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

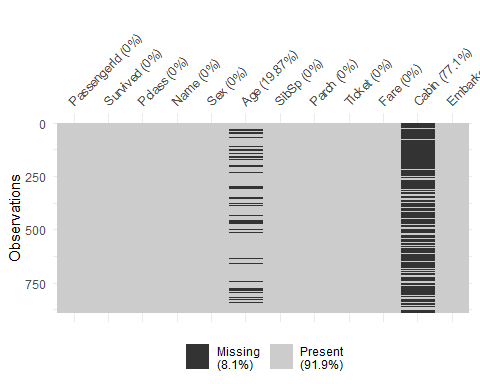
Before constructing a model, we should examine the relationships present between the various variables, survival and their various interactions. With this we can choose which parameters would be suitable. With this in mind the following visuals are created as part of the EDA to provide insight on what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

### 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.

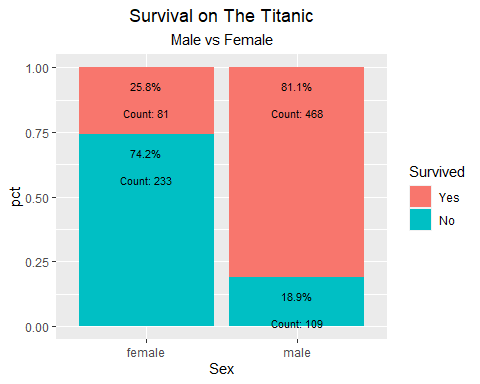
With the missing data acknowledged, lets will proceed with looking at the relationships with various variables and passenger survival.

### Questions

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

1. Is there a higher probability of survival depending on a passenger’s sex?
2. Is there a higher probability of survival depending on ticket class?
3. Is age a determining factor for survival?
4. Does having a spouse, sibling(s), parent(s) and/or children with a passenger improve their probability for survival?
5. How does the interaction of these variables relate to survival on the Titanic?

### 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.

# THIS IS A MESS HERE DON’T LOOK

# Exploratory Data Analysis

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

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

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

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

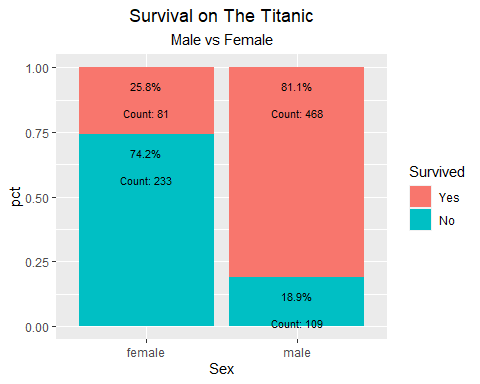
ParchProp<-df %>% group\_by(Parch,Survived)%>%   
 summarize(count=n()) %>%   
 mutate(pct=count/sum(count),Count=count)

## `summarise()` regrouping output by 'Parch' (override with `.groups` argument)

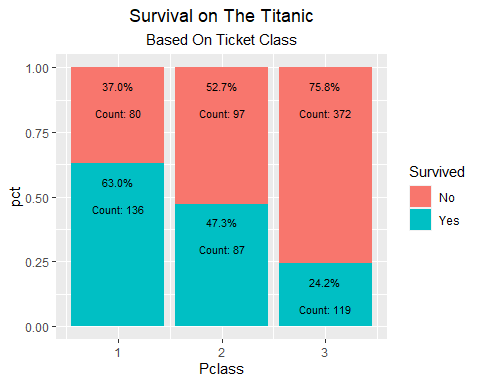
SibSpProp<-df %>% group\_by(SibSp,Survived)%>%   
 summarize(count=n()) %>%   
 mutate(pct=count/sum(count),Count=count)

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

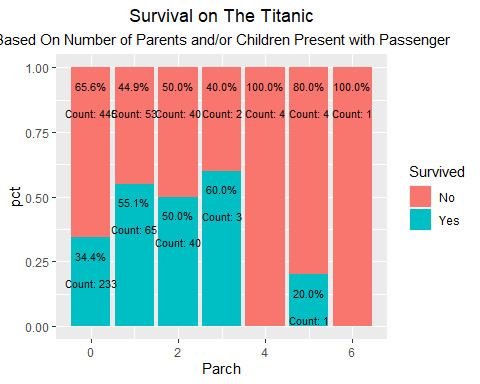
ggplot(data=survivalProp,mapping=aes(x=Sex,y=pct,fill=as.factor(Survived)))+  
 geom\_bar(stat="identity",position="fill")+  
 geom\_text(mapping= aes(label=scales::percent(pct)),position="stack",size=3, vjust=+2.5)+  
 geom\_text(mapping= aes(label=paste("Count:",Count)),position="stack",size=3, vjust=+5.5)+  
 scale\_fill\_manual(name="Survived",  
 values=c("#F8766D","#00BFC4"),  
 labels=c("Yes","No"))+  
 labs(title="Survival on The Titanic",subtitle = "Male vs Female")+  
 theme(plot.title = element\_text(hjust=0.5),plot.subtitle = element\_text(hjust=0.5))



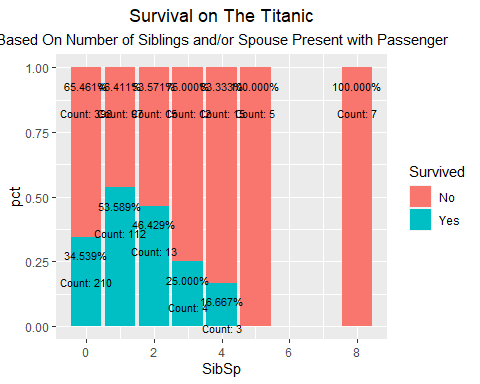
ggplot(data=PclassProp,mapping=aes(x=Pclass,y=pct,fill=as.factor(Survived)))+  
 geom\_bar(stat="identity",position="fill")+  
 geom\_text(mapping= aes(label=scales::percent(pct)),position="stack",size=3, vjust=+2.5)+  
 geom\_text(mapping= aes(label=paste("Count:",Count)),position="stack",size=3, vjust=+5.5)+  
 scale\_fill\_manual(name="Survived",  
 values=c("#F8766D","#00BFC4"),  
 labels=c("No","Yes"))+  
 labs(title="Survival on The Titanic",subtitle = "Based On Ticket Class")+  
 theme(plot.title = element\_text(hjust=0.5),plot.subtitle = element\_text(hjust=0.5))



ggplot(data=ParchProp,mapping=aes(x=Parch,y=pct,fill=as.factor(Survived)))+  
 geom\_bar(stat="identity",position="fill")+  
 geom\_text(mapping= aes(label=scales::percent(pct)),position="stack",size=3, vjust=+2.5)+  
 geom\_text(mapping= aes(label=paste("Count:",Count)),position="stack",size=3, vjust=+5.5)+  
 scale\_fill\_manual(name="Survived",  
 values=c("#F8766D","#00BFC4"),  
 labels=c("No","Yes"))+  
 labs(title="Survival on The Titanic",subtitle = "Based On Number of Parents and/or Children Present with Passenger")+  
 theme(plot.title = element\_text(hjust=0.5),plot.subtitle = element\_text(hjust=0.5))



ggplot(data=SibSpProp,mapping=aes(x=SibSp,y=pct,fill=as.factor(Survived)))+  
 geom\_bar(stat="identity",position="fill")+  
 geom\_text(mapping= aes(label=scales::percent(pct)),position="stack",size=3, vjust=+2.5)+  
 geom\_text(mapping= aes(label=paste("Count:",Count)),position="stack",size=3, vjust=+5.5)+  
 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")+  
 theme(plot.title = element\_text(hjust=0.5),plot.subtitle = element\_text(hjust=0.5))



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

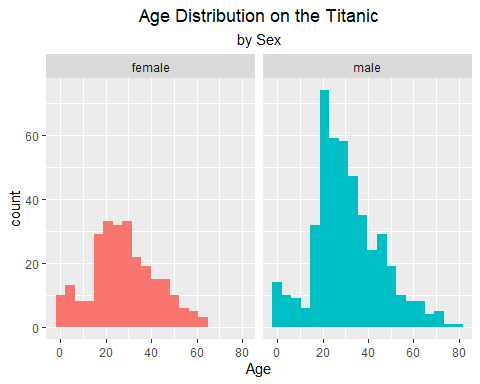
## `summarise()` regrouping output by 'Sex', 'Pclass' (override with `.groups` argument)

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)

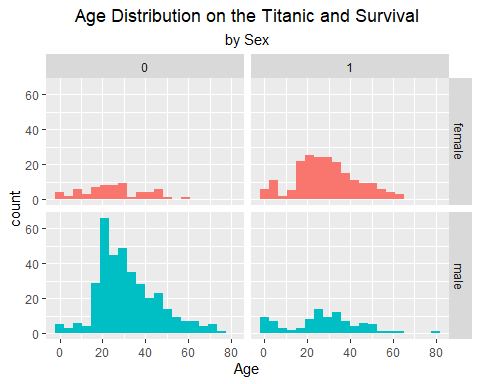
ggplot(data=df,mapping=aes(x=Age,fill=Sex))+  
 geom\_histogram(bins=20)+  
 facet\_grid(~Sex)+  
 labs(title="Age Distribution on the Titanic",subtitle = "by Sex")+  
 theme(legend.position = "none")+  
 theme(plot.title = element\_text(hjust=0.5),plot.subtitle = element\_text(hjust=0.5))

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



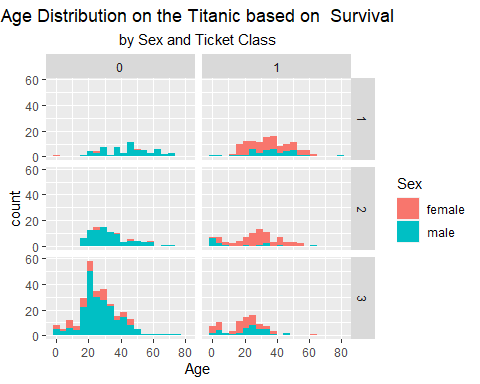
ggplot(data=df,mapping=aes(x=Age,fill=Sex))+  
 geom\_histogram(bins=20)+  
 facet\_grid(Sex~Survived)+  
 labs(title="Age Distribution on the Titanic and Survival",subtitle = "by Sex")+  
 theme(legend.position = "none")+  
 theme(plot.title = element\_text(hjust=0.5),plot.subtitle = element\_text(hjust=0.5))

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

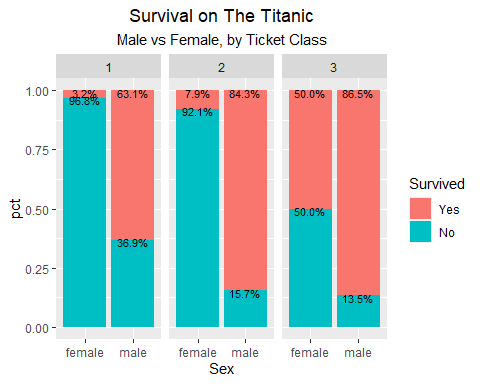


ggplot(data=df,mapping=aes(x=Age,fill=Sex))+  
 geom\_histogram(bins=20)+  
 facet\_grid(Pclass~Survived)+  
 labs(title="Age Distribution on the Titanic based on Survival",subtitle = "by Sex and Ticket Class")+  
 theme(plot.title = element\_text(hjust=0.5),plot.subtitle = element\_text(hjust=0.5))

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



ggplot(data=survivalProp,mapping=aes(x=Sex,y=pct,fill=as.factor(Survived)))+  
 geom\_bar(stat="identity",position="fill")+  
 facet\_grid(~Pclass)+  
 geom\_text(mapping= aes(label=scales::percent(pct)),position="stack",size=3, vjust=0.75)+  
 scale\_fill\_manual(name="Survived",  
 values=c("#F8766D","#00BFC4"),  
 labels=c("Yes","No"))+  
 labs(title="Survival on The Titanic",subtitle = "Male vs Female, by Ticket Class")+  
 theme(plot.title = element\_text(hjust=0.5),plot.subtitle = element\_text(hjust=0.5))



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))

