CentraleSupélec

Data Visualisation: T2 - 2023

Data Visualisation Assignment 1: “Sinking of the RMS Titanic”

Team Xiaolongbao

Team members:

Hongyang YE

Proud CHAREESRI

Ru YI

Wenjing ZHAO

Content page

1. Data Description

1.1. Topic and Significance

The dataset of this project is obtained from a classic machine learning disaster prediction problem - The sinking of the RMS Titanic in 1912. This tragic event resulted in the loss of 1,502 passengers and crew out of the total 2,224 onboard, marking it as the deadliest peacetime maritime disaster in history.

In addition to its historical significance, the dataset comprises information about individuals from various socio-economic backgrounds, age, gender, etc, to allow for a wide range of analysis between different variables on survival rate. It also offers further insights into the criterias that were considered significant at that time for classifying individuals' lives as more or less valuable, in the opinion of those onboard the vessel.

1.2. Name and Description of data

The dataset "Titanic - Machine Learning from Disaster" was obtained from Kaggle and contains information of 891 passengers with the following variables: Passengerid, Survival, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked. The goal of this project is to visualise the relationship between predictor variables on survival.

1.2.1. Data collection

The principal source for data on passengers of the Titanic is the Encyclopedia Titanica. One of the most original sources of Titanic datasets used to predict the effect of different variables on survival was compiled by Eaton & Haas (1994) "Titanic: Triumph and Tragedy" and published by Patrick Stephens Ltd. This compilation included a passenger list created by multiple researchers and edited by Michael A. Findlay.

1.2.2. Variables

Variable	**Type**	**Definition**	**Keys**	**Comments**
Survival	Categorical (binary)	Survival	0 = No; 1 = Yes	-

Pclass	Categorical (ordinal)	Ticket class	1 = 1st;
			2 = 2nd;
			3 = 3rd
				Proxy for socioeconomic status

Sex	Categorical	Sex	-	-
Age	Numerical	Age in years	-	Age is fractional if less than and estimated age is in the form of xx.5

SibSp	Numerical	Number of siblings or spouses onboard	-	Sibling = brother/sister, stepbrother/
				stepsister;

				Spouse = husband/wife

Parch	Numerical	Number of parents or children onboard	-	Parent = mother/father;
				Child = daughter/son, stepdaughter/stepson

Ticket	Categorical	Ticket identification number	-	-
Fare	Numerical	Ticket fare	-	-
Cabin	Categorical	Cabin number	-	-

Embarked	Categorical	Port of embarkation	C= Cherbourg;
			Q = Queenstown;
			S = Southampton
				-

1.2.3. Data cleaning

- Age: 177 missing values, drop rows with missing values as using the mean value would skew visualisation
- Cabin: 687 missing values, column dropped from analysis
- Embarked: 2 missing values, input mode value input
- Created Family Members variable from SibSp and Parch

1.3 Future steps

With the cleaned dataset, four main hypotheses were formulated to facilitate the subsequent stage of exploratory analysis, where a visualisation would be created for each of the following hypotheses:

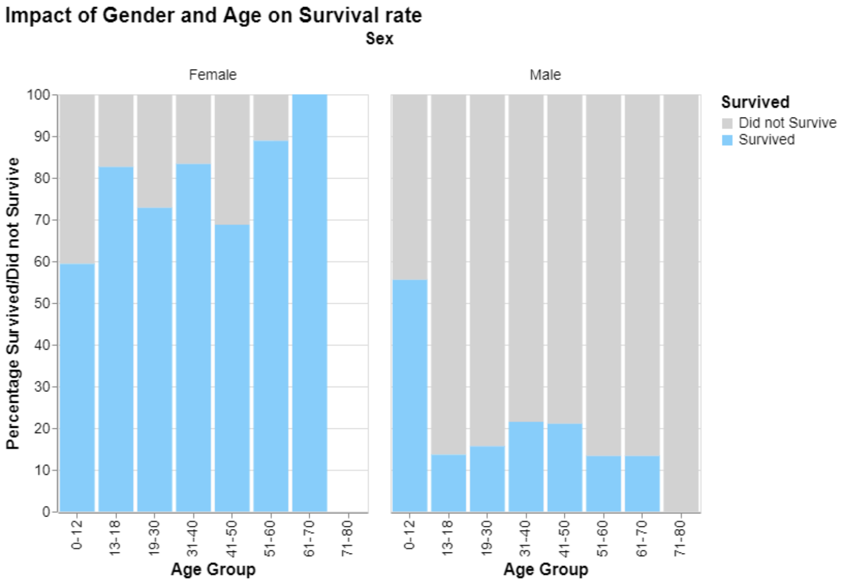
- Impact of gender and age on survival: Larger percentage of women and children survived the shipwreck
- Impact of fare and class on survival: Survival rate increases with fare and class
- Impact of class and age on survival: Survival rate increases with passenger class and decreases with age
- Impact of the presence of siblings/spouse and parents/children on survival: more family members will increase the chance of survival

2. Exploratory Analysis

2.1. Impact of gender and age on survival (Wenjing ZHAO)

The purpose of this visualisation is to examine the effect of gender and age on a passenger’s chance of survival. With reference to the commonly observed protocol of “Women and children first” during disasters, the hypothesis is that there would be a larger percentage of women and children who survived the shipwreck.

2.1.1. Visualisation

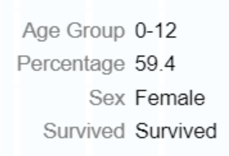


2.1.2. Visualisation description

Passengers are first categorised into 8 distinct age groups, with ‘0-12’ representing young children and ‘13-18’ representing older children. This categorisation allows for analysis of the impact of age on survival rate. Next, the data is normalised to adjust for gender imbalance, enabling fair comparison of relative chances of survival across age and sex. Lastly, the visualisation is segmented by gender to highlight the difference in survival rate with females on the left and males on the right. Survivors are represented in light blue while non-survivors are represented in light grey.

2.1.3. Design rationale

This visualisation is a normalised stack bar chart created using Python’s Altair. The stacked bar chart, normalised data in percentage, and colour scheme provides viewers with a visually intuitive comparison of the proportion of survivors classified by age group and sex.

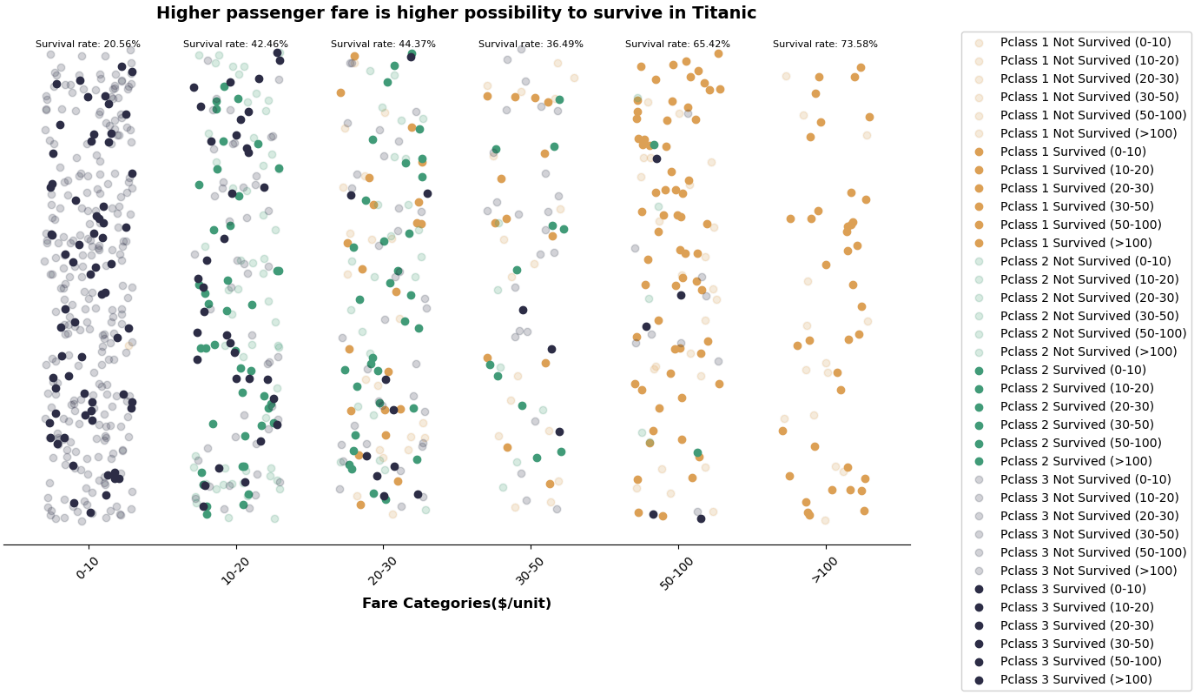


Age Group	0-12
Percentage	59.4
Sex	Female
Survived	Survived

Interactive tooltips shown on the left were also used to provide information on the specific percentage of survivors in each age group. This feature keeps the visualisation clean and allows viewers to focus on the main patterns of survival relative to age and gender.

2.2. Impact of fare and class on survival (Ru YI)

2.2.1. Visualisation



2.2.2. Visualisation description

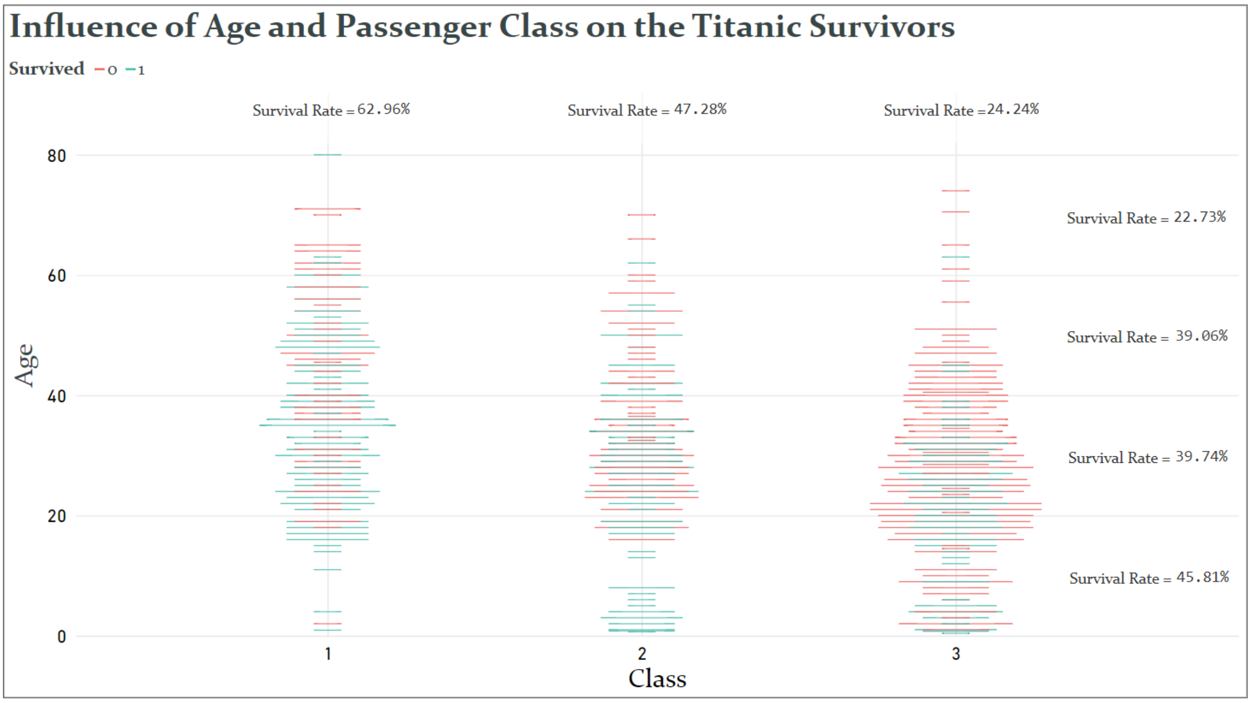
I want to show the relationship between passenger class, fare, and the survival rate in an institutive way. Through the difference of transparency and deep level of colour, people can see different class passengers buying which level of fare, and whether buying higher passenger fare will have higher possibility to survive.

2.2.3. Design rationale

I use matplotlib, seaborn, numpy and pandas in python. Because I do not want to use traditional methods, which need to be analysed by themselves further. (eg. boxplot and violin chart). And python is flexible and functional software to help me achieve it. And each dot in the chart represents each passenger's lives, which show the large number of passengers and the severity of the accident. Each category has the same length and width, it provides subject measure to compare.

2.3. Impact of class and age on survival (Proud CHAREESRI)

2.3.1. Visualisation



2.3.2. Visualisation description

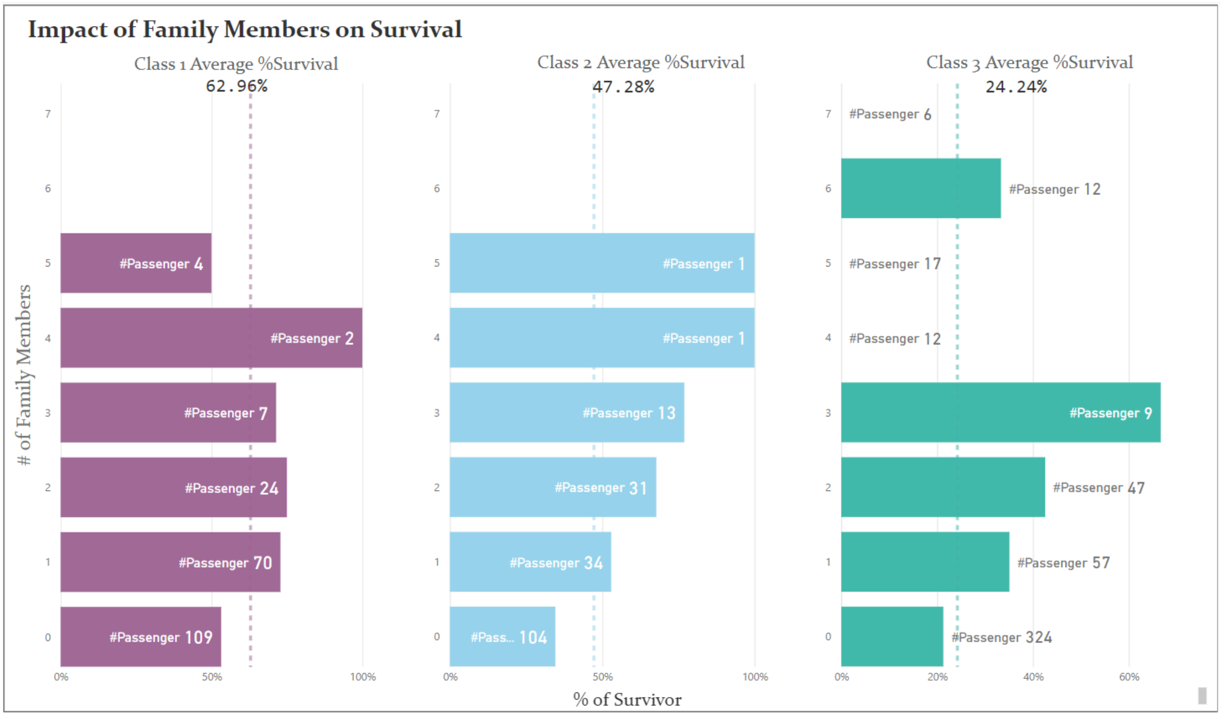
The previous graph shows the effect of fare and class on survival. We can see the correlation of fare and class as well as the effect of fare on survival rate. This graph's purpose is to dive deeper in the impact of class on survival rate. The assumption is that there might be some correlation between age and class as age can impact passenger affordability. If age and class are highly correlated, age might be the more important factor on survival. We should be able to identify if the passenger strength as assumed by age played a more important role in passenger survival than the passenger wealth as assumed by class in this graph.

2.3.3. Design rationale

To see the effect of class and age on survival, a scatter plot was selected. This scatter plot created by PowerBI provides a clear view of how passengers of different classes are distributed across different age ranges as well as the proportion of survivors in different classes and age. The colour green and red were used so that the information can be interpreted quickly and intuitively. Marker style was chosen to be dash instead of default circle dot due to the frequency of age data as well as interpretability of length in comparison to size.

2.4. Impact of the presence of Family Members on survival (Hongyang YE)

2.4.1. Visualisation



2.4.2. Visualisation description

This graph aims to show the impact of family sizes on survival. SibSp and Parch were combined to get # of Family Members and then % of Survivor was calculated for passengers with the same family size. Passengers in different classes are analysed separately to remove the impact of Pclass on survival. We should be able to see if having more family members on Titanic can increase the chances of survival for the passenger.

2.4.3. Design rationale

This bar chart was created using PoweBI. We want to analyse the impact of family size on survival without the impact of the passenger's class. Therefore, different colours were used to differentiate each class. The percentage of survivors was used instead of the count of survivors since the number of passengers in each class varied and using count will make it difficult to interpret the graph.

3. Findings

3.1. Impact of gender and age on survival

From the first visual, females across all age groups exhibited higher survival rates than males within corresponding age groups. The survival rate for females ranged from 59% to 100%, with those aged 0-12 having the lowest survival rate. Compared to males in other age groups whose survival rates ranged from 0% to 22%, male children aged 0-12 had a noticeably higher survival rate of 56%. In line with the initial hypothesis, the visualisation successfully demonstrated that the survival rate of women and children were much higher than those of adult men.

3.2. Impact of fare and class on survival

Passengers with fares between £0 to £10 are all in third class while those with fare above £100 are exclusively in first class. The visualisation indicates that passengers with higher fare prices have a higher survival rate. However, from the visualisation of mixed fare categories between £20 to £50, it is noticeable that a higher percentage of first class passengers survived as compared to second or third class passengers, despite paying the same fare. These observations suggest that perhaps the fare paid does not affect survival rate as much as socioeconomic status does.

3.3. Impact of class and age on survival

From this plot, first class passengers are more likely to survive than any other class across all ages. The rate of survival for passengers in first class recorded at 63%, in second class at 47%, and in third class at 24% exhibiting decreasing survival rate for passengers in lower class. Moreover, the survival rate of passengers of different ages exhibit an inverse relationship with the survival rate. However, it can be observed that socioeconomic status plays a more important role than age as for passengers aged 40-60, almost all surviving passengers are in first class and almost none third class survived.

3.4 Impact of the presence of siblings/spouse and parents/children on survival

The graph suggests an overall positive correlation between the number of family members and the rate of survival for all passenger classes. The majority of passengers did not have any family member travelling with them and the survival rate for these passengers is significantly lower than the class average. This indicates that family size plays an important role in passenger survival. However, there are exceptions to this conclusion in the third class passengers. There are no survivors for third class passengers with 4, 5, and 7 family members.

References

ML Frank. (n.d.). Titanic Datasets. Lakeforest. https://campus.lakeforest.edu/frank/FILES/MLFfiles/Bio150/Titanic/TitanicMETA.pdf

Will Cukierski. (2012). Titanic - Machine Learning from Disaster. Kaggle. https://kaggle.com/competitions/titanic