THE LEGEND OF THE TITANIC

DATA VISUALIZATION REPORT

Introduction

The Legend of the Titanic is a data visualization project based on the Titanic dataset which encapsulates a rich repository of information chronicling the passengers aboard the ill-fated RMS Titanic. This dataset is an invaluable resource, offering a glimpse into the demographics, ticket specifics, and ultimately, the survival outcomes of over 890 passengers. This project aims to leverage this dataset for a comprehensive visual analysis, uncovering intricate patterns and insights regarding survival rates based on various influential factors.

Dataset

The dataset comprises a multitude of attributes for each passenger, ranging from socio-demographic information to ticket details and, crucially, survival status. It includes variables like passenger names, ages, genders, ticket classes, fares, cabin allocations, and the pivotal indicator of whether the passenger survived or perished in the tragedy. The dataset contains 891 entries (rows) and 12 variables (columns) as shown in figure 1 below.

```
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)
```

Figure 1 Information about the Titanic Dataset.

Implementation

The implementation was done in python to preprocess, feature engineer and visualize the data. A simple web page was then created, using html, CSS and JavaScript to display the visualizations in a visually appealing manner.

- Preprocessing
  The preprocessing stage involved filling cells with missing values for Cabin and Age variables using e (empty) for the former and the mean value for the latter.
- Feature Engineering
  After preprocessing the data, we sought to look for new relationships between different variables to explore the possibility of creating new variables that might give more understanding of the data. New variables like FamilyCount (number of family members of an individual) which was derived from adding two variables SibSp (Siblings/Spouses) and Parch (Parents/children). Other variables like Title and Survival Rate were also created.
- Visualizations

  Various visualizations (9) were made on the dataset to present a clearer perspective on the influences different aspects of the titanic had on the survival of the individuals that were onboard the ill-fated RMS Titanic. Six of them are shown below.

  Survival Rate by Passenger Class

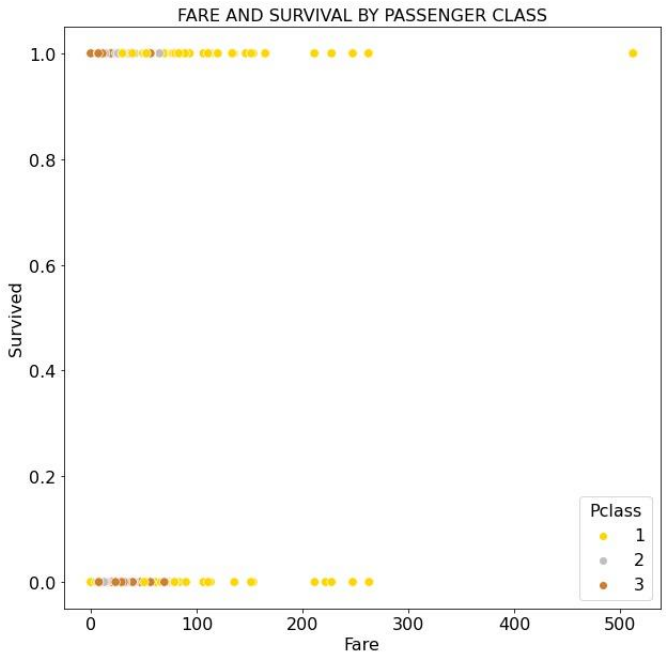

Figure 2 Survival by Fare and Passenger Class

In figure 2, it is evident that people in first (1) class had a higher survival rate than people in second (2) and third (3) class meaning there's a positive correlation between class and survival rate.

Survival by Gender

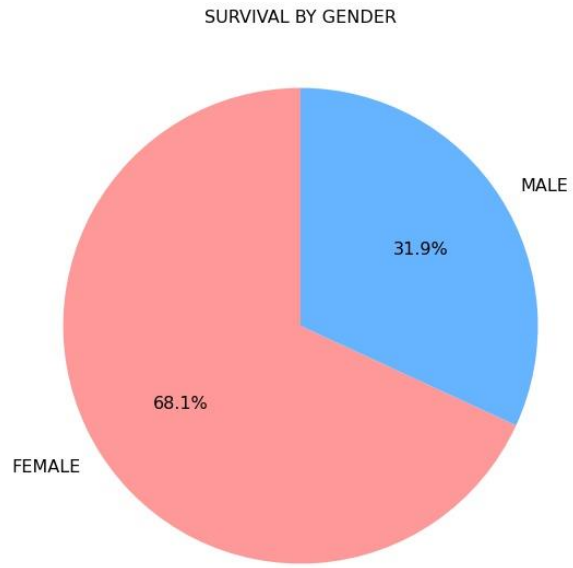

Figure 3 Survival by Gender

Figure 3 above shows that more than half of the survivors were women at 68.1 % compared to men at 31.9%. This could have due to the common moral gesture of putting women and children first.

Age Distribution and Survival

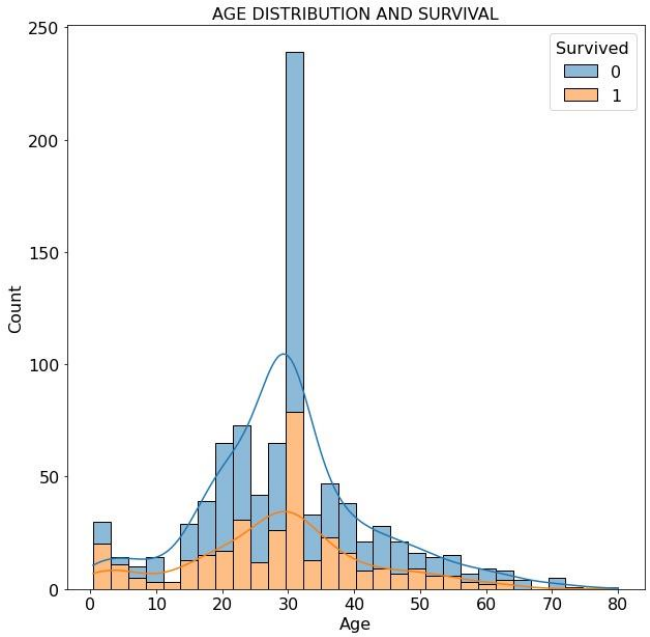

Figure 4 Survival by Age

Figure 4 shows that 30-year-olds had the highest number of people and survivors on the ship. However, people in their 70's and older had the highest survival rate because they weren't many and most of them survived.

Family Size and Survival

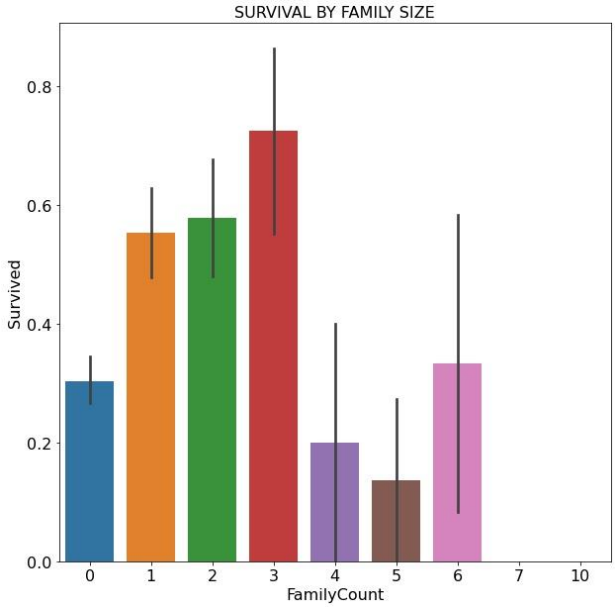

Figure 5 Survival by Family Size

Figure 5 shows that survival rate gradually increases with family size between 0 and 3, but drastically drops for families of sizes above 3. This could be because smaller groups are easier to manage in a "rush hour" than larger groups.

Port of Embarkation

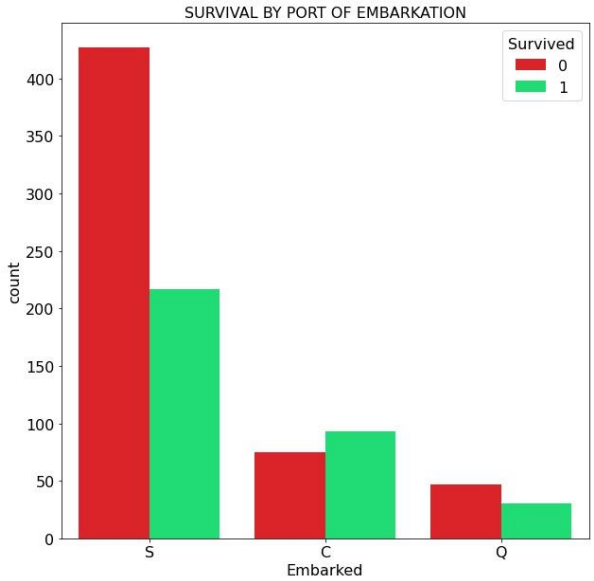

Figure 6 Survival by Port of Embarkation

Figure 6 above shows the relationship between number of survivors and the port of embarkation (boarding point). It is evident that most of the people boarded the ship from port S, but more than half of them did not survive. On the other hand, more than half of the people who boarded from port C survived.

Survival by Title

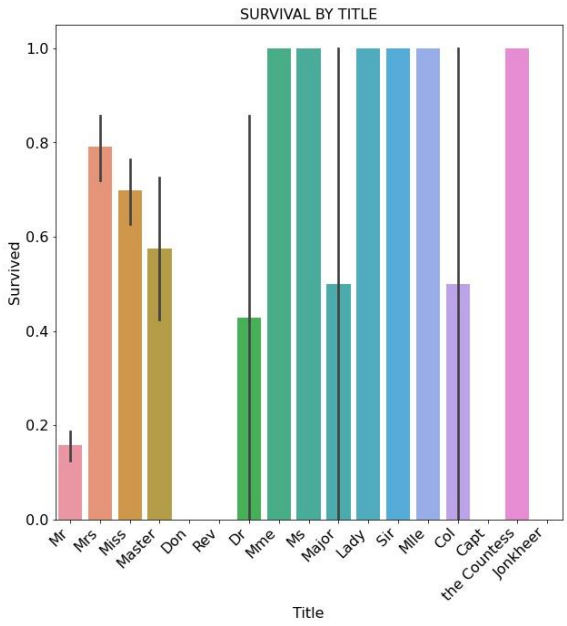

Figure 7 Survival by Title

Figure 7 illustrates the relationship between survival rate of individuals aboard the ship and their titles. From the figure we can tell that, women had a much higher rate of survival than men and that people with high value titles like Col, their survival rate was quite high too.

Conclusion

The Titanic dataset is quite interesting in that it gives us clarity about the ship's tragedy that has been sought after for years, and making visualizations of the data helps us understand even more why some individuals survived and others didn't. The visualizations show that aspects like gender, boarding class, and family size had a high influence on the survival of individuals aboard the RMS Titanic.

Sources

Titanic Dataset: https://www.kaggle.com/datasets/yasserh/titanic-dataset

Project link (GitHub): https://github.com/DAVIDS4A/Data_Visualization