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

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

# First few rows
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

```
# Shape of dataset
df.shape

# Info about datatypes and nulls
df.info()

# Summary statistics
df.describe()

# Check missing values
df.isnull().sum()
```

```
<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
```

	0
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

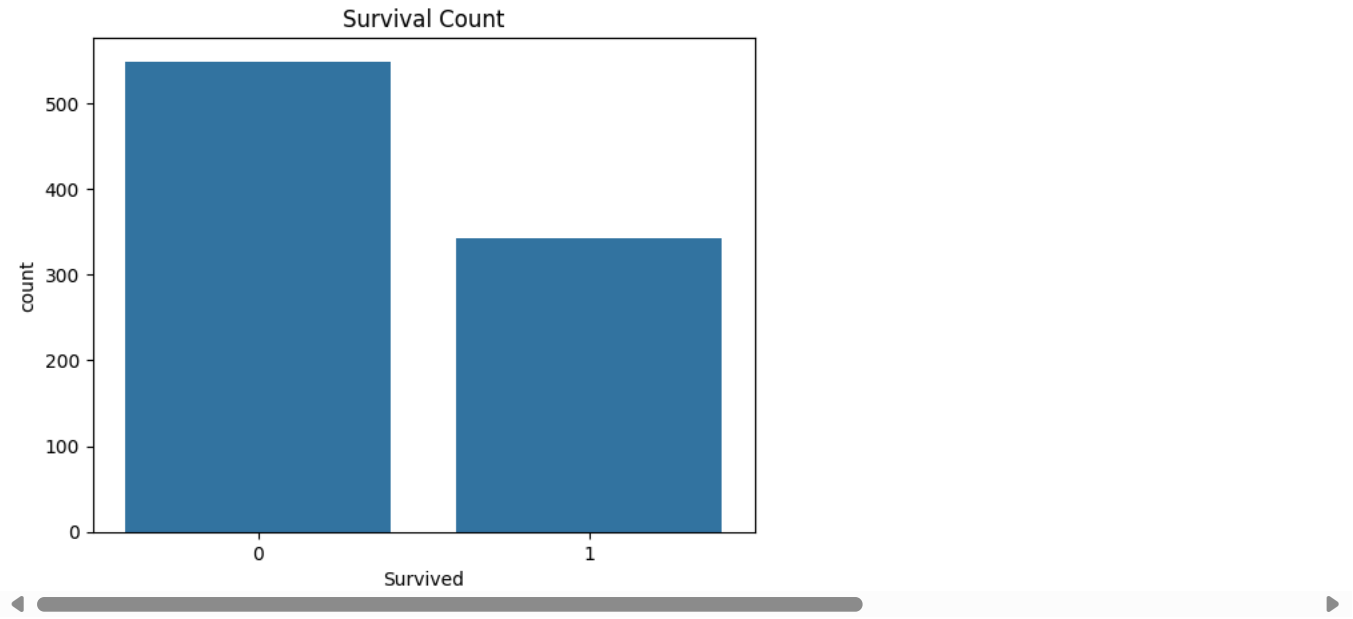
Univariate Analysis

🎯 Target Variable: Survived

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.countplot(x='Survived', data=df)
plt.title("Survival Count")
```

```
Text(0.5, 1.0, 'Survival Count')
```



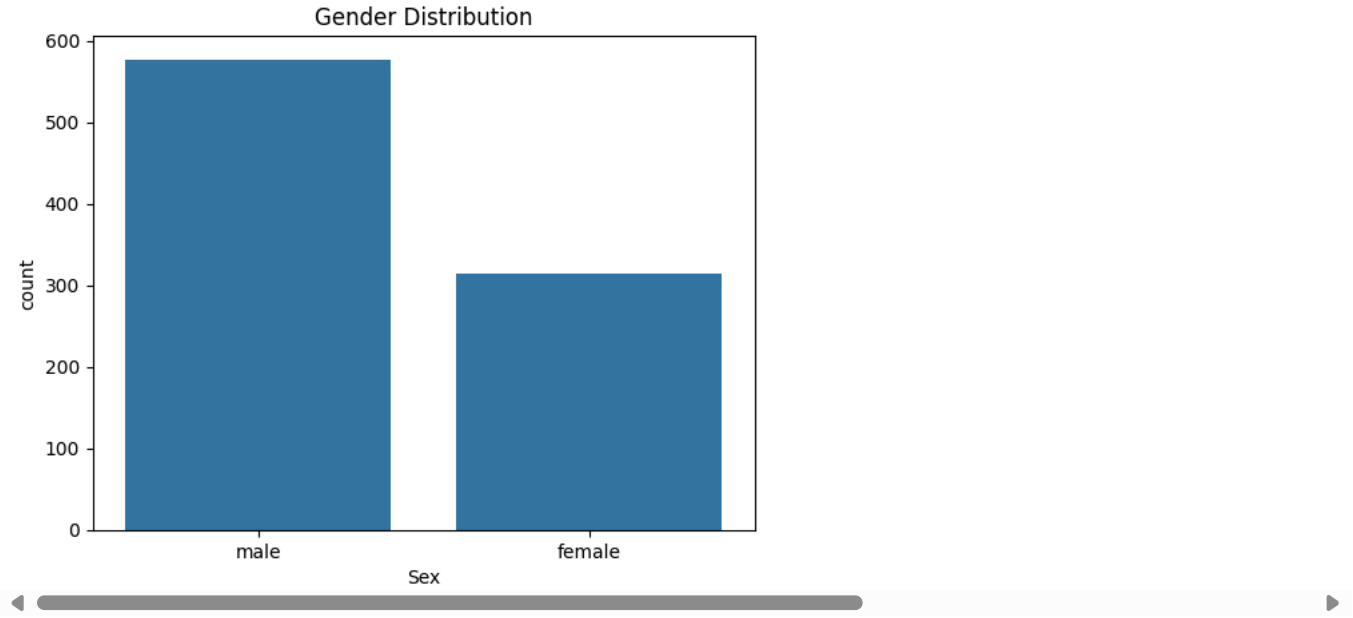
Interpretation: Survived = 0 (Not Survived): Around 550+ passengers did not survive the Titanic disaster.

Survived = 1 (Survived): Around 340+ passengers survived.

Gender Distribution

```
sns.countplot(x='Sex', data=df)
plt.title("Gender Distribution")
```

```
Text(0.5, 1.0, 'Gender Distribution')
```



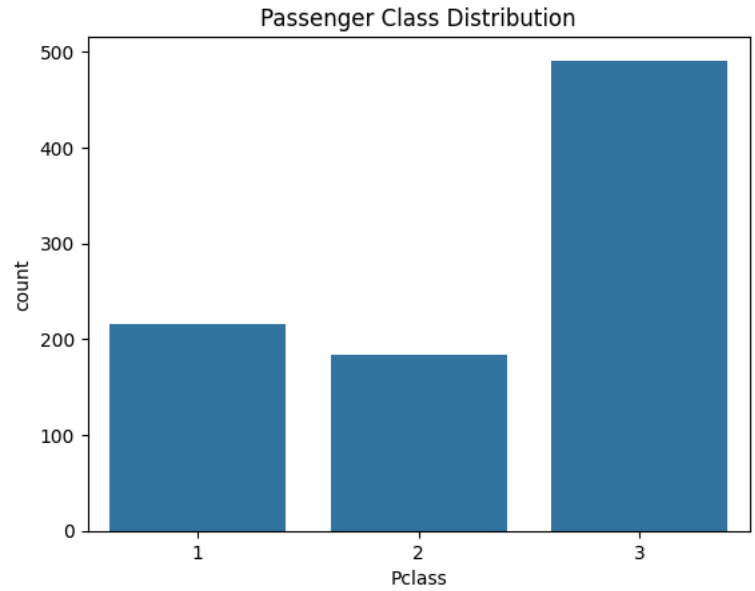
Interpretation: Male Passengers: Approximately 575–590 male passengers were on board.

Female Passengers: Approximately 315–330 female passengers were on board.

Passenger Class

```
sns.countplot(x='Pclass', data=df)
plt.title("Passenger Class Distribution")
```

```
Text(0.5, 1.0, 'Passenger Class Distribution')
```



Interpretation: 3rd Class (Pclass = 3) had the highest number of passengers — nearly 500.

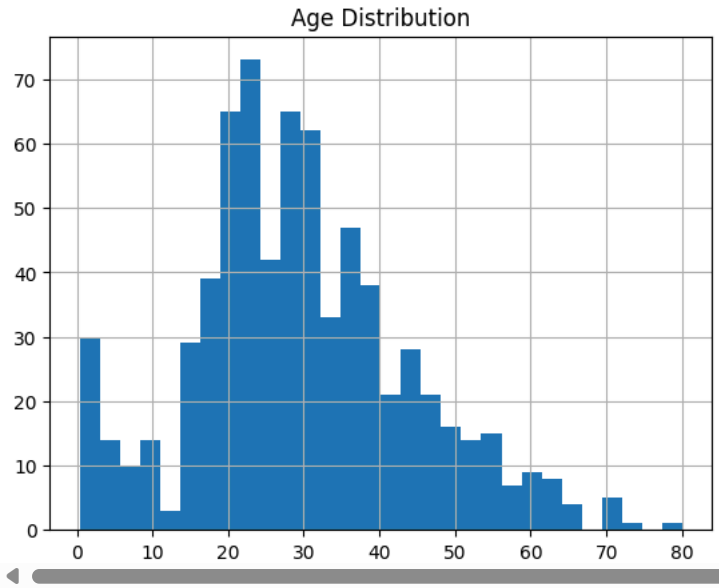
1st Class (Pclass = 1) had around 215 passengers.

2nd Class (Pclass = 2) had the fewest passengers, roughly 185.

Age Distribution

```
df['Age'].hist(bins=30)
plt.title("Age Distribution")
```

```
Text(0.5, 1.0, 'Age Distribution')
```



Interpretation: The majority of passengers were aged between 20 to 40 years, with peaks around:

20–25 years (highest frequency)

30–35 years (second-highest)

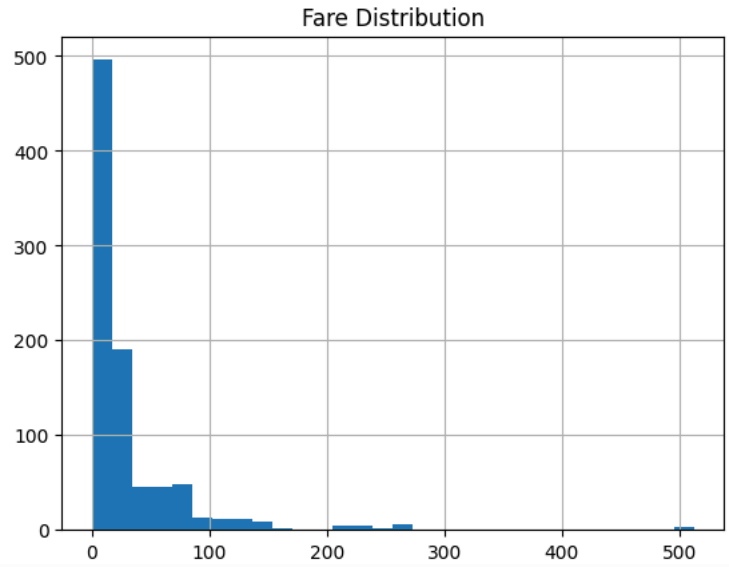
There is also a noticeable number of children under the age of 10.

Fewer elderly passengers (above 60 years) were on board.

Fare Distribution

```
df['Fare'].hist(bins=30)
plt.title("Fare Distribution")
```

```
Text(0.5, 1.0, 'Fare Distribution')
```



Interpretation: The majority of fares are concentrated between $0 and 50$.

The distribution is highly right-skewed (positively skewed):

A few passengers paid very high fares (over $100, even upto 500$).

These are likely 1st-class or luxury cabin tickets.

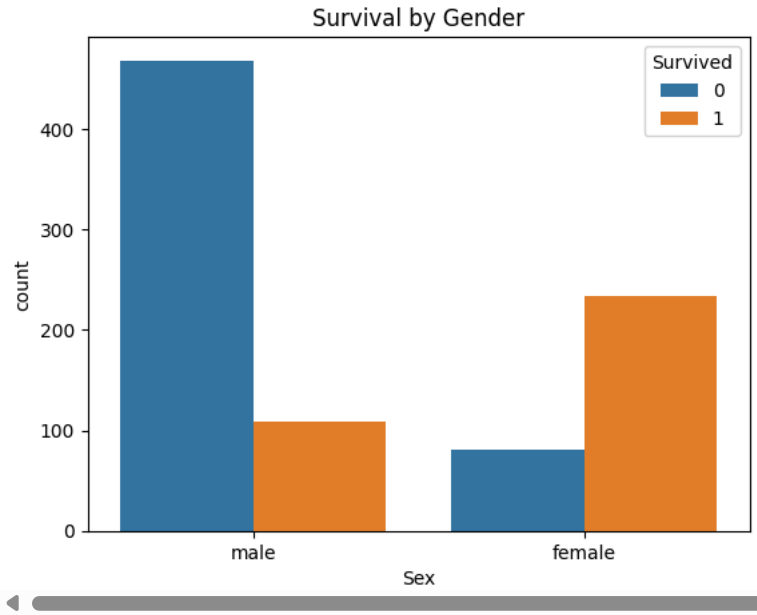
Most passengers likely belonged to 3rd or 2nd class, as reflected in the lower fare values.

Bivariate Analysis

Survival vs Gender

```
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title("Survival by Gender")
```

```
Text(0.5, 1.0, 'Survival by Gender')
```



Interpretation: Males:

Majority did not survive (tall blue bar).

Only a small portion survived (short orange bar).

Females:

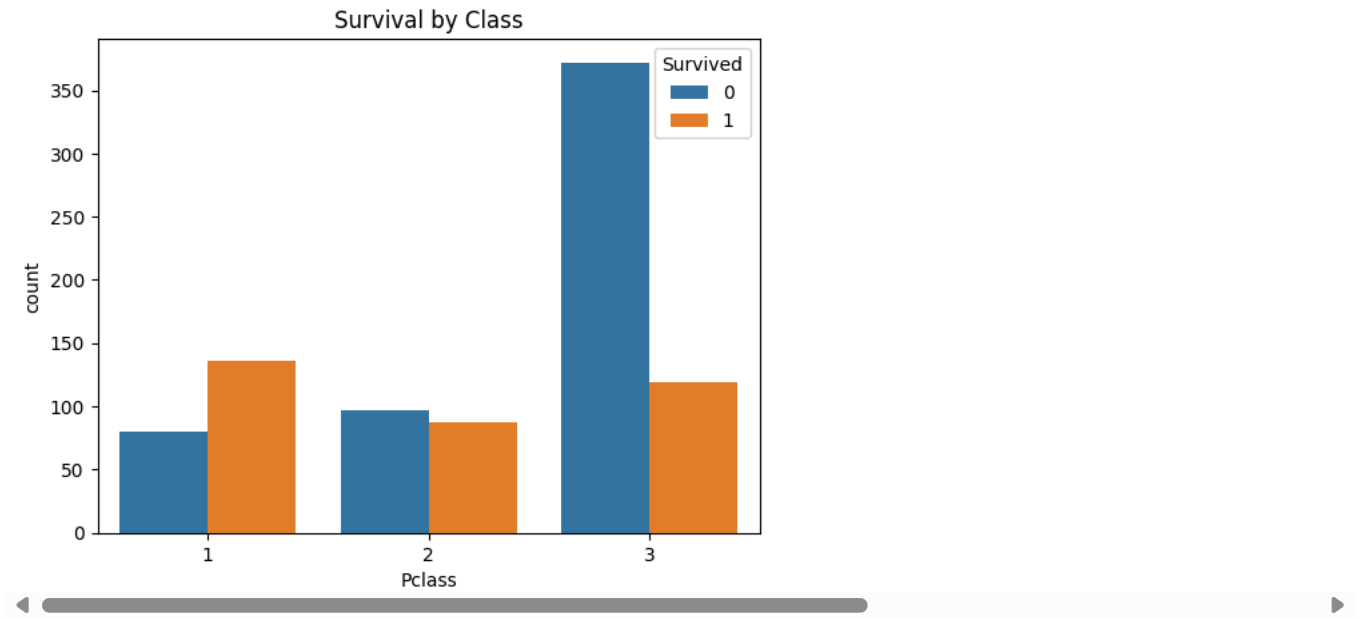
Majority survived (tall orange bar).

Fewer females did not survive (short blue bar).

Survival vs Passenger Class

```
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title("Survival by Class")
```

```
Text(0.5, 1.0, 'Survival by Class')
```



Interpretation: 1st Class:

Majority survived (tall orange bar).

Fewer deaths (short blue bar).

2nd Class:

Survival and death counts are almost equal (orange ≈ blue).

3rd Class:

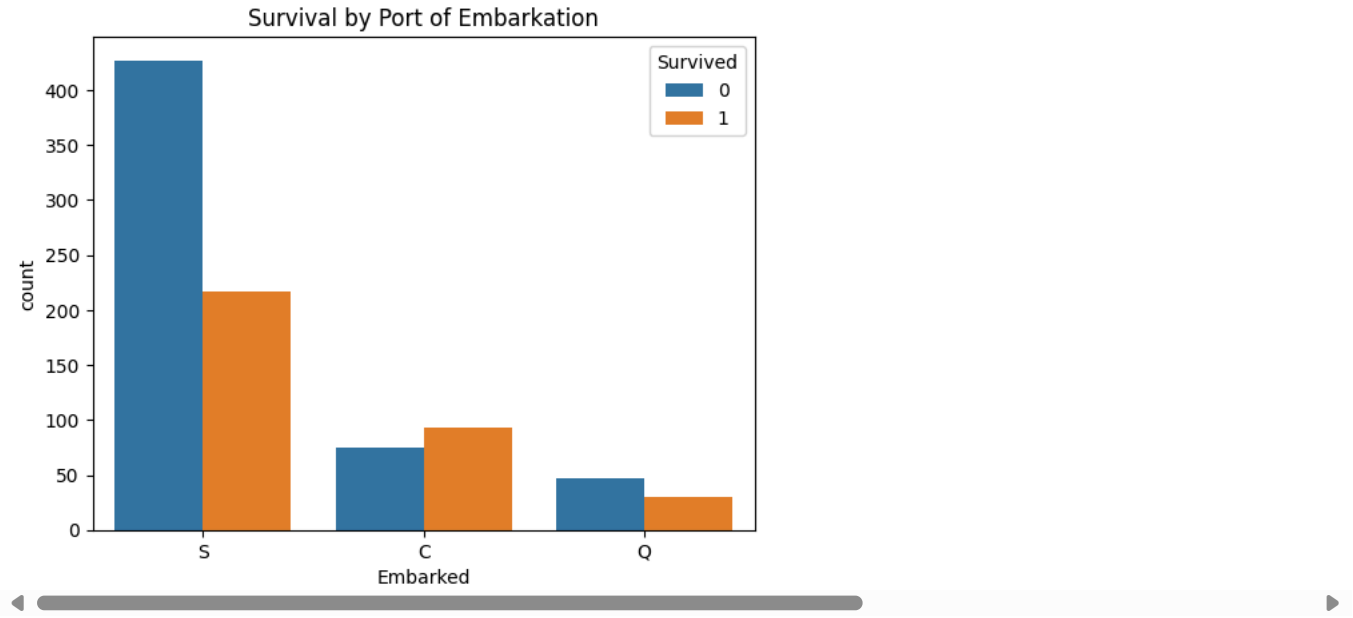
Majority did not survive (very tall blue bar).

Much fewer survivors (short orange bar).

Survival vs Embarked

```
sns.countplot(x='Embarked', hue='Survived', data=df)
plt.title("Survival by Port of Embarkation")
```

```
Text(0.5, 1.0, 'Survival by Port of Embarkation')
```



Interpretation: The fare distribution is right-skewed (positively skewed).

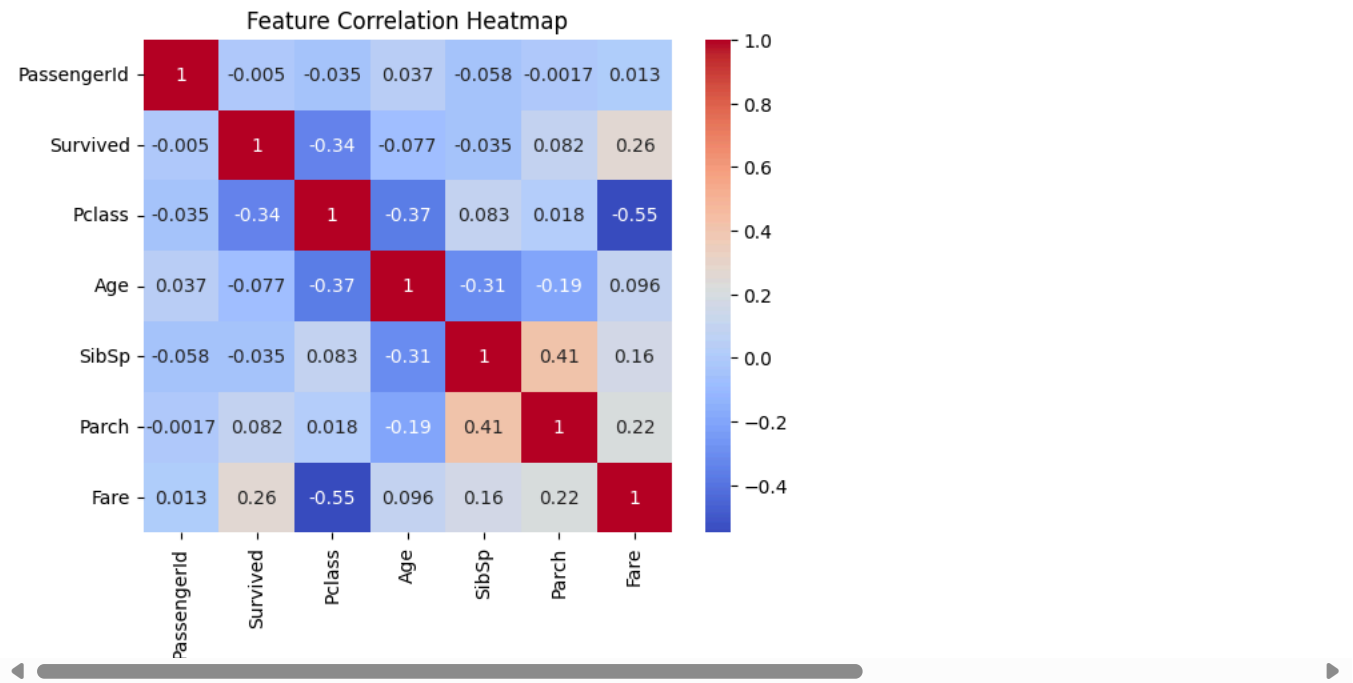
Most passengers paid low fares (clustered around $0–50).

A few outliers paid very high fares (up to over $500), but these are rare.

Multivariate Analysis

Heatmap of Correlation

```
sns.heatmap(df.select_dtypes(include='number').corr(), annot=True, cmap='coolwarm')
plt.title("Feature Correlation Heatmap")
plt.show()
```



Observations on the Feature Correlation Heatmap PassengerId:

Shows near-zero correlations with all features (values close to 0), as expected for an identifier variable.

Survived:

Moderate negative correlation with Pclass (-0.34): Lower-class passengers (higher Pclass values) were less likely to survive.

Positive correlation with Fare (0.26): Higher fare-paying passengers had better survival rates, likely linked to higher socio-economic status (e.g., first-class passengers).

Pclass:

Strong negative correlation with Fare (-0.55): Confirms that lower-class passengers paid lower fares.

Negative correlation with Age (-0.37): Suggests older passengers were more likely to be in higher classes (e.g., first class).

Age:

Negative correlation with SibSp (-0.31): Older passengers tended to have fewer siblings/spouses aboard.

Weak correlation with `Survived (-0.077)**: Age alone is not a strong predictor of survival in this dataset.

SibSp & Parch:

Positive correlation (0.41): Passengers with more siblings/spouses (SibSp) often had more parents/children (Parch) aboard, indicating family groups.

Fare:

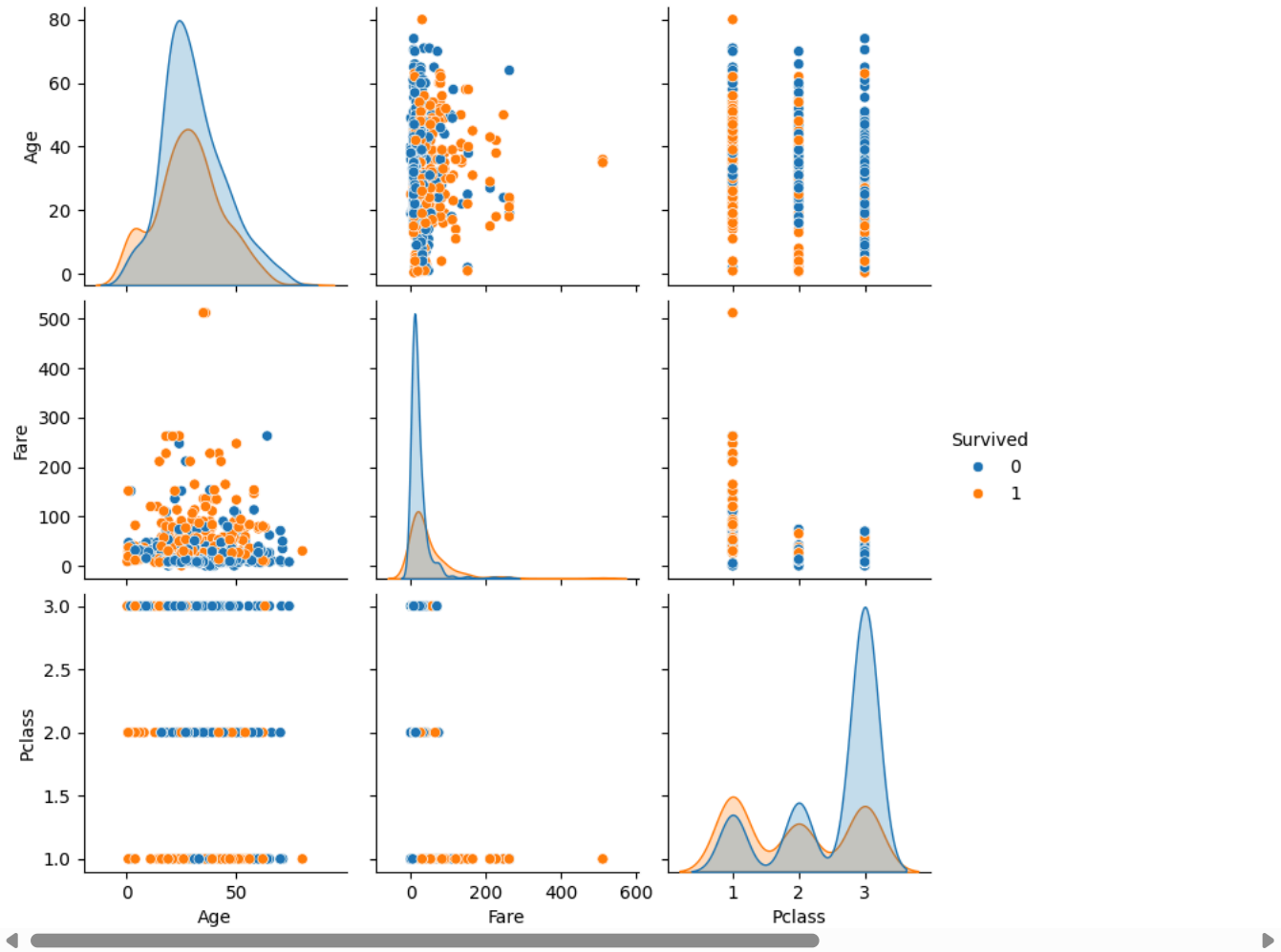
Strong negative correlation with Pclass (-0.55): Reinforces the link between fare and class.

Positive correlation with Survived (0.26): Further supports the connection between fare and survival outcomes.

Pairplot for selected features

```
sns.pairplot(df[['Survived', 'Age', 'Fare', 'Pclass']], hue='Survived')
```

```
<seaborn.axisgrid.PairGrid at 0x7d5151ab5210>
```



Survival Distribution by Variables:

Age:

The histograms for Age show overlapping distributions for survivors (1) and non-survivors (0), suggesting age alone is not a strong predictor of survival.

Scatterplots for Age vs Survived (vertical/horizontal clusters) confirm weak correlation, aligning with the heatmap findings.

Fare:

Survivors (1) cluster more densely at higher Fare values, supporting the earlier observation that higher fares correlate with better survival rates.

The histogram for Fare is heavily right-skewed, with most fares concentrated below 200.

Pclass:

Survivors (1) are more frequent in lower Pclass values (e.g., 1st class), while non-survivors dominate in higher classes (e.g., 3rd class).

Relationships Between Features:

Fare vs Pclass:

Lower classes (Pclass = 3) cluster at lower fares, while higher classes (Pclass = 1) show a wider fare range, including very high values (up to ~600).

Age vs Pclass:

Older passengers are slightly more common in lower Pclass (1st class), though the trend is not strong.

Age vs Fare:

No clear linear relationship, but younger passengers (<20) tend to have lower fares.

Key Insights

Women had a much higher survival rate than men.

Passengers in 1st class had higher survival probability.

Younger passengers had slightly higher survival.