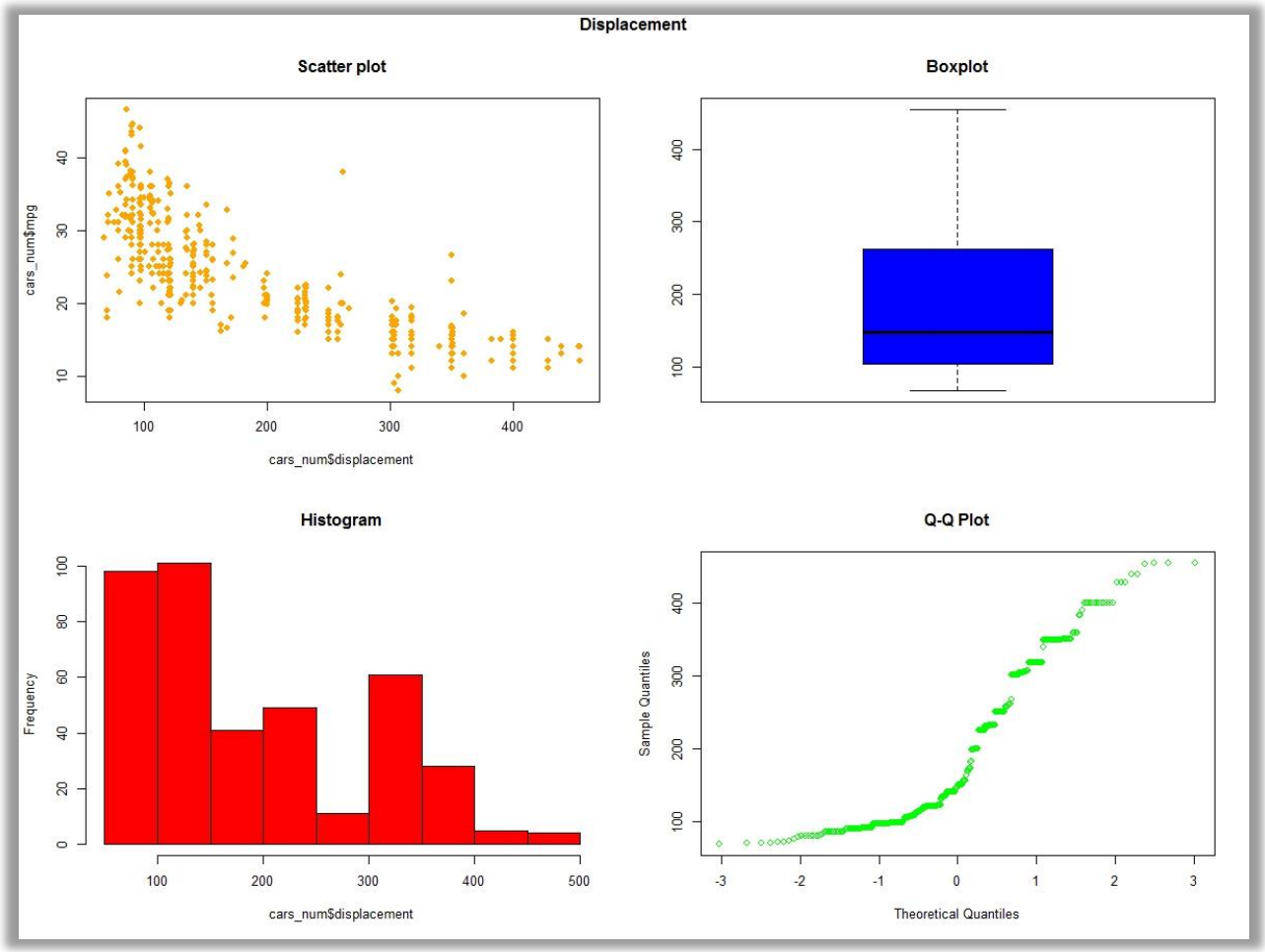
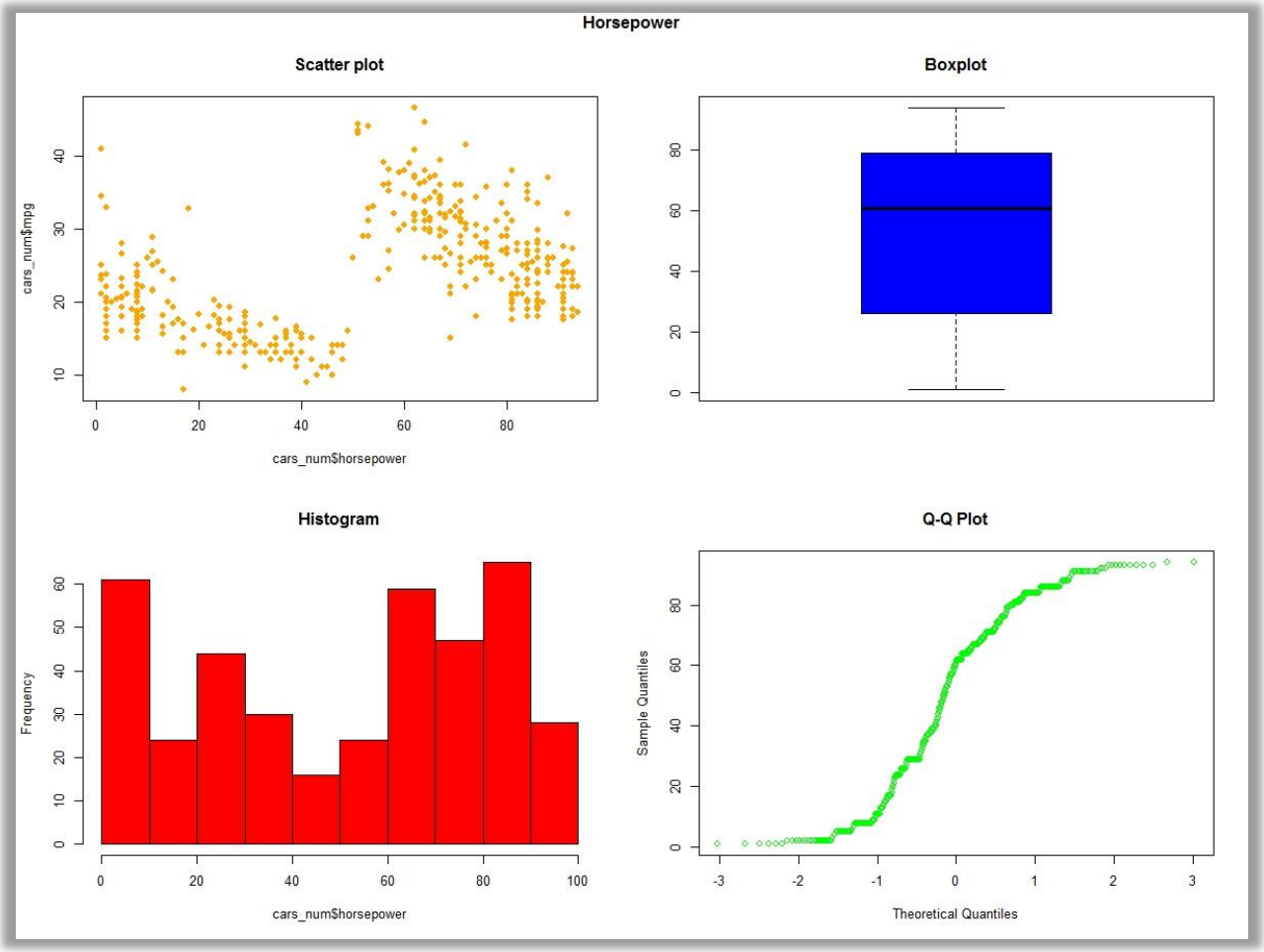
```
> ##Reading dataset
> cars<-read.csv("Car Mileage Dataset.csv")
>
> ##Normalizing the variable names
> ##install.packages("rattle")
> library(rattle)
> names(cars)
 [1] "MPG"          "Cylinders"    "Displacement" "Horsepower"   "Weight"
"Acceleration"
 [7] "Model_year"   "Year_03_06"   "Year_07_11"   "Year_12_15"   "Origin"
"Car_Name"
> names(cars)<-normVarNames(names(cars))
>
> ##Understanding the datastructure for data preparation
> str(cars)
'data.frame':  398 obs. of  12 variables:
 $ mpg         : num  8 15 18 16 17 15 14 14 14 15 ...
 $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
 $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
 $ horsepower  : Factor w/ 94 levels "?","100","102",..: 17 35 29 29 24 42 47
46 48 40 ...
 $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ model_year  : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
 $ year_03_06  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ year_07_11  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ year_12_15  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ car_name    : Factor w/ 305 levels "amc ambassador brougham",..: 50 37 232
15 162 142 55 224 242 2 ...
```

```
> ##Splitting dataset for further analysis
> cars_num<- subset(cars,select = c(mpg,
+                                   displacement,
+                                   horsepower,
+                                   acceleration,
+                                   weight))
>
> cars_date<-subset(cars, select=c(year_03_06, year_07_11, year_12_15))
```

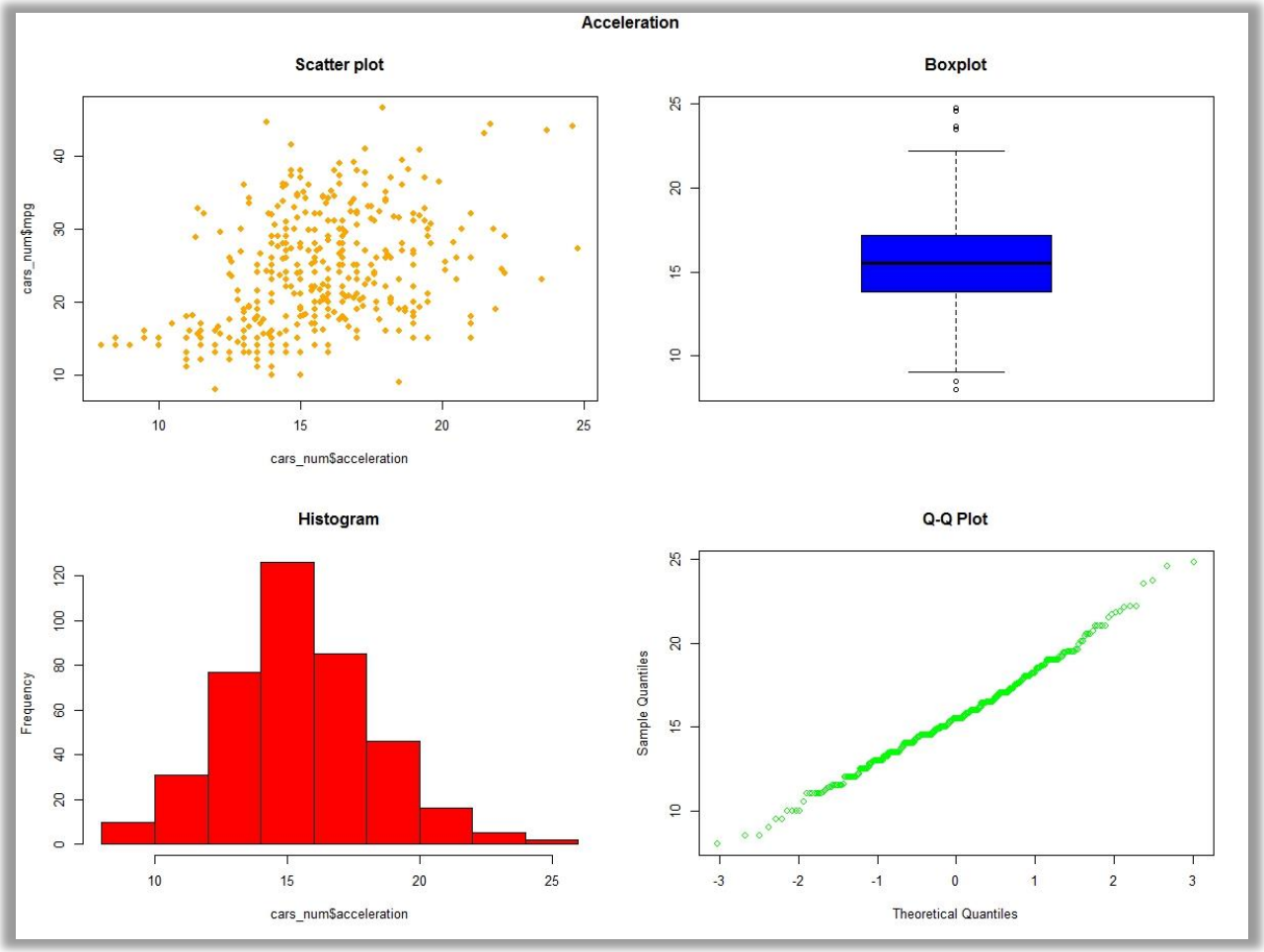
```
> ##Data visualization
> ##Displacement
> par(mfrow=c(2,2), oma=c(0,0,1,0))
> plot(cars_num$displacement,cars_num$mpg, pch=19,main="Scatter plot", col="o
range")
> boxplot(cars_num$displacement, main="Boxplot", col="blue")
> hist(cars_num$displacement, main="Histogram", col="red")
> qqnorm(cars_num$displacement, main="Q-Q Plot", col="green")
> title("Displacement", outer=TRUE)
```



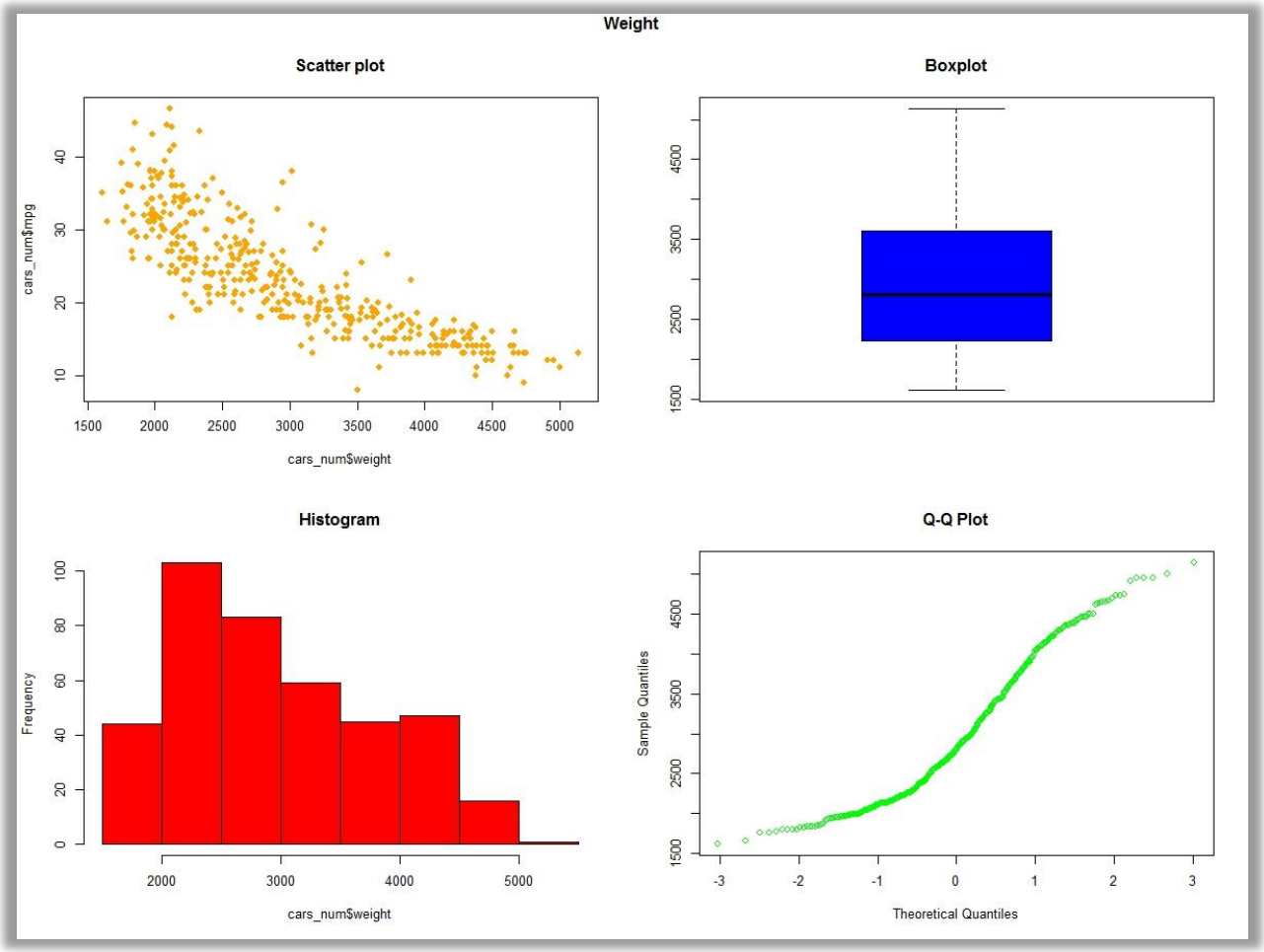
```
> ##Horsepower
> plot(cars_num$horsepower,cars_num$mpg, pch=19,main="Scatter plot", col="ora
nge")
> boxplot(cars_num$horsepower, main="Boxplot", col="blue")
> hist(cars_num$horsepower, main="Histogram", col="red")
> qqnorm(cars_num$horsepower, main="Q-Q Plot", col="green")
> title("Horsepower", outer=TRUE)
```



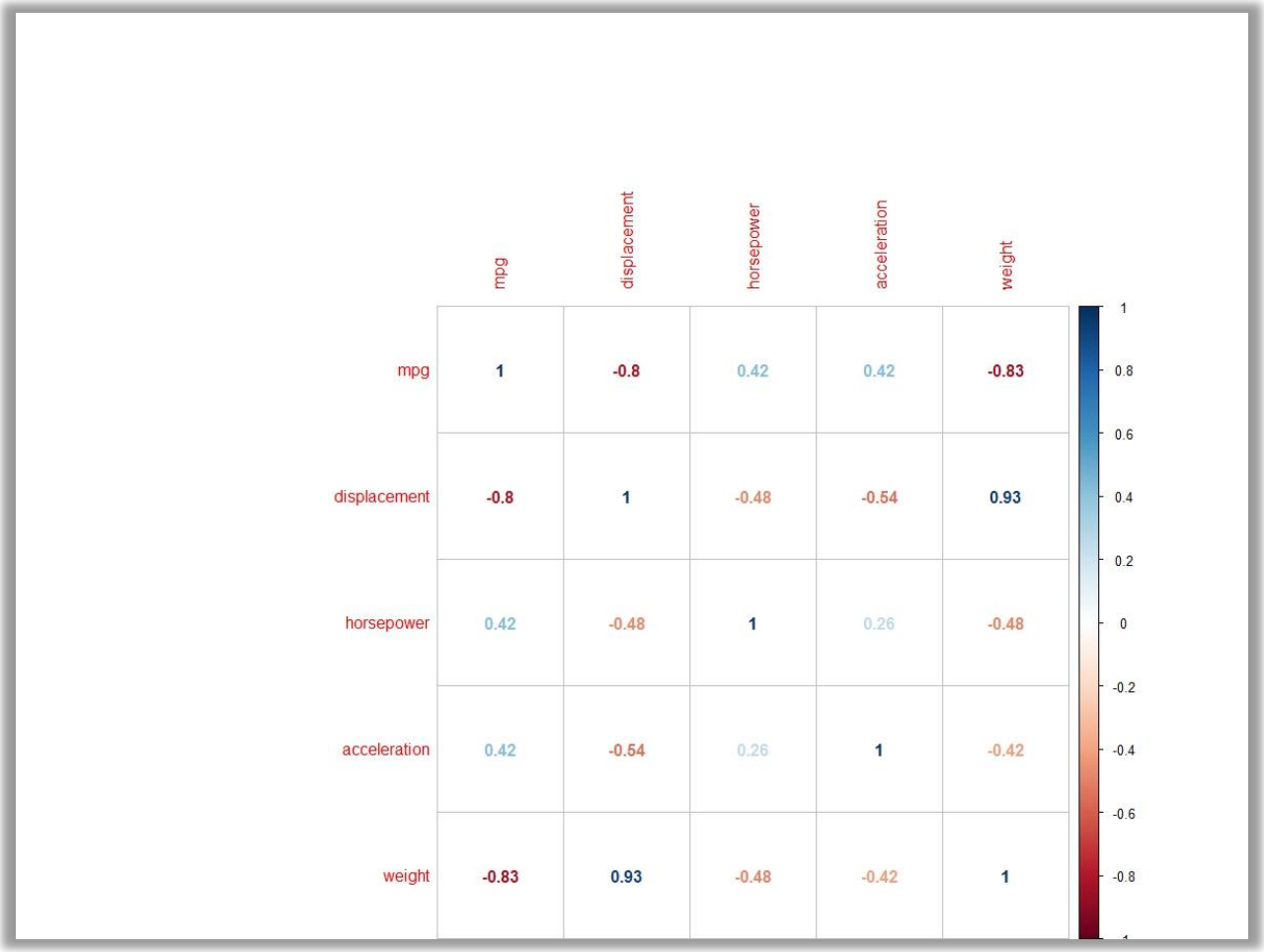
```
> plot(cars_num$acceleration,cars_num$mpg, pch=19,main="Scatter plot", col="o
range")
> boxplot(cars_num$acceleration, main="Boxplot", col="blue")
> hist(cars_num$acceleration, main="Histogram", col="red")
> qqnorm(cars_num$acceleration, main="Q-Q Plot", col="green")
> title("Acceleration", outer=TRUE)
```



```
> ##Weight
> plot(cars_num$weight,cars_num$mpg, pch=19,main="Scatter plot", col="orange"
)
> boxplot(cars_num$weight, main="Boxplot", col="blue")
> hist(cars_num$weight, main="Histogram", col="red")
> qqnorm(cars_num$weight, main="Q-Q Plot", col="green")
> title("Weight", outer=TRUE)
```



```
> ##Correlation plot
> ###install.packages("corrplot")
> library(corrplot)
> cor_cars<-cor(cars_num)
> corrplot(cor_cars, method="number")
```




```
> ##Create dummy variable
> ###install.packages("caret")
> library(caret)
> dummy_cyl<-(predict(dummyVars(mpg~cylinders, data=cars), newdata=cars))
> dummy_cyl<-dummy_cyl[,-1]
>
> dummy_org<-(predict(dummyVars(mpg~origin, data=cars), newdata=cars))
> dummy_org<-dummy_org[,-1]
>
> ##Arranging the required dataset in one dataframe
> data<-cbind(cars_num, dummy_org,dummy_cyl, cars_date )
> head(data)
  mpg displacement horsepower acceleration weight origin.2 origin.3 cylinders
.4 cylinders.5 cylinders.6
1   8          307         17         12.0   3504        0        0
0           0           0
```

```
2  15           350          35          11.5   3693         0          0
0            0            0
3  18           318          29          11.0   3436         0          0
0            0            0
4  16           304          29          12.0   3433         0          0
0            0            0
5  17           302          24          10.5   3449         0          0
0            0            0
6  15           429          42          10.0   4341         0          0
0            0            0
  cylinders.8 year_03_06 year_07_11 year_12_15
1           1          0          0          1
2           1          0          0          1
3           1          0          0          1
4           1          0          0          1
5           1          0          0          1
6           1          0          0          1
```

```
> ##Fitting regression model
> model<-lm(mpg~.,data=train)
> summary(model)

Call:
lm(formula = mpg ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2704 -1.7165  0.1033  1.5487 11.6399

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.1123357  2.6816916   9.737  < 2e-16 ***
displacement  0.0142230  0.0085933   1.655 0.099076 .
horsepower   -0.0038425  0.0078439  -0.490 0.624627
acceleration  0.2938442  0.0981134   2.995 0.003004 **
weight       -0.0064594  0.0006714  -9.620  < 2e-16 ***
origin.2      2.1541302  0.6622924   3.253 0.001292 **
origin.3      2.4503120  0.6507149   3.766 0.000205 ***
cylinders.4   6.4891373  2.0182131   3.215 0.001464 **
cylinders.5   5.8946738  2.8331962   2.081 0.038430 *
cylinders.6   3.6959694  2.2153652   1.668 0.096426 .
cylinders.8   5.4180474  2.6318099   2.059 0.040499 *
year_03_06    7.6164528  0.5574875  13.662  < 2e-16 ***
year_07_11    2.4499365  0.5062980   4.839 2.21e-06 ***
year_12_15           NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.226 on 266 degrees of freedom
Multiple R-squared:  0.8465,  Adjusted R-squared:  0.8396
F-statistic: 122.3 on 12 and 266 DF,  p-value: < 2.2e-16
```

```
> ##Step wise regression
> ##install.packages("MASS")
> library(MASS)
> step<-stepAIC(model, direction="both")
Start:  AIC=666.24
mpg ~ displacement + horsepower + acceleration + weight + origin.2 +
    origin.3 + cylinders.4 + cylinders.5 + cylinders.6 + cylinders.8 +
    year_03_06 + year_07_11 + year_12_15


Step:  AIC=666.24
mpg ~ displacement + horsepower + acceleration + weight + origin.2 +
    origin.3 + cylinders.4 + cylinders.5 + cylinders.6 + cylinders.8 +
    year_03_06 + year_07_11

               Df Sum of Sq    RSS    AIC
- horsepower    1      2.50 2770.8 664.49
<none>                     2768.3 666.24
- displacement  1     28.51 2796.8 667.10
- cylinders.6   1     28.97 2797.2 667.14
- cylinders.8   1     44.11 2812.4 668.65
- cylinders.5   1     45.05 2813.3 668.74
- acceleration  1     93.35 2861.6 673.49
- cylinders.4   1    107.59 2875.8 674.88
- origin.2      1    110.10 2878.4 675.12
- origin.3      1    147.57 2915.8 678.73
- year_07_11    1    243.68 3011.9 687.78
- weight        1    963.19 3731.4 747.54
- year_03_06    1   1942.50 4710.8 812.56

Step:  AIC=664.49
mpg ~ displacement + acceleration + weight + origin.2 + origin.3 +
    cylinders.4 + cylinders.5 + cylinders.6 + cylinders.8 + year_03_06 +
    year_07_11

               Df Sum of Sq    RSS    AIC
<none>                     2770.8 664.49
- displacement  1     27.64 2798.4 665.26
- cylinders.6   1     29.06 2799.8 665.40
+ horsepower    1      2.50 2768.3 666.24
- cylinders.5   1     44.12 2814.9 666.90
- cylinders.8   1     45.55 2816.3 667.04
- acceleration  1     92.44 2863.2 671.65
- cylinders.4   1    105.10 2875.9 672.88
- origin.2      1    113.74 2884.5 673.71
- origin.3      1    147.04 2917.8 676.92
- year_07_11    1    249.55 3020.3 686.55
- weight        1    964.73 3735.5 745.84
- year_03_06    1   1968.55 4739.3 812.25
> step

Call:
lm(formula = mpg ~ displacement + acceleration + weight + origin.2 +
    origin.3 + cylinders.4 + cylinders.5 + cylinders.6 + cylinders.8 +
    year_03_06 + year_07_11, data = train)

Coefficients:
```

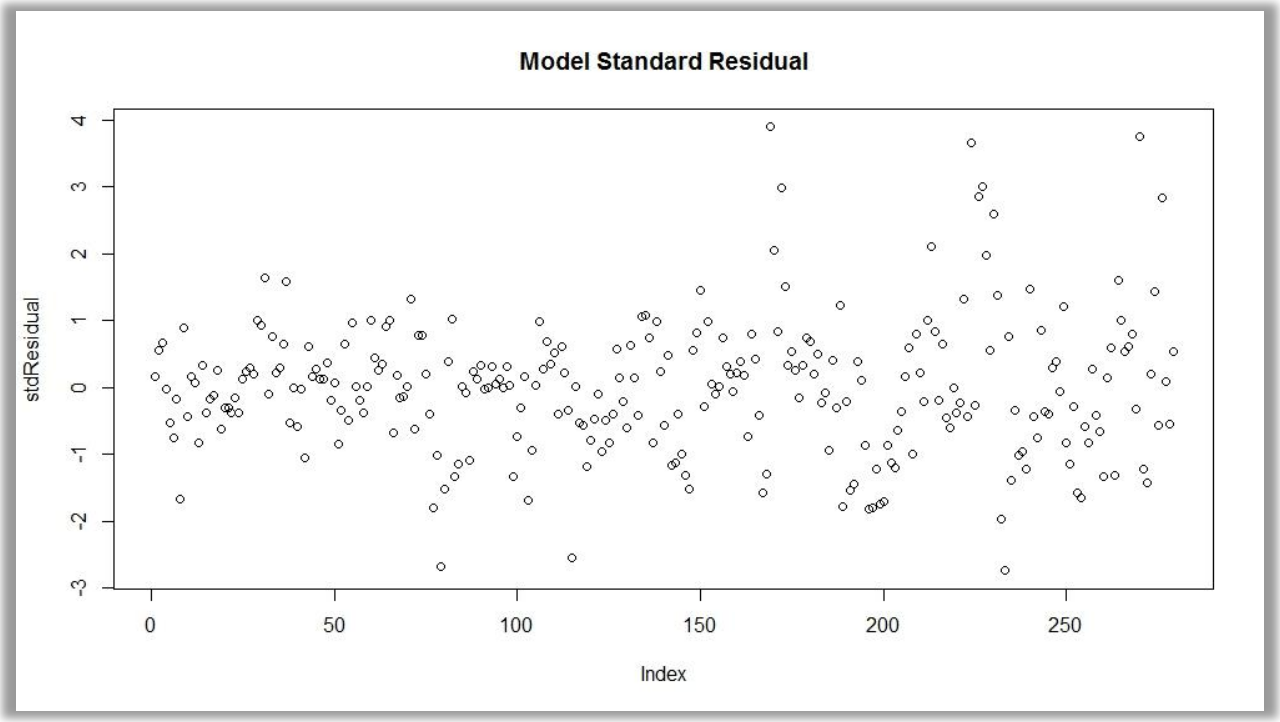
```
 (Intercept)  displacement  acceleration        weight      origin.2      ori
gin.3   cylinders.4
   26.029223      0.013981      0.292243     -0.006464      2.181622      2.4
45725      6.348694
 cylinders.5   cylinders.6   cylinders.8    year_03_06    year_07_11
    5.826458      3.701659      5.496124      7.639648      2.470612
```

```
> ##Final model after several trails
>
> model_F<-lm(formula = mpg ~ weight + origin.2 + origin.3 +
+               cylinders.6 + year_03_06 + year_07_11,
+             data = train)
> summary(model_F)

Call:
lm(formula = mpg ~ weight + origin.2 + origin.3 + cylinders.6 +
    year_03_06 + year_07_11, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0330 -1.8844  0.0389  1.7628 12.8498

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.3210712  1.1739776  31.790  < 2e-16 ***
weight      -0.0057793  0.0003039 -19.020  < 2e-16 ***
origin.2     1.9856199  0.6191600   3.207  0.00150 **
origin.3     1.7604034  0.6392194   2.754  0.00628 **
cylinders.6 -1.8956871  0.5093306  -3.722  0.00024 ***
year_03_06   7.6374228  0.5458576  13.992  < 2e-16 ***
year_07_11   2.4154432  0.4931275   4.898 1.66e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.344 on 272 degrees of freedom
Multiple R-squared:  0.8314,   Adjusted R-squared:  0.8277
F-statistic: 223.6 on 6 and 272 DF,  p-value: < 2.2e-16

> ##install.packages("car")
> library(car)
> vif(model_F)
     weight    origin.2    origin.3 cylinders.6  year_03_06  year_07_11
   1.717807    1.429124    1.591605    1.118948    1.575210    1.449295
```

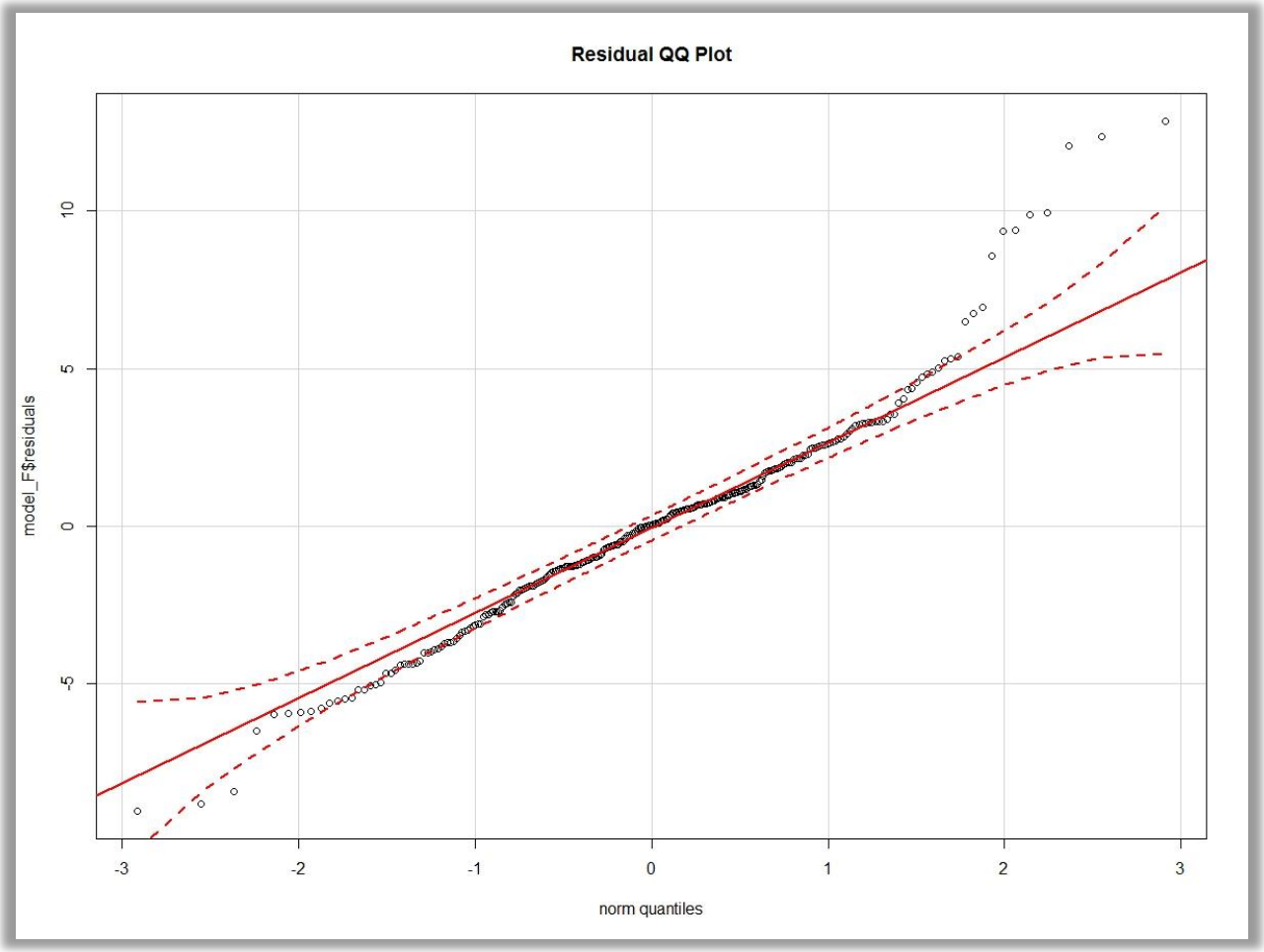
```
> ##Prediction
> predTest<-predict(model_F, test)
```

```
> ##Model validation
> ##MAPE Calculation
> MAPE<-function(actual,predicted) {
+   mean(abs(actual-predicted)/actual)
+ }
>
> ##Testing MAPE
> MAPE(test$mpg,predTest)
[1] 0.1065613
>
> #Correlation
> cor(test$mpg,predTest)
[1] 0.9336085
> cor(test$mpg,predTest)^2
[1] 0.8716248
```

```
> ##Standard residual
> stdResidual = rstandard(model_F)
>
> ##Standard residual plot
> plot(stdResidual, main="Model Standard Residual")
```



```
> ##Q-Q Plot
> qqPlot(model_F$residuals, main="Residual QQ Plot")
```



```
> ##Density on Histogram
> x <-model_F$residuals
> h<-hist(x, breaks=10, col="red", xlab="Residuals",
+         main="Residual Curve")
> xfit<-seq(min(x),max(x),length=40)
> yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
> yfit <- yfit*diff(h$mids[1:2])*length(x)
> lines(xfit, yfit, col="blue", lwd=2)
```