Titanic Challange

The famous data set from Kaggle

Alexander Nyberg

2021-04-01

```
library(tidyverse)
library(corrplot)
```

Load the data sets.

EDA

Training Data Analysis

```
glimpse(titanic_train)
```

```
## Rows: 891
## Columns: 12
## $ PassengerId <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
## $ Survived    <dbl> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1~
## $ Pclass      <dbl> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2, 3, 3~
## $ Name        <chr> "Braund, Mr. Owen Harris", "Cumings, Mrs. John Bradley (Fl~
## $ Sex         <chr> "male", "female", "female", "female", "male", "male", "mal~
## $ Age         <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14, 4, 58, 20, 39, 14, ~
## $ SibSp       <dbl> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4, 0, 1, 0~
## $ Parch       <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1, 0, 0, 0~
## $ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 3101282", "113803", "37~
## $ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, 51.8625,~
## $ Cabin       <chr> NA, "C85", NA, "C123", NA, NA, "E46", NA, NA, NA, "G6", "C~
## $ Embarked    <chr> "S", "C", "S", "S", "S", "Q", "S", "S", "S", "C", "S", "S"~
```

```
summary(titanic_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
```

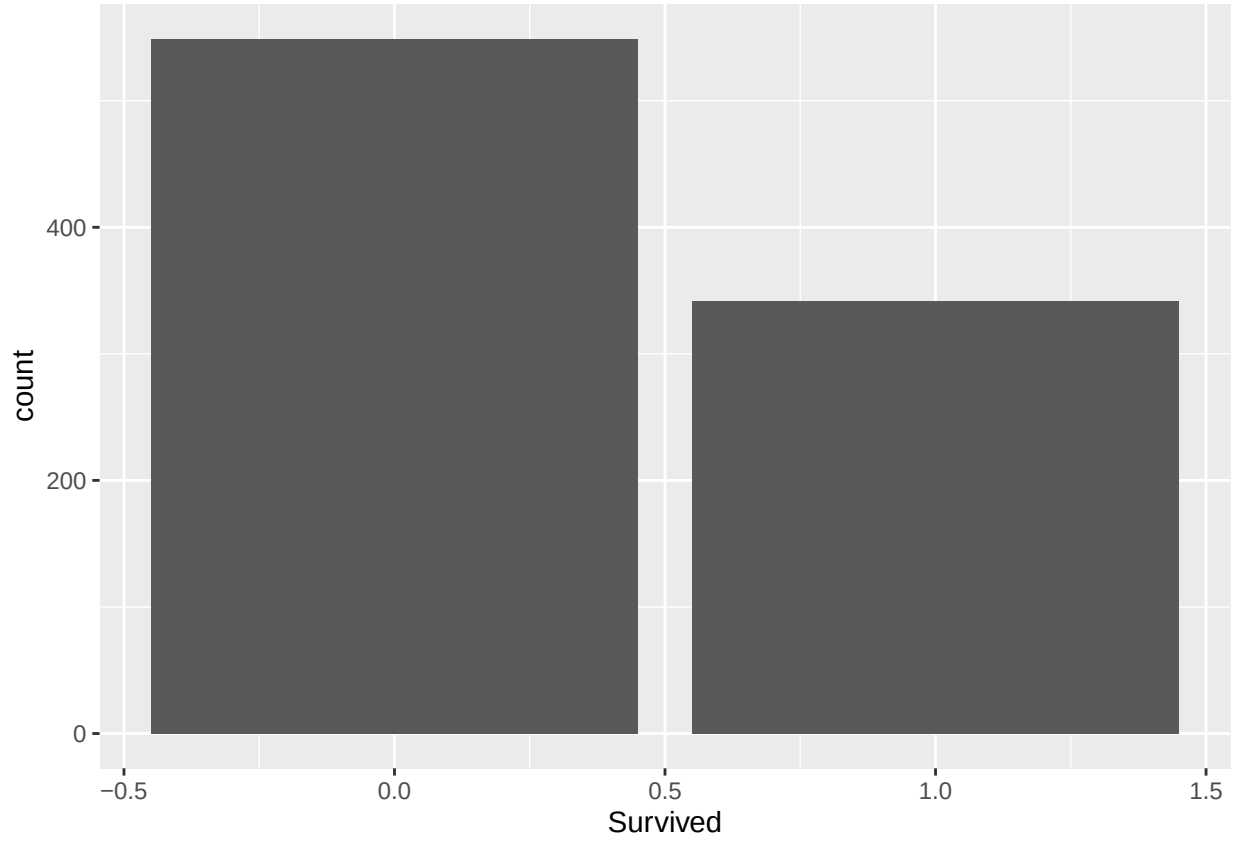
```
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

Mean age for training set is 29.70 for the passengers, youngest being 0.42 and oldest being 80. We also indicate 177 NA's in age. We can see that 38% of the people in the training set survived the disaster.

Plotting the survival rate.

```
SurvDeath <- titanic_train %>%
  select(Survived)

  ggplot() +
    geom_bar(data = SurvDeath, aes(x = Survived))
```



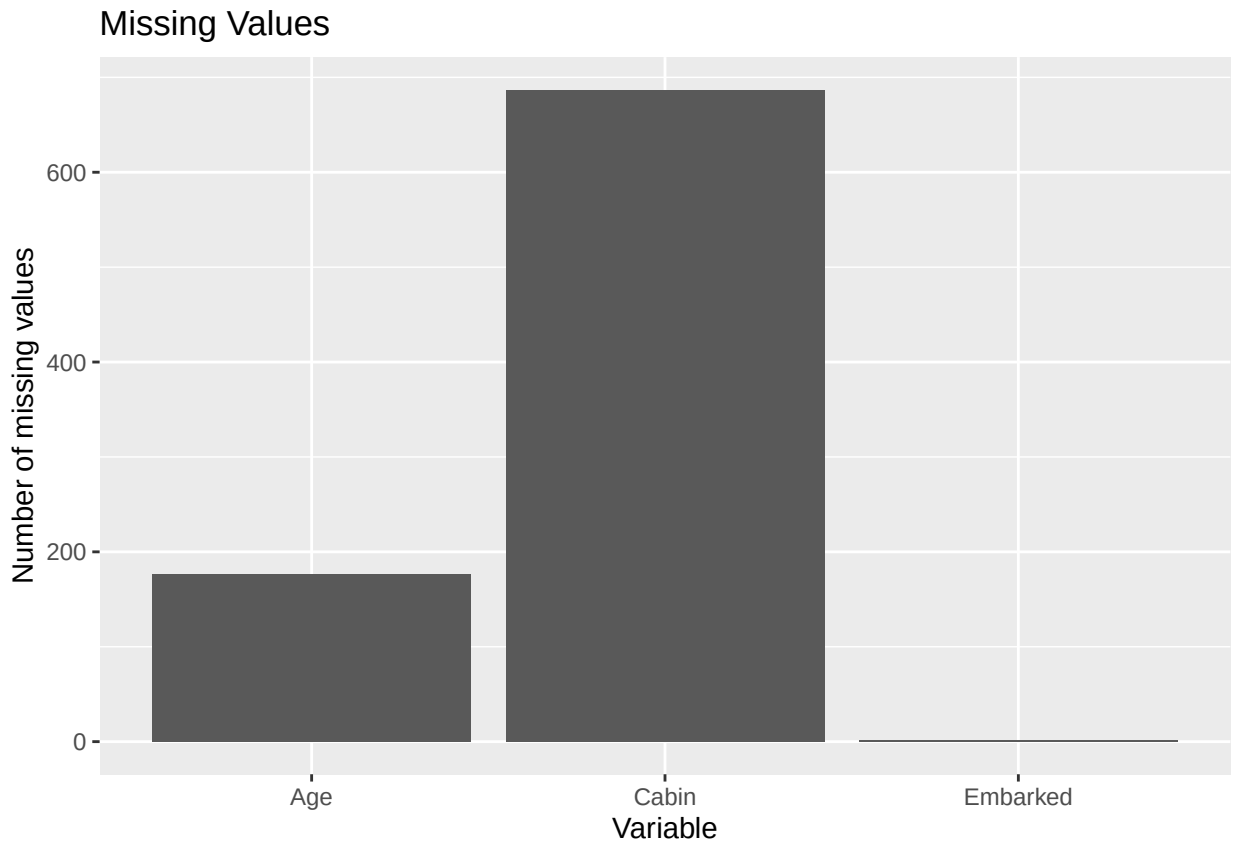

```
# Identifying the missing values
df_missValue <- data.frame(Key = character(1), Value = integer(1), Perc = integer(1))
for (i in 1:ncol(titanic_train)) {
  df_missValue <- rbind(df_missValue,
                        c(colnames(titanic_train[, i]),
                        sum(is.na(titanic_train[, i])),
```

```
                        sum(is.na(titanic_train[, i]))/nrow(titanic_train)))
}
df_missValue <- df_missValue %>%
  mutate(Value = as.integer(Value),
         Perc = as.double(Perc)) %>%
  filter(Value != 0) %>%
  arrange(desc(Value))

knitr::kable(df_missValue)
```

Key	Value	Perc
Cabin	687	0.7710438
Age	177	0.1986532
Embarked	2	0.0022447

```
df_missValue %>%
  ggplot() +
    geom_bar(aes(x = Key, y = Value), stat = 'identity') +
    labs(x = "Variable", y = "Number of missing values", title = "Missing Values")
```



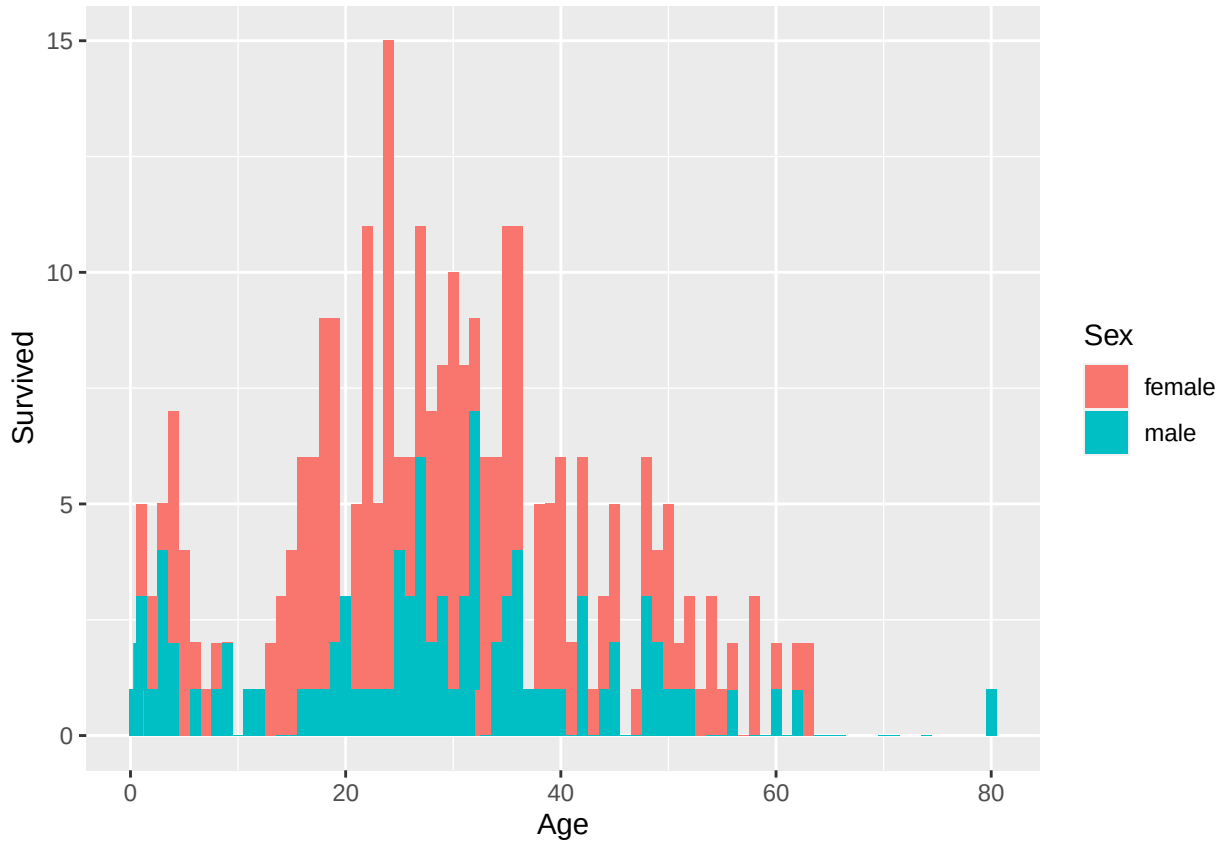

```
# Percentage of missing values
df_missValuePerc <- data.frame(Key = character(1), Value = integer(1), Tot = integer(1))
for (i in 1:ncol(titanic_train)) {
  df_missValuePerc <- rbind(df_missValuePerc,
                        c(colnames(titanic_train[, i]),
```

```
                        sum(is.na(titanic_train[, i])),
                        nrow(titanic_train[, i])))
}

df_missValuePerc %>%
  mutate(Value = as.integer(Value),
         Tot = as.integer(Tot)) %>%
  filter(Value != 0) %>%
  mutate(PercentageMissing = Value/Tot*100) %>%
  knitr::kable()
```

Key	Value	Tot	PercentageMissing
Age	177	891	19.8653199
Cabin	687	891	77.1043771
Embarked	2	891	0.2244669

```
titanic_train %>%
  select(Survived, Sex, Age) %>%
  group_by(Age) %>%
  ggplot() +
    geom_col(aes(x = Age, y = Survived, fill = Sex), width = 1)
```



Majority of survivals is women.

```r
# Total % of females that survived
femalesAlive <- titanic_train %>%
  filter(Sex == "female" & Survived == 1)

totFemales <- titanic_train %>%
  filter(Sex == "female")

knitr::kable(nrow(femalesAlive)/nrow(totFemales),
             col.names = "SurvRateFemale")
```

SurvRateFemale
0.7420382

```r
# Total % males that survived
malesAlive <- titanic_train %>%
  filter(Sex == "male" & Survived == 1)

totMales <- titanic_train %>%
  filter(Sex == "male")

knitr::kable(nrow(malesAlive)/nrow(totMales),
             col.names = "SurvRateMale")
```

SurvRateMale
0.1889081

Test Data Analysis

In the test data, we can see that we are missing some *Age* values again. These will be fixed with the mean of the training data set. Additionally, we have a missing value in the *Fare* column as well.

```r
glimpse(titanic_test)
```

```
## Rows: 418
## Columns: 11
## $ PassengerId <dbl> 892, 893, 894, 895, 896, 897, 898, 899, 900, 901, 902, 903~
## $ Pclass      <dbl> 3, 3, 2, 3, 3, 3, 3, 2, 3, 3, 3, 1, 1, 2, 1, 2, 2, 3, 3, 3~
## $ Name        <chr> "Kelly, Mr. James", "Wilkes, Mrs. James (Ellen Needs)", "M~
## $ Sex         <chr> "male", "female", "male", "male", "female", "male", "femal~
## $ Age         <dbl> 34.5, 47.0, 62.0, 27.0, 22.0, 14.0, 30.0, 26.0, 18.0, 21.0~
## $ SibSp       <dbl> 0, 1, 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0~
## $ Parch       <dbl> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Ticket      <chr> "330911", "363272", "240276", "315154", "3101298", "7538",~
## $ Fare        <dbl> 7.8292, 7.0000, 9.6875, 8.6625, 12.2875, 9.2250, 7.6292, 2~
## $ Cabin       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "B45", NA,~
## $ Embarked    <chr> "Q", "S", "Q", "S", "S", "S", "Q", "S", "C", "S", "S", "S"~
```

```r
summary(titanic_test)
```

```
##   PassengerId         Pclass          Name               Sex
##  Min.   : 892.0   Min.   :1.000   Length:418         Length:418
##  1st Qu.: 996.2   1st Qu.:1.000   Class :character   Class :character
```

```
##  Median :1100.5   Median :3.000   Mode  :character   Mode  :character
##  Mean   :1100.5   Mean   :2.266
##  3rd Qu.:1204.8   3rd Qu.:3.000
##  Max.   :1309.0   Max.   :3.000
##
##       Age            SibSp            Parch           Ticket
##  Min.   : 0.17   Min.   :0.0000   Min.   :0.0000   Length:418
##  1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median :27.00   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   :30.27   Mean   :0.4474   Mean   :0.3923
##  3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :76.00   Max.   :8.0000   Max.   :9.0000
##  NA's   :86
##       Fare            Cabin             Embarked
##  Min.   :  0.000   Length:418         Length:418
##  1st Qu.:  7.896   Class :character   Class :character
##  Median : 14.454   Mode  :character   Mode  :character
##  Mean   : 35.627
##  3rd Qu.: 31.500
##  Max.   :512.329
##  NA's   :1
```

Full Data Set

Firstly, we might consider the specific variables that actually might effect death probability. For now, we'll drop PassengerId, Name and Ticket. We'll also drop the Cabin column since 77% of the data is missing.

```
fullData <- titanic_train %>%
  full_join(titanic_test) %>%
  select(-PassengerId, -Name, -Ticket, -Cabin)

knitr::kable(fullData[1:15, ])
```

Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	male	22	1	0	7.2500	S
1	1	female	38	1	0	71.2833	C
1	3	female	26	0	0	7.9250	S
1	1	female	35	1	0	53.1000	S
0	3	male	35	0	0	8.0500	S
0	3	male	NA	0	0	8.4583	Q
0	1	male	54	0	0	51.8625	S
0	3	male	2	3	1	21.0750	S
1	3	female	27	0	2	11.1333	S
1	2	female	14	1	0	30.0708	C
1	3	female	4	1	1	16.7000	S
1	1	female	58	0	0	26.5500	S
0	3	male	20	0	0	8.0500	S
0	3	male	39	1	5	31.2750	S
0	3	female	14	0	0	7.8542	S

Since we have two different outcomes in the *Sex* column, we can easily change these to numeric.

```
for (i in 1:nrow(fullData)) {
  if (fullData[i, "Sex"] == "male") {
```

```
    fullData[i, "Sex"] <- "0"
  }
  else{
    fullData[i, "Sex"] <- "1"
  }
}

fullData <- fullData %>%
  mutate(Sex = as.double(Sex))
```

Handling Missing Values

With the missing data in age, a good choice could be to find the mean or the median and fill the missing values since we dont have that many missing values. Same goes for embarked.

```
# Finding age mean
meanAge <- fullData %>%
  select(Age) %>%
  summarise(mean(Age, na.rm = TRUE)) %>%
  round(digits = 1)
```

```
# Fill meanAge in missing Age values
for (i in 1:nrow(fullData)) {
  if (is.na(fullData[i, "Age"]) == TRUE) {
    fullData[i, "Age"] <- meanAge
  }
}
```

```
# Fixing embarked data
for (i in 1:nrow(fullData)) {
  if (is.na(fullData[i, "Embarked"]))
    fullData[i, "Embarked"] <- "S"
}
```

```
# Fixing the missing data in Fare
meanFare <- fullData %>%
  select(Fare) %>%
  summarise(median(Fare, na.rm = T)) %>%
  round(digits = 1)

for (i in 1:nrow(fullData)) {
  if (is.na(fullData[i, "Fare"]) == TRUE) {
    fullData[i, "Fare"] <- meanFare
  }
}
```

```
# Double Checking no data missing
df_missValue <- data.frame(Key = character(1), Value = integer(1), Perc = integer(1))
for (i in 1:ncol(fullData)) {
  df_missValue <- rbind(df_missValue,
                        c(colnames(fullData[, i]),
                        sum(is.na(fullData[, i])),
                        sum(is.na(fullData[, i]))/nrow(fullData)))
}
df_missValue <- df_missValue %>%
  mutate(Value = as.integer(Value),
```

```
        Perc = as.double(Perc)) %>%
  filter(Value != 0) %>%
  arrange(desc(Value))

knitr::kable(df_missValue)
```

Key	Value	Perc
Survived	418	0.3193277

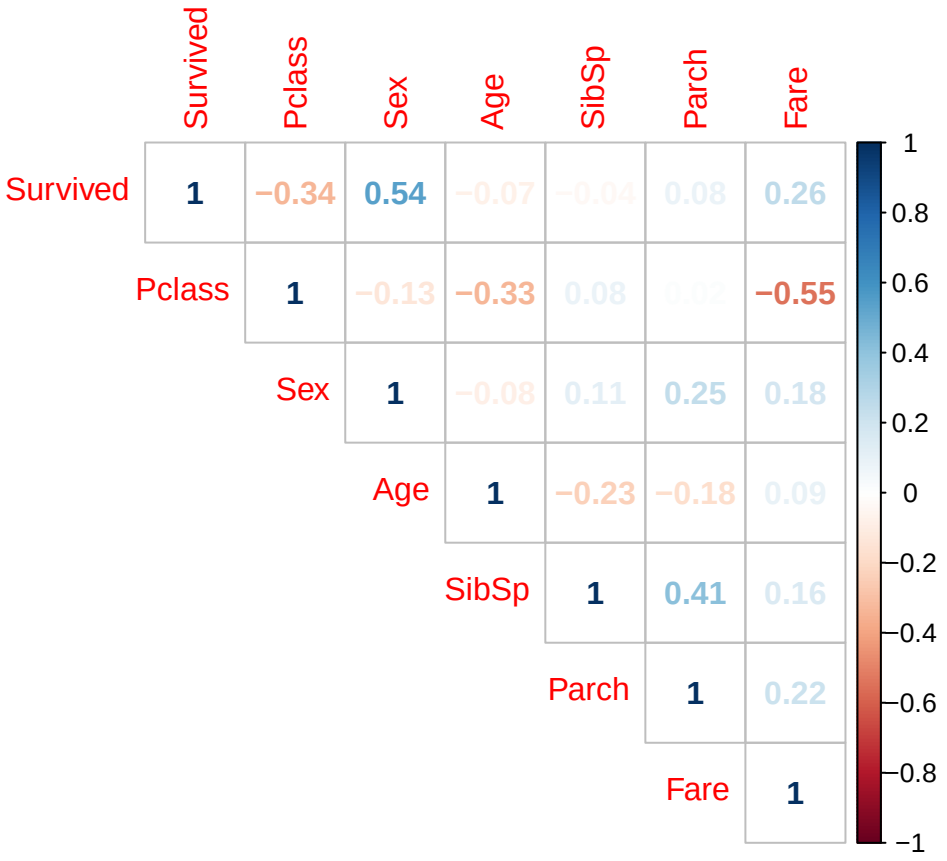
The data set is now complete and we are to predict the 418 test individuals.

Prediction

```
xTrain <- fullData %>%
  filter(!is.na(Survived))
xTest <- fullData %>%
  filter(is.na(Survived))
```

```
tCorr <- cor(xTrain[,-8])

corrplot(tCorr, method = "number", type = "upper")
```




```
# Linear Regression Prediction
```

```
LR <- lm(Survived ~ . - Embarked, data = xTrain)
summary(LR)
```

Linear Regression

```
##
## Call:
## lm(formula = Survived ~ . - Embarked, data = xTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0972 -0.2130 -0.0905  0.2345  0.9835
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.7871591  0.0705543  11.157  < 2e-16 ***
## Pclass      -0.1699054  0.0196752  -8.636  < 2e-16 ***
## Sex          0.5123422  0.0279333  18.342  < 2e-16 ***
## Age         -0.0058724  0.0010743  -5.466 5.97e-08 ***
## SibSp       -0.0433459  0.0130305  -3.326 0.000916 ***
## Parch       -0.0200171  0.0181160  -1.105 0.269484
## Fare         0.0004137  0.0003233   1.280 0.201044
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3798 on 884 degrees of freedom
## Multiple R-squared:  0.395,  Adjusted R-squared:  0.3908
## F-statistic: 96.17 on 6 and 884 DF,  p-value: < 2.2e-16
```

```
# Presenting the results
testSurv <- data.frame("SurvivedAlle" = round(predict(LR, xTest)))
testResult <- testSurv %>%
  mutate(PassengerId = titanic_test$PassengerId) %>%
  select(PassengerId, SurvivedAlle)
row.names(testResult) <- NULL

knitr::kable(head(testResult))
```

PassengerId	SurvivedAlle
892	0
893	0
894	0
895	0
896	1
897	0

```
# Comparison Results 99% Accuracy
compareData <- read_csv("../Data/submit.csv")
```

```
compareDF <- testResult %>%
  left_join(compareData, by = "PassengerId") %>%
  rename("Survived99PercentAcc" = Survived)
knitr::kable(head(compareDF))
```

PassengerId	SurvivedAlle	Survived99PercentAcc
892	0	0
893	0	1
894	0	0
895	0	0
896	1	1
897	0	0

```
identVector <- c()
for (i in 1:nrow(compareDF)) {
  if (compareDF[i, "SurvivedAlle"] == compareDF[i, "Survived99PercentAcc"]) {
    identVector <- append(identVector, T)
  }
  else{
    identVector <- append(identVector, F)
  }
}
compareDF$Alike <- identVector
```

```
compareDF %>%
  group_by(Alike) %>%
  count() %>%
  knitr::kable(col.names = c("Key", "Counts"))
```

Key	Counts
FALSE	16
TRUE	402