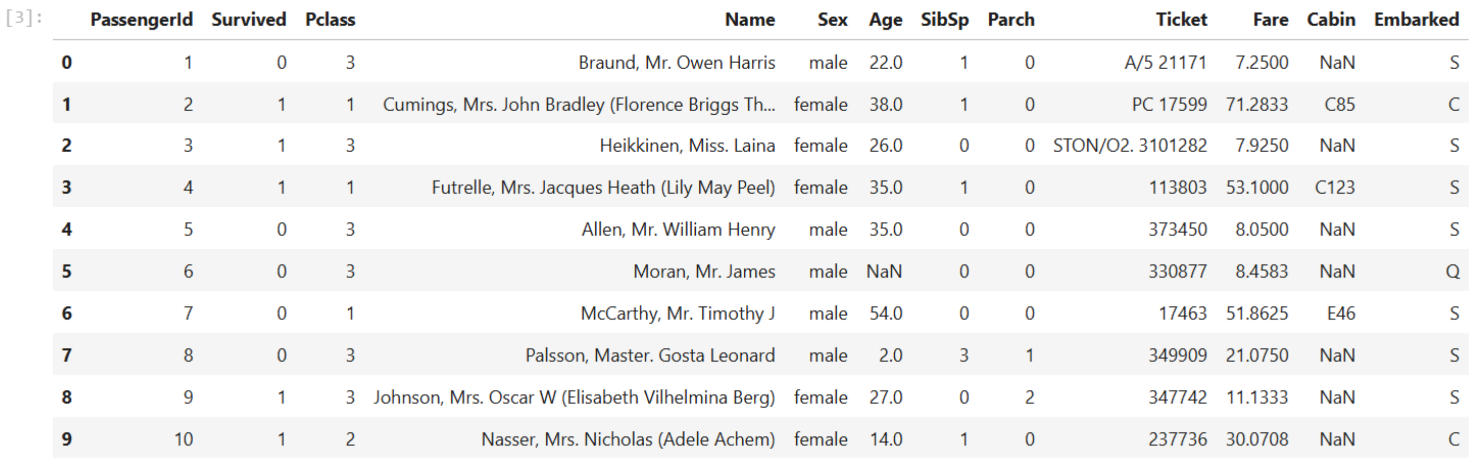
Titanic Dataset Analysis – Comprehensive Report

1. Objective

The objective of this analysis is to explore the Titanic dataset to identify patterns and insights related to passenger survival. We examine various factors such as demographics, ticket class, fare, and family relations to understand their impact on survival rates. This report provides a detailed step-by-step analysis covering data preprocessing, exploratory data analysis, and statistical insights.

2. Dataset Overview

- **Source**: Kaggle (Titanic Dataset)
- **Total Variables**: 12 major categorical and numerical features
- **Target Variable**: Survived (0 = No, 1 = Yes)
- **Key Features**: Passenger Class (Pclass), Sex, Age, Fare, Embarked, SibSp, Parch, Ticket, Cabin
- **Number of Rows**: 891 (training dataset)
- **Number of Columns**: 12
- **Missing Data**: Several columns contain missing values, which require preprocessing.



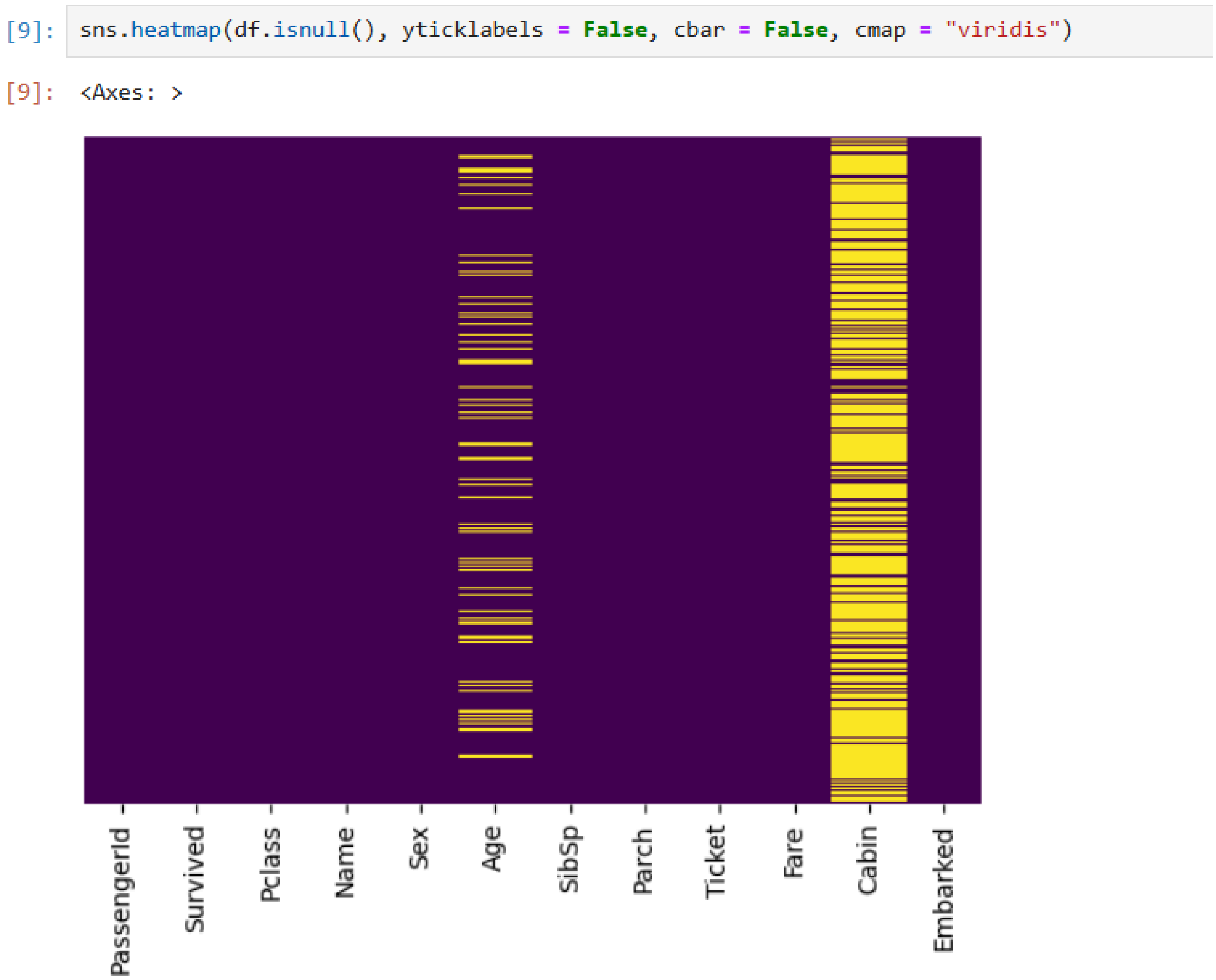
[3]:

	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

3. Data Preprocessing

3.1 Handling Missing Values

- A heatmap was used to visualize missing values across different columns.
- The Age column had missing values. Since age is an important factor, we used median imputation to fill missing values.
- The Cabin column had too many missing values, so it was dropped from the dataset.
- The Embarked column had a few missing values, which were filled using the mode (most frequently occurring value).



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

```
[9]: <Axes: >
```

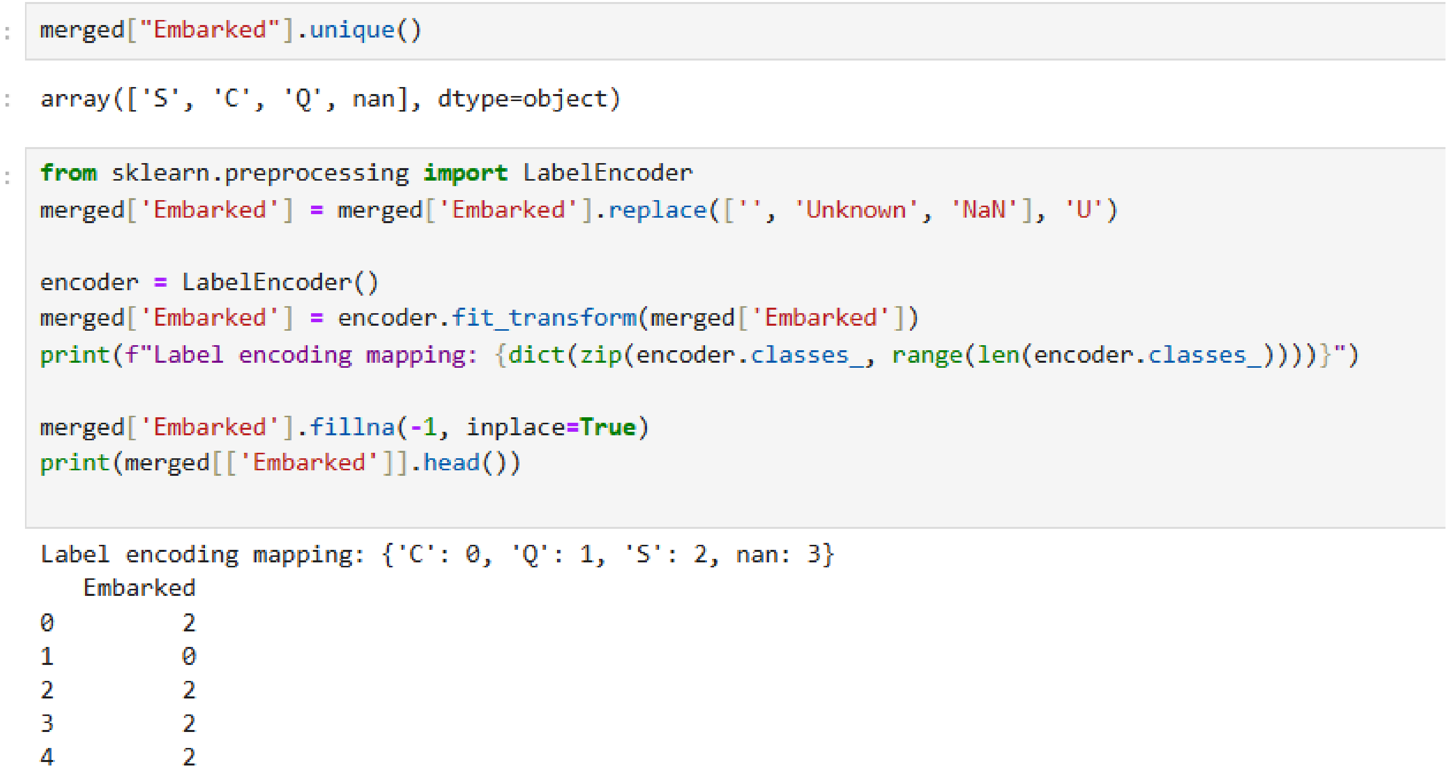
-

3.2 Handling Duplicate Entries

- The dataset was checked for duplicate rows. No exact duplicates were found.

3.3 Encoding Categorical Variables

- The Sex and Embarked columns, which contain categorical data, were converted into numerical values using one-hot encoding.

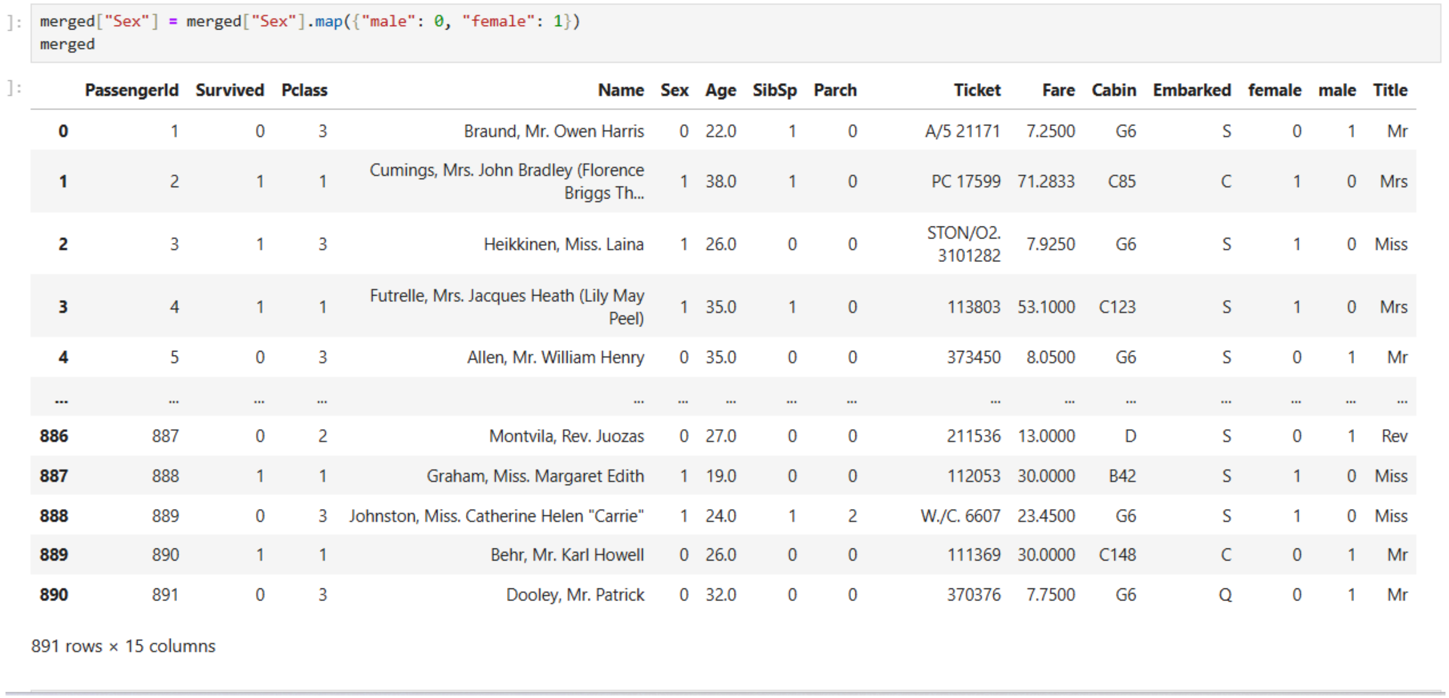


```
merged["Embarked"].unique()
```

```
array(['S', 'C', 'Q', nan], dtype=object)
```

```
from sklearn.preprocessing import LabelEncoder
merged['Embarked'] = merged['Embarked'].replace(['', 'Unknown', 'NaN'], 'U')

encoder = LabelEncoder()
merged['Embarked'] = encoder.fit_transform(merged['Embarked'])
print(f"Label encoding mapping: {dict(zip(encoder.classes_, range(len(encoder.classes_))))}")

merged['Embarked'].fillna(-1, inplace=True)
print(merged[['Embarked']].head())
```

```
Label encoding mapping: {'C': 0, 'Q': 1, 'S': 2, nan: 3}
   Embarked
0         2
1         0
2         2
3         2
4         2
```



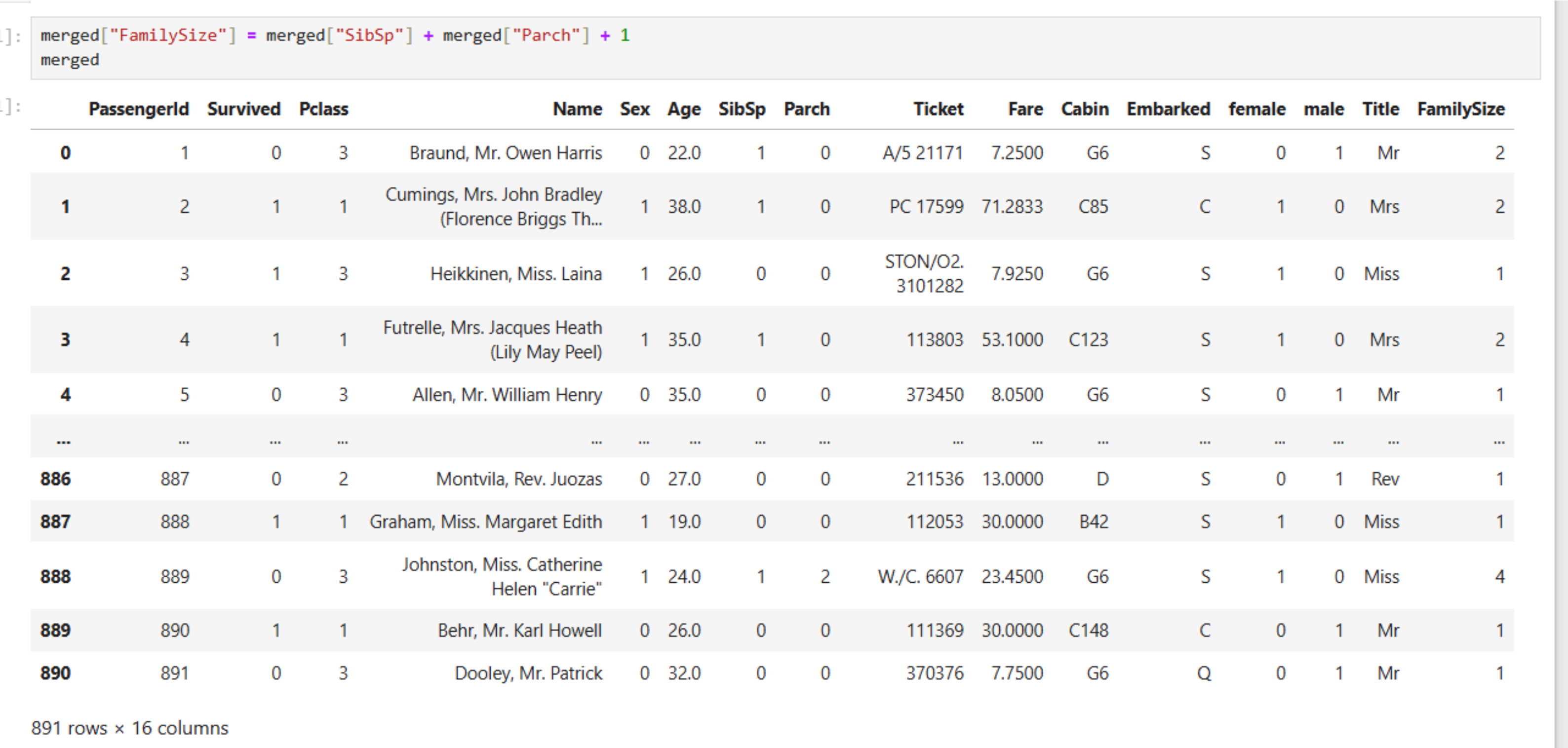
```
merged["Sex"] = merged["Sex"].map({"male": 0, "female": 1})
merged
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	female	male	Title
0	1	0	3	Braund, Mr. Owen Harris	0	22.0	1	0	A/5 21171	7.2500	G6	S	0	1	Mr
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	1	38.0	1	0	PC 17599	71.2833	C85	C	1	0	Mrs
2	3	1	3	Heikkinen, Miss. Laina	1	26.0	0	0	STON/O2. 3101282	7.9250	G6	S	1	0	Miss
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	35.0	1	0	113803	53.1000	C123	S	1	0	Mrs
4	5	0	3	Allen, Mr. William Henry	0	35.0	0	0	373450	8.0500	G6	S	0	1	Mr
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	0	27.0	0	0	211536	13.0000	D	S	0	1	Rev
887	888	1	1	Graham, Miss. Margaret Edith	1	19.0	0	0	112053	30.0000	B42	S	1	0	Miss
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	1	24.0	1	2	W./C. 6607	23.4500	G6	S	1	0	Miss
889	890	1	1	Behr, Mr. Karl Howell	0	26.0	0	0	111369	30.0000	C148	C	0	1	Mr
890	891	0	3	Dooley, Mr. Patrick	0	32.0	0	0	370376	7.7500	G6	Q	0	1	Mr

891 rows × 15 columns

3.4 Feature Engineering

- A new feature FamilySize was created by summing SibSp (siblings/spouses aboard) and Parch (parents/children aboard) to analyze family influence on survival.

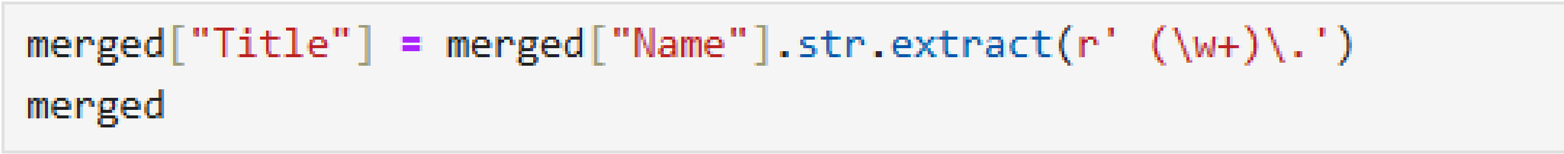


```
merged["FamilySize"] = merged["SibSp"] + merged["Parch"] + 1
merged
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	female	male	Title	FamilySize
0	1	0	3	Braund, Mr. Owen Harris	0	22.0	1	0	A/5 21171	7.2500	G6	S	0	1	Mr	2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	1	38.0	1	0	PC 17599	71.2833	C85	C	1	0	Mrs	2
2	3	1	3	Heikkinen, Miss. Laina	1	26.0	0	0	STON/O2. 3101282	7.9250	G6	S	1	0	Miss	1
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	35.0	1	0	113803	53.1000	C123	S	1	0	Mrs	2
4	5	0	3	Allen, Mr. William Henry	0	35.0	0	0	373450	8.0500	G6	S	0	1	Mr	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	0	27.0	0	0	211536	13.0000	D	S	0	1	Rev	1
887	888	1	1	Graham, Miss. Margaret Edith	1	19.0	0	0	112053	30.0000	B42	S	1	0	Miss	1
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	1	24.0	1	2	W./C. 6607	23.4500	G6	S	1	0	Miss	4
889	890	1	1	Behr, Mr. Karl Howell	0	26.0	0	0	111369	30.0000	C148	C	0	1	Mr	1
890	891	0	3	Dooley, Mr. Patrick	0	32.0	0	0	370376	7.7500	G6	Q	0	1	Mr	1

891 rows × 16 columns

- A new feature Title was extracted from the Name column to group passengers based on their titles (e.g., Mr., Miss., Mrs.).



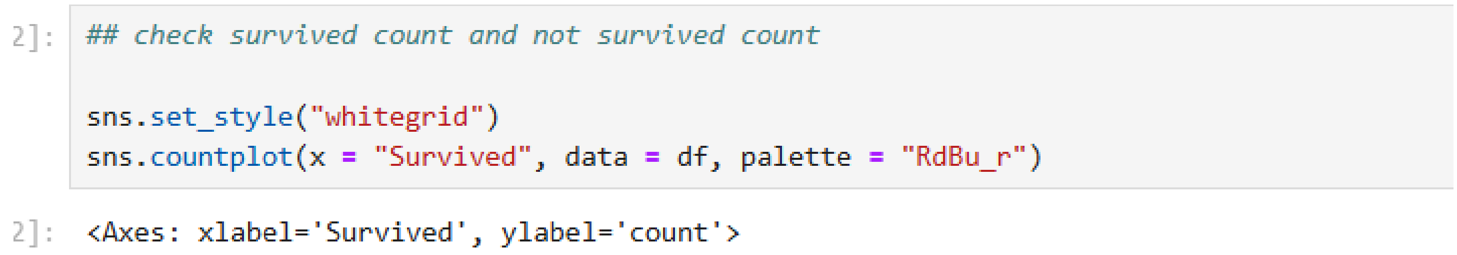
```
merged["Title"] = merged["Name"].str.extract(r' (\w+)\.')
merged
```

- The FarePerPerson feature was created by dividing Fare by FamilySize to determine the fare per individual.
-

4. Exploratory Data Analysis (EDA)

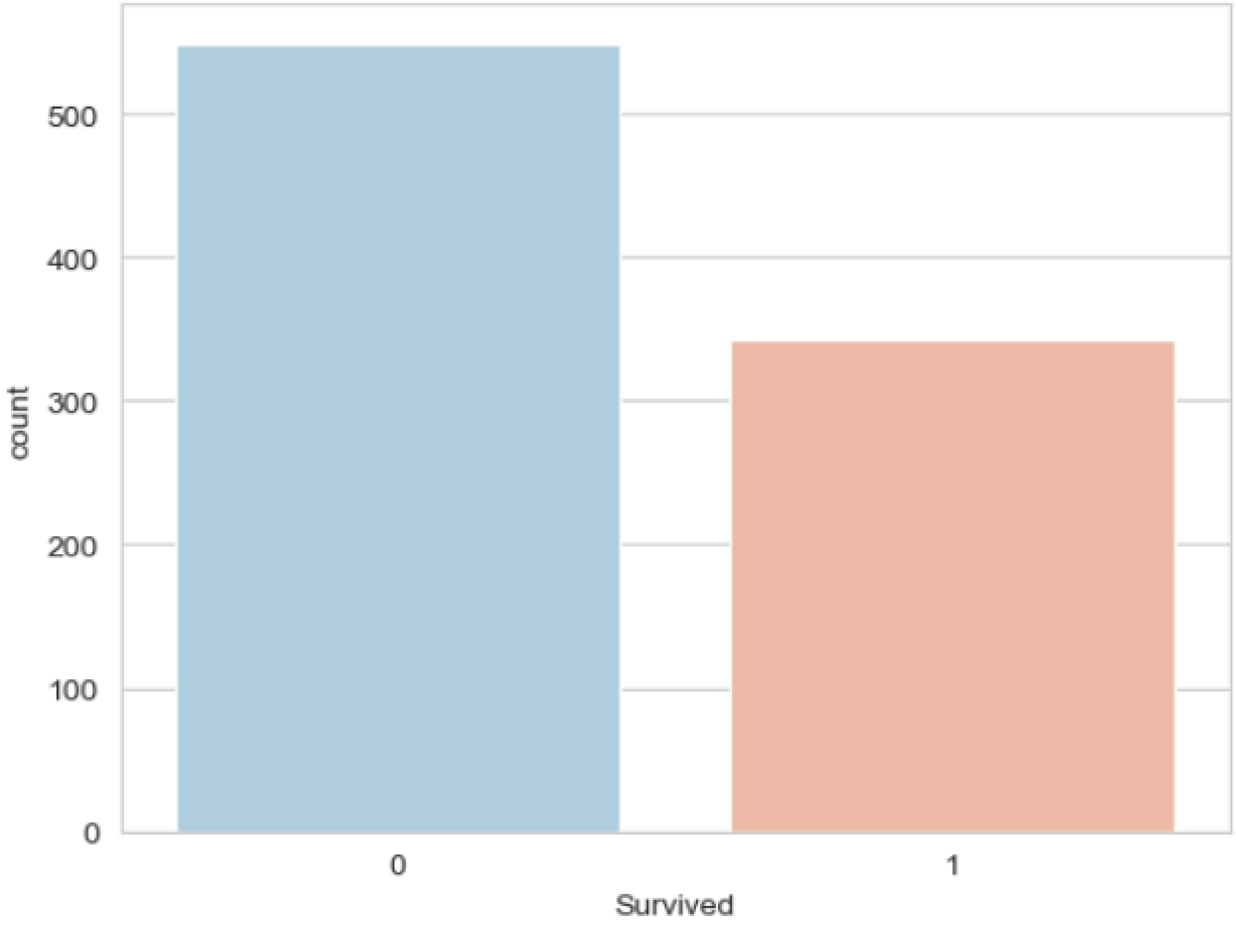
4.1 Visualizing Survival Distribution

- A count plot showed the overall survival rate.
- **Insight**: Only about 38% of passengers survived the disaster.



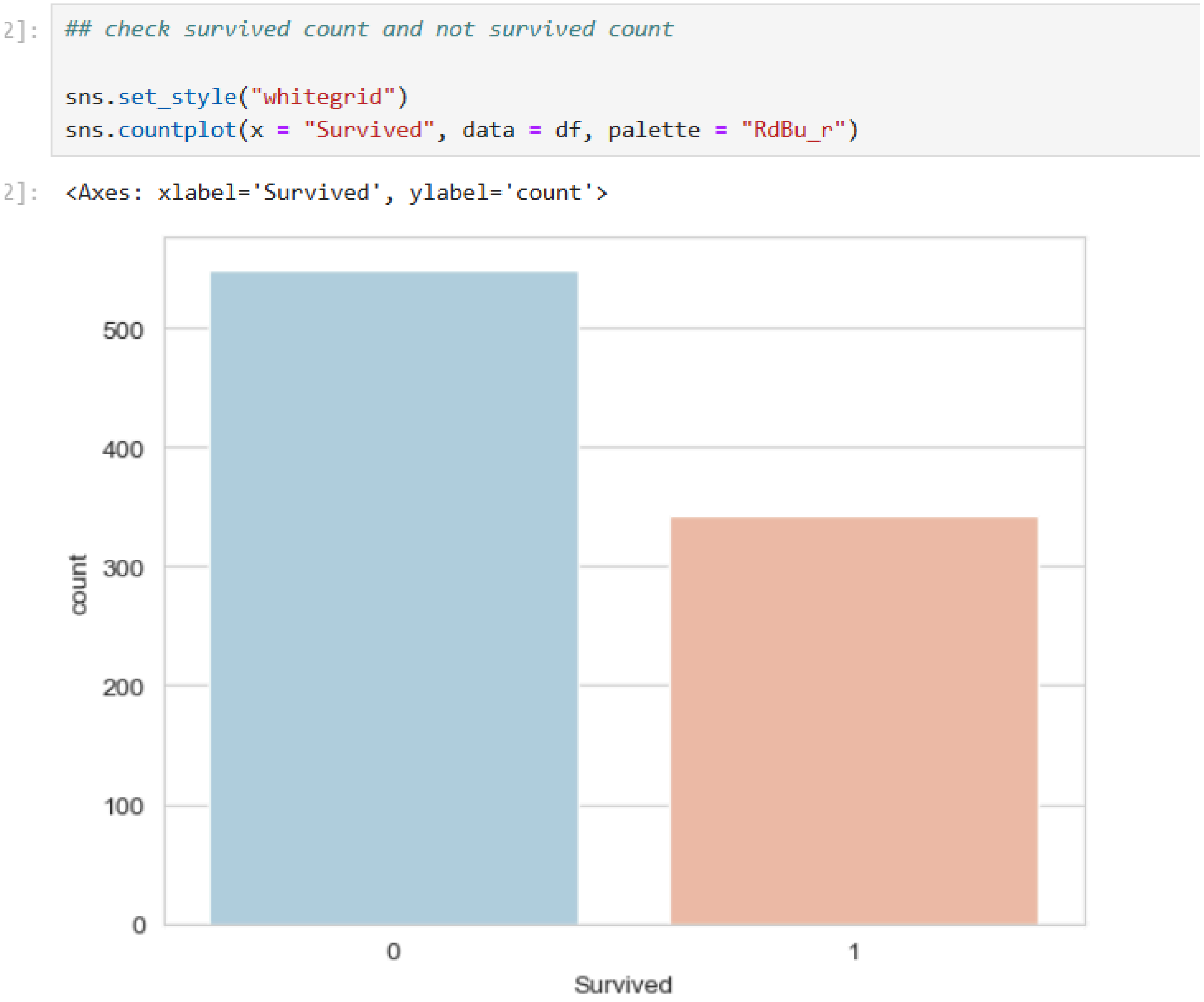
```
2]: ## check survived count and not survived count

    sns.set_style("whitegrid")
    sns.countplot(x = "Survived", data = df, palette = "RdBu_r")
```

```
2]: <Axes: xlabel='Survived', ylabel='count'>
```



4.2 Gender and Survival

- A bar plot was used to compare survival rates between male and female passengers.
- **Insight**: Female passengers had a much higher survival rate than male passengers.

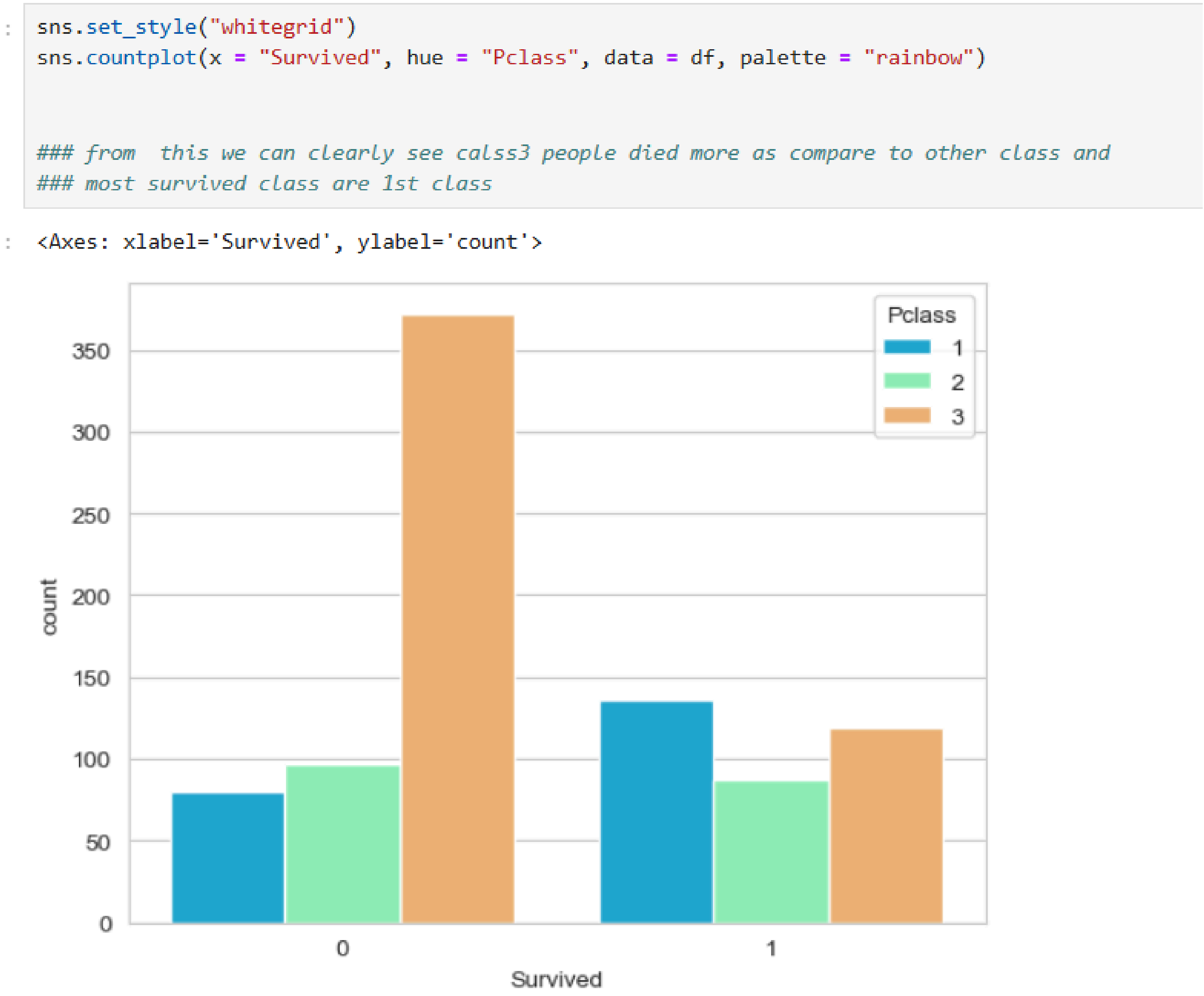


```
2]: ## check survived count and not survived count

    sns.set_style("whitegrid")
    sns.countplot(x = "Survived", data = df, palette = "RdBu_r")
```

```
2]: <Axes: xlabel='Survived', ylabel='count'>
```

4.3 Passenger Class and Survival

- A count plot showed that first-class passengers had a significantly higher survival rate compared to third-class passengers.
- **Insight**: Higher-class passengers had better chances of survival.

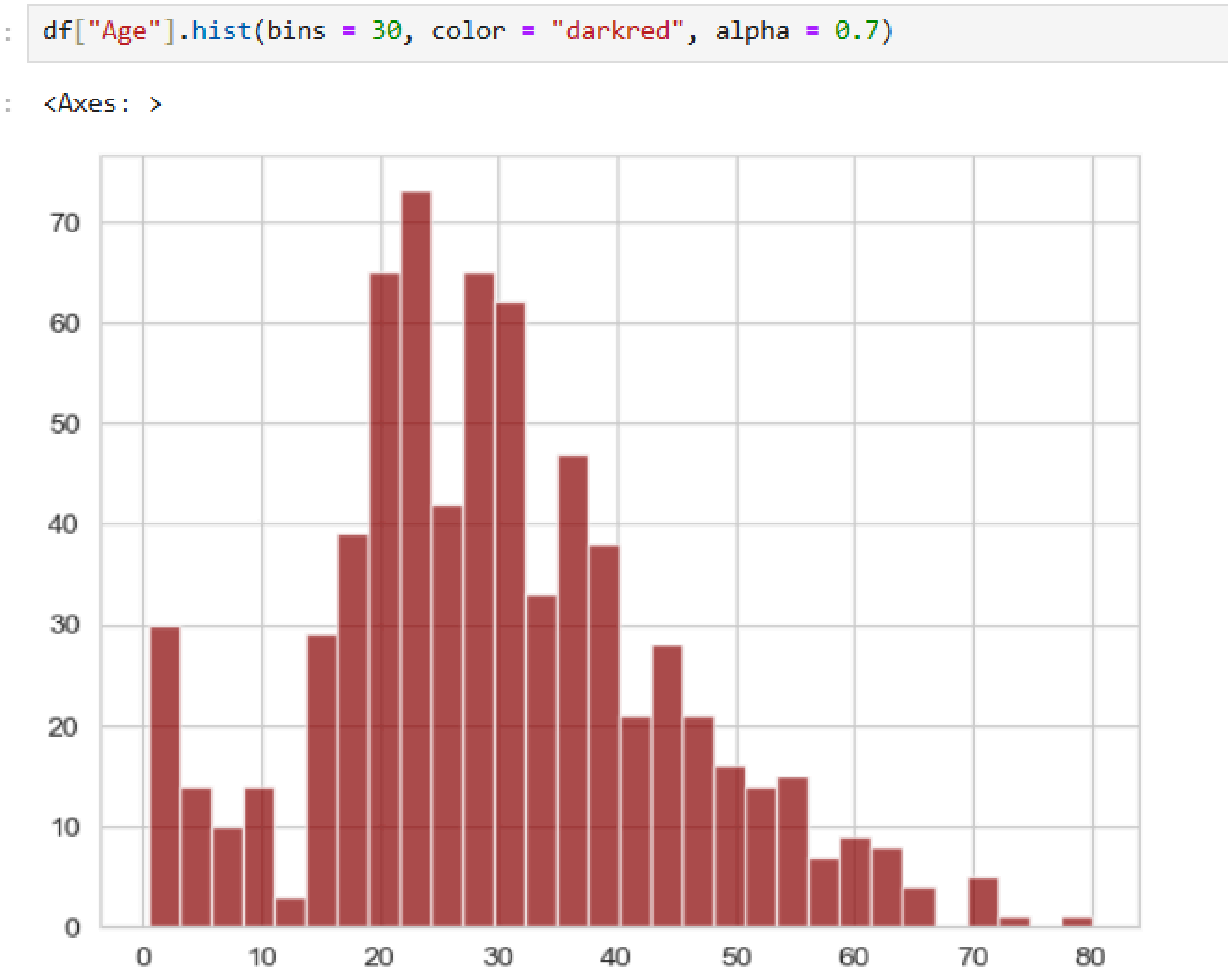


```
sns.set_style("whitegrid")
sns.countplot(x = "Survived", hue = "Pclass", data = df, palette = "rainbow")


### from  this we can clearly see calss3 people died more as compare to other class and
### most survived class are 1st class
```

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

4.4 Age Distribution

- A histogram was plotted to analyze the distribution of age.
- **Insight**: Most passengers were between 20-40 years old.

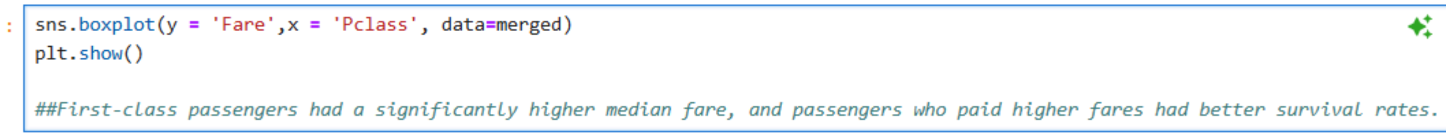


```
df["Age"].hist(bins = 30, color = "darkred", alpha = 0.7)
```

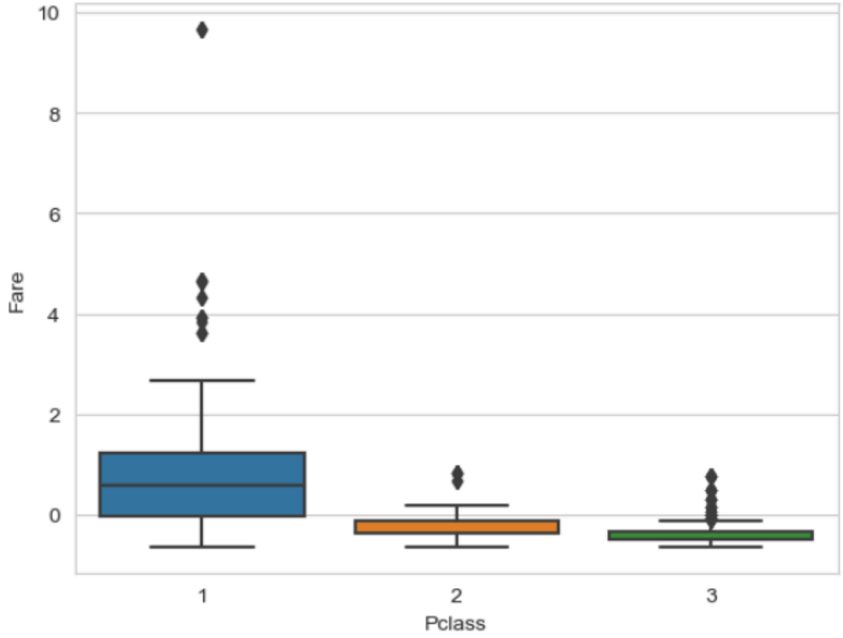
```
<Axes: >
```

4.5 Fare Distribution and Survival

- A boxplot was created to visualize fare distribution across passenger classes.
- **Insight**: First-class passengers had a significantly higher median fare, and passengers who paid higher fares had better survival rates.

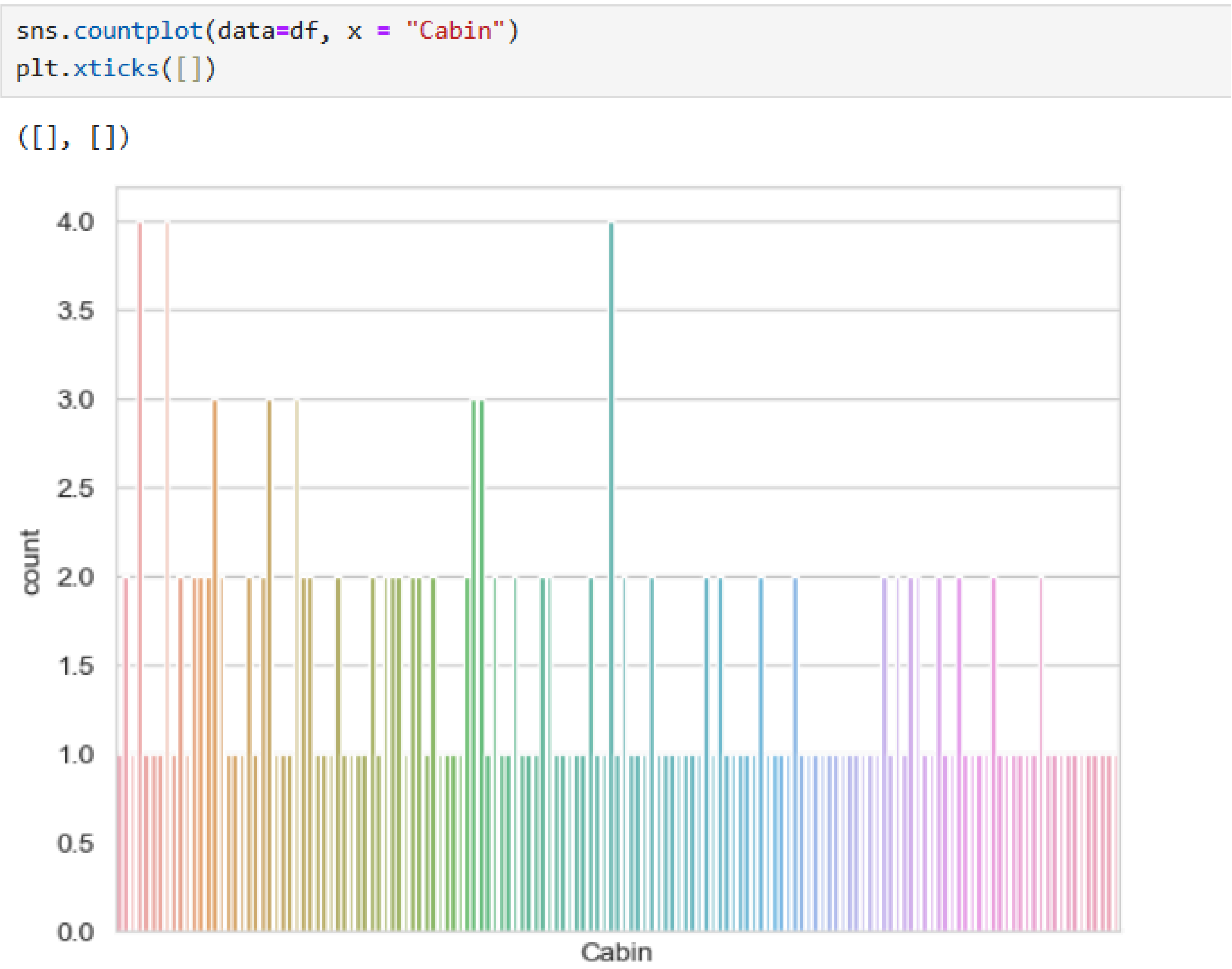


```
sns.boxplot(y = 'Fare',x = 'Pclass', data=merged)
plt.show()

##First-class passengers had a significantly higher median fare, and passengers who paid higher fares had better survival rates.
```



4.6 Cabin Feature Analysis

- A histogram of Cabin values was plotted to understand how many passengers had recorded cabin numbers.
- **Insight**: Most values in the Cabin column were missing, indicating that lower-class passengers had no assigned cabins.
- **Alternative Visualization**: A count plot was used to see the most common recorded cabins.

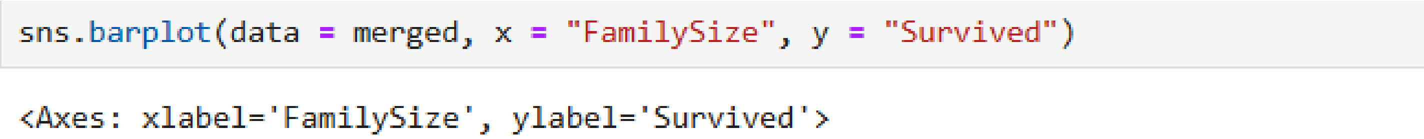


```
sns.countplot(data=df, x = "Cabin")
plt.xticks([])
```

```
([], [])
```

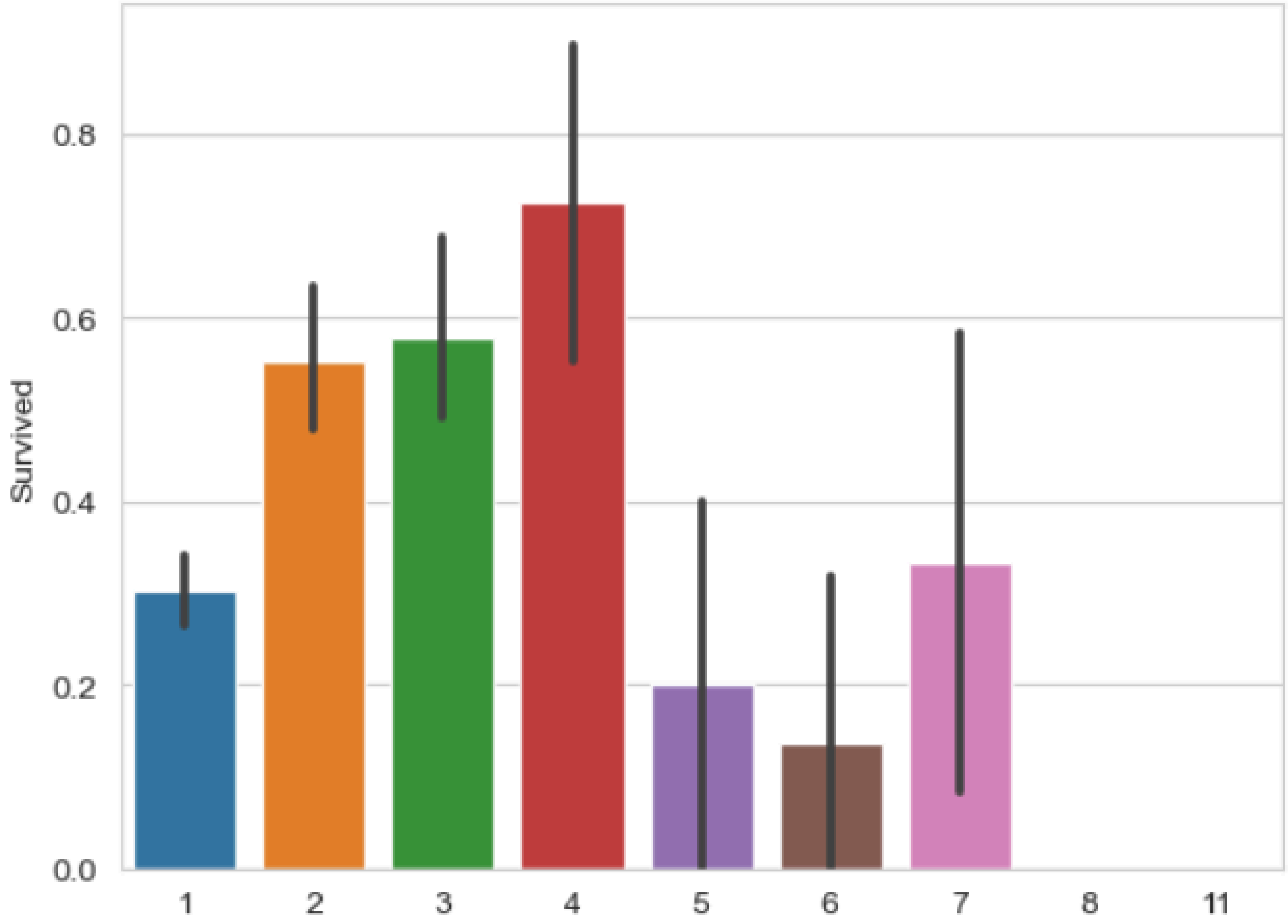
4.7 Family Size and Survival

- A bar plot was created to analyze how family size affected survival.
- **Insight**: Passengers with small families (1-4 members) had higher survival chances, while those traveling alone or in very large families had lower survival rates.



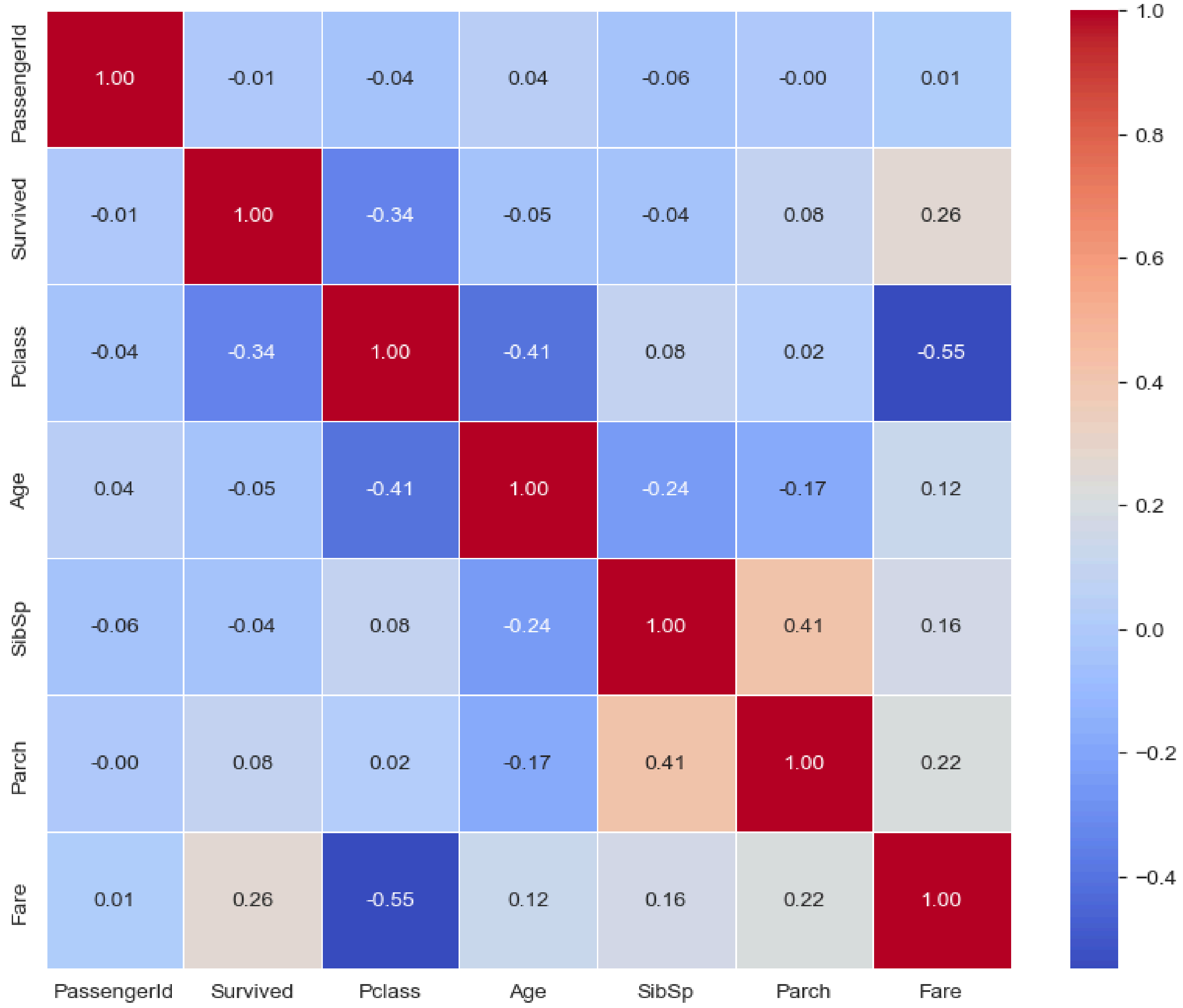
```
sns.barplot(data = merged, x = "FamilySize", y = "Survived")
```

```
<Axes: xlabel='FamilySize', ylabel='Survived'>
```



4.8 Correlation Analysis

- A heatmap was used to visualize correlations between numerical features.
- **Insight**: Pclass had a strong negative correlation with survival, indicating third-class passengers had lower survival rates.
- **Insight**: Fare had a positive correlation with survival, indicating higher-paying passengers had better survival odds.



5. Statistical Analysis

5.1 Descriptive Statistics

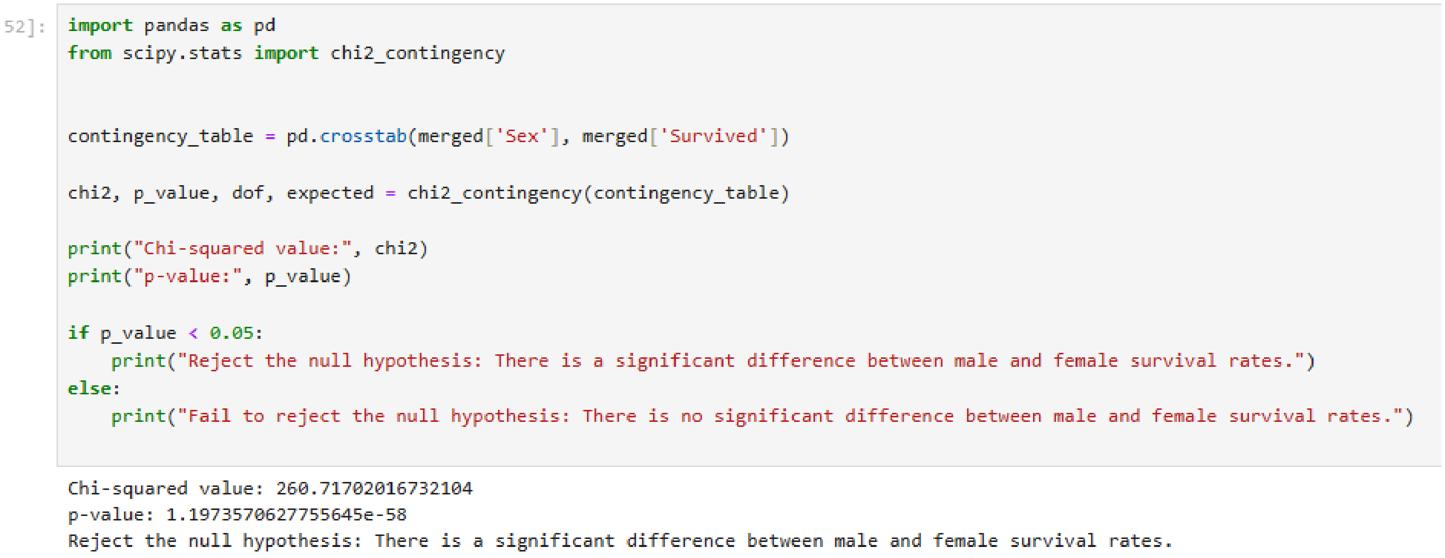
- The mean, median, and standard deviation were computed for numerical features.
- **Insight**: The average fare was significantly higher for first-class passengers.

5.2 Hypothesis Testing

5.2.1 Chi2-Test for Survival Based on Gender

- Null Hypothesis (H_0): There is no significant difference in survival between males and females.
- Alternative Hypothesis (H_1): There is a significant difference in survival between males and females.
- **Result**: The p-value was less than 0.05, leading to the rejection of H_0.

- **Insight**: Gender significantly impacted survival, with females having a higher chance of survival.



52]:

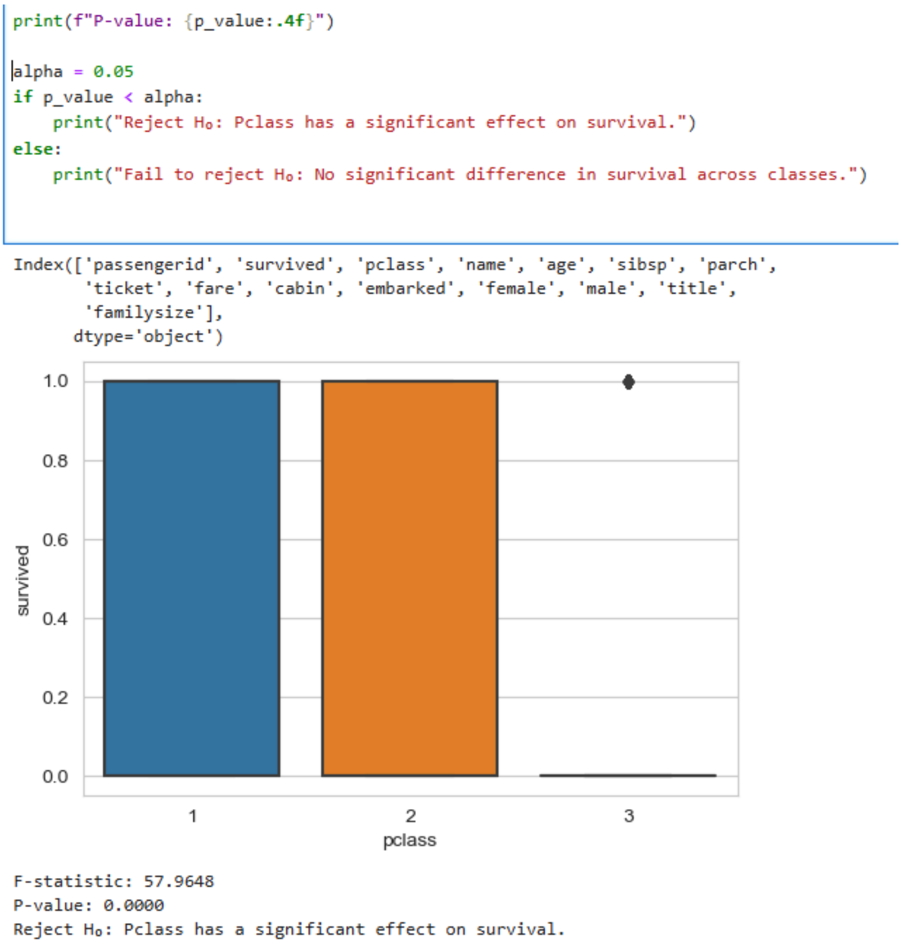
```
import pandas as pd
from scipy.stats import chi2_contingency


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.
```

-

5.2.2 ANOVA Test for Passenger Class and Survival

- Null Hypothesis (H_0): There is no significant difference in survival between passenger classes.
- Alternative Hypothesis (H_1): There is a significant difference in survival between passenger classes.
- **Result**: The test showed a significant p-value, proving survival rates varied across classes.
- **Insight**: First-class passengers had better survival rates.



```
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', 'age', 'sibsp', 'parch',
       'ticket', 'fare', 'cabin', 'embarked', 'female', 'male', 'title',
       'familysize'],
      dtype='object')
```

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

6. Feature Selection

- Features such as Sex, Pclass, Age, Fare, and FamilySize were identified as the most important predictors of survival.
- Cabin was dropped due to too many missing values.
- Ticket was removed as it had no meaningful contribution to survival.

7. Key Insights and Findings

- **Gender Factor**: Females had a significantly higher chance of survival.
- **Passenger Class Influence**: First-class passengers had a much better survival rate than third-class passengers.
- **Family Size Effect**: Passengers with small family sizes (1-4 members) had better survival chances compared to those traveling alone or in very large families.
- **Age Factor**: Children had a better chance of survival compared to older adults.
- **Fare Impact**: Passengers who paid higher fares had better survival rates, correlating with first-class advantages.

8. Conclusion

This extensive analysis of the Titanic dataset provided valuable insights into the factors influencing passenger survival. It highlighted the significant roles of gender, ticket class, and age in survival probability. Further predictive modeling can be conducted using machine learning techniques to enhance accuracy and decision-making.

Future Scope:

- Implement machine learning models like logistic regression, decision trees, and random forests to predict survival.
- Perform deep learning analysis using neural networks.
- Explore additional datasets with more passenger details for further insights.

9. References

- Kaggle Titanic Dataset
- Pandas, Matplotlib, Seaborn, NumPy documentation for data analysis techniques