#Linear regression model

setwd("D:\\DS 1")

data=read.table("auto-mpg1.data",na.strings = "?")

names(data)= c("mpg", "cylinders", "displacement", "horsepower", "weight", "acceleration", "model", "origin", "car")

View(data)

data1=data[-c(45,112,328),-9]

View(data1)

# check missing values

sum(is.na(data1$origin))

sum(is.na(data1$mpg))

sum(is.na(data1$cylinders))

sum(is.na(data1$displacement))

sum(is.na(data1$horsepower))

sum(is.na(data1$weight))

sum(is.na(data1$acceleration))

sum(is.na(data1$model))

attach(data1)

hist(horsepower)

median(horsepower,na.rm = T)

data1$horsepower[is.na(data1$horsepower)]=93.5

sum(is.na(horsepower))

View(data1)

summary(data1)

# find outliers and cap it

bx=boxplot(mpg)

boxplot(cylinders)

boxplot(displacement)

bx=boxplot(horsepower)

quantile(horsepower,seq(0,1,0.02))

bx$stats

horsepower=ifelse(horsepower>198,190,horsepower)

boxplot(horsepower)

boxplot(weight)

bx=boxplot(acceleration)

quantile(acceleration,seq(0,1,0.02))

bx$stats

acceleration=ifelse(acceleration>22,21,acceleration)

acceleration=ifelse(acceleration<9,10,acceleration)

boxplot(acceleration)

boxplot(model)

boxplot(origin)

library(car)

par(mfrow=c(1,2))

names(data1)

# split the data

split= sample.split(data1$mpg,SplitRatio = 0.8)

train\_data= subset(data1, split=="TRUE")

test\_data= subset(data1, split=="FALSE")

# correlation plot

library(corrplot)

cor= cor(train\_data)

corrplot(cor, type="lower")

corrplot(cor,method="number")

# build a model

mod=lm(mpg~ cylinders+displacement+horsepower+weight+acceleration+model+origin, data=train\_data)

summary(mod)

vif(mod)

mod1=lm(mpg~ cylinders+horsepower+weight+acceleration+model+origin, data=train\_data)

summary(mod1)

vif(mod1)

step(mod1)

mod3=lm(mpg ~ weight + model + origin,data=data1)

summary(mod3)

library(lmtest)

par(mfrow=c(1,1))

library(car)

plot(mod3)

quantile(train\_data$mpg,seq(0,1,0.05))

finaldata = train\_data[(train\_data$mpg >=13 & train\_data$mpg <=37),]

mod4=lm(mpg ~ weight + model + origin,data=finaldata)

summary(mod4)

par(mfrow=c(1,2))

plot(mod4,which = 1)

plot(mod4,which = 2)

plot(mod4,which = 3)

plot(mod4,which = 4)

# for checking autocorrelation

durbinWatsonTest(mod4)

hist(residuals(mod4))

plot(finaldata$mpg,residuals(mod4))

library(predictmeans)

cooksd= CookD(mod4)

predict=predict(mod4,data=newdata)

View(predict)

library(Metrics)

rmse(finaldata$mpg,predict)

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

lines(predict,type="l",col="blue")