Titanic Survival

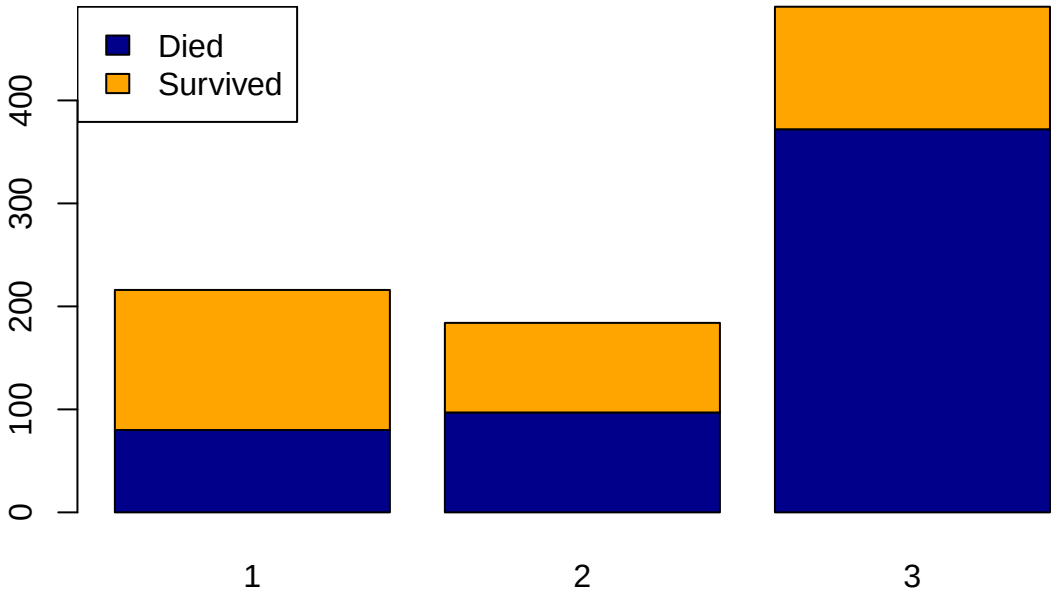
Travis Barton

7/24/2018

1. Data Clean

Pclass

This represents the passenger class of the ship. This should directly coorilate with socioeconomic status. Lets take a look at how they fared in terms of survuval (pun intented).



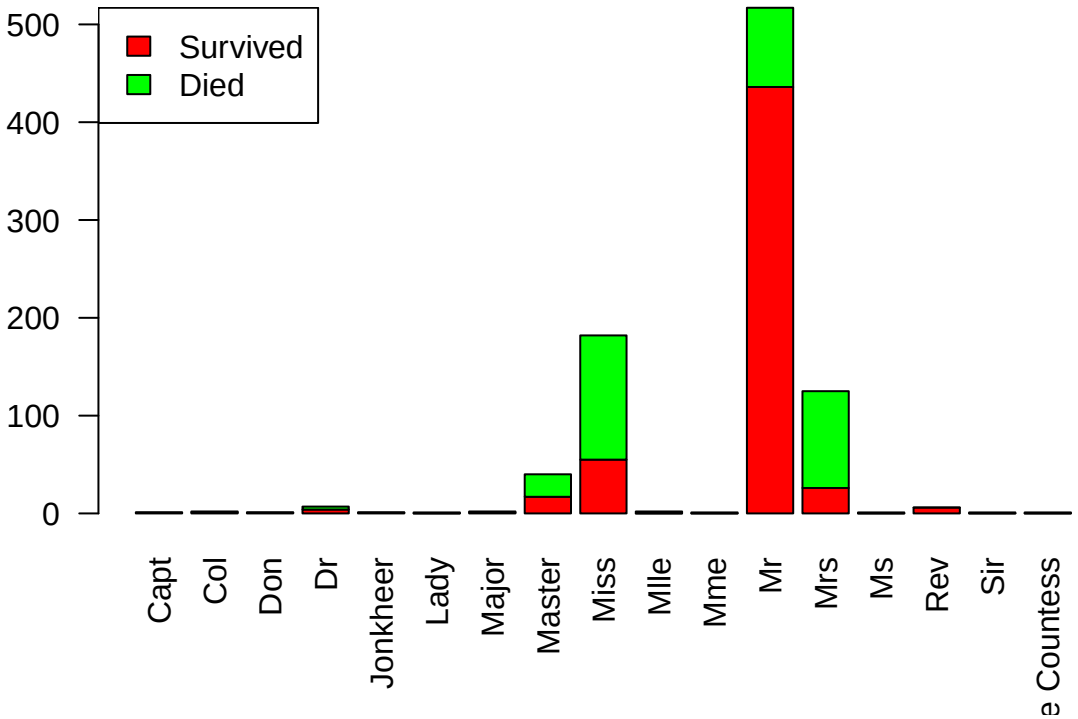
So it seems that a better passanger class corresponds with better survival. This means we should keep the variable for our final model.

Name

'Name' will be likely be the least helpful variable of the bunch. But thats not to say that there is nothing to be learned from it. The names themselves are not all that informative, but the titles that go along with the names will be. Their usefullness will be seen more in the 'age' stage, but for now, lets extract the titles.

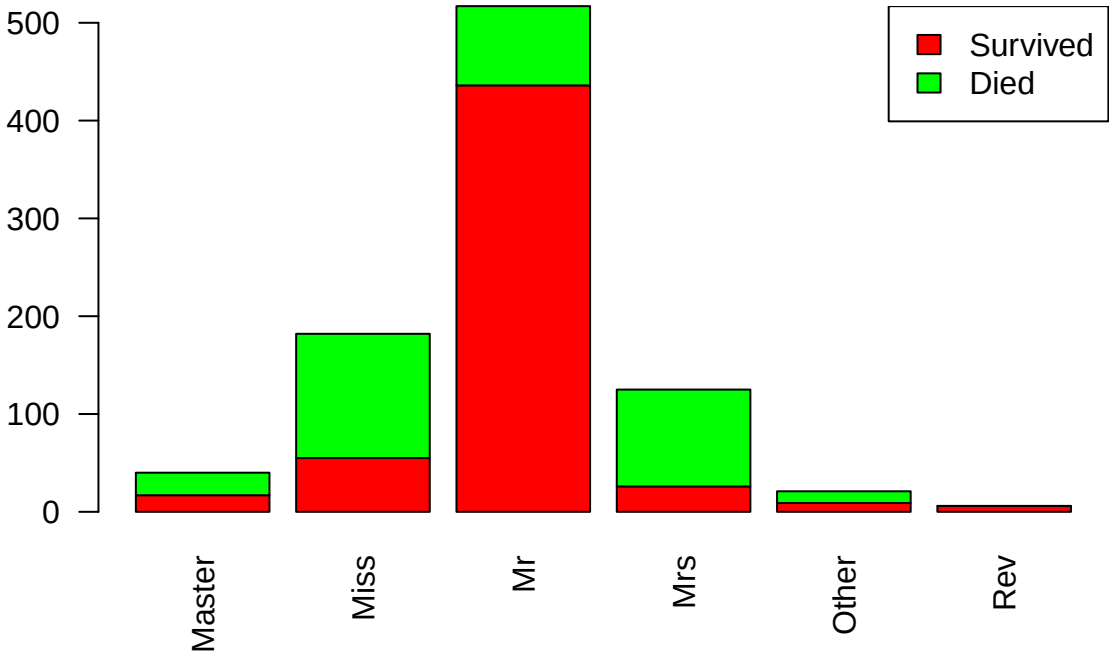
```
train$title <- str_sub(train$Name, str_locate(train$Name, ",")[ , 1] + 2, str_locate(train$Name, "\\.")

barplot(table( train$Survived,train$title), col = rainbow(3), las = 2)
legend("topleft", fill=rainbow(3), legend=c("Survived", "Died"))
```



So while many may be hard to see, Mr. seems to correlate with low survival rates, and Mrs/Miss seem to mean higher survival rates. This will be useful to us. In fact, the rest are so low, that we might as well combine them into their own catagory

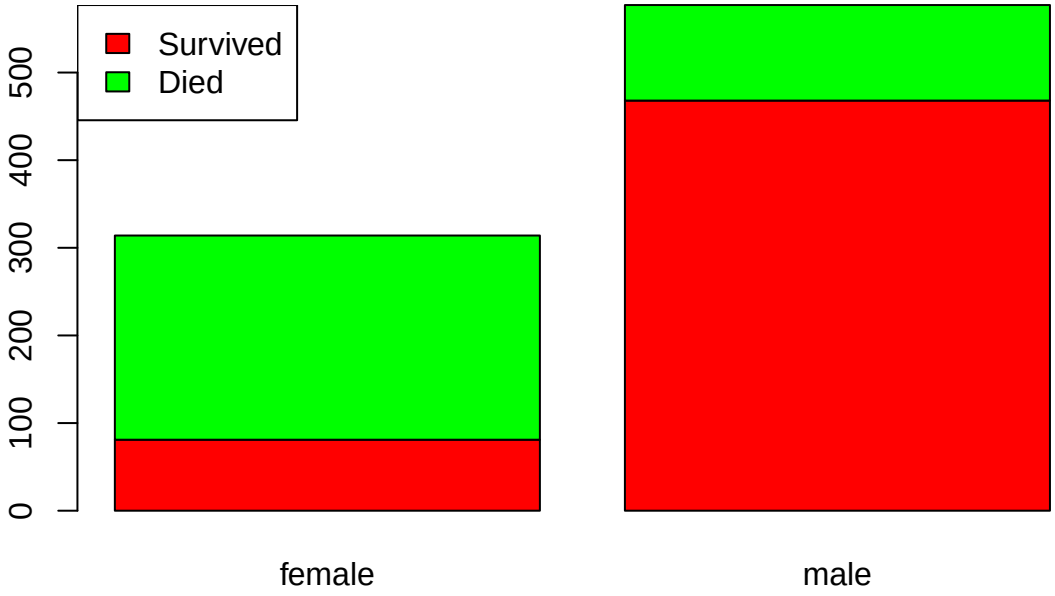
```
index <- which(train$title %in% c("Master", "Miss", "Mr", "Mrs", "Rev"))
train$title[-index] = "Other"
barplot(table( train$Survived,train$title), col = rainbow(3), las = 2)
legend("topright", fill=rainbow(3), legend=c("Survived", "Died"))
```



Sex

Lets see how sex relates to survival, and how many missing values we have.

```
barplot(table(train$Survived, train$Sex), col = rainbow(3))
legend("topleft", fill=rainbow(3), legend=c("Survived", "Died"))
```



Well it looks like being female on the titanic increased your likelyhood of survuval. This means that it will be a usefull predictor. What is more, we have no empty sex's.

Age

Age will be a troublesome variable. 19.86% of the ages are missing. With so many missing values, it would not be practical to just give them the same value (whether mean or median imputaion). Doing so would cause us to miss much of the information that can be gleamed.

This is where the title variable will come in. We can get a more accurate estimate of the ages of the passengers if we imoute the age based on their title.

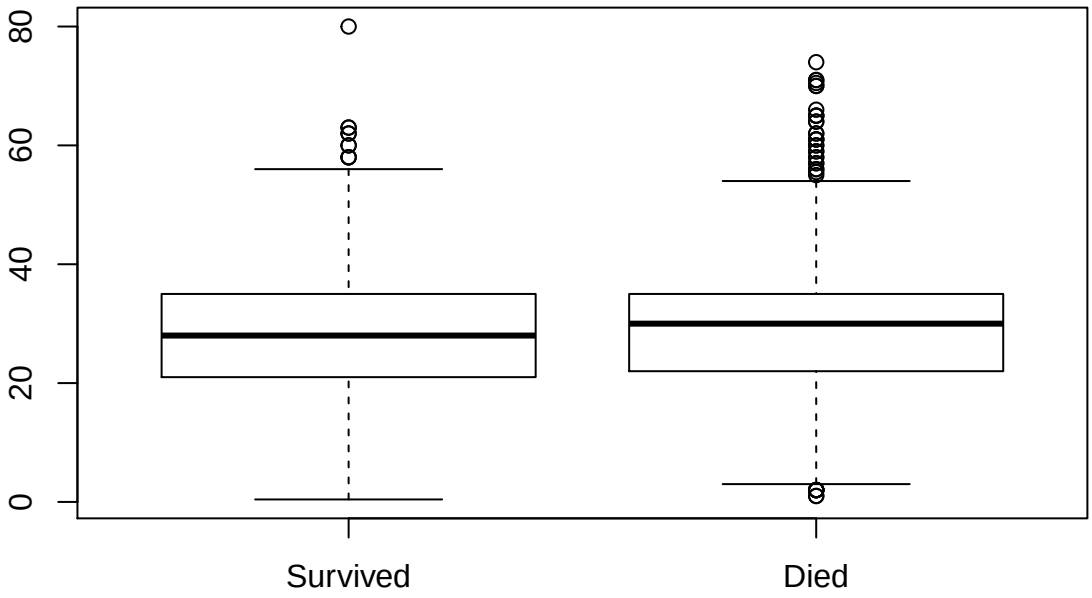
```
Master_age <- median(train$Age[which(train$title == "Master" & is.na(train$Age) == F)])
Miss_age <- median(train$Age[which(train$title == "Miss" & is.na(train$Age) == F)])
Mr_age <- median(train$Age[which(train$title == "Mr" & is.na(train$Age) == F)])
Mrs_age <- median(train$Age[which(train$title == "Mrs" & is.na(train$Age) == F)])
Other_age <- median(train$Age[which(train$title == "Other" & is.na(train$Age) == F)])
Rev_age <- median(train$Age[which(train$title == "Rev" & is.na(train$Age) == F)])

for(i in 1:891)
{
  if(is.na(train$Age[i]) == T)
  {
    if(train$title[i] == "Master")
    {
      train$Age[i] = Master_age
    }
    else if(train$title[i] == "Miss")
    {
      train$Age[i] = Miss_age
    }
    else if(train$title[i] == "Mr")
```

```
    {
      train$Age[i] = Mr_age
    }
    else if(train$title[i] == "Mrs")
    {
      train$Age[i] = Mrs_age
    }
    else if(train$title[i] == "Other")
    {
      train$Age[i] = Other_age
    }
    else
    {
      train$Age[i] = Rev_age
    }
  }
}
```

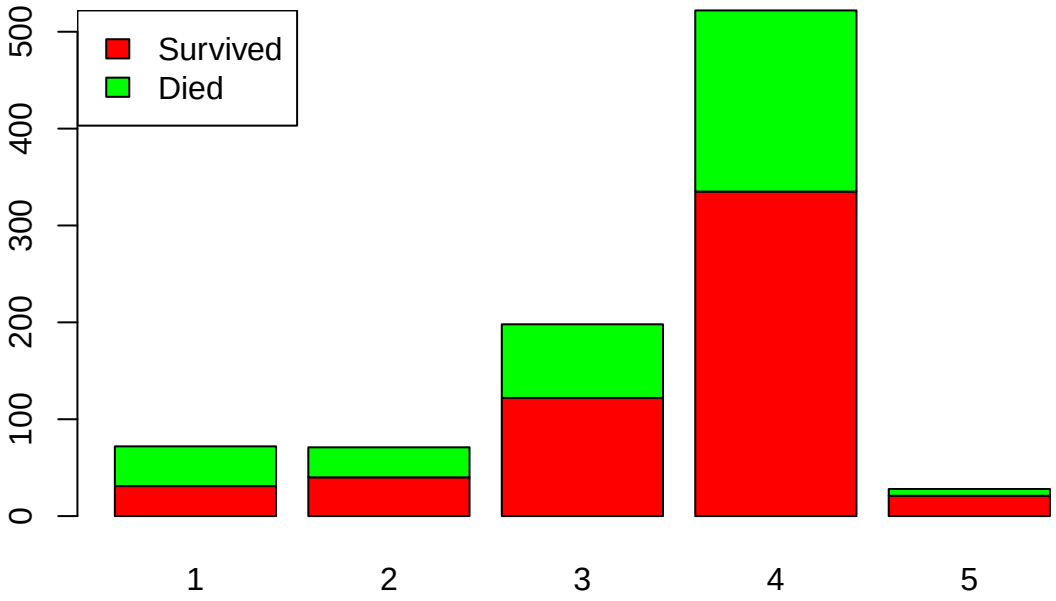
Now that we have a more accurate estimate of the ages, lets take a look at how they do in terms of survival.

```
index <- which(train$Survived == 1)
boxplot(train$Age[index], train$Age[-index], names = c("Survived", "Died"))
```



It seems that the people that survived had a slightly lower age then those who lived. But the picture is not clear. Lets try binning the ages and see if that can clarify what is happening.

```
barplot(table(train$Survived,train$agebin), col = rainbow(3))
legend("topleft", fill=rainbow(3), legend=c("Survived", "Died"))
```

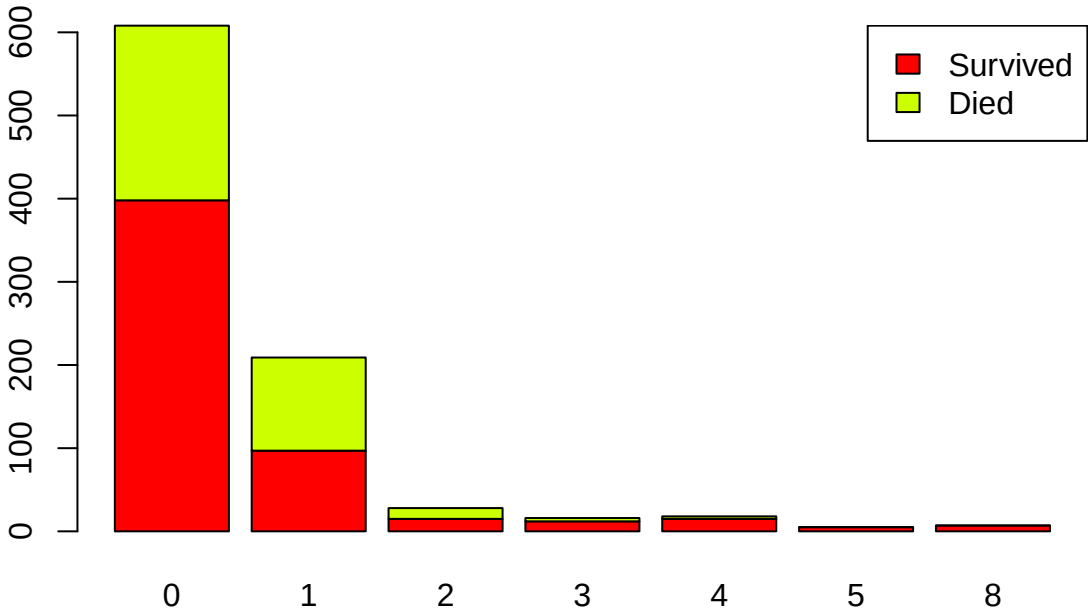


This is still not looking very useful, however. Age may have to be dropped.

SibSp

This variable represents the number of siblings that a person had on board. My intuition tells me that larger families will be able to get off of the boat more easily, but Lets see what the data has to say.

```
barplot(table(train$Survived,train$SibSp), col = rainbow(5))
legend("topright", fill=rainbow(5), legend=c("Survived", "Died"))
```



My intuition seems to be wrong. Lone passangers seem to be better off than passangers in a family unit. However dark that implication may be, it will give our model valuable insight.

Parch

Parch measures the number of parents or children that were abord the ship. Since SibSp surprised me by implying being alone was the best for surviving, I suspect that Parch will agree, but lets take a look at the numbers first.

Models

Results

```
#https://www.tutorialspoint.com/r/r_mean_median_mode.htm
#https://stackoverflow.com/questions/9981929/how-to-display-all-x-labels-in-r-barplot
#https://www.kaggle.com/tysonni/extracting-passenger-titiles-in-r
#https://www.kaggle.com/nadintamer/titanic-survival-predictions-beginner
#https://stat.ethz.ch/R-manual/R-devel/library/stats/html/ksmooth.html
```