Titanic_survival_analysis_report

Student Name - Misha Aggarwal

Roll Number - 101803590

Group - COE27

Dataset Name - Titanic Dataset

Dataset Source - https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv (https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv)

Importing all the libraries used

```
library(ggplot2)
library(corrgram)
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'lattice'
```

```
## The following object is masked from 'package:corrgram':
##
##     panel.fill
```

```
library(ROCR)
library(dummies)
```

```
## dummies-1.5.6 provided by Decision Patterns
```

DATA EXPLORATION

We are importing the datasets

```
setwd("/Users/mishaaggarwal/Downloads")
titanic_data <- read.csv("titanic.csv",stringsAsFactors=FALSE)  #loading dataset
head(titanic_data)
```

	Survived	Pclass	Name	Sex	...
	<int>	<int>	<chr>	<chr>	<dbl>
1	0	3	Mr. Owen Harris Braund	male	22

	Survived	Pclass	Name	Sex	...
	<int>	<int>	<chr>	<chr>	<dbl>
2	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cumings	female	38
3	1	3	Miss. Laina Heikkinen	female	26
4	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35
5	0	3	Mr. William Henry Allen	male	35
6	0	3	Mr. James Moran	male	27

6 rows | 1-6 of 9 columns

We will explore the data that is contained in the titanic dataframe. We will then proceed to explore the other components of this dataframe –

```
summary(titanic_data)
```

```
##     Survived          Pclass          Name               Sex
##  Min.   :0.0000   Min.   :1.000   Length:887         Length:887
##  1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character
##  Median :0.0000   Median :3.000   Mode  :character   Mode  :character
##  Mean   :0.3856   Mean   :2.306
##  3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :1.0000   Max.   :3.000
##       Age        Siblings.Spouses.Aboard Parents.Children.Aboard
##  Min.   : 0.42   Min.   :0.0000          Min.   :0.0000
##  1st Qu.:20.25   1st Qu.:0.0000          1st Qu.:0.0000
##  Median :28.00   Median :0.0000          Median :0.0000
##  Mean   :29.47   Mean   :0.5254          Mean   :0.3833
##  3rd Qu.:38.00   3rd Qu.:1.0000          3rd Qu.:0.0000
##  Max.   :80.00   Max.   :8.0000          Max.   :6.0000
##       Fare
##  Min.   :  0.000
##  1st Qu.:  7.925
##  Median : 14.454
##  Mean   : 32.305
##  3rd Qu.: 31.137
##  Max.   :512.329
```

```
dim(titanic_data)
```

```
## [1] 887   8
```

DATA MANIPULATION

converting values to vectors

```
titanic_data$name=as.character(titanic_data$Name)
titanic_data$survived = as.factor(titanic_data$Survived)
titanic_data$sex = as.factor(titanic_data$Sex)
titanic_data$pclass = as.factor(titanic_data$Pclass)
```

To know number of males and females

```
sex<-table(titanic_data$Sex)
sex
```

```
##
## female   male
##    314    573
```

To differentiate between class of travel

```
pclass<-table(titanic_data$Pclass)
pclass
```

```
##
##   1   2   3
## 216 184 487
```

Survived out of total people

```
survived<-table(titanic_data$Survived)
survived
```

```
##
##   0   1
## 545 342
```

Survival percentage

```
x<-dim(titanic_data)[1]
y<-survived[2]
z<-y/x
survival_percent<-100*z
survival_percent
```

```
##        1
## 38.55693
```

Death precentage

```
x<-dim(titanic_data)[1]
y<-survived[1]
z<-y/x
death_percent<-100*z
death_percent
```

```
##        0
## 61.44307
```

Treating missing values

```
colSums(is.na(titanic_data))
```

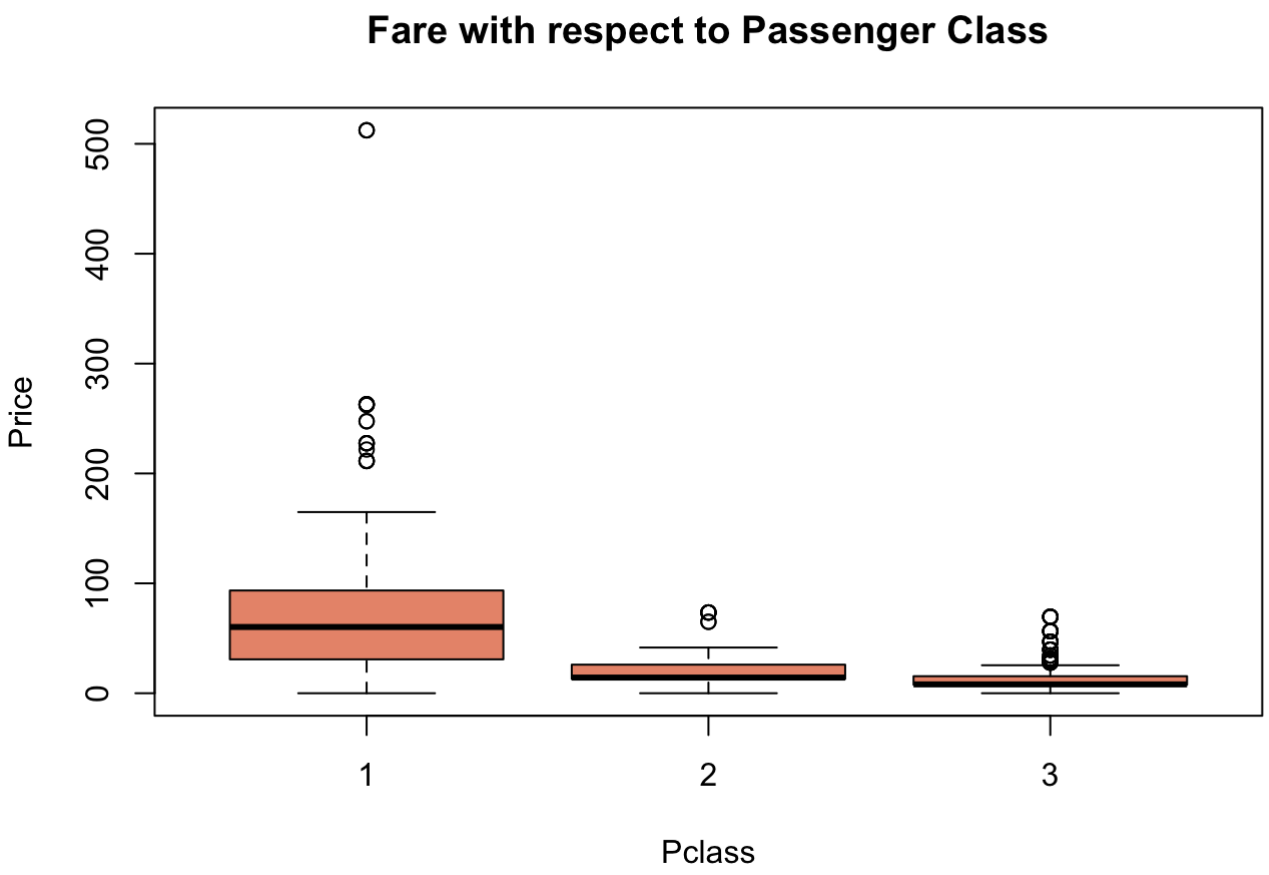
```
##                 Survived                   Pclass                     Name
##                        0                        0                        0
##                      Sex                      Age Siblings.Spouses.Aboard
##                        0                        0                        0
## Parents.Children.Aboard                     Fare                     name
##                        0                        0                        0
##                 survived                      sex                   pclass
##                        0                        0                        0
```

DATA VISUALISATION

```
##Plots
```

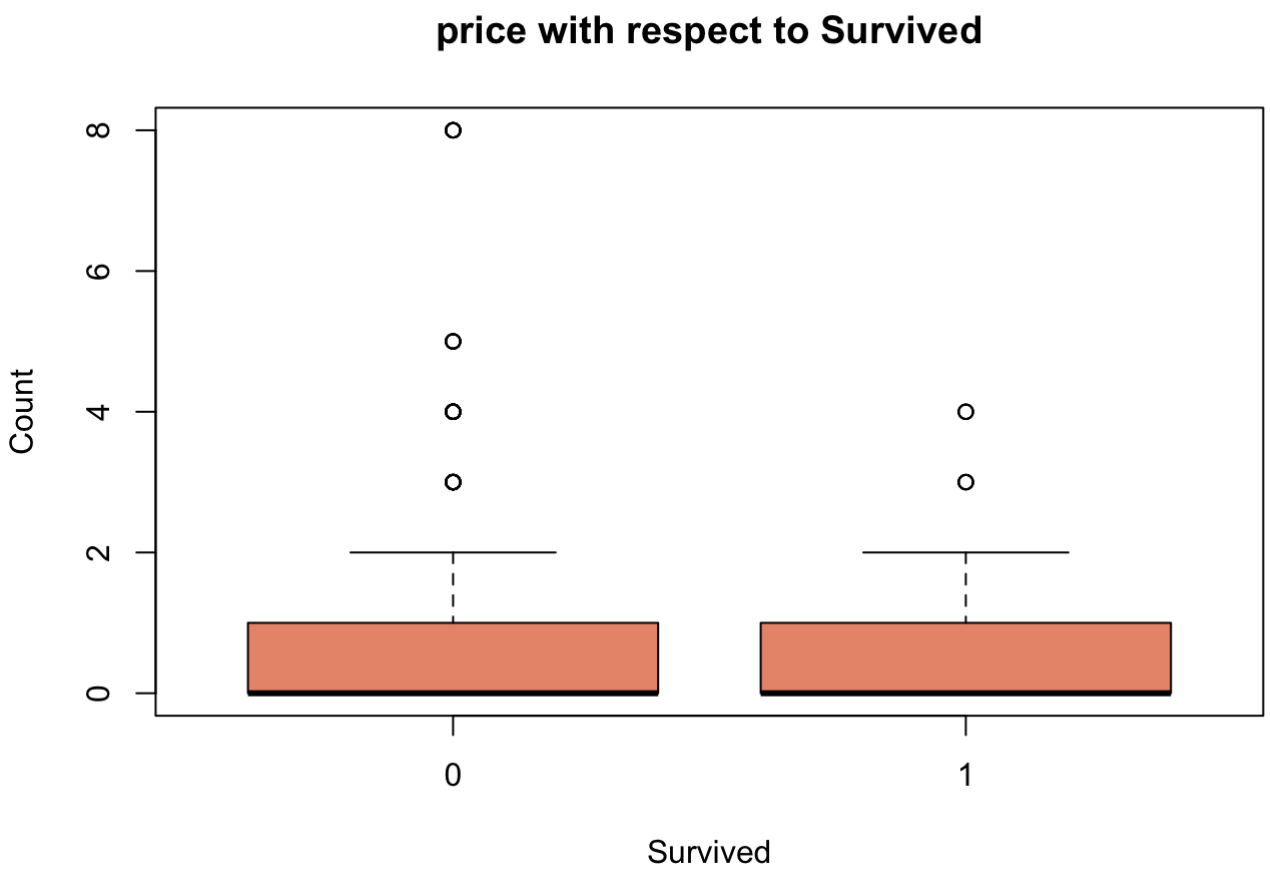
Price and pclass

```
boxplot(Fare ~ Pclass, data = titanic_data,
        main = 'Fare with respect to Passenger Class', ylab = 'Price', col = 'darksalmo
n')
```



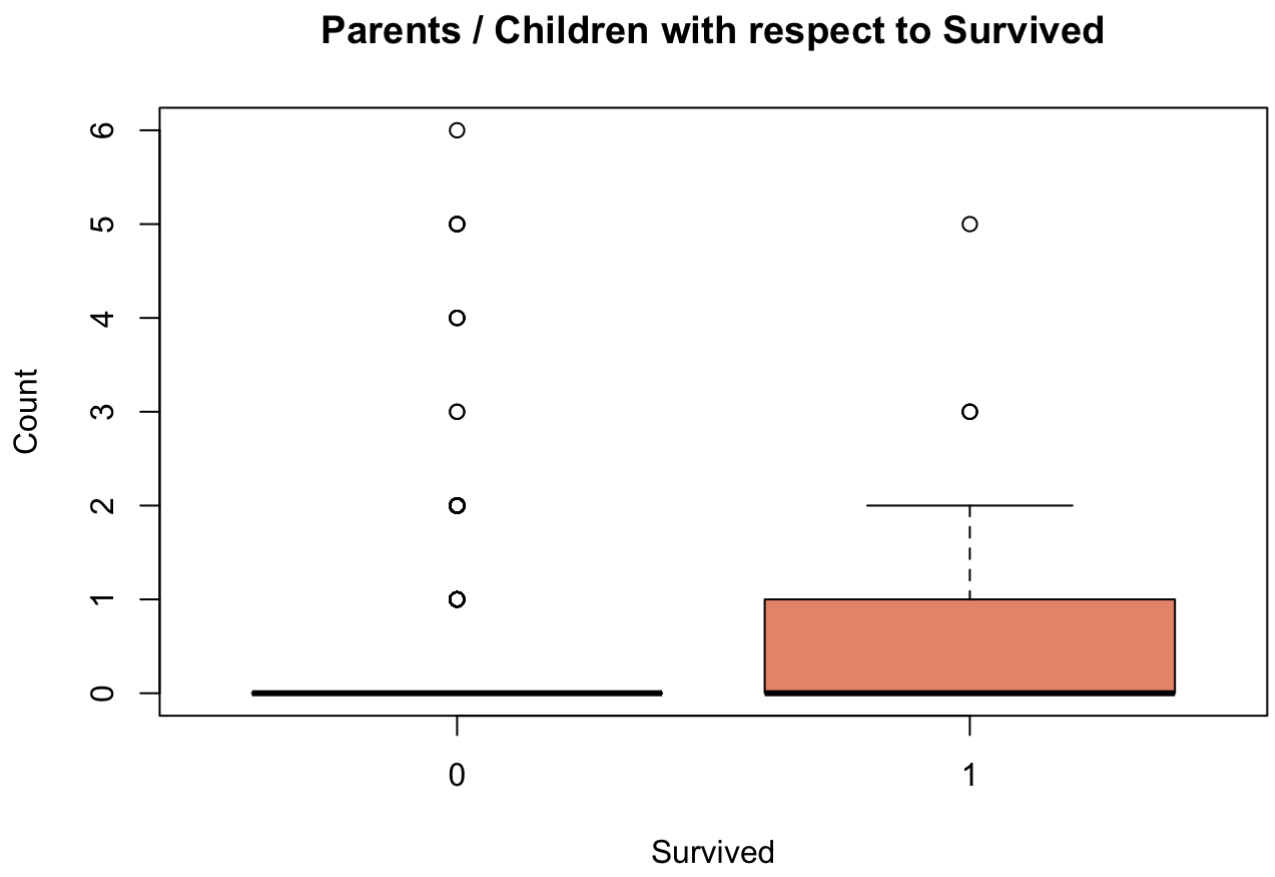
Siblings.Spouses.Aboard vs survived

```
boxplot(Siblings.Spouses.Aboard ~ Survived, data = titanic_data,
        main = 'price with respect to Survived', ylab = 'Count', col = 'darksalmon')
```



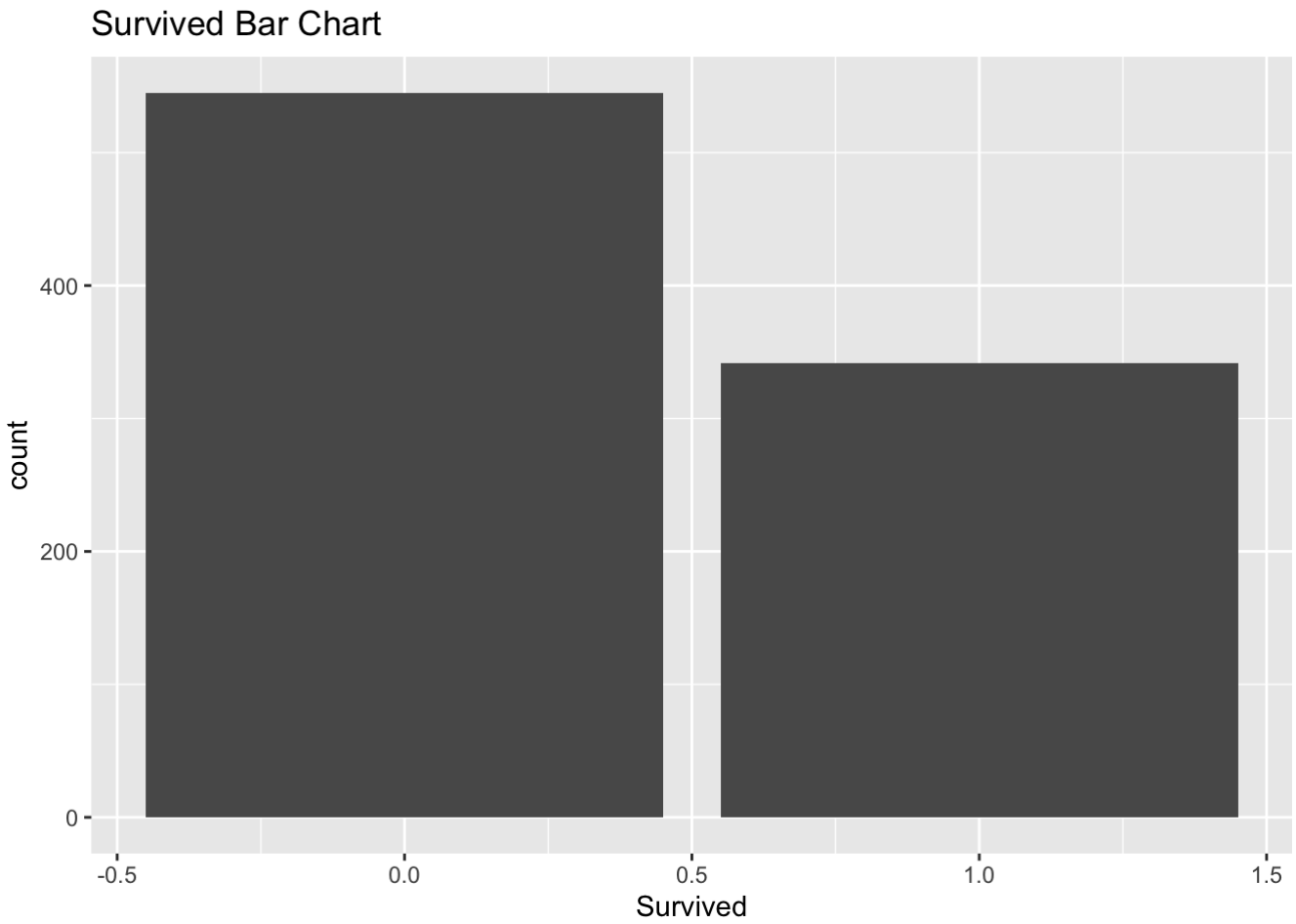
Parents.Children.Aboard vs survived

```
boxplot(Parents.Children.Aboard ~ Survived, data = titanic_data,
        main = 'Parents / Children with respect to Survived', ylab = 'Count', col = 'dar
ksalmon')
```



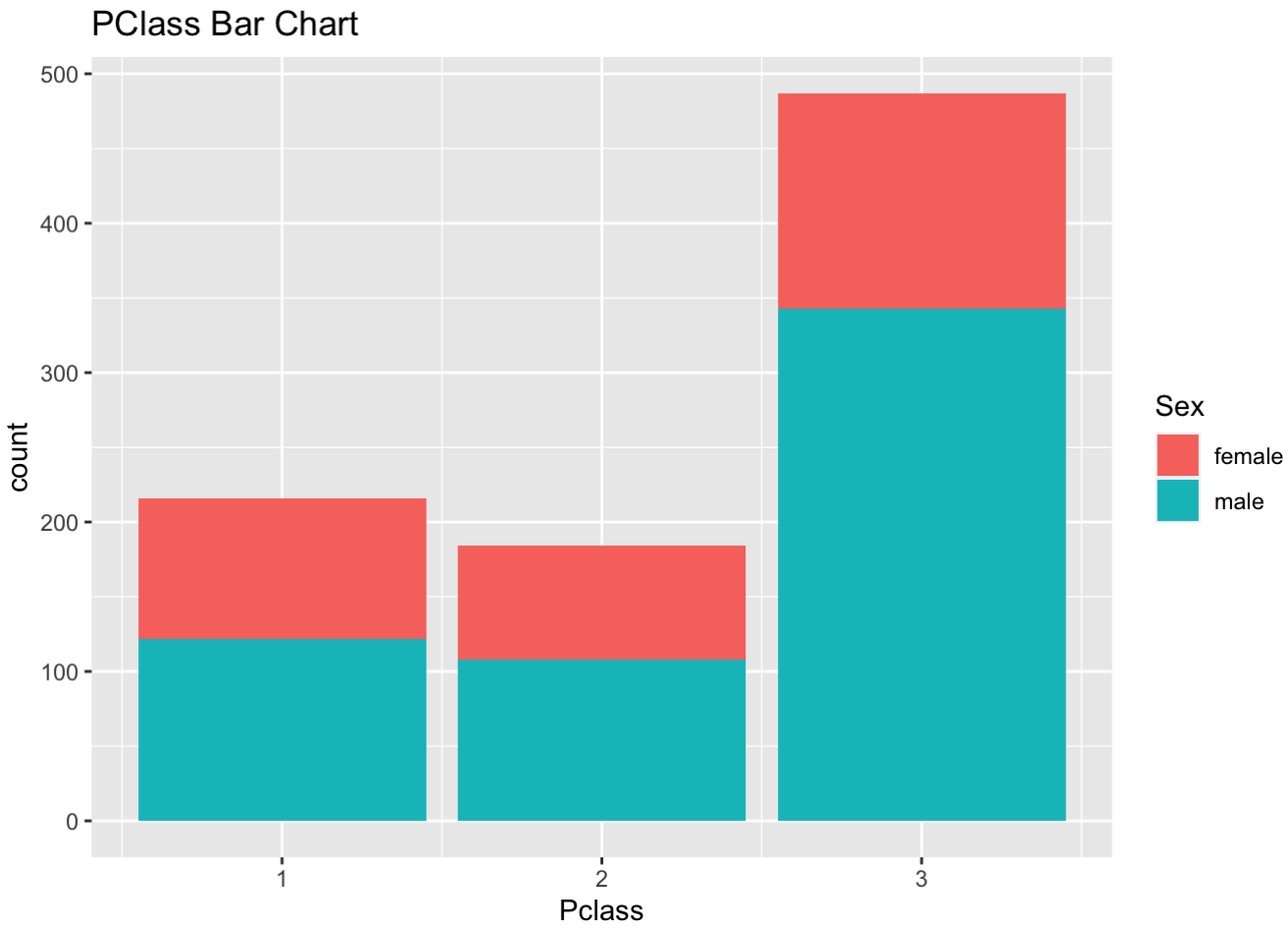
Comparison of survived vs died

```
ggplot.relation.object <- ggplot(titanic_data, aes(x=Survived))
ggplot.relation.object <-ggplot.relation.object+geom_bar()+ggtitle("Survived Bar Chart")
ggplot.relation.object
```



Check the PClass/survived_died/sex Distribution bar chart

```
ggplot.relation.object <- ggplot(titanic_data, aes(x=Pclass, fill = Sex))
ggplot.relation.object <-ggplot.relation.object+geom_bar()+ggtitle("PClass Bar Chart")
ggplot.relation.object
```

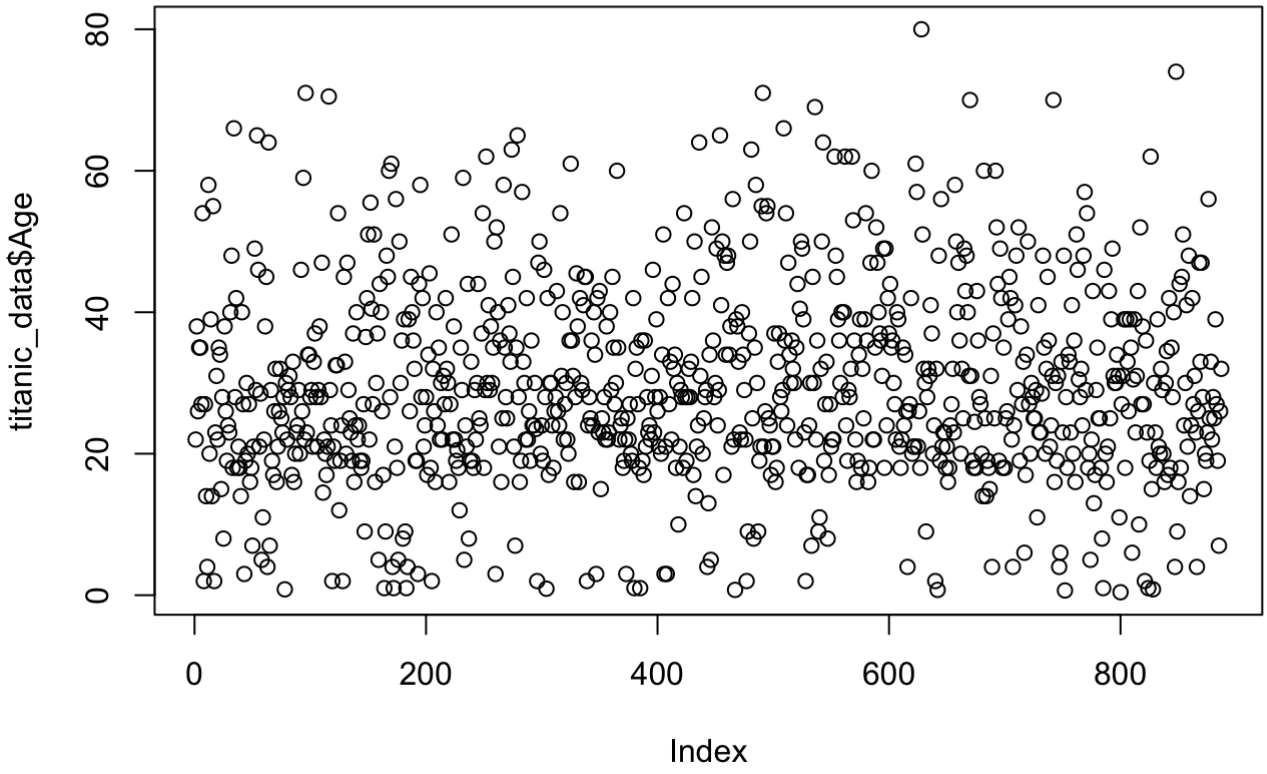


Age distribution table

```
summary(titanic_data$Age)
```

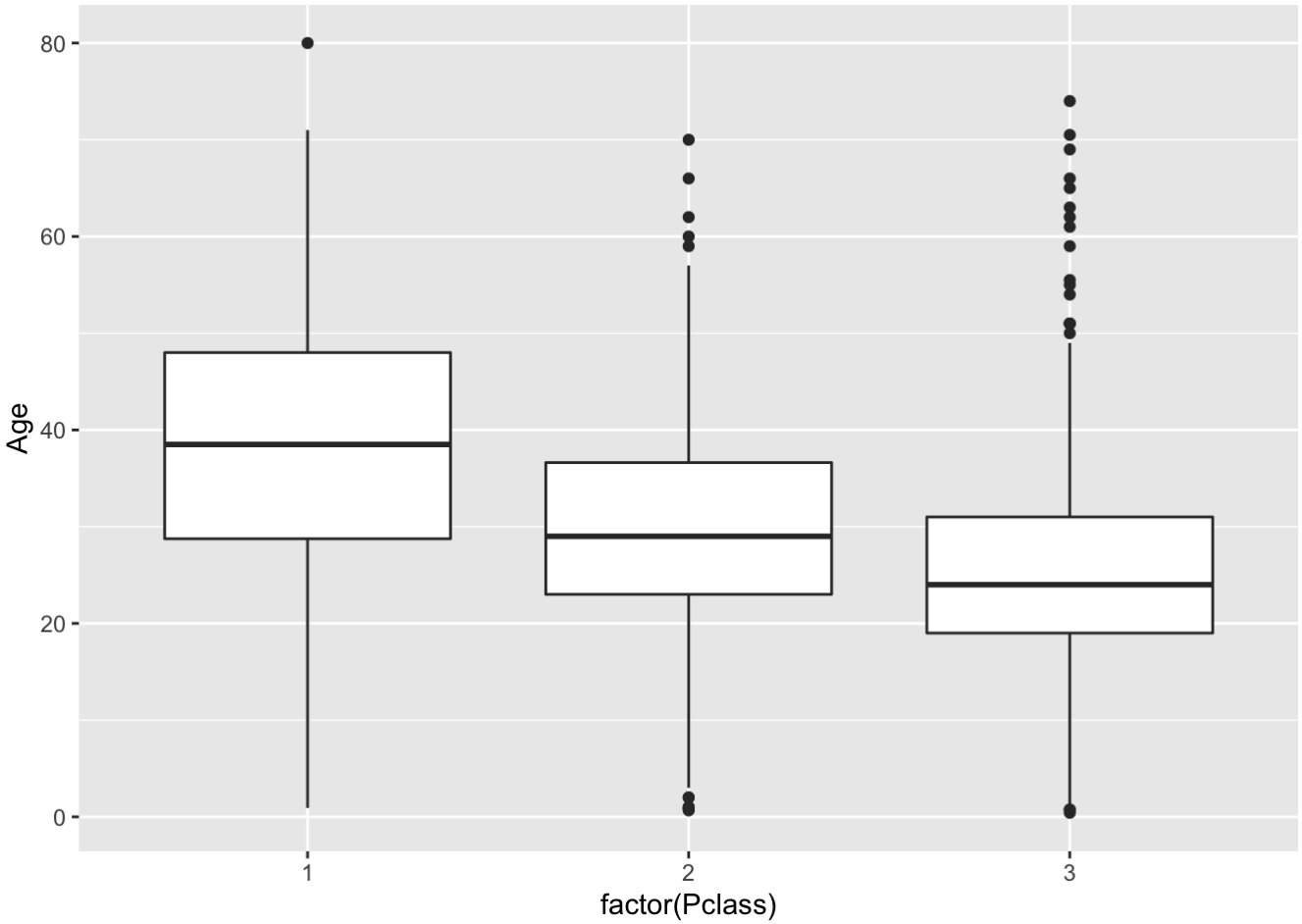
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   20.25   28.00   29.47   38.00   80.00
```

```
plot(titanic_data$Age)
```



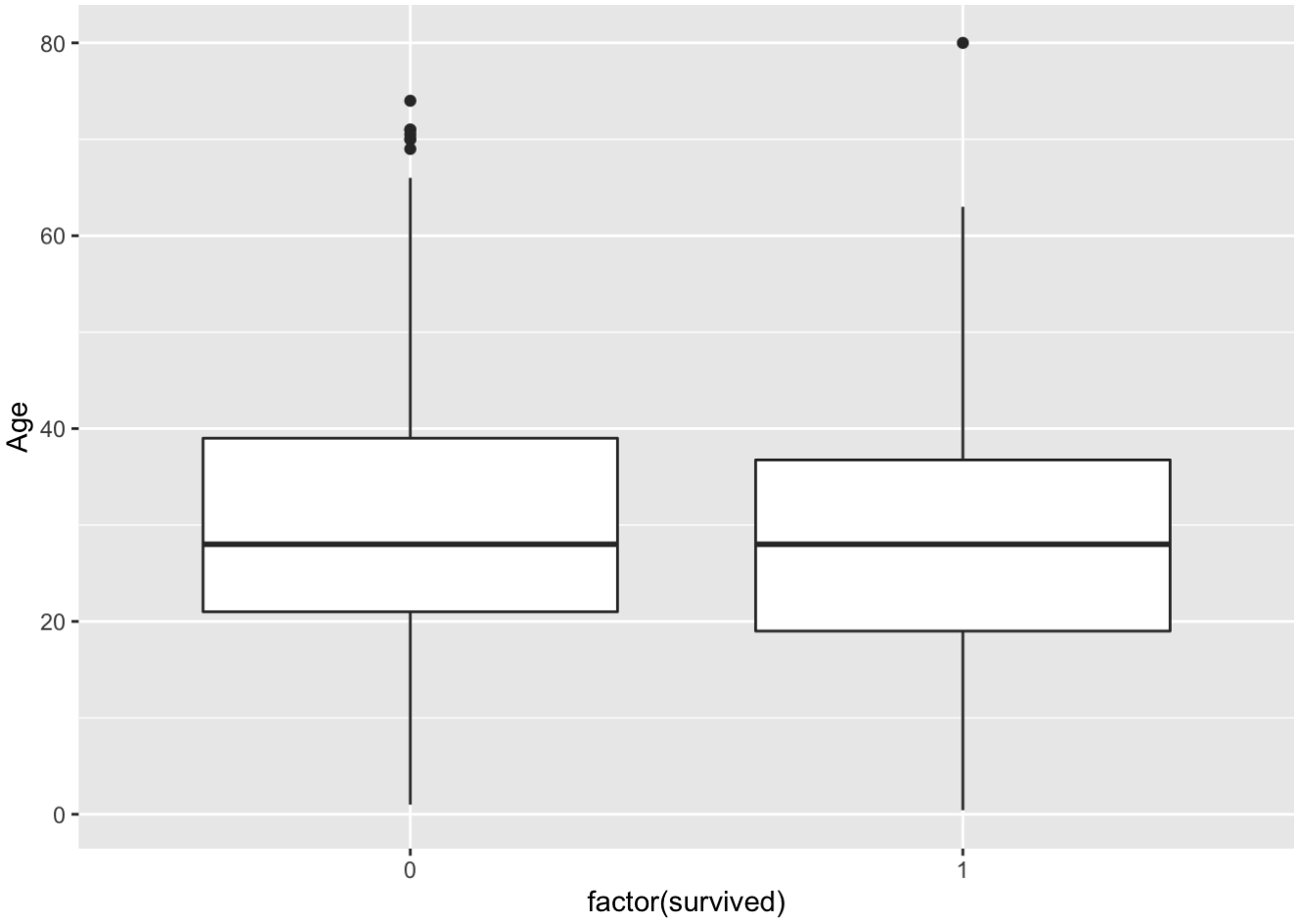
Pclass box vs Age

```
qplot(factor(Pclass), Age, data = titanic_data, geom = "boxplot")
```



Age vs survival box

```
qplot(factor(survived), Age, data = titanic_data, geom = "boxplot")
```

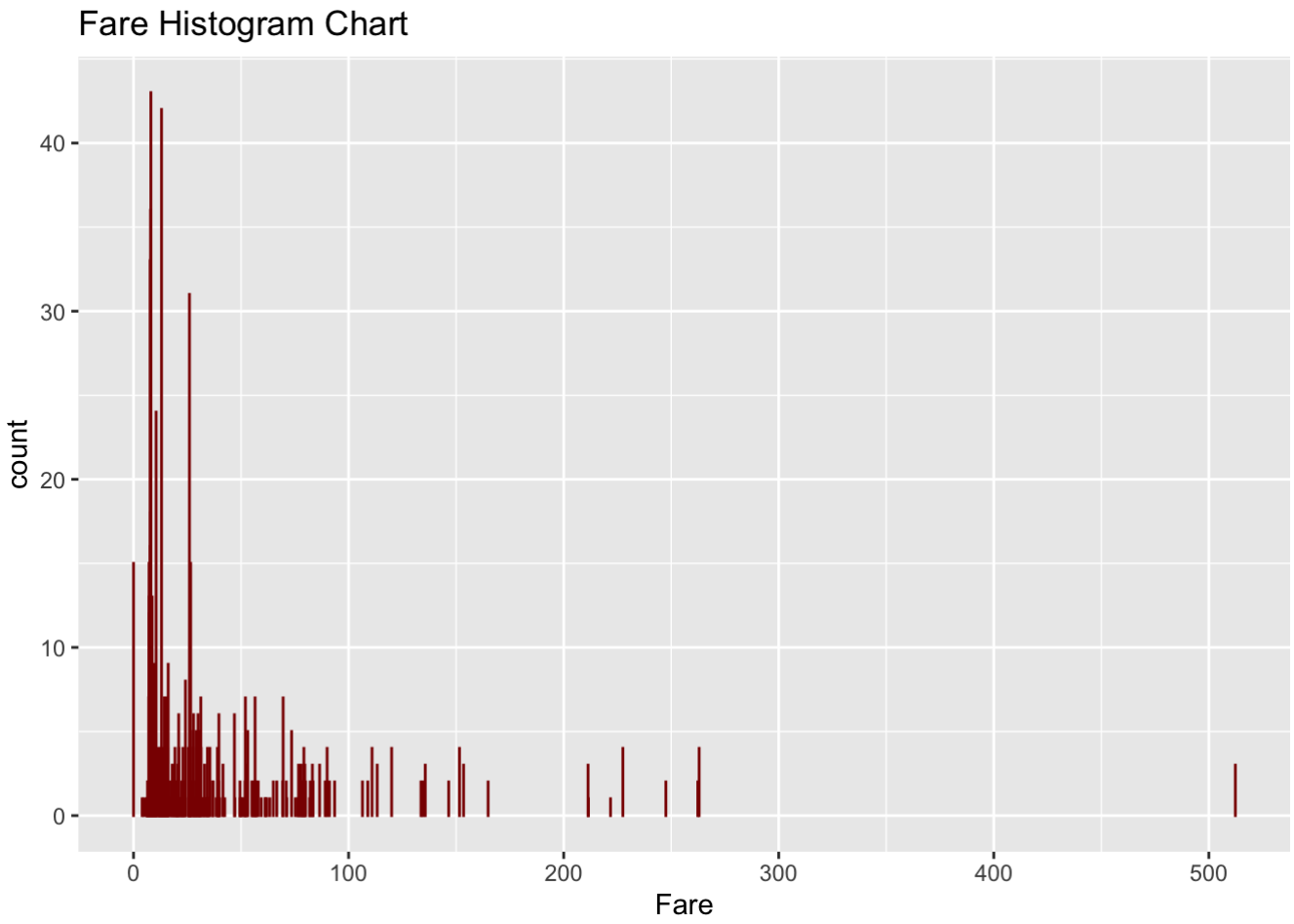


Fare distribution Histogram

```
summary(titanic_data$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.925  14.454  32.305  31.137 512.329
```

```
ggplot.relation.object <- ggplot(titanic_data, aes(x=Fare))
ggplot.relation.object <-ggplot.relation.object+geom_bar(colour="darkred", fill="white")
+ggtitle("Fare Histogram Chart")
ggplot.relation.object
```



Create correlogram that depicts correlation between variables.

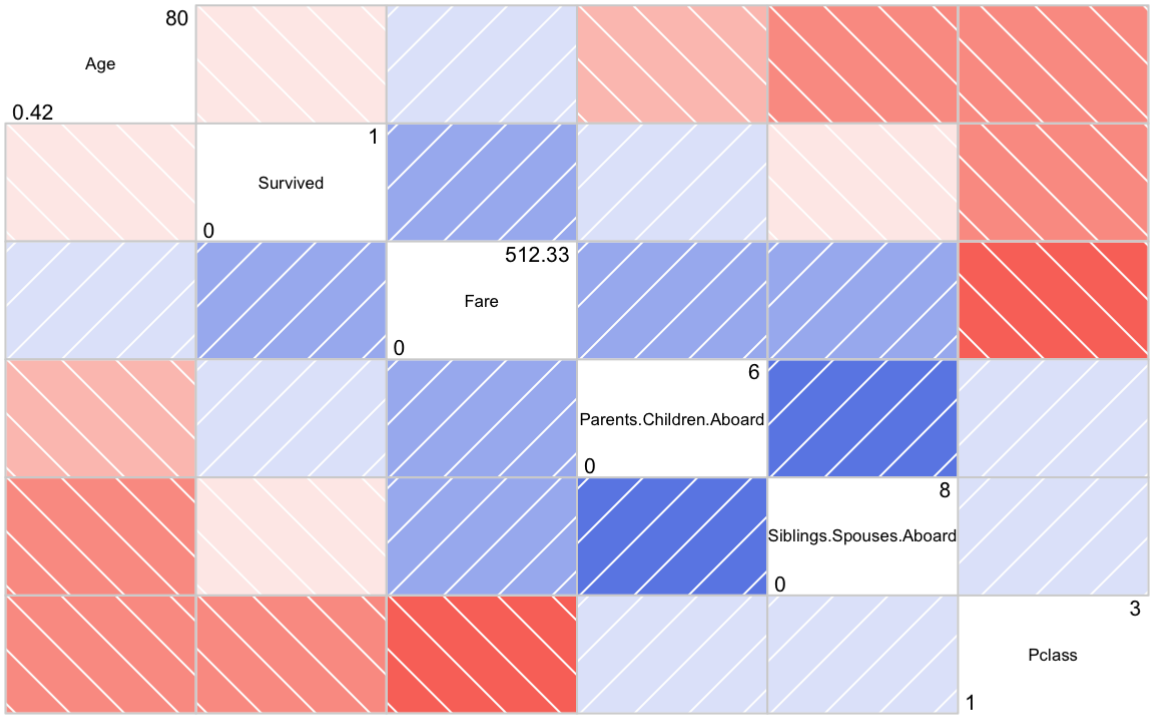
All variables need to be numeric

```
corrgram.data <- titanic_data
## generate correlogram
corrgram.vars <- c("Survived", "Pclass", "Age","Siblings.Spouses.Aboard", "Fare","Parent
s.Children.Aboard")
```

The positive correlations are shown in blue, while the negative correlations are shown in red. The darker the hue, the greater the magnitude of the correlation.

```
corrgram(corrgram.data[,corrgram.vars], order=TRUE,
         text.panel=panel.txt,diag.panel = panel.minmax, main="Titanic Data")
```

Titanic Data



Data Preprocessing

```
titanic_data <- dummy.data.frame(titanic_data, names=c("Pclass","Sex"), sep="_")
```

```
## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):
## non-list contrasts argument ignored

## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):
## non-list contrasts argument ignored
```

BUILDING THE MODEL

Splitting training and test data

carving out the training and testing data sets. we will split our dataset into training set as well as test set with a split specified. This means that 667 rows of our data will be attributed to the train_data whereas the remaining will be attributed to the test data.

```
train <- titanic_data[1:667,]
test <- titanic_data[668:887,]
## Set a random seed
set.seed(754)
```

Model Creation (note: not all possble variables are used)

Next, we feed X_train and y_train into an instance of the Binomial Logistic Regression model class and train the model:

```
model <- glm(factor(Survived) ~ pclass + sex + Age + Siblings.Spouses.Aboard + Fare + Pa
rents.Children.Aboard,family=binomial(link='logit'),data=train)
## Model Summary
summary(model)
```

```
##
## Call:
## glm(formula = factor(Survived) ~ pclass + sex + Age + Siblings.Spouses.Aboard +
##     Fare + Parents.Children.Aboard, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5343  -0.6478  -0.4143   0.6200   2.3873
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               3.9422978  0.5447660   7.237 4.60e-13 ***
## pclass2                  -1.1046497  0.3580968  -3.085  0.00204 **
## pclass3                  -2.3254319  0.3644502  -6.381 1.76e-10 ***
## sexmale                  -2.7324134  0.2258015 -12.101  < 2e-16 ***
## Age                      -0.0358490  0.0087764  -4.085 4.41e-05 ***
## Siblings.Spouses.Aboard -0.3205378  0.1254375  -2.555  0.01061 *
## Fare                     -0.0007928  0.0032100  -0.247  0.80493
## Parents.Children.Aboard -0.1299233  0.1410223  -0.921  0.35690
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 892.88  on 666  degrees of freedom
## Residual deviance: 605.28  on 659  degrees of freedom
## AIC: 621.28
##
## Number of Fisher Scoring iterations: 5
```

TESTING

A chi-square (χ2) statistic is a test that measures how a model compares to actual observed data. χ2 provides a way to test how well a sample of data matches the (known or assumed) characteristics of the larger population that the sample is intended to represent. If the sample data do not fit the expected properties of the population that we are interested in, then we would not want to use this sample to draw conclusions about the larger population.

Using anova() to analyze the table of deviance

```
anova(model, test="Chisq")
```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
	<int>	<dbl>	<int>	<dbl>	<dbl>
NULL	*NA*	*NA*	666	892.8835	*NA*
pclass	2	66.5597518	664	826.3238	3.521547e-15
sex	1	198.3881888	663	627.9356	4.694314e-45
Age	1	11.0011715	662	616.9344	9.105432e-04
Siblings.Spouses.Aboard	1	10.5632050	661	606.3712	1.153610e-03
Fare	1	0.2205716	660	606.1506	6.386047e-01
Parents.Children.Aboard	1	0.8662944	659	605.2843	3.519832e-01

7 rows

Predicting Test Data

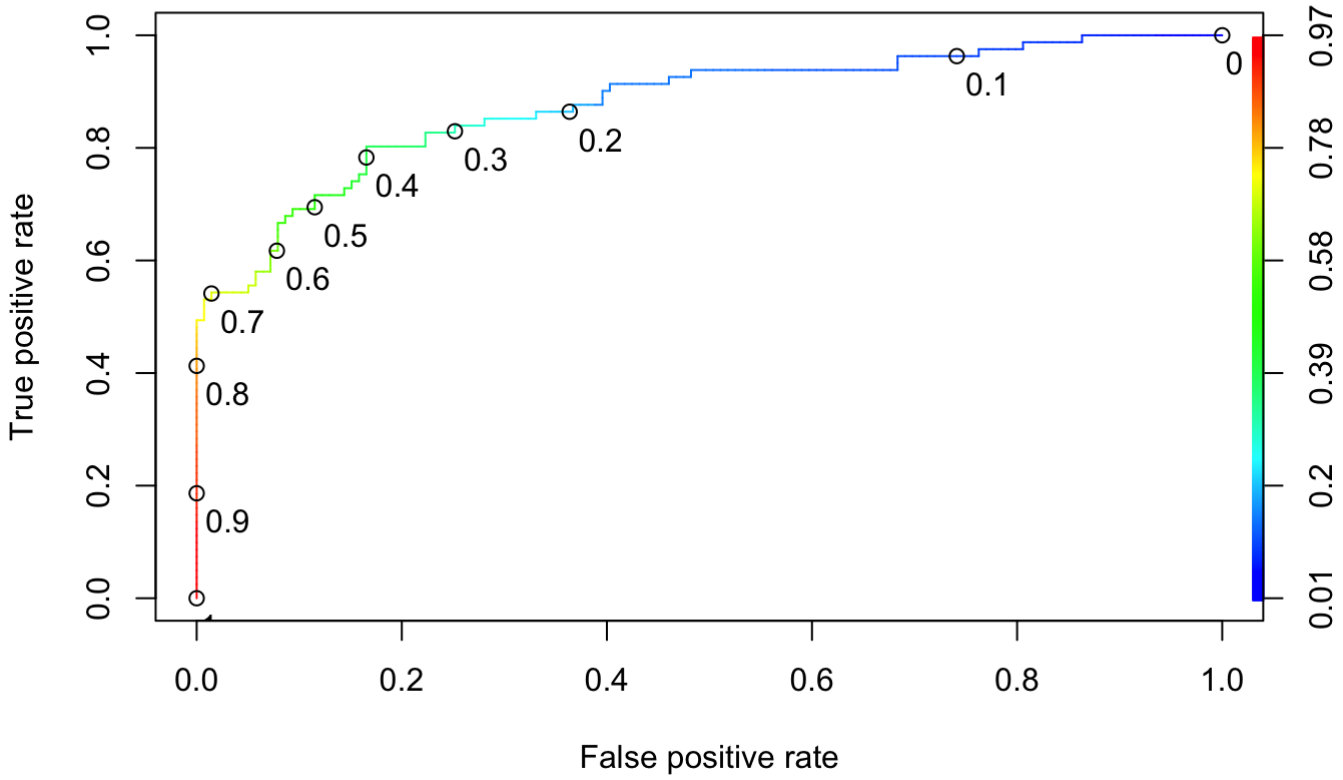
```
result <- predict(model,newdata=test,type='response')
result <- ifelse(result > 0.5,1,0)

## Confusion matrix and statistics
confusionMatrix(as.factor(result),as.factor(test$Survived))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 123  25
##          1  16  56
##
##                Accuracy : 0.8136
##                  95% CI : (0.7558, 0.8628)
##     No Information Rate : 0.6318
##     P-Value [Acc > NIR] : 3.246e-09
##
##                   Kappa : 0.5899
##
##  Mcnemar's Test P-Value : 0.2115
##
##             Sensitivity : 0.8849
##             Specificity : 0.6914
##          Pos Pred Value : 0.8311
##          Neg Pred Value : 0.7778
##              Prevalence : 0.6318
##          Detection Rate : 0.5591
##    Detection Prevalence : 0.6727
##       Balanced Accuracy : 0.7881
##
##        'Positive' Class : 0
##
```

ROC Curve and calculating the area under the curve(AUC)

In order to assess the performance of our model, we will delineate the ROC curve. ROC is also known as Receiver Optimistic Characteristics. For this, we will first import the ROCR package and then plot our ROC curve to analyze its performance. ##### plot for sensitivity vs specificity

```
predictions <- predict(model, newdata=test, type="response")
ROCRpred <- prediction(predictions, test$Survived)
ROCRperf <- performance(ROCRpred, measure = "tpr", x.measure = "fpr")
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7), print.cutoffs.at = seq(0,1,0.1))
```



Area under the ROC curve

AUC provides an aggregate measure of performance across all possible classification thresholds. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

```
auc <- performance(ROCRpred, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

```
## [1] 0.8781419
```

SUMMARY

Concluding our R Data Science project, we developed our titanic survival detection model using machine learning. We used a variety of ML algorithms to implement this model, plotted the respective performance curves for the models and analyzed and visualized dataset from all types of data.

A Binomial Regression model can be used to predict the odds of an event. The Binomial Regression model is a member of the family of Generalized Linear Models which use a suitable link function to establish a relationship between the conditional expectation of the response variable y with a linear combination of explanatory variables X. The Logistic Regression model is a special case of the Binomial Regression model in the situation where the size of each group of explanatory variables in the data set is one.