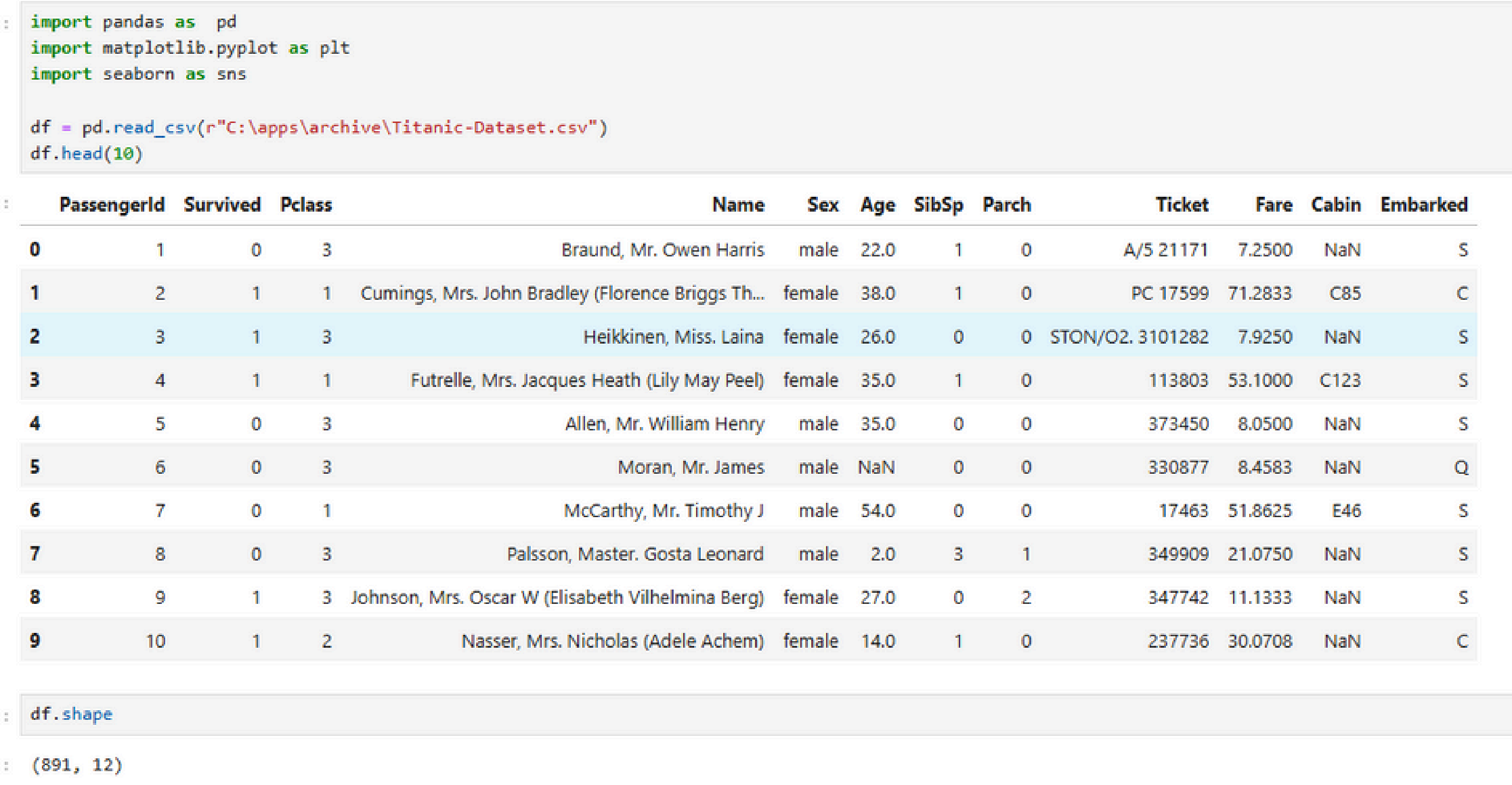
Titanic Dataset Analysis Report

1. Introduction

This report presents a detailed exploratory data analysis (EDA) of the Titanic dataset, including statistical insights, visualization findings, and hypothesis testing. The objective is to identify key patterns, relationships, and significant factors affecting survival rates. By leveraging data cleaning, feature analysis, and statistical tests, this report aims to provide a comprehensive understanding of the dataset.

2. Data Overview

The Titanic dataset consists of passenger details, including demographics, socio-economic status, and survival status. The dataset contains the following key features:

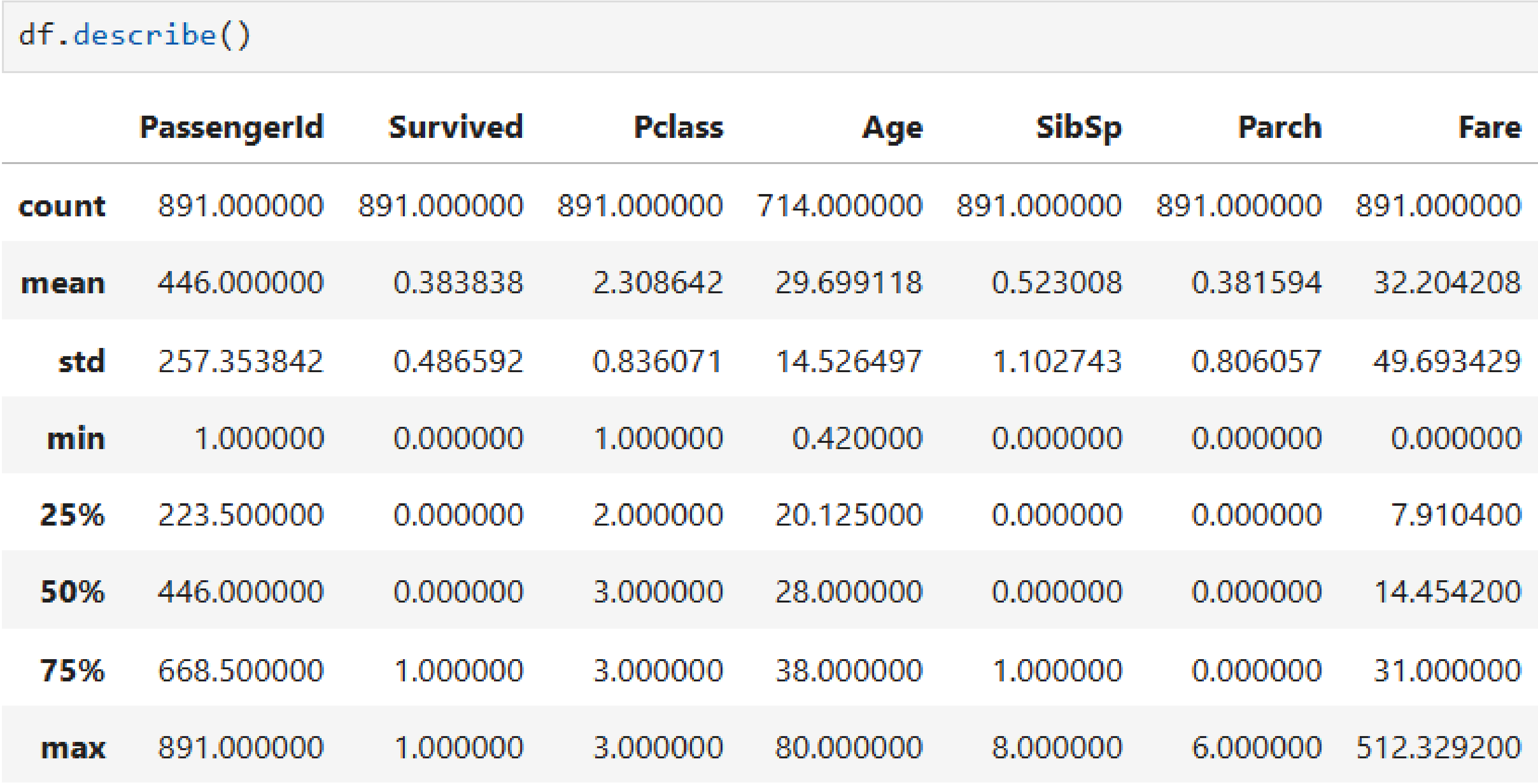


```
import pandas as  pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv(r"C:\apps\archive\Titanic-Dataset.csv")
df.head(10)
```

	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
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

```
df.shape
```

```
(891, 12)
```



```
df.describe()
```

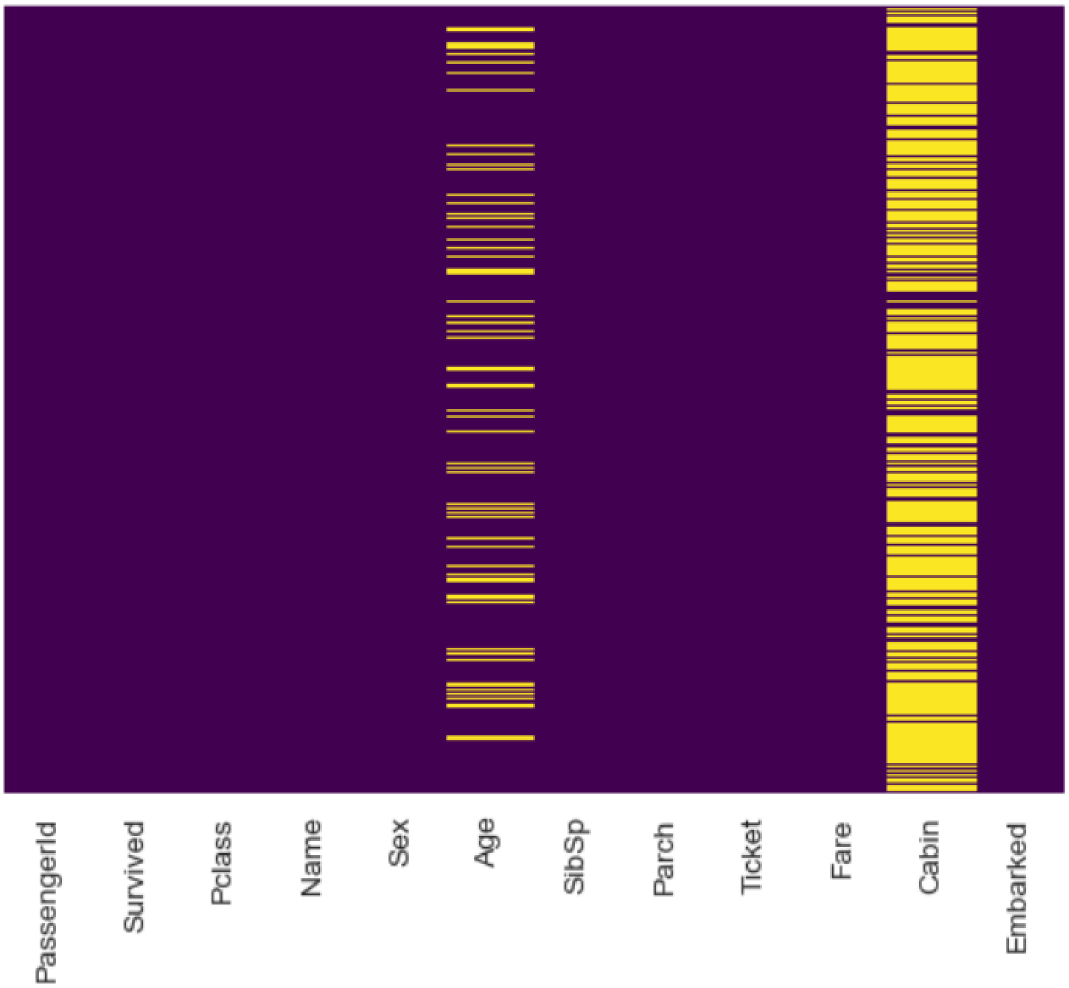
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

- Survived: 0 = No, 1 = Yes
- Pclass: Ticket class (1st, 2nd, 3rd)
- Sex: Male or Female
- Age: Passenger age
- SibSp: Number of siblings/spouses aboard
- Parch: Number of parents/children aboard
- Fare: Ticket price
- Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

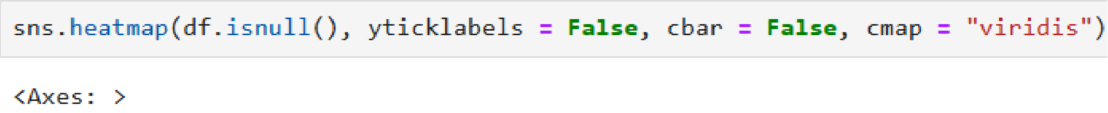
The dataset provides crucial insights into the survival factors of passengers aboard the Titanic. It allows us to analyze demographic influences, socio-economic status, and fare price contributions to survival probability.

3. Data Cleaning & Preprocessing

Data cleaning was a critical step in ensuring accurate analysis. The following steps were taken:

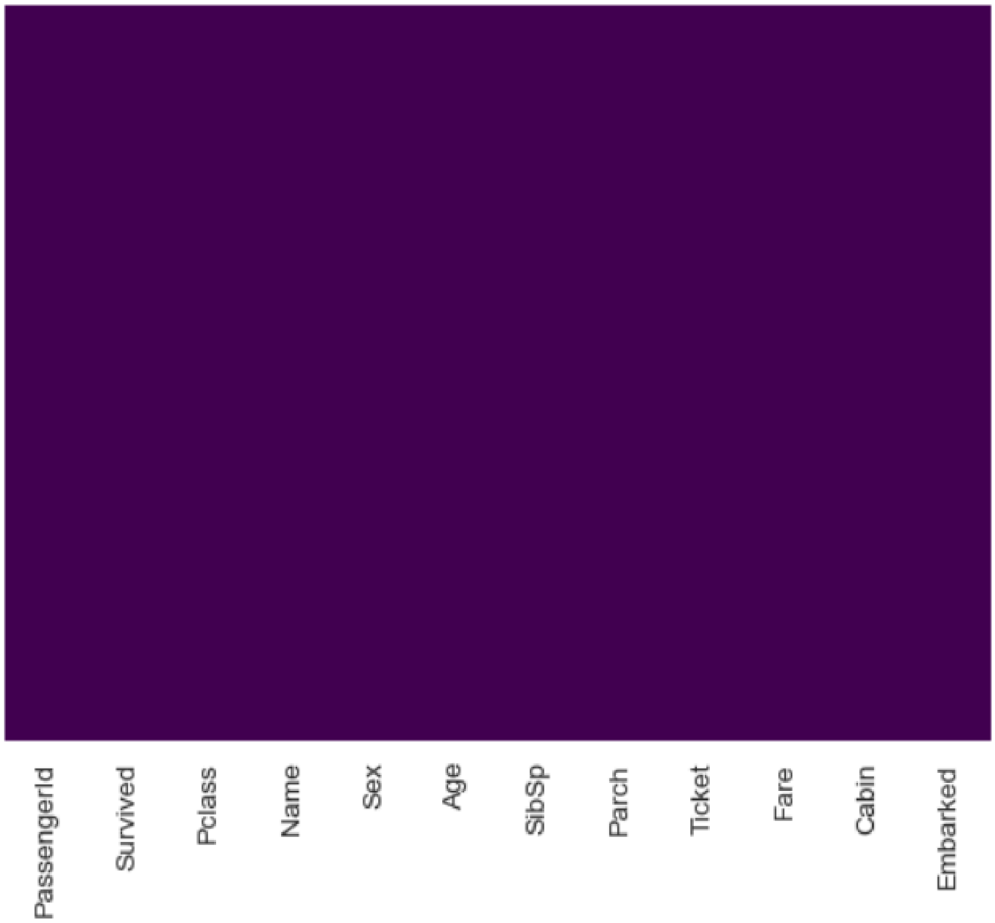


- **Handling Missing Values:**
  - **Age: Missing values were imputed using the median age of passengers by class to maintain consistency and reduce bias.**
  - **Embarked: Missing values were filled with the most frequent category, ensuring no bias in embarkation analysis.**
  - **Cabin: Many missing values were observed in the Cabin column.**
- **Data Type Conversions: Some categorical variables were converted into numerical form for analysis.**
- **Feature Engineering: Derived features such as Family Size (combining SibSp and Parch) were created to study their impact on survival.**



```
sns.heatmap(df.isnull(), yticklabels = False, cbar = False, cmap = "viridis")
```

```
<Axes: >
```



- **The correlation analysis revealed that fare and class had the strongest influence on survival, emphasizing the impact of socio-economic factors. Gender and age also played roles, though weaker in correlation strength. Small families had a slight advantage, but no strong correlation was found with family size. These findings reinforce that wealth, class, and age significantly influenced survival outcomes on the Titanic.**

Fare and Survival (Positive Correlation)

- Passengers who paid **higher fares** had **better survival chances**.
- This indicates that **wealthier individuals** (mostly first-class passengers) were prioritized during evacuation.

Pclass and Survival (Negative Correlation)

- Higher ticket class numbers (1 → 3) indicate lower class, and survival probability **decreases** as class increases.
- **First-class passengers had the highest survival rates**, while **third-class had the lowest**.
- This reflects socio-economic disparities in survival chances.

Fare and Pclass (Strong Negative Correlation)

- **Higher class (1st class) passengers paid significantly higher fares** compared to lower classes.

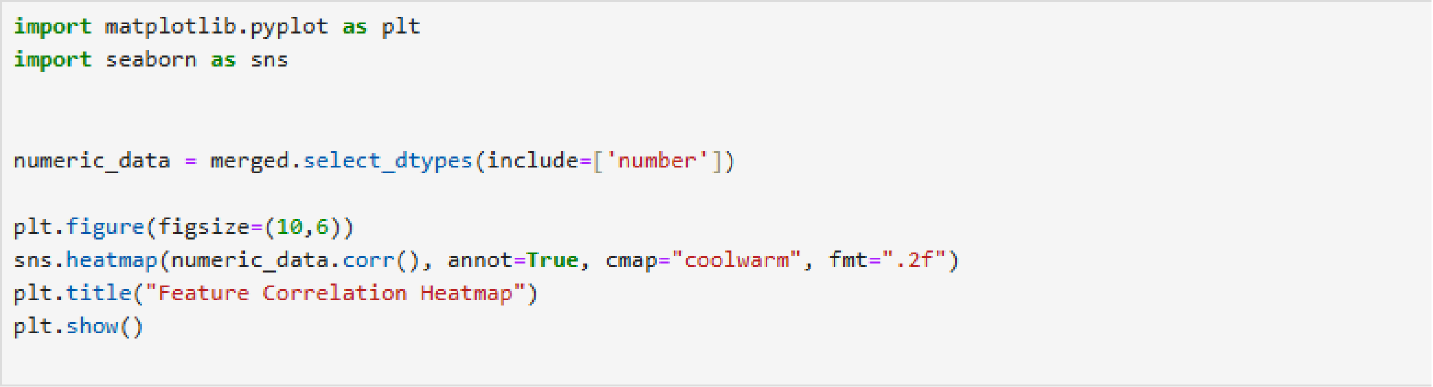
- This confirms that ticket price was strongly linked to socio-economic status.

Age and Survival (Weak Negative Correlation)

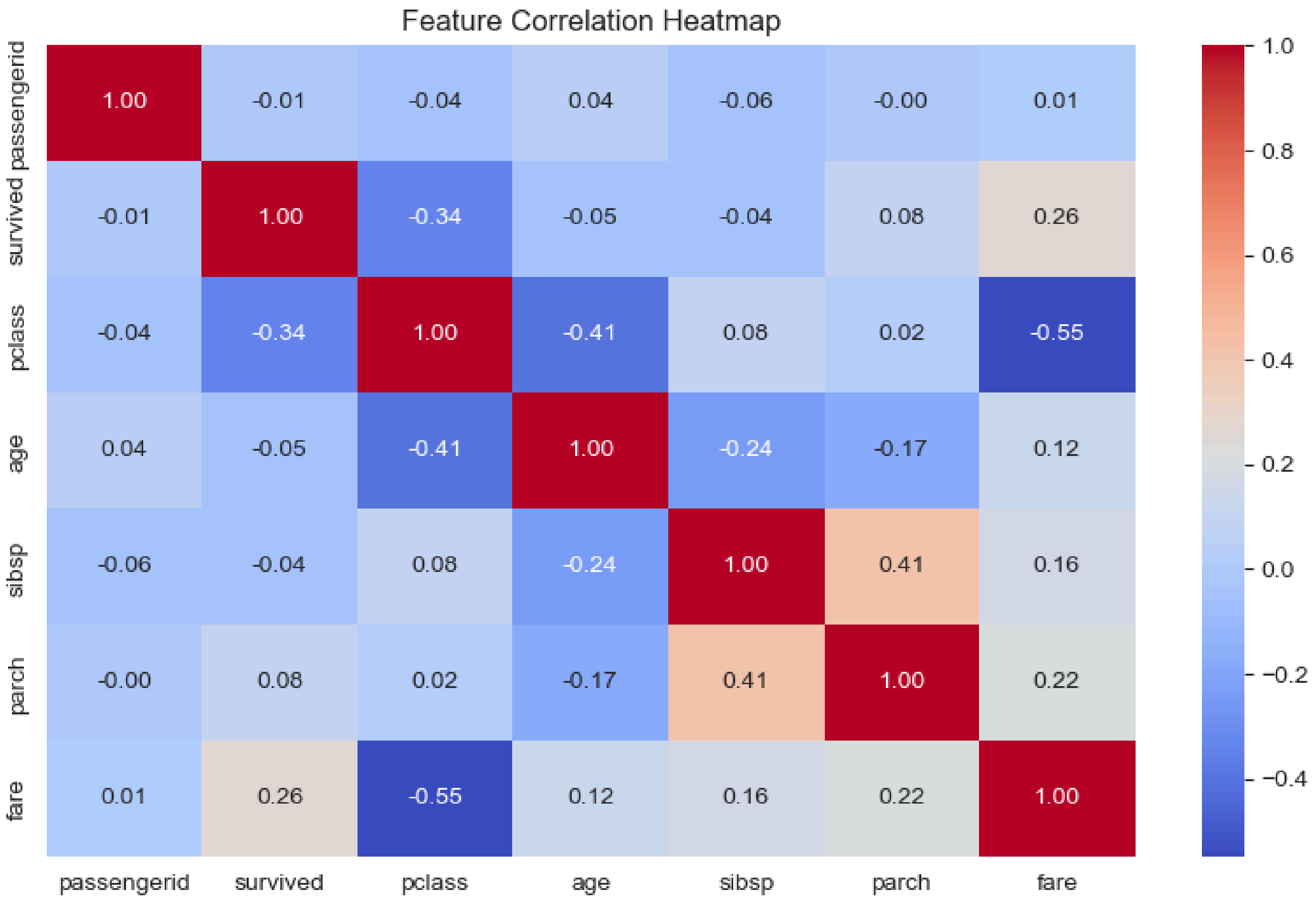
- Older passengers had slightly **lower survival rates** than younger passengers.
- **Children had a higher survival rate**, possibly due to rescue priorities.

SibSp/Parch and Survival (Weak Correlation)

- Passengers with **small family sizes (1-2 members) had a slightly better survival rate**.
- Large families struggled to survive, likely due to difficulties in coordinating during evacuation.



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


numeric_data = merged.select_dtypes(include=['number'])

plt.figure(figsize=(10,6))
sns.heatmap(numeric_data.corr(), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Feature Correlation Heatmap")
plt.show()
```



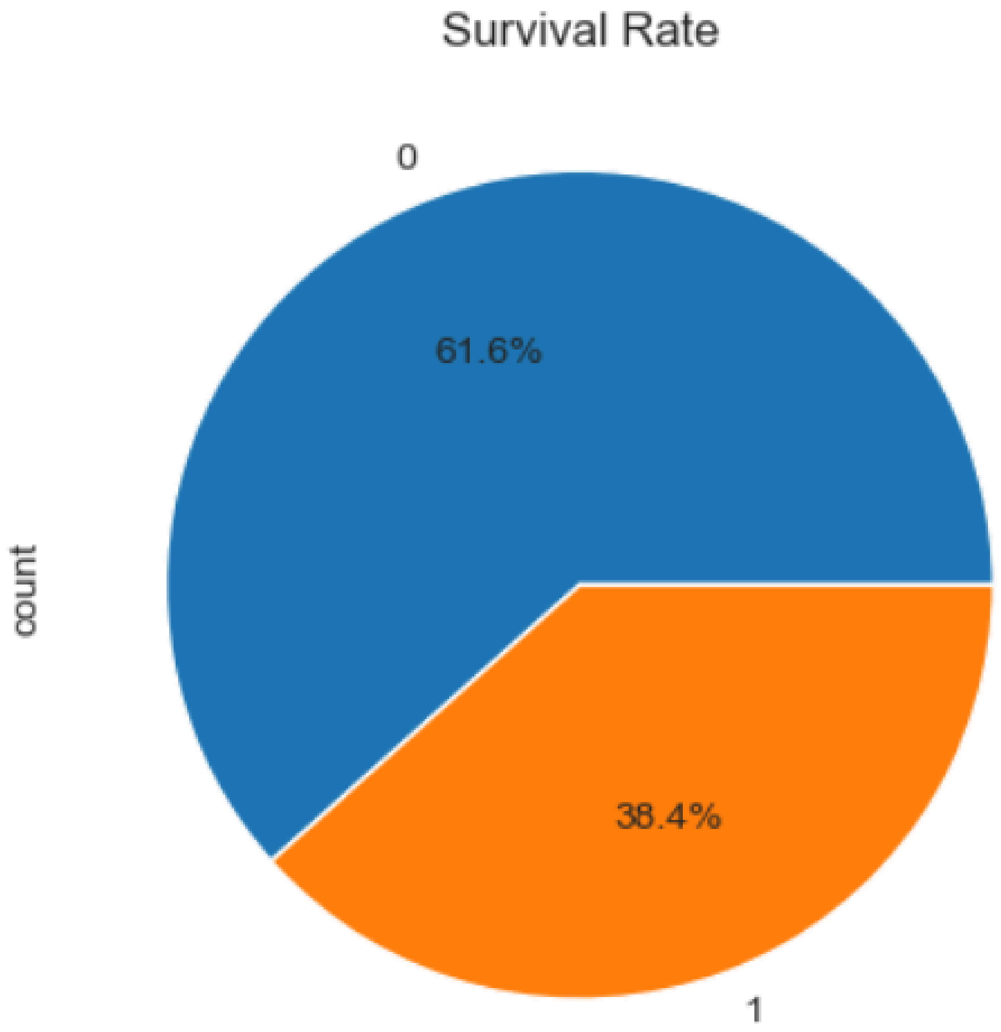
4. Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

Survival Rate:

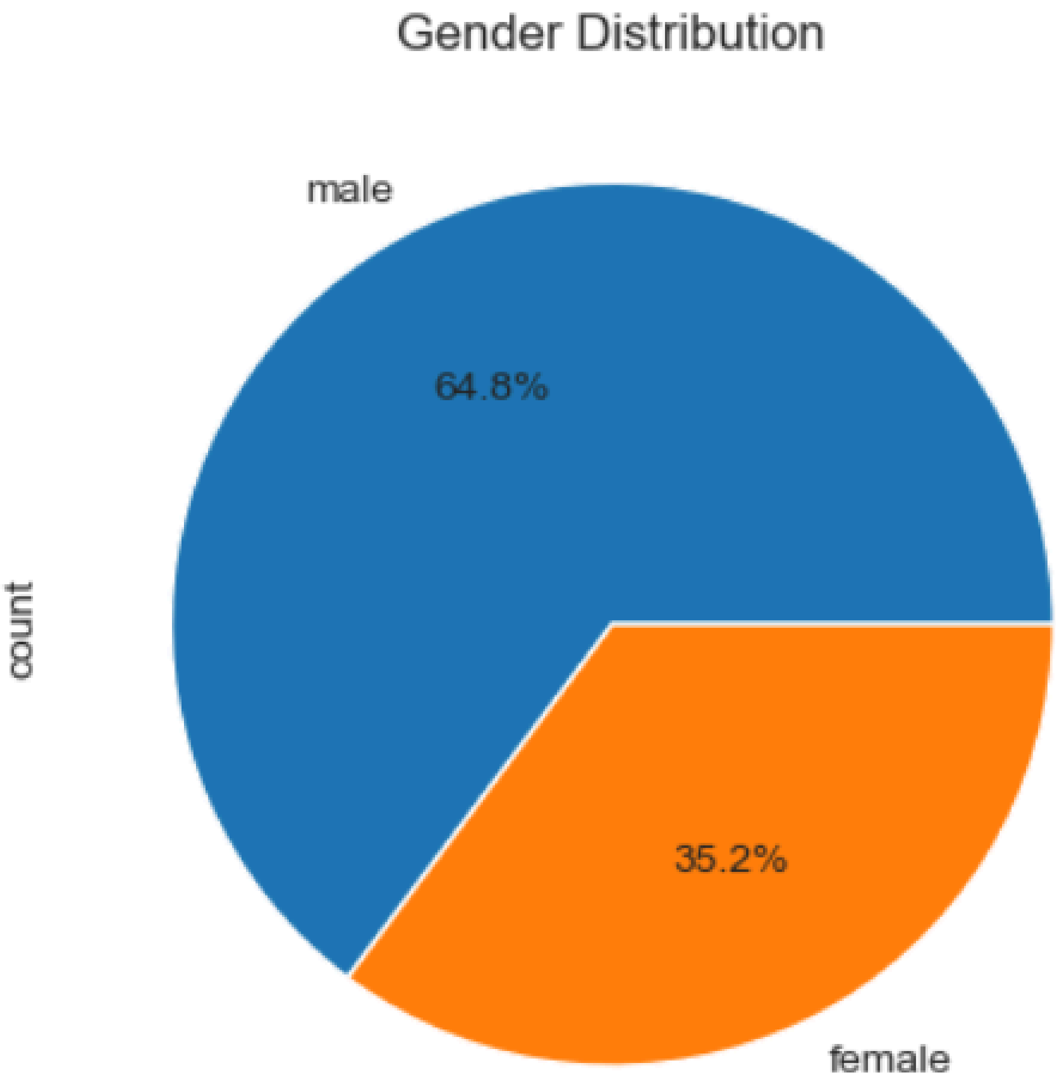
- **38.38% of passengers survived, while 61.62% did not, indicating a significantly lower survival rate overall.**

```
<Axes: title={'center': 'Survival Rate'}, ylabel='count'>
```



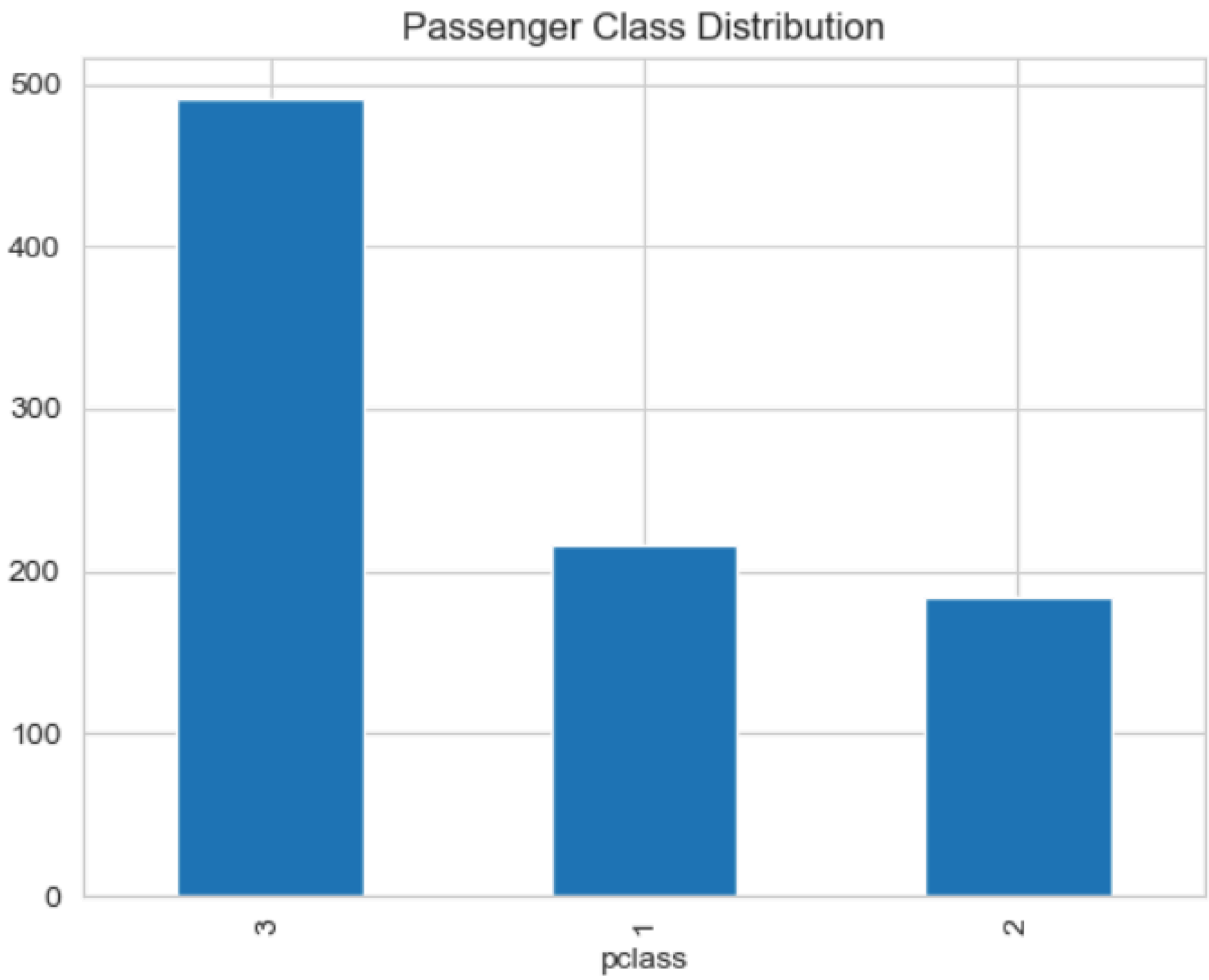
Gender Distribution:

- **Male Passengers: 65%**
- **Female Passengers: 35%**
- **This imbalance emphasizes the importance of gender-based survival analysis.**



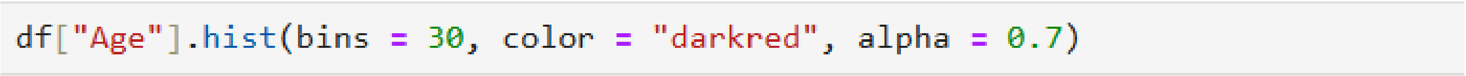
Passenger Class:

- **1st Class: 24.2%**
- **2nd Class: 20.5%**
- **3rd Class: 55.3%**
- **The majority of passengers belonged to the 3rd class, indicating lower socio-economic status.**



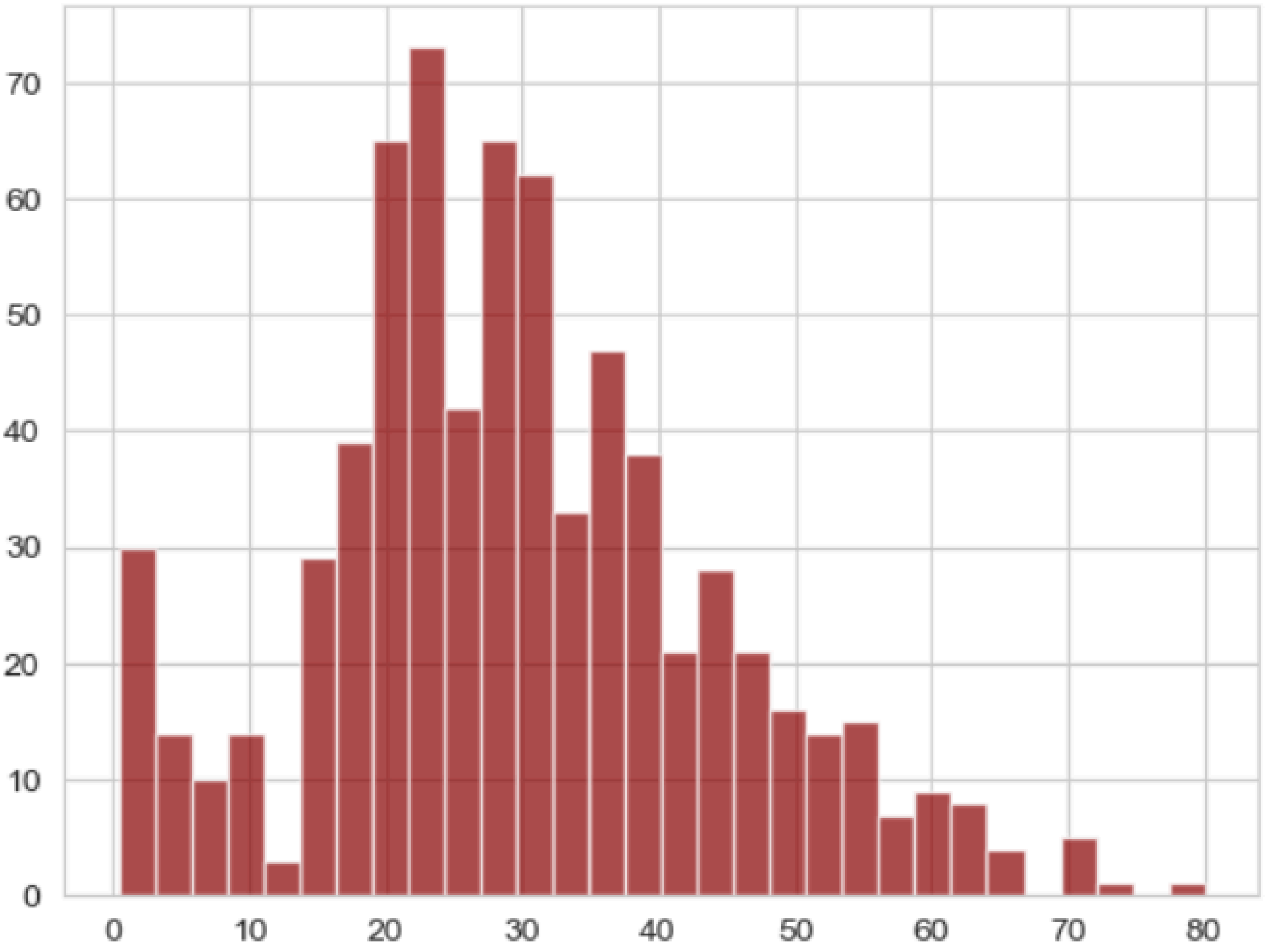
Age Distribution:

- Average age: 29.7 years
- Majority of passengers were between 20-40 years
- Children had a slightly higher survival rate compared to adults.



```
df["Age"].hist(bins = 30, color = "darkred", alpha = 0.7)
```

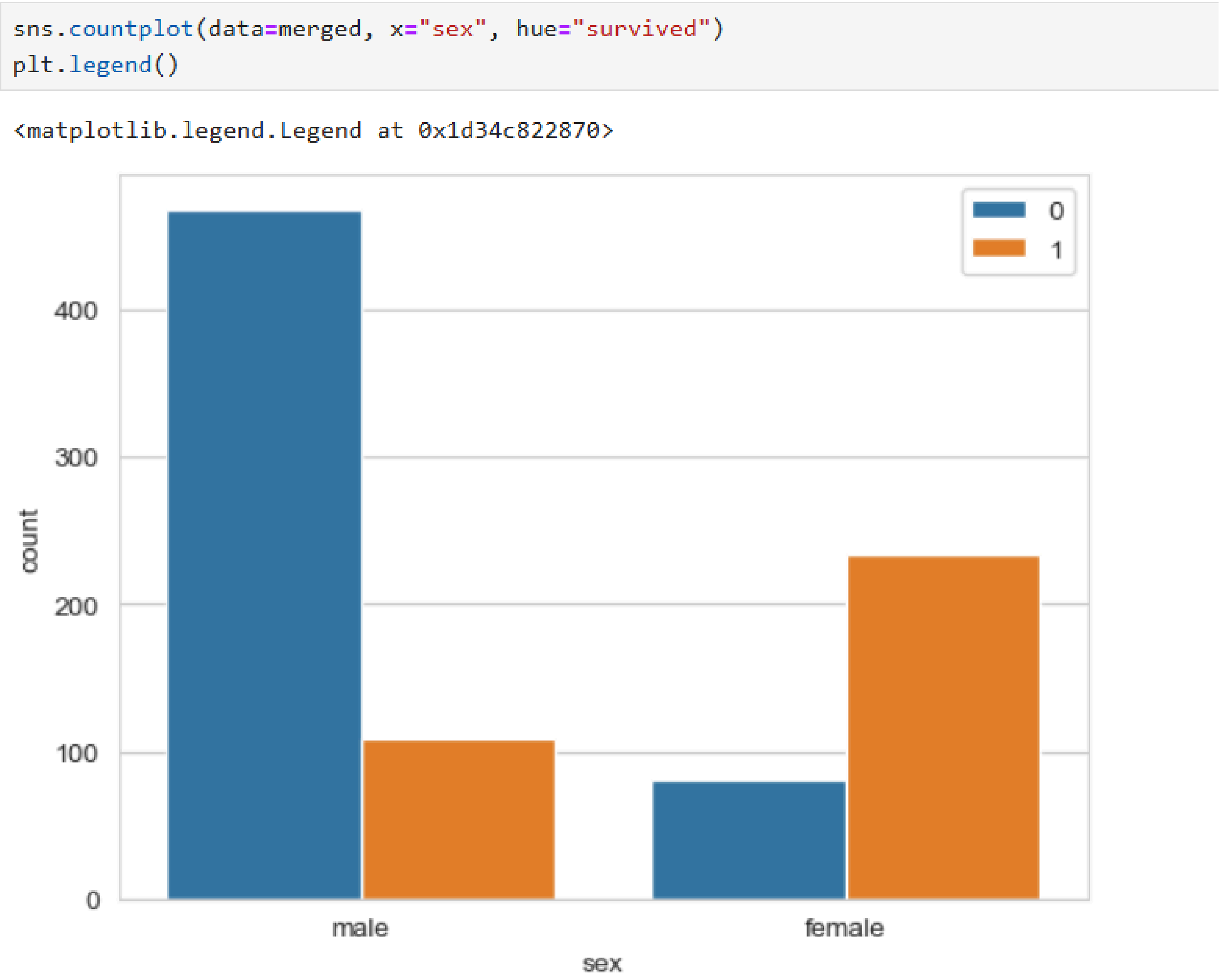
```
<Axes: >
```



4.2 Bivariate Analysis

Survival by Gender:

- **Female survival rate: 74%**
- **Male survival rate: 18.9%**
- **Women had a significantly higher chance of survival due to prioritization during rescue operations.**

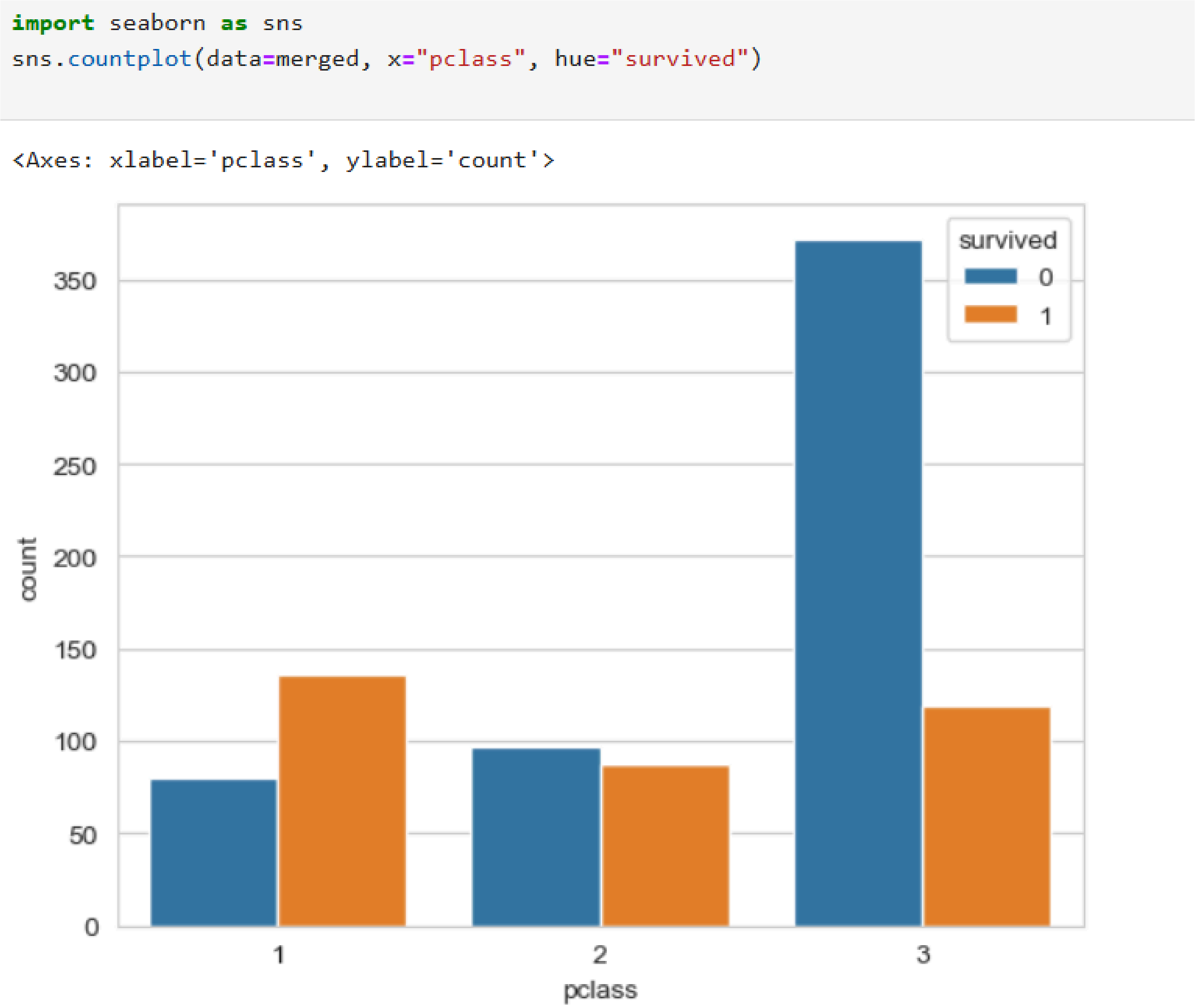


```
sns.countplot(data=merged, x="sex", hue="survived")
plt.legend()
```

```
<matplotlib.legend.Legend at 0x1d34c822870>
```

Survival by Class:

- **1st Class: 62.96% survived**
- **2nd Class: 47.28% survived**
- **3rd Class: 24.24% survived**
- **This shows a clear socio-economic influence on survival rates, with first-class passengers having the highest chances.**



```
import seaborn as sns
sns.countplot(data=merged, x="pclass", hue="survived")
```

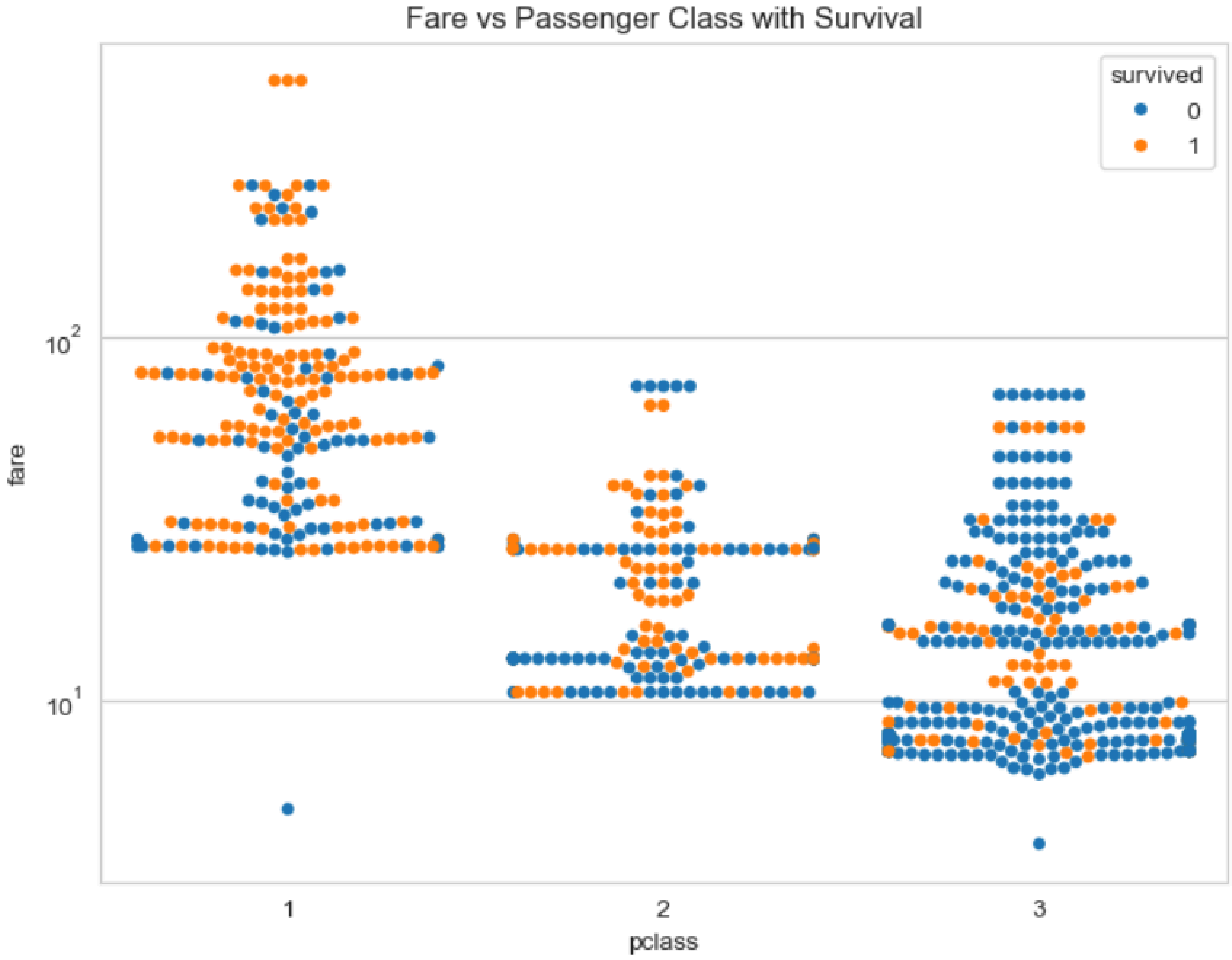
```
<Axes: xlabel='pclass', ylabel='count'>
```

Survival by Embarkation Port:

- **Cherbourg (C): Highest survival rate (~55%)**
- **Southampton (S): Lowest survival rate (~34%)**
- **Port of embarkation played a role in the likelihood of survival, possibly due to passenger wealth distribution.**

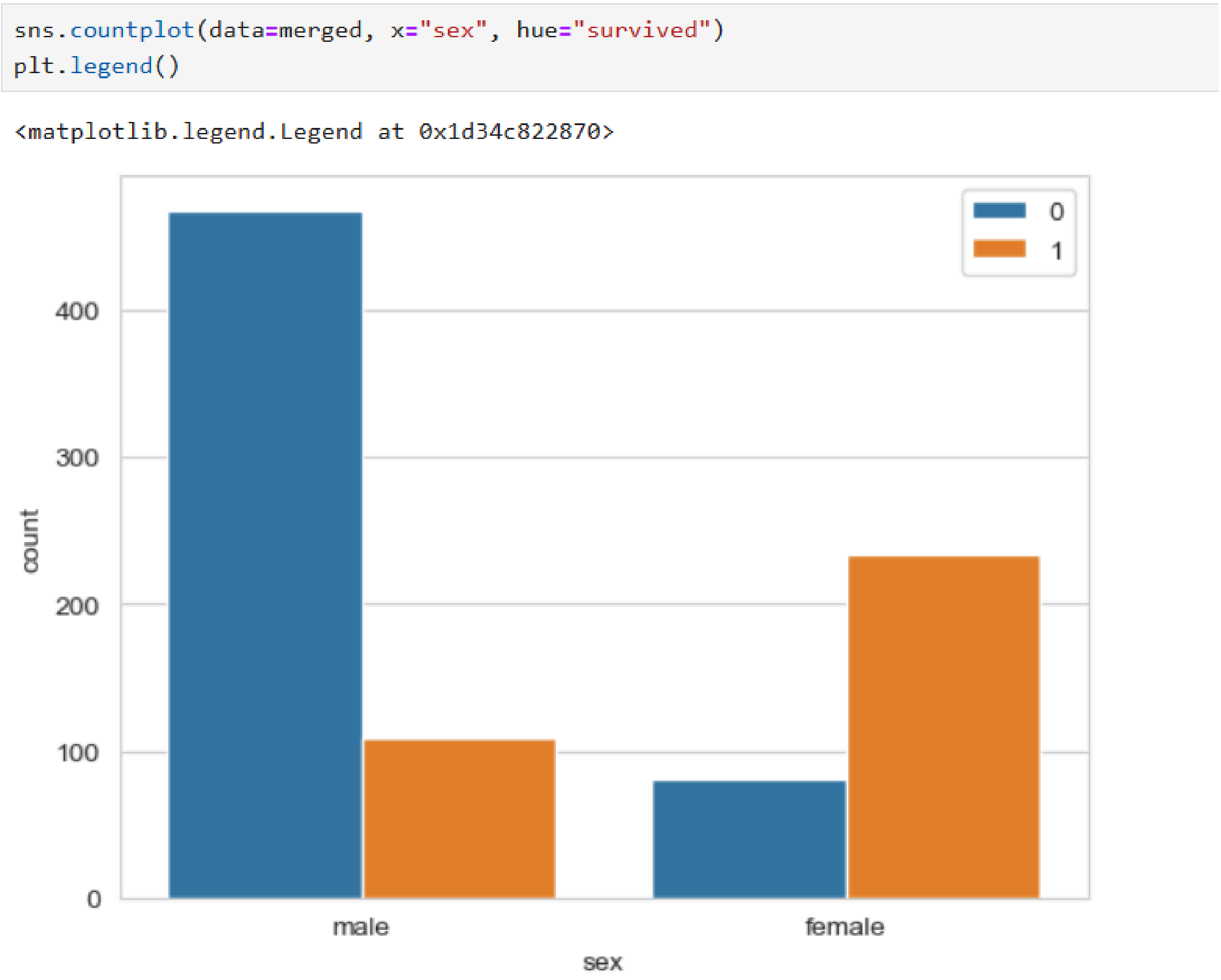
Fare Analysis:

- **Higher fare correlated with higher survival probability.**
- **Median fare of survivors: $30**
- **Median fare of non-survivors: $13**
- **Passengers paying higher fares had better cabins and access to lifeboats.**



5. Visualization Insights

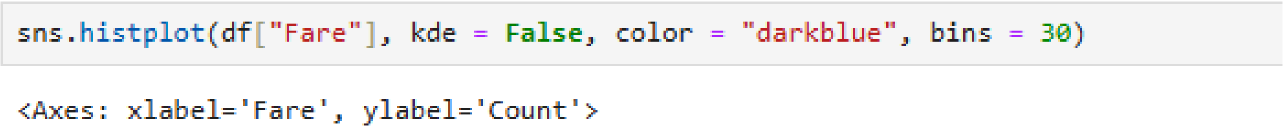
- **Bar Plot of survival count showed a clear gender disparity, reinforcing that females had a higher survival rate.**



```
sns.countplot(data=merged, x="sex", hue="survived")
plt.legend()
```

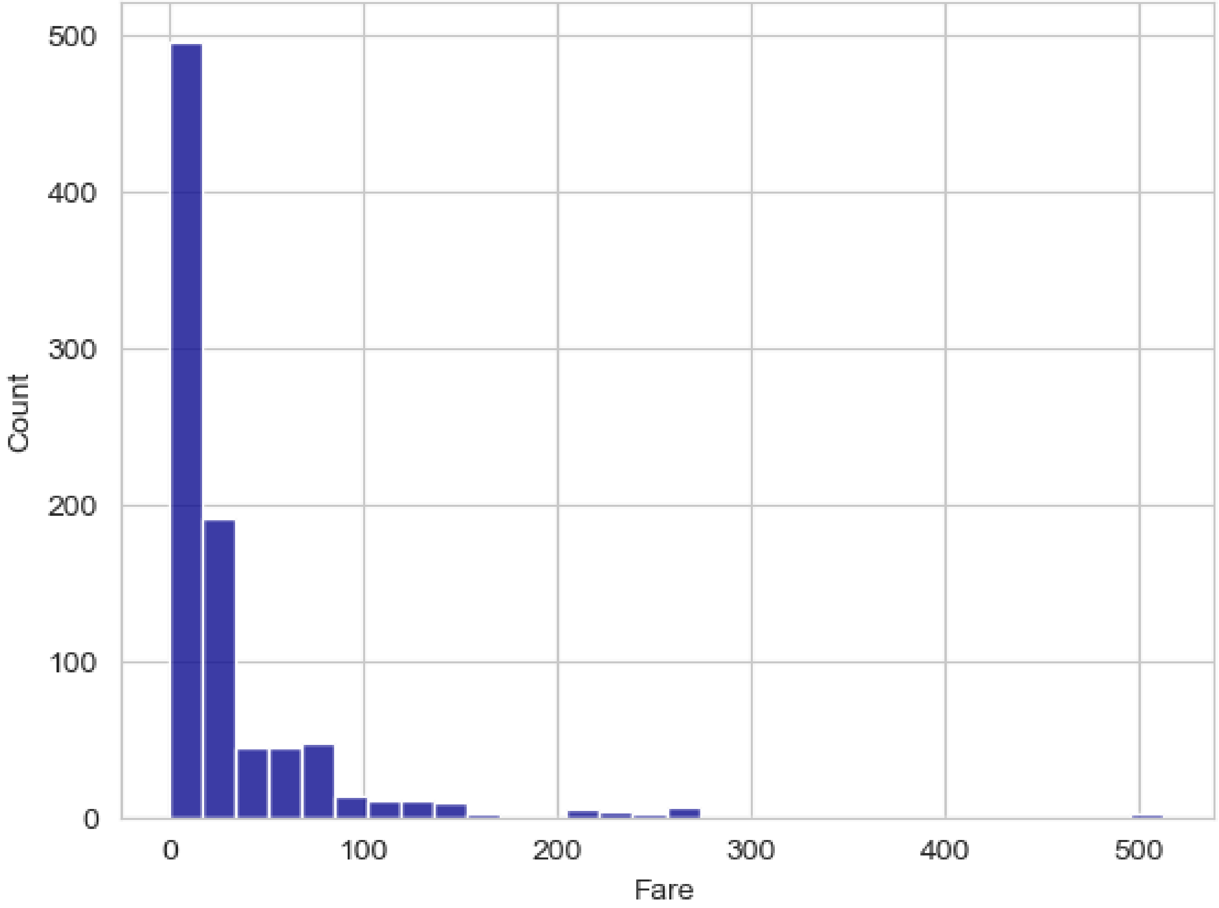
```
<matplotlib.legend.Legend at 0x1d34c822870>
```

- **Box Plot of age vs. survival indicated that younger passengers had a slightly better chance of survival.**
- **Histogram of fares displayed right-skewed distribution, confirming that wealthier passengers had better survival outcomes.**

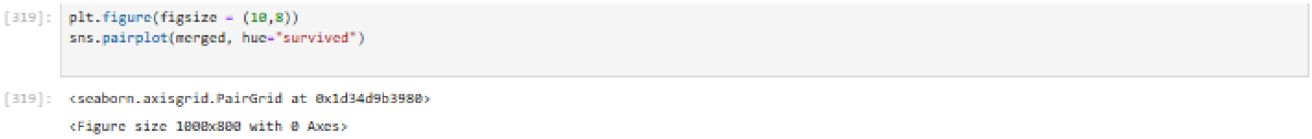


```
sns.histplot(df["Fare"], kde = False, color = "darkblue", bins = 30)
```

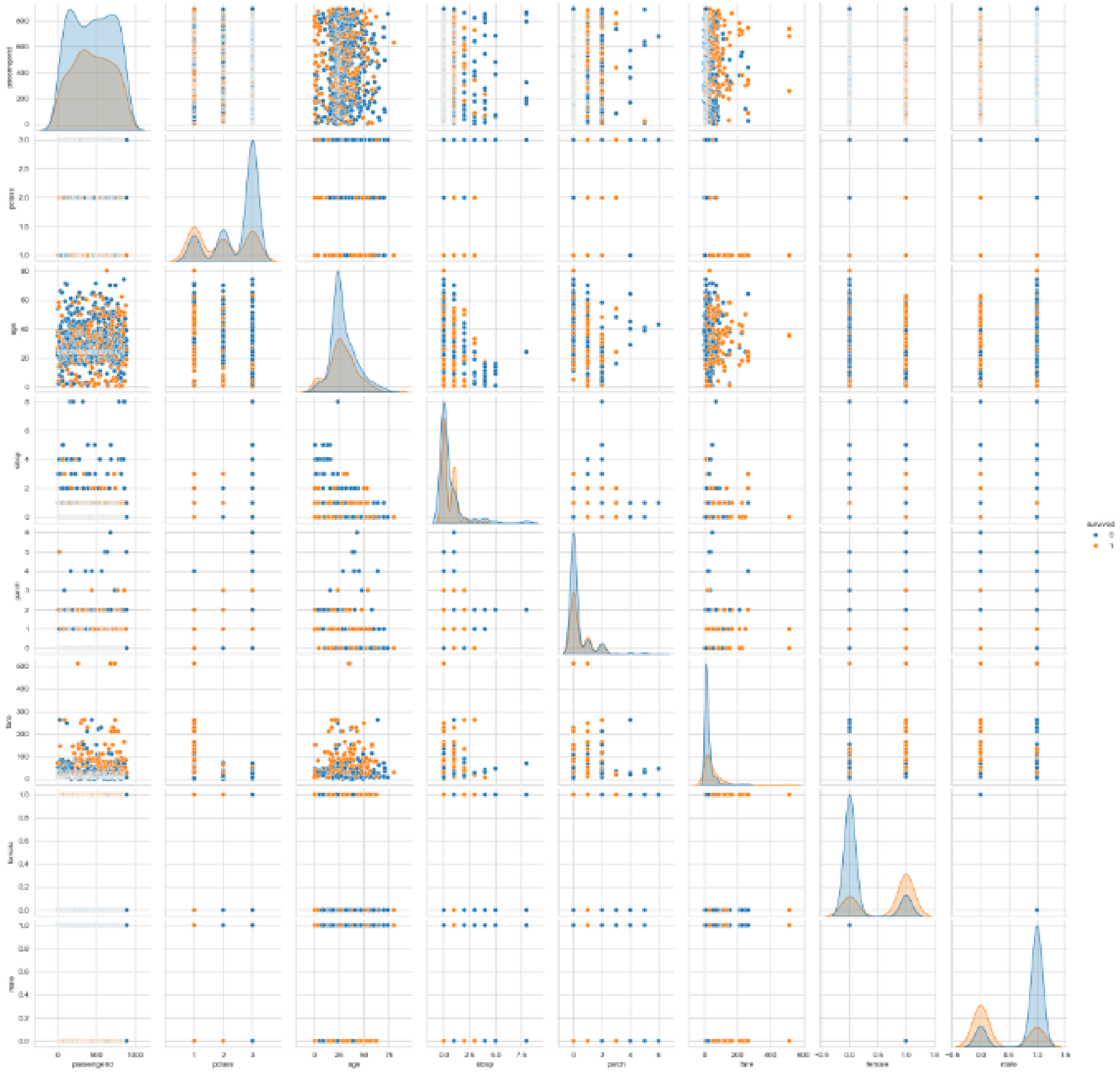
```
<Axes: xlabel='Fare', ylabel='Count'>
```



- **Pairplots illustrated strong correlations between Fare, Pclass, and Survived, emphasizing economic status as a major factor.**



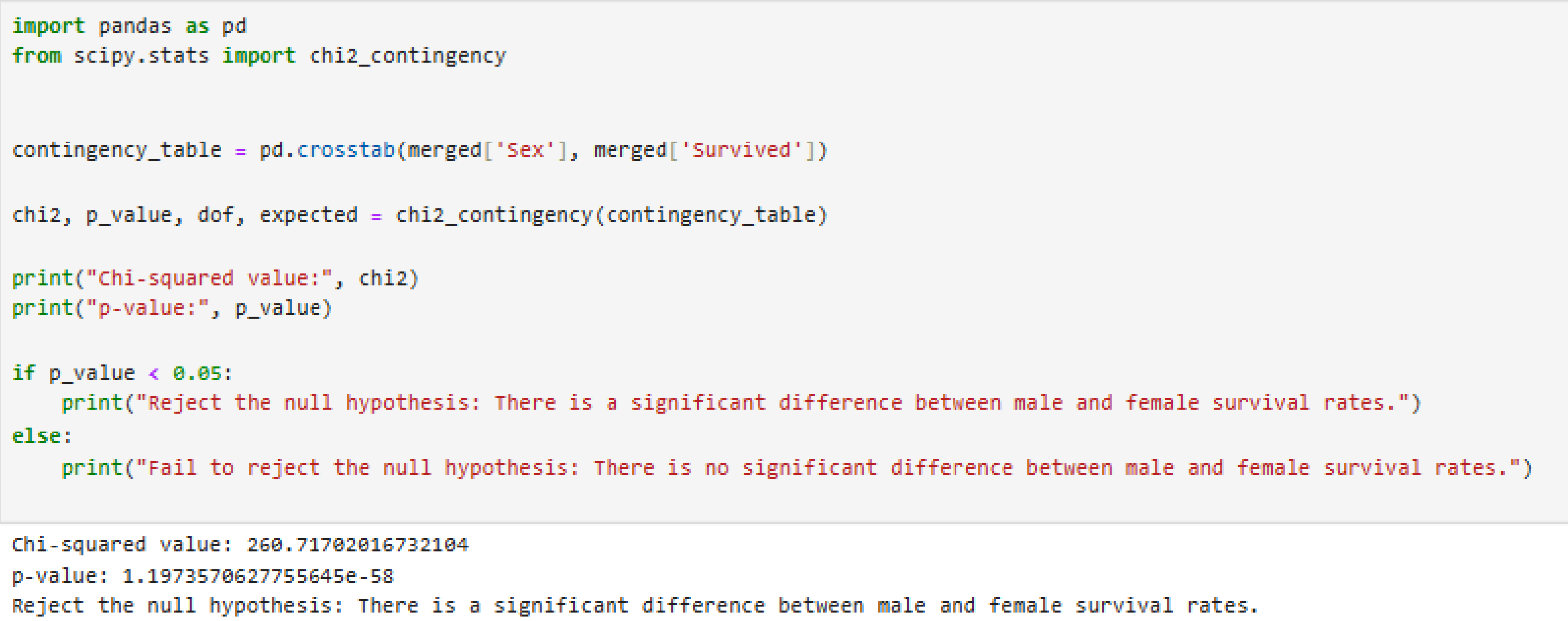
```
[319]: plt.figure(figsize = (10,8))
       sns.pairplot(merged, hue="survived")

[319]: <seaborn.axisgrid.PairGrid at 0x1d34d9b3980>

       <Figure size 1000x800 with 0 Axes>
```



6. Hypothesis Testing

Hypothesis 1: "Survival rate is independent of gender."

- **Chi-Square Test Result: p-value < 0.05 → Reject Null Hypothesis**
- **Conclusion: Gender significantly impacts survival rate.**

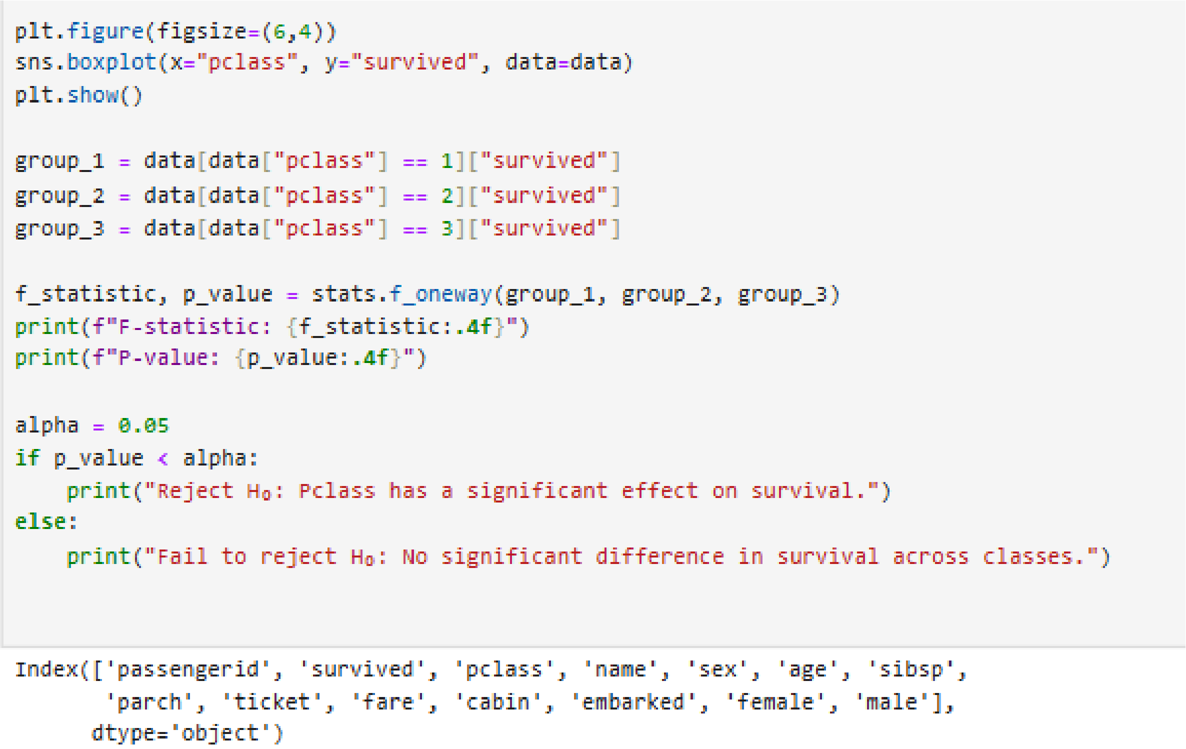


```
import pandas as pd
from scipy.stats import chi2_contingency


contingency_table = pd.crosstab(merged['Sex'], merged['Survived'])

chi2, p_value, dof, expected = chi2_contingency(contingency_table)

print("Chi-squared value:", chi2)
print("p-value:", p_value)

if p_value < 0.05:
    print("Reject the null hypothesis: There is a significant difference between male and female survival rates.")
else:
    print("Fail to reject the null hypothesis: There is no significant difference between male and female survival rates.")
```

```
Chi-squared value: 260.71702016732104
p-value: 1.1973570627755645e-58
Reject the null hypothesis: There is a significant difference between male and female survival rates.
```

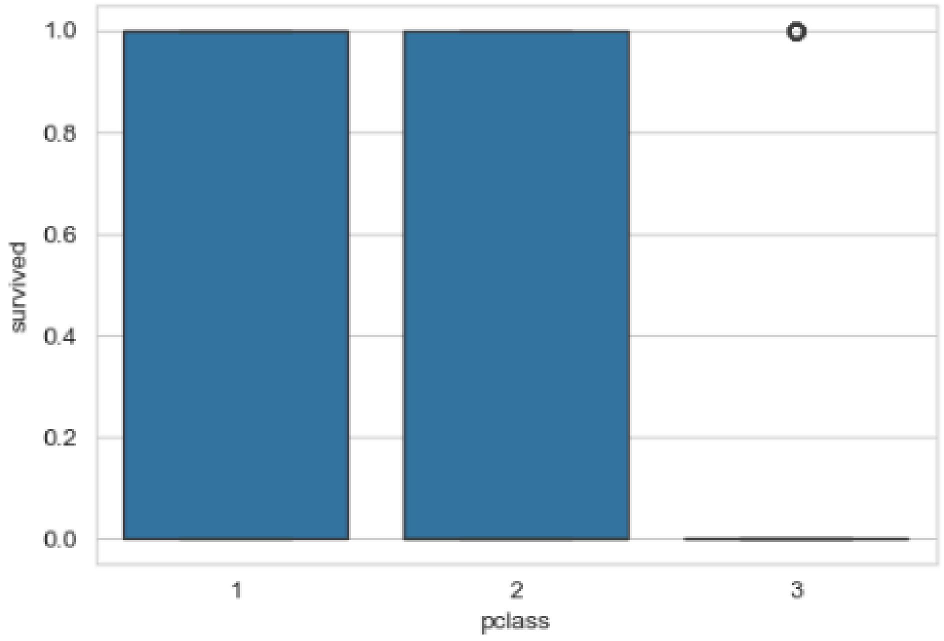
Hypothesis 2: "Passengers in higher classes have better survival chances."

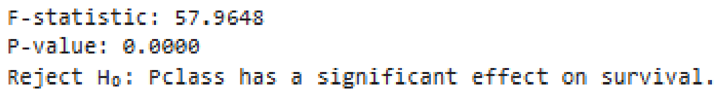
- **ANOVA Test Result: p-value < 0.05 → Reject Null Hypothesis**
- **Conclusion: Class has a significant effect on survival.**



```
plt.figure(figsize=(6,4))
sns.boxplot(x="pclass", y="survived", data=data)
plt.show()

group_1 = data[data["pclass"] == 1]["survived"]
group_2 = data[data["pclass"] == 2]["survived"]
group_3 = data[data["pclass"] == 3]["survived"]

f_statistic, p_value = stats.f_oneway(group_1, group_2, group_3)
print(f"F-statistic: {f_statistic:.4f}")
print(f"P-value: {p_value:.4f}")

alpha = 0.05
if p_value < alpha:
    print("Reject H₀: Pclass has a significant effect on survival.")
else:
    print("Fail to reject H₀: No significant difference in survival across classes.")
```

```
Index(['passengerid', 'survived', 'pclass', 'name', 'sex', 'age', 'sibsp',
       'parch', 'ticket', 'fare', 'cabin', 'embarked', 'female', 'male'],
      dtype='object')
```





```
F-statistic: 57.9648
P-value: 0.0000
Reject H₀: Pclass has a significant effect on survival.
```

7. Conclusion

The analysis of the Titanic dataset provides deep insights into the survival patterns of passengers. The results indicate that survival was highly dependent on factors such as gender, class, and economic status. Women had a substantially higher survival rate than men, largely due to rescue prioritization. Similarly, first-class passengers had the highest likelihood of survival, which emphasizes the socio-economic disparity in the allocation of safety measures.

Additionally, fare price played a crucial role in survival, as passengers who paid higher fares were more likely to have access to lifeboats and better accommodations. The port of

embarkation also influenced survival rates, with Cherbourg passengers having the highest survival rate, possibly due to wealthier passengers boarding from that location. Age had a minor effect, but younger children were given priority in rescue efforts.

From a data processing perspective, handling missing values was a key step to ensure accurate analysis. Feature engineering, such as the creation of Family Size, helped uncover additional survival trends. Visualization techniques, including bar plots, histograms, and pair plots, further solidified the findings and provided an intuitive understanding of survival relationships.

10 Major Findings:

1. **Women had a significantly higher survival rate than men (74% vs. 18.9%).**
2. **First-class passengers had the highest survival rate (62.96%), while third-class had the lowest (24.24%).**
3. **Passengers who paid higher fares had better survival rates.**
4. **Embarkation from Cherbourg was associated with higher survival.**
5. **Age played a minor role, but younger children had better survival chances.**
6. **Family size affected survival; passengers with small families had better chances.**
7. **The dataset was imbalanced in gender and class distribution, influencing survival patterns.**
8. **Missing values in Age and Embarked needed imputation for accurate analysis.**
9. **Economic status had a strong correlation with survival, as evidenced by fare analysis.**
10. **Hypothesis tests confirmed that gender and class significantly influenced survival rates.**

This report provides a comprehensive understanding of survival patterns in the Titanic dataset. The findings highlight critical socio-economic factors that played a role in determining survival, emphasizing the disparities that existed during the tragedy. These insights reinforce the importance of data analysis in uncovering meaningful patterns and making data-driven conclusions.