Que. 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")]

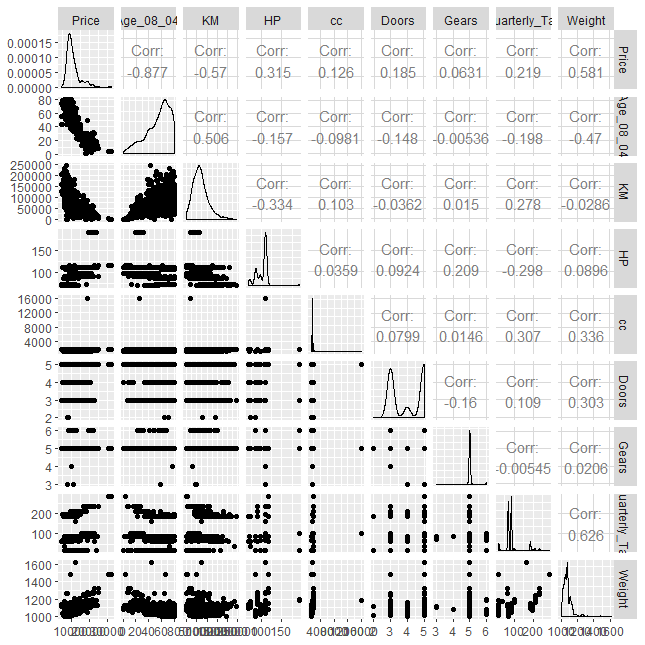
Ans:

> ToyotaCorolla <- read\_csv("ToyotaCorolla.csv")

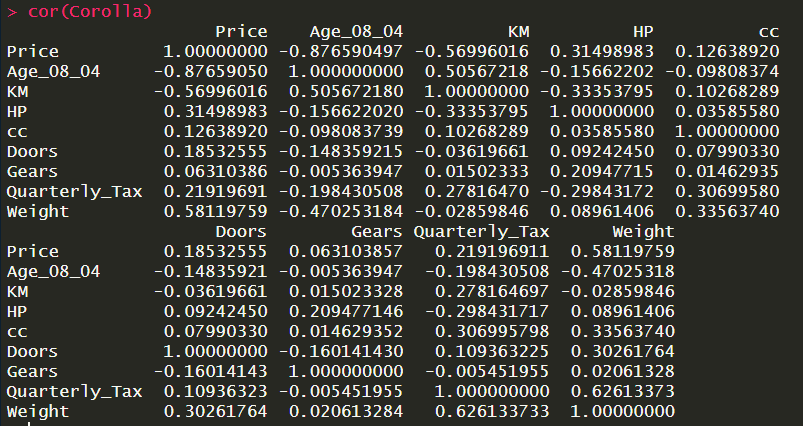
> Corolla <- ToyotaCorolla[c("Price", "Age\_08\_04","KM","HP","cc","Doors", "Gears", "Quarterly\_Tax", "Weight")]

>windows()

>ggpairs(Corolla)



> cor(Corolla)



From Above Correlation Function there is no strong correlation between input variables.

SO now we will begin with Building our model.

**1. Model 1**

> attach(Corolla)

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

> summary(model.Corolla)

Call:

lm(formula = Price ~ ., data = Corolla)

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

**From above summary of model, all P values are significant except Doors and cc variables.**

> vif(model.Corolla)

Age\_08\_04 KM HP cc Doors

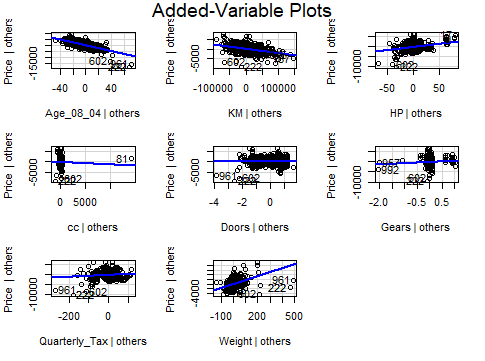
1.884620 1.756905 1.419422 1.163894 1.156575

Gears Quarterly\_Tax Weight

1.098723 2.311431 2.516420

**As VIF value is less than 10 it isn’t involved in collinearity.**

> avPlots(model.Corolla)



**From above plot there seems no relation between Price and Doors variable.**

**So we will drop the Doors variable from the input and build new model.**

**2. Model 2**

> model.Corolla1 <- lm(Price ~ Age\_08\_04+KM + HP + cc + Gears + Quarterly\_Tax+ Weight)

> summary(model.Corolla1)

Call:

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

Weight)

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

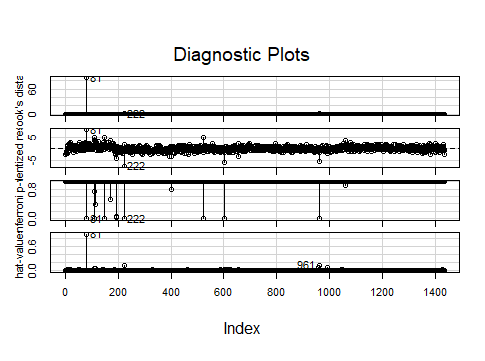
Residual standard error: 1342 on 1428 degrees of freedom

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

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

**All input variable show significant p value except cc, so we will check the variable with influence Index plot function.**

> influenceIndexPlot(model.Corolla1)



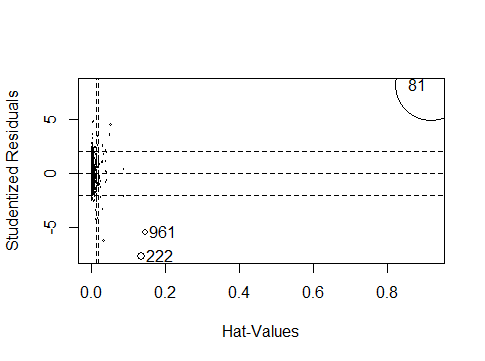
> influencePlot(model.Corolla1)

StudRes Hat CookD

81 8.136545 0.9177109 88.2598571

222 -7.645571 0.1338990 1.0859440

961 -5.409408 0.1439311 0.6030376



**From above plot observation no. 81 961 and 222 are highly influencing the model output. So we will drop out these observations and build the new Model.**

**3. Model 3**

> model.Corolla2 <- lm(Price ~ Age\_08\_04+KM + HP + cc + Gears + Quarterly\_Tax+ Weight, data = Corolla[-c(81,961,222),])

> summary(model.Corolla2)

Call:

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

Weight, data = Corolla[-c(81, 961, 222), ])

Residuals:

Min 1Q Median 3Q Max

-8448.5 -741.7 -23.4 727.2 6347.5

Coefficients:

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

(Intercept) -1.450e+04 1.436e+03 -10.098 < 2e-16 \*\*\*

Age\_08\_04 -1.124e+02 2.485e+00 -45.237 < 2e-16 \*\*\*

KM -1.722e-02 1.202e-03 -14.328 < 2e-16 \*\*\*

HP 3.594e+01 2.747e+00 13.084 < 2e-16 \*\*\*

cc -3.667e+00 3.006e-01 -12.198 < 2e-16 \*\*\*

Gears 5.793e+02 1.782e+02 3.251 0.00118 \*\*

Quarterly\_Tax 5.537e+00 1.372e+00 4.034 5.77e-05 \*\*\*

Weight 2.930e+01 1.227e+00 23.872 < 2e-16 \*\*\*

---

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

Residual standard error: 1236 on 1425 degrees of freedom

Multiple R-squared: 0.8843, Adjusted R-squared: 0.8838

F-statistic: 1556 on 7 and 1425 DF, p-value: < 2.2e-16

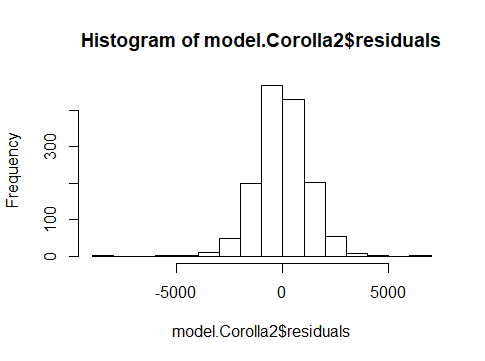
**So from above summary p values for all predictors are significant as well as R sqr values are also high. model.Corolla2 will be final model.**

**Model Evaluation:**

> mean(model.Corolla2$residuals)

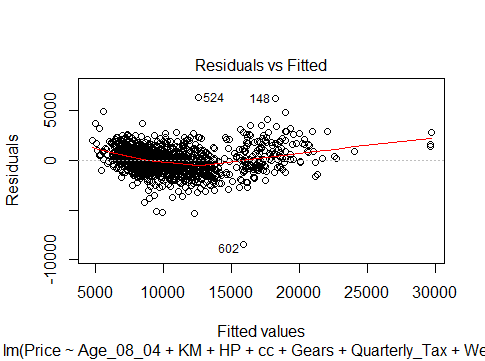
[1] 1.024411e-14

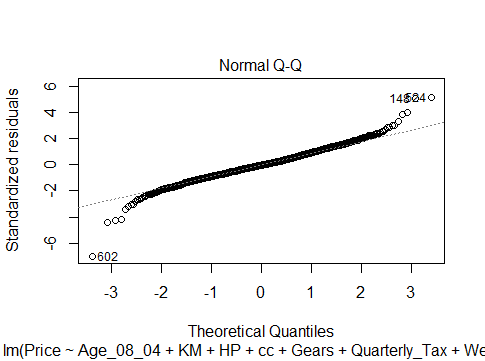
> hist(model.Corolla2$residuals)

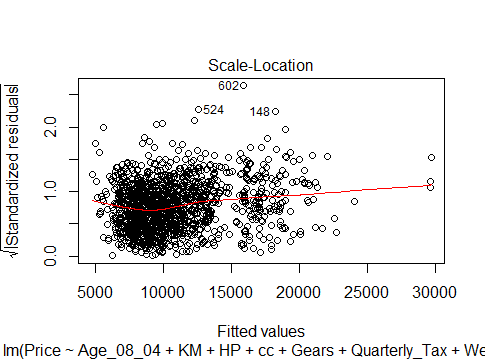


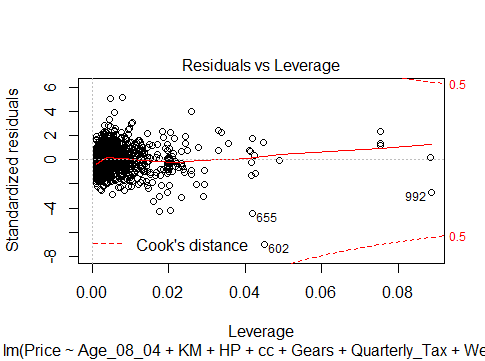
**Histogram of above model shows good normal distribution.**

> plot(model.Corolla2)









**So all above plots represent good characteristics of final model.**

**Cross Validation:**

> split <- sample.split(Corolla1$Price, SplitRatio = 0.70)

> table(split)

split

FALSE TRUE

411 1022

> corolla.train <- subset(Corolla1, split== T)

> corolla.test <- subset(Corolla1, split== F)

> model.train <- lm(Price ~ ., data = corolla.train)

> summary(model.train)

Call:

lm(formula = Price ~ ., data = corolla.train)

Residuals:

Min 1Q Median 3Q Max

-5171.6 -771.7 -47.1 731.1 6327.6

Coefficients:

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

(Intercept) -1.681e+04 1.661e+03 -10.119 < 2e-16 \*\*\*

Age\_08\_04 -1.074e+02 2.899e+00 -37.060 < 2e-16 \*\*\*

KM -1.816e-02 1.358e-03 -13.371 < 2e-16 \*\*\*

HP 3.559e+01 3.101e+00 11.479 < 2e-16 \*\*\*

cc -3.534e+00 3.434e-01 -10.291 < 2e-16 \*\*\*

Doors -2.039e+02 4.424e+01 -4.610 4.53e-06 \*\*\*

Gears 2.652e+02 2.152e+02 1.233 0.218

Quarterly\_Tax 2.940e+00 1.568e+00 1.876 0.061 .

Weight 3.356e+01 1.502e+00 22.350 < 2e-16 \*\*\*

---

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

Residual standard error: 1218 on 1013 degrees of freedom

Multiple R-squared: 0.8931, Adjusted R-squared: 0.8922

F-statistic: 1058 on 8 and 1013 DF, p-value: < 2.2e-16

> model.test <- lm(Price ~ ., data = corolla.test)

> summary(model.test)

Call:

lm(formula = Price ~ ., data = corolla.test)

Residuals:

Min 1Q Median 3Q Max

-6744.6 -763.2 -1.3 686.9 6909.5

Coefficients:

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

(Intercept) -9.293e+03 2.786e+03 -3.335 0.000932 \*\*\*

Age\_08\_04 -1.229e+02 4.680e+00 -26.253 < 2e-16 \*\*\*

KM -1.407e-02 2.489e-03 -5.651 3.02e-08 \*\*\*

HP 3.888e+01 5.862e+00 6.632 1.07e-10 \*\*\*

cc -4.396e+00 6.305e-01 -6.972 1.29e-11 \*\*\*

Gears 7.869e+02 3.273e+02 2.404 0.016651 \*

Quarterly\_Tax 9.785e+00 2.816e+00 3.474 0.000568 \*\*\*

Weight 2.403e+01 2.498e+00 9.622 < 2e-16 \*\*\*

---

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

Residual standard error: 1229 on 402 degrees of freedom

Multiple R-squared: 0.8719, Adjusted R-squared: 0.8694

F-statistic: 342 on 8 and 402 DF, p-value: < 2.2e-16