## Predicting the Flight Arrival Delays

#Read the data from the working directory, create your own working directly to read the dataset.

setwd("C:/Users/Deep/Desktop/data")

data1 <- read.csv ("C:/Users/Deep/Desktop/data/

flight\_delay.csv",header=TRUE,sep=",")

## Performing Data exploration

#perform exploratory data analysis to know about the data

# display top 6 rows of dataset to see how data look like

head (data1)

# display bottom 6 rows

tail(data1)

# describe the structure of data

str(data1)

#display the column name of the data

names(data1)

#display the summary or descriptive statistics of the data

summary(data1$Arr\_Delay)

#Let’s check the missing values present in the data

sum(is.na(data2))

#To find out the correlation between the variables

corr <- cor.test(data1$Arr\_Delay,data1$Number\_of\_flights, method = "pearson" )

corr

#to add four charts in one window or plotting panel

par(mfrow = c(2,2))

#To plot the dependent and independent variable

plot(data1$Arr\_Delay,data1$Number\_of\_flights)

plot(data1$Arr\_Delay,data1$Security\_o)

plot(data1$Arr\_Delay,data1$Support\_Crew\_Available)

plot(data1$Arr\_Delay,data1$Airport\_Distance)

#To drop the first variable from data1

data2 <- data1[-c(1)]

#To check the correlation between the variables

cor(data2)

#Splitting dataset into training and testing dataset

#Install caTools package for splitting the data

install.packages("caTools")

library(caTools)

#To reproduce the sample

set.seed(1000)

sample <- sample.split(data2$Arr\_Delay,SplitRatio=0.70)

sample

#split of the data using subset command

train\_data <- subset(data2,sample==TRUE)

test\_data <- subset(data2,sample==FALSE)

### Model Building & Interpretation on Training and Testing Data

#Building multiple linear regression model using lm on train\_data set

model <- lm(Arr\_Delay ~., data = train\_data)

summary(model)

#Building final multiple linear regression model with significant variables on train\_data set

model\_sig<-lm(Arr\_Delay~Airport\_Distance+Number\_of\_flights

+Weather+Support\_Crew\_Available+Baggage\_loading\_time +Late\_Arrival\_o, data= train\_data)

model\_sig

#To look at the names of model\_sig

names(model\_sig)

#number of the fitted values

length(model\_sig$fitted.values)

#predicting fitted values of train\_data set

pred\_train<- model\_sig$fitted.values

head(pred\_train)

pred\_train1 <- data.frame(pred\_train)

#residual values

resed\_train <- model\_sig$residuals

head(resed\_train)

resed\_train1<-data.frame(resed\_train)

#Prediction on the unforeseen dataset i.e on test data

pred\_test<- predict(model\_sig,newdata = test\_data)

head(pred\_test)

pred\_test1<- data.frame(pred\_test)

#Plotting Actual versus Predicted outcome

plot (test\_data$Arr\_Delay,col="red",type ="l",lty=1.8 )

lines(pred\_test1,col="blue",type ="l",lty=1.4)

#Linear Regression Diagnostics statistics and plot

# plot function to view four different diagnostic plots

# To view first plot

plot(model\_sig,which=1)

# Let's look at our second plot now.

plot(model\_sig,which=2)

# Let's look at our third plot now.

plot(model\_sig,which=3)

# Let's look at our final plot(fourth) now.

#Note that this specific plot ID is considered '5' in R.hence mentioned as which=5.

plot(model\_sig,which=5)

#Install car package

install.packages("car")

library(car)

# Test for Independence Assumption

durbinWatsonTest(model\_sig)

# Statistic Test for Homoscedasticity Assumption

ncvTest(model\_sig)

#VIF Test for Collinearity

vif(model\_sig)