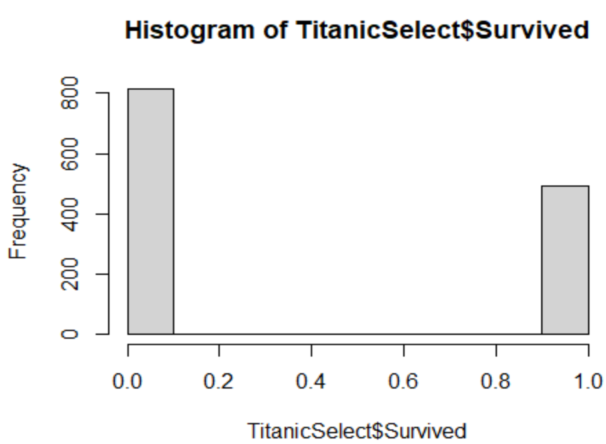
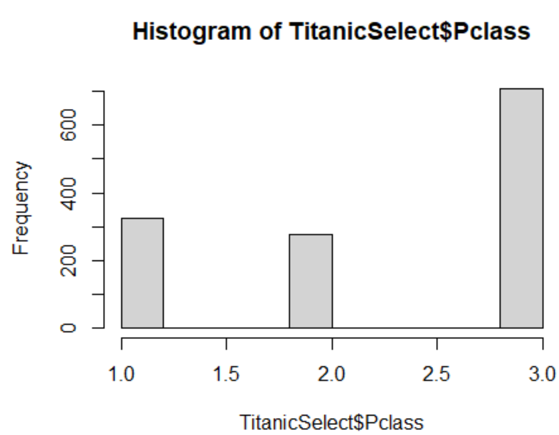
Prediction of Survival using Logistic Regression in R

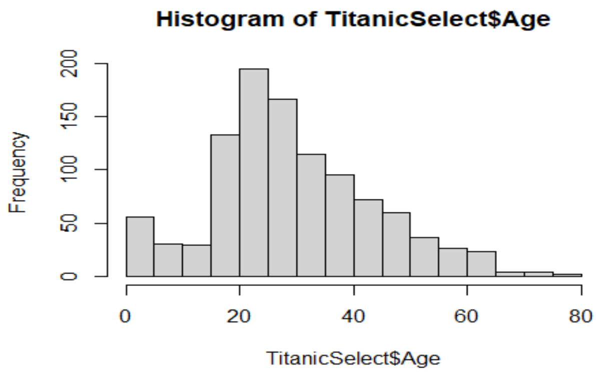
Data Source- Titanic - Machine Learning from Disaster | Kaggle

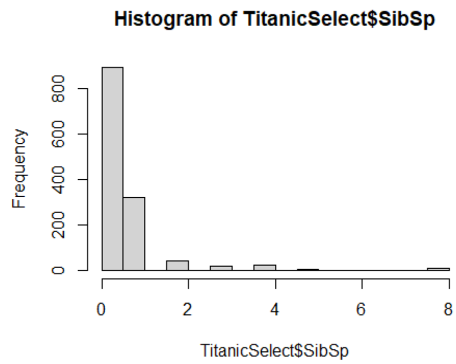
Libraries-(readxl)(tidyverse)(mi)(dplyr)(car)(readr)(ggplot2)(lattice)(caret)(GGally)(ROCR)(pROC)

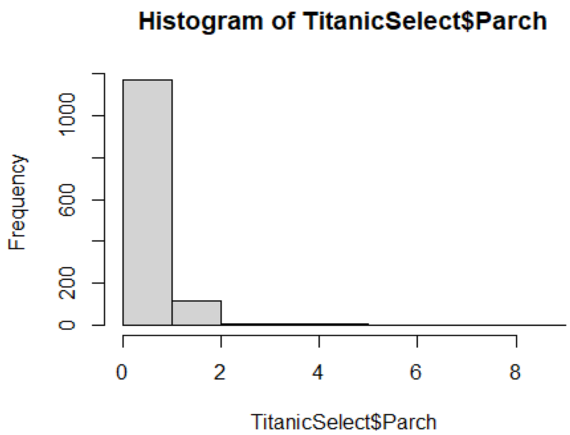
```
TitanicRaw <- read_excel("Titanic_data.xlsx")
nrow(TitanicRaw)
summary(TitanicRaw)
##   PassengerId     Survived        Pclass        Name
## Min.   :   1   Min.   :0.000   Min.   :1.00   Length:1309
## 1st Qu.: 328   1st Qu.:0.000   1st Qu.:2.00   Class :character
## Median : 655   Median :0.000   Median :3.00   Mode  :character
## Mean   : 655   Mean   :0.377   Mean   :2.29
## 3rd Qu.: 982   3rd Qu.:1.000   3rd Qu.:3.00
## Max.   :1309   Max.   :1.000   Max.   :3.00
##
##     Sex                Age            SibSp           Parch
## Length:1309        Min.   : 0.17   Min.   :0.000   Min.   :0.000
## Class :character   1st Qu.:21.00   1st Qu.:0.000   1st Qu.:0.000
## Mode  :character   Median :28.00   Median :0.000   Median :0.000
##                    Mean   :29.88   Mean   :0.499   Mean   :0.385
##                    3rd Qu.:39.00   3rd Qu.:1.000   3rd Qu.:0.000
##                    Max.   :80.00   Max.   :8.000   Max.   :9.000
##                    NA's   :263
##    Ticket              Fare          Cabin            Embarked
## Length:1309        Min.   :  0.0   Length:1309        Length:1309
## Class :character   1st Qu.:  7.9   Class :character   Class :character
## Mode  :character   Median : 14.4   Mode  :character   Mode  :character
##                    Mean   : 33.3
##                    3rd Qu.: 31.3
##                    Max.   :512.3
##                    NA's   :1
TitanicSelect <- select(TitanicRaw, c(2,3,4,5,6,7,8,10,12))
#Histogram for numerical features
hist(TitanicSelect$Survived)
```

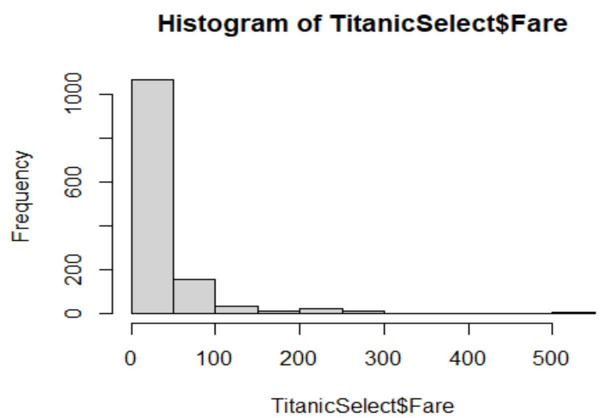




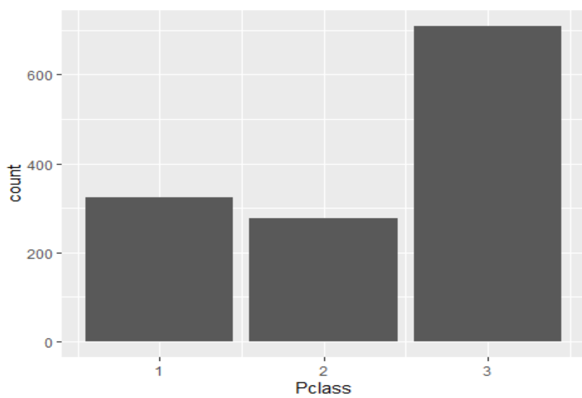


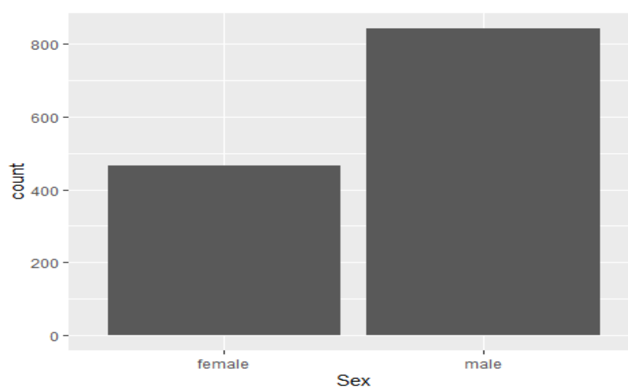


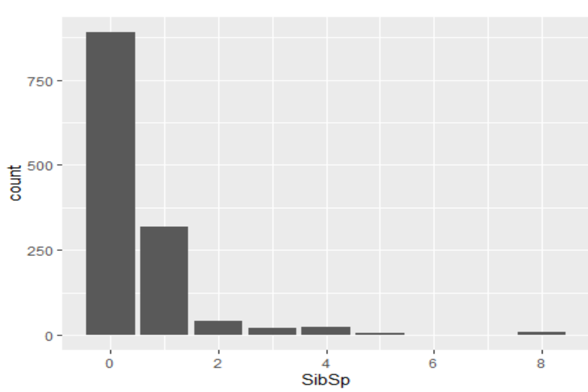


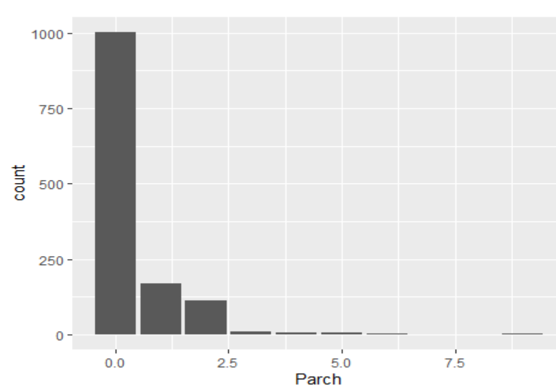


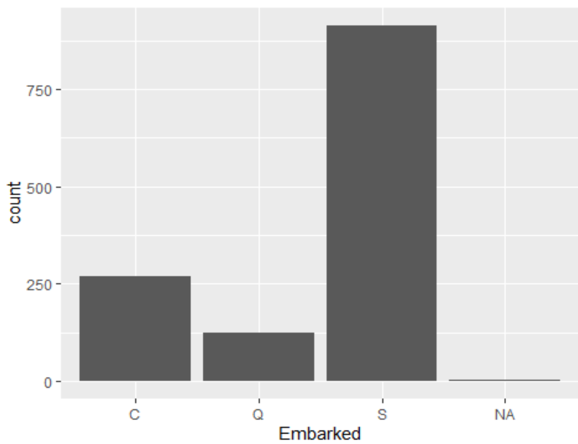
#Bar charts for categorical values







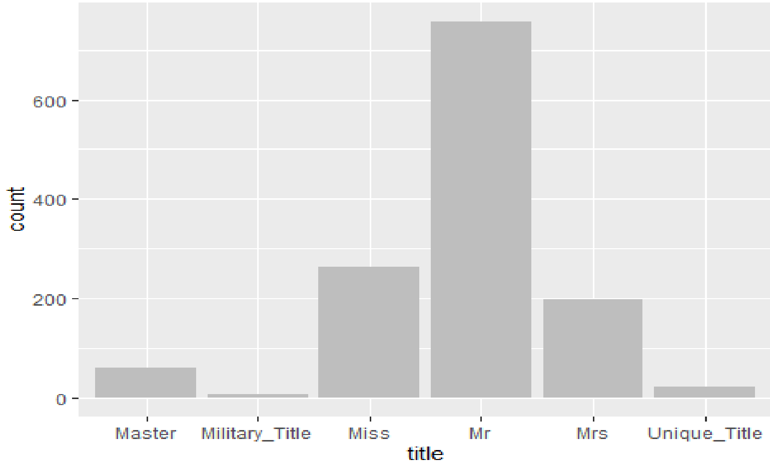





```
#Select title from passenger names:
colnames(TitanicSelect)

## [1] "Survived" "Pclass"   "Name"     "Sex"      "Age"      "SibSp"    "Parch"
## [8] "Fare"     "Embarked"

#Showing number of title counts by sex
TitanicSelect$title<-gsub('(.*, )|(\\..*)', '', TitanicSelect$Name)
table(TitanicSelect$Sex, TitanicSelect$title)

##         Capt Col Don Dona  Dr Jonkheer Lady Major Master Miss Mlle Mme  Mr Mrs
##   female    0   0   0    1   1        0    1     0      0  260    2   1   0 197
##   male      1   4   1    0   7        1    0     2     61    0    0   0 757   0
##          Ms Rev Sir the Countess
##   female   2   0   0            1
##   male     0   8   1            0

#Transformation of title into various categories based on similarities:
#Titles in low numbers are combined as rare_title:
Rare_Title <- c('Don', 'Dona', 'Dr', 'Jonkheer', 'Lady',  'Rev', 'Sir', 'the Countess')
Military_Title <- c('Capt', 'Col', 'Major' )

TitanicSelect$title[TitanicSelect$title=='Mlle'] <- 'Miss'
TitanicSelect$title[TitanicSelect$title=='Ms'] <- 'Miss'
TitanicSelect$title[TitanicSelect$title=='Mme'] <- 'Mrs'
TitanicSelect$title[TitanicSelect$title %in% Military_Title] <- 'Military_Title'
TitanicSelect$title[TitanicSelect$title %in% Rare_Title] <- 'Unique_Title'
str(TitanicSelect)
## tibble [1,309 x 10] (S3: tbl_df/tbl/data.frame)
##  $ Survived: num [1:1309] 0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass  : num [1:1309] 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name    : chr [1:1309] "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
##  $ Sex     : chr [1:1309] "male" "female" "female" "female" ...
##  $ Age     : num [1:1309] 22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp   : num [1:1309] 1 1 0 1 0 0 0 3 0 1 ...
```

```
## $ Parch   : num [1:1309] 0 0 0 0 0 0 0 1 2 0 ...
## $ Fare    : num [1:1309] 7.25 71.28 7.92 53.1 8.05 ...
## $ Embarked: chr [1:1309] "S" "C" "S" "S" ...
## $ title   : chr [1:1309] "Mr" "Mrs" "Miss" "Mrs" ...
```

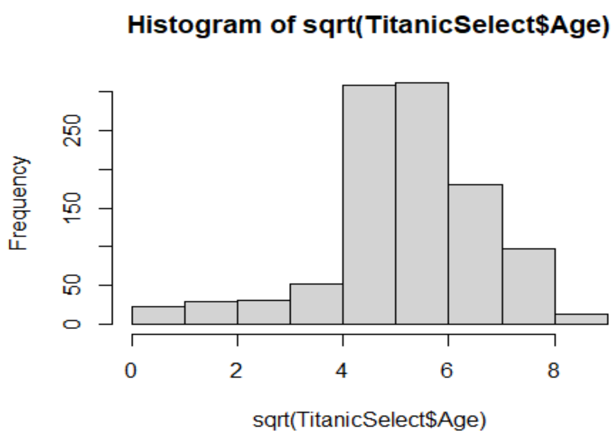
```
#to know the final count based on sex:
table(TitanicSelect$Sex, TitanicSelect$title)
ggplot(data=TitanicSelect, aes(x = title)) + geom_bar(fill = "grey")
##         Master Military_Title Miss  Mr Mrs Unique_Title
##   female      0              0  264   0 198            4
##   male       61              7    0 757   0           18
```

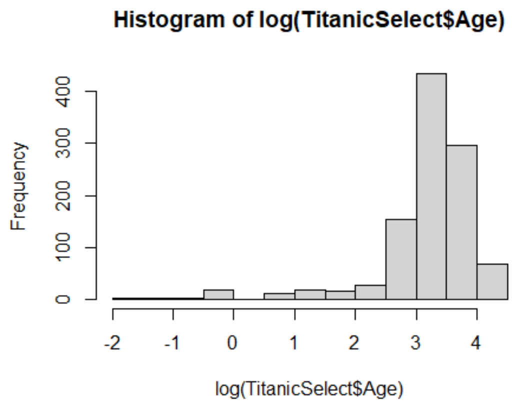



```
#Missing values:
sapply(TitanicSelect, function(x) sum(is.na(x)))
## Survived   Pclass     Name      Sex          Age    SibSp    Parch     Fare  Embarked    title
##        0        0        0        0          263        0        0        1         2        0

#Fare based on mean
avg.Fare=mean(TitanicSelect$Fare, na.rm=T)
TitanicSelect$Fare[is.na(TitanicSelect$Fare)] = avg.Fare

#Embarked to Q
TitanicSelect$Embarked[is.na(TitanicSelect$Embarked)] = 'Q'

#Age based on mean of title
Titanicnew <- TitanicSelect %>% group_by(title) %>% mutate(Age=if_else(is.na(Age), mean(Age, na.rm = TRUE),
Age))

sapply(TitanicSelect, function(x) sum(is.na(x)))
## Survived   Pclass     Name      Sex      Age    SibSp    Parch     Fare
##        0        0        0        0      263        0        0        0
## Embarked    title
##        0        0

Titanicnew<-  Titanicnew[complete.cases(Titanicnew), ]
nrow(Titanicnew)
```

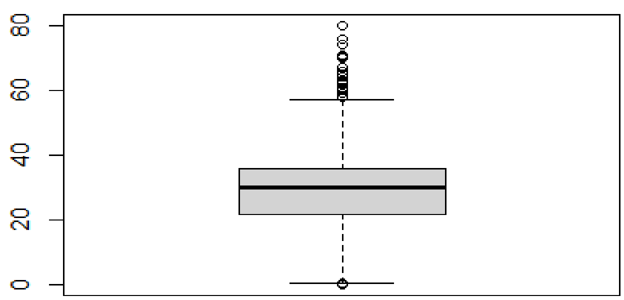
```
## [1] 1309
```

```
#Age transform
hist(sqrt(TitanicSelect$Age)),  hist(log(TitanicSelect$Age))
```





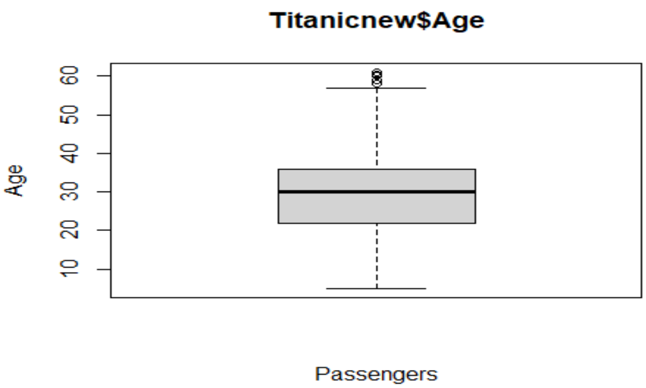

```
# Age Outlier check using boxplot:
Age_plot = boxplot(Titanicnew$Age)
```

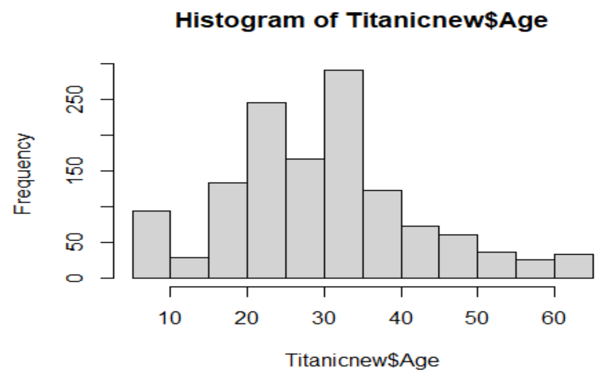



```
Age_plot$stats
##       [,1]
## [1,]  0.670
## [2,] 21.824
## [3,] 30.000
## [4,] 36.000
## [5,] 57.000
quantile(Titanicnew$Age, seq(0, 1, 0.02))
##     0%     2%     4%     6%     8%    10%    12%    14%    16%    18%    20%
##  0.170  2.000  5.000  8.000 13.000 16.000 17.000 18.000 19.000 20.000 21.000
##    22%    24%    26%    28%    30%    32%    34%    36%    38%    40%    42%
## 21.000 21.824 21.824 22.000 22.700 23.280 24.000 25.000 25.000 26.000 27.000
##    44%    46%    48%    50%    52%    54%    56%    58%    60%    62%    64%
## 28.000 29.000 29.840 30.000 31.000 32.000 32.252 32.252 32.252 32.252 32.252
```

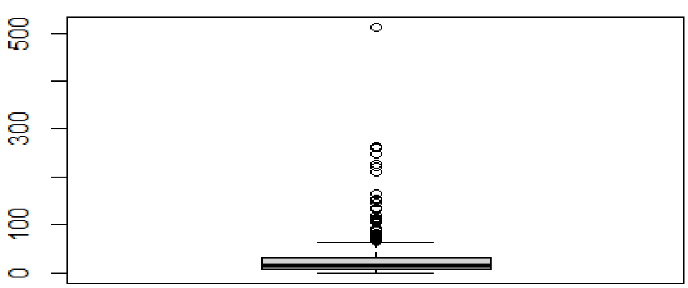
```
## 66% 68% 70% 72% 74% 76% 78% 80% 82% 84% 86%
## 32.252 32.252 33.000 35.000 36.000 36.918 37.000 39.000 40.000 42.000 44.000
## 88% 90% 92% 94% 96% 98% 100%
## 45.000 48.000 50.000 53.000 57.000 61.840 80.000
```

```
#Age Outlier treatment
Titanicnew$Age = ifelse(Titanicnew$Age>=61, 61, Titanicnew$Age) #98% in 61 (61>57 which is 75th percentile)
Titanicnew$Age = ifelse(Titanicnew$Age<=5, 5, Titanicnew$Age) #2% in 5 (5 is in 25th percentile)
boxplot(Titanicnew$Age, ylab= "Age", xlab = "Passengers", main = "Titanicnew$Age", title = TRUE)
hist(Titanicnew$Age)
```





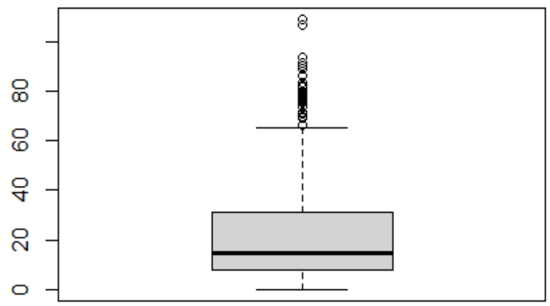

```
# Fare outlier check
Fare_plot = boxplot(Titanicnew$Fare)
```




```
Fare_plot$stats
##         [,1]
## [1,]  0.0000
## [2,]  7.8958
## [3,] 14.4542
## [4,] 31.2750
## [5,] 65.0000
quantile(Titanicnew$Fare, seq(0, 1, 0.02))
##      0%     2%     4%     6%     8%    10%    12%    14%
##  0.0000 6.4958 7.1303 7.2292 7.2500 7.5700 7.7500 7.7500
```

```
## 16% 18% 20% 22% 24% 26% 28% 30%
## 7.7570 7.7768 7.8542 7.8938 7.8958 7.9250 8.0500 8.0500
## 32% 34% 36% 38% 40% 42% 44% 46%
## 8.0500 8.6625 9.4980 10.5000 10.5000 12.3500 13.0000 13.0000
## 48% 50% 52% 54% 56% 58% 60% 62%
## 13.5000 14.4542 15.2458 15.7820 17.5920 20.5250 21.6792 24.0000
## 64% 66% 68% 70% 72% 74% 76% 78%
## 26.0000 26.0000 26.2875 27.0000 28.5285 30.5000 31.5143 36.7500
## 80% 82% 84% 86% 88% 90% 92% 94%
## 41.5792 51.6938 55.4417 61.0340 69.5500 78.0200 83.1583 108.9000
## 96% 98% 100%
## 146.5208 221.7792 512.3292
```

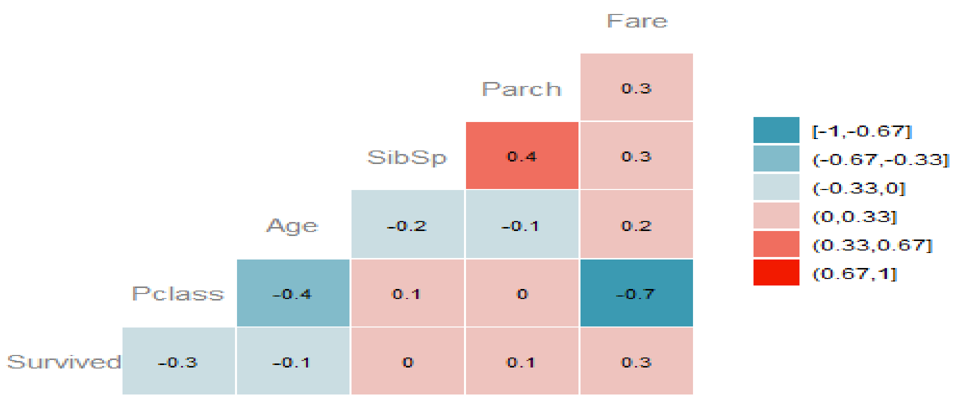
```
#Fare outlier treatment
Titanicnew$Fare = ifelse(Titanicnew$Fare>=109, 109, Titanicnew$Fare) #(95%)
boxplot(Titanicnew$Fare)
```




```
quantile(Titanicnew$Fare, seq(0, 1, 0.02))
```

```
## 0% 2% 4% 6% 8% 10% 12% 14%
## 0.0000 6.4958 7.1303 7.2292 7.2500 7.5700 7.7500 7.7500
## 16% 18% 20% 22% 24% 26% 28% 30%
## 7.7570 7.7768 7.8542 7.8938 7.8958 7.9250 8.0500 8.0500
## 32% 34% 36% 38% 40% 42% 44% 46%
## 8.0500 8.6625 9.4980 10.5000 10.5000 12.3500 13.0000 13.0000
## 48% 50% 52% 54% 56% 58% 60% 62%
## 13.5000 14.4542 15.2458 15.7820 17.5920 20.5250 21.6792 24.0000
## 64% 66% 68% 70% 72% 74% 76% 78%
## 26.0000 26.0000 26.2875 27.0000 28.5285 30.5000 31.5143 36.7500
## 80% 82% 84% 86% 88% 90% 92% 94%
## 41.5792 51.6938 55.4417 61.0340 69.5500 78.0200 83.1583 108.9000
## 96% 98% 100%
## 109.0000 109.0000 109.0000
```

```
#Correlation matrix
ggcorr(Titanicnew, nbreaks = 6, label = TRUE, label_size = 3, color = 'grey50')
```




```
#dummy variable creation and factorizing the categorical variables:
Titanicnew$Sex2 <- ifelse(Titanicnew$Sex == 'male', 1,0)

Titanicnew$Embarked2 <- ifelse(Titanicnew$Embarked == 'C', 1,
                  ifelse(Titanicnew$Embarked == 'S',2,0))

Titanicnew$title2 <- ifelse(Titanicnew$title == 'Mr', 1,
                ifelse(Titanicnew$title == 'Mrs',2,
                ifelse(Titanicnew$title == 'Miss',3,
                ifelse(Titanicnew$title == 'Master',4,
                ifelse(Titanicnew$title == 'Unique_Title',5,0)))))

Titanicnew$Sex2 <- factor(Titanicnew$Sex2)
Titanicnew$Embarked2 <- factor(Titanicnew$Embarked2)
Titanicnew$title2 <- factor(Titanicnew$title2)

colnames(Titanicnew)
"Survived"  "Pclass"  "Name" "Sex"  "Age" "SibSp" "Parch"  "Fare"   "Embarked"  "title"  "Sex2"  "Embarked2" "title2"

#Columns for model
TitanicFinal <- select(Titanicnew,c("title", "Survived", "Pclass", "Sex", "Age", "SibSp", "Parch", "Fare",  "Sex2",
"Embarked2", "title2"))

#Data Partitioning
set.seed(12345)
t= sample(1:nrow(TitanicFinal), 0.7*nrow(TitanicFinal))
Titanictrain = TitanicFinal[t,]
Titanictest = TitanicFinal[-t,]
nrow(Titanictrain)
nrow(Titanictest)
## [1] 916
## [1] 393
```

```
#Model1
trainmodel1<-glm(Survived~., data=Titanictrain, family=binomial(link = logit))
summary(trainmodel1)
exp(cbind(Odds_Ratio_SurviveOrNot=coef(trainmodel1), confint(trainmodel1)))
## Call:
## glm(formula = Survived ~ ., family = binomial(link = logit),
##     data = Titanictrain)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.575  -0.491  -0.313   0.453   2.564
##
## Coefficients: (6 not defined because of singularities)
##                    Estimate Std. Error z value   Pr(>|z|)
## (Intercept)        18.87306  501.36535    0.04     0.9700
## titleMilitary_Title  -2.41303    1.30963   -1.84     0.0654 .
## titleMiss         -14.09151  501.36484   -0.03     0.9776
## titleMr            -2.33512    0.49454   -4.72 0.00000234 ***
## titleMrs          -13.50296  501.36493   -0.03     0.9785
## titleUnique_Title   -2.48169    0.94624   -2.62     0.0087 **
## Pclass             -0.95430    0.19367   -4.93 0.00000083 ***
## Sexmale           -15.70177  501.36467   -0.03     0.9750
## Age                -0.02007    0.01114   -1.80     0.0715 .
## SibSp              -0.30869    0.11105   -2.78     0.0054 **
## Parch              -0.33908    0.14239   -2.38     0.0173 *
## Fare                0.00665    0.00591    1.13     0.2602
## Sex21                    NA         NA      NA         NA
## Embarked21         -0.08448    0.43755   -0.19     0.8469
## Embarked22         -0.37360    0.38122   -0.98     0.3271
## title21                  NA         NA      NA         NA
## title22                  NA         NA      NA         NA
## title23                  NA         NA      NA         NA
## title24                  NA         NA      NA         NA
## title25                  NA         NA      NA         NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1215.50  on 915  degrees of freedom
## Residual deviance:  645.63  on 902  degrees of freedom
## AIC: 673.6
##
## Number of Fisher Scoring iterations: 13
##                 Odds_Ratio_SurviveOrNot                  2.5 %
## (Intercept)         157205350.06877976656 0.00000000000000000023046
## titleMilitary_Title          0.08954396692 0.00370697763516999250875
## titleMiss                    0.00000075881                        NA
## titleMr                      0.09679844014 0.03656665823834219058877
## titleMrs                     0.00000136691                        NA
## titleUnique_Title            0.08360149735 0.01012157990517860255141 9
## Pclass                       0.38508034909 0.26219951842805405695102 6
```

```
## Sexmale                    0.00000015164                    NA
## Age                      0.98012733153 0.958698635128709786279444
## SibSp                    0.73440753530 0.584754679320565795563880
## Parch                    0.71242809226 0.535254194149181961037698
## Fare                     1.00667548566 0.995098169308445967828902
## Sex21                              NA                NA
## Embarked21               0.91898717671 0.390928533110673159800541
## Embarked22               0.68824916415 0.327474716960308942503843
## title21                            NA                NA
## title22                            NA                NA
## title23                            NA                NA
## title24                            NA                NA
## title25                            NA                NA
##                                97.5 %
## (Intercept)                        NA
## titleMilitary_Title           0.97789
## titleMiss        9291782352204721029820.00000
## titleMr                       0.25561
## titleMrs        14660843339530912137666.00000
## titleUnique_Title             0.46992
## Pclass                        0.56097
## Sexmale          2343370165629642342400.00000
## Age                           1.00154
## SibSp                         0.90579
## Parch                         0.93872
## Fare                          1.01845
## Sex21                              NA
## Embarked21                    2.17683
## Embarked22                    1.46177
## title21                            NA
## title22                            NA
## title23                            NA
## title24                            NA
## title25                            NA
```

```
#Model2
trainmodel2<-glm(Survived~Pclass + Age + SibSp + Sex2 ,data=Titanictrain, family = binomial(link = logit))
summary(trainmodel2)
exp(cbind(Odds_Ratio_SurviveOrNot=coef(trainmodel2), confint(trainmodel2)))
## Call:
## glm(formula = Survived ~ Pclass + Age + SibSp + Sex2, family = binomial(link = logit),
##     data = Titanictrain)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.526  -0.520  -0.344   0.461   2.604
##
## Coefficients:
##             Estimate Std. Error z value            Pr(>|z|)
## (Intercept)  5.38527    0.53947    9.98 < 0.0000000000000002 ***
## Pclass      -1.10416    0.13496   -8.18  0.00000000000000028 ***
## Age         -0.03410    0.00882   -3.87              0.00011 ***
## SibSp       -0.28014    0.09630   -2.91              0.00363 **
## Sex21       -3.76875    0.22366  -16.85 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1215.50  on 915  degrees of freedom
## Residual deviance:  678.15  on 911  degrees of freedom
## AIC: 688.1
##
## Number of Fisher Scoring iterations: 5
## Waiting for profiling to be done...
##             Odds_Ratio_SurviveOrNot      2.5 %     97.5 %
## (Intercept)              218.168675 77.829372 646.646927
## Pclass                     0.331488  0.252919   0.429687
## Age                        0.966474  0.949729   0.983194
## SibSp                      0.755677  0.619232   0.906879
## Sex21                      0.023081  0.014674   0.035311
```

```
#Model3
trainmodel3<-glm(Survived~Pclass + Age + SibSp +Parch +Fare + Sex2 + Embarked2 ,data=Titanictrain, family =
binomial(link = logit))
summary(trainmodel3)
exp(cbind(Odds_Ratio_SurviveOrNot=coef(trainmodel3), confint(trainmodel3)))
## Call:
## glm(formula = Survived ~ Pclass + Age + SibSp + Parch + Fare +
##     Sex2 + Embarked2, family = binomial(link = logit), data = Titanictrain)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.556  -0.500  -0.338   0.475   2.572
##
## Coefficients:
##             Estimate Std. Error z value             Pr(>|z|)
## (Intercept)  5.13135    0.77790    6.60       0.000000000042 ***
## Pclass      -0.94420    0.18457   -5.12       0.000000312721 ***
## Age         -0.03418    0.00885   -3.86              0.00011 ***
## SibSp       -0.29426    0.10869   -2.71              0.00678 **
## Parch       -0.10859    0.12843   -0.85              0.39781
## Fare         0.00674    0.00560    1.20              0.22909
## Sex21       -3.75675    0.23146  -16.23 < 0.0000000000000002 ***
## Embarked21  -0.01553    0.43074   -0.04              0.97124
## Embarked22  -0.32849    0.37901   -0.87              0.38611
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1215.50  on 915  degrees of freedom
## Residual deviance:  673.27  on 907  degrees of freedom
## AIC: 691.3
##
## Number of Fisher Scoring iterations: 5
## Waiting for profiling to be done...
##             Odds_Ratio_SurviveOrNot     2.5 %     97.5 %
## (Intercept)              169.245284 37.787757 801.221656
## Pclass                     0.388993  0.269782   0.556856
## Age                        0.966397  0.949590   0.983162
## SibSp                      0.745086  0.596607   0.916887
## Parch                      0.897097  0.695829   1.155227
## Fare                       1.006761  0.995812   1.017948
## Sex21                      0.023359  0.014615   0.036265
## Embarked21                 0.984592  0.423448   2.293058
## Embarked22                 0.720013  0.342967   1.514550
```

```
#Model4
trainmodel4<-glm(Survived~Pclass + Age + SibSp + Parch + Sex2 ,data=Titanictrain, family = binomial(link =
logit))
summary(trainmodel4)
exp(cbind(Odds_Ratio_SurviveOrNot=coef(trainmodel4), confint(trainmodel4)))
## Call:
## glm(formula = Survived ~ Pclass + Age + SibSp + Parch + Sex2,
##     family = binomial(link = logit), data = Titanictrain)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.489  -0.514  -0.346   0.458   2.585
##
## Coefficients:
##             Estimate Std. Error z value            Pr(>|z|)
## (Intercept)  5.43481    0.54527    9.97 < 0.0000000000000002 ***
## Pclass      -1.10411    0.13492   -8.18  0.00000000000000028 ***
## Age         -0.03435    0.00884   -3.88               0.0001 ***
## SibSp       -0.25975    0.10074   -2.58               0.0099 **
## Parch       -0.08059    0.12205   -0.66               0.5091
## Sex21       -3.79900    0.22923  -16.57 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1215.50  on 915  degrees of freedom
## Residual deviance:  677.71  on 910  degrees of freedom
## AIC: 689.7
##
## Number of Fisher Scoring iterations: 5
## Waiting for profiling to be done...
##             Odds_Ratio_SurviveOrNot     2.5 %     97.5 %
## (Intercept)              229.249115 80.852952 687.110375
## Pclass                     0.331506  0.252965   0.429693
## Age                        0.966231  0.949447   0.982987
## SibSp                      0.771242  0.626709   0.934026
## Parch                      0.922574  0.724734   1.173803
## Sex21                      0.022393  0.014073   0.034615
```

```
#Model5
trainmodel5<-glm(Survived~Pclass + Age + SibSp + Sex2 + title2 ,data=Titanictrain, family = binomial(link = logit))
summary(trainmodel5)
exp(cbind(Odds_Ratio_SurviveOrNot=coef(trainmodel5), confint(trainmodel5)))
## Call:
## glm(formula = Survived ~ Pclass + Age + SibSp + Sex2 + title2,
##     family = binomial(link = logit), data = Titanictrain)
##
## Deviance Residuals:
##    Min     1Q  Median     3Q    Max
## -2.561  -0.526  -0.320   0.440   2.708
##
## Coefficients:
##             Estimate Std. Error z value         Pr(>|z|)
## (Intercept)  16.8514   507.5147    0.03          0.97351
## Pclass       -1.0903     0.1418   -7.69 0.000000000000015 ***
## Age          -0.0179     0.0109   -1.64          0.10022
## SibSp        -0.3452     0.0990   -3.49          0.00049 ***
## Sex21       -15.8809   507.5130   -0.03          0.97504
## title21      -0.0730     1.1800   -0.06          0.95069
## title22     -11.7299   507.5144   -0.02          0.98156
## title23     -12.0026   507.5145   -0.02          0.98113
## title24       2.0543     1.2936    1.59          0.11229
## title25      -0.2567     1.4029   -0.18          0.85483
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1215.50  on 915  degrees of freedom
## Residual deviance:  655.15  on 906  degrees of freedom
## AIC: 675.1
##
## Number of Fisher Scoring iterations: 13
##             Odds_Ratio_SurviveOrNot
## (Intercept)   20820457.83457517624
## Pclass               0.33612891313
## Age                  0.98230314656
## SibSp                0.70805087977
## Sex21                0.00000012677
## title21              0.92962842294
## title22              0.00000804982
## title23              0.00000612827
## title24              7.80128437128
## title25              0.77362363068
```

#Scoring the prediction rate:

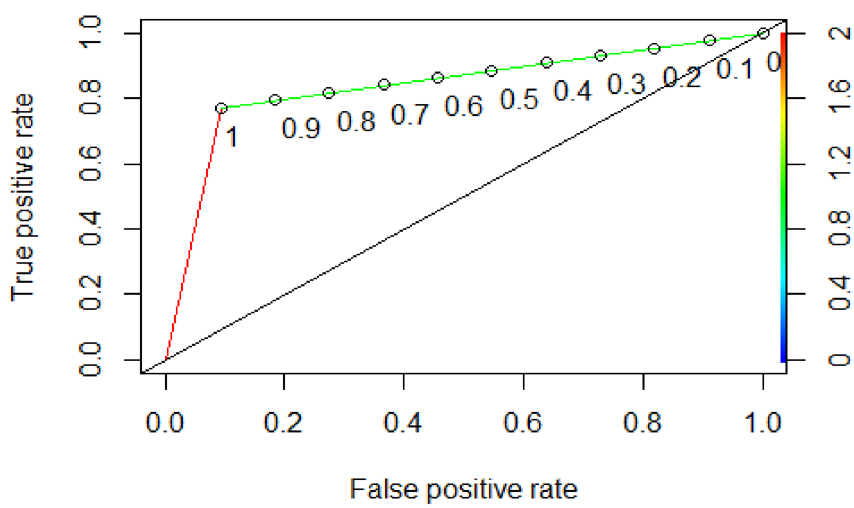
```
Titanictrain$score1 <- predict(trainmodel2, newdata=subset(Titanictrain, select=c('Pclass', 'Age', 'SibSp', 'Sex2')), type="response")
head(Titanictrain$score1)
##        1        2        3        4        5        6
## 0.789609 0.044992 0.056185 0.719125 0.230477 0.051000
```

#Train Model Confusion Matrix

```
Titanictrain$prediction1 <- ifelse(Titanictrain$score1>=0.5, 1, 0)
table(factor(Titanictrain$prediction1),
    factor(Titanictrain$Survived))
##       0   1
##   0 516  79
##   1  53 268
```

#Train Model ROC

```
ROCRpred1 <- prediction(Titanictrain$prediction1, Titanictrain$Survived)
ROCRperf1 <- performance(ROCRpred1, measure = "tpr", x.measure = "fpr")
plot(ROCRperf1, colorize = TRUE, text.adj = c(-0.2,1.7), print.cutoffs.at = seq(0,1,0.1))
abline(a=0, b= 1)
```



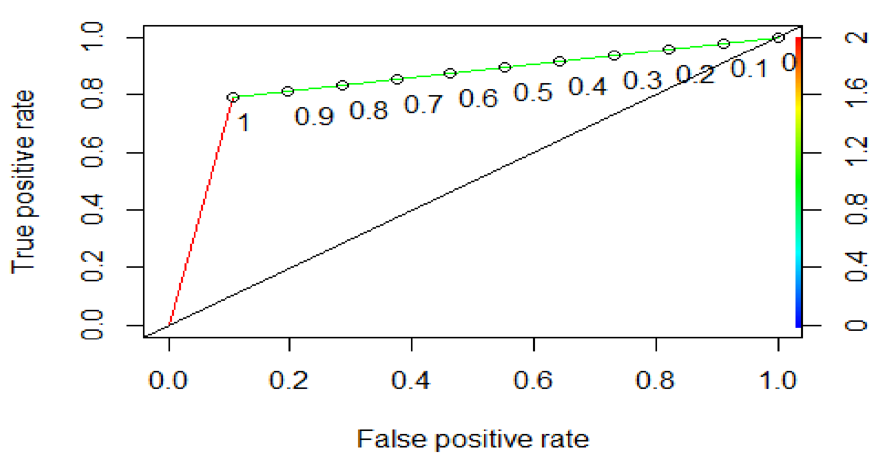

```
auc(Titanictrain$Survived, Titanictrain$prediction1)
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
## Area under the curve: 0.84
```

Model 5- trainmodel5 with Pclass + Age + SibSp + Sex2 + title2 gives the better prediction on train model, but does not deliver clear results on statistical significance of predictor values. Model 2- trainmodel2 with Pclass + Age + SibSp + Sex2 gives almost same accuracy with proper allocation of statistical significance and CI values.

#Test data Confusion Matrix

```
Titanictest$score_test<-predict(trainmodel2, Titanictest, type = "response")
```

```
Titanictest$prediction <- ifelse(Titanictest$score_test>=0.5, 1, 0)

Titanictest$Survived2 <- as.factor(Titanictest$Survived)
Titanictest$prediction2 <- as.factor(Titanictest$prediction)
confusionMatrix(Titanictest$prediction2,Titanictest$Survived2)
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 220  31
##          1  26 116
##
##                Accuracy : 0.855
##                  95% CI : (0.816, 0.888)
##     No Information Rate : 0.626
##     P-Value [Acc > NIR] : <0.0000000000000002
##
##                   Kappa : 0.688
##  Mcnemar's Test P-Value : 0.596
##
##             Sensitivity : 0.894
##             Specificity : 0.789
##          Pos Pred Value : 0.876
##          Neg Pred Value : 0.817
##              Prevalence : 0.626
##          Detection Rate : 0.560
##    Detection Prevalence : 0.639
##       Balanced Accuracy : 0.842
##        'Positive' Class : 0
#Test Data AUC
ROCRpred_test <- prediction(Titanictest$prediction, Titanictest$Survived)
ROCRperf_test <- performance(ROCRpred_test, measure = "tpr", x.measure = "fpr")
plot(ROCRperf_test, colorize = TRUE, text.adj = c(-0.2,1.7), print.cutoffs.at = seq(0,1,0.1))
abline(a=0, b= 1)
```



```
auc(Titanictest$Survived, Titanictest$prediction)
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
## Area under the curve: 0.842

#Train Model Residual Plots
titanic.res <-residuals(trainmodel2)

ggplot(data=Titanictrain, aes(x=Pclass, y=titanic.res))+geom_point()
ggplot(data=Titanictrain, aes(x=Age, y=titanic.res))+geom_point()
ggplot(data=Titanictrain, aes(x=SibSp, y=titanic.res))+geom_point()
ggplot(data=Titanictrain, aes(x=Sex2, y=titanic.res))+geom_point()
```