Analyzing Titanic Survival Rates

Machine Learning Course

Amirhossein Abolhassani

Introduction

Exploratory Data Analysis (EDA) involves statistically examining data, cleaning it, and visualizing it, making it one of the most critical steps in a data science project.
In this report, we analyze a dataset containing information about passengers on the Titanic.

1 Dataset Description

The Titanic dataset consists of 11 features and 1309 samples.
Below is the description of each column:

Mapping	Description	Feature
0: Dead, 1: Alive	Survival	*Survived*
1: First Class, 2: Second Class, 3: Third Class	Ticket Type	*Pclass*
0: Female, 1: Male	Gender	*Sex*
	Age in years, expressed as decimals for children under one year	*Age*
	Number of spouses, siblings on board	*SibSp*
	Number of parents and children on board	*ParCh*
	Ticket number	*Ticket*
	Passenger fare	*Fare*
	Cabin number	*Cabin*
C = Cherbourg, Q = Queenstown, S = Southampton	Port of embarkation	*Embarked*

Table 1: Description of dataset features and their meanings

2 Data Collection and Cleaning

Since the target column is separated from the *test.csv* file and placed in $gender_submission.csv$, we first perform an inner join to combine these two dataframes. The resulting dataframe is then concatenated with $train.csvtoconsolidatealldataintoo$

Next, the features Cabin, Ticket, and Name are removed from the dataset, as they are not directly related to a passenger's survival or dea

2.1 Missing Values

As shown in Figures 1 and 2, the columns *Age*, *Fare*, *Embarked*, and *Cabin* contain missing values.

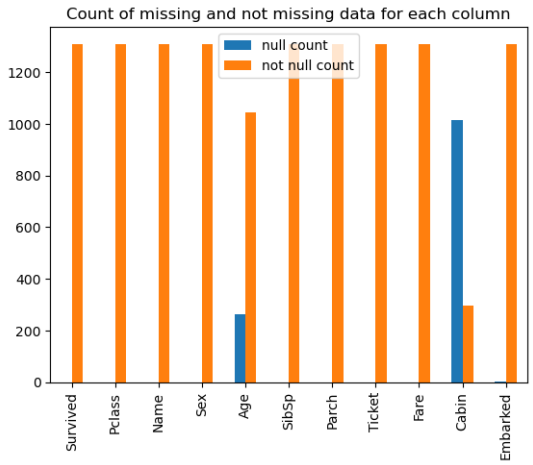

Figure 1: Frequency of missing versus non-missing data for each feature

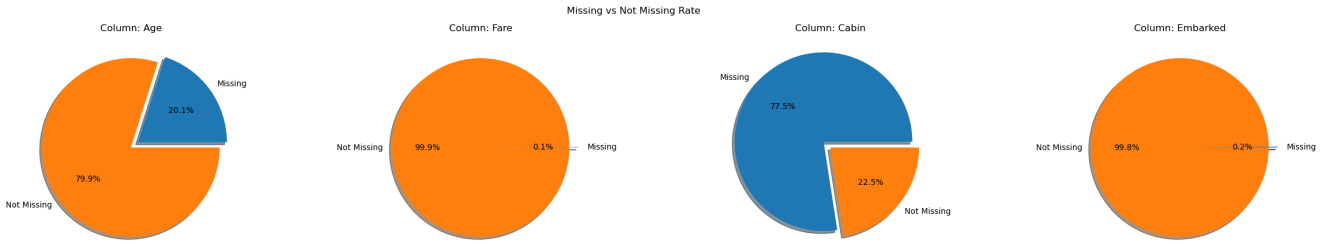

Figure 2: Ratio of missing to non-missing data for columns with missing values

The methods used to handle missing values for each column are as follows:

- *Cabin*: This column was removed in the previous step.
- *Fare*: Since the most influential factors for fare are ticket type and gender, and the number of missing values is small, we group the data by these two features (equivalent to GroupBy in SQL) and calculate the mean for each unique group. Then, for each sample with a missing value in the *Fare* column, we assign the corresponding mean based on the combination of *Pclass* and *Sex*.
- *Embarked*: To fill missing values in the *Embarked* column, the dataset is grouped by *Pclass*, and the frequency of each port is calculated for each group. For each sample with a missing value in this column, the port is assigned based on the most common port for passengers with the same *Pclass*.
- *Age*: For this column, which has the highest number of missing values, the Iterative Imputation method is used. This method treats the *Age* feature as a function of other features in the dataset and estimates missing values using regression to establish relationships between them.

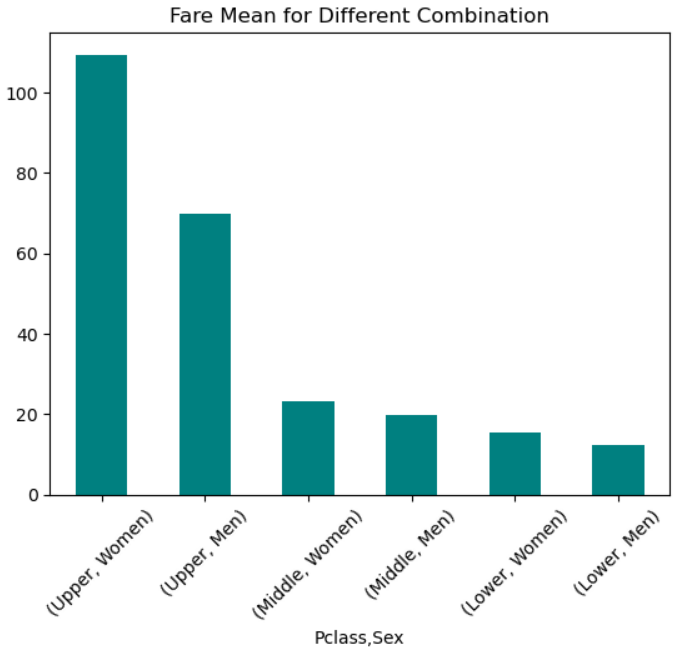

Figure 3: Mean value of the *Fare* feature for unique combinations of *Pclass* and *Sex*

3 Statistical Analysis and Visualization

3.1 Population

Here, we observe the distribution of the population across key groups.

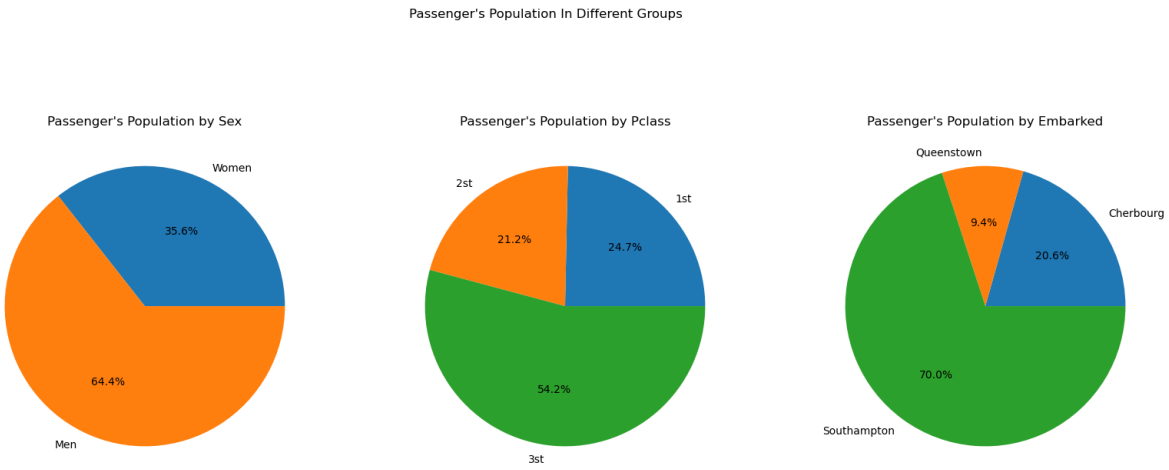

Figure 4: Distribution of passengers based on the *Sex*, *Pclass*, and *Embarked* features

3.2 Survival

As evident from Figure 5, this section examines the frequency of survivors and non-survivors across groups defined by the *Sex* and *Pclass* features, as well as overall.
Since the port of embarkation is assumed to have no impact on survival, grouping by this column is omitted.

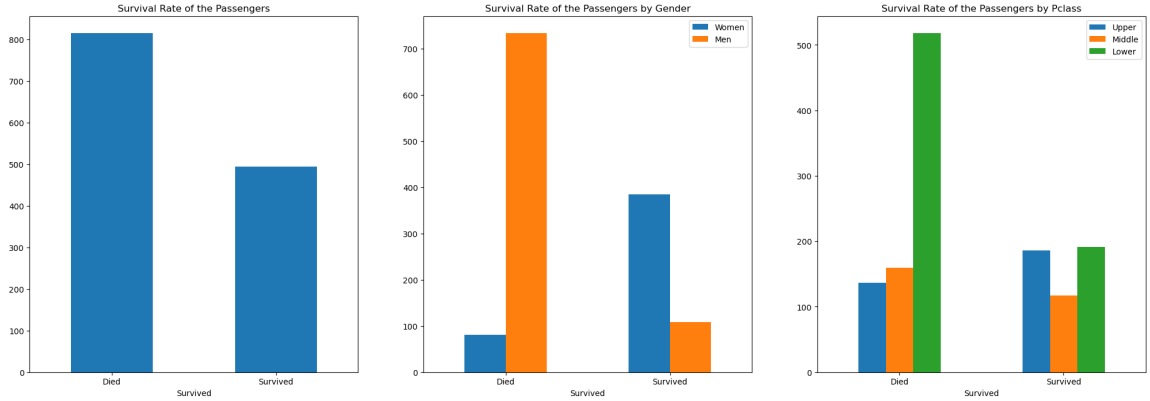

Figure 5: Distribution of survivors and non-survivors based on *Sex*, *Pclass*, and across all samples

plin As shown in Figure 6, the age distribution is analyzed for survivors and non-survivors to better understand trends related to age and survival status.

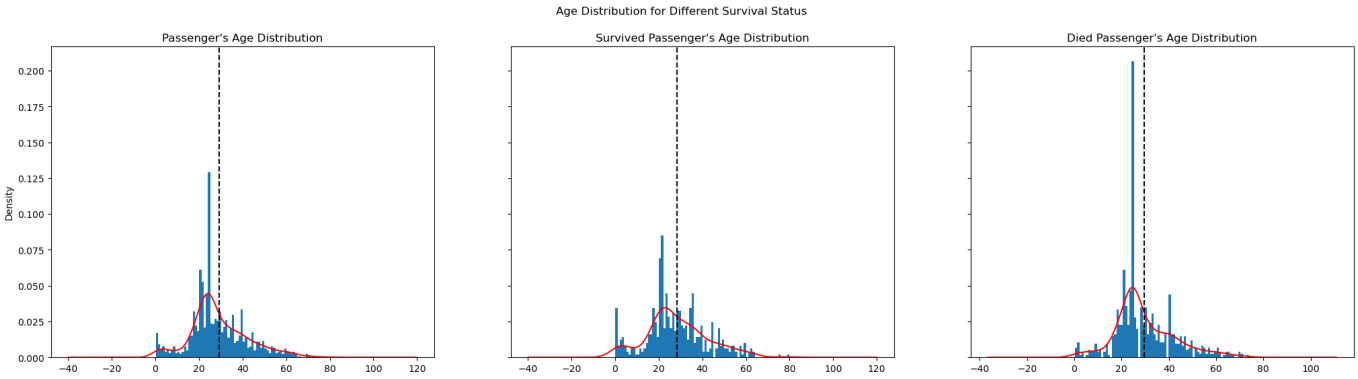

Figure 6: Age distribution across groups in the *Survived* column

The *SibSp* and *ParCh* features can be used to determine the family size of each passenger and explore its relationship with survival. For this, these two columns are summed to create a new column called *Family Population*.

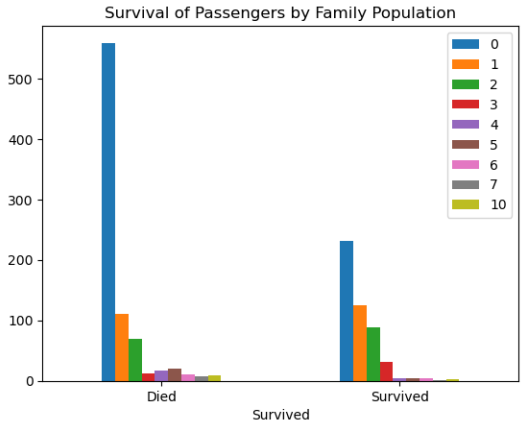

Figure 7: Frequency of family size in groups of the *Survived* column

3.3 Age

The spread of data for the age feature, visualized using box plots across all passengers and grouped by *Survived* and *Pclass*, provides better insight into the data (Figure 8).

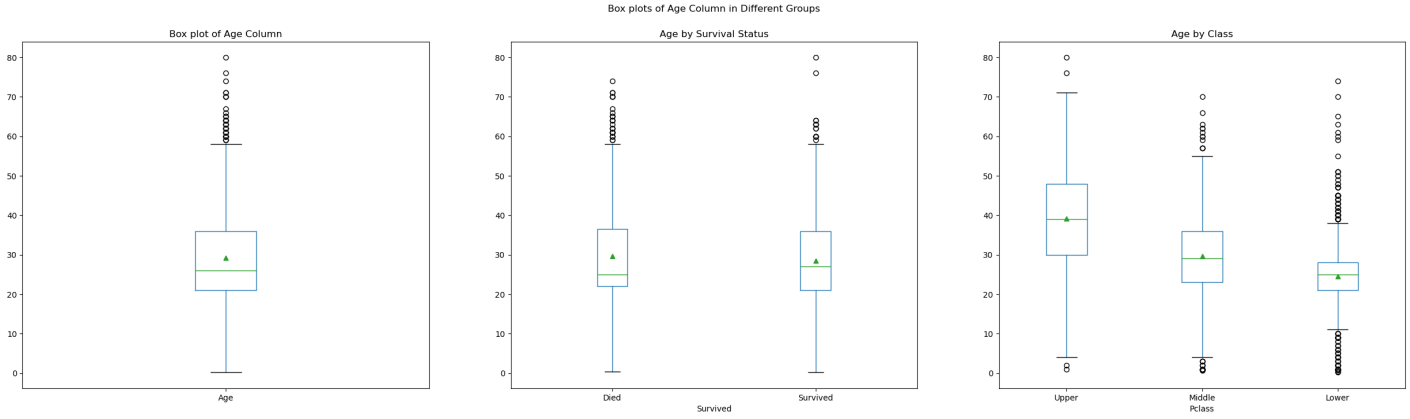

Figure 8: Box plots for the age feature based on values of *Survived* and *Pclass*

Conclusion

Based on Section 3, we can validate hypotheses about relationships between dataset features and survival.

From Figure 5, it is clear that over half of the passengers (approximately 800) lost their lives, with the majority being male. This is reasonable given the higher number of male passengers compared to females (Figure 4).

According to Figure 7, the size of a passenger's family on board has an almost inverse relationship with their survival, which is one of the most intriguing findings from this analysis.

Another aspect examined is the age distribution among survivors and non-survivors. Based on Figure 6, comparing the 0–2 year age range shows that a significant number of children, especially infants under one year, were saved. The distribution of other age groups is roughly similar in both categories, with a notable peak at age 25 for non-survivors, which is reasonable given the high population of this age group on the ship (Figure 8).

The age spread, visualized through box plots in Figure 8, helps identify the mean and median across groups. It can be inferred that individuals over 60 years old are more prevalent among non-survivors, as also seen in Figure 6. Additionally, for ticket classes, passengers in the Upper (1st) class were mostly in their late 30s, those in the middle class were in their late 20s, and those in the lower class were in their mid-20s. Another observation is that most children up to age 15 belonged to the upper and middle classes, making data for the lower class appear noisy. Similarly, individuals over 40 were predominantly from the upper and middle classes. These box plots reveal interesting relationships between age, travel class, and survival.

References

1. "Exploratory Data Analysis" Wikipedia.org
2. "Titanic - Machine Learning from Disaster", Data Dictionary, Kaggle.com
3. "Handling missing values in dataset — 9 methods that you need to know", medium.com