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?

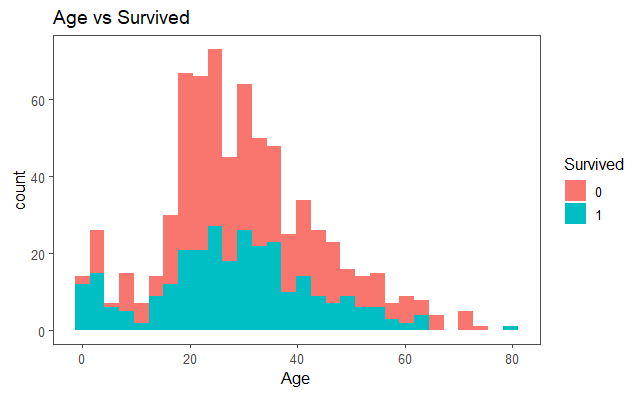
1. EDA(Exploratory Data Analysis)

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?

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!

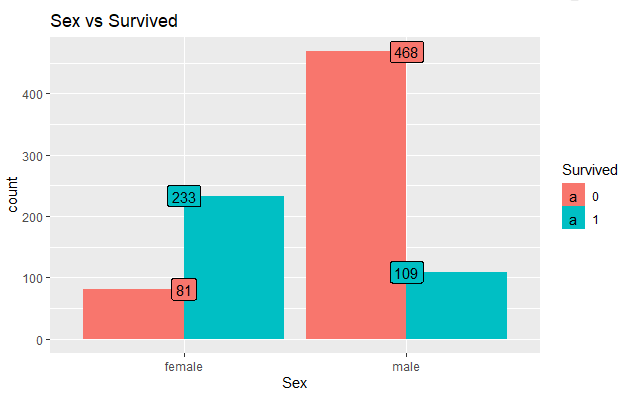
Second, let’s take a look at the univariate relationship between age and survived:

### 2.1 Age

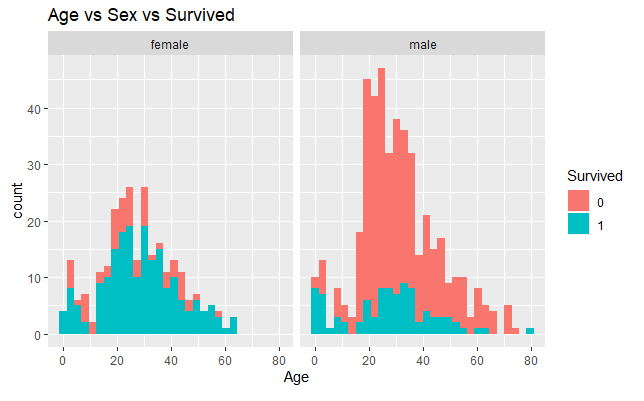


### 2.2 Sex Vs Survive

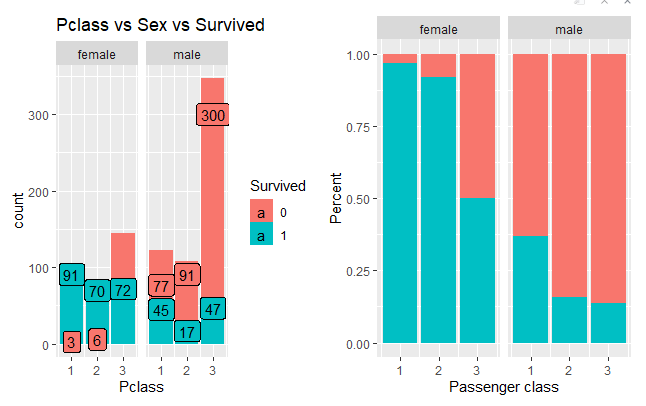
1. Model Building



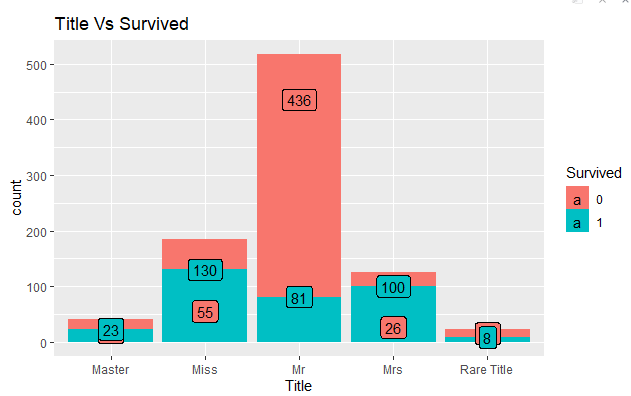
### 2.3 Age Vs Sex Vs Survived



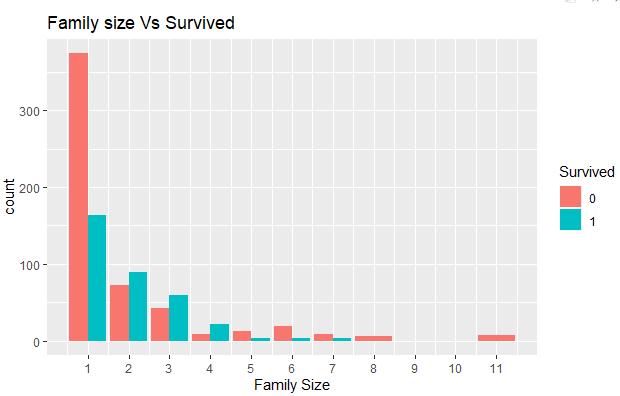
### 2.4. Pclass vs Sex



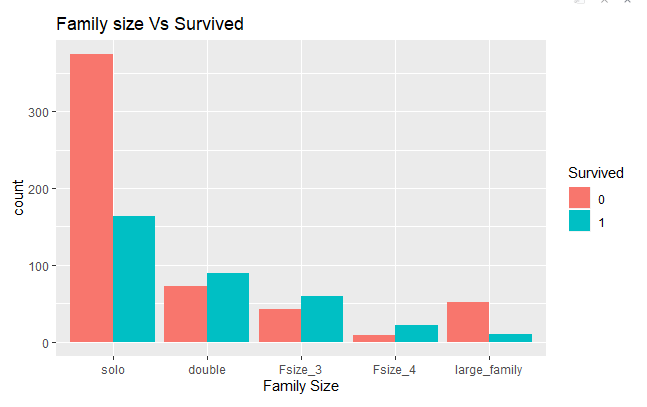
### 3.2 Title Vs Survived



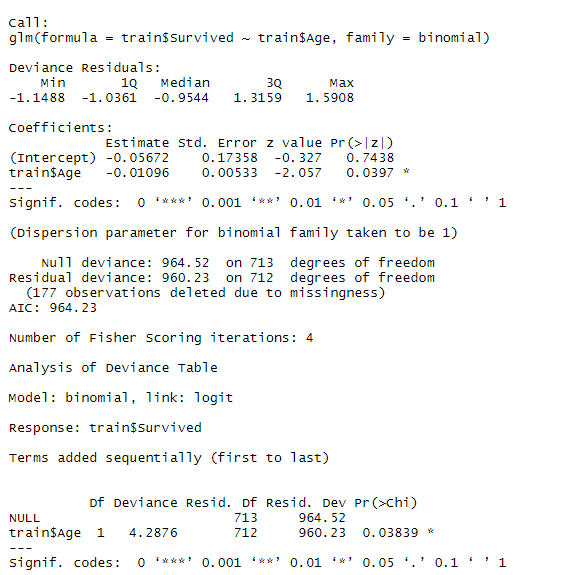
### 3.3 Family size Vs Survived



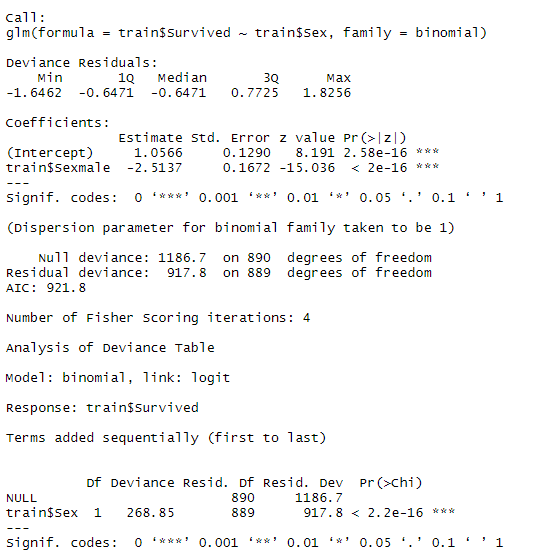
### 4.1 Redefined the Fsize to factor:



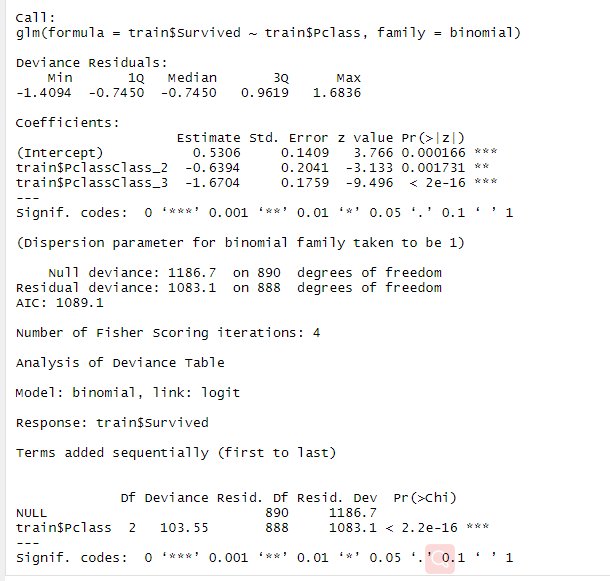
### 4.2 SLLR on Survived~Age



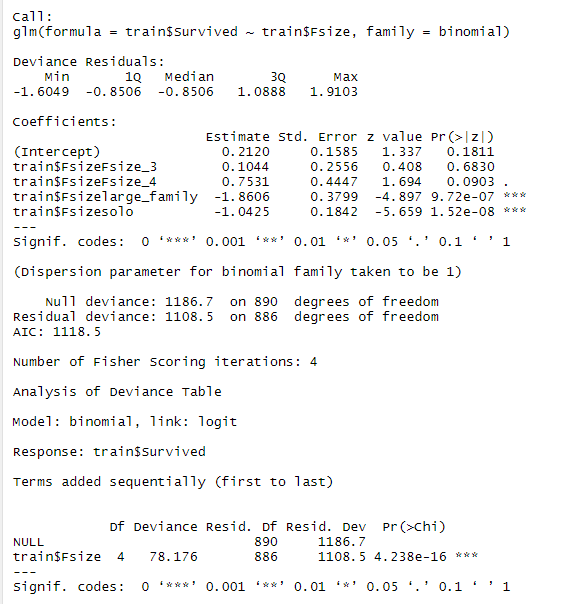
### 4.3 SLLR on Survived~Sex



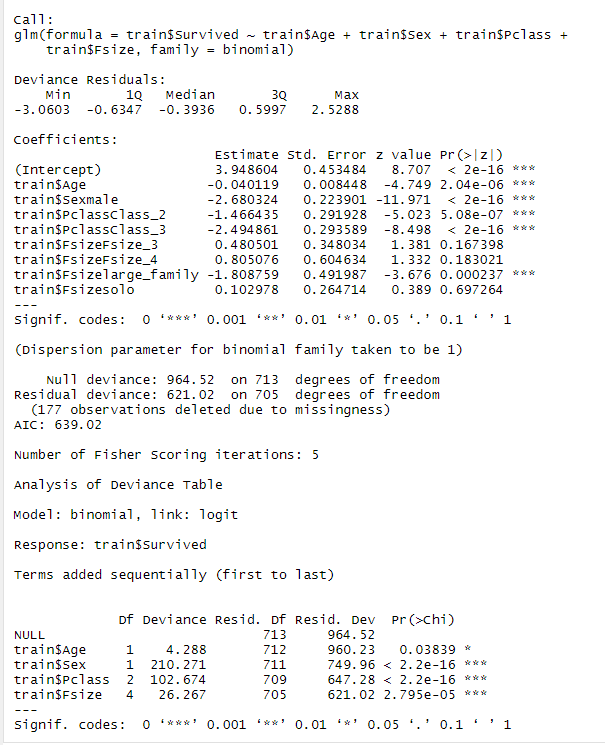
### 4.4 SLLR on Survived~Pclass



### 4.5 SLLR on Survived~Fsize



### 4.6 SLLR on Survived~Age + Sex + Pclass + Fsize



## 5. SLLP with interaction terms

