='Survived', y='Age', data=df)
plt.title('Box plot of Age by Survival')
plt.show()

plt.figure(figsize=(10, 6))
sns.boxplot(x='Pclass', y='Fare', data=df)
plt.title('Box plot of Fare by Pclass')
plt.show()

# Scatter plots
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df)
plt.title('Age vs Fare by Survival')
plt.show()

# Correlation heatmap
plt.figure(figsize=(10, 6))
numeric_df = df.select_dtypes(include=['float64', 'int64'])  # Select only numeric columns
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```


Distribution of Age
Frequency
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
10
20
30
40
50
Age

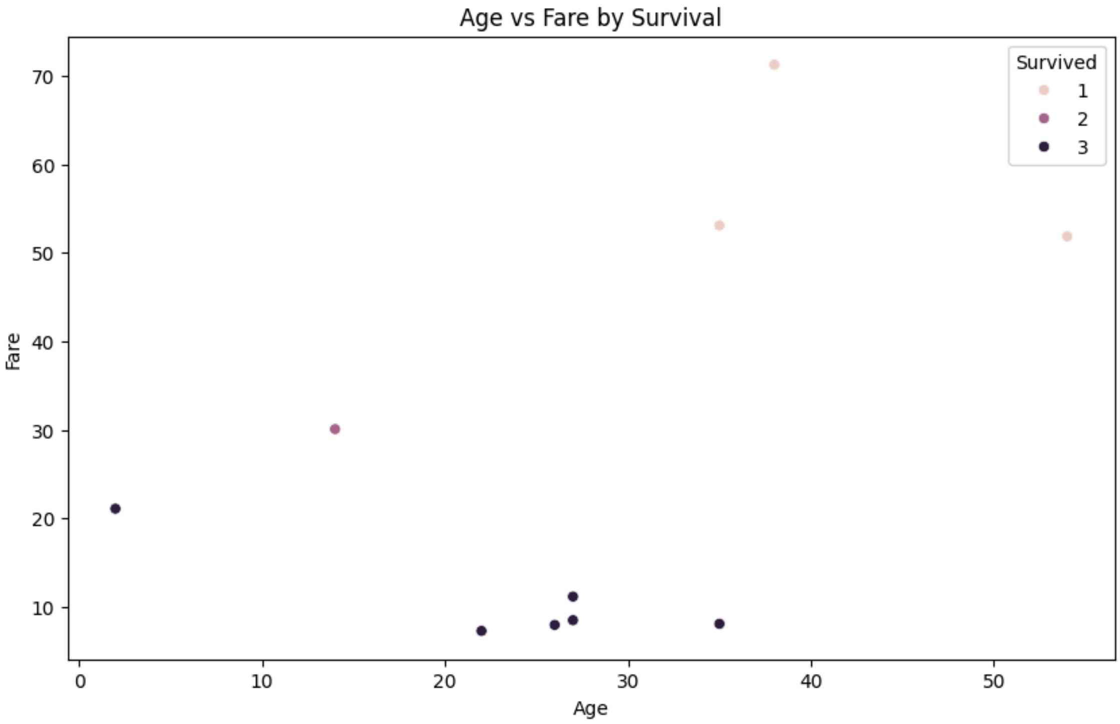

Distribution of Fare
Frequency
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
10
20
30
40
50
60
70
Fare

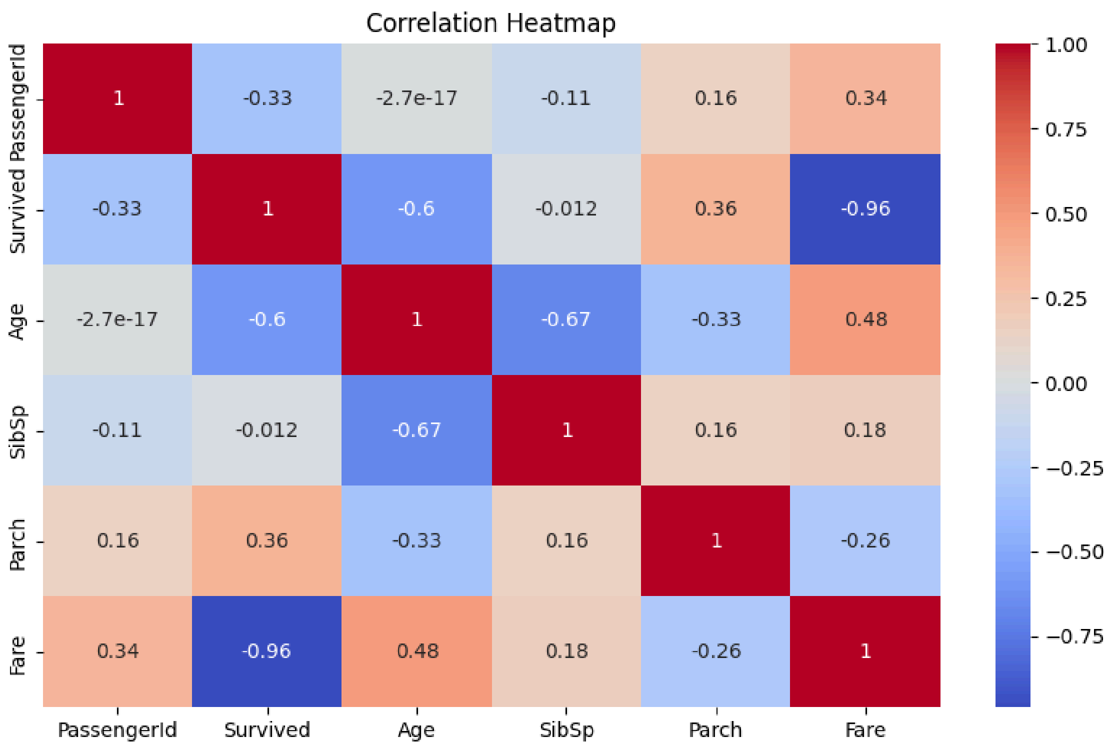


Box plot of Age by Survival
Age
Survived
0
10
20
30
40
50
1
2
3

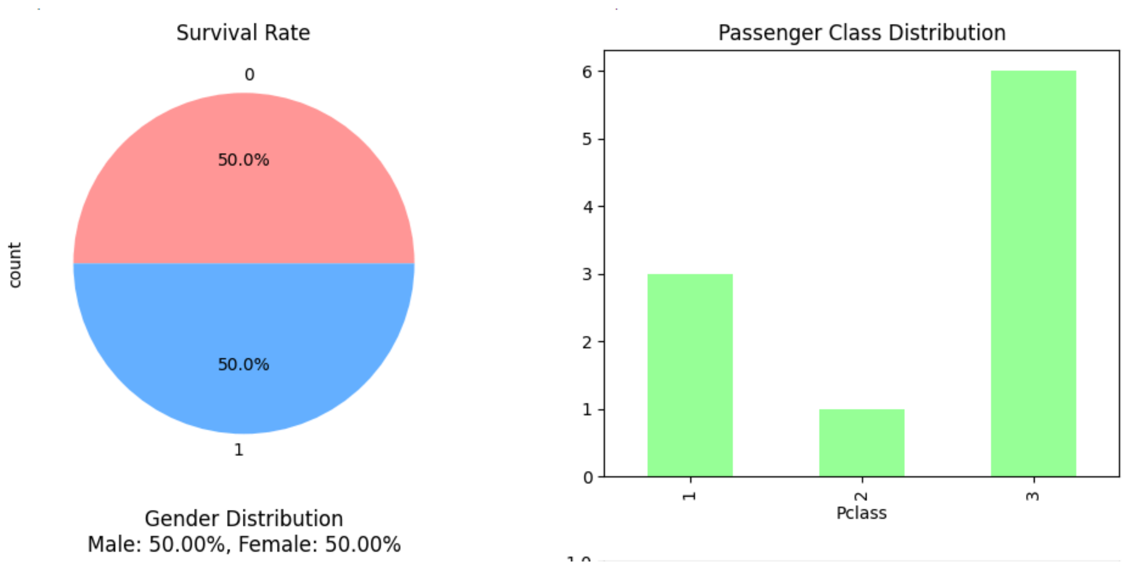


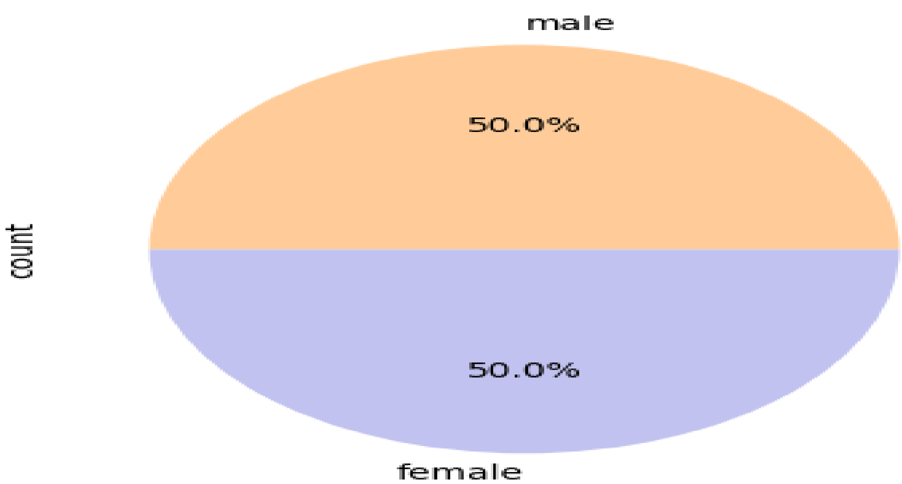
Box plot of Fare by Pclass
Fare
Pclass
10
20
30
40
50
60
70
Braund
Cumings
Heikkinen
Futrelle
Allen
Moran
McCarthy
Palsson
Johnson
Nasser


Age vs Fare by Survival
Survived
1
2
3
Fare
70
60
50
40
30
20
10
0
10
20
30
40
50
Age

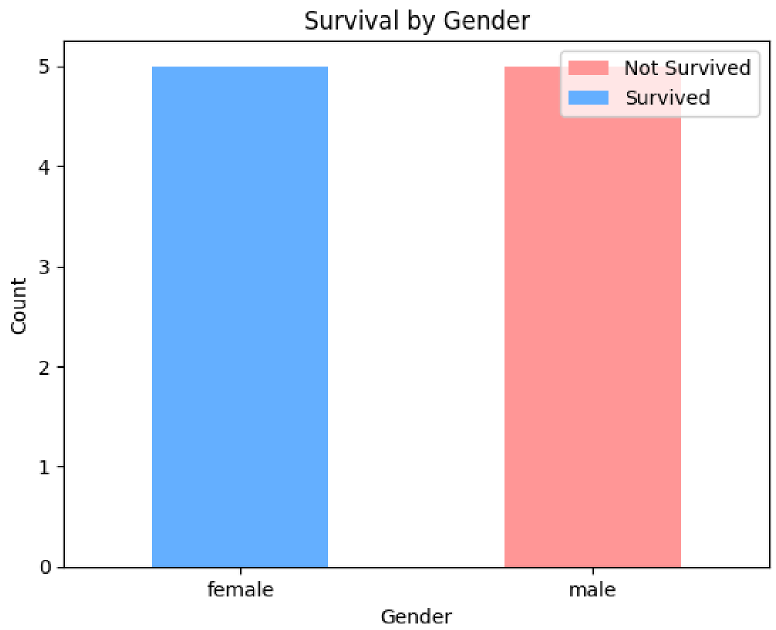

Correlation Heatmap
PassengerId
Survived
Age
SibSp
Parch
Fare
1 -0.33 -2.7e-17 -0.11 0.16 0.34
-0.33 1 -0.6 -0.012 0.36 -0.96
-2.7e-17 -0.6 1 -0.67 -0.33 0.48
-0.11 -0.012 -0.67 1 0.16 0.18
0.16 0.36 -0.33 0.16 1 -0.26
0.34 -0.96 0.48 0.18 -0.26 1
PassengerId
Survived
Age
SibSp
Parch
Fare
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

```python
# Correcting Gender Distribution
gender_counts = df['Sex'].value_counts()
male_percentage = (gender_counts['male'] / gender_counts.sum()) * 100
female_percentage = (gender_counts['female'] / gender_counts.sum()) * 100

# Visualizing key findings
fig, axes = plt.subplots(2, 2, figsize=(12, 10))

# Survival Rate
df['Survived'].value_counts().plot(kind='pie', autopct='%1.1f%%', ax=axes[0, 0], colors=['#ff9999','#66b3ff'])
axes[0, 0].set_title('Survival Rate')

# Passenger Class Distribution
df['Pclass'].value_counts().sort_index().plot(kind='bar', ax=axes[0, 1], color='#99ff99')
axes[0, 1].set_title('Passenger Class Distribution')

# Gender Distribution (Corrected)
gender_counts.plot(kind='pie', autopct='%1.1f%%', ax=axes[1, 0], colors=['#ffcc99','#c2c2f0'])
axes[1, 0].set_title(f'Gender Distribution\nMale: {male_percentage:.2f}%, Female: {female_percentage:.2f}%')
```





```python
df = pd.DataFrame(data)

# Grouping by Sex and Survived, then counting occurrences
survival_by_gender = df.groupby(['Sex', 'Survived']).size().unstack()

# Plotting the bar chart
survival_by_gender.plot(kind='bar', stacked=True, color=['#ff9999', '#66b3ff'])
plt.title('Survival by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.xticks(rotation=0)
plt.legend(['Not Survived', 'Survived'], loc='upper right')
plt.show()
```



Summarize key findings:

- None of the male had survived.
- Passengers in 1st class had higher survival rates than those in 2nd and 3rd classes.
- Age Between 20-30 Survived most.

Conclusion : the Titanic dataset provides valuable insights into the dynamics of survival during the tragic event. Factors such as gender, passenger class, and age played significant roles in determining the likelihood of survival. Further analysis could delve deeper into these factors and their interactions to gain a more comprehensive understanding of the Titanic disaster and its impact on passengers.

References:https://www.kaggle.com/competitions/titanic