Titanic Data Analysis

Ahmed Ashraf Mohamed

Contents

1 Preliminary Look at the data

We need first to define the data we have.

Variable	Definition	Key
survival	Survival	0 = No, 1 = yes
pclass	ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	sex	
age	Age in year	
sibsp	Number of siblings/spouses aboard the titanic	
parch	Number of parents/children aboard the Titanic	
ticket	ticket number(unique)	
fare	Passenger fare	
cabin	Cabin number	
embarked	port of embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

```
# Loading Packages
library(tidyverse)
library(viridis)
library(ggplot2)
library(ggcorrplot)
library(ggthemes)
library(hrbrthemes)
library(e1071)
library(mice)
library(statsr)
```

```
# Loading Data
train <- read_csv("data/train.csv")
test <- read_csv("data/test.csv")
```

2 Exploration Of The Data

2.1 Summary of Data

```
summary(train)
```

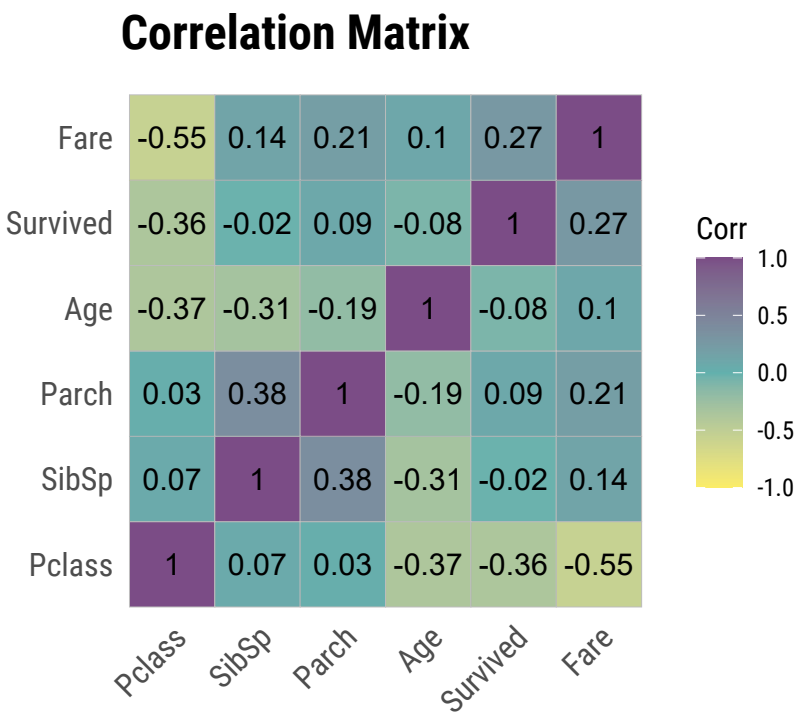
```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   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
##
```

2.2 Plotting The Data

2.2.1 Correlation Matrix (numerical analysis)

We are going to use correlation matrix of the numerical data to assess the correlation, which might gives a better idea of which feature might be important

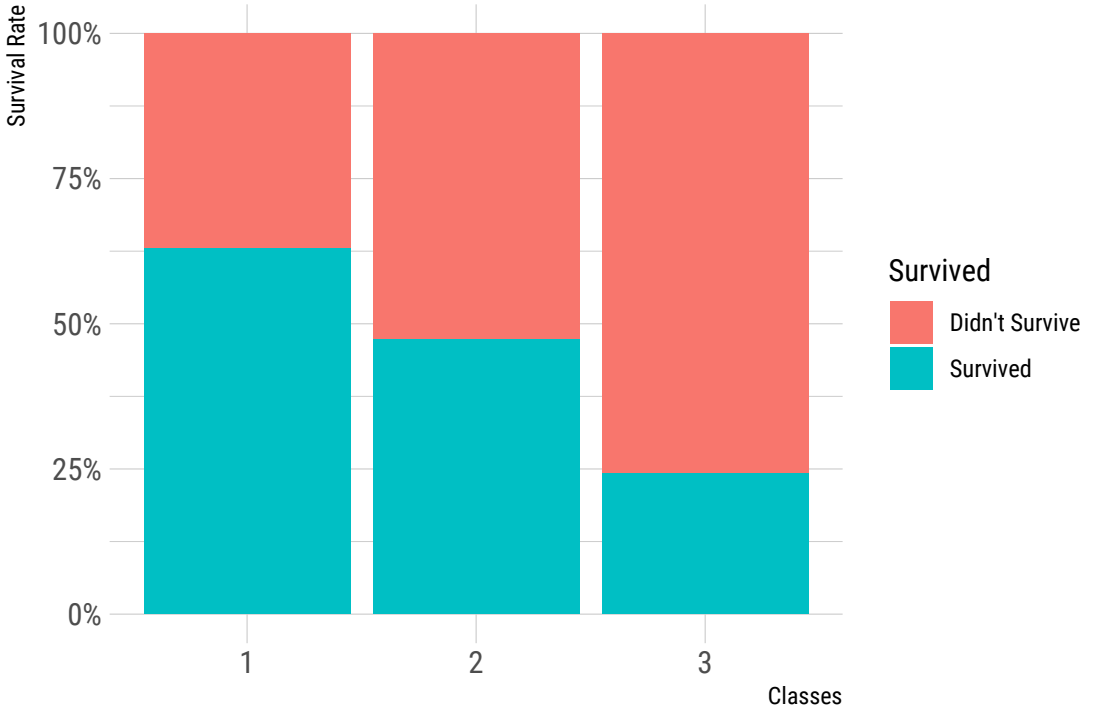
```
train %>%
  filter(!is.na(Age)) %>%
  select(Survived, Pclass,Age,SibSp,Parch,Fare) %>%
  cor() %>%
  ggcorrplot(lab = T,
             ggtheme =theme_ipsum_rc(grid = F),
             title="Correlation Matrix",hc.order=T,
             colors =rev(viridis(3,alpha=0.7)),
             digits = 2)
```



The fare features seems to be the most correlated feature to survival of the passengers, but it doesn't negate the importance of the other features in the data. Which means that we will start by comparing the each that we consider to be important against survival feature

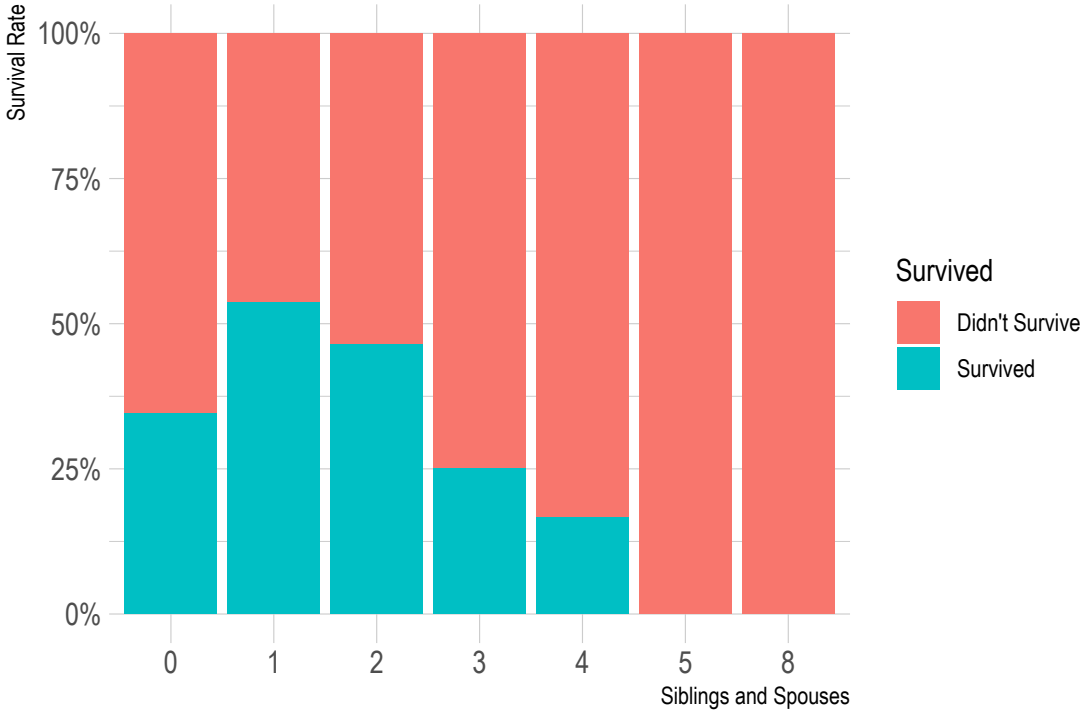
2.2.2 Class of Passenger

```
train %>%
  select(Pclass,Survived) %>%
  ggplot(aes(as_factor(Pclass),fill=as_factor(Survived))) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels=scales::percent) +
  theme_ipsum_rc() +
  labs(x = "Classes",y = "Survival Rate")+
  scale_fill_discrete(name = "Survived", labels = c("Didn't Survive","Survived"))
```



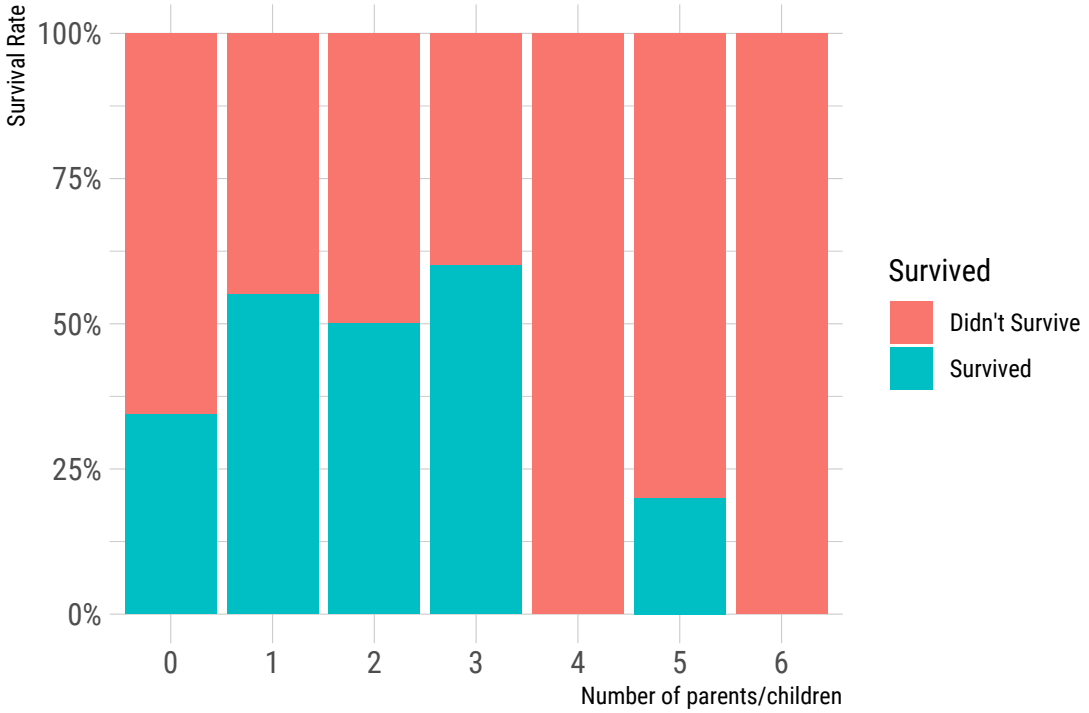
2.2.3 Siblings and Spouses

```
train %>%
  select(SibSp,Survived) %>%
  ggplot(aes(as_factor(SibSp),fill=as_factor(Survived))) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Siblings and Spouses",y = "Survival Rate")+
  scale_fill_discrete(name = "Survived", labels = c("Didn't Survive","Survived"))+
  theme_ipsum()
```



2.2.4 Number of children/parents

```
train %>%
  select(Parch,Survived) %>%
  ggplot(aes(as_factor(Parch),fill=as_factor(Survived))) +
  geom_bar(position = "fill") +
  scale_y_continuous(label = scales::percent)+
  labs(x = "Number of parents/children",y = "Survival Rate")+
  scale_fill_discrete(name = "Survived", labels = c("Didn't Survive","Survived")) +
  theme_ipsum_rc()
```




```
head(train)
```

```
## # A tibble: 6 x 12
##   PassengerId Survived Pclass Name    Sex     Age SibSp Parch Ticket  Fare Cabin
##         <dbl>    <dbl>  <dbl> <chr>   <chr> <dbl> <dbl> <dbl> <chr>  <dbl> <chr>
## 1           1        0      3 Braund~ male     22     1     0 A/5 2~  7.25 <NA>
## 2           2        1      1 Cuming~ fema~    38     1     0 PC 17~ 71.3  C85
## 3           3        1      3 Heikki~ fema~    26     0     0 STON/~  7.92 <NA>
## 4           4        1      1 Futrel~ fema~    35     1     0 113803 53.1  C123
## 5           5        0      3 Allen,~ male     35     0     0 373450  8.05 <NA>
## 6           6        0      3 Moran,~ male     NA     0     0 330877  8.46 <NA>
## # ... with 1 more variable: Embarked <chr>
```

```
train %>%
  group_by(Sex) %>%
  summarise(Age_mean = mean(Age,na.rm=TRUE),
            age_sd = sd(Age,na.rm=T),
            surival_mean = mean(Survived,na.rm =T),
            surival_sd = sd(Survived,na.rm = T))
```

```
## # A tibble: 2 x 5
##   Sex    Age_mean age_sd surival_mean surival_sd
##   <chr>     <dbl>  <dbl>        <dbl>      <dbl>
## 1 female     27.9   14.1        0.742      0.438
## 2 male       30.7   14.7        0.189      0.392
```

3 Description of the data

3.1 Categorical Features

3.2 Numerical Features