**Titanic**

# **Introduction:**

In the Titanic dataset, we will perform several preprocessing steps, including imputing null values with the median and mode, checking whether all columns are in the correct data types, examining correlations, identifying outliers, conducting statistical analysis, and applying standard scaling. We will also perform Principal Component Analysis (PCA) and visualize outliers using box plots and histograms. Next, we will connect Python with MySQL to identify trends and patterns, and finally, visualize the results of SQL queries in Streamlit.

Handling Missing Data:

In three columns, there is missing data. The "Cabin" column has more than 50% missing values, so I dropped this column. The "Age" column has some missing values, which were imputed using the median. The "Embarked" column has very few missing values, which were imputed using the mode.

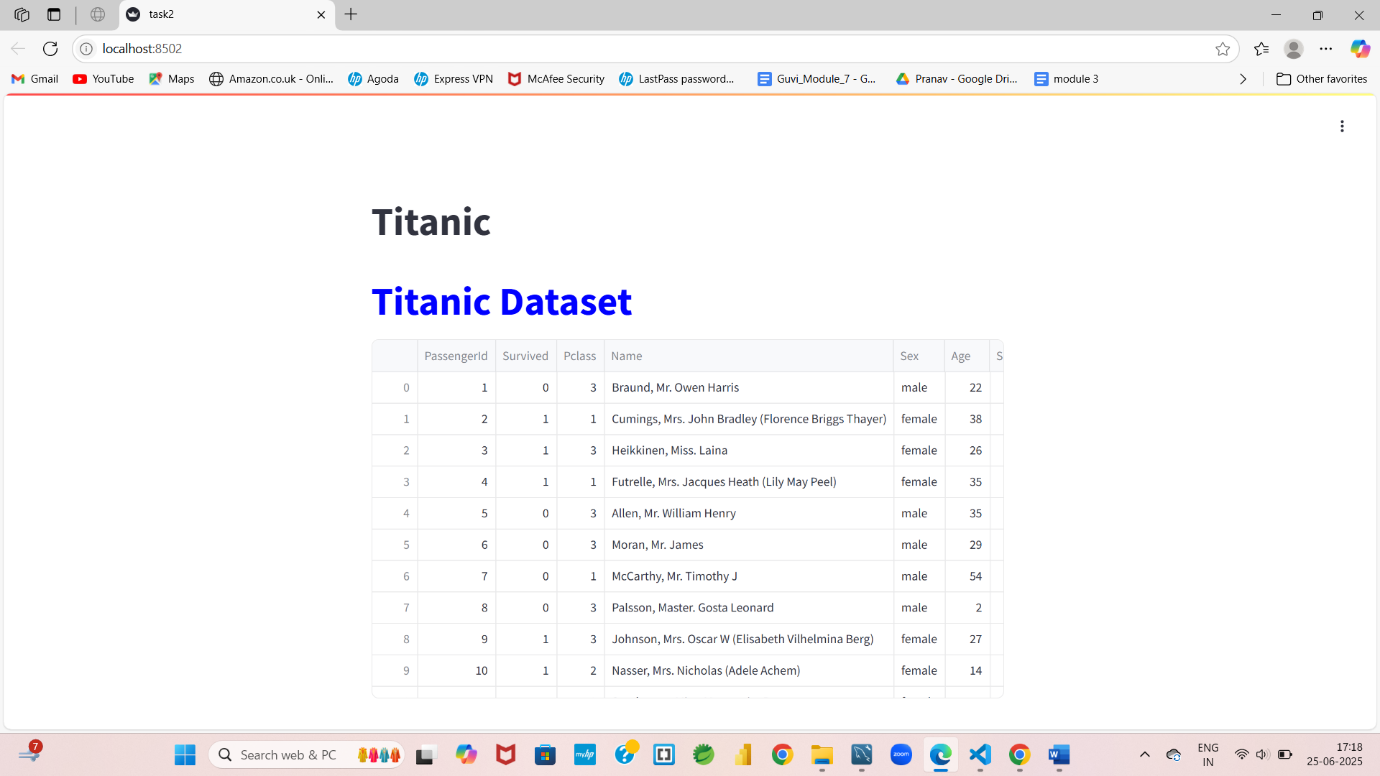
Encoding:

There are some categorical columns, and I performed encoding for only two of them. For the 'Sex' column, I used One-Hot Encoding, and for the 'Embarked' column, I applied One-Hot Encoding to create separate binary columns for each category.

Standardizing and Visualizing Outliers:  
 Except for the target variable, "Name," and the "Ticket" column, I standardized the other columns. There are no outliers in the target column, but the "Age" column and some other columns contain many outliers. If I were to remove these outliers, nearly half of the data would be lost, so I decided not to remove them.

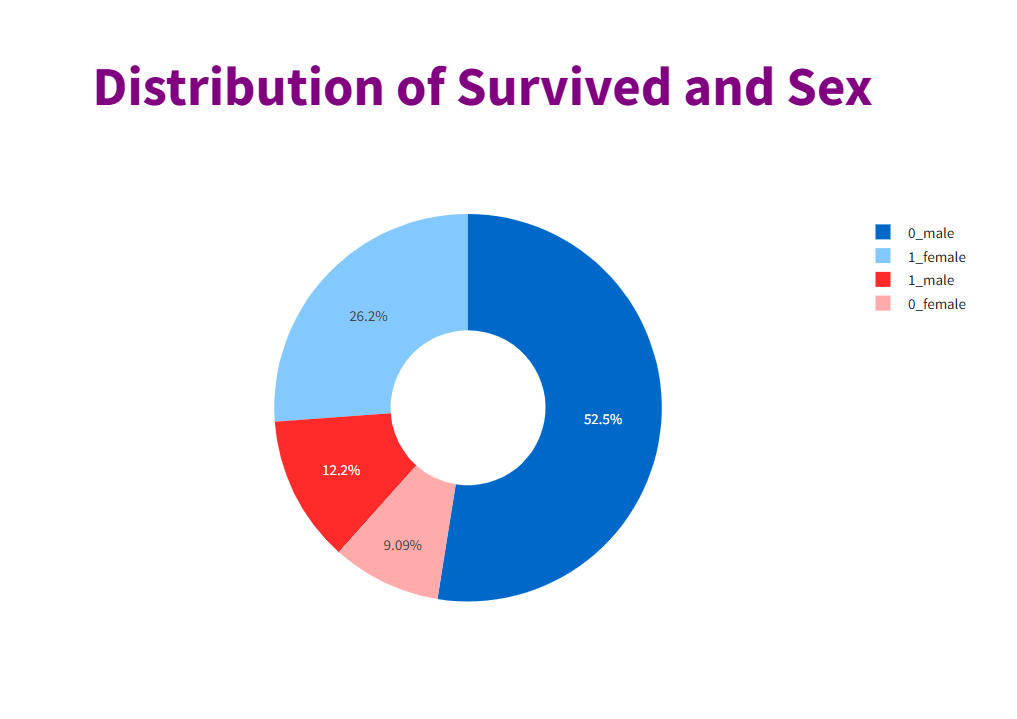
# **Creating MySQL queries and visualizing the results in Streamlit:**

Connecting Python with MySQL to create a database named 'Titanic2' and a table also named 'Titanic2', and then inserting the values from the DataFrame into the MySQL table. Connecting MySQL to Streamlit to visualize the query results.

The home page of Streamlit visualizes the preprocessed dataset.

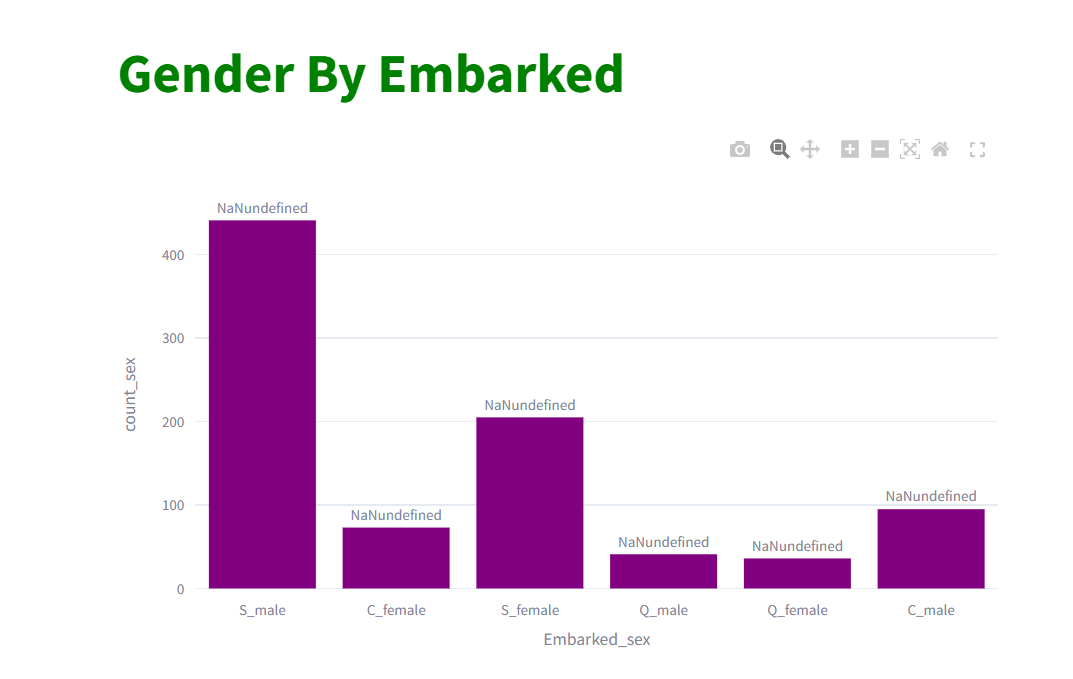
## **First Query :**

Based on gender, we can see how many males and females survived and did not survive. Among them, 52% of males did not survive, while 9% of females did not survive. Additionally, 12% of males survived, and 26% of females survived.



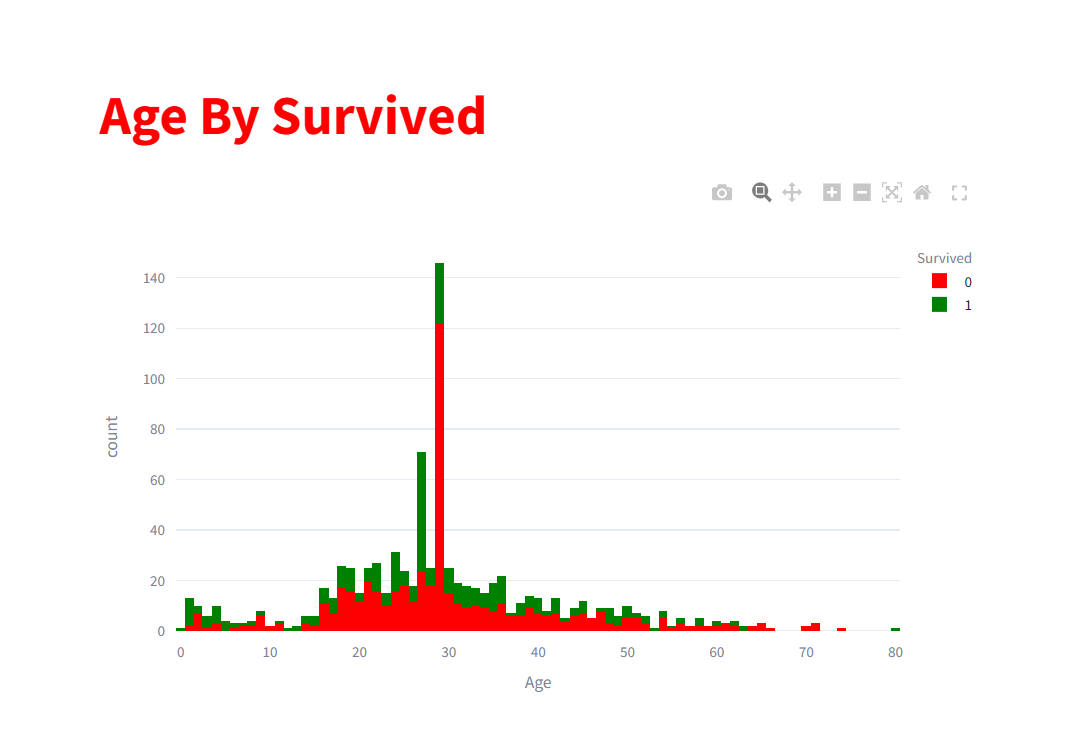
## **Second Query:**

Based on gender and the 'Embarked' category, we can identify which category has more people. In the 'S' category, there are 441 males and 205 females; in the 'C' category, there are 95 males and 73 females; and in the 'Q' category, there are 41 males and 36 females.



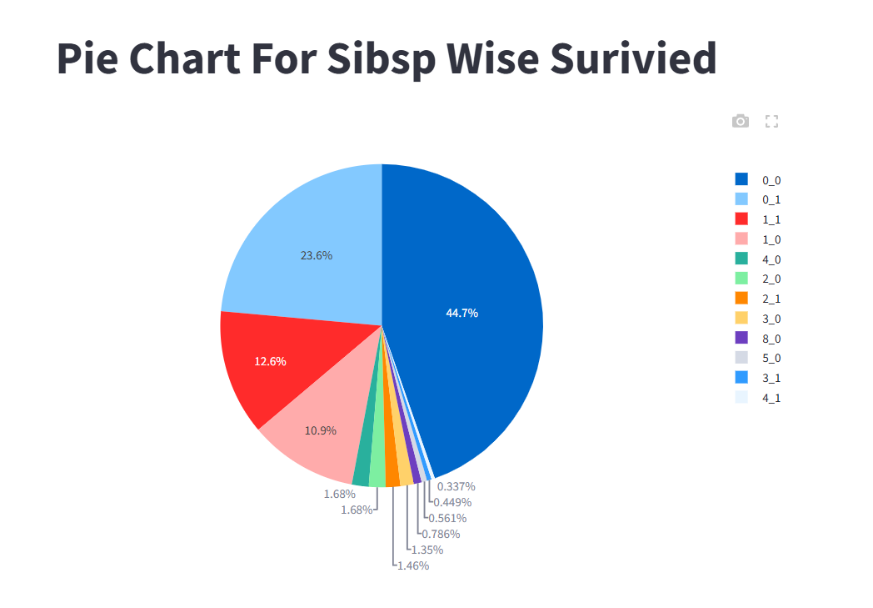
## **Third Query :**

Based on the age and survival columns, we can say that most of the people aged 20 to 35 on the Titanic, there were more individuals who did not survive.



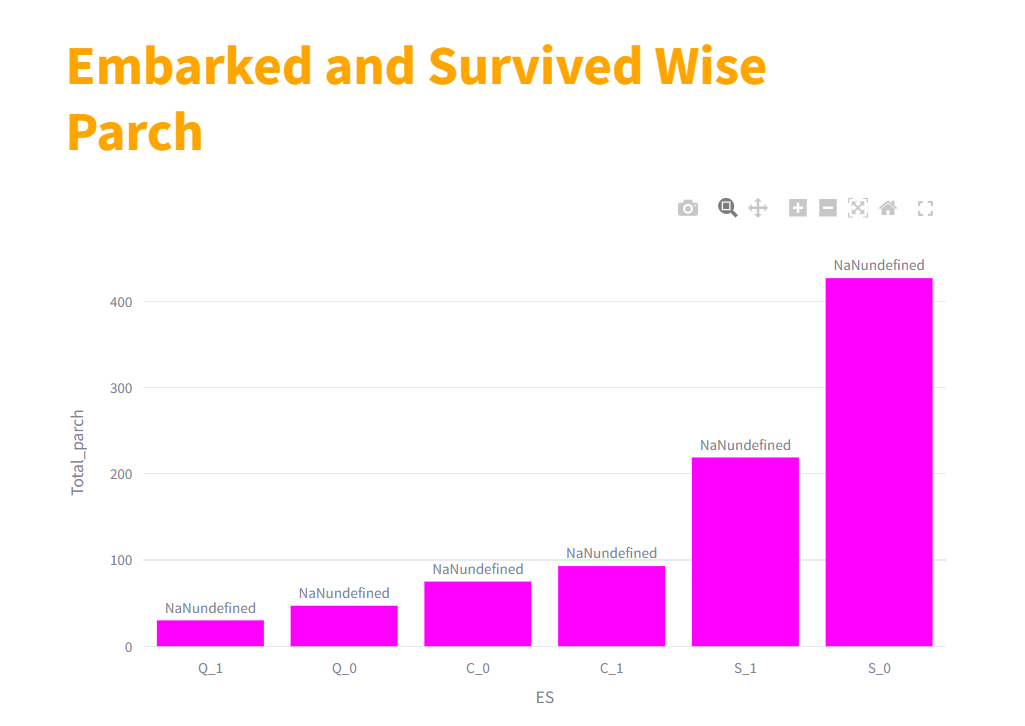
* **Fourth Query:**

In the below pie we can say that non survived and 0 sibsp has 44% and other are less compared to it.



## **Fifth Query:**

Based on the categories of "Parch," "Embarked," and "Survived," there are 427 people in the "S" category who did not survive, while 219 people in the "S" category did survive. The numbers for the other categories are significantly lower.



**Conclusion**:

This analysis of the Titanic dataset involved preprocessing steps such as dropping the 'Cabin' column due to missing data and imputing missing values for 'Age' and 'Embarked.' Categorical variables were transformed using one-hot encoding, and numerical features were standardized, except for outliers in the 'Age' column, which were retained to maintain data integrity. Using MySQL and Streamlit, an interactive querying and visualization system was built to explore the dataset dynamically.