Final Project

Ziyun Wang

12/15/2018

Introduction and Background

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.

Data Sources & Variables

Data Source:

Kaggle (https://www.kaggle.com/c/titanic/data)

Response Variable:

Survival (0 = died, 1 = survived)

Expanatory Variables:

Name: Name of Passenger

Pclass: Ticket Class of Travel (1 = First, 2 = Second, 3 = Third)

Age: Passenger Age in years (continuous)

Sex: Gender (Male & Female)

Sibsp: Number of Sibing/Spouse on aboard (continuous)

Parch: Number of Parent/Child on aboard (continuous)

Ticket: Ticket Number

Fare: Price of the Ticket (continuous)

Cabin: Cabin Number where Passenger Stayed

Embarked: Port in which Passenger Embarked (C - Cherbourg, S - Southampton, Q = Queenstown)

Question of Interest

How do factors such as Class, Gender, Age, Title, Number of Family Members on aboard affect one's chance of survival on Titanic?

Data Analysis

Data and R setup

```
library(ggplot2, quietly = TRUE)
library(dplyr, quietly = TRUE)
install.packages('ggthemes',repos = "http://cran.us.r-project.org")
```

```
##
##   There is a binary version available but the source version is
##   later:
##          binary source needs_compilation
## ggthemes  3.4.0  4.0.1             FALSE
```

```
library(ggthemes, quietly = TRUE)
install.packages('corrplot',repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/nl/h3gc994s3cv2mhd_11yv8zh80000gn/T//RtmpxWCmi0/downloaded_packages
```

```
library(corrplot, quietly = TRUE)
titanic <- read.csv("titanic.csv", header = T, stringsAsFactors = F)
```

Summary of Dataset

```
head(titanic)[1:5,]
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
##                                                   Name    Sex Age SibSp
## 1                              Braund, Mr. Owen Harris   male  22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
## 3                               Heikkinen, Miss. Laina female  26     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
## 5                             Allen, Mr. William Henry   male  35     0
##   Parch           Ticket    Fare Cabin Embarked
## 1     0        A/5 21171  7.2500              S
## 2     0         PC 17599 71.2833   C85        C
## 3     0 STON/O2. 3101282  7.9250              S
## 4     0           113803 53.1000  C123        S
## 5     0           373450  8.0500              S
```

```r
summary(titanic)
```

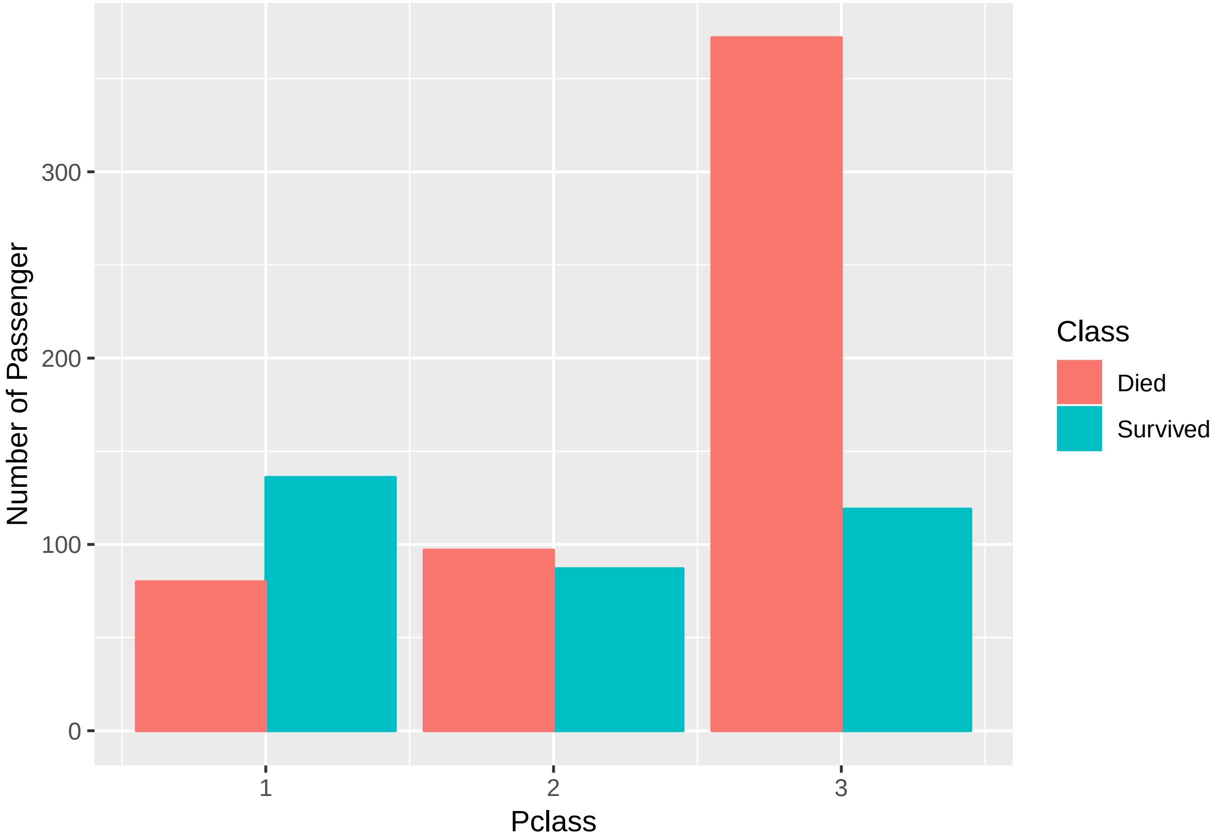
```
##   PassengerId       Survived          Pclass          Name          
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891        
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character  
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character  
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309                     
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000                     
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000                     
##                                                                     
##      Sex                 Age            SibSp           Parch       
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000  
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000  
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000  
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816  
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000  
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000  
##                     NA's   :177                                     
##     Ticket               Fare            Cabin             Embarked        
##  Length:891         Min.   :  0.00   Length:891         Length:891        
##  Class :character   1st Qu.:  7.91   Class :character   Class :character  
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character  
##                     Mean   : 32.20                                        
##                     3rd Qu.: 31.00                                        
##                     Max.   :512.33                                        
##
```

Number of Survivals with One Factor Considered

```r
survival_factor <- function(df, variable)
{
  survive <- df %>%
    mutate(Class = ifelse(Survived == 0, 'Died', 'Survived')) %>%
     group_by_(variable, 'Class') %>%
    summarise(count = n())

  p <- ggplot(survive, aes(y=survive$count, x=survive[[variable]],
                           color=survive$Class,
                           fill=survive$Class))
  p <- p + geom_bar(stat="identity", position="dodge") + labs(color="Class", fill="Class", y="Number of
  print(p)
}
```

Number of Survivals with Class that Passengers Stayed in

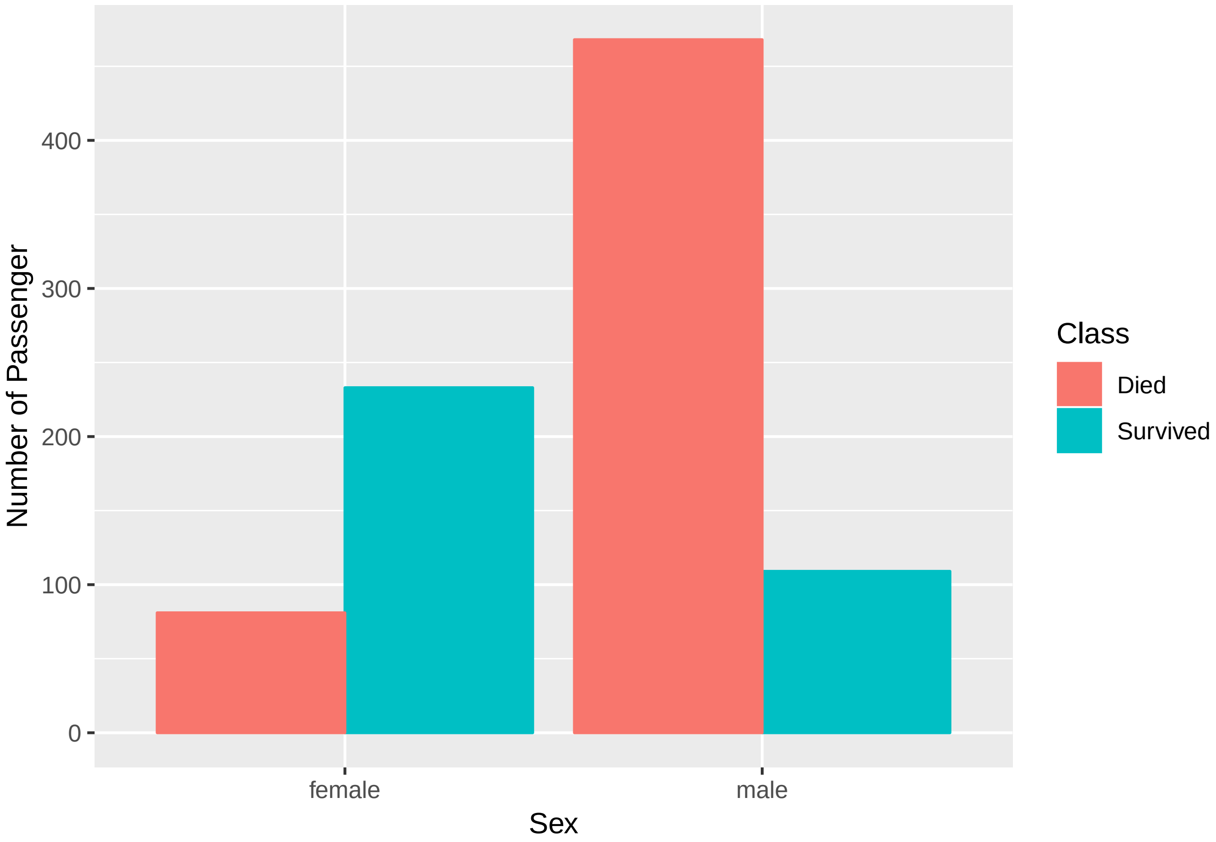
```
survival_factor(titanic, 'Pclass')
```



Insight As we can see from the graph, passengers who stayed in the third class had a much lower chance to survive compared with the first and secons class passengers.

Number of Survivals with Gender of Passengers

```r
survival_factor(titanic, 'Sex')
```



Insight As we can see from the graph, female passengers had a higher chance to survive than male passengers on Titanic. This might result from the cultural behavior, which is allowing women children to leave first when something bad happens.

Average Ticket Price of Each Class

```r
class_fare <- function(df) {
  m <- rep(NA, 3)
  for (i in 1:length(m)) {
    class_df <- dplyr::filter(df, Pclass == i)
    m[i] <- mean(class_df$Fare)
  }
  fare_df <- data.frame(Class=1:3, Avg.Fare=m, stringsAsFactors = F)

  return (fare_df)
}

class_fare(titanic)
```

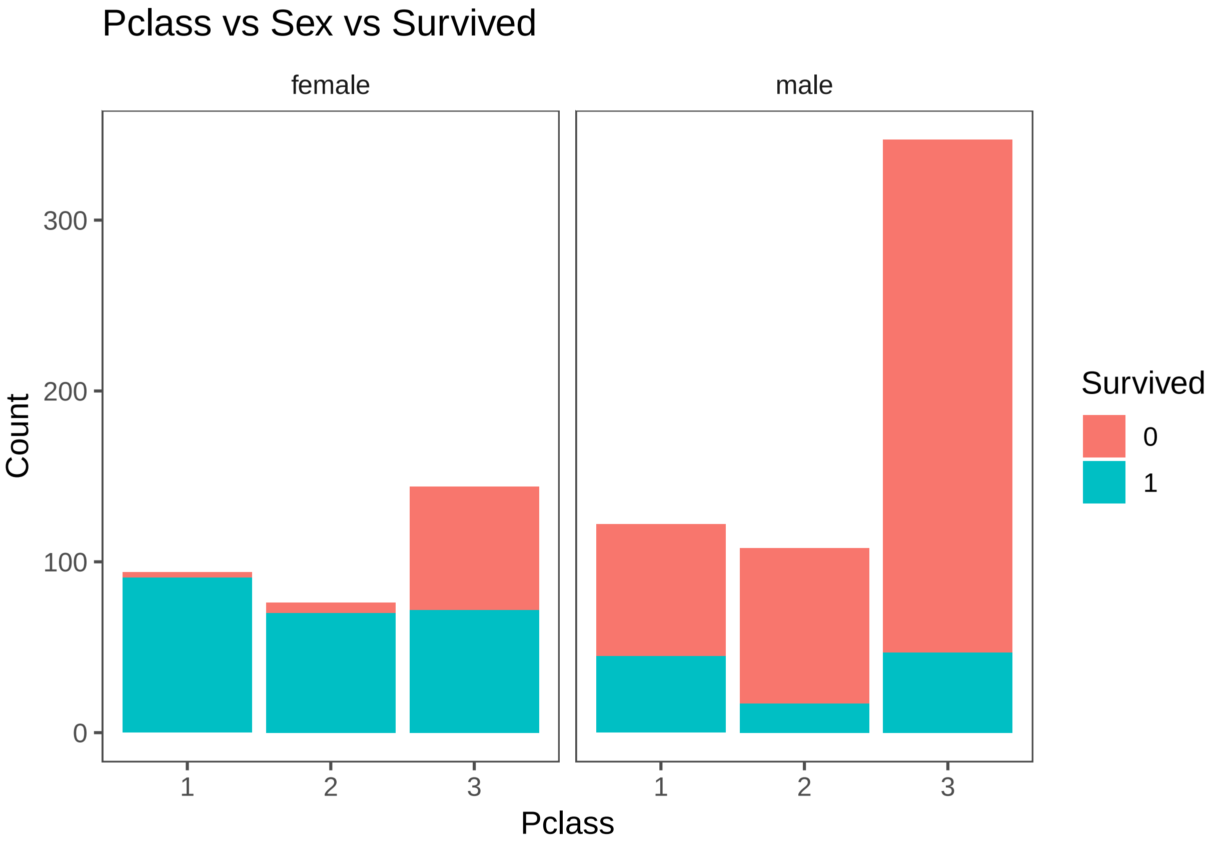
```
##   Class Avg.Fare
## 1     1 84.15469
## 2     2 20.66218
## 3     3 13.67555
```

Number of Survivals with Two Factor Considered

```r
survival_two_factors <- function(df, f1, f2) {
  ggplot(df, aes(df[[f1]], fill = factor(df$Survived))) +
    geom_bar(stat = "count")+
    theme_few() +
    xlab("Pclass") +
    facet_grid(.~df[[f2]])+
    ylab("Count") +
    scale_fill_discrete(name = "Survived") +
    ggtitle("Pclass vs Sex vs Survived")
}
```

Number of Survivals with Gender of Passengers and Class Stayed In

```r
survival_two_factors(titanic, 'Pclass', 'Sex')
```



Insight

From the graph with two factors considered, we can say that most females from the first and second class suvived. Sadly, most males in the thrid class died in the shipwreck of Titanic.

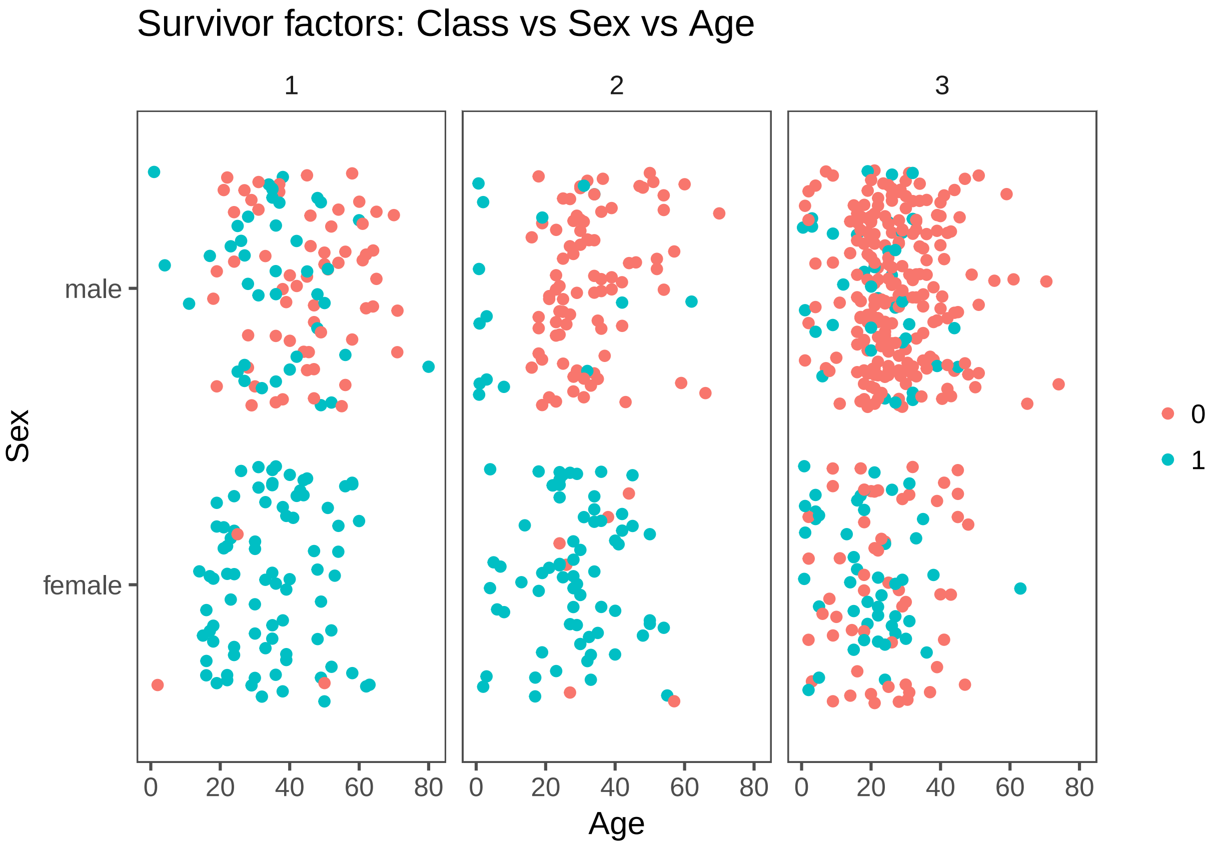
Number of Survivals with Three Factor Considered

```
survival_three_factors <- function(df, Age, Sex, Pclass){
  p <- ggplot(df, aes(x = Age, y = Sex)) +
    geom_jitter(aes(colour = factor(Survived))) +
    theme_few()
  p <- p + theme(legend.title = element_blank())+
    facet_wrap(~Pclass)
  p <- p + labs(x = "Age", y = "Sex", title = "Survivor factors: Class vs Sex vs Age")
  p <- p + scale_fill_discrete(name = "Survived") +
    scale_x_continuous(name="Age",limits=c(0, 81))
  print(p)
}
```

Number of Survivals with Gender and Age of Passenger, and Class Stayed In

```
survival_three_factors(titanic, 'Age', 'Sex', 'Pclass')
```

```
## Warning: Removed 177 rows containing missing values (geom_point).
```

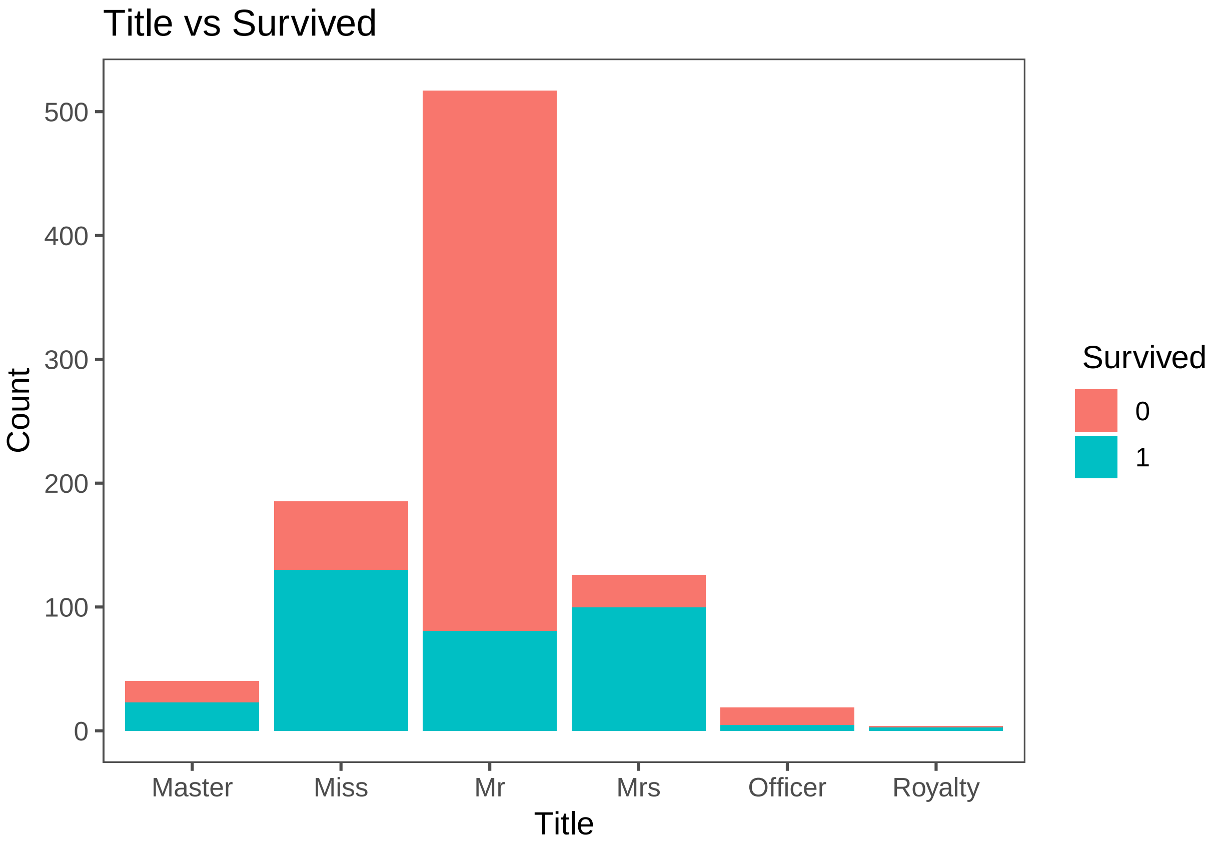


Insight

In this graph, which age factor was incorporated, it gives us an idea how age was affecting the survival chance.

Number of Survivals with the Title of the Passenger

```
survival_title <- function(df) {
  df$Title <- gsub('(.*, )|(\\..*)', '', df$Name)

  officer <- c('Capt', 'Col', 'Don', 'Dr', 'Major', 'Rev')
  royalty <- c('Dona', 'Lady', 'the Countess','Sir', 'Jonkheer')

  # Reassign titles
  df$Title[df$Title == 'Mlle']        <- 'Miss'
  df$Title[df$Title == 'Ms']          <- 'Miss'
  df$Title[df$Title == 'Mme']         <- 'Mrs'
  df$Title[df$Title %in% royalty]  <- 'Royalty'
  df$Title[df$Title %in% officer]  <- 'Officer'

  df$Surname <- sapply(df$Name,
                       function(x) strsplit(x, split = '[,.]')[[1]][1])
  p <- ggplot(df, aes(Title,fill = factor(Survived))) +
    geom_bar(stat = "count")+ xlab('Title') + ylab("Count")
  p <- p + scale_fill_discrete(name = " Survived") +
    ggtitle("Title vs Survived")+
    theme_few()
  print(p)
}
survival_title(titanic)
```

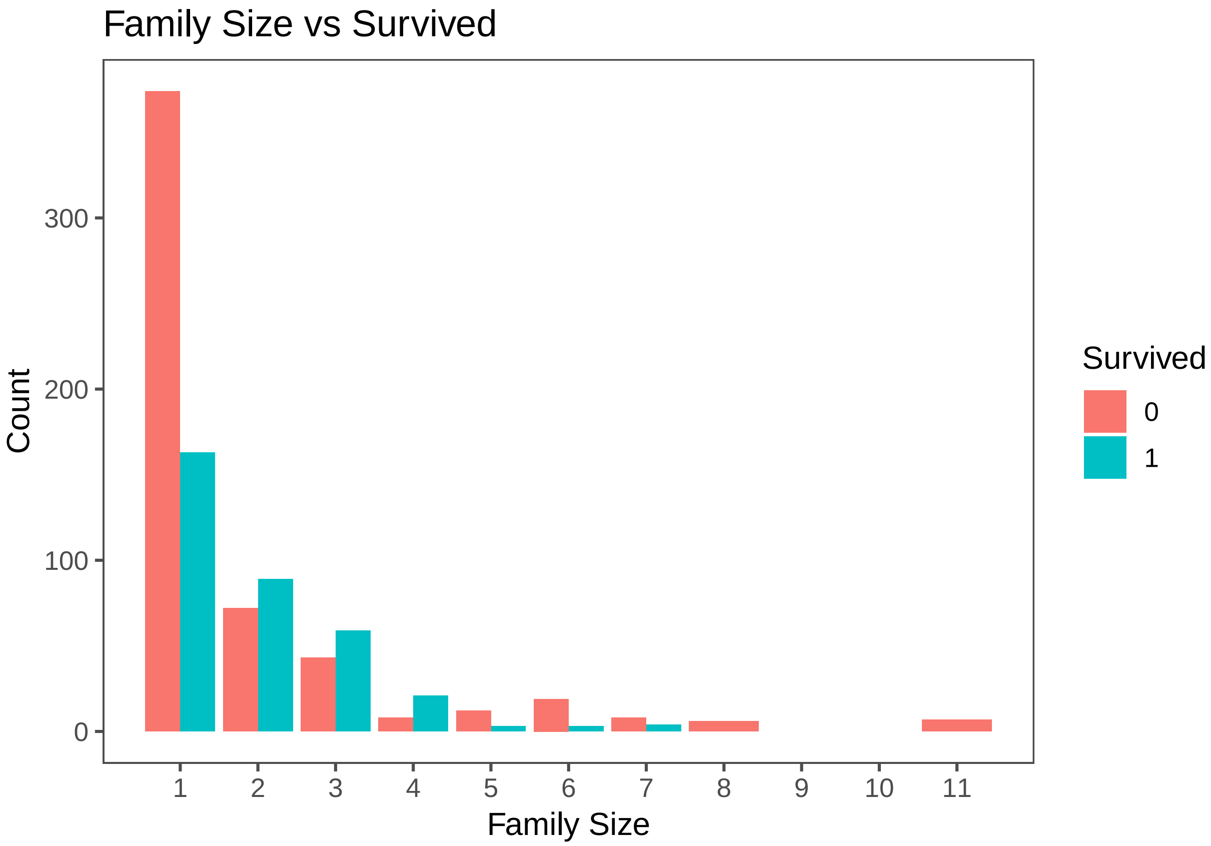


Insight

In this graph, we have a rough idea on how one's title affects one's chance to survive from the shipwreck of Titanic. All passengers who were royal memebers survived. Also, married women had a high chance to survive, which might because their children were with them and so they were identified as the first priority to leave the ship.

Number of Survivals with the Family Size of the Passenger

```
survival_familysize <- function(df) {
  df$Fsize <- df$SibSp + df$Parch + 1

  p <- ggplot(df, aes(x = Fsize, fill = factor(Survived))) +
    geom_bar(stat='count', position='dodge')
  p <- p + scale_x_continuous(breaks=c(1:11)) + xlab('Family Size') +
    ylab("Count")
  p <- p + theme_few()+scale_fill_discrete(name = "Survived")
  p <- p + ggtitle("Family Size vs Survived")
  print(p)
}
survival_familysize(titanic)
```



Insight

From this graph, it shows us how number of family members one has would affect one's chance of survival from the shipwreck. People with more than 4 family members on aboard with them were less likely to survive, which might becuase they used more time looking for their family while the tragedy happened. However, people has no family members also had a lower chance to survive. Surpirsingly, people with one to two family members had the highest chance to survive among others.

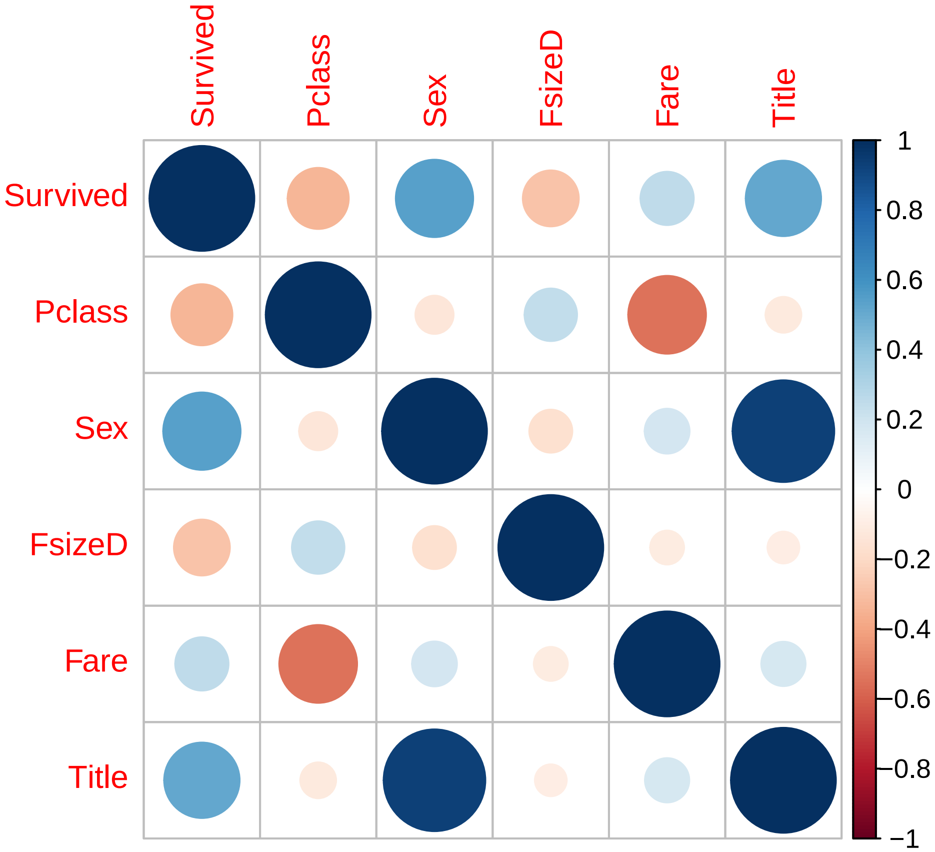
Correlation Maytix of Number of Survivals and All factors

```
#detach("package:dplyr", unload=TRUE)
library(plyr)
```

```
## -------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## -------------------------------------------------------------------------
##
## Attaching package: 'plyr'
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
survival_correlation <- function(df) {
  df <- titanic
  df$Fsize <- df$SibSp + df$Parch + 1
  df$FsizeD[df$Fsize == 1] <- 'Alone'
  df$FsizeD[df$Fsize < 5 & df$Fsize > 1] <- 'Small'
  df$FsizeD[df$Fsize > 4] <- 'Big'
  df$Title <- gsub('(.*, )|(\\..*)', '', df$Name)

  officer <- c('Capt', 'Col', 'Don', 'Dr', 'Major', 'Rev')
  royalty <- c('Dona', 'Lady', 'the Countess','Sir', 'Jonkheer')

  # Reassign titles
  df$Title[df$Title == 'Mlle']        <- 'Miss'
  df$Title[df$Title == 'Ms']          <- 'Miss'
  df$Title[df$Title == 'Mme']         <- 'Mrs'
  df$Title[df$Title %in% royalty]  <- 'Royalty'
  df$Title[df$Title %in% officer]  <- 'Officer'

  corr_data <- df
  corr_data$Sex <- revalue(corr_data$Sex,
                           c("male" = 1, "female" = 2))
  corr_data$Title <- revalue(corr_data$Title,
                             c("Mr" = 1, "Master" = 2,"Officer" = 3,
                               "Mrs" = 4,"Royalty" = 5,"Miss" = 6))
  corr_data$FsizeD <- revalue(corr_data$FsizeD,
                              c("Small" = 1, "Alone" = 2, "Big" = 3))
  corr_data$FsizeD <- as.numeric(corr_data$FsizeD)
  corr_data$Sex <- as.numeric(corr_data$Sex)
  corr_data$Title <- as.numeric(corr_data$Title)
  corr_data$Pclass <- as.numeric(corr_data$Pclass)
  corr_data$Survived <- as.numeric(corr_data$Survived)
  corr_data <-corr_data[,c("Survived", "Pclass", "Sex",
                           "FsizeD", "Fare", "Title")]

  str(corr_data)
  mcorr_data <- cor(corr_data)

  corrplot(mcorr_data,method="circle")
}

survival_correlation(titanic)
```

```
## 'data.frame':    891 obs. of  6 variables:
##  $ Survived: num  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass  : num  3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : num  1 2 2 2 1 1 1 1 2 2 ...
##  $ FsizeD  : num  1 1 2 1 2 2 2 3 1 1 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Title   : num  1 4 6 4 1 1 1 2 4 4 ...
```



Conclusion

To sum up, the passenger's gender, title and the class stayed in play influencial roles in determining tht passenger's chance to survive from the shipwreck of the Titanic.