Chai Grindean

Blog Assignment

For the blog assignment I want to improve my Kaggle score for the Titantic Dataset. Seeing as how this was an early assignment in the first couple weeks, I am interested in seeing just how much better I can improve my score. This would 1) allow me to improve my predictive modeling, while 2) be able to actively see if my score is improving by turning it into Kaggle.

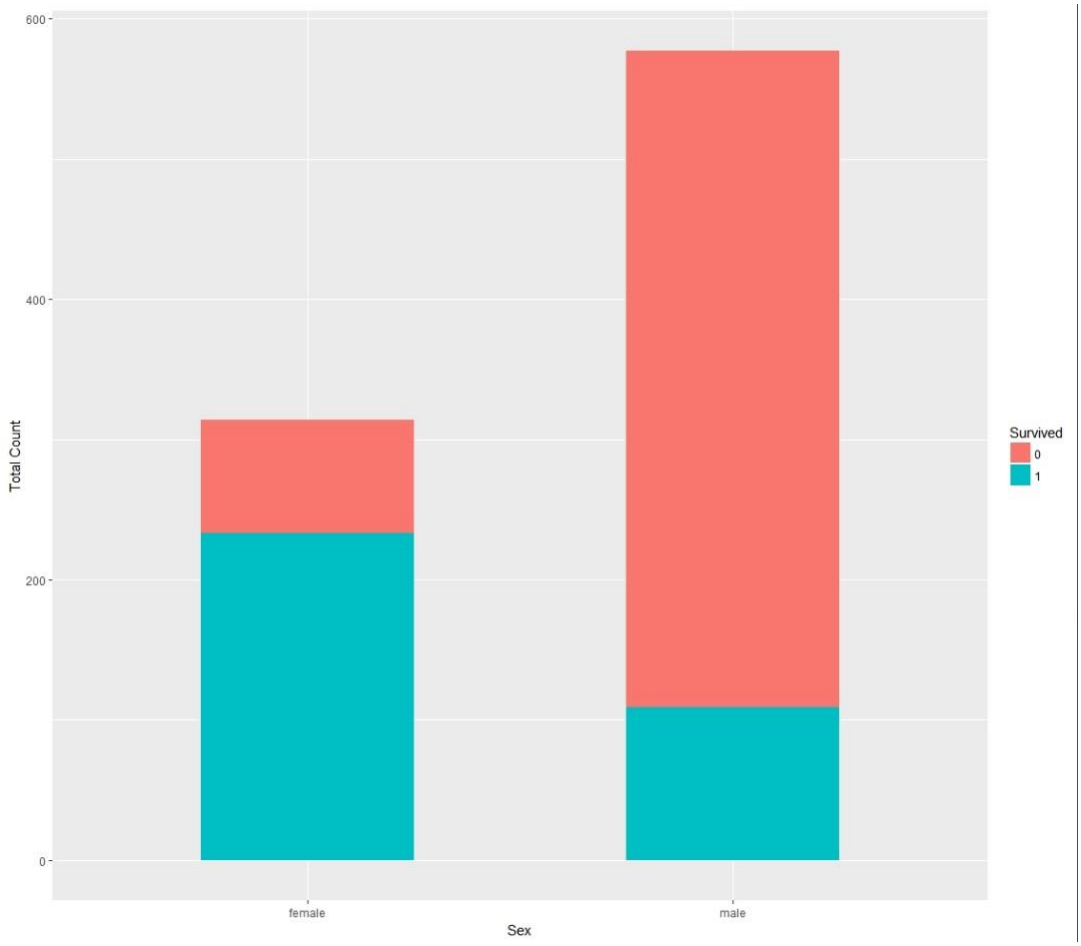
For this competition we have to try and accurately predict who will survive / who will perish in the event of the Titanic Sinking. Before I made simple assumptions for my analysis, such that men were more likely to pass away due to chivalry; given more time to explore the data, I would like to see more of the hidden factors that might correlate to an increased rate in mortality. The data can be found from the Kaggle Competition page.

https://www.kaggle.com/c/titanic/data

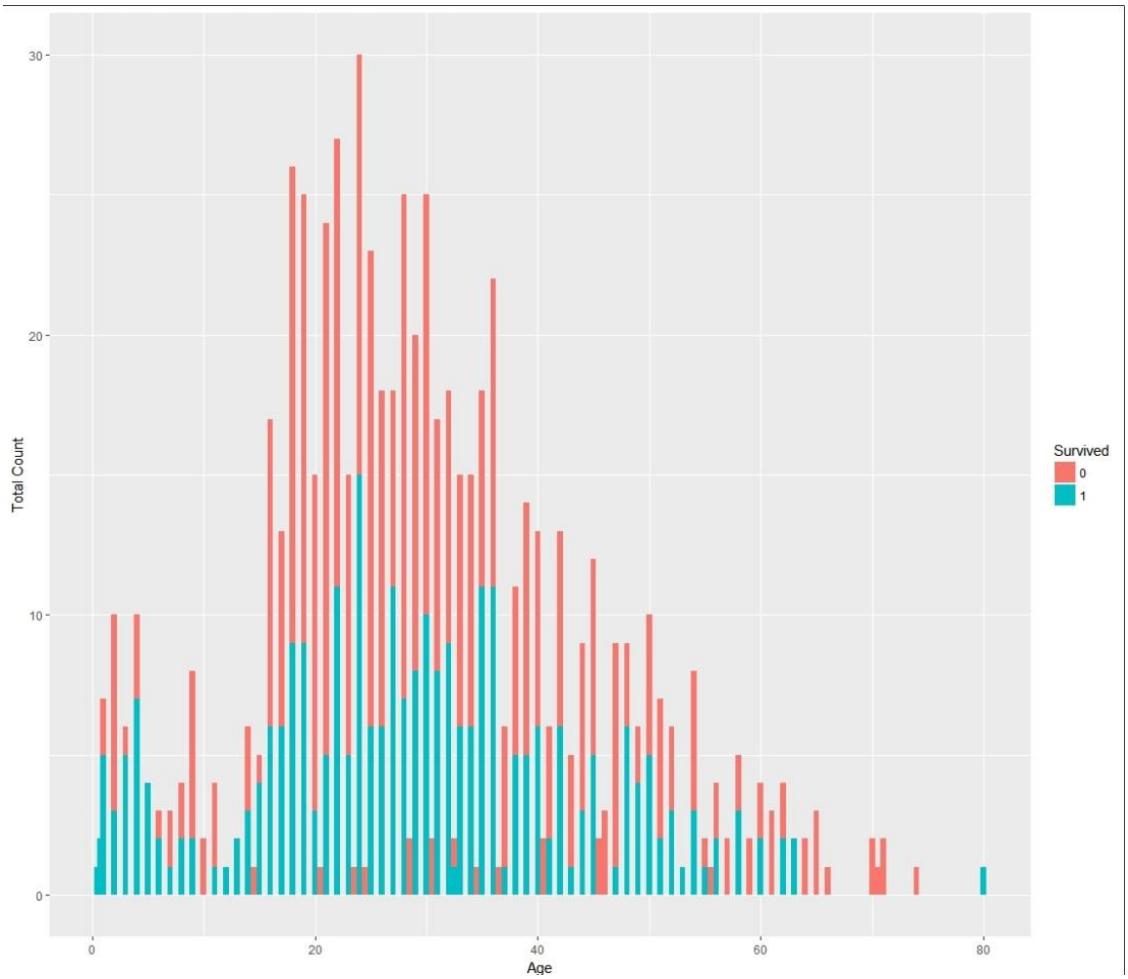
Before we are even able to begin creating a predictive model, we need to organize the data and get rid of NA values. There are many ways of trying to tackle this initial problem; some will give better solutions than others. As a starter, the first time I attempted this competition, what I did to fill in the missing values, was rbind both the test and train data, and take the median. I then used the median and filled it with the NA or missing values. I did this for the Fare and Age. This is an extremely inefficient way to try and find the missing Age values; because the median Age for all around, might be different to the median age for woman, or the median age for men. What I did differently this time around was create predictive models to predict more accurately the missing NA values. One other quick and easy technique I used to fill the data was

for the ‘Embarked’ There were 2 NA values, because it was just a low count, I assigned it to the mode of Embarked.

Before we create our predictive model, let’s look at what the data has to say.

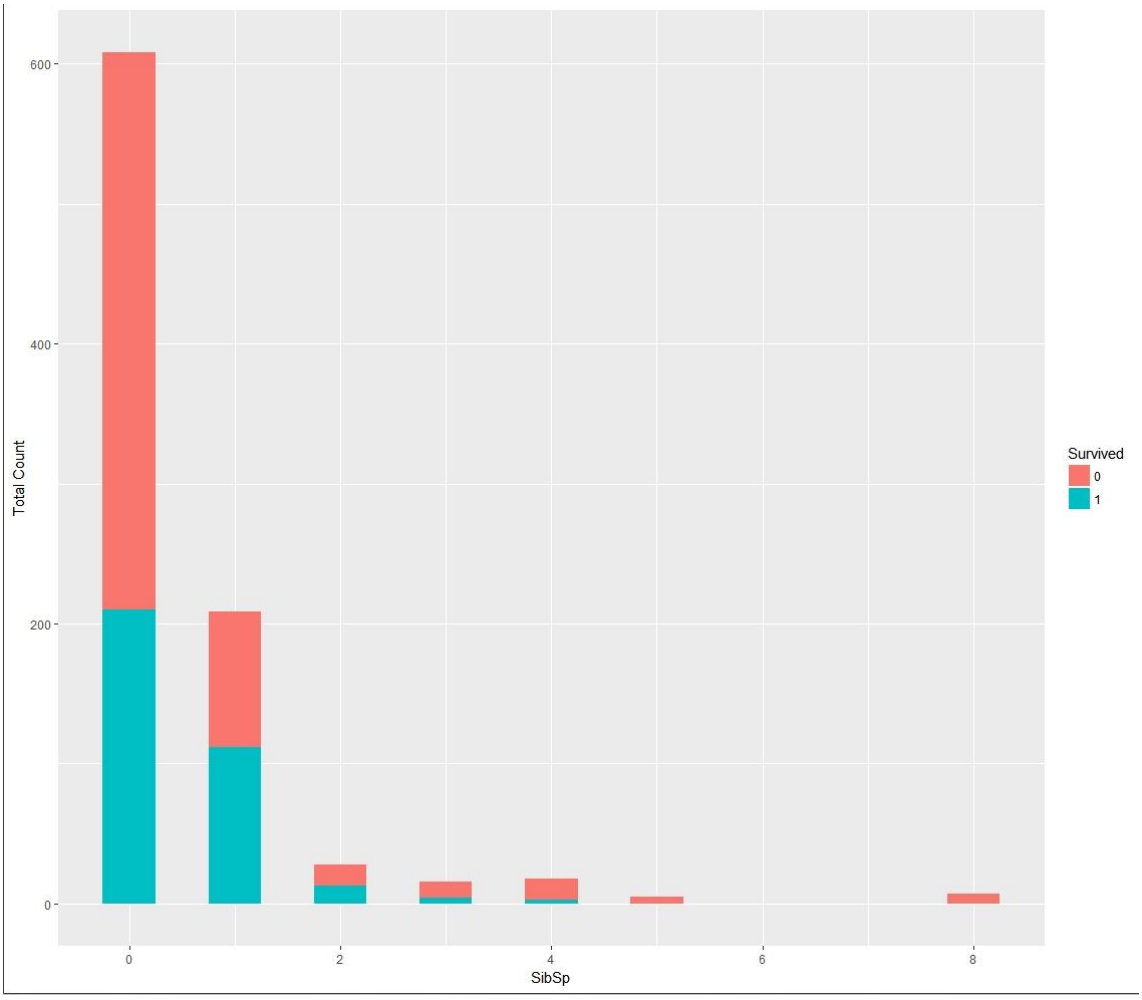


Right away we are reminded of the time period the Titanic happened in, a more chivalrous time period. The data is heavily skewed towards women surviving and men perishing. These figures help fall in line, with the concept of wanting to save the women and children first. Which brings me to my next thought, let’s see how Age, affected survival rates.



After look at this, this does in fact support the idea of saving women and children first. Here the magical number is roughly at the age of 6. If you were younger than 6 years old, your chances of survival go up significantly. This is interesting because I have a couple different theories. One, being that because the child is so young, people prioritized on saving the weak children who can't help themselves. However this idea of trying to save the weak, doesn't apply to senior citizens. As we can see, there isn't the same chance for survival when you look at the opposite end of the spectrum. Because of these finding, I am now curious to know, did the number of Siblings or Parents affect your survival chances? Since we know that being extremely

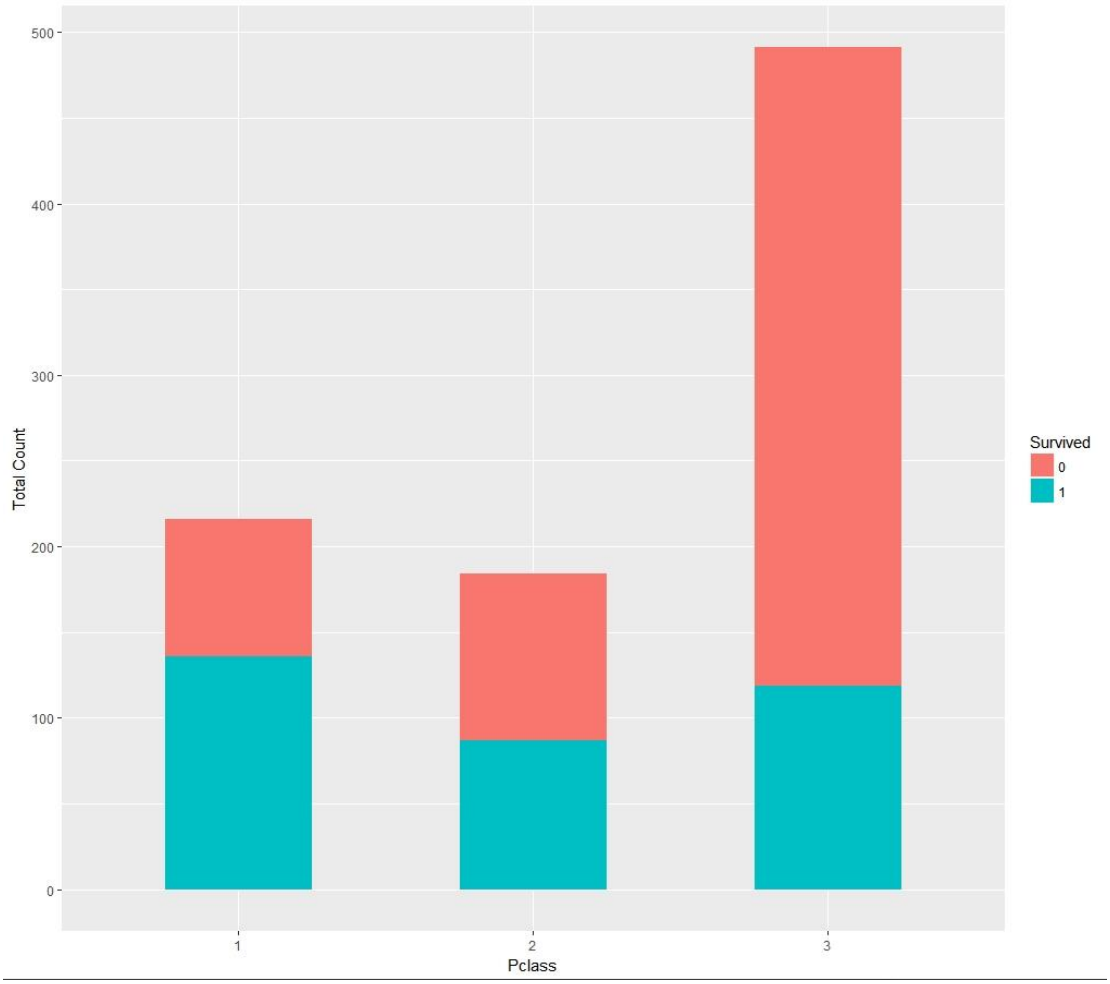
young increases your chances, it would make sense that they came on board with other siblings and parents. Perhaps the number of family member you are with will be a big factor.



Now from this we see that the most people that survived were lone adults who were not accompanied by their parents or siblings, but this is just purely because this category had the most people. Per percentage, your best chances were if you had 1-2 other siblings or parents with you. To try and make sense of this, here is my theory. We now know that to have the best chances for survival, it would be in your favor, if you were a young female child; being a child implies that you are not independent and need to rely on others. This would mean that

as a child you are being accompanied by at least 1 parent, potentially your mother. If this was that case, and both the mother and child survived, then this would increase the stats of the survival rates when 'SibSp' is between 1-2. Unfortunately there aren't any distinctions between which family members you were accompanied by, as this would be another interesting factor to look at.

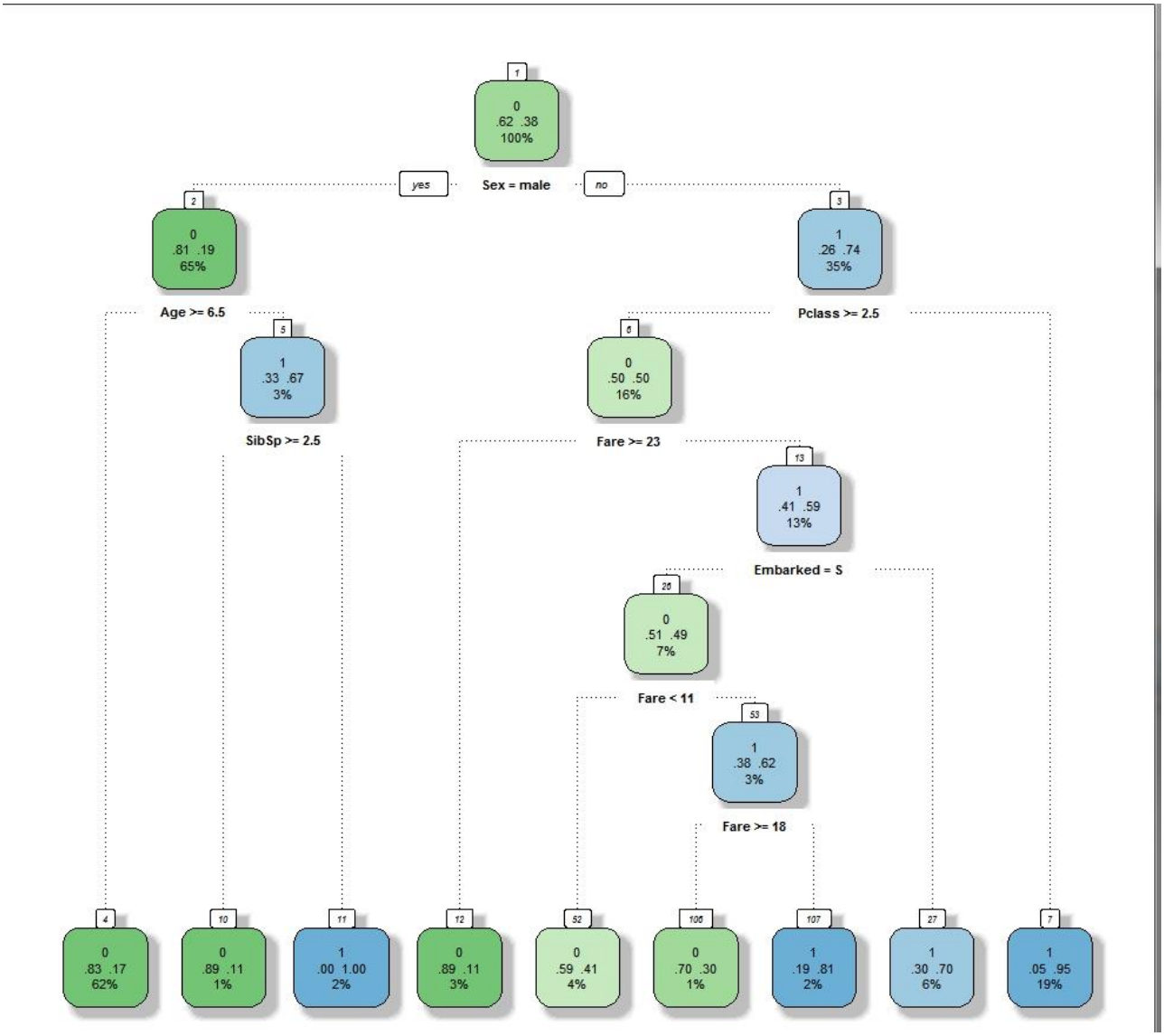
Last, but certainly not least. Let's take a at the 'Pclass' or Passenger Class factor.



This ggplot for Pclass does a great job of showing just how drastic your chances for survival were if you were a third-class passenger. Looking at this image, a couple things stand

out. One, there were actually more survivors from 1st class, than there were people who perished. This is extremely important to notice because as most of us know, through trivial knowledge, more people did not survive the Titanic crash. Another thing that stands out to me, is that there are also more 1st class passengers, than 2nd class. Generally you would except for there to be the most people in the lowest costing class, and the least in the more expensive 1st class. This could be of some significance since we see that 1st and 2nd class passengers had a strong fighting chance, but only 32% of the people in 3rd class survived. I personally have been on a cruise before, and if the Titanic was in any way the same, it could be a fare assumption that people in 1st and 2nd class were higher up on the ship, this means closer to life-rafts. That could help explain the massive increased survival rates for people in 1st and 2nd class. Another sad truth could be that people just valued your life more if you were wealthy and were a higher class passenger.

Now that we have filled in the missing NA values, and have spent some time looking through the data, we can create our predictive model. I used randomForests for my prediction models. Here is an rpart tree to see the data in a different presentation.



Here we can see the different survival rates. What we would expect, from the data, is that you would have the best chances for survival if you were a young (<6) female child in 1st class, being accompanied with her parent/s. Following this tree, it agrees with us in that those factors play a big role but, surprisingly, it predicted that a young male child, being accompanied with his parent/s, was just of likely to have survived, as the female child in first class. This is crucial detail because now we know that even though women had a much higher chance to survive, this wasn't a factor if you were a young child.

Overall I enjoyed my findings and am pleased with my improvements from the first time I attempted this competition. Some things I was able to do differently was actually take the time to look through the data, make my own predictions and then reconfirm them with r. This helped me greatly with my r programming skills. I also made predictive models to fill in the missing values instead of omitting them or just trying to fill it with the median. If I were to continue with this competition, I think the next step to try and make a more accurate reading, would be to try and cross validate some of the data. I noticed there were a couple of outliers in the data, I applied some filters to help my predictive model, but there is always room for more improvements.

VERSION 2 CHANGES

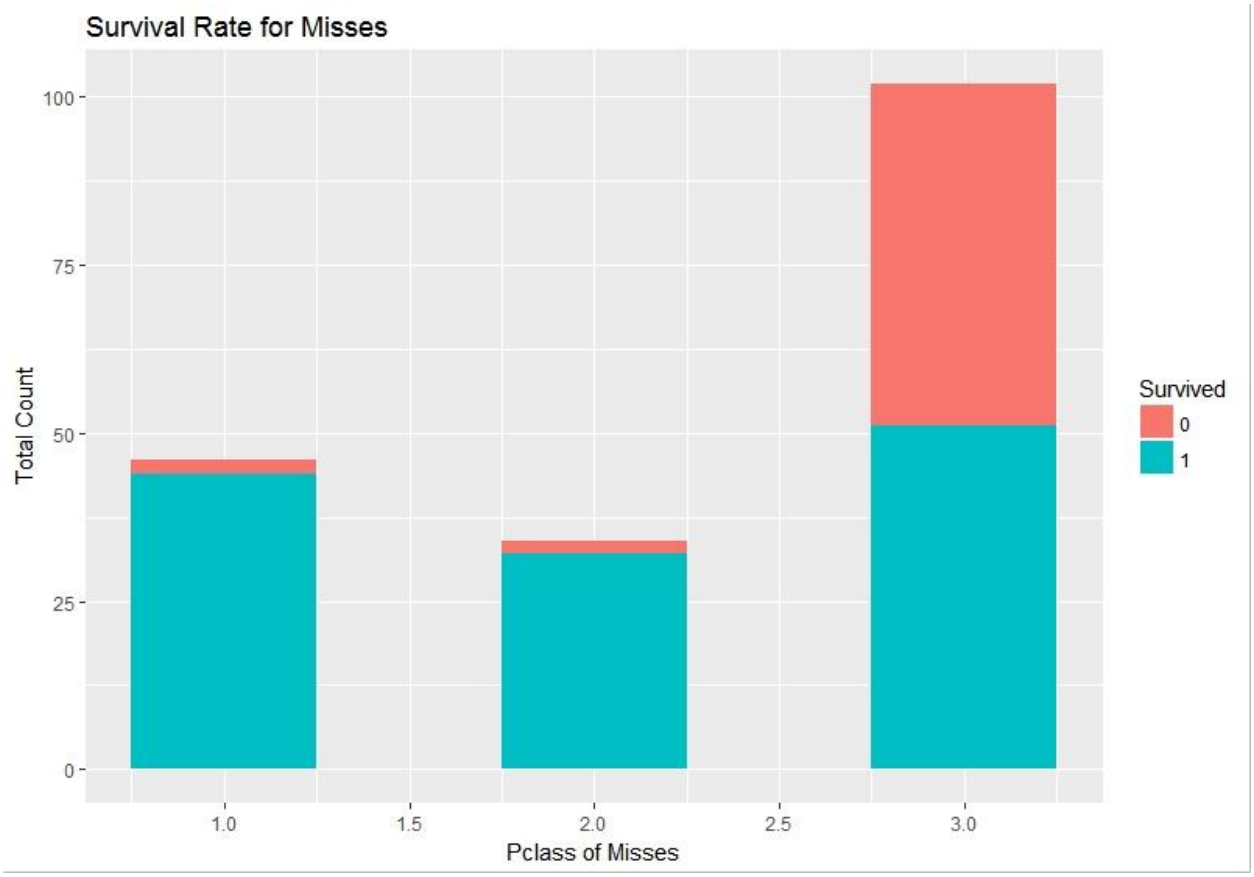
Overall I changed some of the source code to fix some minor readability issues and some unneeded code.

Added some feature engineering with title extractions. Just to help the model I pulled the titles and changed them to fit into 3 groupings, Mr., Mrs., and Miss. I choose to break it down into 3 groups just to 1 help simplify all the different titles that could be used to describe the same type of honorifics. The reason I left an extra grouping for Miss, was because I wanted to see if being a Miss (often used to denote a single, unmarried lady) affected survival rates.

Capt.	Col.	Countess.	Don.	Dona.	Dr.	Jonkheer.	Lady.	Major.	Master.	Miss.	Mlle.
1	4	1	1	1	8	1	1	2	61	260	2

Mme.	Mr.	Mrs.	Ms.	Rev.	Sir.
1	757	197	2	8	1

```
> titles.lookup
       Title New.Title
1        Mr.        Mr.
2      Capt.        Mr.
3       Col.        Mr.
4       Don.        Mr.
5        Dr.        Mr.
6  Jonkheer.        Mr.
7     Major.        Mr.
8       Rev.        Mr.
9    Master.        Mr.
10       Sir.        Mr.
11       Mrs.       Mrs.
12      Dona.       Mrs.
13      Lady.       Mrs.
14       Mme.       Mrs.
15 Countess.       Mrs.
16      Miss.      Miss
17      Mlle.      Miss
18        Ms.      Miss
```



Shockingly here, we can see that Passenger Class played a massive role in deciding if a single female survives. There were hardly any casualties lost from the 1st and 2nd class.

Sources Used

Langer, David. "Introduction to Data Science with R – Data Analysis Part 1." *YouTube.* YouTube, 08 Nov. 2014. https://www.youtube.com/watch?v=32o0DnuRjfg

Langer, David. "Introduction to Data Science with R – Data Analysis Part 2." *YouTube.* YouTube, 08 Nov. 2014. https://www.youtube.com/watch?v=u6sahb7Hmog

Langer, David. "Introduction to Data Science with R – Data Analysis Part 3." *YouTube.* YouTube, 08 Nov. 2014. https://www.youtube.com/watch?v=aMV_6LmCs4Q

Data Science Dojo. "My First Kaggle Submission." *YouTube.* YouTube, 14 Nov. 2016 https://www.youtube.com/watch?v=68l47Zu4Yxg

Lamouri, Saad. "Saad Lamouri: Kaggle Tutorial with R | Introduction to Data Science." *YouTube*. YouTube, 05 Mar. 2015 https://www.youtube.com/watch?v=tDsiiUYRphM

IBM, Armand Ruiz. "Getting Started with R: Titanic Competition in Kaggle." *Getting Started with R: Titanic Competition in Kaggle* . N.p., 26 Mar. 2015 http://armandruiz.com/kaggle/Titanic_Kaggle_Analysis.html