Math 6620 Project

Benjamin Smith and Omesh Chandan

2/25/2021

# Introduction

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there were not enough lifeboats for everyone on board, resulting in the death of 1502 out of 2224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this report we will explore the data set on the passengers of the Titanic and determine what are the key factors for predicting passenger survival. After this we will proceed with showing a developed model and examine the necessary diagnostics.

# The Data

The data set is an openly available on [Kaggle.com](https://www.kaggle.com/c/titanic/overview) as one of their available competitions. However, make this analysis as reproducible as possible we have downloaded the the data with the readr package from the gist posted by Micheleen Harris ([here](https://gist.github.com/michhar/2dfd2de0d4f8727f873422c5d959fff5)).

The data set consists of 12 variables. The data dictionary is posted with the [data description on the Kaggle site](https://www.kaggle.com/c/titanic/data). For convenience it will be written here to provide a description each variable.

## Data Dictionary

* PassengerId : Id assigned to each passenger
* Survived: Survival 0 = No, 1 = Yes
* Pclass: Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd
* Name: Name of the passenger
* Sex: Sex of the passenger
* Age: Age of the passenger 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

# Exploratory Data Analysis

Before constructing a model, we should examine the relationships present between the various variables, survival and their various interactions. With this we can choose which parameters would be suitable. With this in mind the following visuals are created as part of the EDA to provide insight on what variables would be suitable for constructing a model for predicting passenger survival.

## Questions

The questions we are interesting in answering with this exploratory data analysis are:

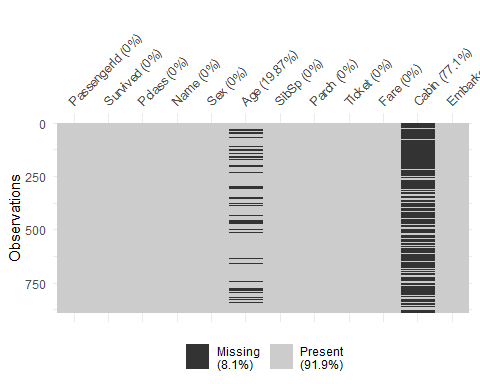
1. How much data is missing in this dataset?
2. Is there a higher probability of survival depending on a passenger’s sex?
3. Is there a higher probability of survival depending on ticket class?
4. Is age a determining factor for survival?
5. Does a passenger having an immediate family member (i.e. a spouse, sibling(s), parent(s) and/or children) with them improve their probability for survival?
6. How does the interaction of these variables relate to survival on the Titanic?

## Missing Data

Before examining the data it is important to first see what data is missing. Even if there is no treatment done to the missing data, awareness of missing data for a given variable is important to know about before adding it as a variable to our model.

Using the naniar package we are able to visualize the proportion of data missing in each variable field.

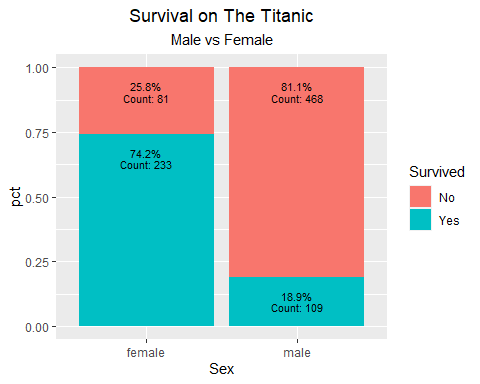
## # A tibble: 12 x 3  
## variable n\_miss pct\_miss  
## <chr> <int> <dbl>  
## 1 Cabin 687 77.1   
## 2 Age 177 19.9   
## 3 Embarked 2 0.224  
## 4 PassengerId 0 0   
## 5 Survived 0 0   
## 6 Pclass 0 0   
## 7 Name 0 0   
## 8 Sex 0 0   
## 9 SibSp 0 0   
## 10 Parch 0 0   
## 11 Ticket 0 0   
## 12 Fare 0 0



In the entire data set, 8.1% of the data is missing. Missing data is present in the Cabin (77.1%), Age (19.87%) and Embarked (0.22%) fields. While it is possible to use the data present in Cabin if we understand its context (see [here](https://titanic.fandom.com/wiki/First_Class_Staterooms) and [here](https://www.encyclopedia-titanica.org/cabins.html)) it will require some data wrangling and classification which is beyond the scope of this report.

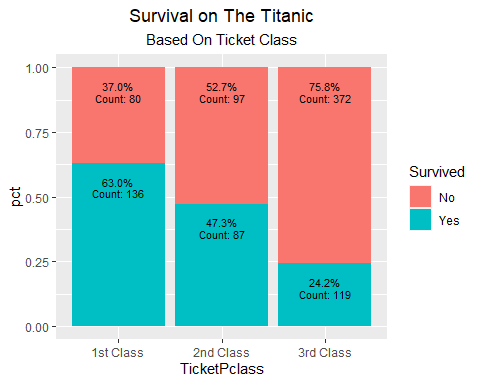
With the missing data acknowledged, lets will proceed with looking at the relationships with various variables and passenger survival.

## Sex of Passenger and Survival



Overall, women on the Titanic were 3.9 times more likely to survive than men. Men on the Titanic were 3.1 times more likely to die than women.

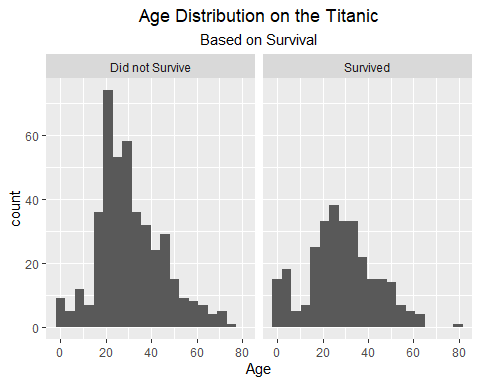
## Ticket Class of Passenger and Survival

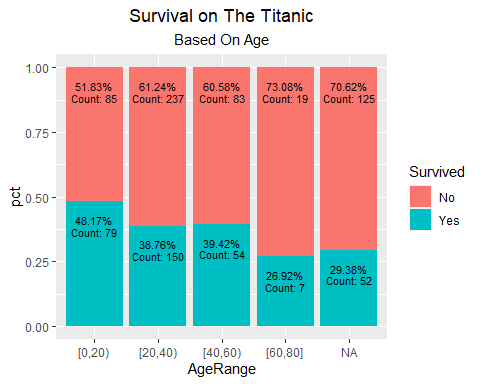


Passengers in first class had the highest probability of surviving followed by those in second and third class. Passengers in first class were 1.3 times more likely to survive than passengers in second class and were 2.6 times more likely to survive than passengers in third class.

Passengers in third class were 1.4 times more likely to die than passengers on second class and were over 2 times more likely to die than passengers in first class.

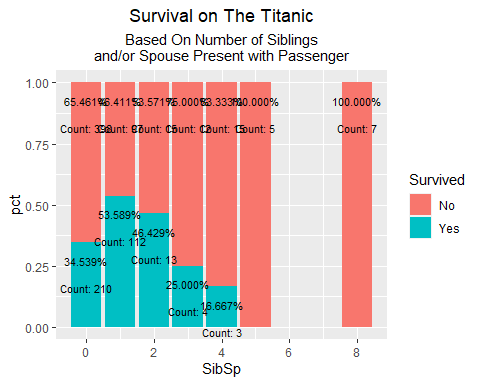
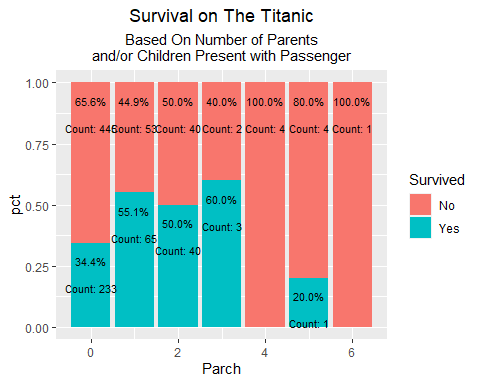
## Age and Survival

 Looking at the distribution of age alone. There does not seem to be any outstanding relationship between age and survival. This is because age is being treated as a continuous variable. Lets investigate the relationship between survival and age if we were to bin the ages.



With this binning we see that passengers under the age of 20 gad the largest probability of survival (46.17%) followed by passengers in the range of 40-59 (39.42%) and 20-40 (39.42). It is with this binning that age will be included in the model.

## Number of passenger relatives and survival.

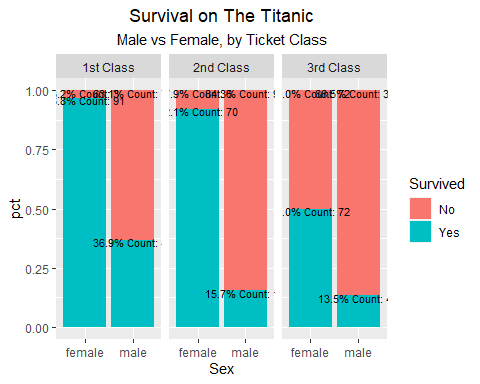
 Based on both these charts having/not having a spouse, sibling parent or child with an individual has an effect on the probability of survival. Individuals who came alone and with 1-3 parents/children had a higher probability of survival than those who came with more. However the number of individuals with 3-6 parents and/or children with them is small, however because the numbers for individuals who came alone and with 1-2 parents/children is significant for the data set, it is worthwhile to consider this in a model.

Similarly with presence of siblings and/or a spouse individuals who came alone or had 1-4 siblings and/or a spouse with them had a higher probability than those who came with more (however the numbers for people who came with more than 5 siblings and/or a spouse is small). In particular those who came with one other person had the highest probability of survival.

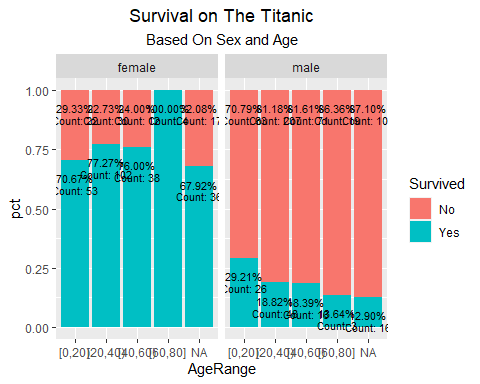
## Interaction of Variables

### Sex of Passenger, Ticket class and Survival

Because there are significant differences in probability of survival based on the sex and ticket class of a passenger independently it makes sense to explore their interaction as well.

 The probability of survival for women on the Titanic in first and second class was over 90 percent. However male passengers in first class were 19.7 times more likely to die than than females in the same class. Men in 3rd class were 1.37 times more likely to die than men in first class. Probability of survival for men in first class is more than double that of men in second and third class.

### Sex and Age of Passenger and Survival.

 Looking at the interaction of the binned age, women had an overall higher probability of suvival than men. Men who were younger had a larger probability of survival than men who were older.

### Sex of Passenger,

# THIS IS A MESS HERE DON’T LOOK

# This is altered from original  
survivalProp <- df %>% group\_by(Sex,Pclass,Survived) %>%   
 summarize(count=n()) %>%   
 mutate(pct=count/sum(count),Count=count)

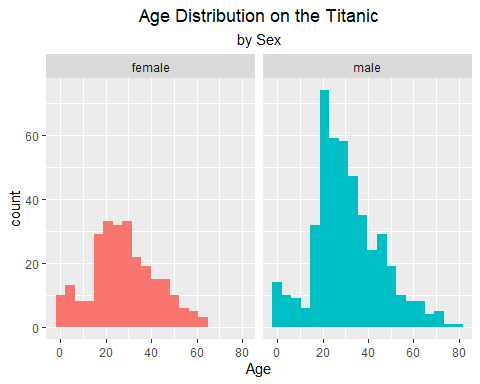
## `summarise()` regrouping output by 'Sex', 'Pclass' (override with `.groups` argument)

SibSpProp<-df %>% group\_by(SibSp,Sex,Survived)%>%   
 summarize(count=n()) %>%   
 mutate(pct=count/sum(count),Count=count)

## `summarise()` regrouping output by 'SibSp', 'Sex' (override with `.groups` argument)

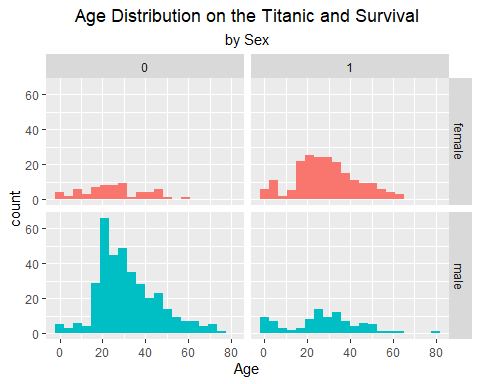
ggplot(data=df,mapping=aes(x=Age,fill=Sex))+  
 geom\_histogram(bins=20)+  
 facet\_grid(~Sex)+  
 labs(title="Age Distribution on the Titanic",subtitle = "by Sex")+  
 theme(legend.position = "none")+  
 theme(plot.title = element\_text(hjust=0.5),plot.subtitle = element\_text(hjust=0.5))

## Warning: Removed 177 rows containing non-finite values (stat\_bin).



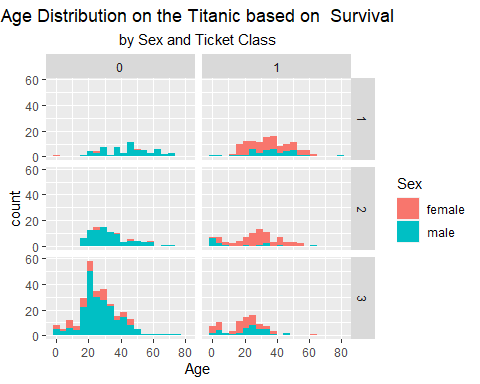
ggplot(data=df,mapping=aes(x=Age,fill=Sex))+  
 geom\_histogram(bins=20)+  
 facet\_grid(Sex~Survived)+  
 labs(title="Age Distribution on the Titanic and Survival",subtitle = "by Sex")+  
 theme(legend.position = "none")+  
 theme(plot.title = element\_text(hjust=0.5),plot.subtitle = element\_text(hjust=0.5))

## Warning: Removed 177 rows containing non-finite values (stat\_bin).



ggplot(data=df,mapping=aes(x=Age,fill=Sex))+  
 geom\_histogram(bins=20)+  
 facet\_grid(Pclass~Survived)+  
 labs(title="Age Distribution on the Titanic based on Survival",subtitle = "by Sex and Ticket Class")+  
 theme(plot.title = element\_text(hjust=0.5),plot.subtitle = element\_text(hjust=0.5))

## Warning: Removed 177 rows containing non-finite values (stat\_bin).



ggplot(data=SibSpProp,mapping=aes(x=SibSp,y=pct,fill=as.factor(Survived)))+  
 geom\_bar(stat="identity",position="fill")+  
 facet\_grid(~Sex)+  
 geom\_text(mapping= aes(label=Count),position="stack",size=3, vjust=+1.5)+  
 scale\_fill\_manual(name="Survived",  
 values=c("#F8766D","#00BFC4"),  
 labels=c("No","Yes"))+  
 labs(title="Survival on The Titanic",subtitle = "Based On Number of Siblings and/or Spouse Present with Passenger and Sex")+  
 theme(plot.title = element\_text(hjust=0.5),plot.subtitle = element\_text(hjust=0.5))

