EXPLORATORY DATA ANALYSIS (EDA)

Titanic Dataset Analysis

Submitted By:

Sharvari Kulkarni

Data Analyst Intern

Tools Used:

- Python
- Pandas
- Matplotlib
- Seaborn
- Jupyter Notebook

Dataset Source:

Titanic Dataset from Kaggle

Date of Submission:

19 February 2026

Introduction

In the field of Data Analytics, understanding the dataset is the most crucial step before building any predictive model. Exploratory Data Analysis (EDA) is a systematic approach used to analyze and summarize datasets using statistical techniques and visualizations. It helps in identifying patterns, detecting anomalies, understanding relationships between variables, and preparing data for further modeling.

This report presents an exploratory data analysis of the Titanic dataset obtained from Kaggle. The dataset contains information about passengers aboard the Titanic, including demographic details, ticket class, fare, family information, and survival status. The primary goal of this analysis is to uncover meaningful insights and understand the factors that influenced passenger survival.

What is Exploratory Data Analysis (EDA)?

Exploratory Data Analysis (EDA) is the process of examining and investigating datasets to summarize their main characteristics, often using visual methods. It is an essential step in the data analysis pipeline because it helps analysts:

- Understand the structure of the dataset
- Identify missing or inconsistent data
- Detect outliers and anomalies
- Discover patterns and trends
- Analyze relationships between variables
- Make informed decisions for feature selection

EDA typically involves three types of analysis:

1. Univariate Analysis – Analysis of a single variable
2. Bivariate Analysis – Analysis of the relationship between two variables
3. Multivariate Analysis – Analysis involving more than two variables

By performing EDA, data analysts gain deeper insights into the dataset, which improves the accuracy and reliability of future predictive models.

Objective of the Analysis

The main objectives of this exploratory data analysis on the Titanic dataset are:

1. To understand the structure and key characteristics of the dataset.
2. To identify and handle missing values.
3. To analyze the distribution of individual variables such as age, gender, passenger class, and fare.
4. To examine the relationship between survival and other factors like gender, passenger class, age, and fare.
5. To identify patterns, trends, and anomalies in the dataset.
6. To determine the most influential features affecting passenger survival.

Through this analysis, we aim to extract meaningful insights that explain survival patterns and provide a strong foundation for building predictive machine learning models in future stages.

Dataset Description

The dataset used for this analysis is the Titanic Dataset, obtained from Kaggle. It contains information about passengers who were aboard the RMS Titanic and whether they survived or not.

Number of Rows and Columns

- Total Rows (Observations): 891
- Total Columns (Features): 12

Each row represents an individual passenger, and each column represents a specific attribute related to that passenger.

Important Features

The dataset consists of the following key features:

- PassengerId – Unique identifier for each passenger
- Survived – Passenger survived or not (Yes = 1, No = 0)
- Pclass – Passenger class (1 = First, 2 = Second, 3 = Third)
- Name – Name of the passenger
- Sex – Gender of the passenger
- Age – Age in years
- 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)

Among these features, the most influential variables for survival analysis are:

- Sex
- Pclass
- Age
- Fare
- Family-related features (SibSp, Parch)

These features are particularly important because they help in understanding demographic, social, and economic factors influencing survival.

Target Variable

The target variable in this dataset is:

Survived

- 0 → Did Not Survive
- 1 → Survived

This is a binary categorical variable, and the primary objective of this analysis is to understand the factors that influenced passenger survival.

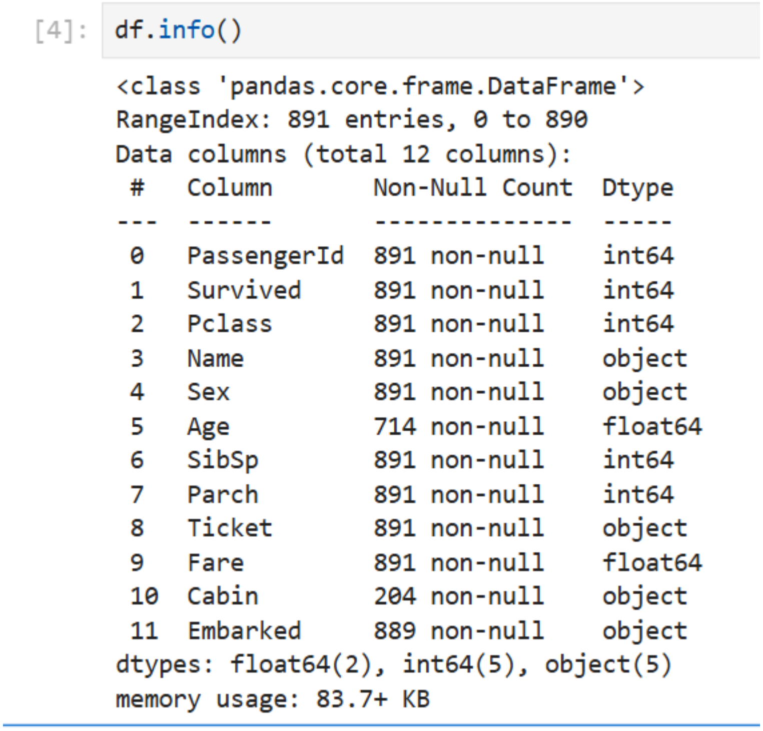
Dataset Nature

- The dataset contains both numerical variables (Age, Fare, SibSp, Parch) and categorical variables (Sex, Pclass, Embarked).
- Some features contain missing values, particularly Age and Cabin, which require data preprocessing.

Initial Data Exploration

Using .info() revealed:

- 891 entries and 12 columns
- Missing values in Age and Cabin
- Presence of both numerical and categorical variables

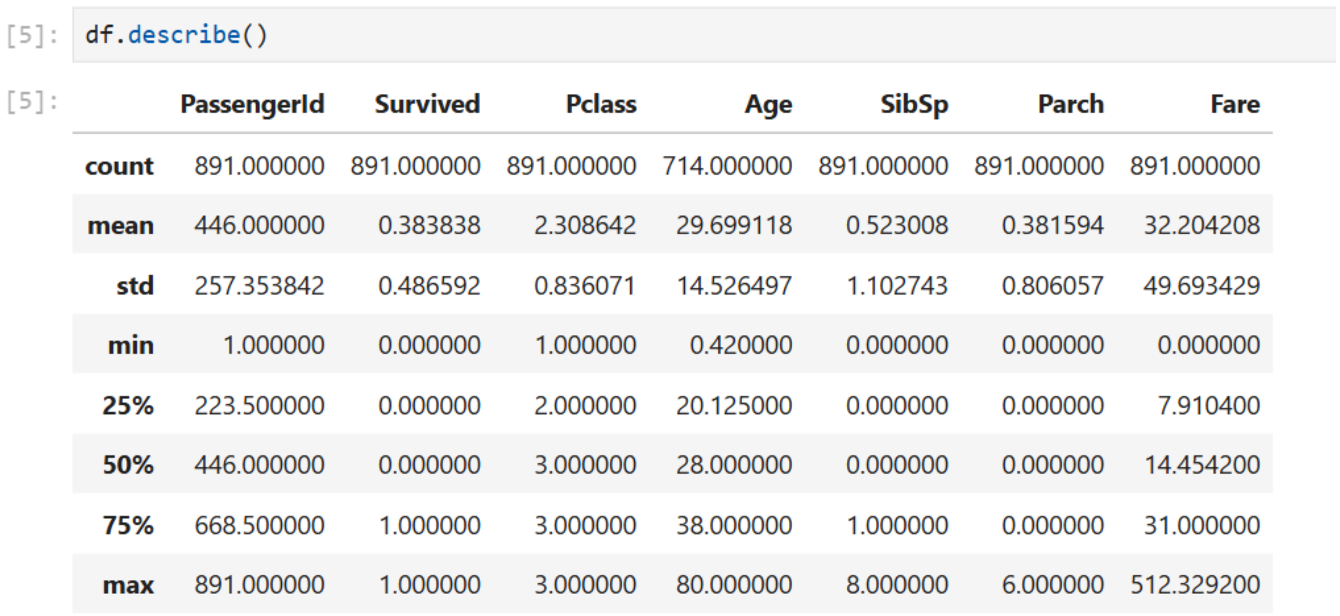


```
[4]: df.info()
```

```
<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
```

Using .describe() revealed:

- Average age is approximately 29 years
- Fare distribution is highly skewed
- Approximately 38% passengers survived



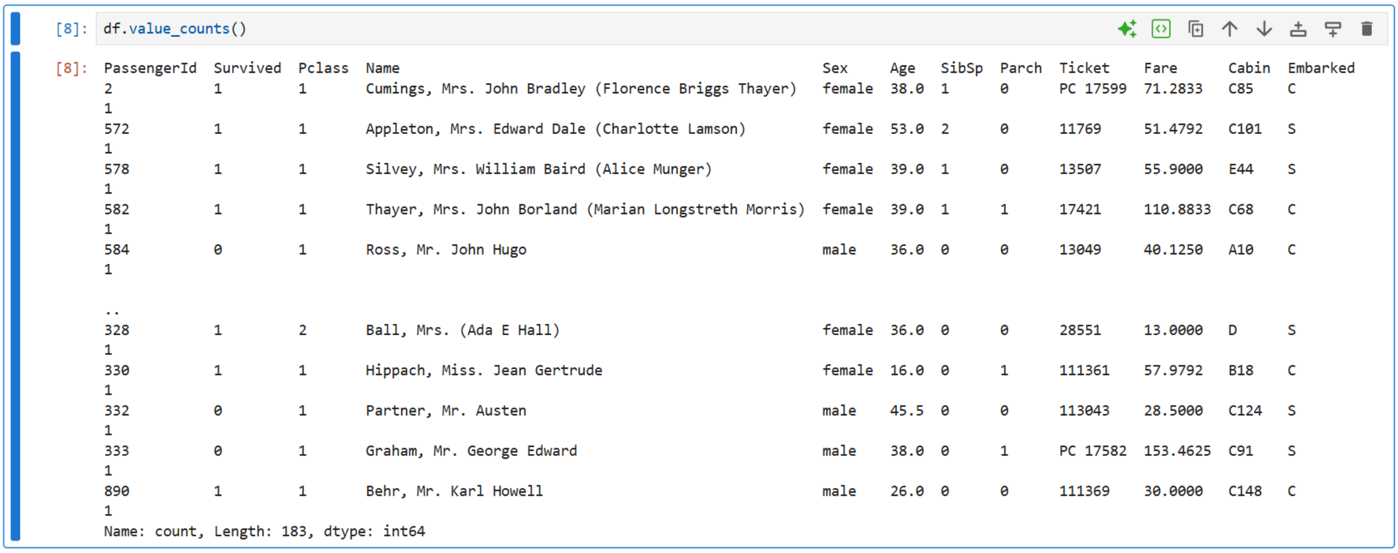
```
[5]: df.describe()
```

[5]:

	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

Using .value_counts():

- More males than females
- Majority passengers were in 3rd class
- More non-survivors than survivors



```
[8]: df.value_counts()
```

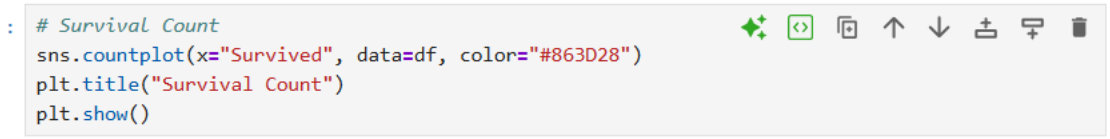
```
[8]: PassengerId  Survived  Pclass  Name                                                Sex     Age   SibSp  Parch  Ticket    Fare      Cabin  Embarked
     2            1         1       Cumings, Mrs. John Bradley (Florence Briggs Thayer)  female  38.0  1      0      PC 17599  71.2833   C85    C
     1
     572          1         1       Appleton, Mrs. Edward Dale (Charlotte Lamson)        female  53.0  2      0      11769     51.4792   C101   S
     1
     578          1         1       Silvey, Mrs. William Baird (Alice Munger)            female  39.0  1      0      13507     55.9000   E44    S
     1
     582          1         1       Thayer, Mrs. John Borland (Marian Longstreth Morris) female  39.0  1      1      17421     110.8833  C68    C
     1
     584          0         1       Ross, Mr. John Hugo                                  male    36.0  0      0      13049     40.1250   A10    C
     1

     ..
     328          1         2       Ball, Mrs. (Ada E Hall)                              female  36.0  0      0      28551     13.0000   D      S
     1
     330          1         1       Hippach, Miss. Jean Gertrude                         female  16.0  0      1      111361    57.9792   B18    C
     1
     332          0         1       Partner, Mr. Austen                                  male    45.5  0      0      113043    28.5000   C124   S
     1
     333          0         1       Graham, Mr. George Edward                            male    38.0  0      1      PC 17582  153.4625  C91    S
     1
     890          1         1       Behr, Mr. Karl Howell                                male    26.0  0      0      111369    30.0000   C148   C
     1
     Name: count, Length: 183, dtype: int64
```

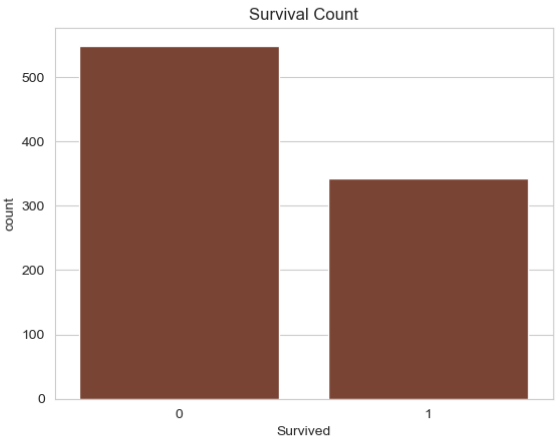
Univariate Analysis

Univariate analysis involves examining one variable at a time to understand its distribution, central tendency, and overall pattern.

1. Survival Distribution



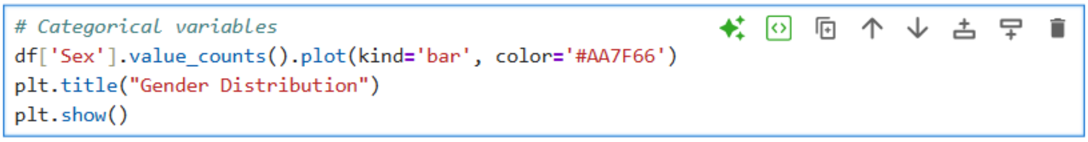
```
# Survival Count
sns.countplot(x="Survived", data=df, color="#863D28")
plt.title("Survival Count")
plt.show()
```



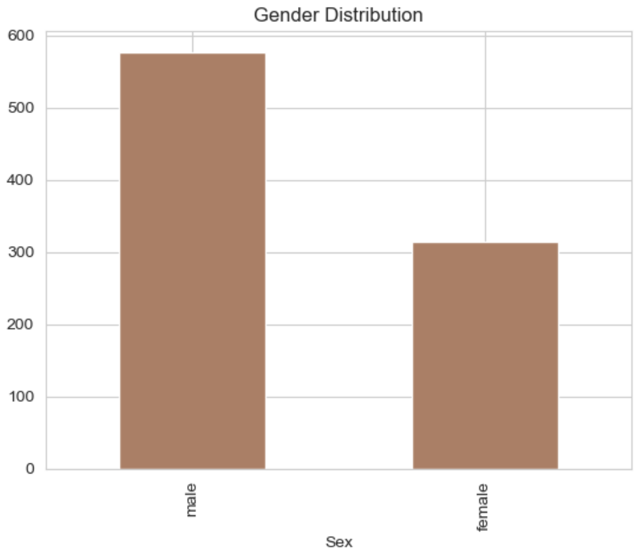
Observation:

- The number of passengers who did not survive is higher than those who survived.
- This indicates class imbalance in the dataset.

2. Gender Distribution



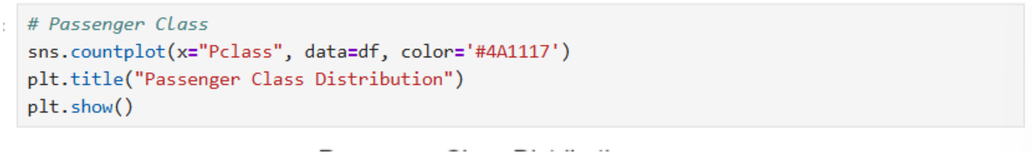
```
# Categorical variables
df['Sex'].value_counts().plot(kind='bar', color='#AA7F66')
plt.title("Gender Distribution")
plt.show()
```



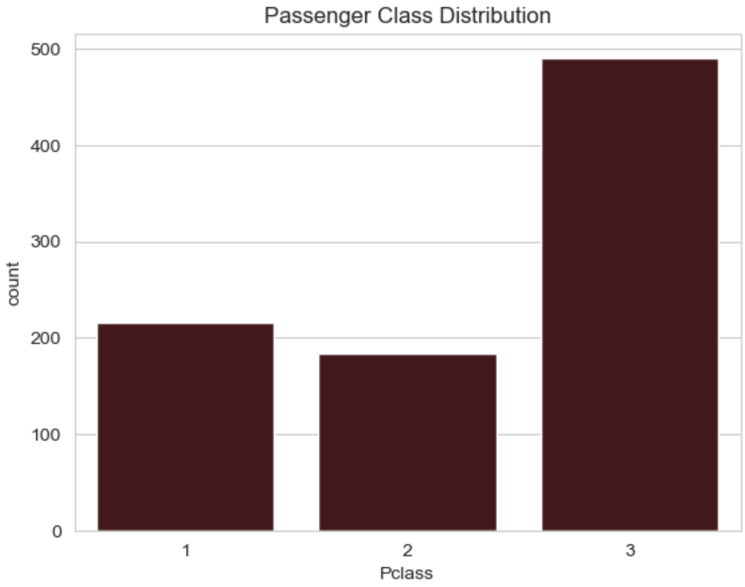
Observation:

- The dataset contains more male passengers than female passengers.
- Gender distribution may significantly influence survival analysis.

3. Passenger Class Distribution



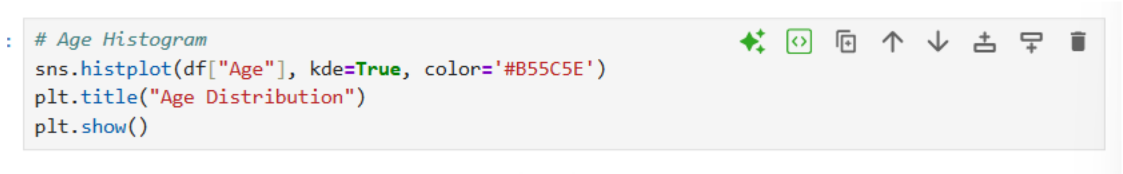
```
# Passenger Class
sns.countplot(x="Pclass", data=df, color='#4A1117')
plt.title("Passenger Class Distribution")
plt.show()
```



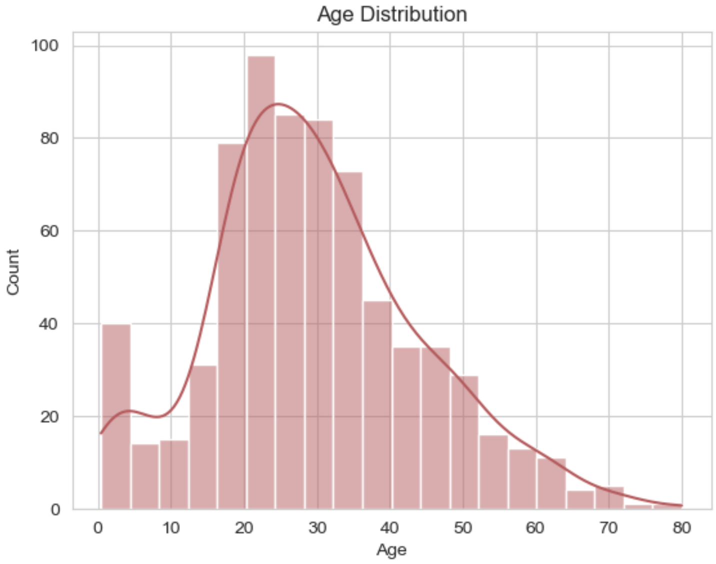
Observation**:**

- Most passengers traveled in 3rd class.
- First-class passengers represent the smallest proportion.

4. Age Distribution



```
# Age Histogram
sns.histplot(df["Age"], kde=True, color='#B55C5E')
plt.title("Age Distribution")
plt.show()
```



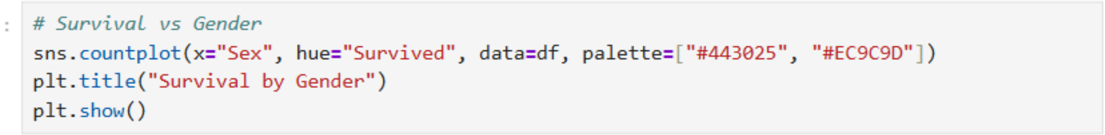
Observation:

- Most passengers were between 20 and 40 years old.
- The distribution shows slight right skewness.

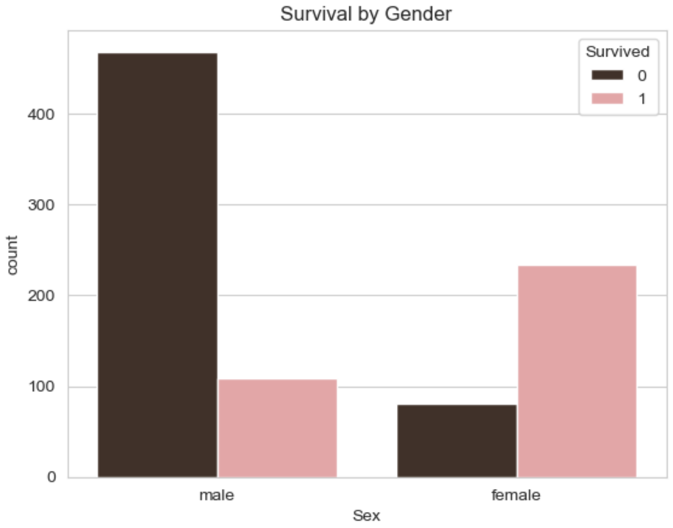
Bivariate Analysis

Bivariate analysis examines the relationship between two variables. In this analysis, the primary focus is to understand how different features influence the target variable Survived in the Titanic dataset from Kaggle.

1. Survival vs Gender

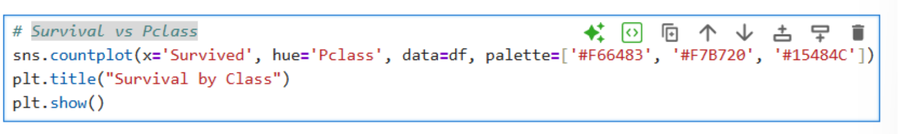


```
# Survival vs Gender
sns.countplot(x="Sex", hue="Survived", data=df, palette=["#443025", "#EC9C9D"])
plt.title("Survival by Gender")
plt.show()
```

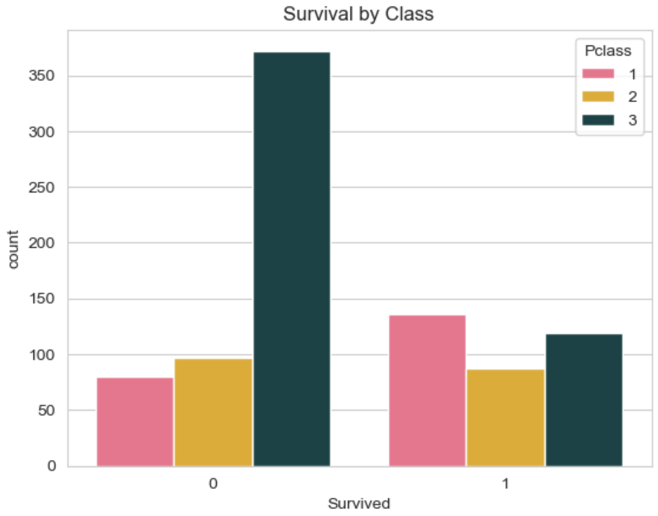


Observation: Females had higher survival rate than males.

2. Survival vs Pclass



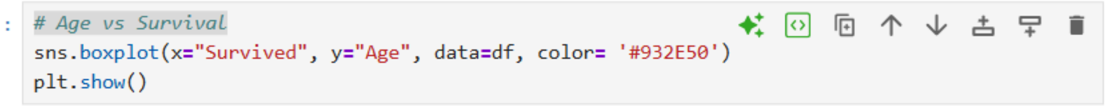
```
# Survival vs Pclass
sns.countplot(x='Survived', hue='Pclass', data=df, palette=['#F66483', '#F7B720', '#15484C'])
plt.title("Survival by Class")
plt.show()
```



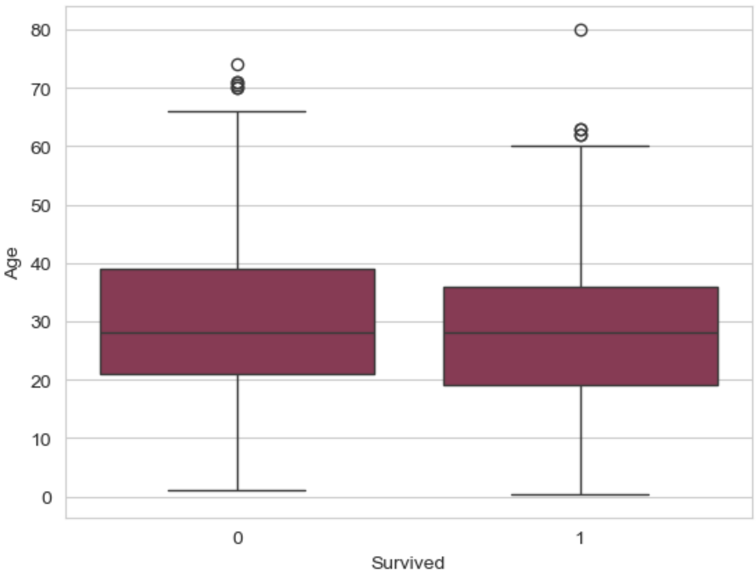
Observation:

- 1st class passengers survived more than 3rd class.
- More people died from 3rd class than 1st class.

3. Age vs Survival



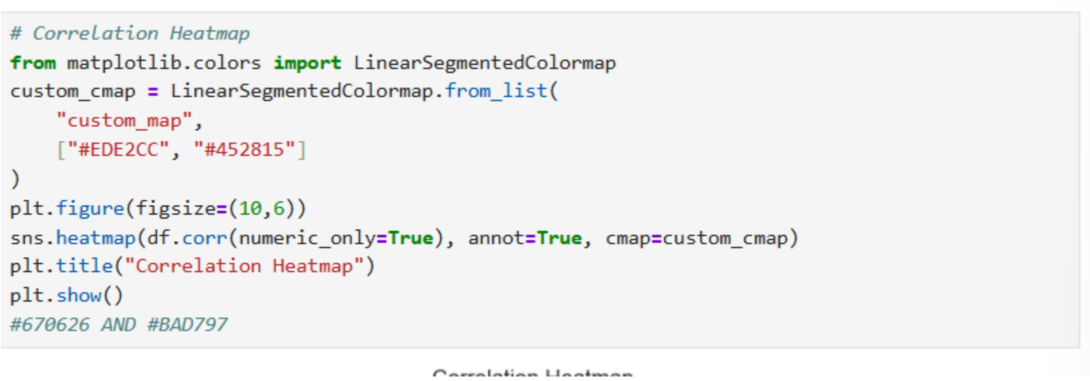
```
# Age vs Survival
sns.boxplot(x="Survived", y="Age", data=df, color= '#932E50')
plt.show()
```



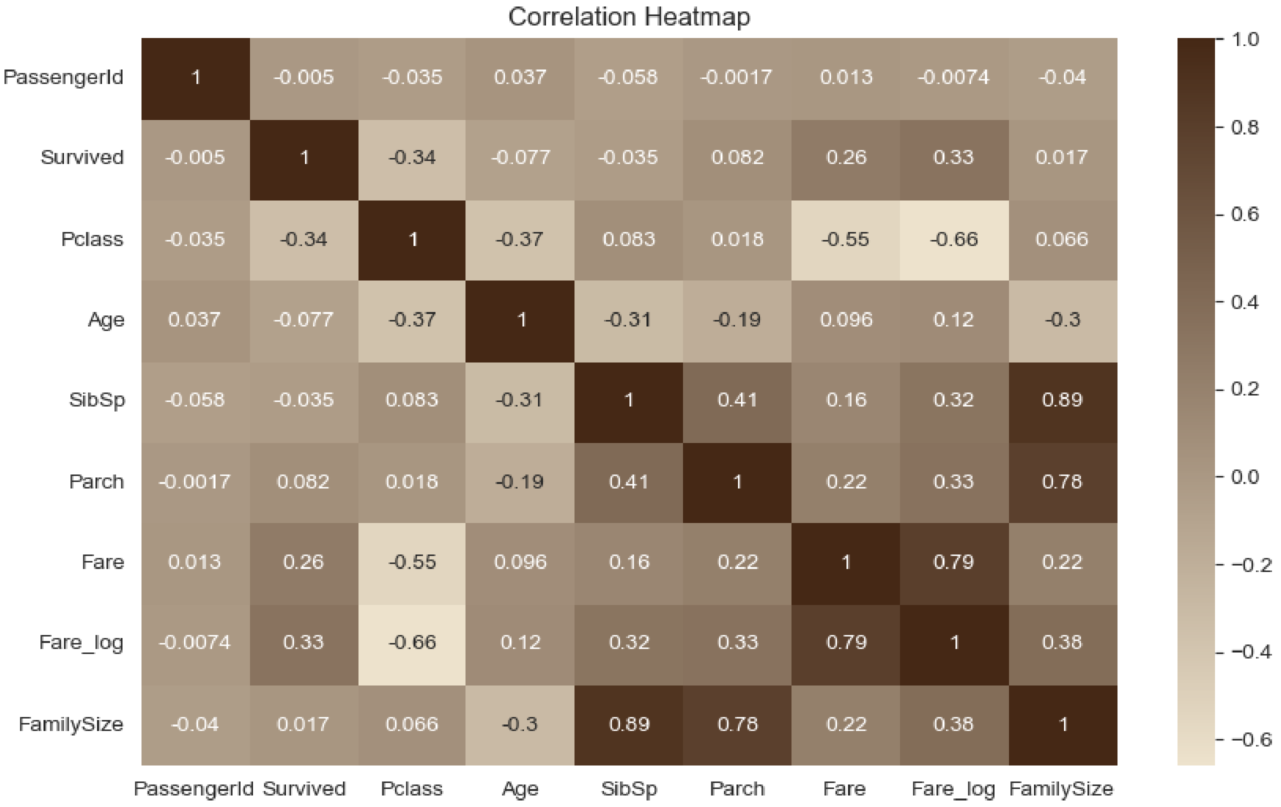
Observation: Younger passengers had slightly higher survival.

Multivariate Analysis

1. Correlation Heatmap



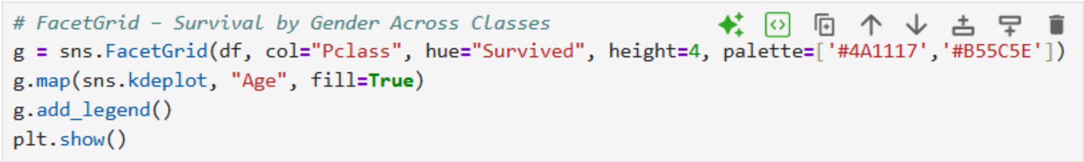
```
# Correlation Heatmap
from matplotlib.colors import LinearSegmentedColormap
custom_cmap = LinearSegmentedColormap.from_list(
    "custom_map",
    ["#EDE2CC", "#452815"]
)
plt.figure(figsize=(10,6))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap=custom_cmap)
plt.title("Correlation Heatmap")
plt.show()
#670626 AND #BAD797
```



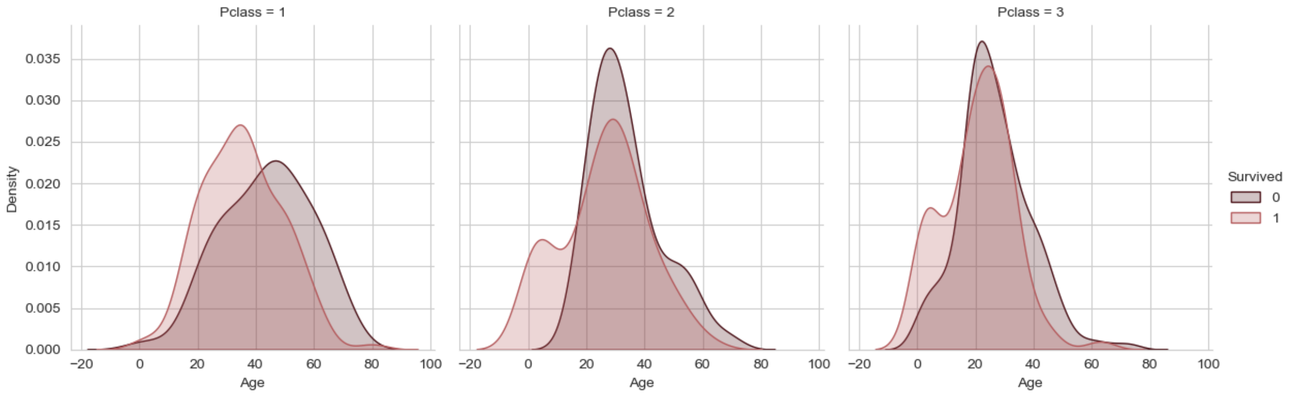
Observation:

- Pclass negatively correlated with Survival.
- Fare positively correlated with Survival.

2. FacetGrid – Survival by Gender Across Classes



```
# FacetGrid – Survival by Gender Across Classes
g = sns.FacetGrid(df, col="Pclass", hue="Survived", height=4, palette=['#4A1117','#B55C5E'])
g.map(sns.kdeplot, "Age", fill=True)
g.add_legend()
plt.show()
```



Observation:

- In 1st class, survival is high across many age groups.
- In 3rd class, most age groups show low survival density.

Key Findings

After performing univariate, bivariate, and multivariate analysis on the Titanic dataset obtained from Kaggle, the following key insights were identified:

- Gender is the Strongest Predictor

  Female passengers had a significantly higher survival rate compared to male passengers. This indicates that gender played a major role in survival probability.
- Passenger Class Influenced Survival

  First-class passengers had the highest survival rate, while third-class passengers had the lowest. This suggests that socioeconomic status impacted access to lifeboats and safety measures.
- Fare is Positively Correlated with Survival

  Passengers who paid higher fares were more likely to survive. Since fare is closely related to passenger class, it reinforces the importance of socioeconomic factors.
- Age Has Moderate Influence

  Younger passengers showed slightly better survival rates compared to older individuals. However, age alone was not a strong determining factor.
- Family Size Affected Survival

Passengers traveling in small families (2–4 members) had higher survival rates. Very large families and individuals traveling alone had comparatively lower survival probabilities.

- Data Imbalance Observed

  The dataset contains more non-survivors than survivors, indicating class imbalance, which is important for future predictive modeling.

Conclusion

The exploratory data analysis revealed that survival on the Titanic was strongly influenced by gender, passenger class, and fare. Female passengers and first-class travelers had significantly higher survival probabilities, highlighting the impact of social and economic status during the disaster.

Although age and family size showed some influence, they were not as dominant as gender and class. Proper handling of missing values and visualization techniques helped uncover meaningful relationships within the dataset.

Overall, this analysis provides valuable insights into survival patterns and establishes a solid foundation for building predictive machine learning models in future stages.

The findings from this exploratory analysis demonstrate how demographic and socioeconomic factors shaped survival outcomes, emphasizing the importance of thorough data exploration before model development.