data.df <-read.csv(paste("C:/Users/raja/Downloads/SixAirlines.csv", sep=""))#Loading the Data set

> attach(data.df)#Attaching the Data ste

> View(data.df)#General view of the entire Data frme

>

>

> #Simple Statistical Analysis using describe funtion

> library(psych)

> describe(data.df)

vars n mean sd median trimmed mad min max range skew kurtosis se

AIRLINE\* 1 462 3.02 1.65 2.00 2.90 1.48 1.00 6.00 5.00 0.59 -0.95 0.08

AIRCRAFT 2 462 0.33 0.47 0.00 0.28 0.00 0.00 1.00 1.00 0.74 -1.46 0.02

FLIGHT\_DURATION 3 462 7.55 3.54 7.75 7.54 4.82 1.25 14.66 13.41 -0.05 -1.12 0.16

MONTH 4 462 1.67 1.05 2.00 1.71 1.48 0.00 3.00 3.00 -0.16 -1.20 0.05

INTERNATIONAL 5 462 0.91 0.28 1.00 1.00 0.00 0.00 1.00 1.00 -2.93 6.60 0.01

SEATS\_ECONOMY 6 462 200.71 77.96 185.00 193.76 85.99 17.00 389.00 372.00 0.61 -0.26 3.63

SEATS\_PREMIUM 7 462 33.54 13.26 36.00 33.20 11.86 8.00 66.00 58.00 0.25 -0.46 0.62

PITCH\_ECONOMY 8 462 31.21 0.66 31.00 31.25 0.00 30.00 33.00 3.00 -0.03 -0.38 0.03

PITCH\_PREMIUM 9 462 37.92 1.32 38.00 38.06 0.00 34.00 40.00 6.00 -1.48 3.43 0.06

WIDTH\_ECONOMY 10 462 17.83 0.56 18.00 17.81 0.00 17.00 19.00 2.00 -0.03 -0.12 0.03

WIDTH\_PREMIUM 11 462 19.48 1.10 19.00 19.54 0.00 17.00 21.00 4.00 -0.09 -0.34 0.05

PRICE\_ECONOMY 12 462 1317.06 989.81 1224.00 1231.30 1163.84 65.00 3593.00 3528.00 0.52 -0.88 46.05

PRICE\_PREMIUM 13 462 1832.35 1289.97 1710.00 1782.94 1852.51 86.00 7414.00 7328.00 0.51 0.41 60.01

PRICE\_RELATIVE 14 462 0.49 0.45 0.38 0.43 0.42 0.02 1.89 1.87 1.14 0.61 0.02

N 15 462 234.25 86.88 227.00 227.69 90.44 38.00 441.00 403.00 0.61 -0.44 4.04

LAMBDA 16 462 0.15 0.06 0.13 0.14 0.03 0.05 0.55 0.50 2.70 14.02 0.00

QUALITY 17 462 6.72 1.78 7.00 6.79 0.00 2.00 10.00 8.00 -0.51 1.67 0.08

>

> library(ggplot2)

Attaching package: ‘ggplot2’

The following objects are masked from ‘package:psych’:

%+%, alpha

> ## Loading required package: ggplot2

> #Seggregating different flights

> ggplot(data.df, aes(x = AIRLINE, fill = AIRLINE)) + geom\_bar()

>

> #Seggregating international and domestic flights

> ggplot(data.df, aes(x = INTERNATIONAL))+ geom\_bar()

>

> #Prices of Economy and Premium tickets

> ggplot(data.df, aes(x = PRICE\_ECONOMY)) + geom\_density()

> ggplot(data.df, aes(x = PRICE\_PREMIUM)) + geom\_density()

>

>

>

> # A Scatterplot of price economy vs flight hours of travel

> # ==========

> plot(FLIGHT\_DURATION,PRICE\_ECONOMY,

+ col="blue",

+ main="Price economy vs flight hours",

+ xlab="Hours", ylab="Price")

>

> # Add the sample means to the Scatterplot

> # ==========

> abline(h=mean(PRICE\_ECONOMY), col="dark blue", lty="dotted")

> abline(v=mean(FLIGHT\_DURATION), col="dark blue", lty="dotted")

>

> # Add a regression line

> # ==========

> abline(lm(PRICE\_ECONOMY ~ FLIGHT\_DURATION))

>

>

> # A Scatterplot of price premium vs flight hours of travel

> # ==========

> plot(FLIGHT\_DURATION,PRICE\_PREMIUM,

+ col="blue",

+ main="Price economy vs flight hours",

+ xlab="Hours", ylab="Price")

>

> # Add the sample means to the Scatterplot

> # ==========

> abline(h=mean(PRICE\_PREMIUM), col="dark blue", lty="dotted")

> abline(v=mean(FLIGHT\_DURATION), col="dark blue", lty="dotted")

>

> # Add a regression line

> # ==========

> abline(lm(PRICE\_PREMIUM ~ FLIGHT\_DURATION))

>

>

> #Correlation and Correlation Matrix for Price Economy

>

> library(corrplot)

Warning message:

package ‘corrplot’ was built under R version 3.4.1

> library(gplots) # for color interpolation

Attaching package: ‘gplots’

The following object is masked from ‘package:stats’:

lowess

> par(mfrow=c(1, 1))

> corrplot.mixed(corr=cor(data.df[ , c(2:6, 8,10,12,15,17)], use="complete.obs"),

+ upper="ellipse", tl.pos="lt",

+ col = colorpanel(50, "red", "gray60", "blue4"))

>

>

> #Correlation and Correlation Matrix for Price Premium

>

> par(mfrow=c(1, 1))

> corrplot.mixed(corr=cor(data.df[ , c(2:6, 7,9,11,15,17)], use="complete.obs"),

+ upper="ellipse", tl.pos="lt",

+ col = colorpanel(50, "red", "gray60", "blue4"))

>

>

>

> #Scatter Plot Matrix for Price Economy and Price Premium

> library(car)

Attaching package: ‘car’

The following object is masked from ‘package:psych’:

logit

> scatterplotMatrix(formula = ~ SEATS\_ECONOMY + PITCH\_ECONOMY + WIDTH\_ECONOMY + PRICE\_ECONOMY, cex=0.6,

+ data=data.df, diagonal="histogram")

Warning message:

In smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE, spread = spread, :

could not fit smooth

>

> scatterplotMatrix(formula = ~ SEATS\_PREMIUM + PITCH\_PREMIUM + WIDTH\_PREMIUM + PRICE\_PREMIUM, cex=0.6,

+ data=data.df, diagonal="histogram")

Warning message:

In smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE, spread = spread, :

could not fit smooth

>

> #Calculating correlations between Prices of Economy and Premium in correlation to other factors

> cor.test(PRICE\_ECONOMY, PITCH\_ECONOMY)

Pearson's product-moment correlation

data: PRICE\_ECONOMY and PITCH\_ECONOMY

t = 8.8003, df = 460, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.2987210 0.4550742

sample estimates:

cor

0.379605

> cor.test(PRICE\_ECONOMY, WIDTH\_ECONOMY)

Pearson's product-moment correlation

data: PRICE\_ECONOMY and WIDTH\_ECONOMY

t = 1.764, df = 460, p-value = 0.0784

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.009330795 0.171911298

sample estimates:

cor

0.0819679

> cor.test(PRICE\_PREMIUM, PITCH\_PREMIUM)

Pearson's product-moment correlation

data: PRICE\_PREMIUM and PITCH\_PREMIUM

t = 1.5338, df = 460, p-value = 0.1258

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.02002801 0.16150915

sample estimates:

cor

0.07133125

> cor.test(PRICE\_PREMIUM, WIDTH\_PREMIUM)

Pearson's product-moment correlation

data: PRICE\_PREMIUM and WIDTH\_PREMIUM

t = 1.0592, df = 460, p-value = 0.2901

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.04209336 0.13992426

sample estimates:

cor

0.04932498

>

> #Using the Boruta package to calcuate the effectiveness of different variables in Calculating the price of Economy class tickets

> eco.df <-read.csv(paste("C:/Users/USER/Downloads/Economy.csv", sep=""))

Error in file(file, "rt") : cannot open the connection

In addition: Warning message:

In file(file, "rt") :

cannot open file 'C:/Users/USER/Downloads/Economy.csv': No such file or directory

> library(Boruta)

Loading required package: ranger

Warning messages:

1: package ‘Boruta’ was built under R version 3.4.1

2: package ‘ranger’ was built under R version 3.4.1

> set.seed(1234) # for code reproducibility

> response <- data.df$PRICE\_ECONOMY

> bor.results <- Boruta(eco.df,response,maxRuns=101,doTrace=0)

Error in Boruta(eco.df, response, maxRuns = 101, doTrace = 0) :

object 'eco.df' not found

> plot(bor.results)

Error in plot(bor.results) : object 'bor.results' not found

>

> #Using the Boruta package to calcuate the effectiveness of different variables in Calculating the price of Premium class tickets

> pre.df <-read.csv(paste("C:/Users/USER/Downloads/prem.csv", sep=""))

Error in file(file, "rt") : cannot open the connection

In addition: Warning message:

In file(file, "rt") :

cannot open file 'C:/Users/USER/Downloads/prem.csv': No such file or directory

> library(Boruta)

> set.seed(1234) # for code reproducibility

> response <- data.df$PRICE\_PREMIUM

> bor.results <- Boruta(pre.df,response,maxRuns=101,doTrace=0)

Error in Boruta(pre.df, response, maxRuns = 101, doTrace = 0) :

object 'pre.df' not found

> plot(bor.results)

Error in plot(bor.results) : object 'bor.results' not found

>

>

> #Dividing the Data set into Test and Training Data ste

> ratio = sample(1:nrow(data.df), size = 0.25\*nrow(data.df))

> Test = data.df[ratio,] #Test dataset 25% of total

> Training = data.df[-ratio,] #Train dataset 75% of total

> dim(Training)

[1] 347 17

> dim(Test)

[1] 115 17

>

> #Generating A Multi Variable Linear Regressional Model for Price of Economy Flights

> linear.mod<- lm(PRICE\_ECONOMY~ PITCH\_ECONOMY + WIDTH\_ECONOMY + FLIGHT\_DURATION + QUALITY + PRICE\_RELATIVE, data = Training)

> summary(linear.mod)

Call:

lm(formula = PRICE\_ECONOMY ~ PITCH\_ECONOMY + WIDTH\_ECONOMY +

FLIGHT\_DURATION + QUALITY + PRICE\_RELATIVE, data = Training)

Residuals:

Min 1Q Median 3Q Max

-1581.47 -485.24 33.63 552.87 1609.05

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -7284.17 3101.37 -2.349 0.0194 \*

PITCH\_ECONOMY 504.58 94.34 5.348 1.63e-07 \*\*\*

WIDTH\_ECONOMY -525.36 70.38 -7.464 7.04e-13 \*\*\*

FLIGHT\_DURATION 182.93 11.65 15.701 < 2e-16 \*\*\*

QUALITY 190.28 33.13 5.744 2.05e-08 \*\*\*

PRICE\_RELATIVE -913.89 86.78 -10.532 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 632.4 on 341 degrees of freedom

Multiple R-squared: 0.5674, Adjusted R-squared: 0.5611

F-statistic: 89.47 on 5 and 341 DF, p-value: < 2.2e-16

>

> #the t value of Pitch\_economy and quality is positive indicating that these predictors are associated with

> #Price\_economy. A larger t-value indicates that that it is less likely that the coefficient is not equal to zero purely by chance.

> #Again, as the p-value for Flight\_Duration and Price\_Relative is less than 0.05 they are both statistically significant in the multiple linear regression model for Price\_Economy response variable.

> #The model's, p-value: < 2.2e-16 is also lower than the statistical significance level of 0.05, this indicates that we can safely reject the null hypothesis that the value for the coefficient is zero

> #(or in other words, the predictor variable has no explanatory relationship with the response variable).

> #The model has a F Statistic of 90, which is considerably high

> library(rpart)

> library(randomForest)

randomForest 4.6-12

Type rfNews() to see new features/changes/bug fixes.

Attaching package: ‘randomForest’

The following object is masked from ‘package:ranger’:

importance

The following object is masked from ‘package:ggplot2’:

margin

The following object is masked from ‘package:psych’:

outlier

Warning message:

package ‘randomForest’ was built under R version 3.4.1

> model.forest <- randomForest(PRICE\_ECONOMY~ PITCH\_ECONOMY + WIDTH\_ECONOMY + FLIGHT\_DURATION + QUALITY + PRICE\_RELATIVE, data = Training, method = "anova",

+ ntree = 300,

+ mtry = 2, #mtry is sqrt(6)

+ replace = F,

+ nodesize = 1,

+ importance = T)

>

> varImpPlot(model.forest)

> #From the VIF plot we see that Flight Duration and Price Relative are most important factors in predicitng Price Economy.

>

> #We test the model using Random Forest

> prediction <- predict(model.forest,Test)

> rmse <- sqrt(mean((log(prediction)-log(Test$PRICE\_ECONOMY))^2))

> round(rmse, digits = 3)

[1] 0.347

>

> # Evaluation metric function

> #A custom root mean Square Function to evaluate the performance of our model

> RMSE <- function(x,y)

+ {

+ a <- sqrt(sum((log(x)-log(y))^2)/length(y))

+ return(a)

+ }

>

> #Implementing the Regression Tree Model

> model <- rpart(PRICE\_ECONOMY~ PITCH\_ECONOMY + WIDTH\_ECONOMY + FLIGHT\_DURATION + QUALITY + PRICE\_RELATIVE, data = Training, method = "anova")

> predict <- predict(model, Test)

> RMSE1 <- RMSE(predict, Test$PRICE\_ECONOMY)

> RMSE1 <- round(RMSE1, digits = 3)

> RMSE1

[1] 0.519

>

> #For Premium Class Tickets

>

> #Generating A Multi Variable Linear Regressional Model for Price of Premium Flights

> linear.mod<- lm(PRICE\_PREMIUM~ PITCH\_PREMIUM + WIDTH\_PREMIUM + FLIGHT\_DURATION + QUALITY + PRICE\_RELATIVE, data = Training)

> summary(linear.mod)

Call:

lm(formula = PRICE\_PREMIUM ~ PITCH\_PREMIUM + WIDTH\_PREMIUM +

FLIGHT\_DURATION + QUALITY + PRICE\_RELATIVE, data = Training)

Residuals:

Min 1Q Median 3Q Max

-2137.3 -558.9 -104.5 733.7 4562.8

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4550.89 4532.40 -1.004 0.31605

PITCH\_PREMIUM 96.55 137.53 0.702 0.48315

WIDTH\_PREMIUM 79.81 74.68 1.069 0.28599

FLIGHT\_DURATION 226.51 16.10 14.072 < 2e-16 \*\*\*

QUALITY -61.46 106.99 -0.574 0.56605

PRICE\_RELATIVE -349.15 132.26 -2.640 0.00867 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 933.5 on 341 degrees of freedom

Multiple R-squared: 0.4428, Adjusted R-squared: 0.4346

F-statistic: 54.2 on 5 and 341 DF, p-value: < 2.2e-16

> #The model has an F Statistic of 48.4 which is mediumly high

> #the t value of Pitch\_premium, width\_premium, Price\_relative and quality is positive indicating that these predictors are associated with

> #Price\_Premium. A larger t-value indicates that that it is less likely that the coefficient is not equal to zero purely by chance.

> #Again, as the p-value for Flight\_Duration is less than 0.05 they are both statistically significant in the multiple linear regression model for Price\_Economy response variable.

> #The model's, p-value: < 2.2e-16 is also lower than the statistical significance level of 0.05, this indicates that we can safely reject the null hypothesis that the value for the coefficient is zero

> #(or in other words, the predictor variable has no explanatory relationship with the response variable).

>

> library(rpart)

> library(randomForest)

> model.forest <- randomForest(PRICE\_PREMIUM~ PITCH\_PREMIUM + WIDTH\_PREMIUM + FLIGHT\_DURATION + QUALITY + PRICE\_RELATIVE, data = Training, method = "anova",

+ ntree = 300,

+ mtry = 2, #mtry is sqrt(6)

+ replace = F,

+ nodesize = 1,

+ importance = T)

>

> varImpPlot(model.forest)

> #From the VIF plot we see that Flight Duration and Price Relative are most important factors in predicitng Price Economy.

>

> # Evaluation metric function

> #A custom root mean Square Function to evaluate the performance of our model

> RMSE <- function(x,y)

+ {

+ a <- sqrt(sum((log(x)-log(y))^2)/length(y))

+ return(a)

+ }

>

> #Implementing the Regression Tree Model

> model <- rpart(PRICE\_ECONOMY~ PITCH\_ECONOMY + WIDTH\_ECONOMY + FLIGHT\_DURATION + QUALITY + PRICE\_RELATIVE, data = Training, method = "anova")

> predict <- predict(model, Test)

> RMSE1 <- RMSE(predict, Test$PRICE\_ECONOMY)

> RMSE1 <- round(RMSE1, digits = 3)

> RMSE

function(x,y)

{

a <- sqrt(sum((log(x)-log(y))^2)/length(y))

return(a)

}