Titanic Data Report

# GitHub URL

https://github.com/CMcN98/UCDPA\_CliodhnaMcNamara

# Abstract

This project is exploring Titanic data and making comparisons based on ticket class, age and sex. The survival rate based on these aspects was analysed. The Titanic event offers a varied and well rounded dataset, making it ideal for this exploratory data analysis. Following the cleaning of the data the methods of analysis used in this project include .groupby function, column manipulation, .sum and np.where function. The results of my analysis are presented in various formats including bar charts and plots.

# Introduction

I chose the Titanic as it is really interesting how much high quality data is available for a renowned historic event that occurred over a hundred years ago and shaped the ship building industry from then on. There is a good spread of data that has been gathered regarding this event. A certain level of data quality is guaranteed with the Titanic as it is a much-studied tragedy. The integrity of the data allows for a more accurate analysis, and this was a big draw. The data available also acted as a social commentary for the period. It was revealing that there were stark differences between ticket classes, age range and sex. Analysing these events of the past gives us a better understanding of a time gone by.

# Dataset

Kaggle is the data source for my dataset. I chose Kaggle as the source as it has a large number of interesting datasets of different shapes and sizes. The datasets it has to offer are opensource data so anybody can take them and use them. Kaggle has high quality datasets to pick from and I trust the soundness of the datasets.

Survived – Integer, 0 = No, 1 = Yes, Whether the passenger survived or not.

Pclass – Integer, 1 = 1st, 2 = 2nd, 3 = 3rd, The ticket class of the passenger.

Name – object, The name of the passenger.

Sex – object, The gender of the passenger.

Age – float, The age of the individual.

SibSp – Integer, The number of siblings and spouses on board.

Parch – Integer, The number of parents and children on board.

Ticket – Object, The ticket number.

Fare – Float, The price of the ticket.

Cabin – Object, The cabin number.

Embarked – Object, C = Cherbourg, Q = Queenstown, S = Southampton, the port from where the passenger embarked.

# Implementation Process

1. I imported all of the necessary libraries that I would need to do my analysis.
2. I imported the dataset naming it Titanic and then created a data frame from the dataset.
3. Showed the first 10 columns of the dataset to ensure the data was imported correctly and displaying in a data frame.
4. Showed the last 10 rows of the dataset in order to check the whole dataset was imported.
5. Checked the size of the dataset to see how many rows in total and how many columns I was going to be working with.
6. Got a view of the different types of data in the dataset.
7. Used the describe() function to generate the descriptive statistics of the dataset.
8. Dropped irrelevant columns from the dataset to clean the data.
9. Renamed some of the columns in the data set to be clearer on what they were representing.
10. Named the index column in the data frame as part of the cleaning process.
11. Added two of the columns together and created one new column populated with the sum of the two merged columns. I wouldn’t be using the columns separately in the analysis, so it made for a cleaner data frame and made it easier to analyse the family size by the one column.
12. Called the top 5 rows of the data frame to check the changes had been implemented correctly i.e., the new column was created and populated with the sum of the two existing columns.
13. Removed the two columns I previously added together as the value for them was captured in the new column, so they were irrelevant in the dataset.
14. Called the top 5 rows of the data frame to check the changes had been implemented correctly.
15. Checked the dataset for any duplicates, in this dataset there are zero but if there were some, I would have removed the duplicates to ensure the integrity of the analysis.
16. Checked for missing/null values – there are a low number of missing values and if I were to remove the rows with missing/null values it could impact the integrity of the dataset.
17. Checked for outliers in the age column.
18. Used describe() function to generate the descriptive statistics of the Age data.
19. Displayed the passengers ages on a histogram to clearly show the spread of ages and the most common age in the data.
20. Created a new column and sorted the ages into three age groups to enable me to some further analysis on the ages of survival.
21. Called the top 5 rows of the data frame to check the changes had been implemented correctly.
22. Checked how many passengers there were in each age range to see the distribution in each bracket.
23. Showed the survival rate based on the age bracket, I have displayed it in a plot to make the data clearer to read and make it more impactful.
24. Called to see how many passengers there were in each class type.
25. I then showed this in a bar chart.
26. Used describe() function to generate the descriptive statistics of all of the data.
27. Calculated the percentage of survived people.
28. Looked for the survival percentage based on the ticket class.
29. I then demonstrated the results of this in a chart.
30. Looked for the survival percentage based on gender.
31. Calculated the percentage based on family size.
32. Ran .groupby function to help show the number survived based on Ticket class and Gender.
33. Calculated the percentage based on place of embarkment.
34. Checked how many people embarked from each port.
35. Called the top 20 rows of the data frame to check the changes previously implemented were still correct.

# Results

1. The max age of passengers was 80 and the mean is 29, although the mean is not extremely small compared to the max age it does show that the max age is an outlier in the overall age bracket of passengers.

Chart

Description automatically generated

Chart, histogram

Description automatically generated

1. There was nearly three times as many people in the 20 – 49 range than in the 0 – 19 range and just under double than in the 20 – 49 range than in the 50 – 80 range. The below shows a higher number died in the 20 – 49 range than the other brackets and an equal amount survived in the 0 – 19 and 50 – 80 ages.

Chart, bar chart

Description automatically generated

# Simple bar chart to display the number of passengers in each class.

# Chart, bar chart Description automatically generated

# Plot shows 3rd class had the highest death rate and 1st class had the highest survival rate.

# Chart, bar chart Description automatically generated

# Insights

1. The age range of 20 – 49 years had the highest death rate and the highest survival rate.
2. 50 – 80 years had the second highest death rate.
3. 0 – 19 years looks to have equal number of deaths as survivals in the range.
4. 50 – 80 years and 0 – 19 years have the same number of survivals.
5. The higher the class the higher the survival rate of a passenger. This would have been due to getting priority for the lifeboats but also being closer to the deck of the boat. The lower the class the lower down in the boat the passengers’ rooms were. When the ship started to sink it would have been the lower class that would have been hit first.
6. There was a higher survival rate in females over males. Women and children would have been prioritised over men for getting onto the lifeboats.
7. In all the different family sizes, the family size of 3 had a higher survival rate.
8. Over 50% of survivors embarked from Cherbourg. Nearly four times the amount of people that embarked from Cherbourg, embarked from Southampton yet there was a higher survival rate from Cherbourg passengers.

Whilst class structures aren’t as prevalent in western cultures anymore if using machine learning we could perform a prediction that there would be less of a gap between 1st class and 3rd class rate of survival. In recent years gender roles have become less subscribed to and I would estimate survival rates would differ more significantly than they did in the period of the Titanic. Although I did not analyse changes in the ship building industry and safety measures since the sinking of the Titanic it would provide an interesting comparison. A study of launch times between traditional wooden lifeboats versus new highly engineered lifeboats and their resultant improved survival rates would provide an interesting juxtaposition.

# References

Kaggle.com. 2022. *Titanic - Machine Learning from Disaster | Kaggle*. [online] Available at: <https://www.kaggle.com/competitions/titanic/data> [Accessed 25 August 2022].

Kaggle.com. 2022. *Titanic Survival Rate*. [online] Available at: <https://www.kaggle.com/code/prakharrathi25/titanic-survival-rate/notebook> [Accessed 25 August 2022].

Kaggle.com. 2022. *Titanic Tutorial*. [online] Available at: <https://www.kaggle.com/code/alexisbcook/titanic-tutorial> [Accessed 25 August 2022].