Stat3302 Final Report

6. You may include R code in an appendix (not counted in the 5-6 page limit), but no R code

or R summaries can be included in the main report. For example, present your results in

tables and make sure to discuss your results in the text of the report.

Background description:

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class. [1] (Titanic: Machine Learning from Disaster, n.d.)



1. Scientific question:

What sorts of people were more likely to survive the Titanic sinking?

1. EDA(Exploratory Data Analysis)

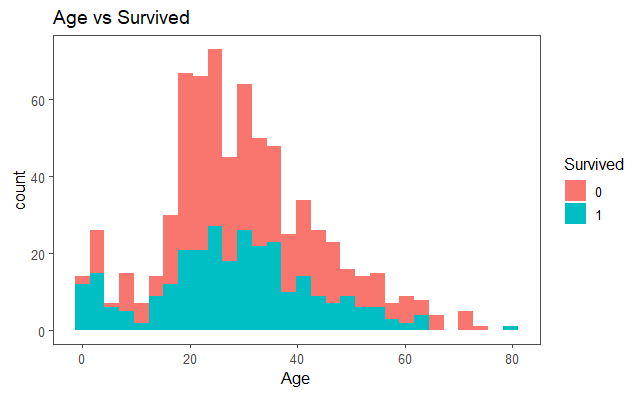
Now, before diving into the data, let’s have some understanding about our data. Specifically, we want to have some information about the structure of data, what variable do we have, what is out target and response variables, what assumption could we say about out model?

| Variable | Description |
| --- | --- |
| PassengerId | (sorted in any order) |
| pclass | Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex |
| Age | Age in years |
| sibsp | # of siblings / spouses aboard the Titanic |
| parch | # of parents / children aboard the Titanic |
| ticket | Ticket number |
| fare | Passenger fare |
| cabin | Cabin number |
| embarked | Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton |

First, we discovered that the training dataset contains 891 observation and 12 variable; testing dataset contains 418 observation with missed variable “Survived” (That's the variable we want to predict with for testing dataset); full dataset contains 1309 observation and 12 variables. And, we know that our target variable is Survived (0=died, 1=survived), and rest of variables that what we need to figure. Now, we want to make some plot, or visualization graph, to help us discover what are those potential variables that may be useful to build a predictive model!

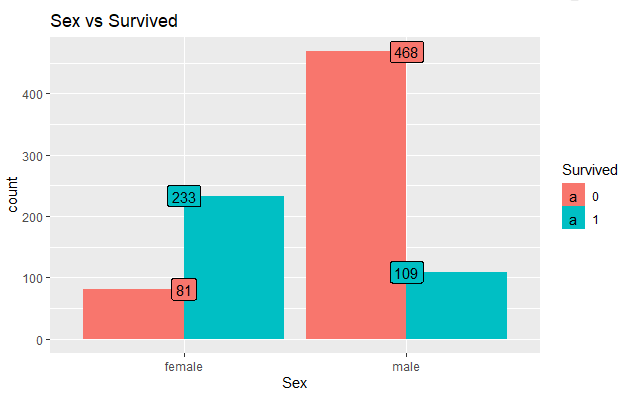
Second, let’s take a look at the univariate relationship between age and survived:

### 2.1 Age



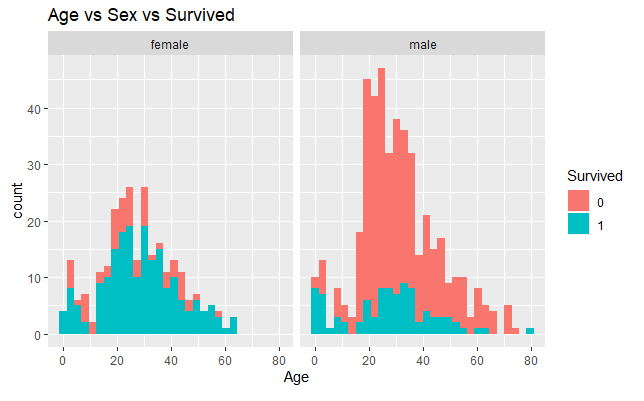
Note: 1)people who are in age between 20 to 50 are less likely being survived, and the people with age below 10 are more likely being survived; 2)the distribution of age looks like right skewed.

### 2.2 Sex Vs Survive



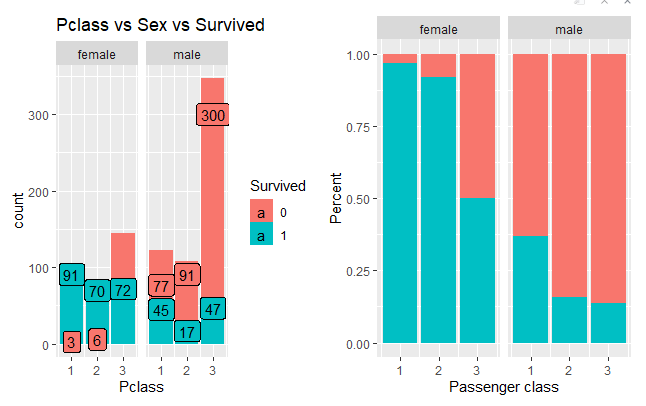
Note: 1) female is more likely being survived than male, female survived rate roughly 75%, and male is roughly 16.7%, so almost 5 time greater!

### 2.3 Age Vs Sex Vs Survived



Note: 1) Again, female is more likely being survived than male. 2) The differences between the number of peoples survived and not survived is largest at roughly age 20to30, and this is true for both female and male.

### 2.4. Pclass vs Sex



Note: 1) female is more likely of being survived than male in average. 2)In female group, majority passengers in class 1 and class 2 are survived, and more people in classed died. However, in male group, the survived rate in class 2 (~18.7%) just as bad as class 3(~15.67%).

### 2.5 Survived Vs Embarked

Levels: C Q S (C = Cherbourg; Q = Queenstown; S = Southampton)

0 1

C 0.441 0.559

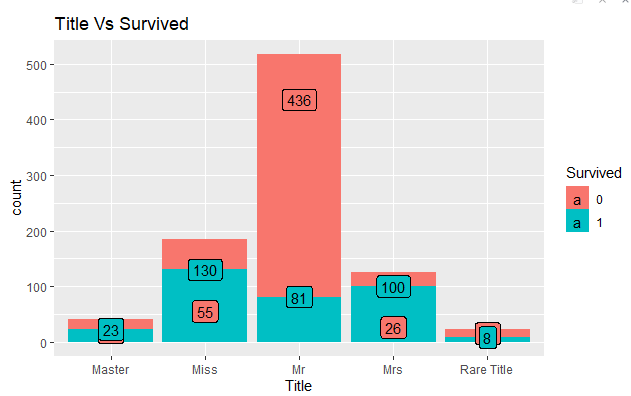
Q 0.610 0.390

S 0.663 0.337

Note: The probability of survived decreased in an order as Embarked port from C > Embarked port from Q > Embarked port from S

### 2.6 Title Vs Survived

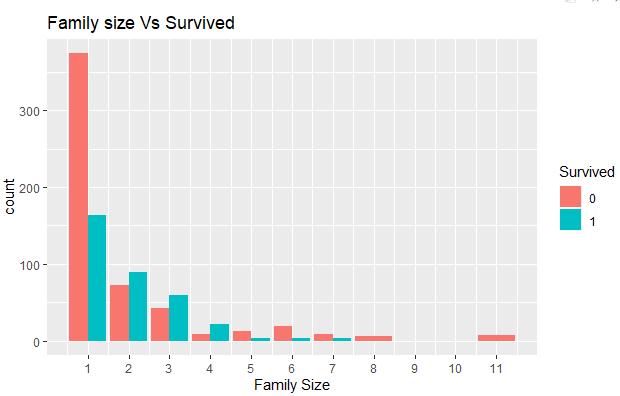
Notices, there are some useful infromation in passenger name, what is it? For example: the passenger title! (e.g. Ms, Miss, Mrs..) So, we can use this information to ask some question like, is there any relationship between the passenger title and probability of survived?



Note: we see the Mr. "Mr" are died pretty badly, which proved our previous observation that male is less likely being survived than female.

### 2.7 Family size Vs Survived

Family size might be a interesting predictor for evaluating the probability of being survived. So, let's use the sum of "sibsp" and "parch" to create another new variable, and then we can analysis their relationship!



Note: By comparing the "family size" and "Survived", we noticed the singleton, familes sizes 1, and large families (size > 5) are less likely being survived than the family with size between 2 and 4. Keep this in mind, that might be something we want to use in building our regression model.

1. Model Building

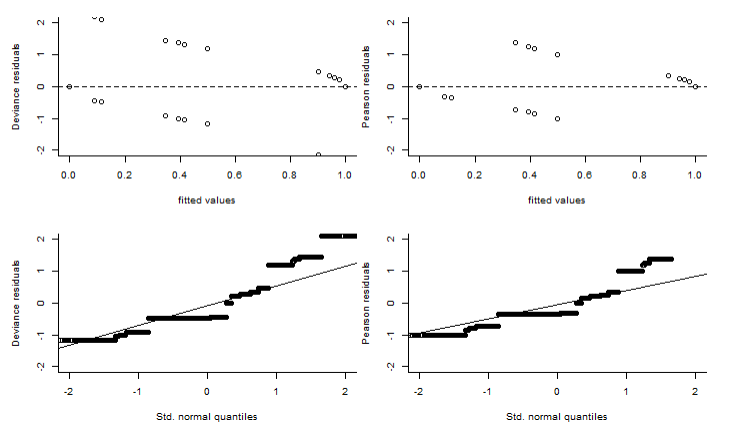
Null Hypothesis: zero line Resid Dev: 1186.7

### 4.2 SLLR on Survived~Age

### 4.3 SLLR on Survived~Sex

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Estimated variable | Estimate coeff | p-value(>|Z|) | AIC: | Resid. Dev | Df |
| Survived\_age\_model | Age | -0.00693 | 0.1616 | 1188.7 | 1184.7 | 1 |
| Survived\_sex\_model | Sex | -2.5137 | < 2e-16 | 921.8 | 917.8 | 1 |
| Survived\_pclass\_model | Pclass(1 = 1st; 2 = 2nd; 3 = 3rd) | -0.6394  -1.6704 | 0.001731  < 2e-16 | 1089.1 | 1083.1 | 2 |
| Survived\_Fsize\_model | Fsize(Double, Fsize\_3, Fsize4, large\_family, Solo,) | 0.1044  0.7531  -1.8606  -1.0425 | 0.6830  0.0903  9.72e-07  1.52e-08 | 1118.5 | 1108.5 | 4 |
| Survived\_title\_model | Title(Miss, Mr, Mrs, Rare) | 0.5579  -1.9855  1.0448  -0.9309 | 0.11915  6.4e-09  0.00713  0.08600 | 896.59 | 886.59 | 4 |
| Survived~Title + Sex |  |  |  | 891.37 | 879.37 |  |
| Survived~Age + Title + Sex |  |  |  | 819.48 | 809.48 | 886 |
| Survived~Title \* Sex \* Pclass |  |  |  | 768.14 | 738.14 | 876 |

1. Model Selection
2. Model diagnostic



Haha, so the residual doesn't work very well in this case, and that's because the number of success(survived), m\_i = 1, for each person(i) is one, which is like Bernoulli distribution, so that's kinda of missleading.

1. Conclusion

Now let’s answer our original scientific question that we proposed at the beginning.

With some interaction terms, we notice that the Deviance Residual had dicreased, but not substantially.

I think the reason is because that there might exist a very high correlation between those variables, so not too much variance can be explained by adding new terms. Also we can see that the degree of freedom is pretty big here(over 800), so if we applied the PCA technique to reduced the amount of features/variables and then picked several most important component as our representative variables, it's possible that we could get a better predictive model!

Appendix: