**##################Multi linear Regression#####################################**

**#Problem Statement:Consider only the below columns and prepare a prediction model for #predicting Price.**

**#Corolla<-Corolla[c("Price","Age\_08\_04","KM","HP","cc","Doors","Gears","Quarterly\_Tax","Weight")]**

**# Data : ToyotaCorolla.csv**

**###########################################################################**

install.packages("psych")

library(psych)

# Diagnostic Plots

install.packages("car")

library(car)

toyotaDetails <- read.csv(file.choose()) #ToyotaCorolla.csv

toyotaDetails<-toyotaDetils[c("Price","Age\_08\_04","KM","HP","cc","Doors","Gears","Quarterly\_Tax","Weight")]

attach(toyotaDetails)

View(toyotaDetails)

summary(toyotaDetails)

#Price Age\_08\_04 KM HP cc Doors

#Min. : 4350 Min. : 1.00 Min. : 1 Min. : 69.0 Min. : 1300 Min. :2.000

#1st Qu.: 8450 1st Qu.:44.00 1st Qu.: 43000 1st Qu.: 90.0 1st Qu.: 1400 1st Qu.:3.000

#Median : 9900 Median :61.00 Median : 63390 Median :110.0 Median : 1600 Median :4.000

#Mean :10731 Mean :55.95 Mean : 68533 Mean :101.5 Mean : 1577 Mean :4.033

#3rd Qu.:11950 3rd Qu.:70.00 3rd Qu.: 87021 3rd Qu.:110.0 3rd Qu.: 1600 3rd Qu.:5.000

#Max. :32500 Max. :80.00 Max. :243000 Max. :192.0 Max. :16000 Max. :5.000

#Gears Quarterly\_Tax Weight

#Min. :3.000 Min. : 19.00 Min. :1000

#1st Qu.:5.000 1st Qu.: 69.00 1st Qu.:1040

#Median :5.000 Median : 85.00 Median :1070

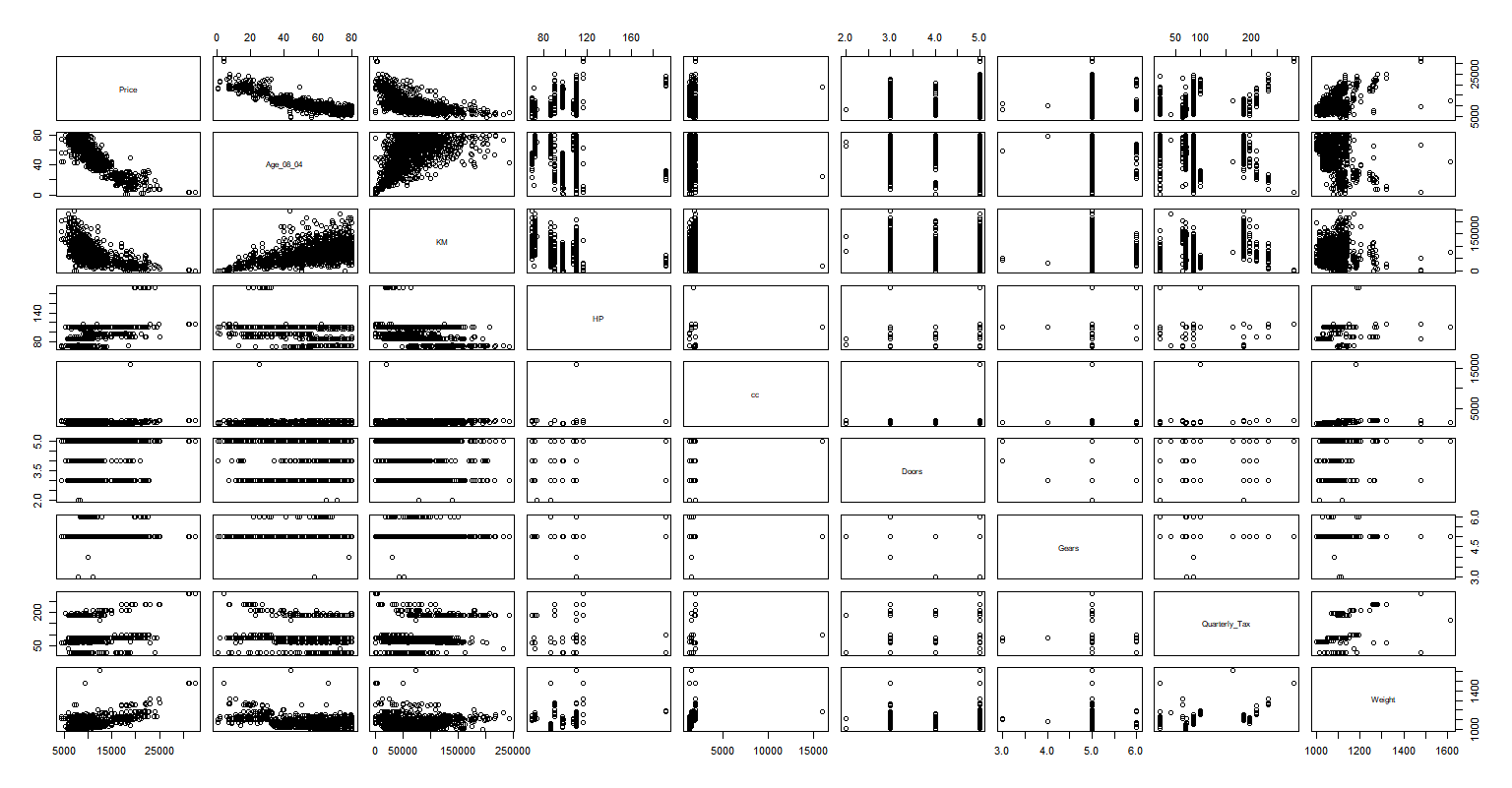
#Mean :5.026 Mean : 87.12 Mean :1072

#3rd Qu.:5.000 3rd Qu.: 85.00 3rd Qu.:1085

#Max. :6.000 Max. :283.00 Max. :1615

#Scatter plot for all pairs of variables.

pairs(toyotaDetails)



##From plots we can infer that the newer the car(Age) is more expensive it is.

#More miles(KM) a car has the cheaper it is.

#More Horse power(Hp) more expensive the car is but not always the case.

#Number of doors dont affect price of the car.

#The heavier the car is more costly it is.

# correlation matrix

cor(toyotaDetails)

# Price Age\_08\_04 KM HP cc Doors Gears

#Price 1.00000000 -0.876590497 -0.56996016 0.31498983 0.12638920 0.18532555 0.063103857

#Age\_08\_04 -0.87659050 1.000000000 0.50567218 -0.15662202 -0.09808374 -0.14835921 -0.005363947

#KM -0.56996016 0.505672180 1.00000000 -0.33353795 0.10268289 -0.03619661 0.015023328

#HP 0.31498983 -0.156622020 -0.33353795 1.00000000 0.03585580 0.09242450 0.209477146

#cc 0.12638920 -0.098083739 0.10268289 0.03585580 1.00000000 0.07990330 0.014629352

#Doors 0.18532555 -0.148359215 -0.03619661 0.09242450 0.07990330 1.00000000 -0.160141430

#Gears 0.06310386 -0.005363947 0.01502333 0.20947715 0.01462935 -0.16014143 1.000000000

#Quarterly\_Tax 0.21919691 -0.198430508 0.27816470 -0.29843172 0.30699580 0.10936323 -0.005451955

#Weight 0.58119759 -0.470253184 -0.02859846 0.08961406 0.33563740 0.30261764 0.020613284

# Quarterly\_Tax Weight

#Price 0.219196911 0.58119759

#Age\_08\_04 -0.198430508 -0.47025318

#KM 0.278164697 -0.02859846

#HP -0.298431717 0.08961406

#cc 0.306995798 0.33563740

#Doors 0.109363225 0.30261764

#Gears -0.005451955 0.02061328

#Quarterly\_Tax 1.000000000 0.62613373

#Weight 0.626133733 1.00000000

#Building multi linear regression model

model <- lm(Price ~., data= toyotaDetails)

summary(model)

#Call:

# lm(formula = Price ~ ., data = toyotaDetails)

#Residuals:

# Min 1Q Median 3Q Max

#-9366.4 -793.3 -21.3 799.7 6444.0

#Coefficients:

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

#(Intercept) -5.573e+03 1.411e+03 -3.949 8.24e-05 \*\*\*

# Age\_08\_04 -1.217e+02 2.616e+00 -46.512 < 2e-16 \*\*\*

# KM -2.082e-02 1.252e-03 -16.622 < 2e-16 \*\*\*

# HP 3.168e+01 2.818e+00 11.241 < 2e-16 \*\*\*

# cc -1.211e-01 9.009e-02 -1.344 0.17909

#Doors -1.617e+00 4.001e+01 -0.040 0.96777

#Gears 5.943e+02 1.971e+02 3.016 0.00261 \*\*

# Quarterly\_Tax 3.949e+00 1.310e+00 3.015 0.00262 \*\*

# Weight 1.696e+01 1.068e+00 15.880 < 2e-16 \*\*\*

# Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Residual standard error: 1342 on 1427 degrees of freedom

#Multiple R-squared: 0.8638, Adjusted R-squared: 0.863

#F-statistic: 1131 on 8 and 1427 DF, p-value: < 2.2e-16

#**Multiple R-squared: 0.8638, Adjusted R-squared: 0.863**, cc and doors are influence to each other.

#Checking significane by building individual models respectively.

**##Model building using cc variable.**

ccmodel <- lm(Price ~ cc, data=toyotaDetails)

summary(ccmodel)

#Call:

# lm(formula = Price ~ cc, data = toyotaDetails)

#Residuals:

# Min 1Q Median 3Q Max

#-7360.2 -2305.8 -855.8 1194.2 21312.1

#Coefficients:

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

#(Intercept) 9027.5548 365.5755 24.694 < 2e-16 \*\*\*

# cc 1.0802 0.2239 4.825 1.55e-06 \*\*\*

#Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Residual standard error: 3599 on 1434 degrees of freedom

#Multiple R-squared: 0.01597, Adjusted R-squared: 0.01529

#F-statistic: 23.28 on 1 and 1434 DF, p-value: 1.551e-06

##cc variable is significant in building model.

##Model building using Doors variable

doorsModel <- lm(Price ~ Doors, data= toyotaDetails)

summary(doorsModel)

#Call:

# lm(formula = Price ~ Doors, data = toyotaDetails)

#Residuals:

# Min 1Q Median 3Q Max

#-7062.8 -2251.7 -915.3 958.0 21087.2

#Coefficients:

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

#(Intercept) 7885.01 409.44 19.258 < 2e-16 \*\*\*

# Doors 705.56 98.79 7.142 1.46e-12 \*\*\*

# Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Residual standard error: 3565 on 1434 degrees of freedom

#Multiple R-squared: 0.03435, Adjusted R-squared: 0.03367

#F-statistic: 51 on 1 and 1434 DF, p-value: 1.461e-12

## Doors variable is also significant.

##Building model using both doors and cc

carModel <- lm(Price~cc+Doors,data = toyotaDetails)

summary(carModel)

#Call:

# lm(formula = Price ~ cc + Doors, data = toyotaDetails)

#Residuals:

# Min 1Q Median 3Q Max

#-7243.9 -2273.6 -821.3 1054.4 20714.1

#Coefficients:

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

#(Intercept) 6509.4211 515.7732 12.621 < 2e-16 \*\*\*

# cc 0.9597 0.2211 4.340 1.52e-05 \*\*\*

# Doors 671.3973 98.5009 6.816 1.37e-11 \*\*\*

#Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Residual standard error: 3543 on 1433 degrees of freedom

#Multiple R-squared: 0.04688, Adjusted R-squared: 0.04555

#F-statistic: 35.24 on 2 and 1433 DF, p-value: 1.15e-15

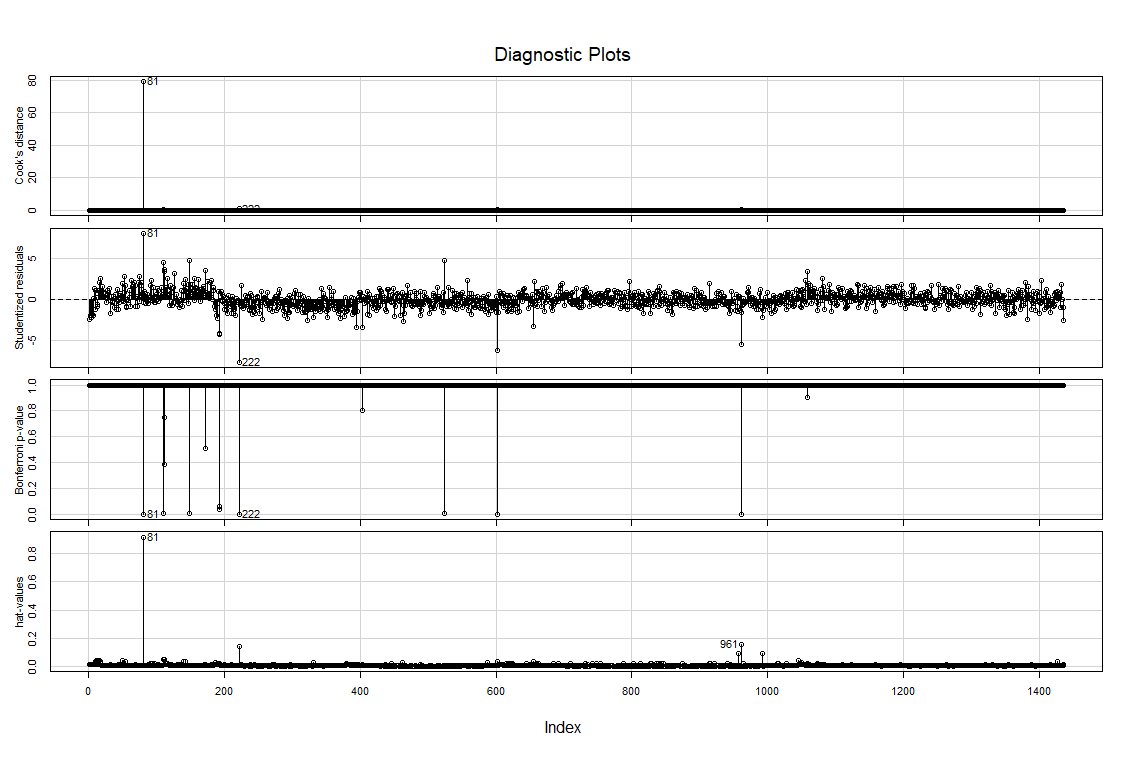
##Both the variables are significant when used together in building model.

##Finding out the influence record.

# Deletion Diagnostics for identifying influential variable

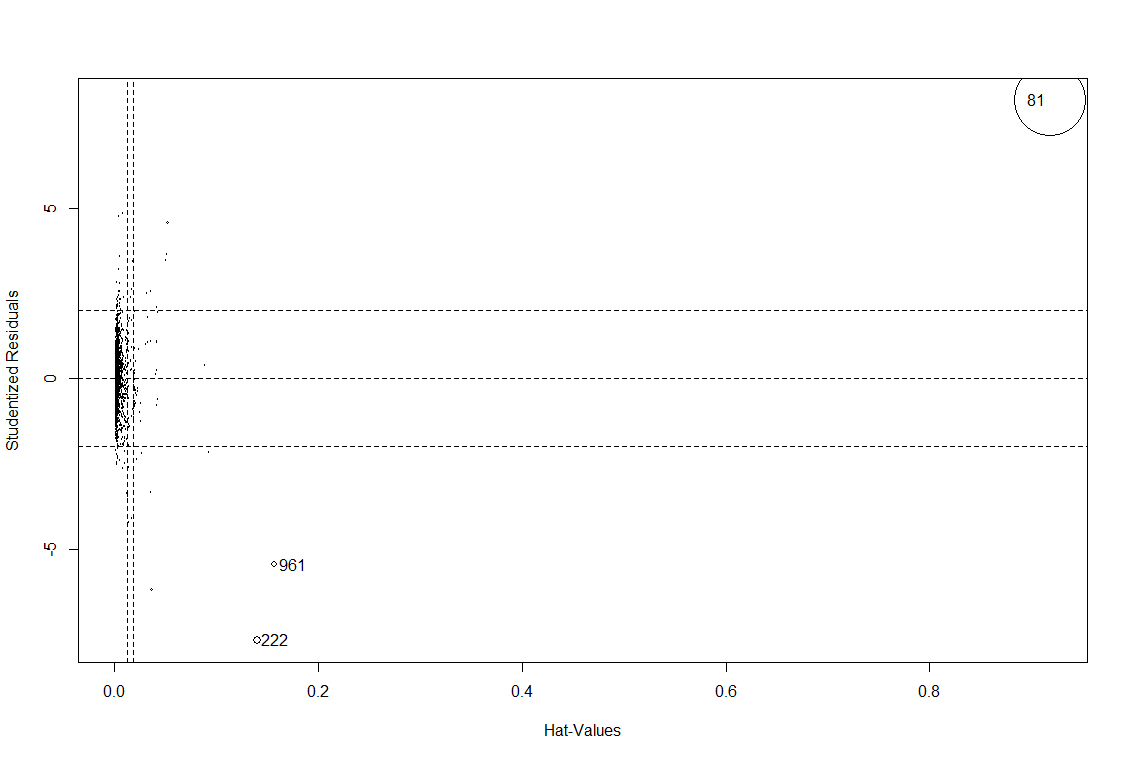
influence.measures(model)

influenceIndexPlot(model) # Index Plots of the influence measures



influencePlot(model)# A user friendly representation of the above

?influencePlot



#Stud Res Hat CookD

#81 8.164500 0.9182368 79.5201062

#222 -7.673262 0.1397116 1.0210312

#961 -5.456195 0.1572484 0.6049996

#Deleting influential record 81 and building the model

model.car1 <- lm(Price~.,data= toyotaDetails[-81,])

summary(model.car1)

#Call:

#lm(formula = Price ~ ., data = toyotaDetails[-81, ])

#Residuals:

# Min 1Q Median 3Q Max

#-11455.7 -761.7 -32.7 739.3 6739.7

#Coefficients:

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

#(Intercept) -6.285e+03 1.383e+03 -4.545 5.95e-06 \*\*\*

# Age\_08\_04 -1.205e+02 2.562e+00 -47.021 < 2e-16 \*\*\*

# KM -1.785e-02 1.277e-03 -13.973 < 2e-16 \*\*\*

# HP 3.935e+01 2.911e+00 13.516 < 2e-16 \*\*\*

# cc -2.524e+00 3.072e-01 -8.216 4.67e-16 \*\*\*

# Doors -2.723e+01 3.924e+01 -0.694 0.48788

#Gears 5.239e+02 1.929e+02 2.717 0.00667 \*\*

# Quarterly\_Tax 9.044e+00 1.425e+00 6.348 2.93e-10 \*\*\*

# Weight 2.017e+01 1.116e+00 18.076 < 2e-16 \*\*\*

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

#Residual standard error: 1313 on 1426 degrees of freedom

#Multiple R-squared: 0.8694, Adjusted R-squared: 0.8686

#F-statistic: 1186 on 8 and 1426 DF, p-value: < 2.2e-16

### Variance Inflation Factors

vif(model.car1)

# VIF is > 10 => collinearity

#Age\_08\_04 KM HP cc Doors

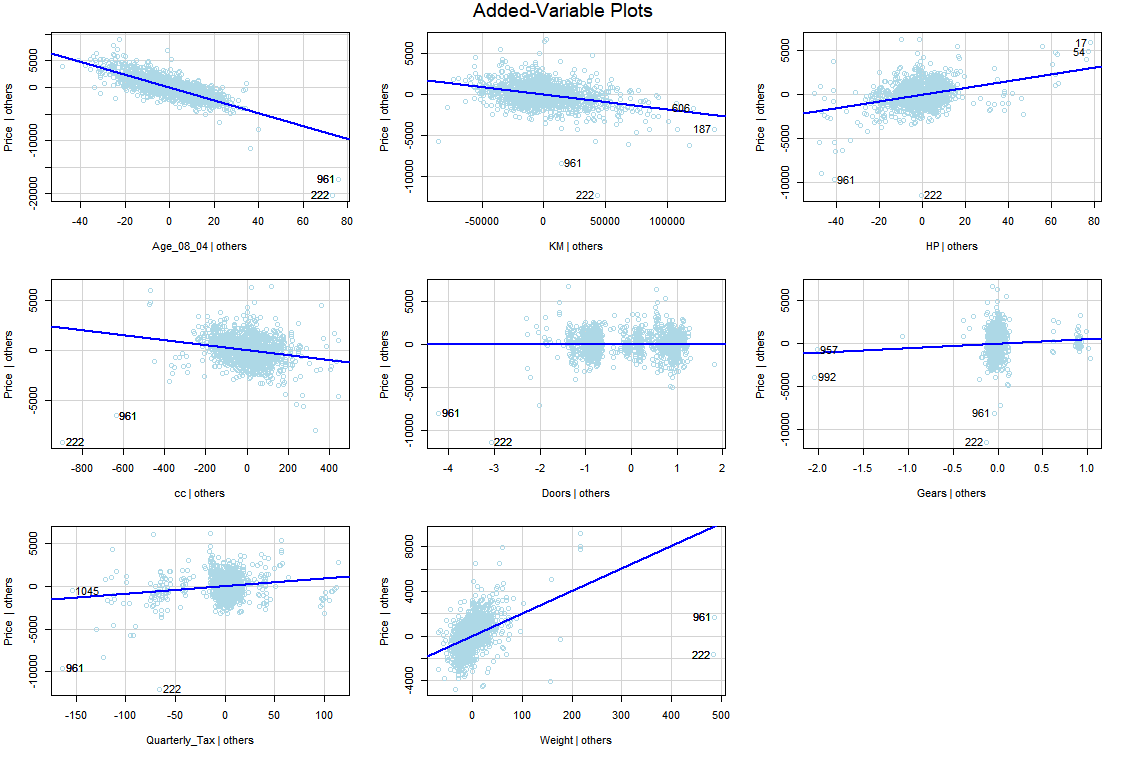
#1.887229 1.909570 1.583835 2.754405 1.163178

#Gears Quarterly\_Tax Weight

#1.100907 2.859861 2.864117

##Added variable plots

avPlots(model.car1, id.n=5, id.cex=100, col="lightblue")



install.packages("MASS")

library("MASS")

stepAIC(model.car1) # backward

#Start: AIC=20614.93

#Price ~ Age\_08\_04 + KM + HP + cc + Doors + Gears + Quarterly\_Tax +

# Weight

# Df Sum of Sq RSS AIC

#- Doors 1 829529 2457762537 20613

#<none> 2456933007 20615

#- Gears 1 12715451 2469648459 20620

#- Quarterly\_Tax 1 69419813 2526352820 20653

#- cc 1 116301260 2573234268 20679

#- HP 1 314730220 2771663227 20786

#- KM 1 336376338 2793309345 20797

#- Weight 1 562939527 3019872535 20909

#- Age\_08\_04 1 3809460567 6266393575 21957

#Step: AIC=20613.41

#Price ~ Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax + Weight

#Df Sum of Sq RSS AIC

#<none> 2457762537 20613

#- Gears 1 14537937 2472300474 20620

#- Quarterly\_Tax 1 69984492 2527747029 20652

#- cc 1 115472997 2573235533 20677

#- HP 1 314440997 2772203534 20784

#- KM 1 338971056 2796733593 20797

#- Weight 1 592472069 3050234606 20921

#- Age\_08\_04 1 3809636409 6267398946 21955

#Call:

# lm(formula = Price ~ Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax +

# Weight, data = toyotaDetails[-81, ])

#Coefficients:

# (Intercept) Age\_08\_04 KM HP cc

#-6.314e+03 -1.205e+02 -1.789e-02 3.916e+01 -2.507e+00

#Gears Quarterly\_Tax Weight

#5.497e+02 9.076e+00 1.996e+01

# Lower the AIC (Akaike Information Criterion) value better is the model. AIC is used only #if you build multiple models.

##Final Model after removing doors variable

model.final <- lm(Price ~ Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax +Weight , data=toyotaDetails)

summary(model.final)

#Call:

# lm(formula = Price ~ Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax +

# Weight, data = toyotaDetails)

#Residuals:

# Min 1Q Median 3Q Max

#-9362.3 -792.5 -21.3 801.2 6446.4

#Coefficients:

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

#(Intercept) -5.575e+03 1.410e+03 -3.954 8.06e-05 \*\*\*

# Age\_08\_04 -1.217e+02 2.615e+00 -46.528 < 2e-16 \*\*\*

# KM -2.082e-02 1.251e-03 -16.636 < 2e-16 \*\*\*

# HP 3.167e+01 2.810e+00 11.270 < 2e-16 \*\*\*

# cc -1.210e-01 9.005e-02 -1.344 0.17909

#Gears 5.958e+02 1.934e+02 3.081 0.00210 \*\*

# Quarterly\_Tax 3.953e+00 1.306e+00 3.027 0.00251 \*\*

# Weight 1.695e+01 1.033e+00 16.401 < 2e-16 \*\*\*

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

#Residual standard error: 1342 on 1428 degrees of freedom

#**Multiple R-squared: 0.8638, Adjusted R-squared: 0.8631**

#F-statistic: 1293 on 7 and 1428 DF, p-value: < 2.2e-16

##Building final model by removing doors variable and 81 observation

model.final1 <- lm(Price ~ Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax +Weight , data=toyotaDetails[-81,])

summary(model.final1)

#Call:

# lm(formula = Price ~ Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax +

# Weight, data = toyotaDetails[-81, ])

#Residuals:

# Min 1Q Median 3Q Max

#-11372.3 -759.9 -26.4 743.5 6777.4

#Coefficients:

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

#(Intercept) -6.314e+03 1.382e+03 -4.569 5.32e-06 \*\*\*

# Age\_08\_04 -1.205e+02 2.561e+00 -47.031 < 2e-16 \*\*\*

# KM -1.789e-02 1.275e-03 -14.029 < 2e-16 \*\*\*

# HP 3.916e+01 2.898e+00 13.512 < 2e-16 \*\*\*

# cc -2.507e+00 3.062e-01 -8.188 5.83e-16 \*\*\*

# Gears 5.497e+02 1.892e+02 2.905 0.00373 \*\*

# Quarterly\_Tax 9.076e+00 1.424e+00 6.374 2.47e-10 \*\*\*

# Weight 1.996e+01 1.076e+00 18.547 < 2e-16 \*\*\*

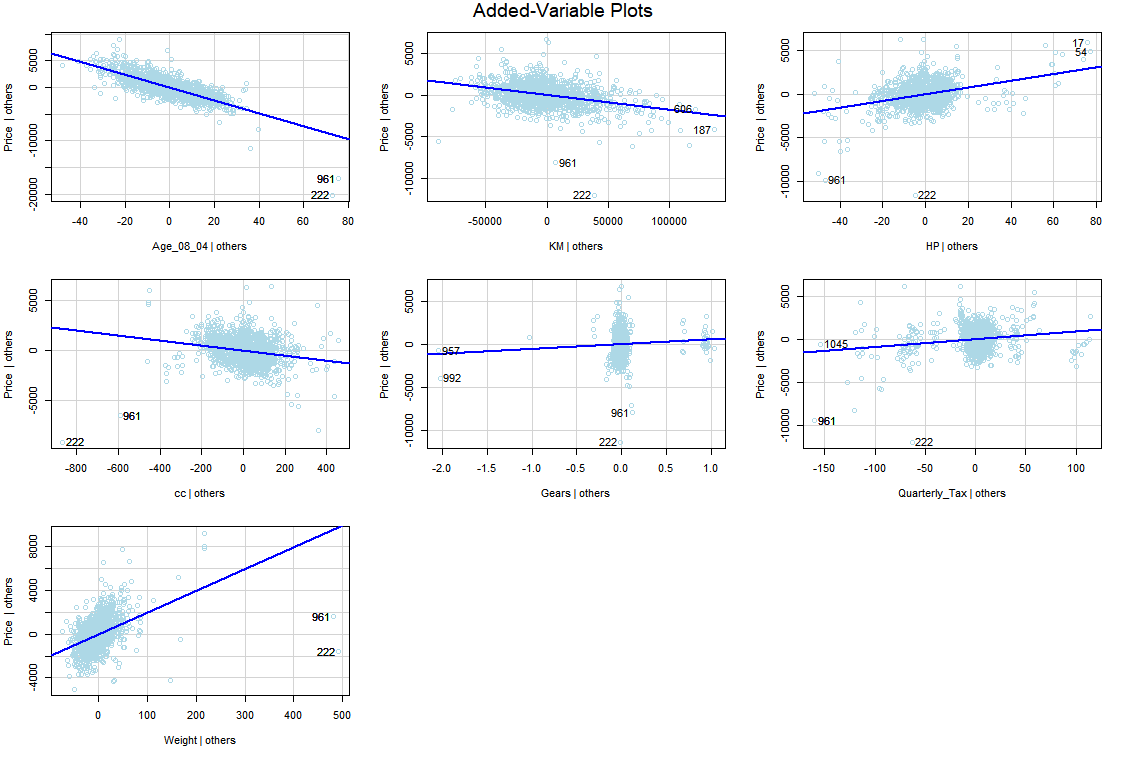
#Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Residual standard error: 1312 on 1427 degrees of freedom

#**Multiple R-squared: 0.8693, Adjusted R-squared: 0.8687**

#F-statistic: 1356 on 7 and 1427 DF, p-value: < 2.2e-16

avPlots(model.final1, id.n=2, id.cex=0.8, col="lightblue")



vif(model.final1)

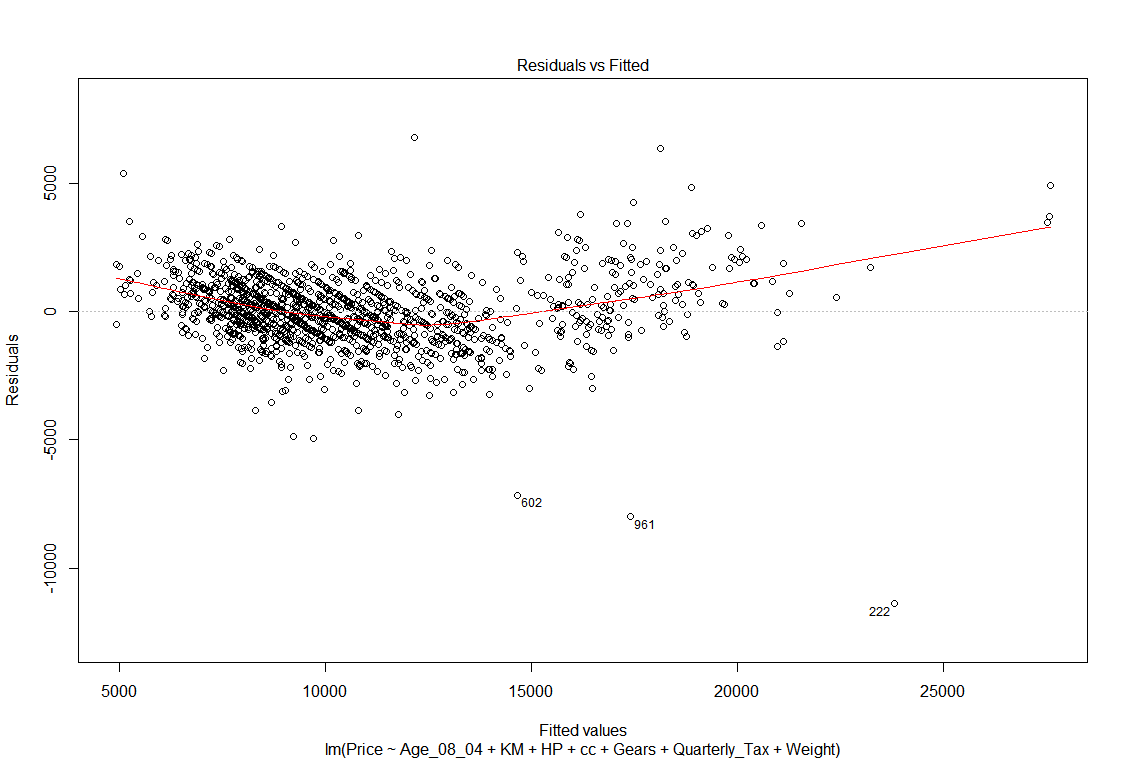
#Age\_08\_04 KM HP cc Gears

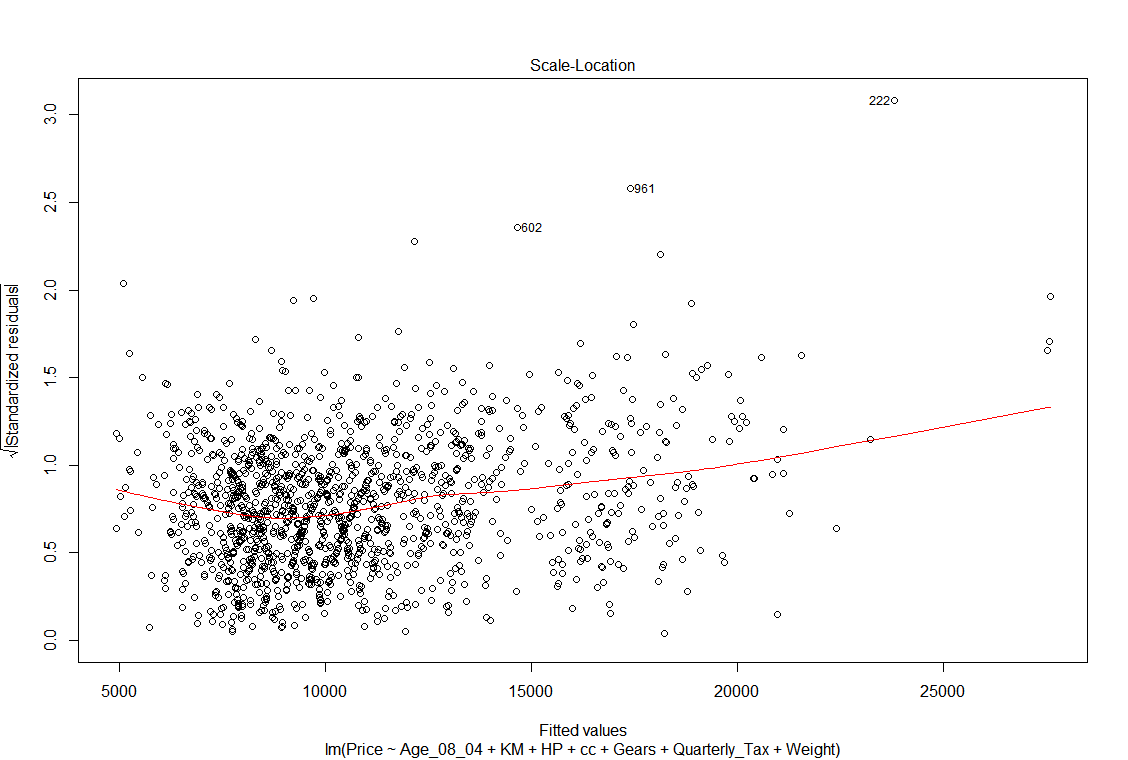
#1.887225 1.904583 1.570256 2.736259 1.060021

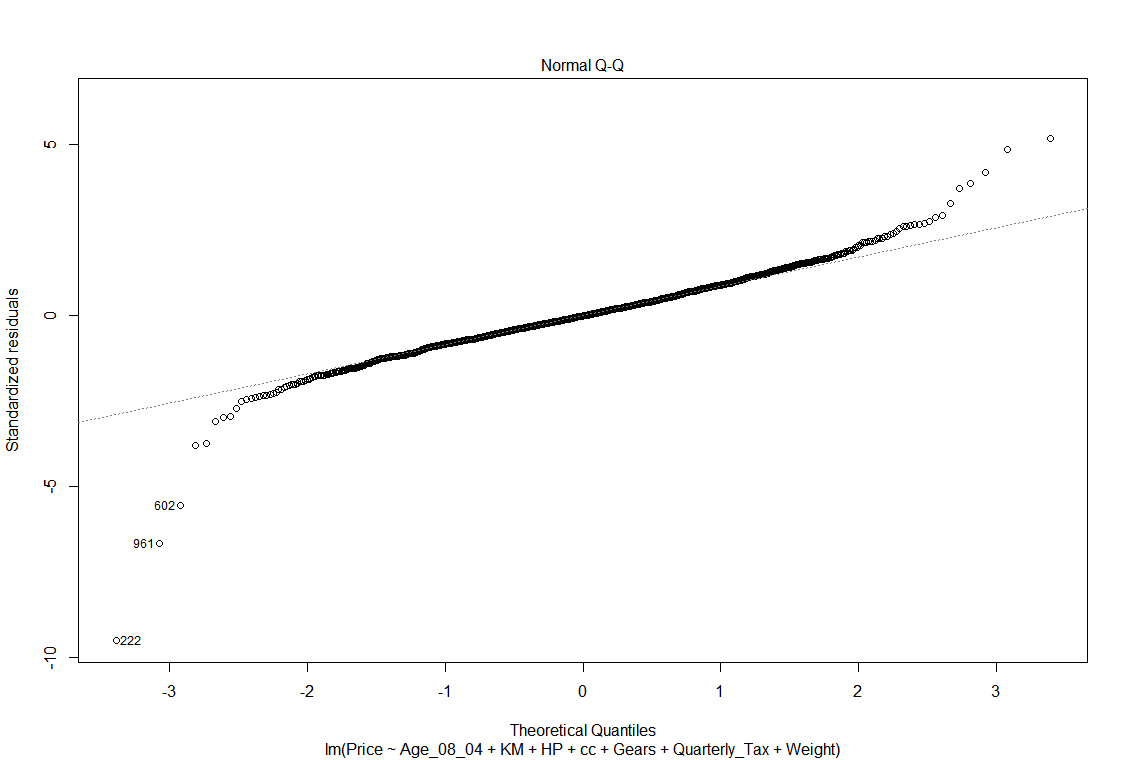
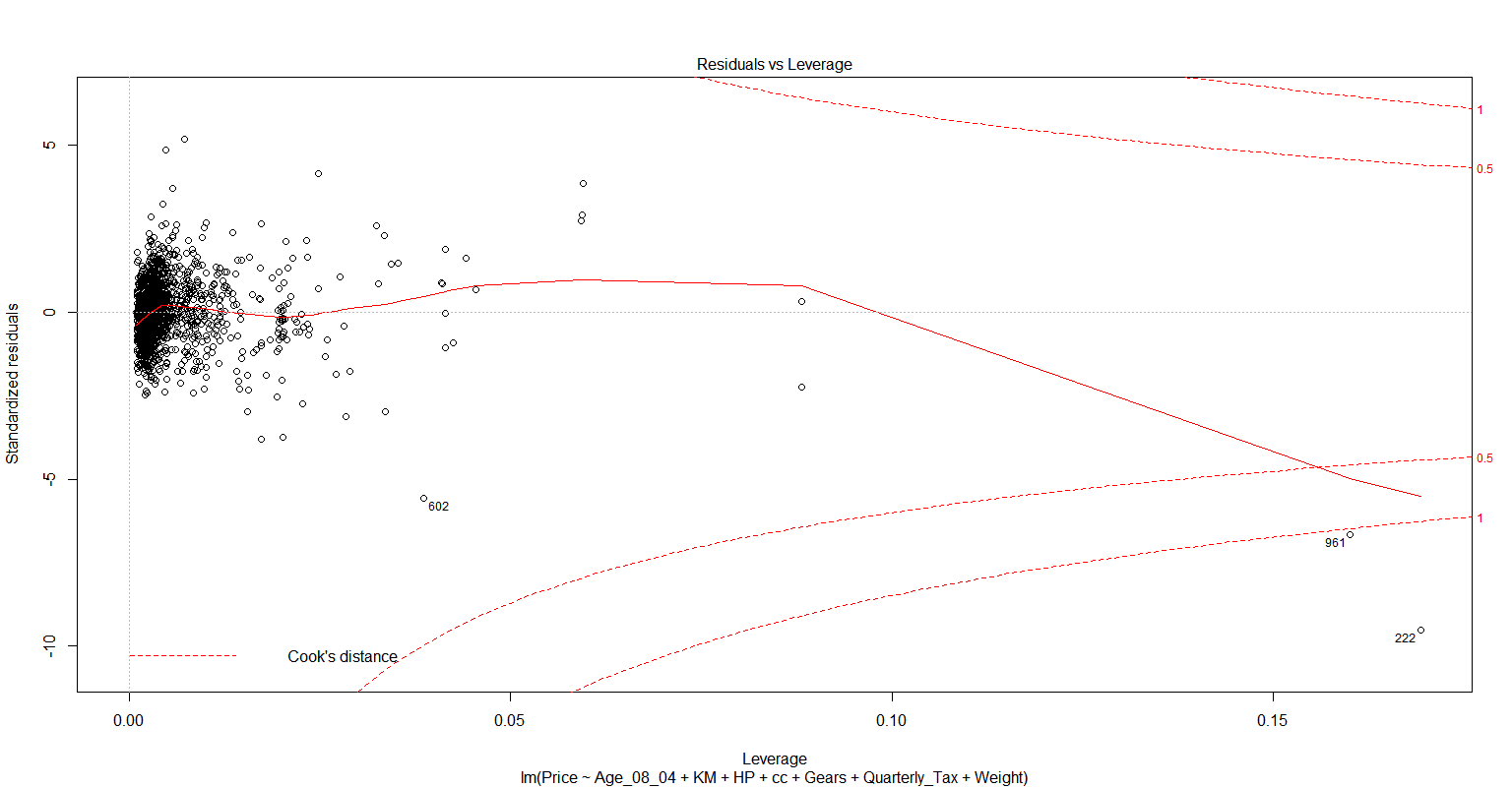
#Quarterly\_Tax Weight

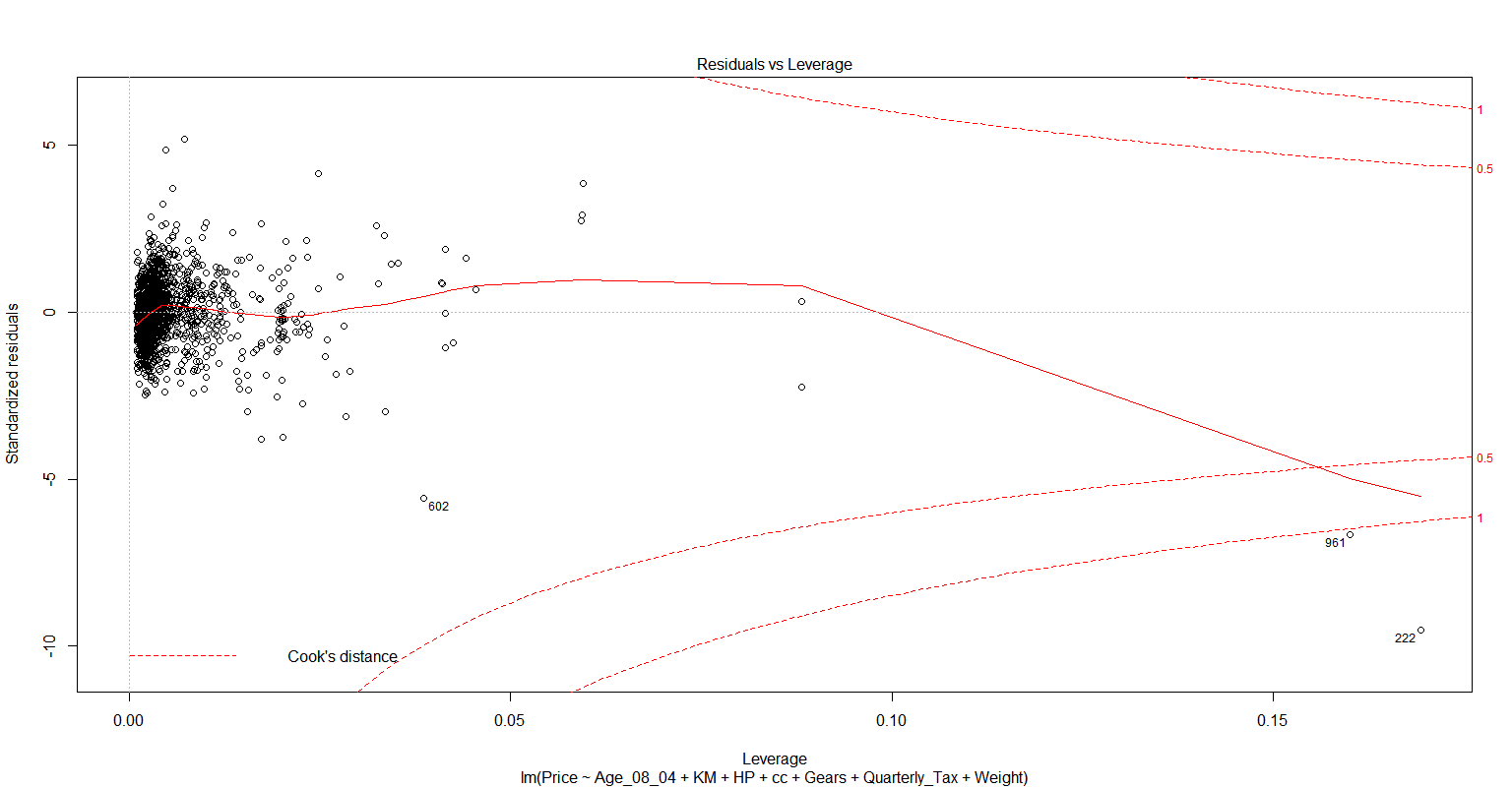
#2.856872 2.666788

plot(model.final1)









**##Hence the final1 model prices 86.9% of the variation in price using** **explanatory variables**

**#Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax +Weight**

**#Age is most significant with t-value of -47.031, followed by weight with t-value 18.54**

**#The least significant variable is number of doors.**