**Exploratory Data Analysis (EDA) Report for the Titanic Dataset**

**Brief Introduction of the Dataset**

The Titanic dataset is a historical dataset derived from the tragic sinking of the RMS Titanic in 1912. It contains data on 891 passengers and includes information such as their demographics, ticket class, fare, and survival status. The dataset is widely used in data science to explore patterns and relationships in real-world data and to demonstrate techniques like data cleaning, exploratory data analysis (EDA), and machine learning.

Key attributes in the dataset include:

* **Survived**: Indicates whether the passenger survived (1) or not (0).
* **Pclass**: The passenger's ticket class (1st, 2nd, or 3rd class).
* **Sex**: The gender of the passenger.
* **Age**: The age of the passenger.
* **SibSp**: The number of siblings or spouses aboard the Titanic.
* **Parch**: The number of parents or children aboard the Titanic.
* **Fare**: The price of the passenger's ticket.
* **Embarked**: The port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

This dataset serves as a classic example for studying relationships between categorical and numerical data, identifying survival patterns, and visualizing trends.

**Data Cleaning Steps Taken**

To ensure the dataset was prepared for analysis and to improve the quality of the insights, the following data cleaning steps were performed:

**Handling Missing Values**

The dataset contained missing data in several columns, including Age, Cabin, and Embarked. These issues were addressed as follows:

* **For Age**, the missing values were replaced with the median age of the passengers. The median was chosen because it is less affected by outliers and provided a robust measure of central tendency for this continuous variable.
* **For Embarked**, the missing values were filled with the mode, which was the most frequent value, representing the majority of passengers who embarked from Southampton.
* **For Cabin**, the column was dropped entirely due to the high percentage of missing values, which exceeded 75%. Retaining the column would have added noise to the dataset without providing significant value to the analysis.

#### ****Encoding Categorical Variables****

Several categorical columns were converted into numerical formats to facilitate analysis and modeling:

* **The Sex column**, which contained values such as 'male' and 'female', was encoded as numeric values, with males represented by 0 and females by 1.
* **The Embarked column**, which indicated the port of embarkation, was transformed into binary variables using one-hot encoding. This process created separate columns for each port of embarkation while eliminating redundancy.

#### ****Removing Irrelevant or Redundant Columns****

Certain columns were deemed irrelevant to the analysis and were removed. These included:

* **Name** and **Ticket**, as they were non-numeric and did not contribute meaningful information to the survival analysis.
* **PassengerId**, which served only as a unique identifier and was not relevant for pattern detection or modeling.

#### ****Addressing Outliers****

The dataset was reviewed for outliers, particularly in continuous variables such as Fare and Age. While extreme values were identified, they were retained in the dataset as they represented plausible real-world scenarios, such as wealthy passengers paying high fares or younger children traveling.

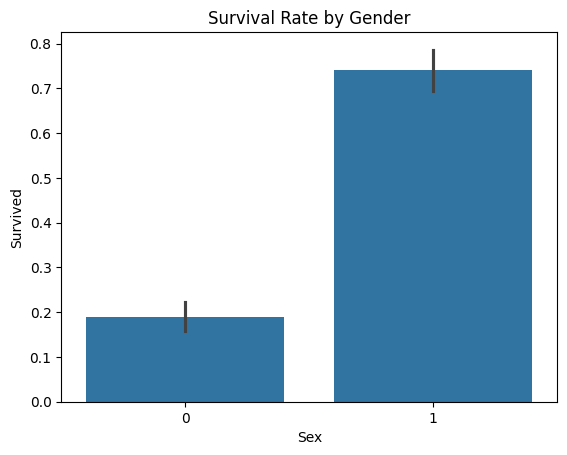
#### ****Ensuring Data Consistency****

The dataset was examined for duplicate rows to maintain data integrity. No duplicates were found, ensuring the data was unique and reliable for analysis.

### ****Key Visualizations and Trends****

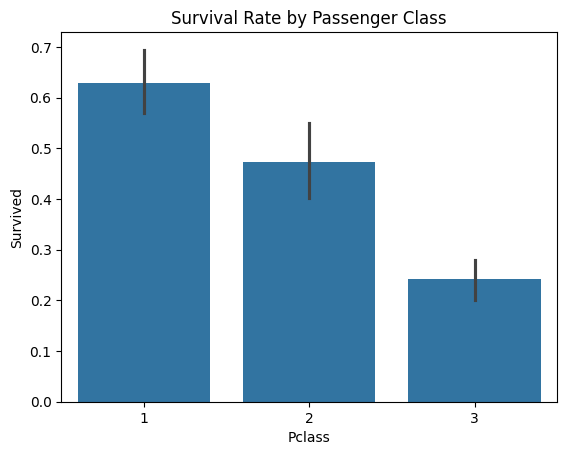
The exploratory data analysis (EDA) involved generating various visualizations to uncover patterns and trends in the Titanic dataset. Here are the key insights obtained:

#### ****Survival Rates by Gender****



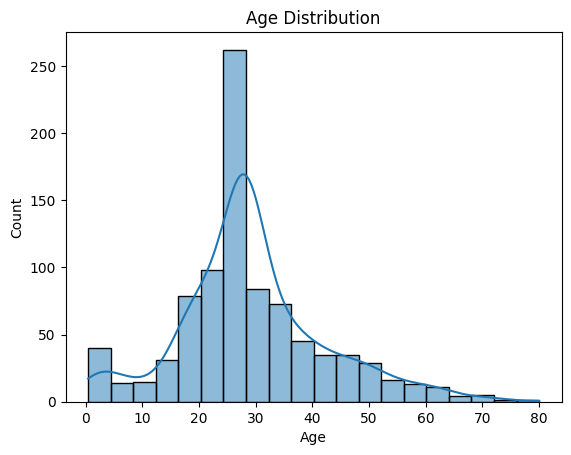
The bar graph illustrates the survival rates of passengers based on gender, revealing a significant disparity. Females had a substantially higher survival rate, with over 70% surviving, compared to less than 20% for males. This trend highlights the prioritization of women during the Titanic evacuation, reflecting the "women and children first" policy. The visual emphasizes the critical role gender played in determining survival outcomes.

#### ****Survival Rates by Passenger Class (Pclass)****



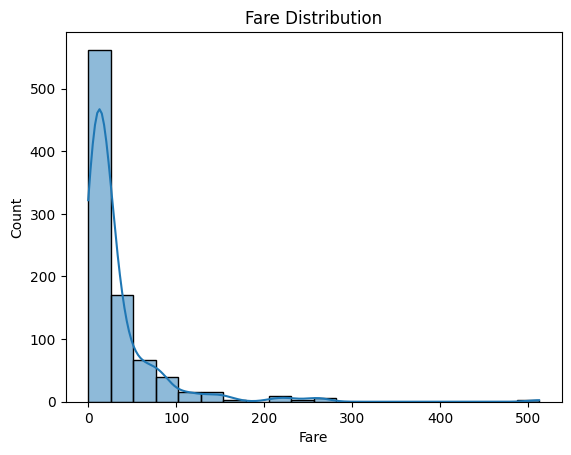
The graph depicts survival rates across the three passenger classes, highlighting clear disparities based on socioeconomic status. First-class passengers had the highest survival rate, with nearly 63% surviving, while second-class passengers showed a moderate survival rate, and third-class passengers had the lowest, with only about 24% surviving. This trend indicates that passengers in higher classes likely had better access to lifeboats and assistance during the evacuation, emphasizing the influence of class and privilege on survival outcomes.

#### ****Age Distribution****



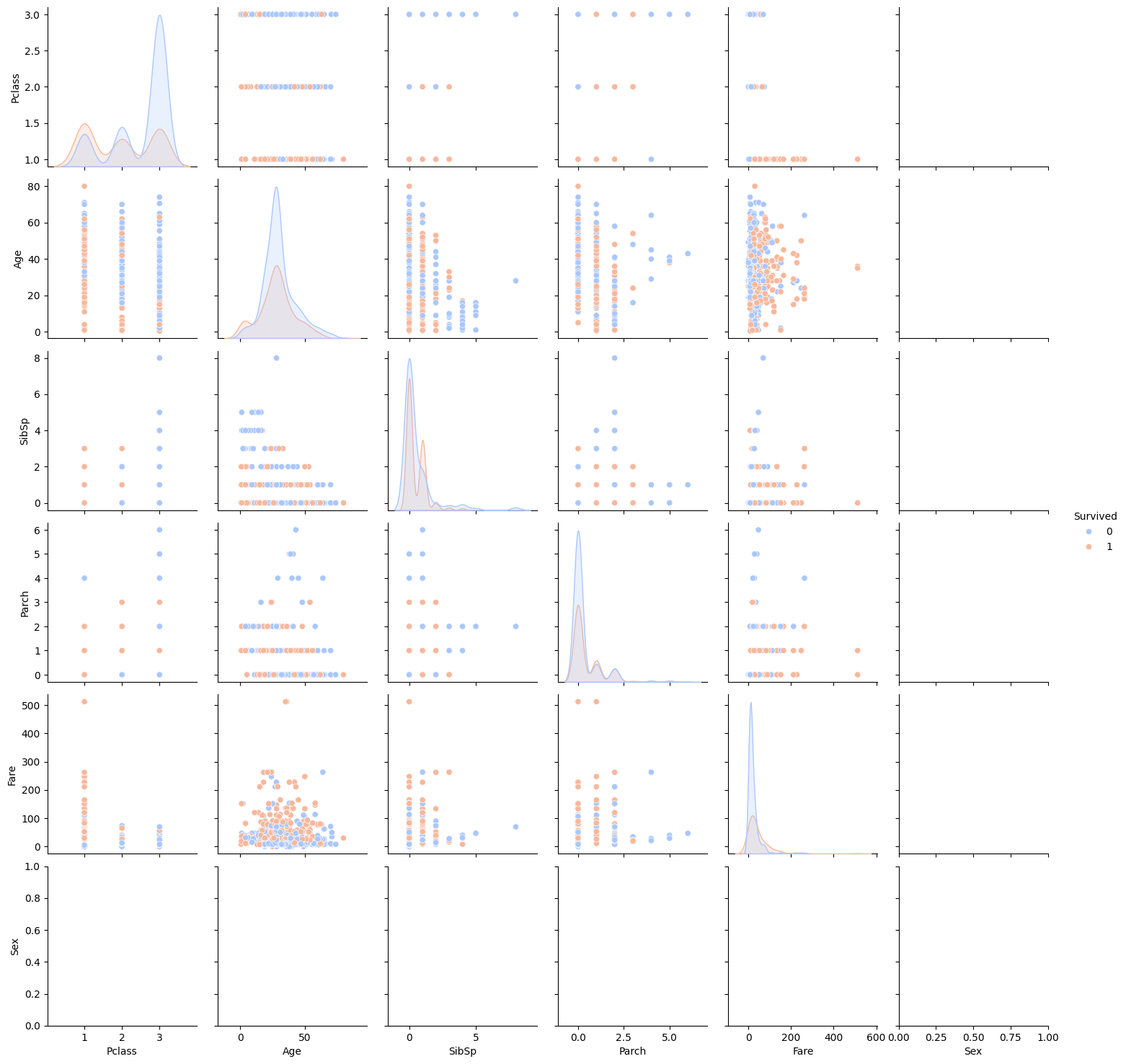
The age distribution plot shows the frequency of passengers' ages aboard the Titanic. The distribution is right-skewed, with a prominent peak around the age of 20, indicating that the majority of passengers were young adults. The density curve further confirms this trend, suggesting that most passengers were between 20 and 30 years old. As the age increases, the number of passengers decreases, with very few elderly passengers and children above the age of 70. This distribution reflects the demographic composition of the passengers, with a higher concentration of younger individuals and a decline in older age groups.

#### ****Fare Distribution****



The fare distribution plot highlights the wide range of ticket prices among Titanic passengers, with fares skewed heavily towards the lower end. A significant portion of passengers paid low fares, corresponding to those in third class. However, a smaller number of passengers paid exceptionally high fares, indicative of first-class ticket prices. This right-skewed distribution reflects the socioeconomic diversity of passengers aboard the Titanic. Additionally, survivors are observed to have generally paid higher fares, further emphasizing the correlation between fare amount and survival likelihood.

#### ****Pairwise Relationships****



The pairplot of selected variables, including Age, Fare, Pclass, and Survived, reveals clear patterns and relationships among these features. Notably, there are distinct clusters of first-class passengers, often associated with higher fares, which show a positive correlation with survival. This indicates that passengers in higher classes, who paid more for their tickets, had a significantly better chance of survival. The visualization underscores the influence of class and fare on survival rates, highlighting the advantage that wealthier passengers had during the Titanic evacuation.

### ****Observations and Conclusions****

The analysis of the Titanic dataset reveals that several key factors influenced passenger survival rates during the disaster. Gender was one of the most significant determinants, with females having a markedly higher survival rate than males. Over 70% of females survived, compared to less than 20% of males. This disparity aligns with the "women and children first" policy that prioritized women and children during the evacuation, highlighting the role of gender in survival outcomes. Similarly, passenger class (Pclass) was also a crucial factor. First-class passengers had the highest survival rates, followed by second-class, and then third-class passengers, who experienced the lowest survival rates. This indicates that passengers in higher classes, who paid more for their tickets, were more likely to survive due to better access to lifeboats and resources.

Age emerged as another important factor in determining survival. Younger passengers, particularly children under the age of 10, had higher survival rates than older passengers, reflecting the prioritization of children during the rescue efforts. The analysis also showed that the median age of survivors was slightly younger compared to non-survivors, suggesting that younger individuals were more likely to be saved. In addition, the fare paid by passengers had a notable correlation with survival. Survivors generally paid higher fares, with first-class passengers—who paid the most—being more likely to survive. This trend suggests that wealthier passengers, who were more likely to be in higher social classes, had a better chance of survival due to their proximity to lifeboats and other safety measures.

Lastly, the embarkation point also played a role in survival outcomes, with passengers who embarked from Cherbourg exhibiting higher survival rates compared to those who boarded in Southampton or Queenstown. This could be attributed to the higher proportion of first-class passengers who embarked from Cherbourg. In conclusion, the data highlights the complex interplay of factors—such as gender, class, age, fare, and embarkation point—that influenced survival during the Titanic disaster. The findings suggest that those who were wealthier or in higher social classes had a greater chance of survival, while women and children were prioritized in the evacuation process, emphasizing the social dynamics at play during this tragic eve