Report Final Project Titanic

Maxence Raveau

391748: Exploratory Data Analysis and Visualization

Stefan Lin

UCLA Extension

06/15/2023

Contents

[Summary Tables 3](#_Toc137764345)

[Pivot table 5](#_Toc137764346)

[Missing Values 5](#_Toc137764347)

[Questions 5](#_Toc137764348)

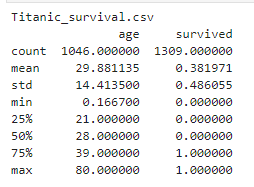
[Graphs 6](#_Toc137764349)

[Logistic Regression 8](#_Toc137764350)

[Kmeans Clustering 8](#_Toc137764351)

# Summary Tables

Titanic\_survival



Titanic\_info

A screenshot of a computer

Description automatically generated with medium confidence

Merged\_df

A screenshot of a computer

Description automatically generated with medium confidence

# Pivot table

A picture containing text, font, screenshot, algebra

Description automatically generated

# Missing Values

We have a lot of missing values, especially in cabin, boat, body, home.dest and cabin level, with respecitvely 1015,824,1189,565 and 1015 missing values.

For columns with not a lot of missing values, i will just remove the rows with N/A in those columns.

For the ones with a lot of missing values, i'm gonna delete them.

I dropped home.dest, boat, body, name, ticket and embarked as they are not impacting the survival rate. I also removed the cabins columns as they have too many missing values. I keep Cabin level, only for visualization, and I will drop it later.

Doing all that, we end up with a dataset with 1043 rows and 8 columns.

# Questions

The survival rate of passenger was 40.7%.

The survival rate for women is 75.13%, while it’s 20.5% for males. Women had more than 3 times more chance to survive compared to men’s.

There are significant differences in survival rates across different passages, going from 63% for P1 to 44% for P2 and 26% for P3.

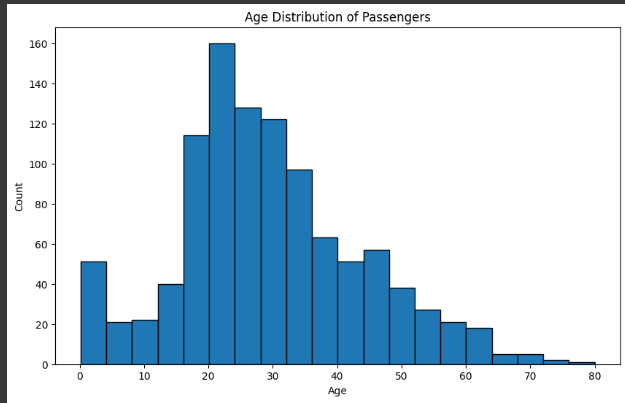
Passengers’ age had an impact on their chances of survival.

Fare and pclass are the highest correlated feature, with -0.56. We also have age and p class that kinda correlated, with -0.41.

A screenshot of a computer screen

Description automatically generated with low confidence

# Graphs



We can see that the titanic was mostly populated with people from 18 to 30 years old.

A picture containing screenshot, text, line

Description automatically generated

Using this, we can visualize that it's not because people were aged that they paid more their ticket(better cabin or better services)

A picture containing text, screenshot, display, number

Description automatically generated

We can see that the most used cabin level was the C, going after with the B and then the D

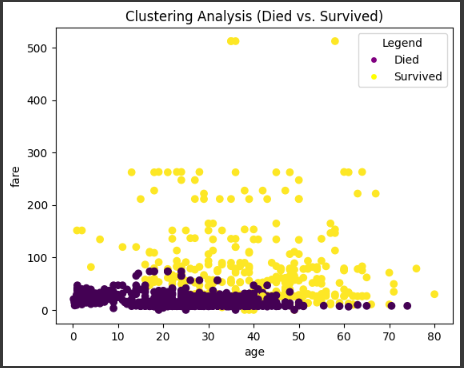
# Logistic Regression

A picture containing text, screenshot, font

Description automatically generated

While running the model, i tried removing sibsp from the features, and it ended up with a higher results, that'why i removed it in the code above. The logistic regression model achieved an accuracy of 78.46%, indicating that it correctly predicted the survival outcome for approximately 78.46% of the test instances. The recall score of 0.63 suggests that the model successfully identified 63% of the actual positive (survived) instances. The F1 score, which combines precision and recall, is 0.69, indicating a reasonable balance between the two metrics. Overall, the logistic regression model performs moderately well in predicting survival outcomes in the Titanic dataset.

# Kmeans Clustering



I used the same variable as for the logistic regression. We can see on this graphic that people who paid a higher fare had more chance to survive.

We achieve an accuracy of 65% on this model, which is worst than for logistic regression. Inertia is equal to 3823.

# Conclusion

Survival Distribution:

The majority of passengers onboard the Titanic did not survive the disaster, with approximately 62% of the passengers in the dataset labeled as "not survived." This indicates that the survival rate was relatively low.

On the other hand, around 38% of the passengers in the dataset managed to survive the tragedy, which accounts for a smaller proportion of the total passengers.

Age Distribution:

The age distribution of the passengers shows a peak in the young adult range, with a higher concentration of passengers between 20 and 40 years old. This suggests that a significant number of passengers onboard the Titanic belonged to this age group.

Furthermore, there were a considerable number of children (below 10 years old) among the passengers, indicating that families with young children were also present on the ship.

Gender Distribution:

The dataset contains a higher number of male passengers compared to female passengers. This implies that there were more men onboard the Titanic than women.

Interestingly, the survival rate among female passengers was higher compared to male passengers. This observation suggests that there may have been a bias in favor of females when it came to survival during the disaster.

Based on the scatter plot, it can be observed that the passengers who paid higher fares had a higher chance of survival. This suggests that those who paid more for their tickets may have been given priority in terms of safety measures or access to lifeboats. We also saw that female and children had higher chance of surviving, I guess because of the “women and children first” rule. Male might have sacrificed themselves so they could leave.