setwd ("D://RPA/Acadgild Assignments/BusinessAnalytics\_Assignment 17")

getwd ()

Dataset <- read.csv ("Dataset.csv", header=TRUE)

head (Dataset)

Dataset\_new <- Dataset

str (Dataset)

#Summary Statistics

summary (Dataset)

#Checking for Missing Values

for (i in 1:12)

{

print(sum(is.na(Dataset[,i])))

}

#Variable transformation

Dataset$Cruise <- as.factor (Dataset$Cruise)

Dataset$Sound <- as.factor (Dataset$Sound)

Dataset$Leather <- as.factor (Dataset$Leather)

#Outlier detection

stats\_out <- boxplot.stats(Dataset$Price)$out

stats\_out

boxplot (Dataset$Price, horizontal=T)

#Plotting Price~Mileage

plot (Dataset$Price~Dataset$Mileage)

#Subsetting dataset with Price > 52000

cars\_r\_out <- Dataset[Dataset$Price>52000,]

cars\_r\_out

#Outlier values- Cars with Price>52000 are on the expensive side and can't be considered as outliers

plot (cars\_r\_out$Price~cars\_r\_out$Mileage)

#Creating a simple Linear Regression Model with Price as DV and Mileage as IV

cars\_r\_lm1 <- lm (Price~Mileage,data=Dataset)

cars\_r\_lm1

summary(cars\_r\_lm1)# R-sq : 0.02046 or 2%

#Creating a simple Linear Regression Model with Price as DV and Mileage as IV for cars with Price>52000

cars\_r\_lm2 <- lm (Price~Mileage, data=cars\_r\_out)

cars\_r\_lm2

summary(cars\_r\_lm2)# R-sq : 0.986 or 98.6%

#Create dummy variables for Make variable

class (Dataset$Make)

Make\_factor <- model.matrix(~Dataset$Make -1)

head(Make\_factor)

#Variable with k levels require only k-1 dummy variables

Make\_factor <- data.frame(Make\_factor)#Converting from matrix to data.frame

factor\_length <- length(names(Make\_factor))-1

Dataset\_new <- cbind (Dataset\_new,Make\_factor[1:factor\_length] )

head (Dataset\_new)

#Multiple Linear Regression

cars\_r\_mlm1 <- lm (Price~Mileage+Cylinder+Liter+Doors+Cruise+Sound+Leather+

Dataset.MakeBuick+Dataset.MakeCadillac+

Dataset.MakeChevrolet+Dataset.MakePontiac+Dataset.MakeSAAB,

data=Dataset\_new)

cars\_r\_mlm1

summary(cars\_r\_mlm1)

#Create dummy variables for Type Variable and including in the next iteration of MLR

class (Dataset\_new$Type)

Type\_factor <- model.matrix(~Dataset\_new$Type -1)

head(Type\_factor)

#Variable with k levels require only k-1 dummy variables

Type\_factor <- data.frame(Type\_factor)#Converting from matrix to data.frame

factor\_length1 <- length(names(Type\_factor))-1

Dataset\_new <- cbind (Dataset\_new,Type\_factor[1:factor\_length1] )

head (Dataset\_new)

#Multiple Linear Regression

cars\_r\_mlm2 <- lm (Price~Mileage+Liter+Doors+

Dataset.MakeBuick+Dataset.MakeCadillac+

Dataset.MakeSAAB+Dataset\_new.TypeConvertible+

Dataset\_new.TypeCoupe+Dataset\_new.TypeHatchback+

Dataset\_new.TypeSedan,

data=Dataset\_new)

cars\_r\_mlm2

summary(cars\_r\_mlm2)

#Second iteration of MLR with added Type variable has resulted in

#better R-sq and Adj R-sq

#All variables included are significant in predicting the Price

#Predicted value from MLR cars\_r\_mlm2

Dataset\_p <- predict (cars\_r\_mlm2)

#Residual value

Dataset\_res <- resid(cars\_r\_mlm2)

#Take the log of Price and generate the regression equation

#Multiple Linear Regression

cars\_r\_mlm3 <- lm (log(Price)~Mileage+Liter+Doors+

Dataset.MakeBuick+Dataset.MakeCadillac+

Dataset.MakeSAAB+Dataset\_new.TypeConvertible+

Dataset\_new.TypeHatchback+

Dataset\_new.TypeSedan,

data=Dataset\_new)

cars\_r\_mlm3

summary(cars\_r\_mlm3)

#Log(Price) is giving a better R-sq and Adj R-sq over the above model

#Therefore we can finalise on this model

#All IV's are significant in predicting the log (Price)

#Predicted value from MLR cars\_r\_mlm2

Dataset\_pl <- predict(cars\_r\_mlm3)

#Residual value

Dataset\_resl <- resid(cars\_r\_mlm3)

plot (Dataset\_new$Price,Dataset\_resl,xlab="Price",ylab="Residuals")

plot (Dataset\_pl,Dataset\_resl,xlab="Pred Price",ylab="Residuals",abline (0,0))

#Distribution above and below are not skewed and near symmetric

#So we can rule out heteroscedacity

#Actual vs Predicted Price Plot

plot (Dataset\_new$Price, col="green",type="l")

par (new=TRUE)

plot (Dataset\_pl,col="red",type="l")

#The actual and predicted price are comparable and close

par(mfrow=c(2,2))

plot (cars\_r\_mlm3)

par(mfrow=c(2,1))

#Comparing the distribution of Residual of Price and Log(Price)

hist (Dataset\_res)

hist (Dataset\_resl)

#Residual of Log(Price) is showing a better normal distribution that residual of Price without

#log transformation