Surviving the Titanic: Multivariate Analysis

Lukas Olenborg
605942

14.4.2023

Contents

1 Introduction

In 1912, the infamous Titanic ship with around 2200 passengers sank on its maiden voyage, losing over 1500 lives. A well-known dataset, "Titanic: Machine Learning from Disaster" (2012), has been extensively used in recent years in research surrounding the Titanic and its passengers. In this report, I use an improved and updated dataset to perform bi- and multivariate analysis on this topic. By analyzing factors that may have contributed to the survival of the passengers, this report aims to gain insight into the circumstances surrounding the disaster and to identify any patterns in the data that may help to explain the outcome of the tragedy.

The dataset is introduced and preliminary preprocessing discussed in Section 2. The data is further processed and fine-tuned in Section 3. Additionally, the significant variables are carefully selected and bivariate analysis is carried out in Sections 3.1 and 3.2. In Section 4, multivariate analysis is implemented using the method MCA (Multiple Correspondence Analysis). The report is concluded in Section 5.

2 Dataset

I am using a full dataset (2019) combining the training and test datasets prepared for a machine learning setting. This dataset is an improved version of the original dataset mentioned in the *Introduction*. It has fewer missing values and some new variables, namely duplicates of the original variables with improved data, added from Wikipedia's datasets.

The dataset has a total of 21 variables and 1310 data points before preprocessing. Two ID and two name variables are removed as they are simply identifications and are irrelevant in terms of survival rate. Similarly, a variable with the ticket number is removed. There remain 7 duplicate variables with data from Kaggle and data from Wikipedia. I keep the variable with fewer missing values and remove the other duplicate. Finally, 3 more variables are removed. Firstly, a variable representing the cabin which contains many missing values. I make an assumption that the most important information regarding the cabin is its location, which is approximately represented by another variable, namely the ticket class of a passenger, as the cabins are located based on ticket classes. Secondly, there is a variable labelling from which lifeboat a passenger was saved. As the values are missing for all the deceased passengers and this directly tells us whether a passenger survived or not, I also remove it. Finally, there is a **Body** variable from Wikipedia with mostly missing values, and I am not sure what it means, so I remove it.

3 Variables

survived	sex	age	class	sibsp	parch	fare	homecountry	embarked	destination
0	M	teen	3	1	0	0-20	UK	S	Canada
1	F	adult	1	1	0	20-100	US	C	US
1	F	adult	3	0	0	0-20	Europe	S	New York City
1	F	adult	1	1	0	20-100	US	S	US
0	M	adult	3	0	0	0-20	UK	S	New York City
0	M	teen	3	0	0	0-20	UK	Q	New York City

Table 1: Preprocessed data.

The preprocessed data has a total of 10 variables and 885 data points. The first six instances can be seen in Table 1 above. I have implemented further preprocessing and fine-tuning, several variables have been binned and/or discretized. As most variables have 2-5 categories I have transformed the other variables by removing or grouping modalities.

The modalities of the continuous variable **age** were originally integers between $[0,74]$ but they have been grouped in the following labeled bins: *toddler* $\in [0,5)$, *child* $\in [5,15)$, *teen* $\in [15,25)$, *adult* $\in [25,50)$, *senior* $\in [50,\infty)$. The choice of these groups was made based on *domain knowledge*, i.e., my estimation of what are meaningful age groups in terms of how they behaved during the tragedy. A similar binning was implemented for the ticket price variable **fare**.

For many other categorical variables, I have joined categories together to get a smaller number of modalities with enough data points in each modality. For example, the variable **homecountry** had originally the home town or address of each passenger. I removed the specificity of a town but still had 30+ modalities for different countries. Finally, I grouped e.g. most of the European countries together with the final modalities being *Europe*, *UK*, *US*, *Africa* and *Other*.

It is questionable whether the variables **homecountry**, **embarked** and **destination** have any correlation with the dependent variable, but I will verify this with statistical tests. Similarly, the variable **fare** might simply be another representation of the variable **class**, as I assume that first-class tickets are the most expensive and third-class tickets the least expensive and so on. It is interesting to see, which variables are used for the final analysis.

The names, modalities and descriptions of the variables are shown in Table 2 below. In Figure A1, some example variables and their distributions are shown.

Variable	Modalities	Description
survived	1/0	Whether passenger survived or not
sex	M/F	Sex of passenger, male or female
age	toddler/child/teen/adult/senior	Age group passenger belongs to
class	1/2/3	Ticket class of passenger
sibsp	0/1/2/3+	Number of siblings or spouses on board
parch	0/1/2/3+	Number of parents or children on board
fare	0-20/20-100/100-200/200+	Ticket fare/price of passenger
homecountry	UK/Europe/US/Africa/Other	Home region of passenger
embarked	C/Q/S	Where passenger embarked on board
destination	US/NYC/Canada/Other	Destination port of passenger

Table 2: Description of preprocessed variables.

3.1 Choice of variables

I used the chi-square test of independence to identify which variables are statistically significant in the survival of passengers. The chi-square test assumes independence of observations which is not entirely true, as children are carried to safety or groups are saved in the same lifeboat, but nevertheless sufficient for the test. The test statistics and p-values are shown in Table 3 below, and the test is inconclusive as the p-values are all very small, indicating strong dependence on the variable of interest **survived**. The highest test statistics appear with the variables **sex** and **class**.

I then proceeded to do a simple Logistic Regression test, where I fit a Logistic Regression model with

the dependent variable and each explanatory variable separately. The coefficients and p-values are also shown in Table 3. With a significance level $\alpha = 0.05$, the null hypothesis, i.e., the variable has no significant effect on survival is accepted for variables **age**, **homecountry** and **destination**. However, given my domain knowledge that age should be a significant factor in survival in such an event, and the possibility of age having an effect when combined with e.g. **sex**, I will not remove this variable. I consider removing the other variables that passed the test, however, keeping in mind that all variables were significant based on the chi-square test.

Variable	C-S. Statistic	C-S. p-value	LR. Coeffiecient	LR. p-value
sex	255.55	1.60e-57	-2.49	3.22e-50
age	22.54	1.57e-04	-0.07	0.806
class	102.04	6.95e-23	-0.85	2.67e-22
sibsp	34.83	1.33e-07	0.78	1.99e-06
parch	24.28	2.18e-05	0.86	2.22e-05
fare	70.04	4.17e-15	2.10	5.25e-07
homecountry	58.98	4.75e-12	-0.098	0.83
embarked	28.83	5.50e-07	-0.70	0.012
destination	21.68	7.50e-05	0.99	0.181

Table 3: Chi-square independence and Logistic Regression test results.

Finally, I use stepwise Logistic Regression with AIC to evaluate the significance of variables. This method begins with a model with all variables and then selects variables which minimize the AIC, while removing variables which increase the AIC. The result is a Logistic Regression model with the most important variables for survival. This test assumes independence of observations, which I covered before, and a linear relationship between the variables and survival. I am suspicious that the relationship is always linear, so the results might be inaccurate. The variables that remain in this model are **age**, **class**, **sex** and **sibsp**. These results seem sensible, taking into account domain knowledge.

Combining all of the above tests, and my domain knowledge I decided to remove the variables **homecountry**, **embarked**, **destination**. I trusted the stepwise AIC logistic regression test and keep the variables **parch** and **fare** as they performed well in the other tests. The removed variables are coloured in red in Table 2.

3.2 Bivariate analysis

Now that I have preprocessed the data and chosen the most important variables, I will carry out some bivariate analysis between these variables.

The proportions of survived passengers in each explanatory variable can be seen in Figure A2 in the Appendix. Interestingly for **age**, *toddlers* have a higher survival rate than *children*, I expected this as I assume that many sufficiently young children have been carried to lifeboats by their parents. The lowest survival rates are with *seniors*. For **sex**, we can clearly see that females had a much higher survival rate than males. For **class**, we can see that first-class passengers have the highest survival rate and third-class passengers have the lowest. For **fare**, we can see that passengers with the cheapest tickets have the lowest survival rates and passengers with the priciest tickets have the highest survival rate, a potential direct correlation with the modalities in **class** is observed here. For **sibsp**, there is an interesting optimum as having one sibling or spouse on board has a higher survival rate than having no or more than one sibling or spouse on board. This could be explained by the fact that having an important sibling/spouse on board helps to stay together and save each other, whereas having too many might lead to more chaos and difficulties in staying alive, and having none might lead to no one looking out for 'you'. For **parch**, a similar phenomenon is seen as having one or two parents or children on board has an increased survival rate to having zero or more than two.

In Figure 1, I have plotted the correlation matrix as a heatmap of different modalities. The correlation matrix has the pairwise correlation of all pairs in the dataset. The most interesting row is the bottom row as this shows the correlation between the modalities of the explanatory variables and the dependent variable, **survived**. We can see that the strongest positive correlations with survival are ***sex**: female* and ***class**: 1st*, indicating that first-class passengers and females were the most likely to survive. A moderate positive correlation can be seen in ***age**: toddler*, ***sibsp**: 1*, ***parch** :1*, ***fare**: 20-200€*.

On the contrary, the strongest negative correlations with survival are ***sex**: male*, ***class**: 3rd* and ***fare**: 0-20€*, indicating that passengers most likely not to survive were male, 3rd class passengers and/or passengers with a cheap ticket. Here comes the issue of high correlation between ticket price and ticket class. A moderate negative correlation can be seen in ***age**: teen*, ***sibsp**: 3+*, ***parch**: 0*.

Several other relationships can be seen, which are listed below. Keep in mind that the statements below are meant to imply correlation, not causation.

> *1st class passengers are likely to be seniors.*
> *Passengers with 3+ siblings on board are likely to be toddlers or children.*
> *Passengers with zero parents/children on board are unlikely to be toddlers.*
> *Passengers with zero parents/children on board are likely to also have zero siblings or spouses on board.*
> *There is almost no correlation between the number of parents or children and the ticket class of a passenger.*
> *Passengers with a cheap ticket are unlikely to be female.*
> *Passengers with a cheap ticket are likely to be in third class.*
> *There is not much correlation between the ticket price and the age of a passenger.*

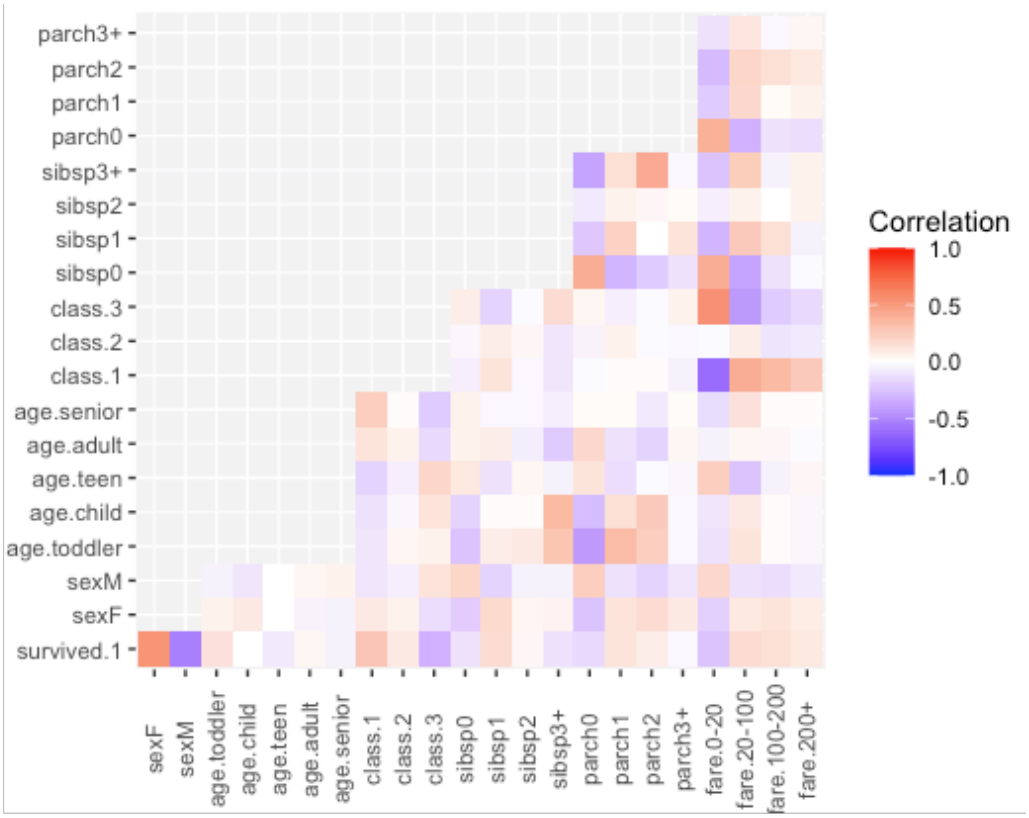


Figure 1: Heatmap of correlation matrix of variables.

4 Multivariate analysis

MCA (Multiple Correspondence Analysis), the multivariate counterpart of PCA is my choice of method for this problem. It is suitable since this method can handle multiple variables that are categorical.

The contribution, i.e., the importance of each modality in the MCA can be seen in Figure 2 below. The most contributing factors in survival according to this are having 3+ siblings or spouses, first and third-class tickets, belonging to age groups toddler or child, having 1 or 2 parents/children on board and having a cheap ticket. The contribution plot does not indicate whether these modalities heighten or lower the chances of survival, only that they contribute to the model.

The variable factor map, shown in Figure 3, displays the variables in two dimensions and with colour. The colour corresponds to the quality of representation of said variable, with red having high quality and blue having low quality of representation. The position of the points indicates which variables are strongly associated. The deceased and survived modalities are opposite of each other and here we can identify whether the contributing modalities of the previous section are associated with survival or not surviving. For example, having a first-class ticket is close to the modality of surviving and having a cheap ticket is closer to the modality of the deceased.

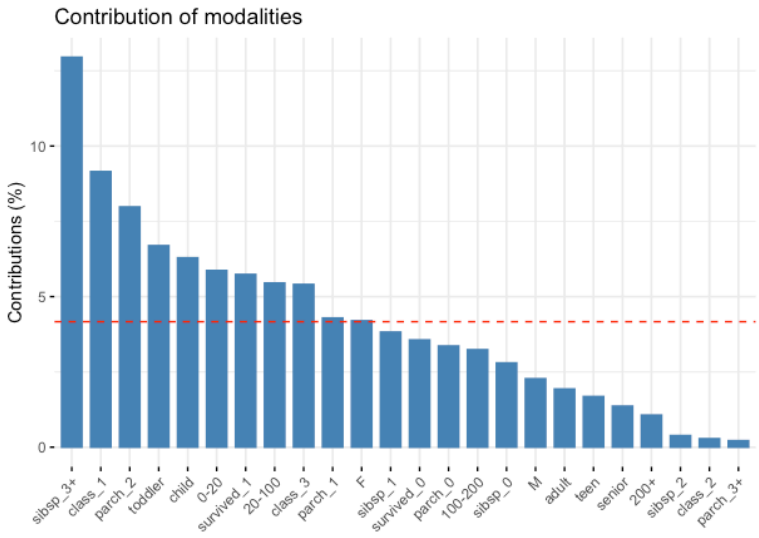


Figure 2: Contribution of modalities in MCA.

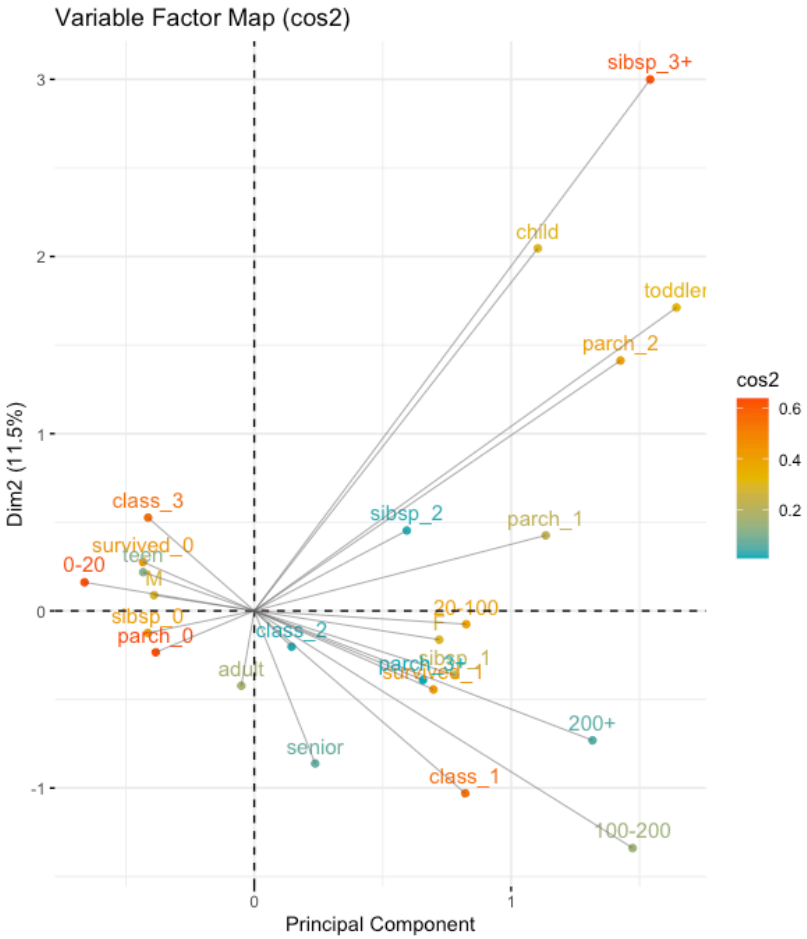


Figure 3: Variable factor map in MCA.

Interestingly, some important modalities according to the contribution plot (Figure 2), ***sibsp****: 3+*, ***age****: toddler/child* or ***parch****: 2* are somewhat orthogonal to both modalities of survival. They are however closer to the direction of the modality of surviving, but it is difficult to draw certain conclusions. On the other hand, blue or green-coloured modalities have a very low quality of representation and no conclusions should be made based on this plot for these modalities.

In Figure 4, I have plotted the observations in the principal component space (same as in plot above). One can identify areas where most points correspond to deceased passengers and other areas where most points correspond to survived passengers, but there are definitely not clear clusters. Combining this plot with the previous plot results in a biplot, but with such a large number of variables and observations, that plot was too cluttered to include in the report.

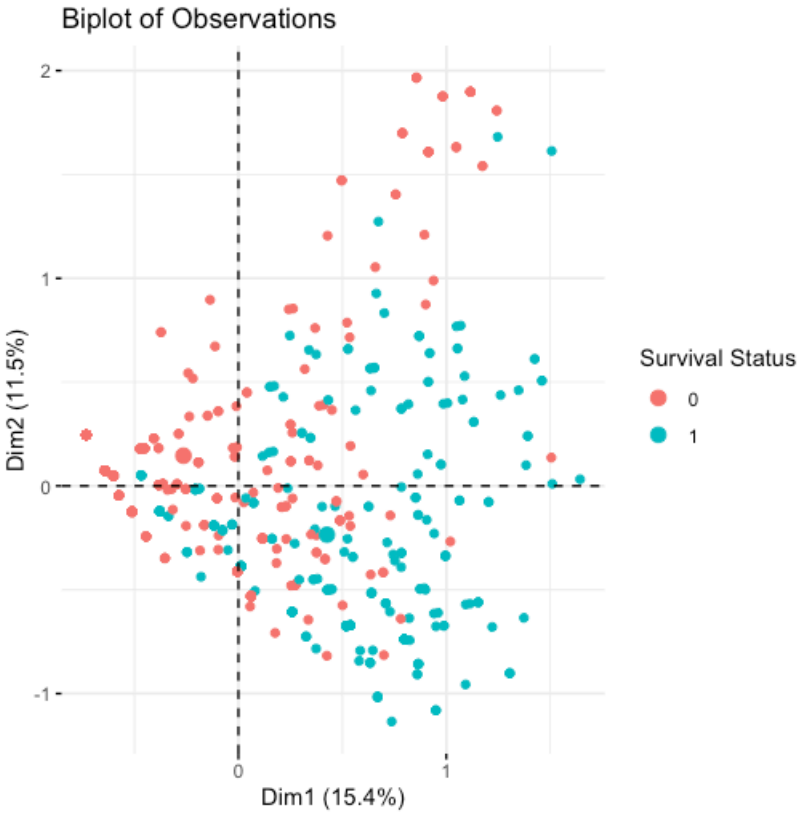


Figure 4: Scatterplot of the observations in MCA.

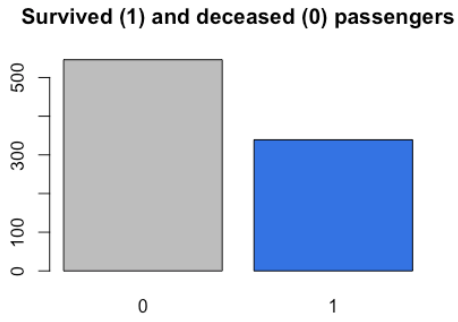
5 Conclusion

Strong conclusions cannot be drawn based on this work. Further multivariate analysis and review of possible bias especially in preprocessing are required. There are however some clear insights, which are immediately obvious from the data, before any multivariate analysis. First-class women had among the best survival rates, whereas third-class men had low survival rates. There are indications of an ”optimal” number of siblings/spouses, or children/parents onboard the Titanic in terms of survival rate, as it seems that having too few or too many relatives lowers the survival rate. There are also indications that in terms of age, toddlers have the best survival rates and (male) seniors have the worst survival rates. The variables that contribute the most in the MCA are **class**, **sibsp**, **parch** and some modalities of **age**. It is interesting how modalities of the same variable may be important or unimportant, e.g., having a second-class ticket does not strongly contribute to survival, whereas the other classes very much affect survival.

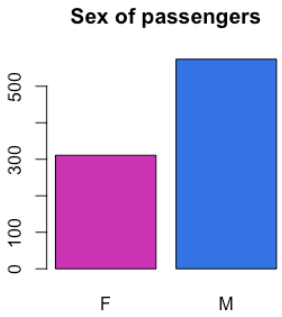
A possible source of bias is the use of my *domain knowledge*. It heavily influenced the way I regrouped modalities in data preprocessing in Sections 2 and 3 and even in the interpretations of independence and other tests performed in order to choose the significant variables in Section 3.1. Another issue is that the observations are not perfectly independent, which might influence the accuracy of many tests or the MCA. The correlation between variables **class** and **fare** should be investigated. There are also some modalities with too few observations: ticket prices over 200€ with 19 observations and the number of parents or children onboard being over 3 with 15 observations, which can reduce the accuracy of MCA. The rest of the modalities have 30+ observations with most modalities having hundreds of observations.

I would like to acknowledge the assistance of ChatGPT in this work, which improved the quality of the plots generated in the R programming language. Additionally, ChatGPT's suggestions inspired certain parts of my text.

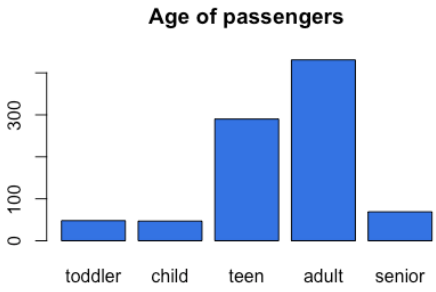
A Appendix



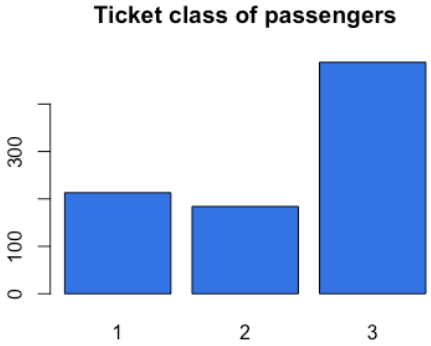
(a) Survivors and deceased passengers.



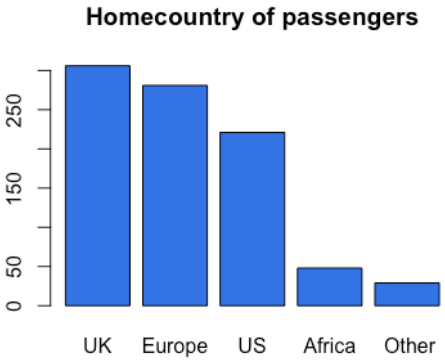
(b) Sex of passengers.



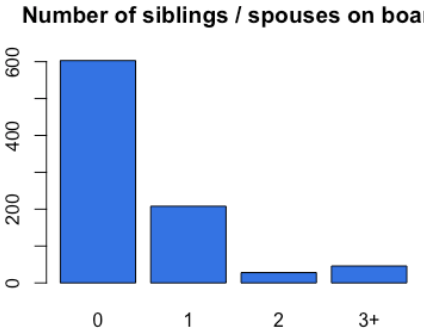
(c) The age groupings of passengers.



(d) The ticket class of passengers.



(e) The homecountry or region of passengers.



(f) The number of siblings or spouses on board, for each passenger.

Figure A1: Example variables and data after preprocessing and regrouped modalities.

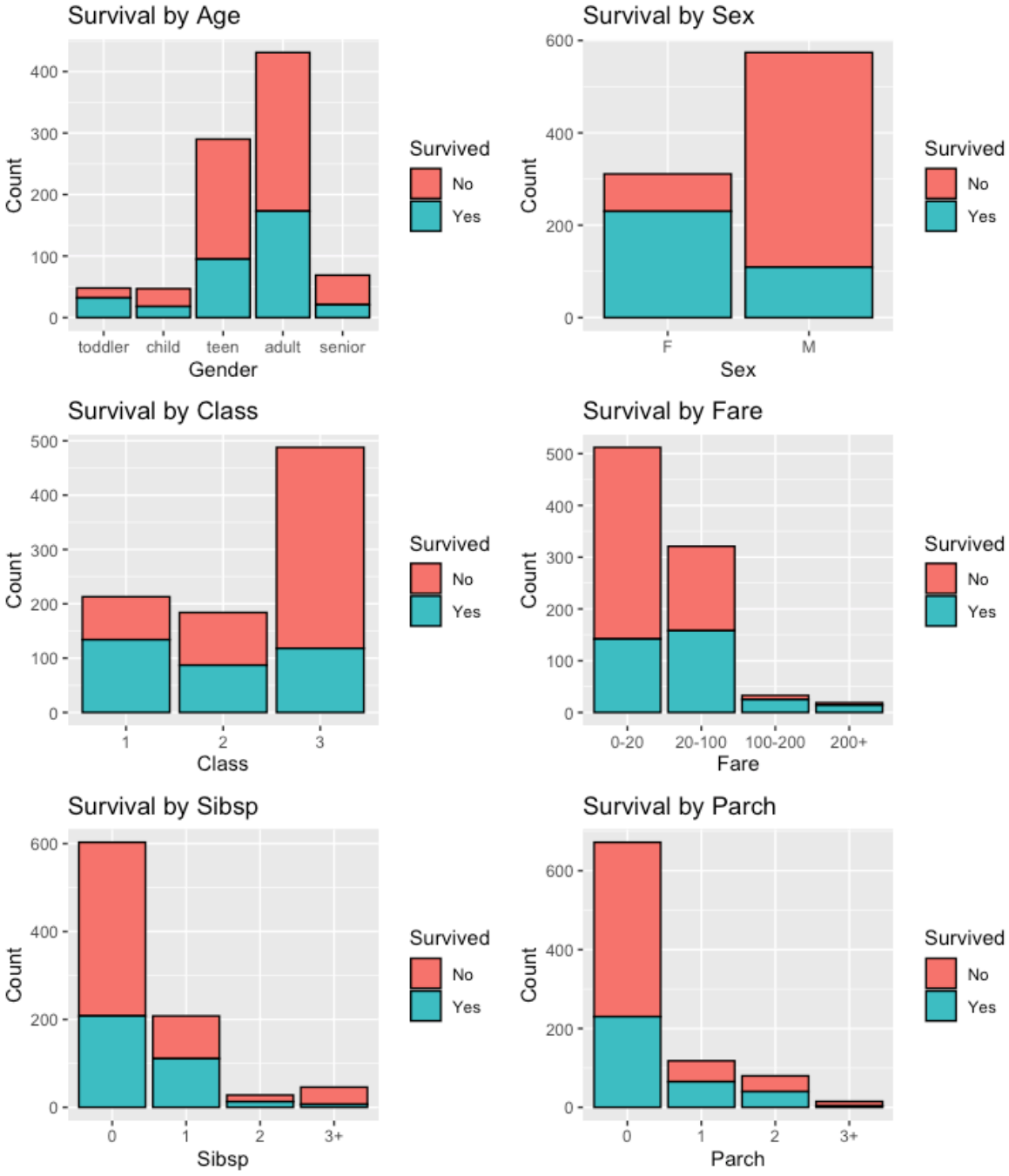


Figure A2: Proportion of survivors in each variable and modality.

References

Titanic extended dataset (kaggle + wikipedia), 2019. URL https://www.kaggle.com/datasets/pavlofesenko/titanic-extended.

Will Cukierski Jessica Li. Titanic - machine learning from disaster, 2012. URL https://kaggle.com/competitions/titanic.