Task 5: Exploratory Data Analysis (EDA) – Titanic Dataset

Objective

The objective of this analysis is to explore the Titanic dataset using statistical and visual techniques in order to identify patterns, trends, and relationships that influence passenger survival.

Tools Used

- Python
- Pandas
- Matplotlib
- Seaborn

In [3]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
```

Dataset Loading

The Titanic dataset is loaded for exploratory data analysis. The dataset contains passenger demographic and travel-related information.

In [2]:
```
import pandas as pd

df = pd.read_csv(r"C:\Users\jeeva\OneDrive\Desktop\task_5\data\train.csv")
df.head()
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0

Data Overview

This section provides a high-level understanding of the dataset, including its dimensions and column structure.

In [5]:
```
df.shape
```

Out[5]:
```
(891, 12)
```

Data Structure and Missing Values

Understanding data types and missing values is essential before performing further analysis.

In [6]:
```
df.info()
```

```
<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
```

Statistical Summary

Descriptive statistics help in understanding the distribution and central tendency of numerical variables.

In [7]:
```
df.describe()
```

Out[7]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329

Target Variable: Survival

Analyzing the distribution of the target variable helps understand class balance.

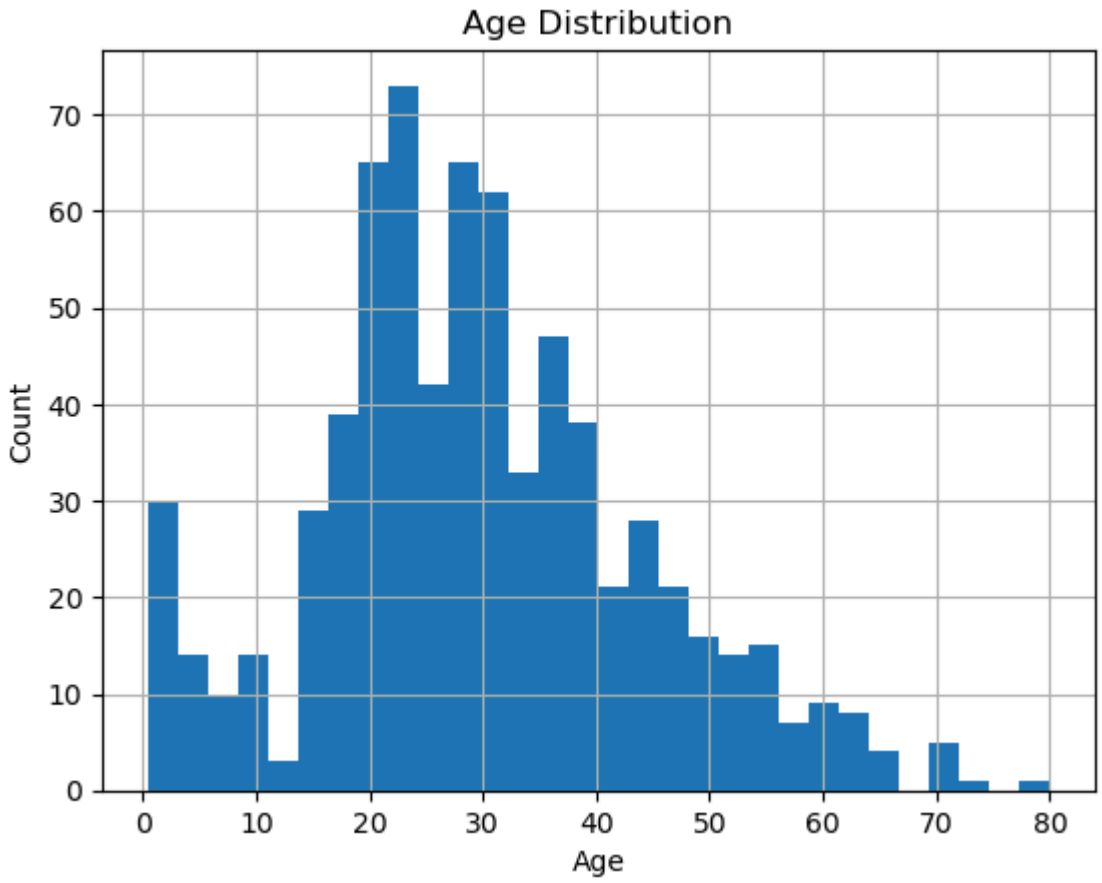
In [8]:
```
df['Survived'].value_counts()
```

Out[8]:
```
Survived
0    549
1    342
Name: count, dtype: int64
```

Univariate Analysis

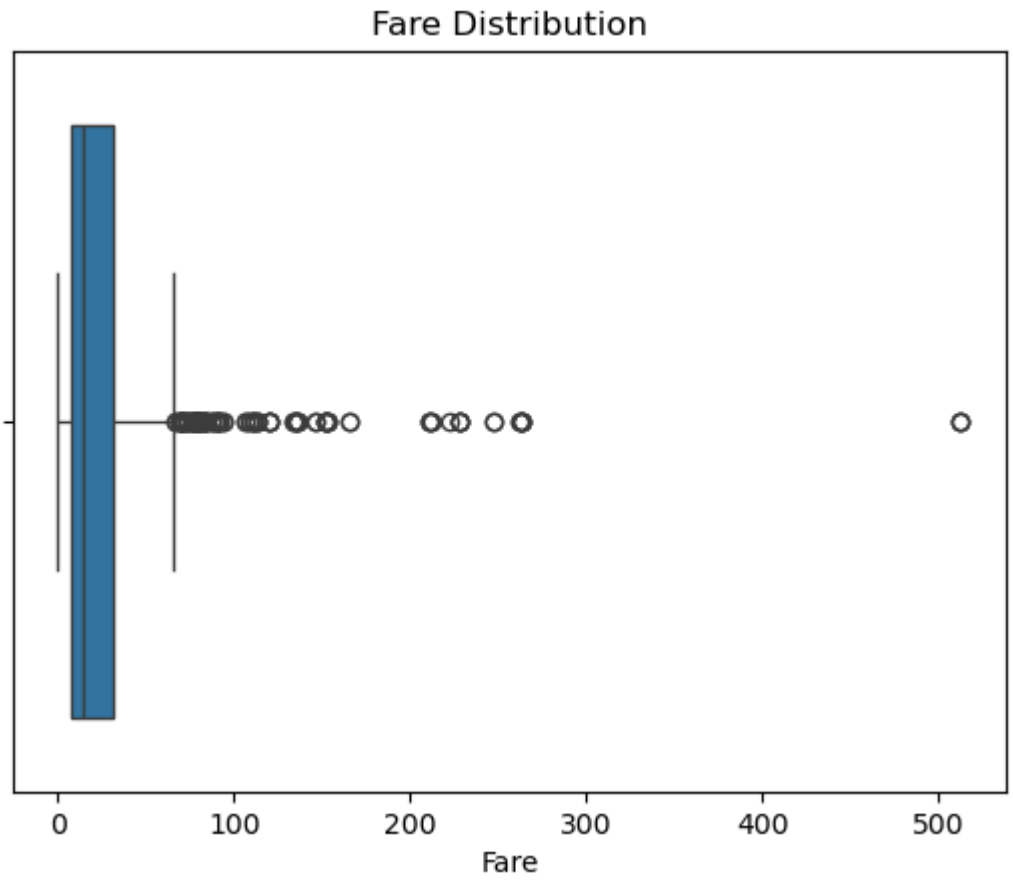
Univariate analysis focuses on examining individual variables independently.

In [9]:
```
df['Age'].hist(bins=30)
plt.title("Age Distribution")
plt.xlabel("Age")
plt.ylabel("Count")
plt.show()
```



- Most passengers fall in the age range of 20–40 years.
- The distribution is slightly right-skewed.

In [10]:
```
sns.boxplot(x=df['Fare'])
plt.title("Fare Distribution")
plt.show()
```

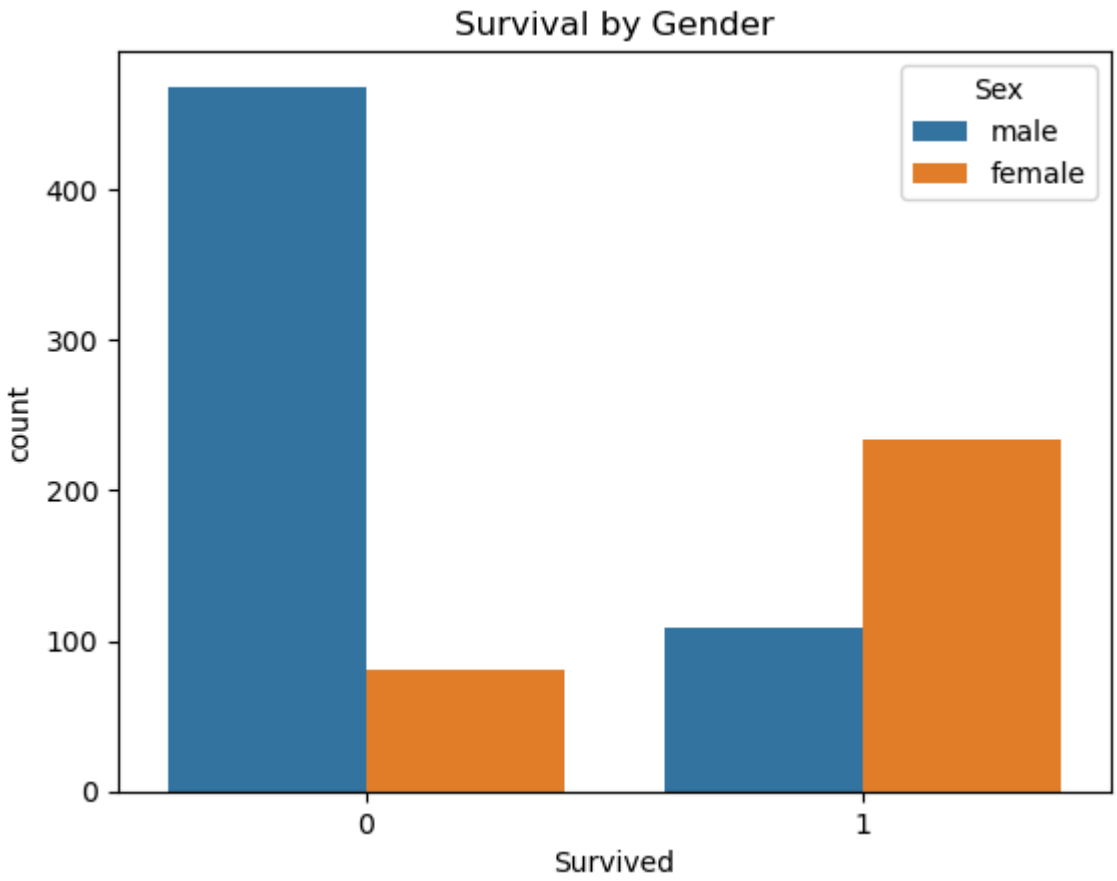


- Fare values are highly skewed.
- Presence of several extreme outliers.

Bivariate Analysis

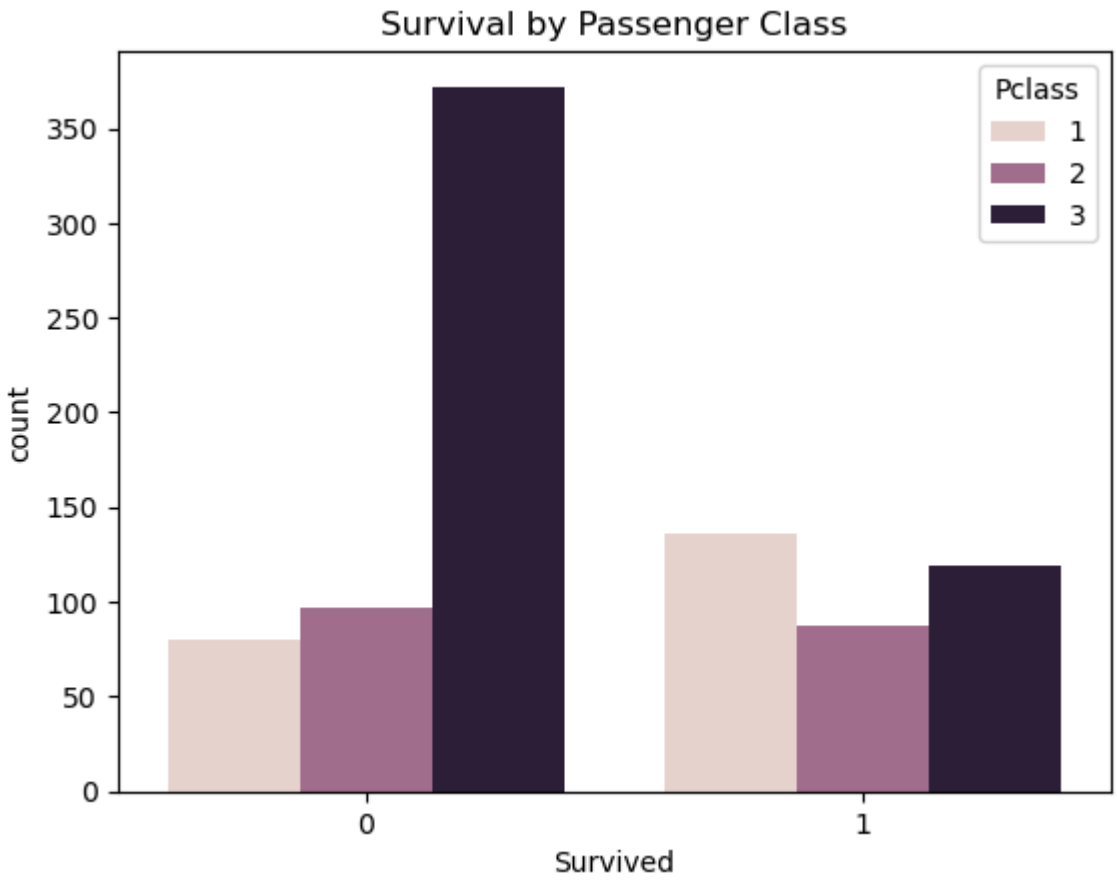
Bivariate analysis explores relationships between two variables.

In [11]:
```
sns.countplot(x='Survived', hue='Sex', data=df)
plt.title("Survival by Gender")
plt.show()
```



- Female passengers had a significantly higher survival rate.
- Gender is a strong predictor of survival.

In [12]:
```
sns.countplot(x='Survived', hue='Pclass', data=df)
plt.title("Survival by Passenger Class")
plt.show()
```



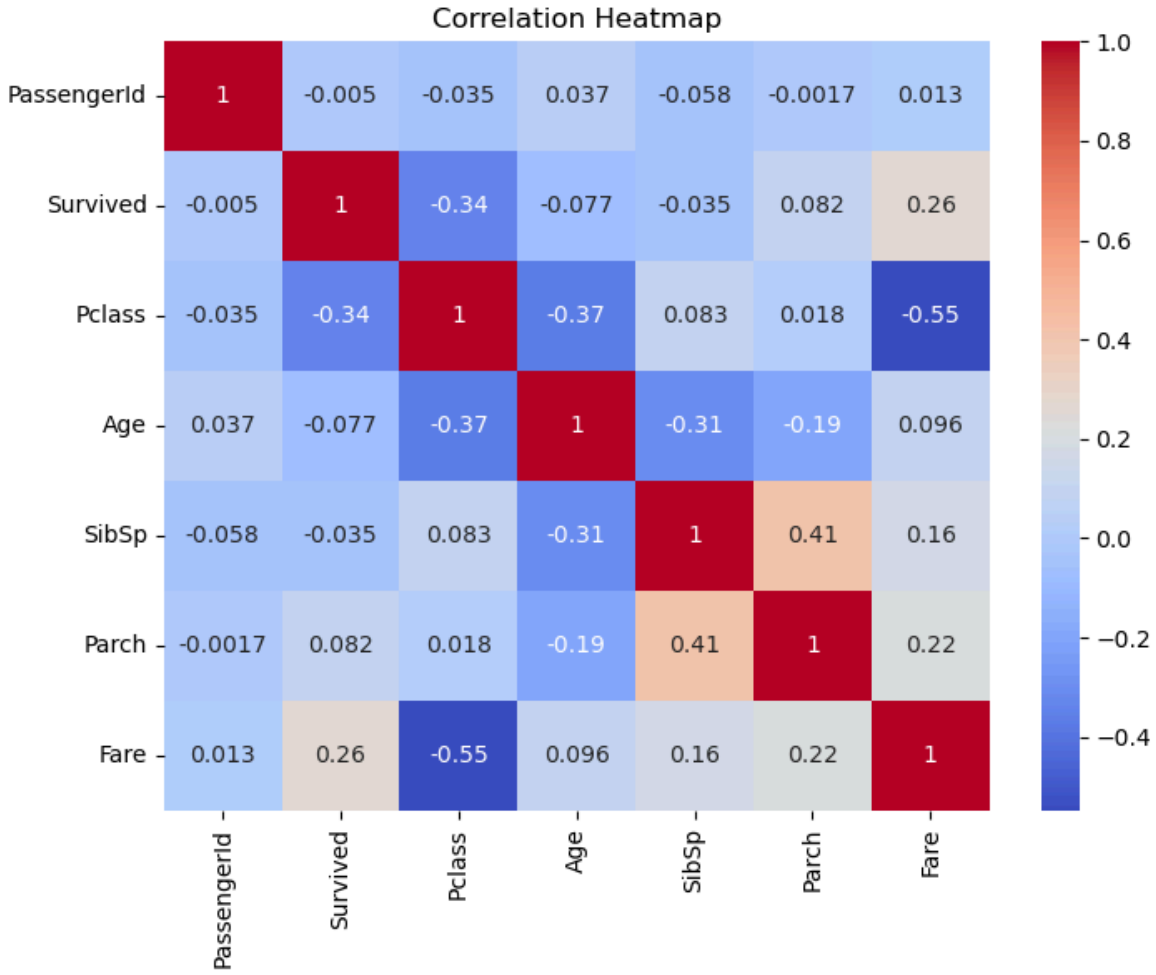
- First-class passengers survived at a higher rate.
- Third-class passengers had the lowest survival rate.

Multivariate Analysis

Multivariate analysis examines interactions between multiple variables simultaneously.

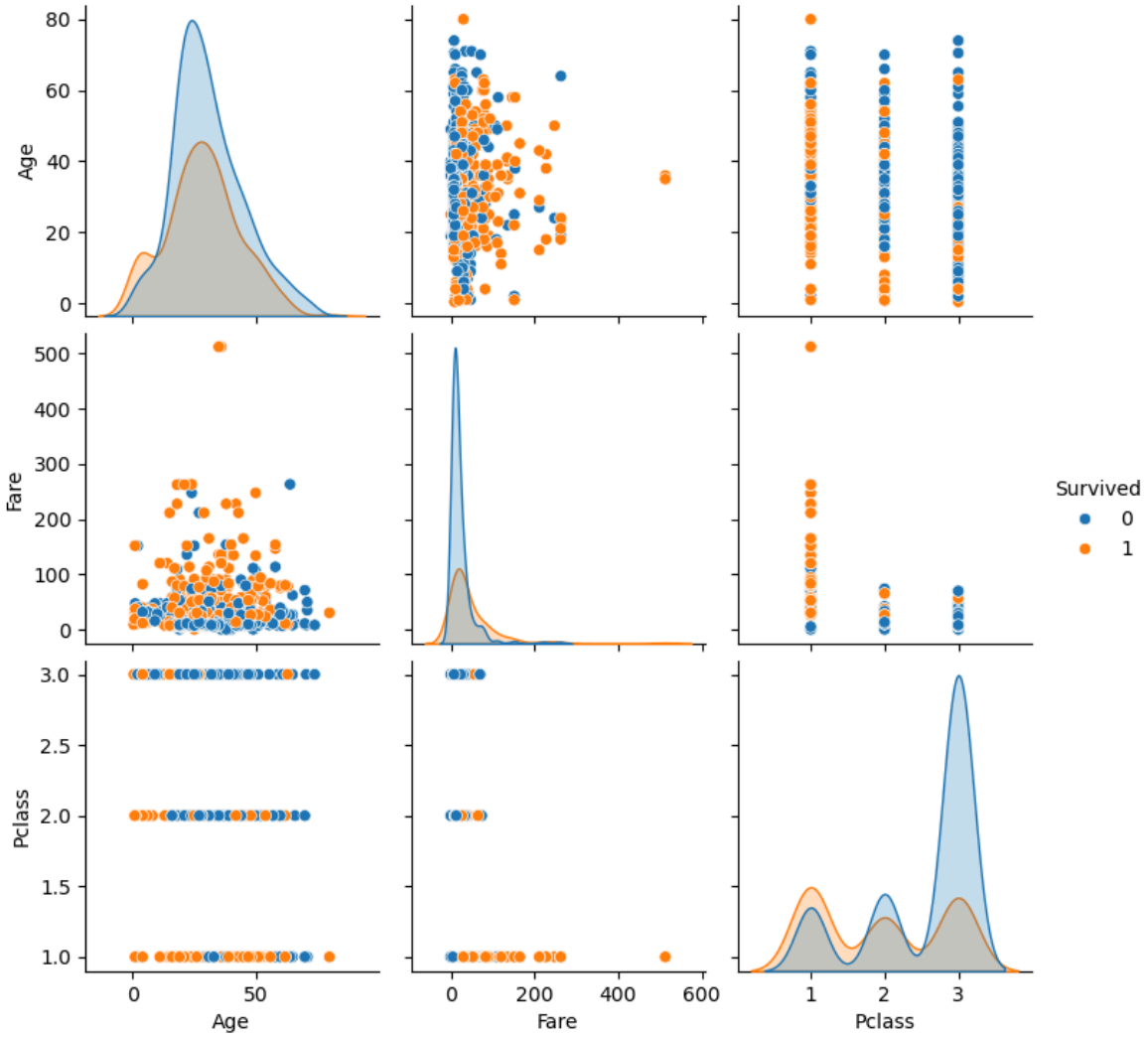
In [14]:
```
plt.figure(figsize=(8,6))
sns.heatmap(df.select_dtypes(include='number').corr(),
            annot=True,
            cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```

```
<Figure size 800x600 with 0 Axes>
```



- Survival is moderately correlated with Fare and Passenger Class.
- Age shows weak correlation with survival.

In [15]:
```
sns.pairplot(df[['Survived','Age','Fare','Pclass']], hue='Survived')
plt.show()
```



- Survivors tend to have higher fares.
- Clear separation observed across passenger classes.

Summary of Findings

- Female and first-class passengers had higher survival rates.
- Fare is highly skewed and contains outliers.
- Passenger class and fare significantly influence survival.
- Age alone is not a strong predictor of survival.

Conclusion

This exploratory data analysis revealed that gender, passenger class, and fare played significant roles in survival outcomes during the Titanic disaster. Female passengers and those traveling in higher classes had higher chances of survival, while age alone showed weak influence. These insights highlight how socio-economic factors influenced survival probability.