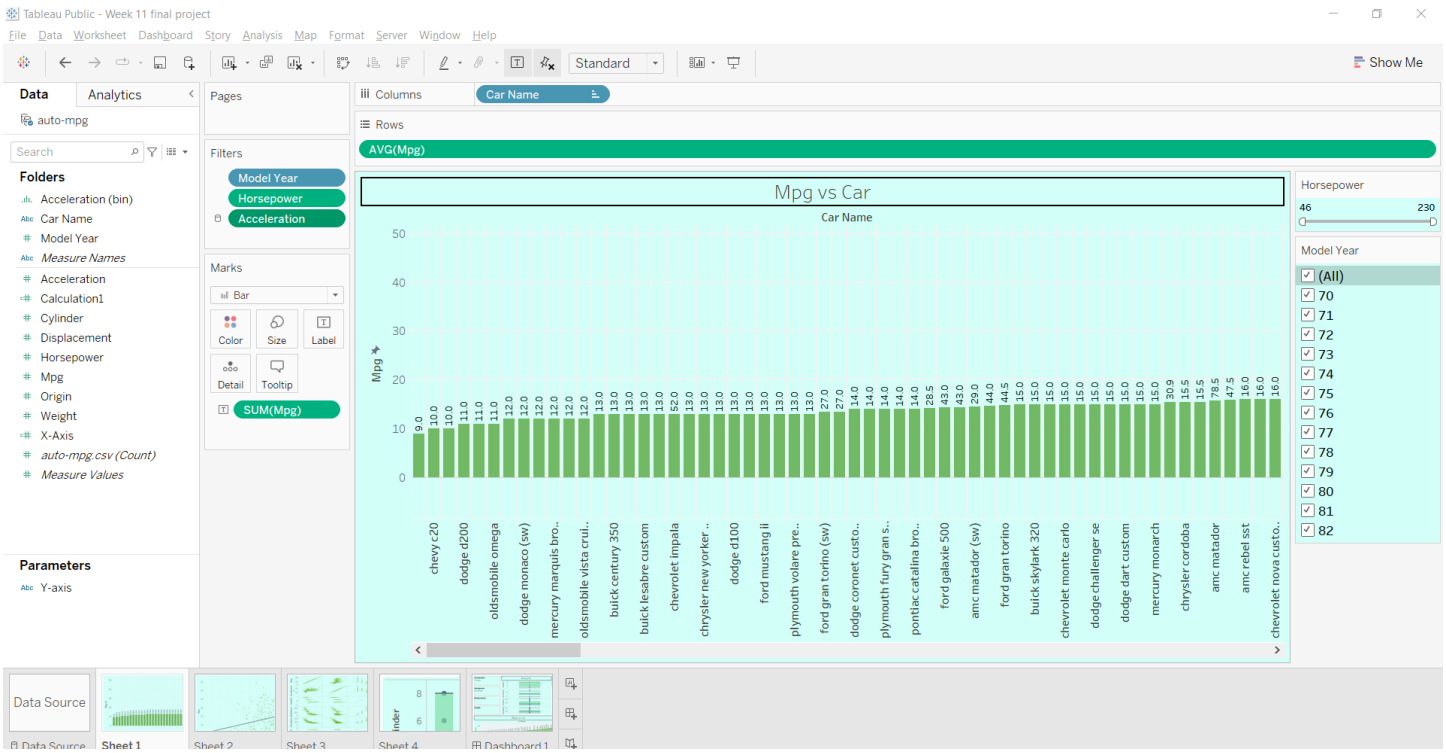
Auto – MPG – Data analysis.

In this project you will investigate the impact of a number of automobile engine factors on the vehicle's mpg. The dataset auto-mpg.csv contains information for 398 different automobile models. Information regarding the number of cylinders, displacement, horsepower, weight, acceleration, model year, origin, and car name as well as mpg are contained in the file.

Perform some initial analysis and create visualizations using Tableau Public

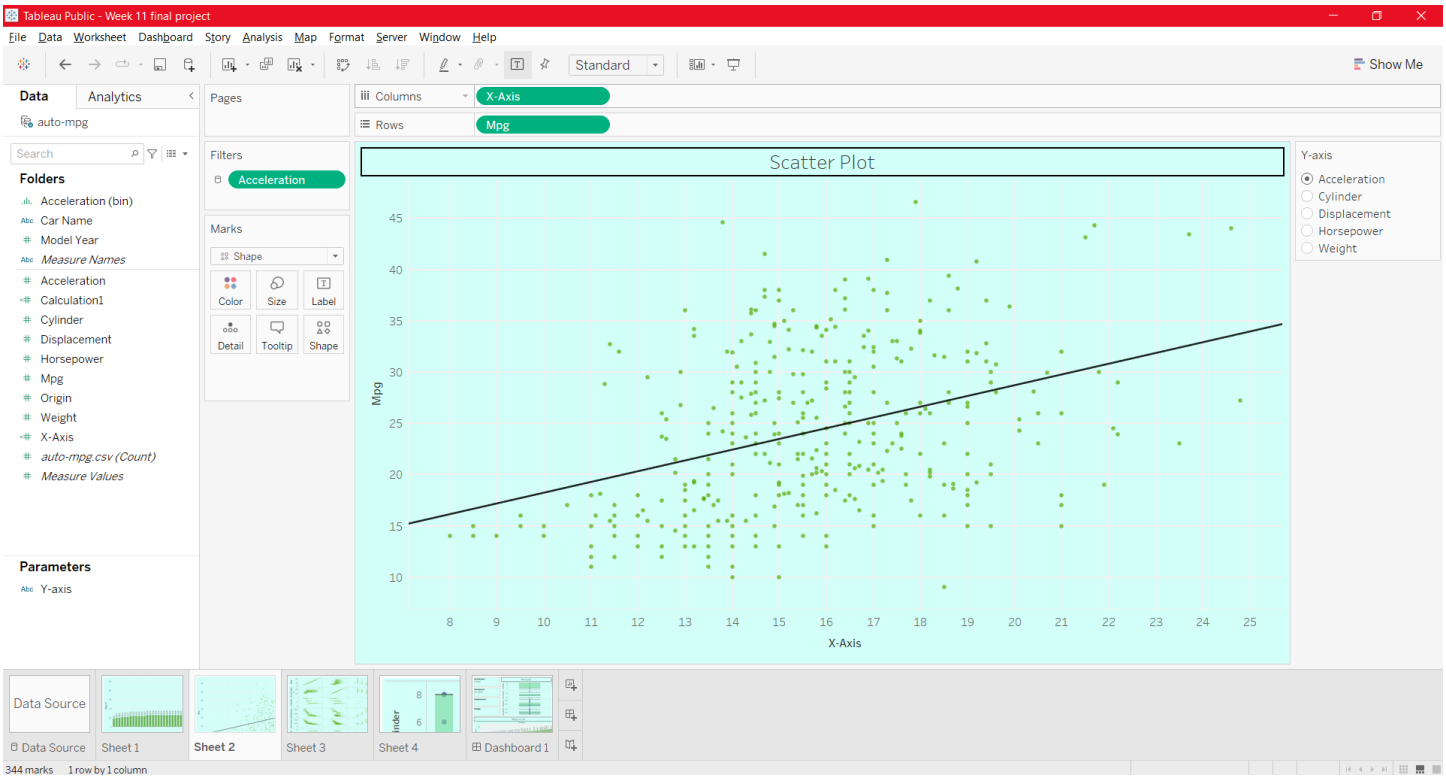
The visualization in done in tableau public and publish to tableau my public account - https://public.tableau.com/app/profile/eslin.kiran.ilangovan/viz/Week11finalproject/Dashboard1?publish=yes

Sheet 1 - MPG vs Car data is compared using box plot



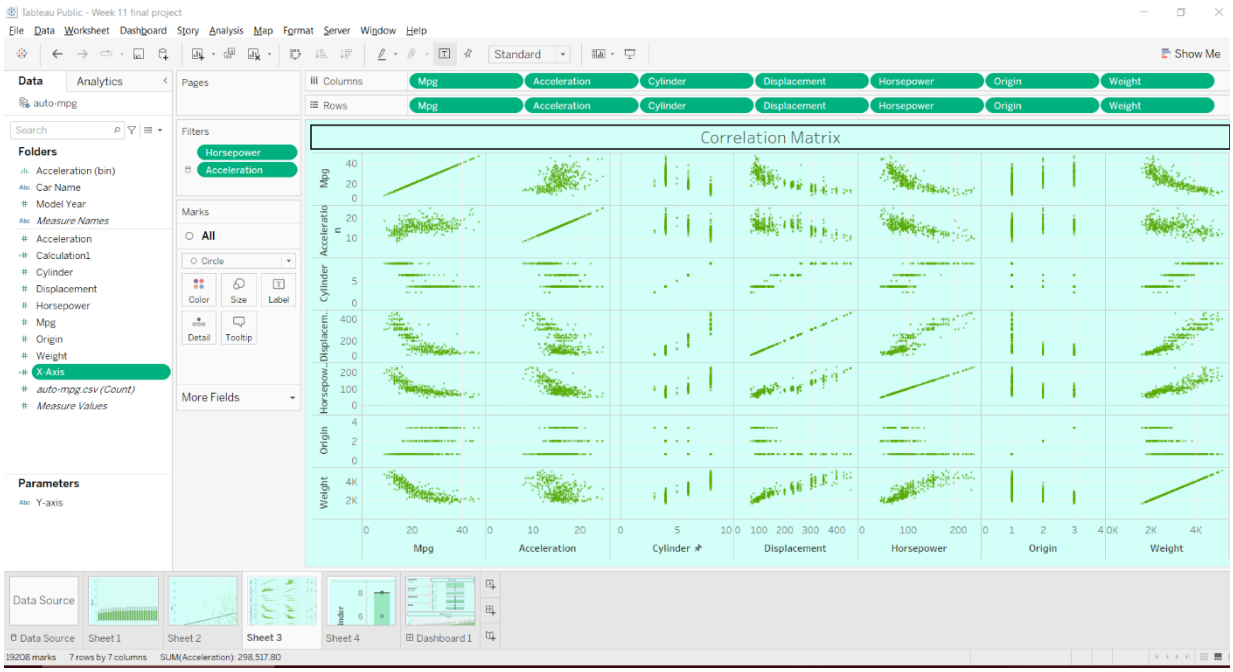
Sheet 2 – Scatter plot for different parameters of X value is given to compare with MGP.

X – Axis : Acceleration, Cylinder, Displacement, Horsepower, Weight Crreated using parameters and parameter calculation

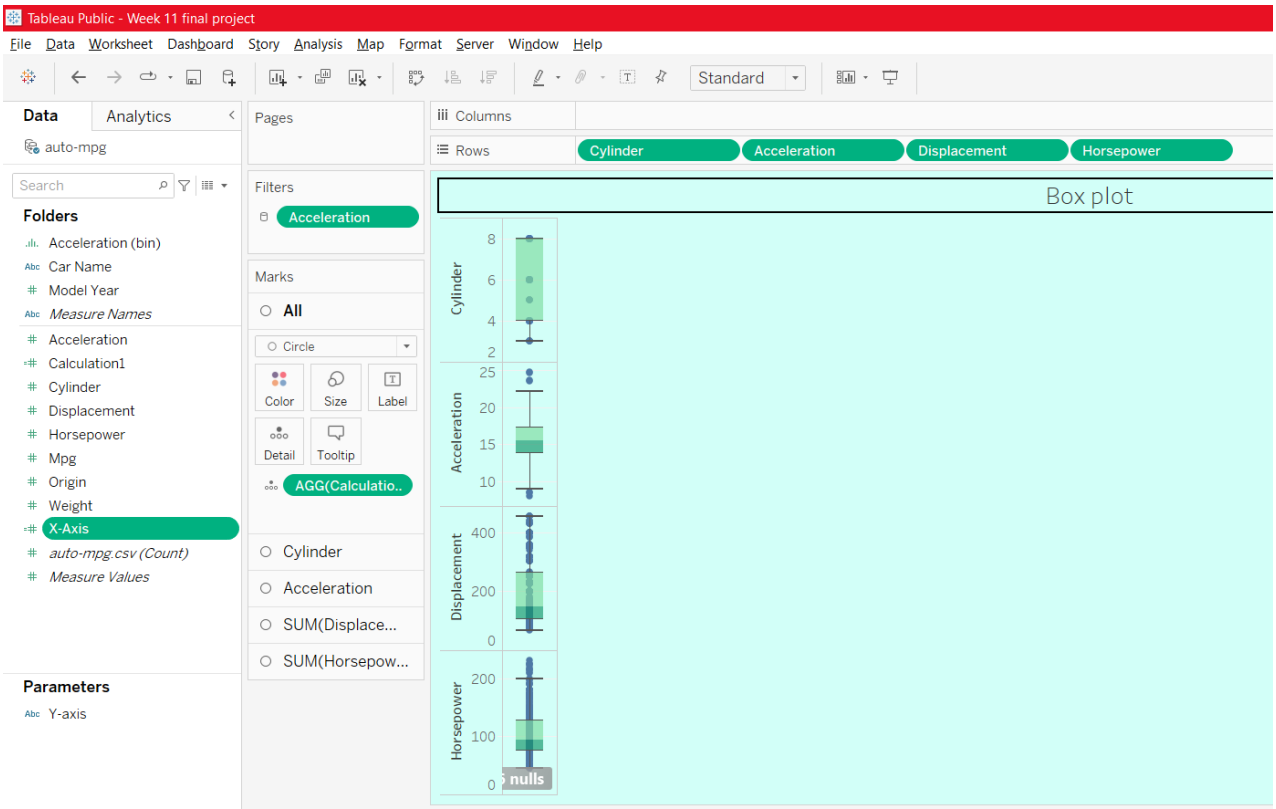


Calculation parameters.

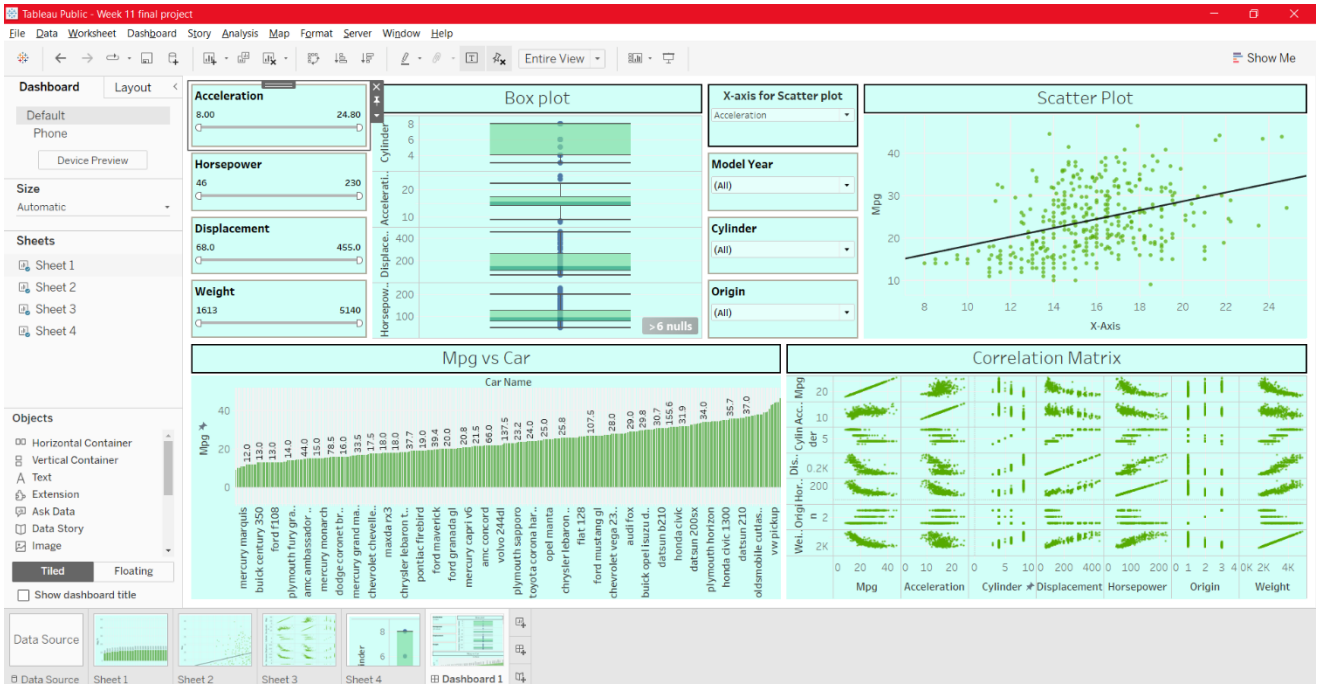
Sheet 3 – Correlation Matrix for the different variable is marked in a sheet to compare correlation coefficient of different fields.



Sheet 4 – Box plot is measured for Acceleration, Cylinder, Displacement, Horsepower to know the upper whisker, lower Whisker, Median and outliers.



Final Dashboard is created by using all the above sheets and filter is added to be sorted for all the values across the dashboard.



And finally publish in the Tableau Public in my account Eslin Kiran Ilangovan as Week 11 final project

Using the first 300 samples in the auto-mpg.csv, run a simple linear regression and multiple linear regression to determine the relationship between mpg and appropriate independent variable/(s). Report all the appropriate information regarding your regression.

RStudio -> File -> Knit Document / Compile Report -> Save as Word / PAUTO_MPG_DATA.

R-code.R

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readxl)
library(ggplot2)
library(ggpubr)
library(qqplotr)
```

```
library(e1071)
library(nortest)
library(BSDA)
```

```
##
## Attaching package: 'qqplotr'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     stat_qq_line, StatQqLine
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
library(psych)

library(caret)
library(leaps)
library(gvlma)

Auto_mpg_data <- read.csv(file.choose())
# structure of the dataframe
str(Auto_mpg_data)
```

```
## 'data.frame':    398 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinder    : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : chr  "130" "165" "150" "150" ...
##  $ 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  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ car.name    : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satell
ite" "amc rebel sst" ...
```

```r
# Converting the horsepower column to a numeric column
Auto_mpg_data$horsepower = as.numeric(Auto_mpg_data$horsepower)
```

```
## Warning: NAs introduced by coercion
```

```r
head(Auto_mpg_data)
```

```
##   mpg cylinder displacement horsepower weight acceleration model.year origin
## 1  18        8          307        130   3504         12.0         70      1
## 2  15        8          350        165   3693         11.5         70      1
## 3  18        8          318        150   3436         11.0         70      1
## 4  16        8          304        150   3433         12.0         70      1
## 5  17        8          302        140   3449         10.5         70      1
## 6  15        8          429        198   4341         10.0         70      1
##                   car.name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

```r
# Summary of the dataframe
summary(Auto_mpg_data)
```

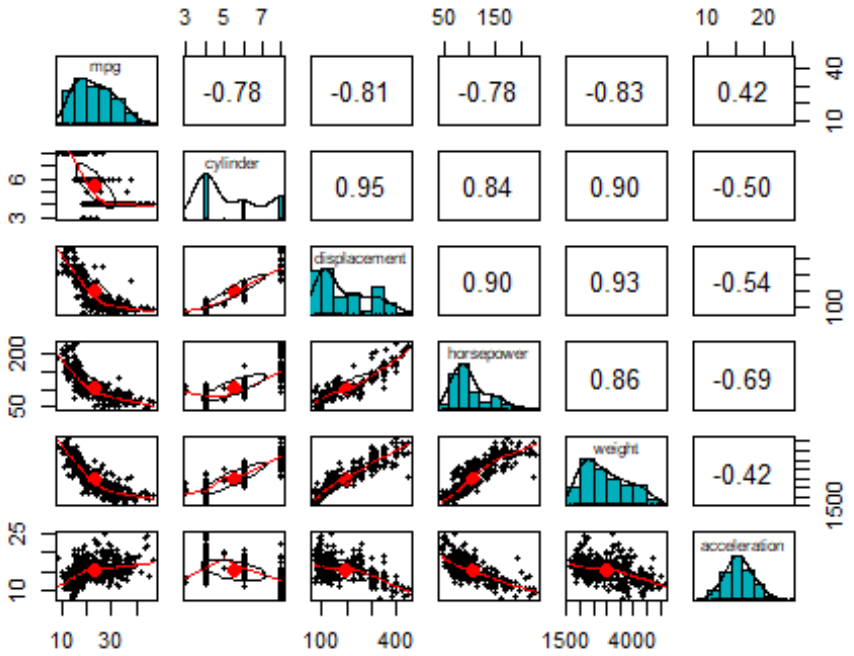
```
##       mpg           cylinder      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.2   1st Qu.: 75.0   1st Qu.:2224
##  Median :23.00   Median :4.000   Median :148.5   Median : 93.5   Median :2804
##  Mean   :23.51   Mean   :5.455   Mean   :193.4   Mean   :104.5   Mean   :2970
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0   3rd Qu.:126.0   3rd Qu.:3608
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##                                                  NA's   :6
##   acceleration     model.year        origin       car.name
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   Length:398
```

```
##  1st Qu.:13.82   1st Qu.:73.00   1st Qu.:1.000   Class :character
##  Median :15.50   Median :76.00   Median :1.000   Mode  :character
##  Mean   :15.57   Mean   :76.01   Mean   :1.573
##  3rd Qu.:17.18   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :24.80   Max.   :82.00   Max.   :3.000
##
```

```
# Removing all case with null value in any columns
Auto_mpg_data <- na.omit(Auto_mpg_data)
summary(Auto_mpg_data)
```

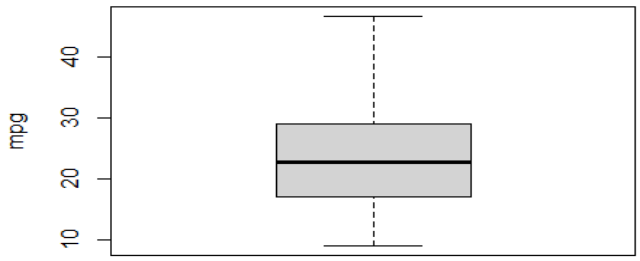
```
##       mpg           cylinder      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##   acceleration     model.year        origin        car.name
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   Length:392
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   Class :character
##  Median :15.50   Median :76.00   Median :1.000   Mode  :character
##  Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :24.80   Max.   :82.00   Max.   :3.000
```

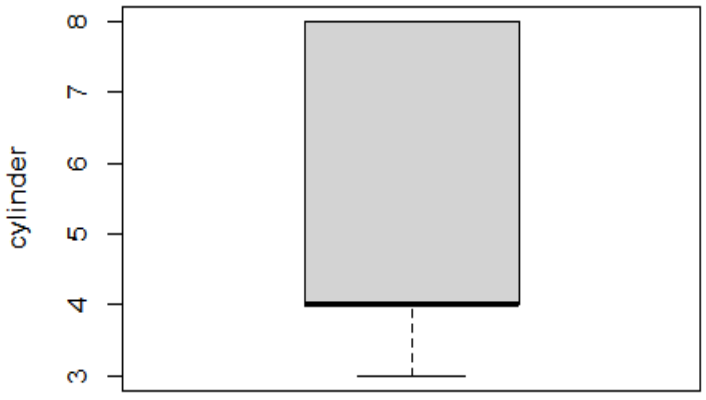
```
# Scatter plot of matrices
pairs.panels(Auto_mpg_data[,1:6],method = "pearson",hist.col ="#00AFBB" ,density = TRUE,e
llipses = TRUE)
```

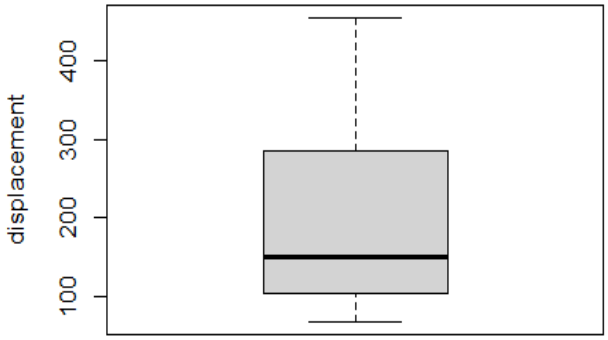


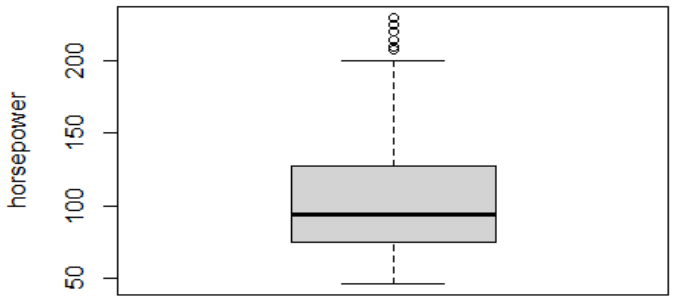
Three separate elements are shown in this output: the correlation between the variables, scatter plots showing the relationships between the variables, and histograms showing the skewness of the data. Notably, there are significant negative relationships between MPG and cylinder, displacement, horsepower, and weight. It is noted that there is multicollinearity among the independent variables. Some of the scatter plots show clear curves, while others indicate that weight and displacement and weight and horsepower have clear linear correlations. We are thinking about using weight and horsepower or weight and displacement in our final multiple linear regression model after looking at the plots of the independent variables against MPG.

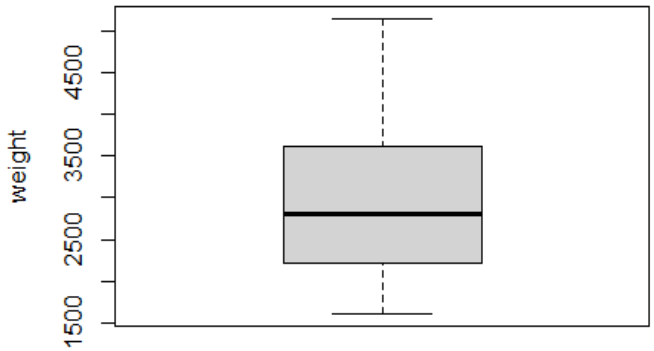
```
for (i in names(Auto_mpg_data[,1:6])) {
  boxplot(Auto_mpg_data[,i], names = "names(Auto_mpg_data[,i])", ylab = i)
```

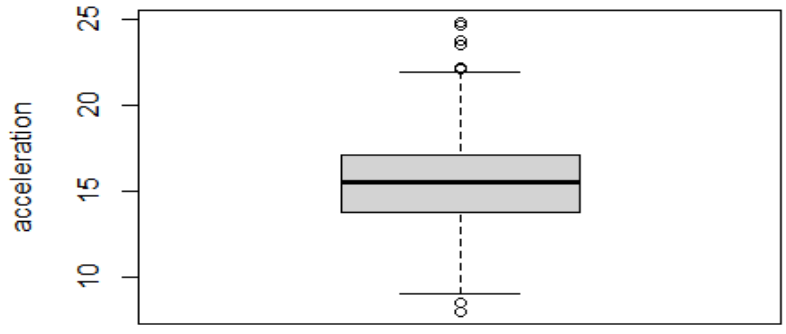






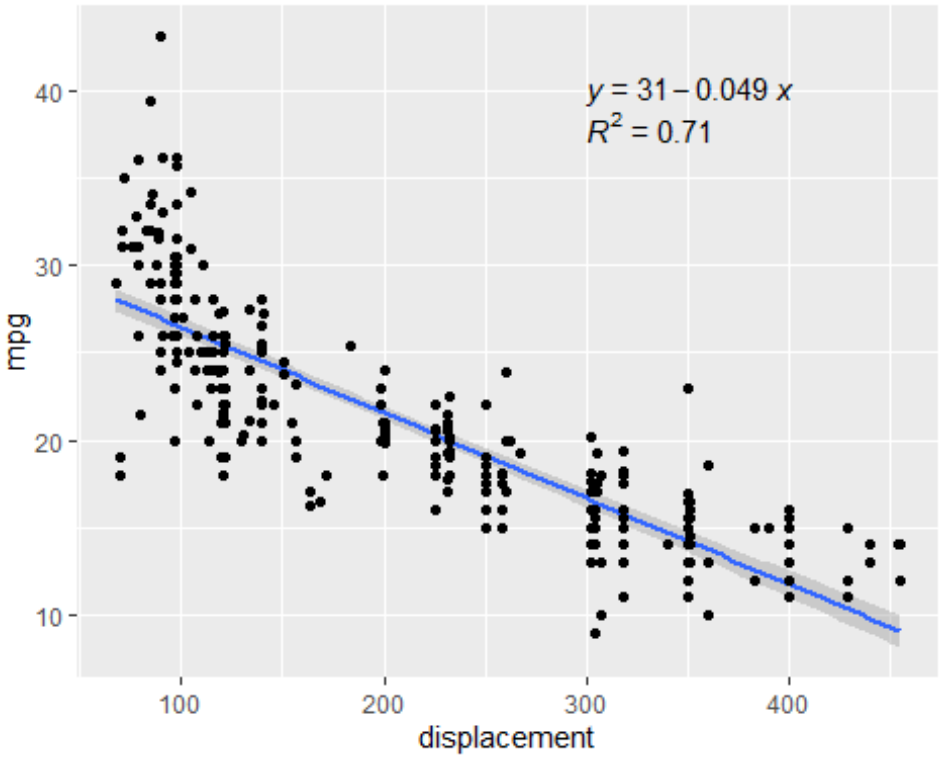






Using Box Plot we have found that some data's in horsepower and acceleration is outlier

```
# To find the association between mpg and an other single variable, perform a basic linea
r regression using the first 300 samples in the data frame.

Auto_mpg_data_train <- Auto_mpg_data[1:300,1:6]
Auto_mpg_data_test <- Auto_mpg_data[301:nrow(Auto_mpg_data),1:6]

ggplot(data=Auto_mpg_data_train, aes(x=displacement, y=mpg)) +
  geom_smooth(method="lm") +
  geom_point() +
  stat_regline_equation(label.x=300, label.y=40) +
  stat_cor(aes(label=..rr.label..), label.x=300, label.y=38)

## Warning: The dot-dot notation (`..rr.label..`) was deprecated in ggplot2 3.4.0.
## ℹ Please use `after_stat(rr.label)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## `geom_smooth()` using formula = 'y ~ x'
```



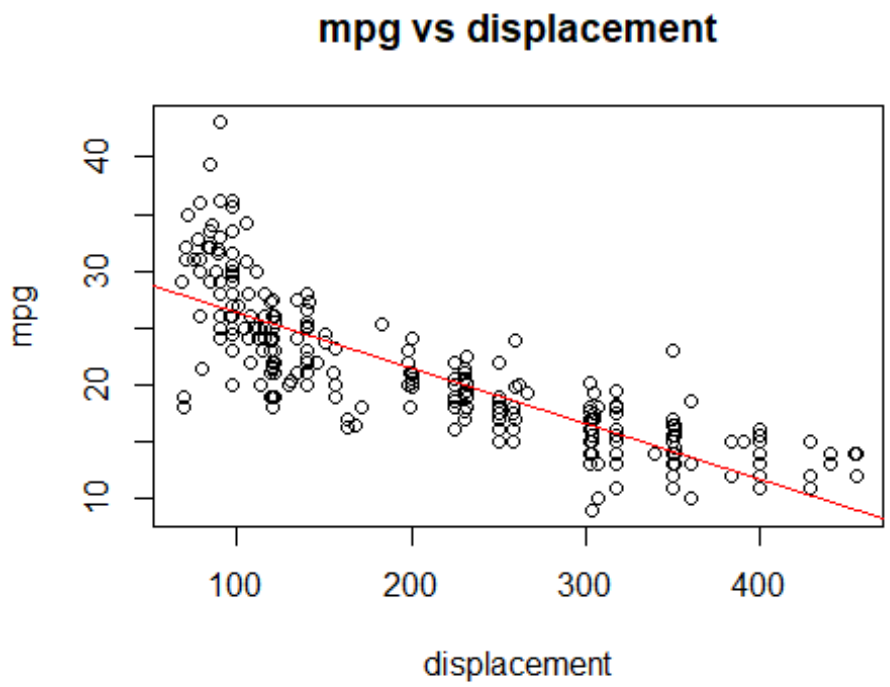

```
#performing regression
Model_1<- lm(mpg~displacement, data=Auto_mpg_data_train)
summary(Model_1)

##
## Call:
## lm(formula = mpg ~ displacement, data = Auto_mpg_data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9282 -2.0043 -0.5401  1.9737 16.1501
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.352035   0.435875   71.93   <2e-16 ***
## displacement -0.048913   0.001809  -27.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.412 on 298 degrees of freedom
## Multiple R-squared:  0.7104, Adjusted R-squared:  0.7094
## F-statistic: 731.1 on 1 and 298 AUTO_MPG_DATA,  p-value: < 2.2e-16
```

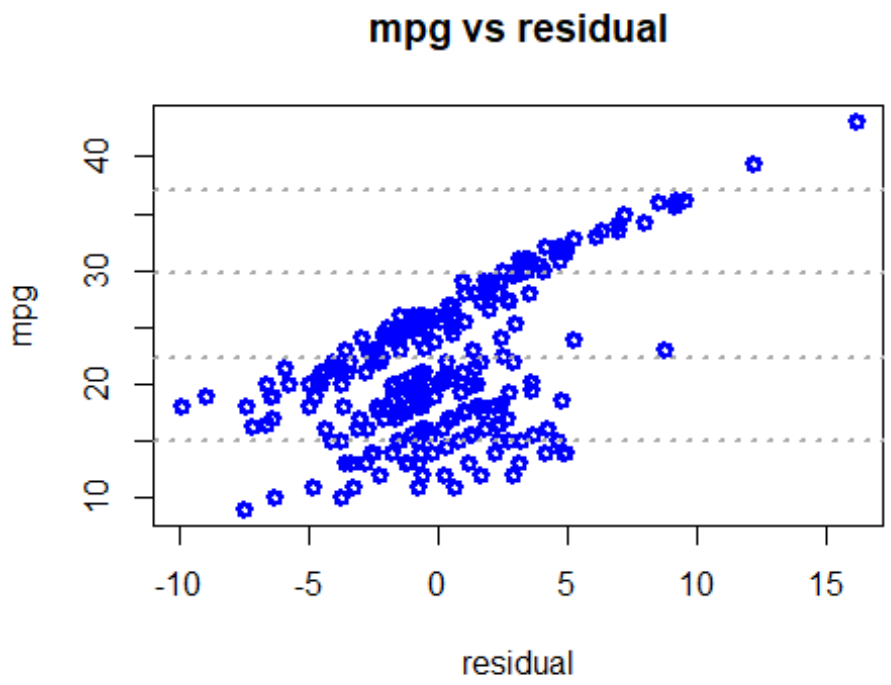
```
# mean of residuals
mean(resid(Model_1))
```

```
## [1] 3.035766e-16
```

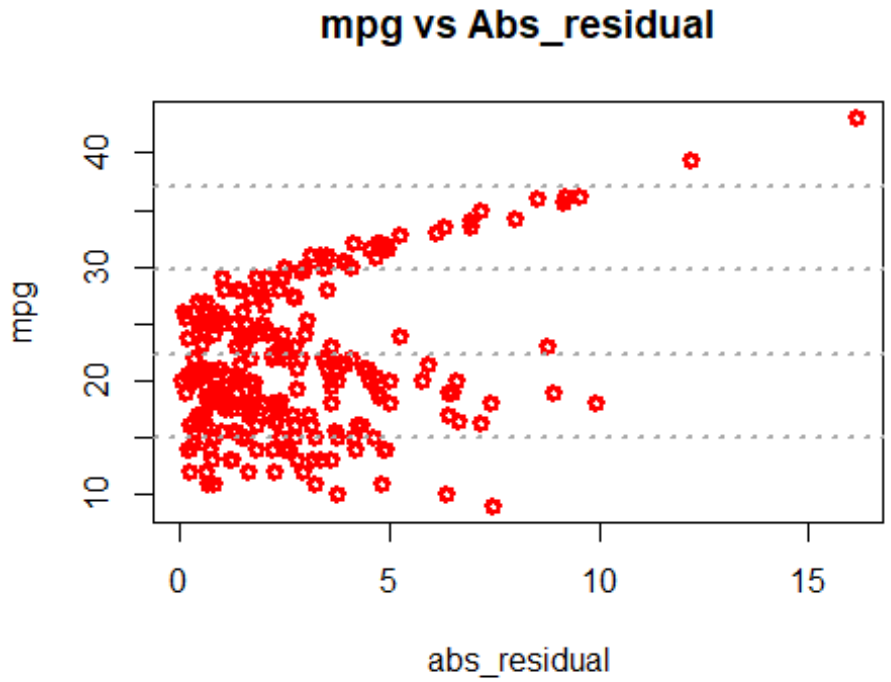
```
plot(Auto_mpg_data_train$mpg~Auto_mpg_data_train$displacement,main="mpg vs displacement",
xlab="displacement",ylab = "mpg")
abline(dis_model,col="red")
```




```
#residuals vs. the predictor variable
residual <- Model_1$residuals
plot(Auto_mpg_data_train$mpg~residual,lwd=3, col="blue",main="mpg vs residual", xlab="res
idual",ylab = "mpg")
grid(NA, 5, lwd = 2,col = "darkgray")
```

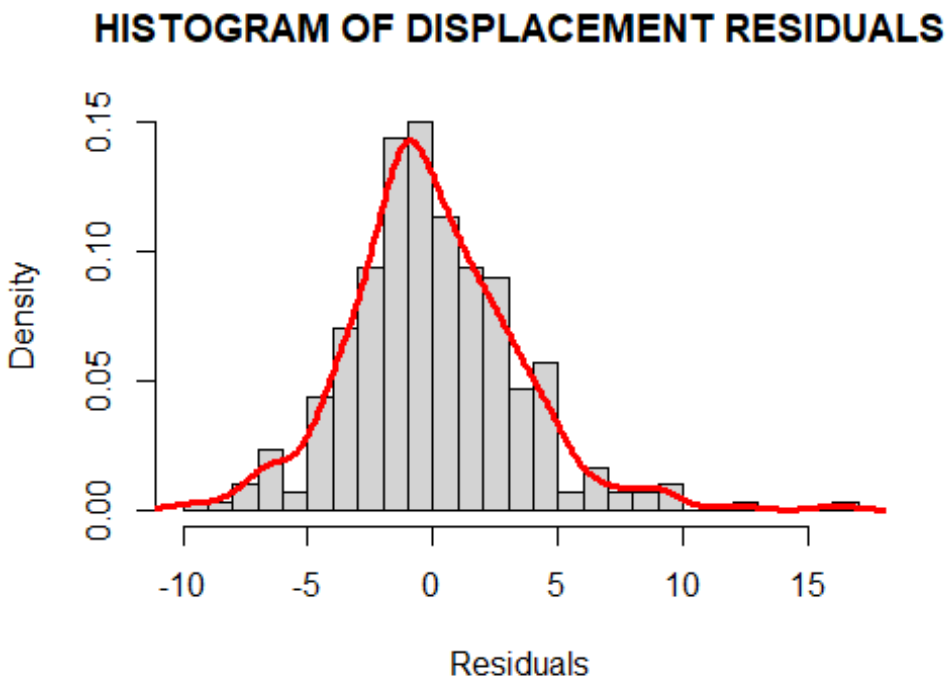



```
#absolute value of the residuals vs. the predictor variable
abs_residual <- abs(residual)
plot(Auto_mpg_data_train$mpg~abs_residual,lwd=3, col="red",main="mpg vs Abs_residual", xl
ab="abs_residual",ylab = "mpg")
grid(NA, 5, lwd = 2,col = "darkgray")
```

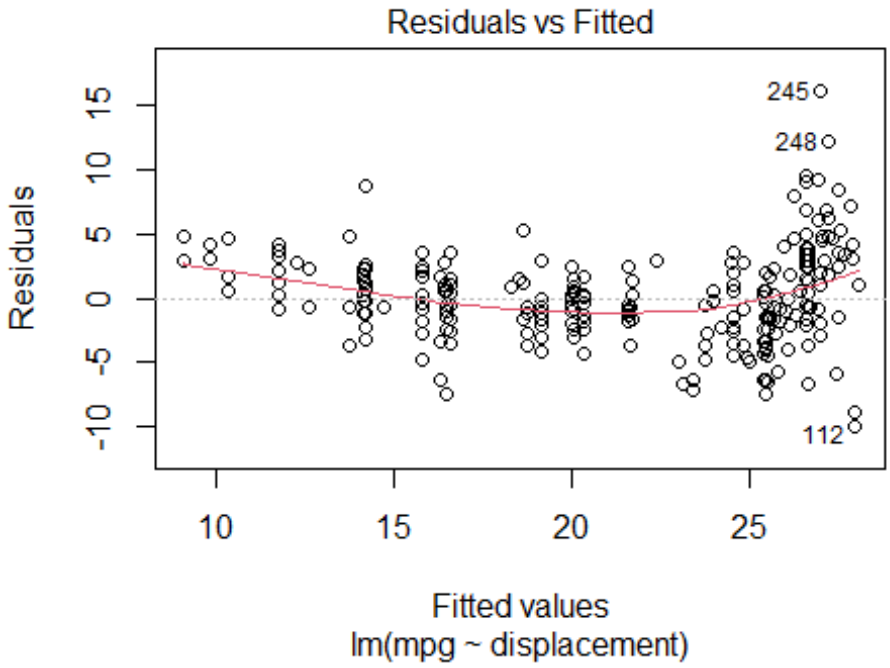


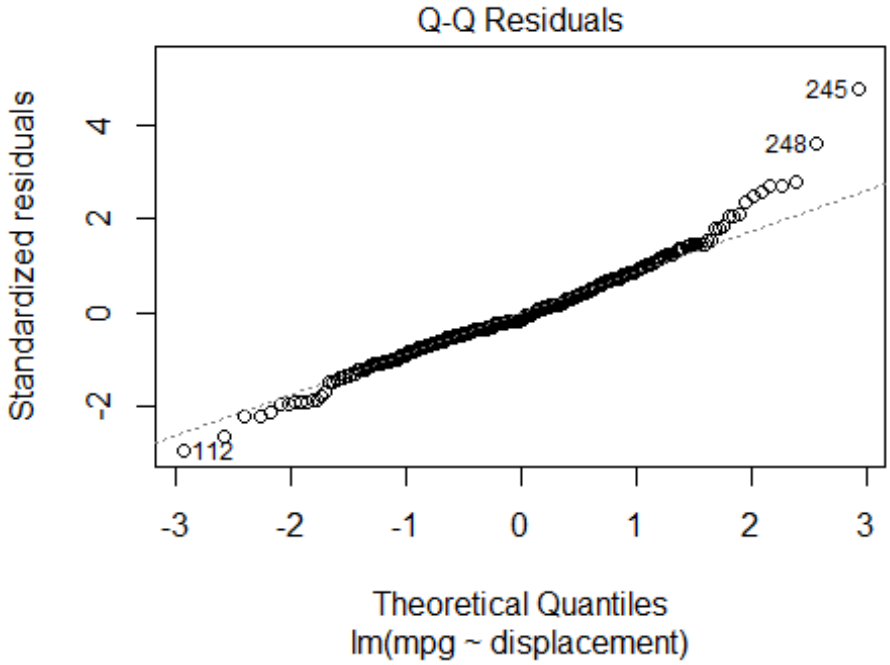

```
#histogram of the residuals
hist(residual,prob=T,breaks=20,main="HISTOGRAM OF DISPLACEMENT RESIDUALS",xlab="Residuals
```

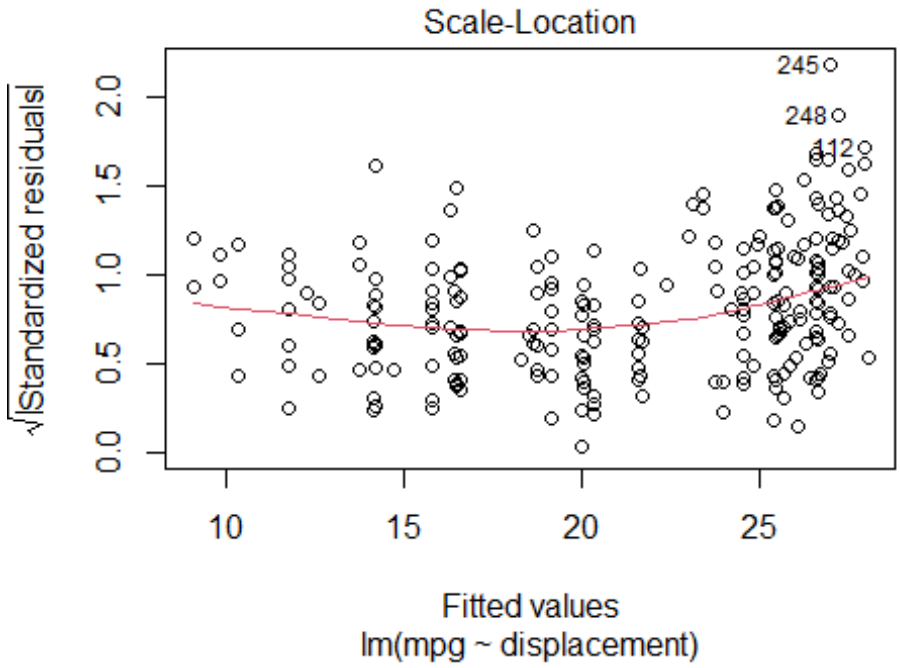
```
")
lines(density(residual),col="red",lwd=3)
```

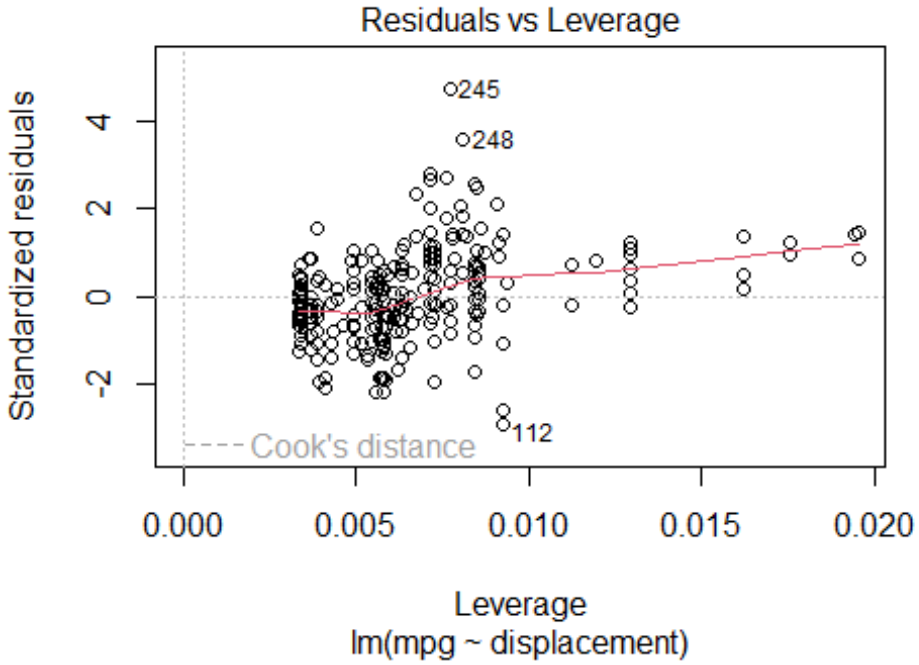



```
plot(dis_model)
```


Residuals vs Fitted
Residuals
245
248
112
-10
-5
0
5
10
15
10
15
20
25
Fitted values
lm(mpg ~ displacement)


Q-Q Residuals
Standardized residuals
245
248
112
-2
0
2
4
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(mpg ~ displacement)






```
# Make predictions and compute the R2, RMSE and MAE
dis_predict <- Model_1%>% predict(Auto_mpg_data_test)
data.frame( R2 = R2(dis_predict, Auto_mpg_data_test$mpg),
            RMSE = RMSE(dis_predict, Auto_mpg_data_test$mpg),
            MAE = MAE(dis_predict, Auto_mpg_data_test$mpg))

##         R2     RMSE      MAE
## 1 0.370789 8.371337 7.050203
```

```r
prediction_error = RMSE(dis_predict, Auto_mpg_data_test$mpg)/mean(Auto_mpg_data_test$mpg)
prediction_error
```

```
## [1] 0.2620493
```

```r
compare_dis = as.data.frame(cbind(Auto_mpg_data_test$mpg,dis_predict),row=FALSE)
names(compare_dis) = c("observed","dis_predict")
head(compare_dis)
```

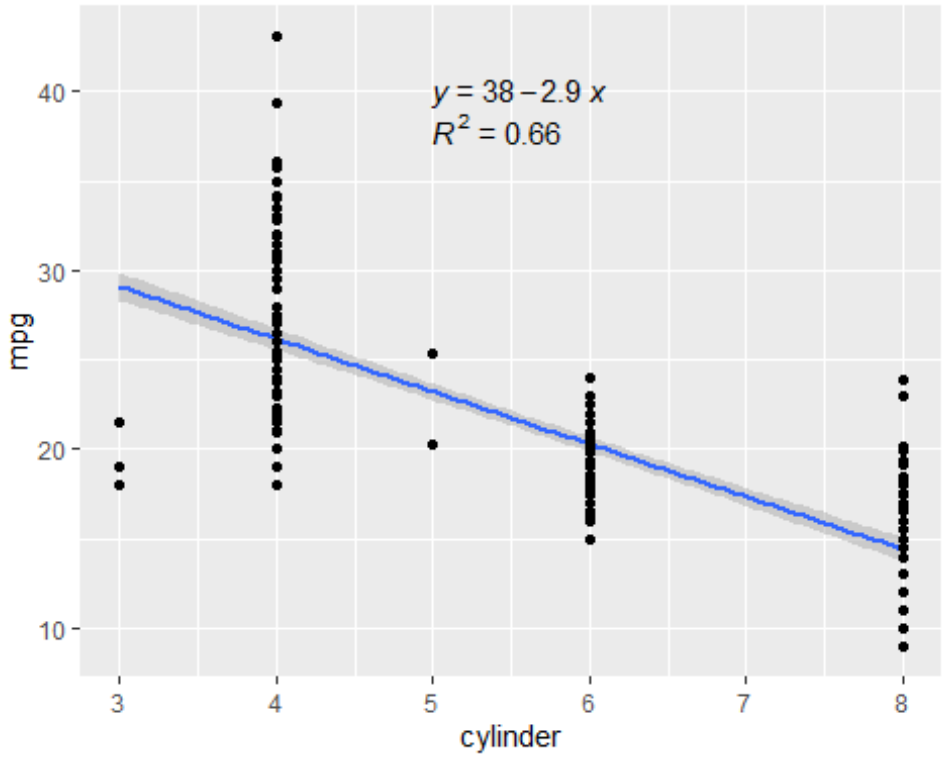
```
##   observed dis_predict
## 1     34.5    26.21621
## 2     31.8    27.19446
## 3     37.3    26.90099
## 4     28.4    23.96623
## 5     28.8    22.89016
## 6     26.8    22.89016
```

All of the estimated values in this model have statistical significance, as shown by p-values smaller than 2e-16. Notably, a non-linear pattern is shown by the MPG vs. Displacement plot, pointing to a complex interaction be tween the variable and the residuals. This model does not meet the optimal standards.

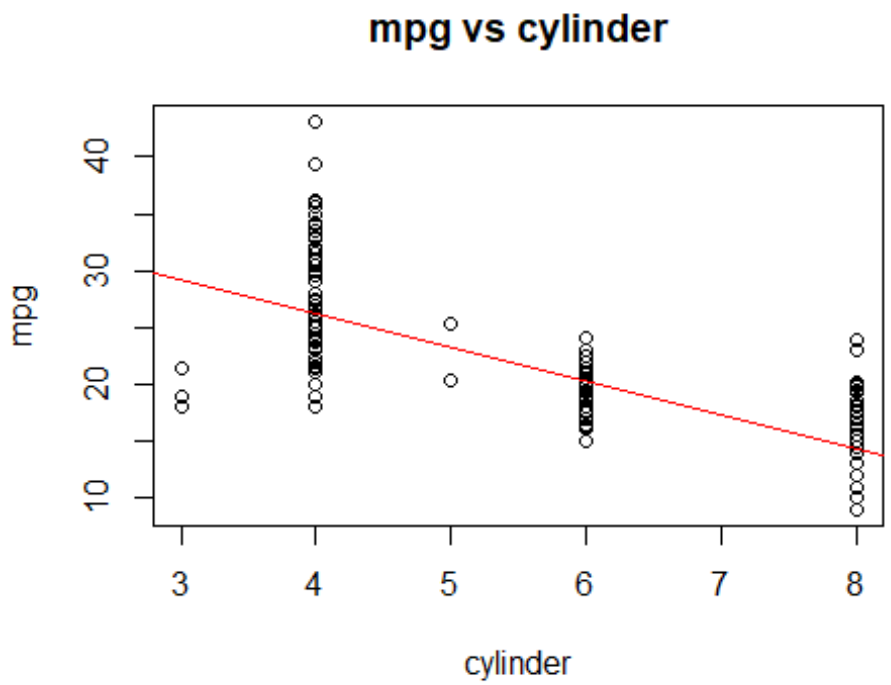
Outliers are detected at data points 112, 245, and 248 via diagnostic plots. Based on the modified R-square, di splacement accounts for roughly 70.94% of the variance in MPG.

```r
#Model with cylinder as explanatory variable

# Linear regression plot
ggplot(data=Auto_mpg_data_train, aes(x=cylinder, y=mpg)) +
  geom_smooth(method="lm") +
  geom_point() +
  stat_regline_equation(label.x=5, label.y=40) +
  stat_cor(aes(label=..rr.label..), label.x=5, label.y=38)
```

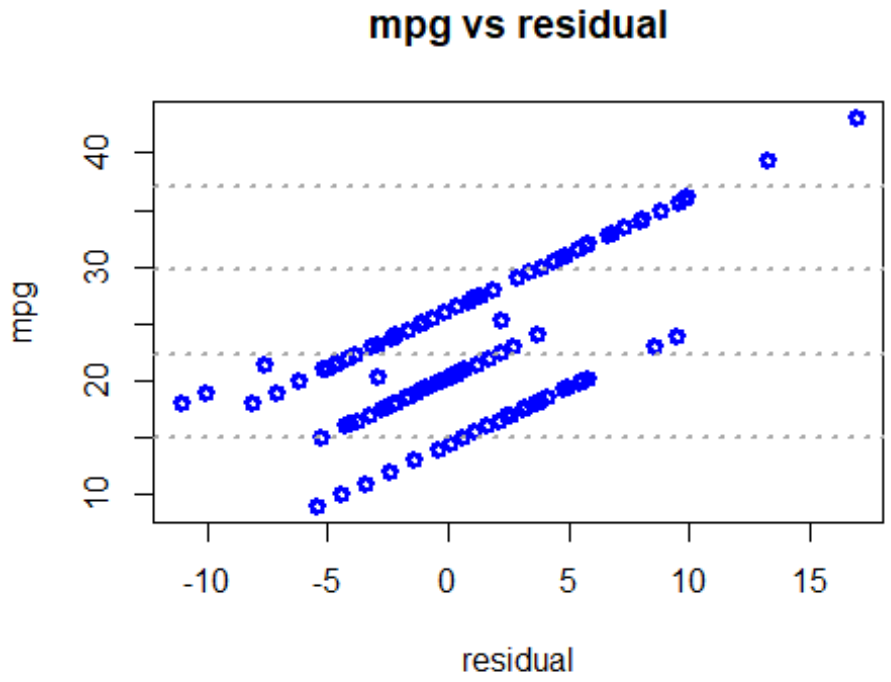
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#performing regression
cylinder_model <- lm(mpg~cylinder, data=Auto_mpg_data_train)
summary(cylinder_model)

##
## Call:
## lm(formula = mpg ~ cylinder, data = Auto_mpg_data_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1071  -2.3012  -0.4306   1.8282  16.9282
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.9130     0.7356   51.54   <2e-16 ***
## cylinder     -2.9353     0.1211  -24.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.678 on 298 degrees of freedom
## Multiple R-squared:  0.6636, Adjusted R-squared:  0.6624
## F-statistic: 587.8 on 1 and 298 AUTO_MPG_DATA,  p-value: < 2.2e-16

# mean of residuals
mean(resid(cylinder_model))

## [1] 2.777408e-16

#plot the variable
plot(Auto_mpg_data_train$mpg~Auto_mpg_data_train$cylinder,main="mpg vs cylinder",xlab="cy
linder",ylab = "mpg")
abline(cylinder_model,col="red")
```

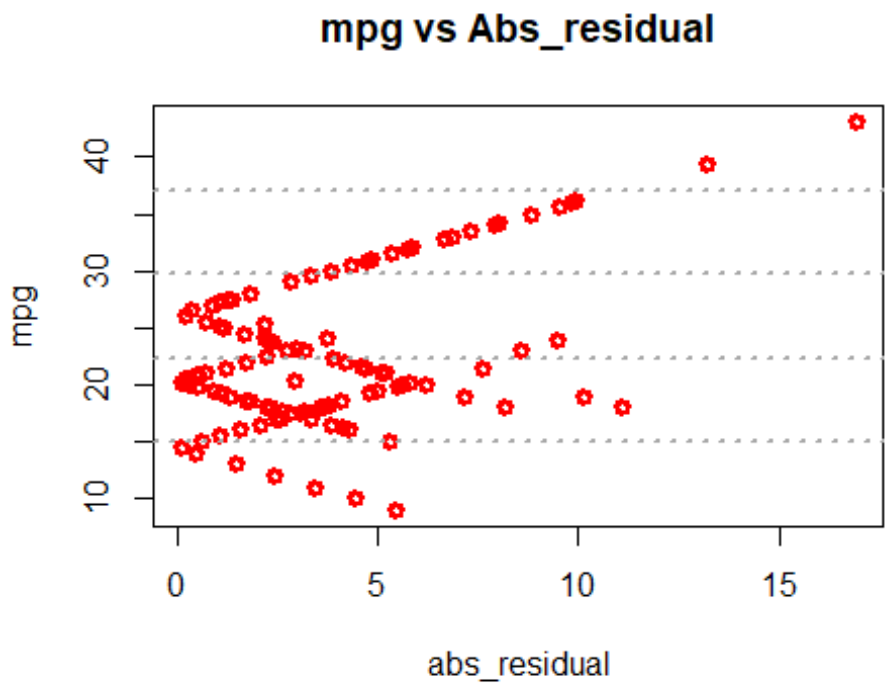



```r
#residuals vs. the predictor variable
residual <- cylinder_model$residuals
plot(Auto_mpg_data_train$mpg~residual,lwd=3, col="blue",main="mpg vs residual", xlab="res
idual",ylab = "mpg")
grid(NA, 5, lwd = 2,col = "darkgray")
```

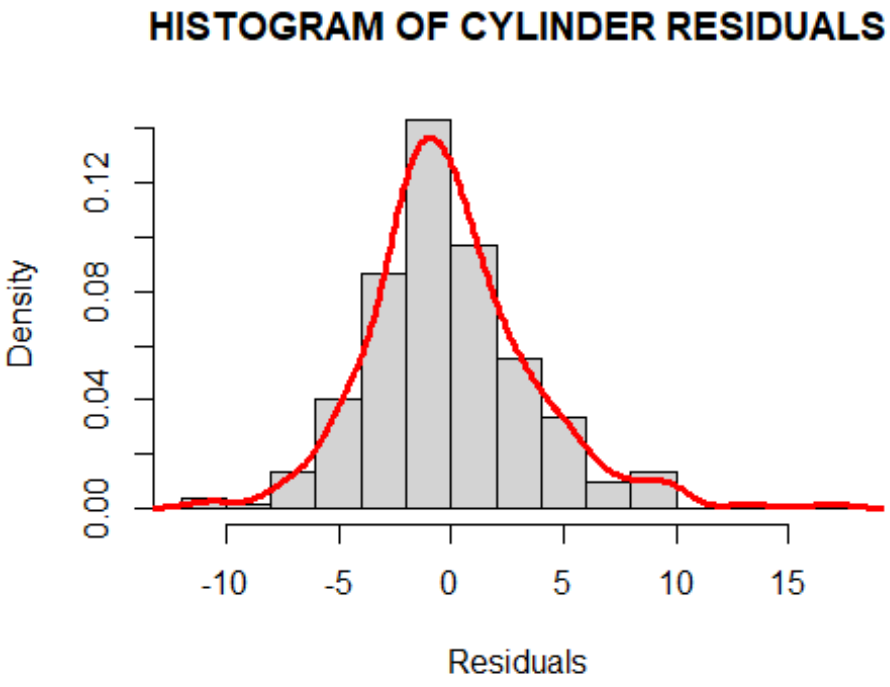



```r
#absolute value of the residuals vs. the predictor variable
abs_residual <- abs(residual)
plot(Auto_mpg_data_train$mpg~abs_residual,lwd=3, col="red",main="mpg vs Abs_residual", xl
```

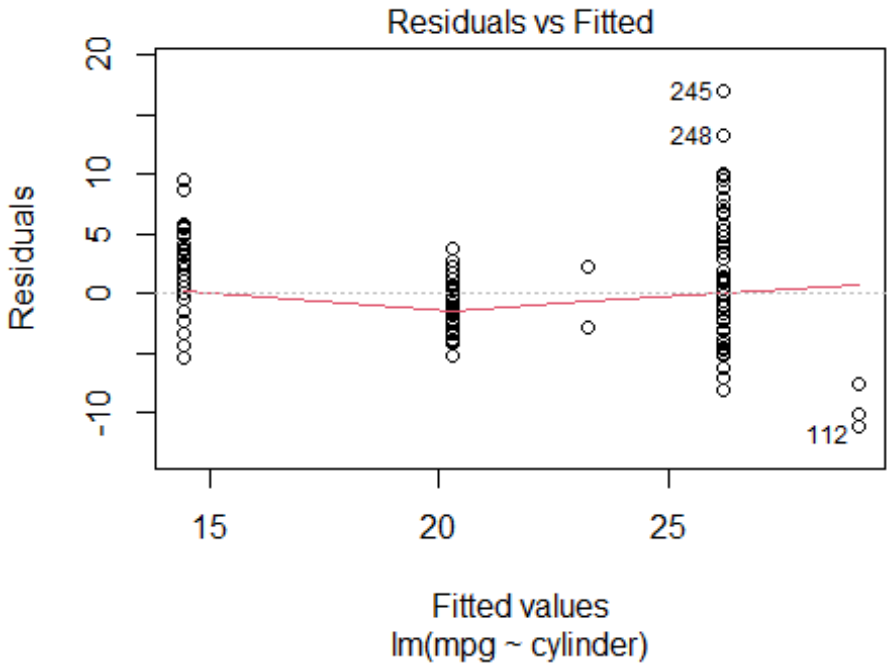
```
ab="abs_residual",ylab = "mpg")
grid(NA, 5, lwd = 2,col = "darkgray")
```

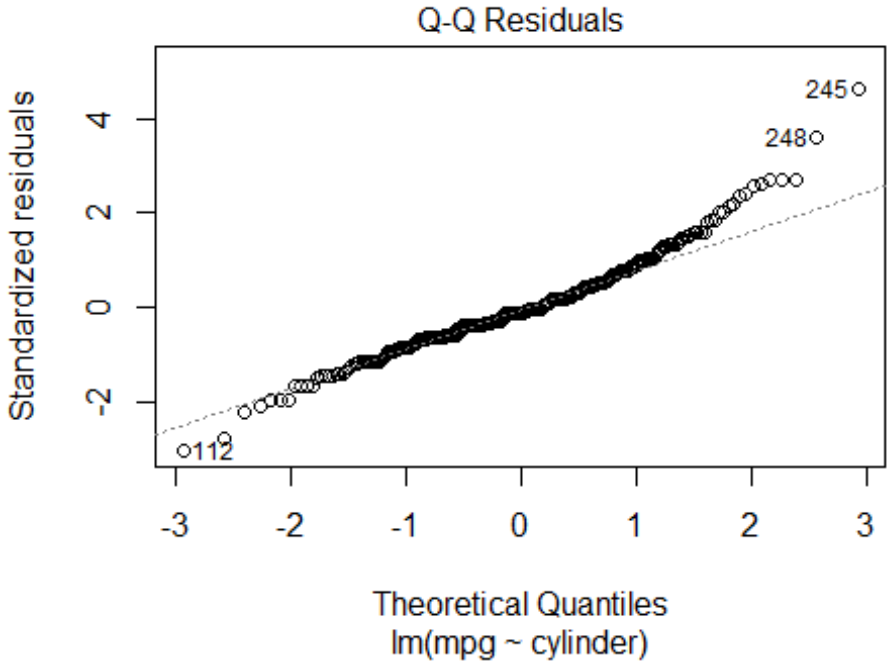


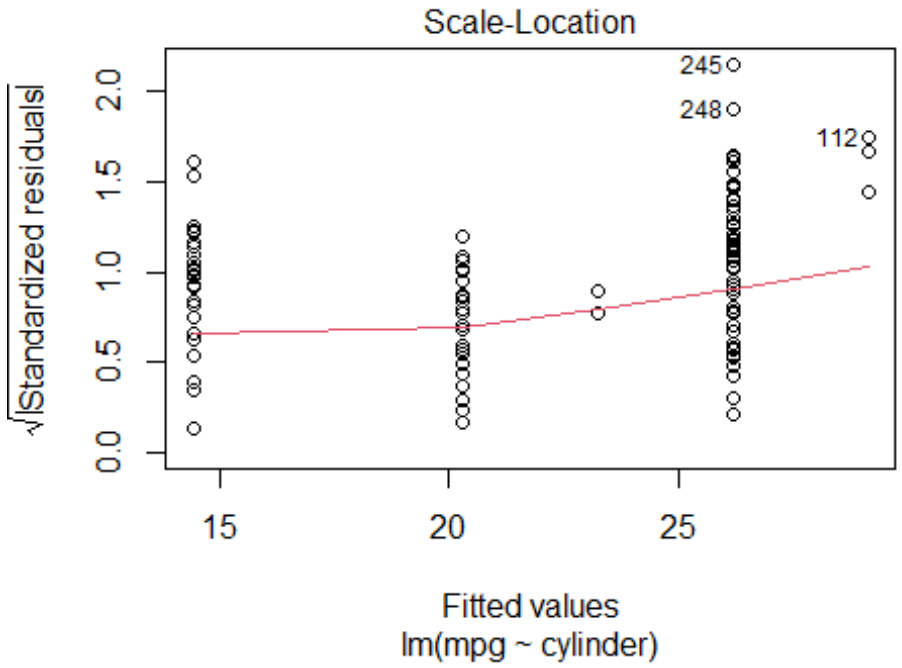

```
#histogram of the residuals
hist(residual,prob=T,breaks=20,main="HISTOGRAM OF CYLINDER RESIDUALS",xlab="Residuals")
lines(density(residual),col="red",lwd=3)
```

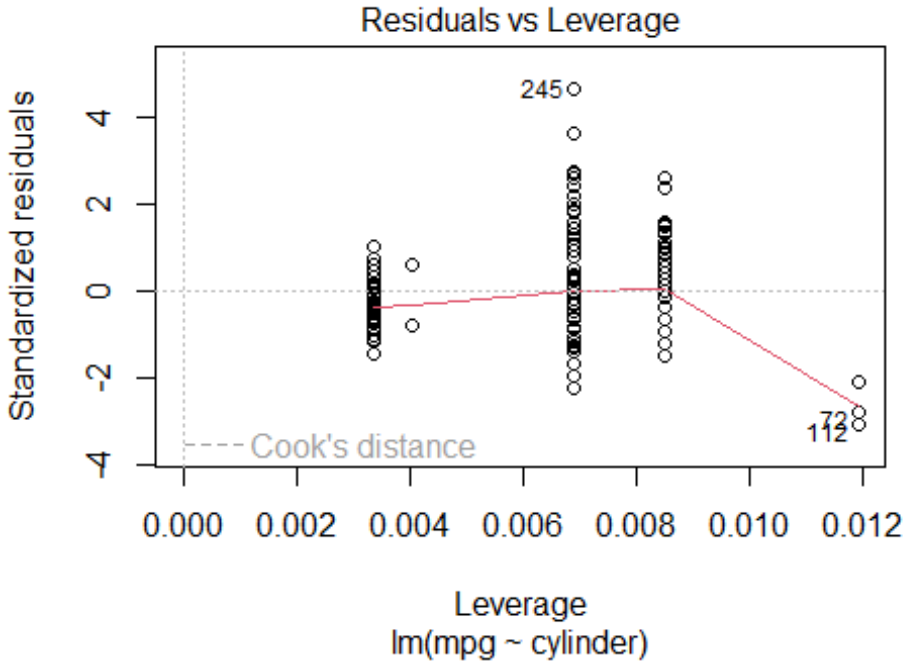



```
plot(cylinder_model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ cylinder)
245
248
112
20
10
5
0
-10
15
20
25


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(mpg ~ cylinder)
245
248
112
4
2
0
-2
-3
-2
-1
0
1
2
3






```
# Make predictions and compute the R2, RMSE and MAE
cyl_predict <- cylinder_model %>% predict(Auto_mpg_data_test)
data.frame( R2 = R2(cyl_predict, Auto_mpg_data_test$mpg),
            RMSE = RMSE(cyl_predict, Auto_mpg_data_test$mpg),
            MAE = MAE(cyl_predict, Auto_mpg_data_test$mpg))

##          R2     RMSE      MAE
## 1 0.1829322 8.611541 7.099133
```

```
compare_cyl = as.data.frame(cbind(Auto_mpg_data_test$mpg,cyl_predict),row=FALSE)
names(compare_cyl) = c("observed","cyl_predict")
head(compare_cyl)
```

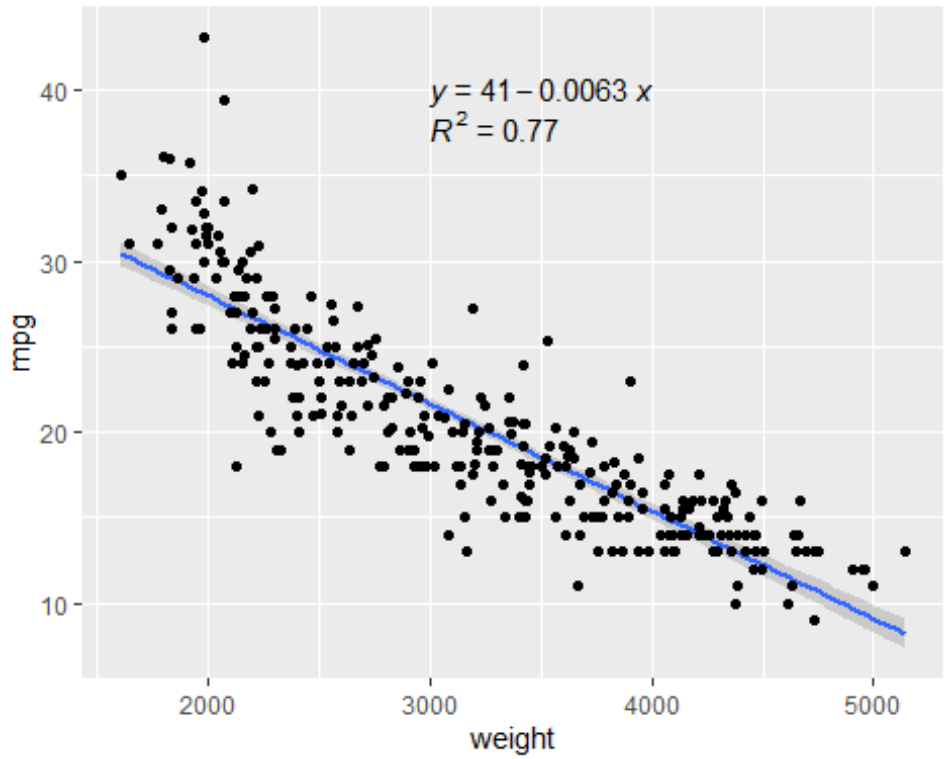
```
##   observed cyl_predict
## 1     34.5    26.17179
## 2     31.8    26.17179
## 3     37.3    26.17179
## 4     28.4    26.17179
## 5     28.8    20.30120
## 6     26.8    20.30120
```

P-values less than 2e-16 imply that every predicted value in this model is statis tically significant.

But the MPG vs. cylinder plot points to a non-linear relationship, which complica tes the variable's link to the residuals. It is believed that this model is not i deal. Outliers are detected using diagnostic plots at data points 245, 248, and 1 12. According to the corrected R-square, the cylinder variable accounts for rough ly 66.24% of the variance in MPG.

```
#weight as explanatory variable

# Linear regression plot
ggplot(data=Auto_mpg_data_train, aes(x=weight, y=mpg)) +
  geom_smooth(method="lm") +
  geom_point() +
  stat_regline_equation(label.x=3000, label.y=40) +
  stat_cor(aes(label=..rr.label..), label.x=3000, label.y=38)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



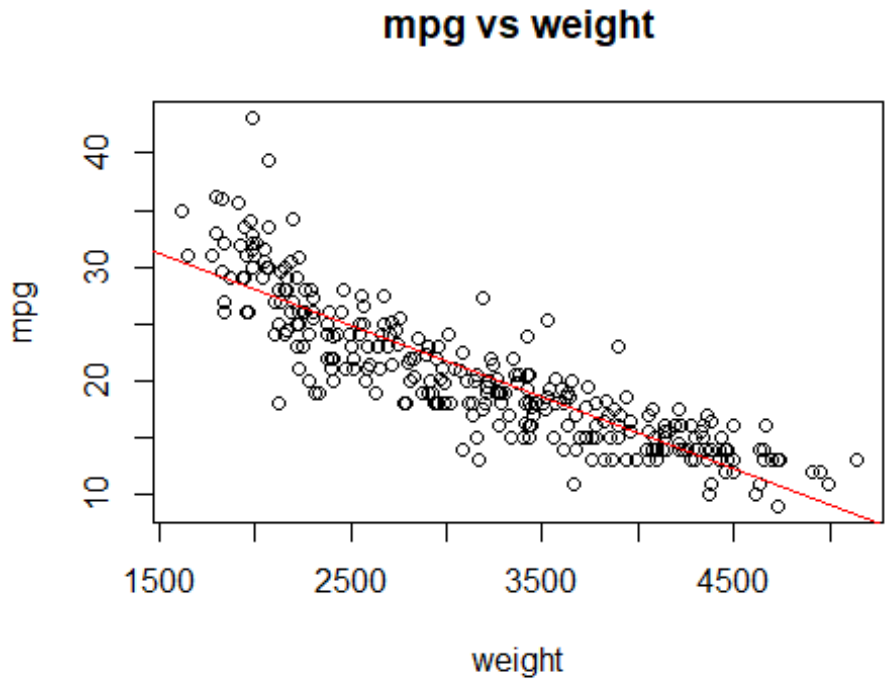
```r
#performing regression
weight_model <- lm(mpg~weight, data=Auto_mpg_data_train)
summary(weight_model)
```

```
##
## Call:
## lm(formula = mpg ~ weight, data = Auto_mpg_data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2011 -1.9157 -0.0812  1.7341 15.0246
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.5619792  0.6461532   62.77   <2e-16 ***
## weight      -0.0062905  0.0001984  -31.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.032 on 298 degrees of freedom
## Multiple R-squared:  0.7714, Adjusted R-squared:  0.7706
## F-statistic:  1005 on 1 and 298 AUTO_MPG_DATA,  p-value: < 2.2e-16
```

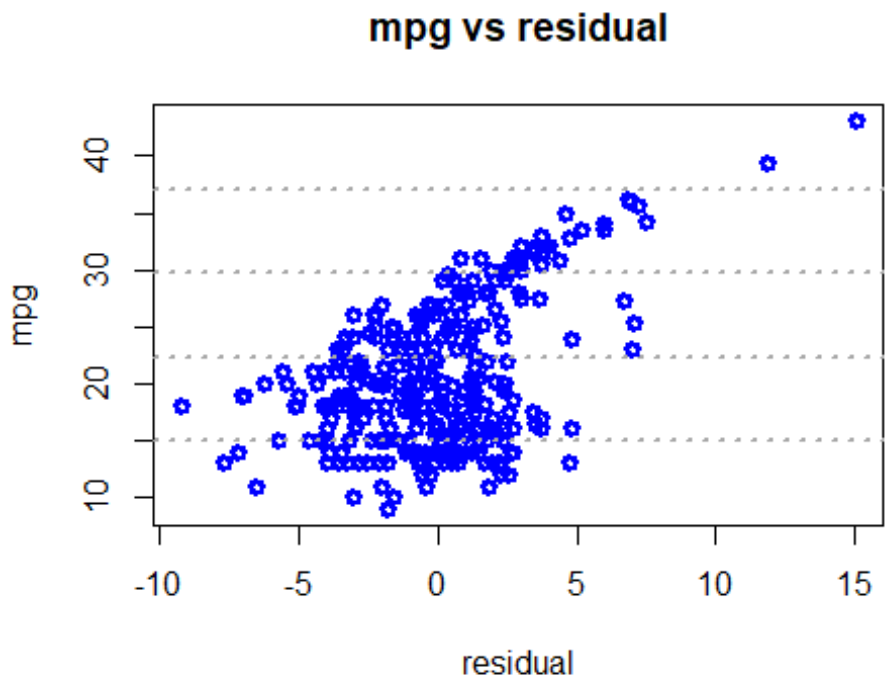
```r
# mean of residuals
mean(resid(weight_model))
```

```
## [1] 2.543538e-16
```

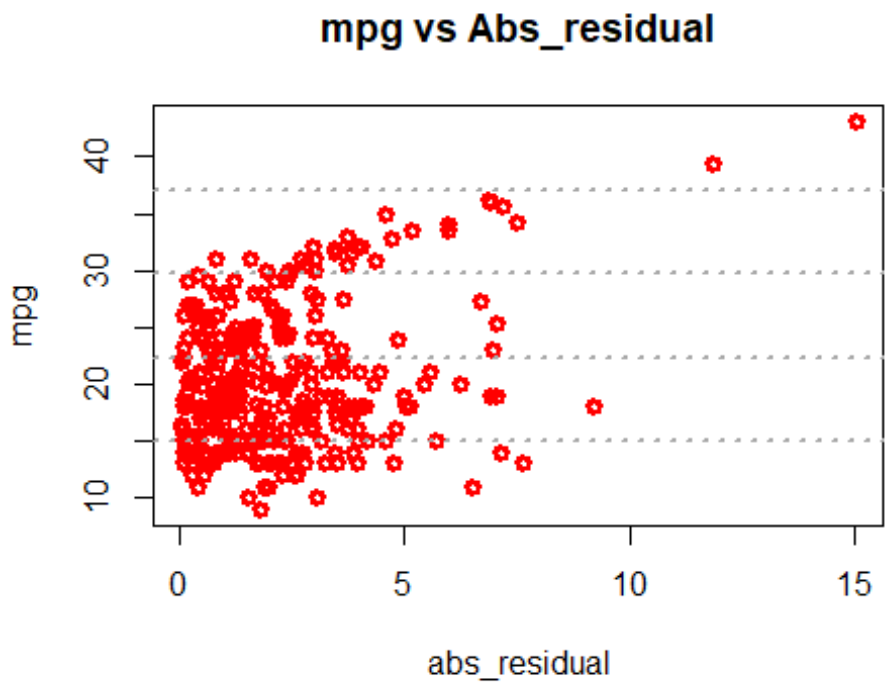
```r
#plot the variable
plot(Auto_mpg_data_train$mpg~Auto_mpg_data_train$weight,main="mpg vs weight",xlab="weight
",ylab = "mpg")
abline(weight_model,col="red")
```



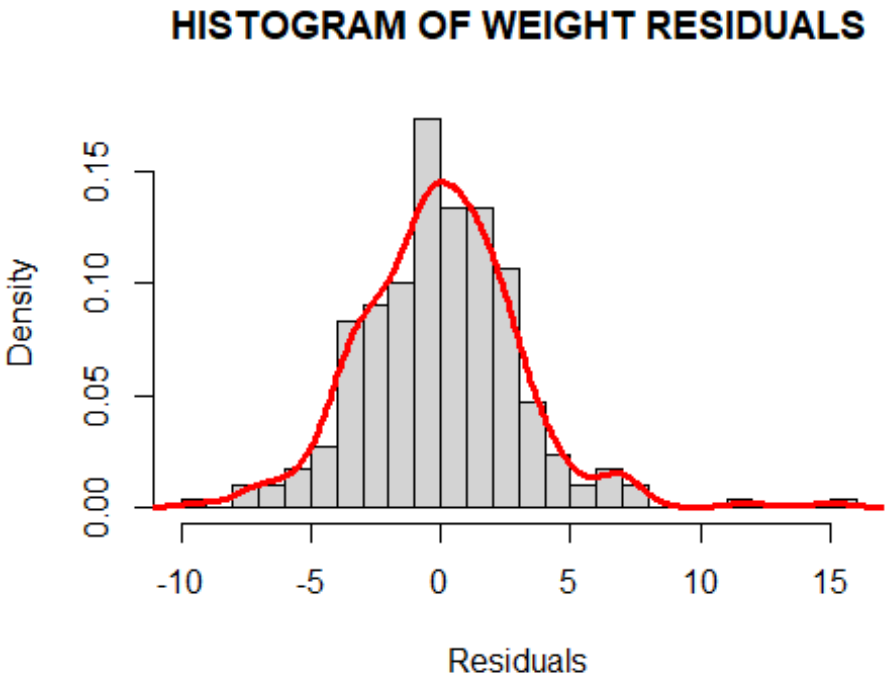
```r
#residuals vs. the predictor variable
residual <- weight_model$residuals
plot(Auto_mpg_data_train$mpg~residual,lwd=3, col="blue",main="mpg vs residual", xlab="res
idual",ylab = "mpg")
grid(NA, 5, lwd = 2,col = "darkgray")
```



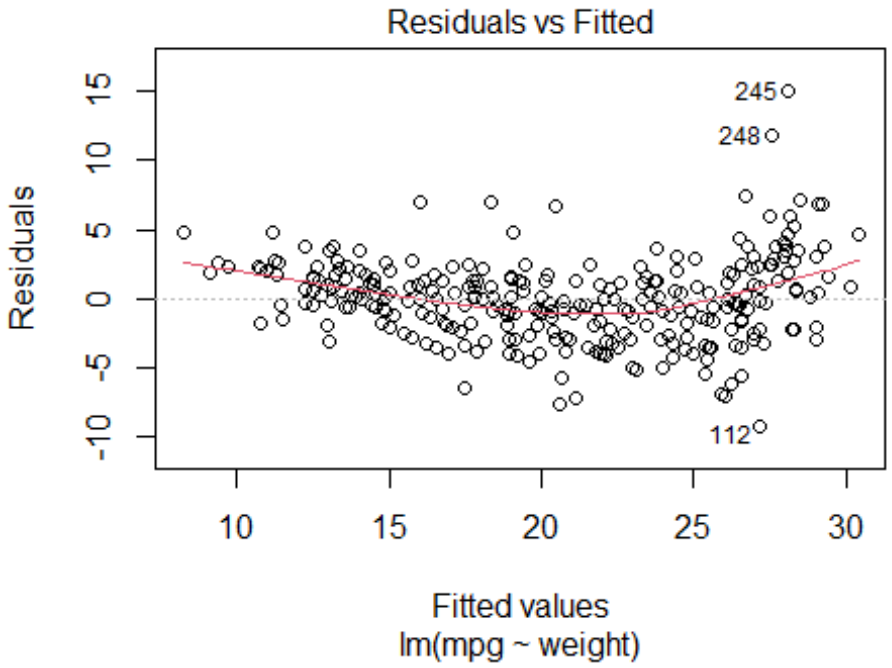

```r
#absolute value of the residuals vs. the predictor variable
abs_residual <- abs(residual)
plot(Auto_mpg_data_train$mpg~abs_residual,lwd=3, col="red",main="mpg vs Abs_residual", xl
ab="abs_residual",ylab = "mpg")
grid(NA, 5, lwd = 2,col = "darkgray")
```

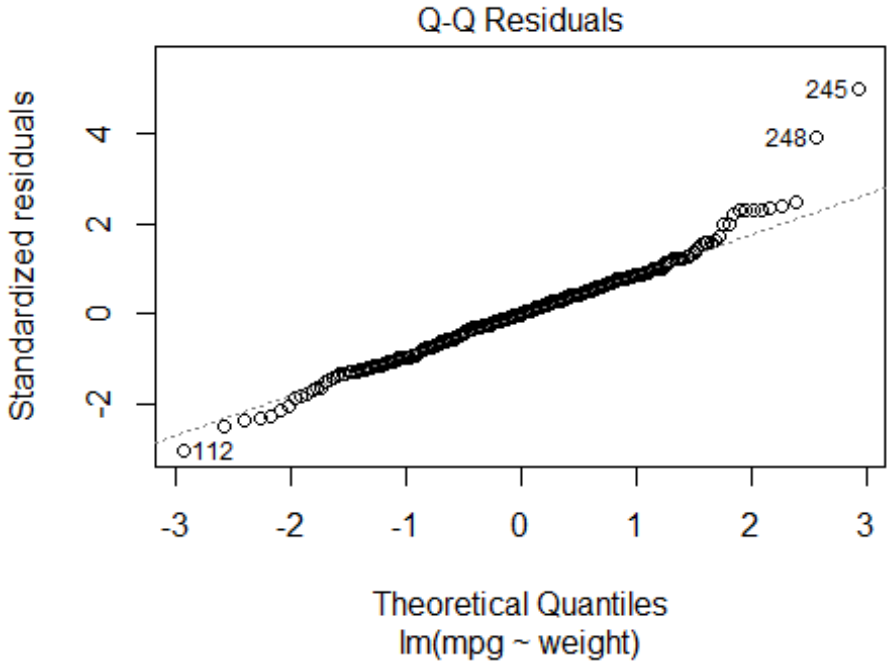


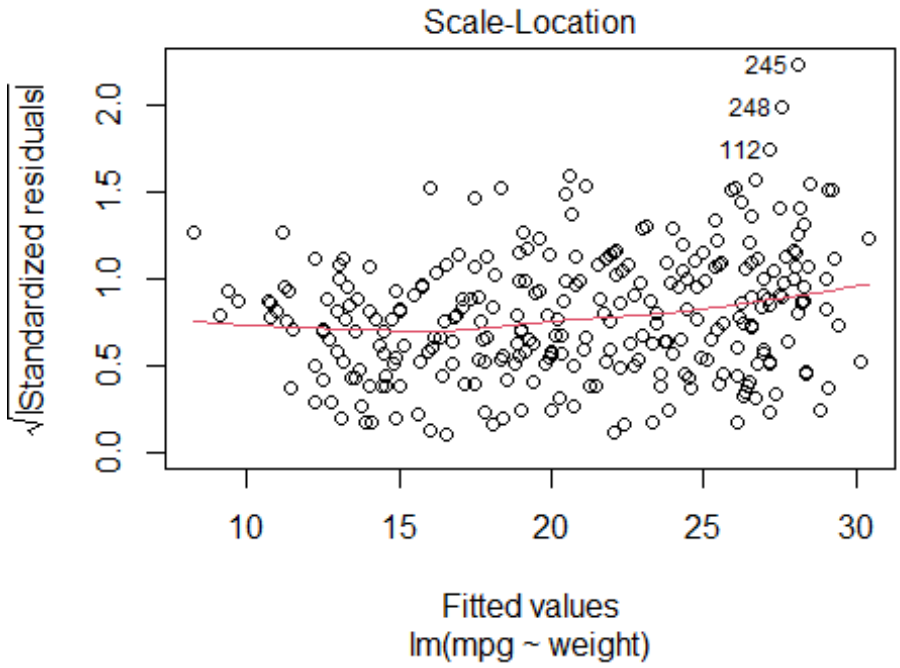

```r
#histogram of the residuals
hist(residual,prob=T,breaks=20,main="HISTOGRAM OF WEIGHT RESIDUALS",xlab="Residuals")
lines(density(residual),col="red",lwd=3)
```

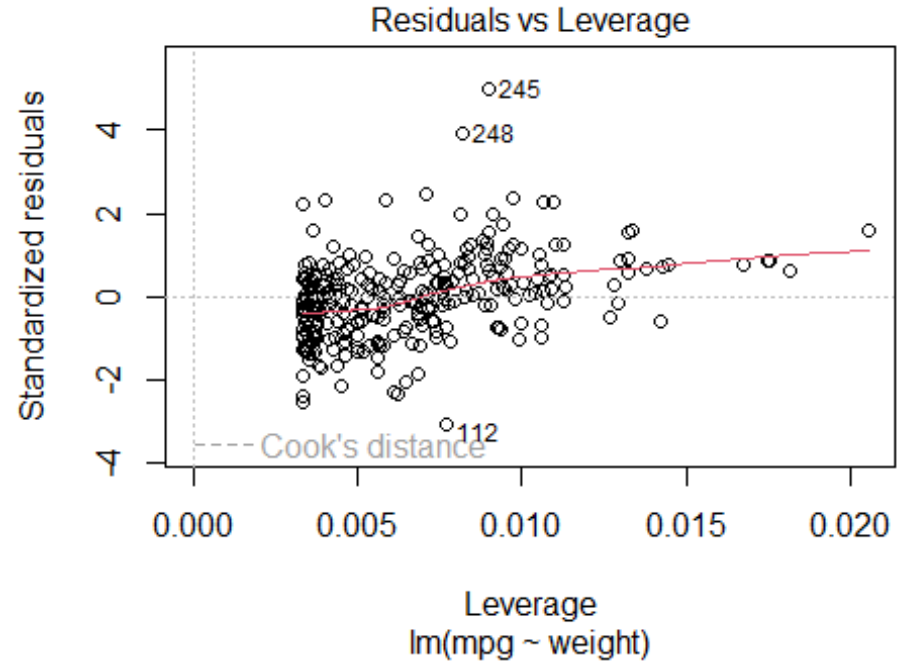



```r
plot(weight_model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ weight)
245
248
112
-10
-5
0
5
10
15
10
15
20
25
30


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(mpg ~ weight)
245
248
112
-2
0
2
4
-3
-2
-1
0
1
2
3






```r
# Make predictions and compute the R2, RMSE and MAE
weight_predict <- weight_model %>% predict(Auto_mpg_data_test)
data.frame( R2 = R2(weight_predict, Auto_mpg_data_test$mpg),
            RMSE = RMSE(weight_predict, Auto_mpg_data_test$mpg),
            MAE = MAE(weight_predict, Auto_mpg_data_test$mpg))

##          R2     RMSE      MAE
## 1 0.5006516 8.157758 6.983514
```

```r
prediction_error = RMSE(weight_predict, Auto_mpg_data_test$mpg)/mean(Auto_mpg_data_test$mpg)
prediction_error
```

```
## [1] 0.2553637
```

```r
compare_wght = as.data.frame(cbind(Auto_mpg_data_test$mpg,weight_predict),row=FALSE)
names(compare_wght) = c("observed","weight_predict")
head(compare_wght)
```

```
##   observed weight_predict
## 1     34.5       27.03751
## 2     31.8       27.85526
## 3     37.3       27.16331
## 4     28.4       23.76647
## 5     28.8       24.23825
## 6     26.8       23.57776
```

All the estimated values in this model are statistically significant, as evidenced by p-values less than 2e-16.

However, the plot of MPG vs. weight indicates a non-linear relationship, suggesting a nuanced association

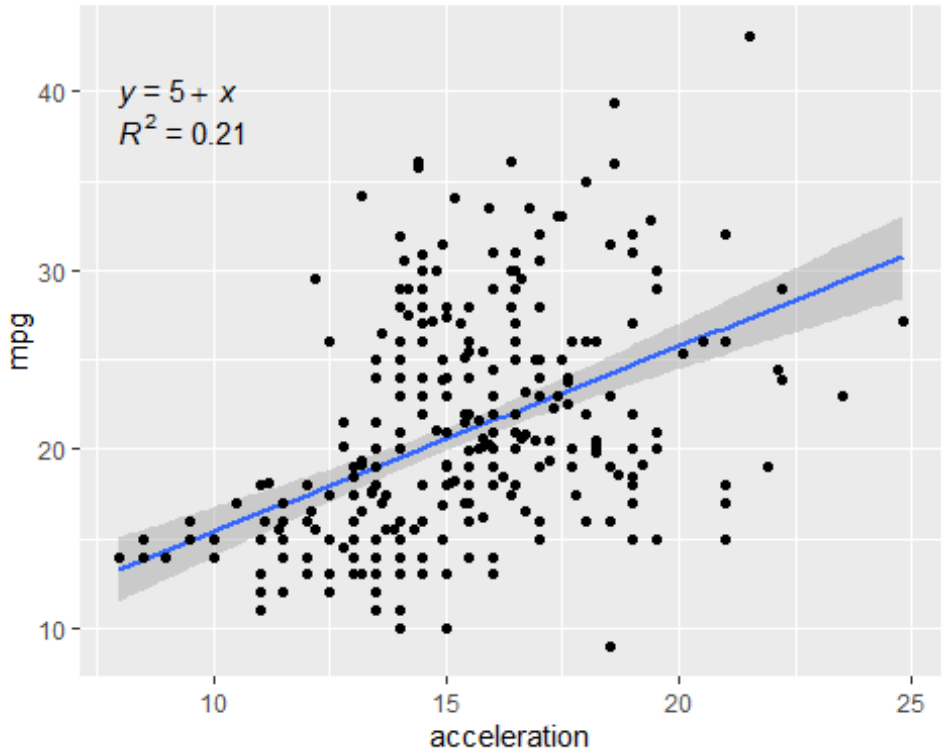
between the variable and the residuals. This model is deemed less than ideal. Diagnostic plots identify

outliers at data points 245, 248, and 112. The adjusted R-square indicates that approximately 77.06% of the

variance in MPG can be explained by the weight variable.

```r
#acceleration as explanatory variable

# Linear regression plot
ggplot(data=Auto_mpg_data_train, aes(x=acceleration, y=mpg)) +
  geom_smooth(method="lm") +
  geom_point() +
  stat_regline_equation(label.x=8, label.y=40) +
  stat_cor(aes(label=..rr.label..), label.x=8, label.y=38)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



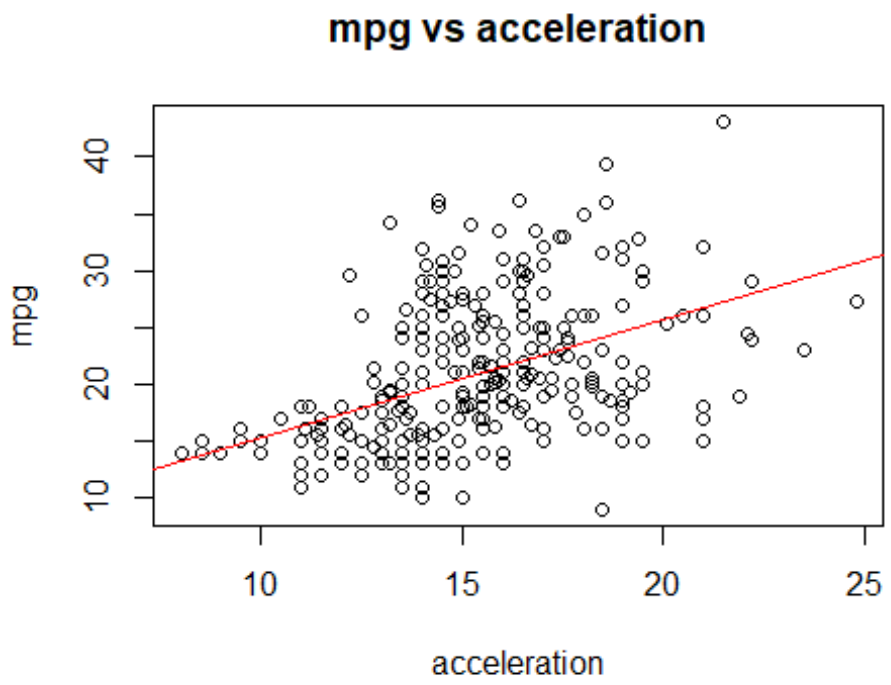

```r
#performing regression
acc_model <- lm(mpg~acceleration, data=Auto_mpg_data_train)
summary(acc_model)
```

```
##
## Call:
## lm(formula = mpg ~ acceleration, data = Auto_mpg_data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.202  -4.126  -1.012   3.268  16.154
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.0012     1.8352   2.725  0.00681 **
## acceleration   1.0379     0.1183   8.770  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.654 on 298 degrees of freedom
## Multiple R-squared:  0.2052, Adjusted R-squared:  0.2025
## F-statistic: 76.91 on 1 and 298 AUTO_MPG_DATA,  p-value: < 2.2e-16
```

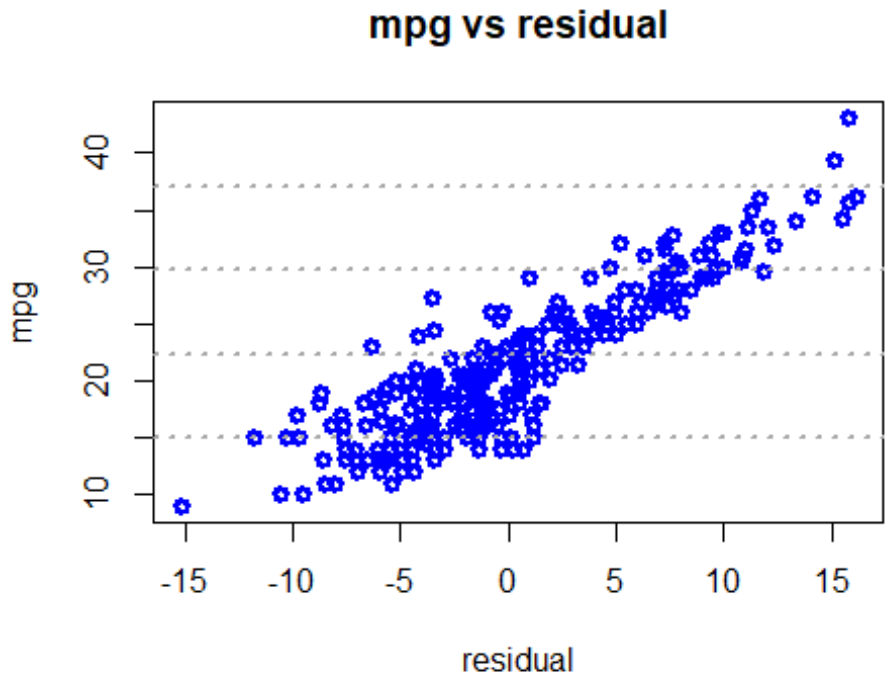
```r
# mean of residuals
mean(resid(acc_model))
```

```
## [1] -5.321611e-16
```

```r
#plot the variable
plot(Auto_mpg_data_train$mpg~Auto_mpg_data_train$acceleration,main="mpg vs acceleration",
xlab="acceleration",ylab = "mpg")
abline(acc_model,col="red")
```

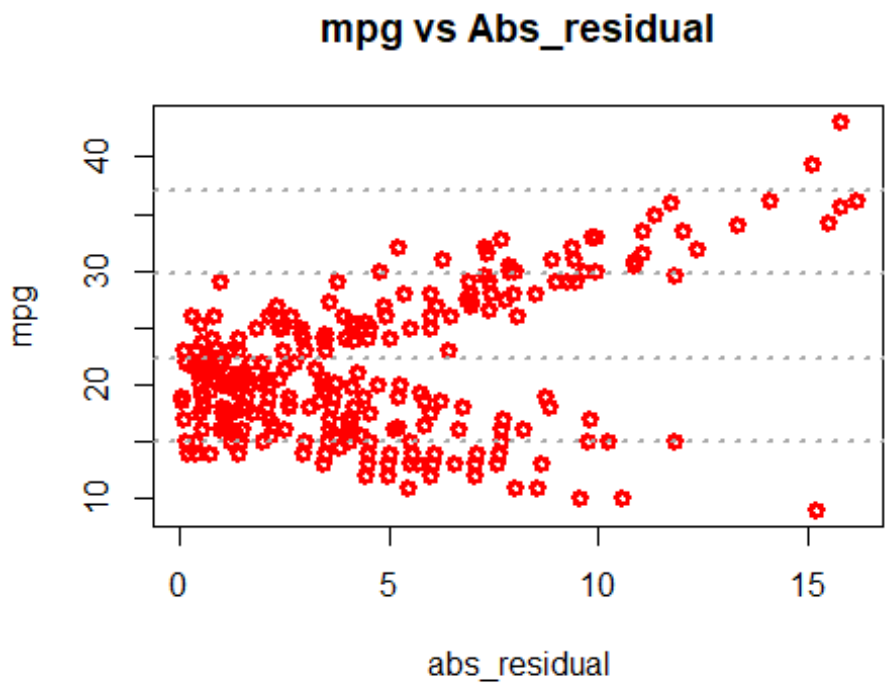



```
#residuals vs. the predictor variable
residual <- acc_model$residuals
plot(Auto_mpg_data_train$mpg~residual,lwd=3, col="blue",main="mpg vs residual", xlab="res
idual",ylab = "mpg")
grid(NA, 5, lwd = 2,col = "darkgray")
```

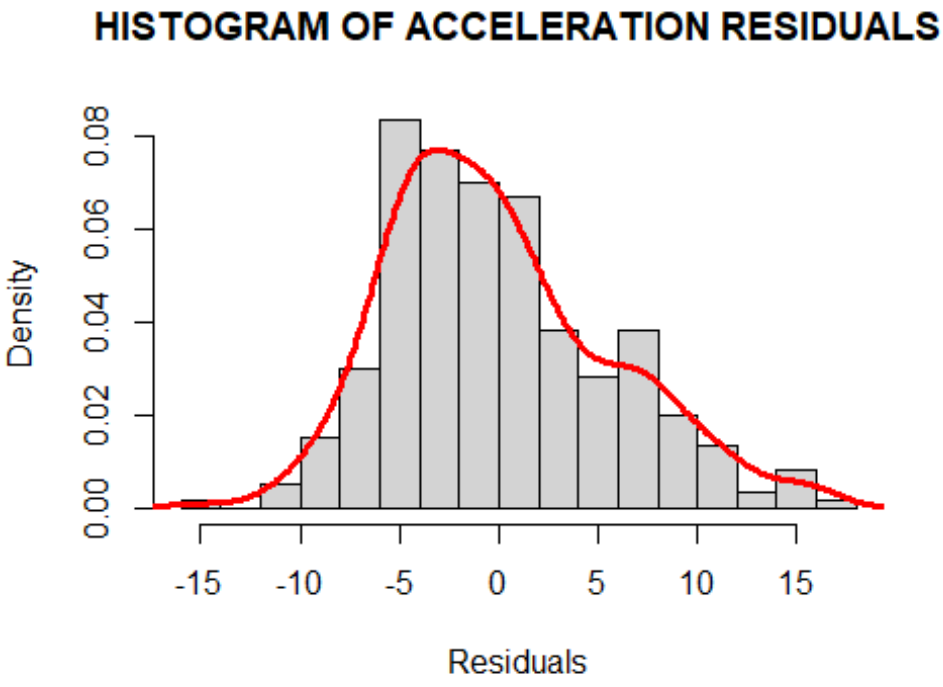



```
#absolute value of the residuals vs. the predictor variable
abs_residual <- abs(residual)
plot(Auto_mpg_data_train$mpg~abs_residual,lwd=3, col="red",main="mpg vs Abs_residual", xl
```

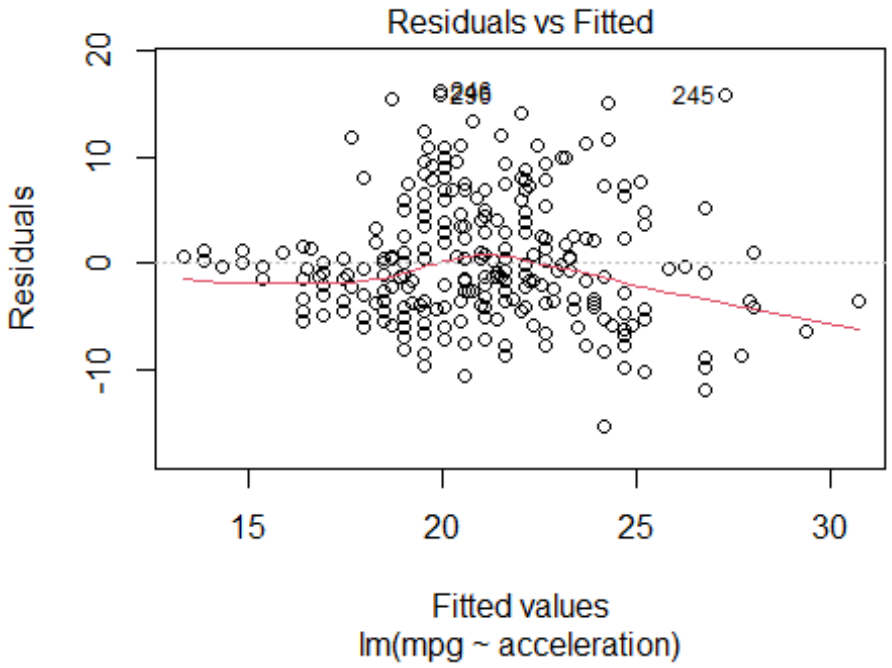
```
ab="abs_residual",ylab = "mpg")
grid(NA, 5, lwd = 2,col = "darkgray")
```

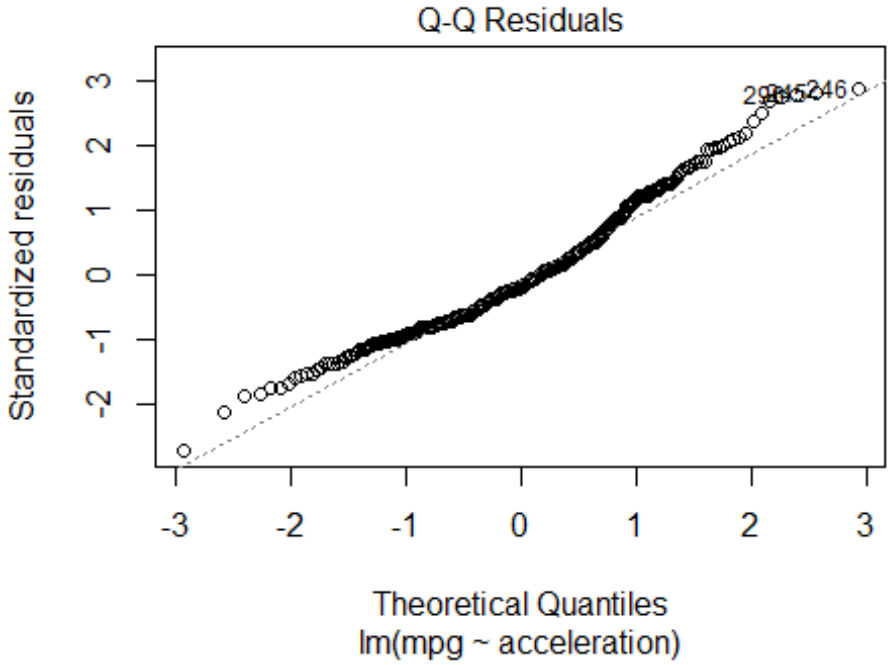


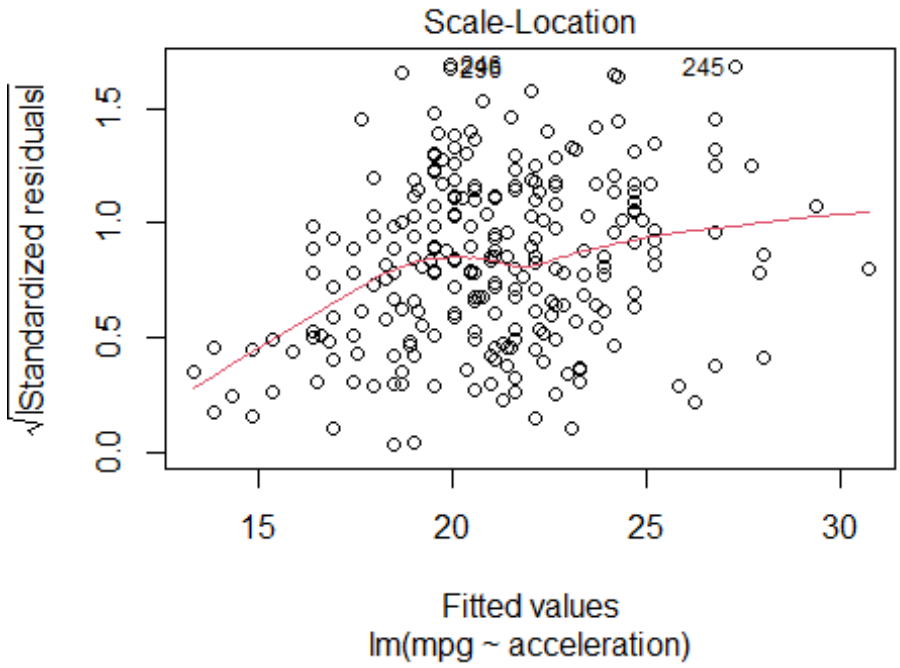

```
#histogram of the residuals
hist(residual,prob=T,breaks=20,main="HISTOGRAM OF ACCELERATION RESIDUALS",xlab="Residuals
")
lines(density(residual),col="red",lwd=3)
```

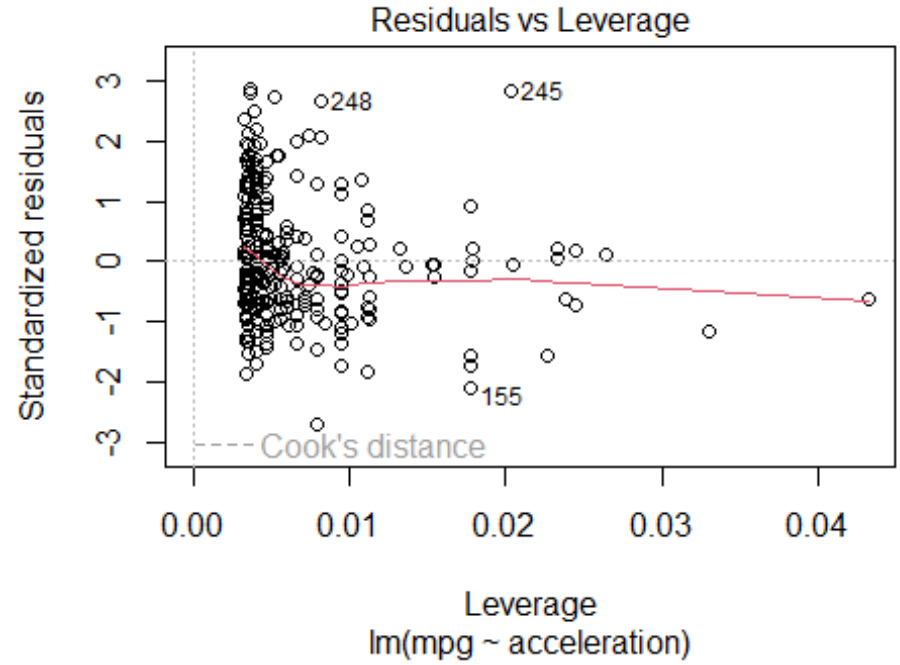



```
plot(acc_model)
```


Residuals vs Fitted
Residuals
20
10
0
-10
245
15
20
25
30
Fitted values
lm(mpg ~ acceleration)


Q-Q Residuals
Standardized residuals
3
2
1
0
-1
-2
246
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(mpg ~ acceleration)






```
# Make predictions and compute the R2, RMSE and MAE
acc_predict <- acc_model %>% predict(Auto_mpg_data_test)
data.frame( R2 = R2(acc_predict, Auto_mpg_data_test$mpg),
            RMSE = RMSE(acc_predict, Auto_mpg_data_test$mpg),
            MAE = MAE(acc_predict, Auto_mpg_data_test$mpg))

##           R2     RMSE      MAE
## 1 0.03597167 11.51665 10.16914
```

```r
prediction_error = RMSE(acc_predict, Auto_mpg_data_test$mpg)/mean(Auto_mpg_data_test$mpg)
prediction_error
```

```
## [1] 0.3605077
```

```r
compare_acc = as.data.frame(cbind(Auto_mpg_data_test$mpg,acc_predict),row=FALSE)
names(compare_acc) = c("observed","acc_predict")
head(compare_acc)
```

```
##   observed acc_predict
## 1     34.5    20.46536
## 2     31.8    24.92818
## 3     37.3    20.25778
## 4     28.4    21.60701
## 5     28.8    16.72904
## 6     26.8    18.38963
```

```r
#Multiple regression


# To find out which independent variable to use in our multiple regression we are going t
o use the step wise regression
null=lm(mpg~1,data=Auto_mpg_data_train)
full=lm(mpg~.,data=Auto_mpg_data_train)
step(null,scope=list(upper=full),data=Auto_mpg_data_train,direction="both")
```

```
## Start:  AIC=1108.25
## mpg ~ 1
##
##                Auto_mpg_data Sum of Sq     RSS     AIC
## + weight         1    9243.7  2739.6  667.54
## + displacement   1    8513.2  3470.1  738.45
## + cylinder       1    7951.8  4031.5  783.43
## + horsepower     1    7680.7  4302.6  802.96
## + acceleration   1    2458.4  9524.9 1041.36
## <none>                       11983.3 1108.25
##
## Step:  AIC=667.54
## mpg ~ weight
##
##                Auto_mpg_data Sum of Sq     RSS     AIC
## + horsepower     1      98.2  2641.5  658.59
## + displacement   1      59.9  2679.7  662.91
## + acceleration   1      41.1  2698.6  665.01
## + cylinder       1      34.9  2704.8  665.70
## <none>                        2739.6  667.54
## - weight         1    9243.7 11983.3 1108.25
##
## Step:  AIC=658.59
## mpg ~ weight + horsepower
##
##                Auto_mpg_data Sum of Sq    RSS    AIC
## <none>                       2641.5 658.59
## + cylinder       1     10.65 2630.8 659.38
## + displacement   1      9.81 2631.7 659.48
## + acceleration   1      1.00 2640.5 660.48
```

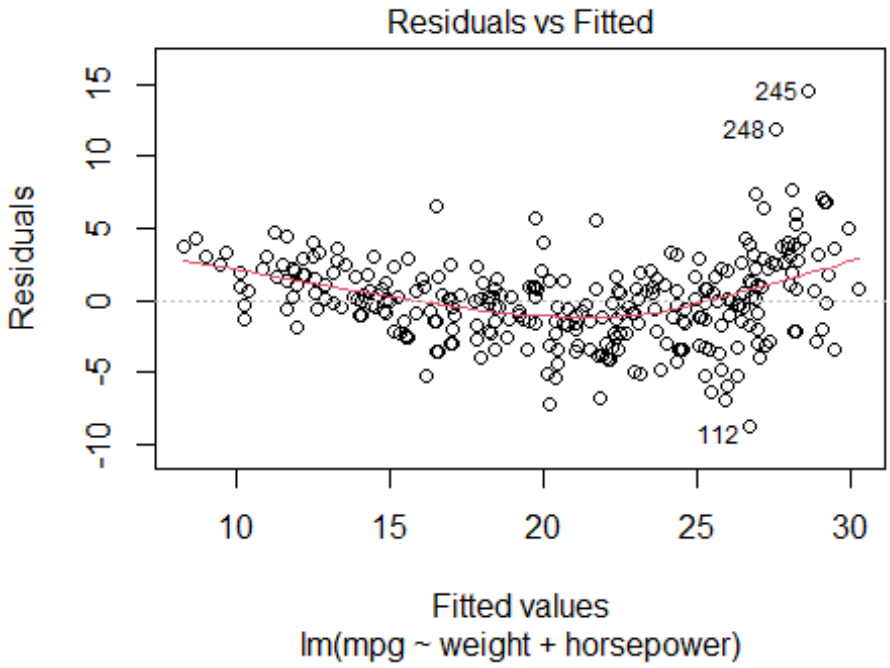
```
## - horsepower    1     98.16 2739.6 667.54
## - weight        1   1661.08 4302.6 802.96

##
## Call:
## lm(formula = mpg ~ weight + horsepower, data = Auto_mpg_data_train)
##
## Coefficients:
## (Intercept)       weight   horsepower
##   40.258743    -0.005204    -0.027759
```

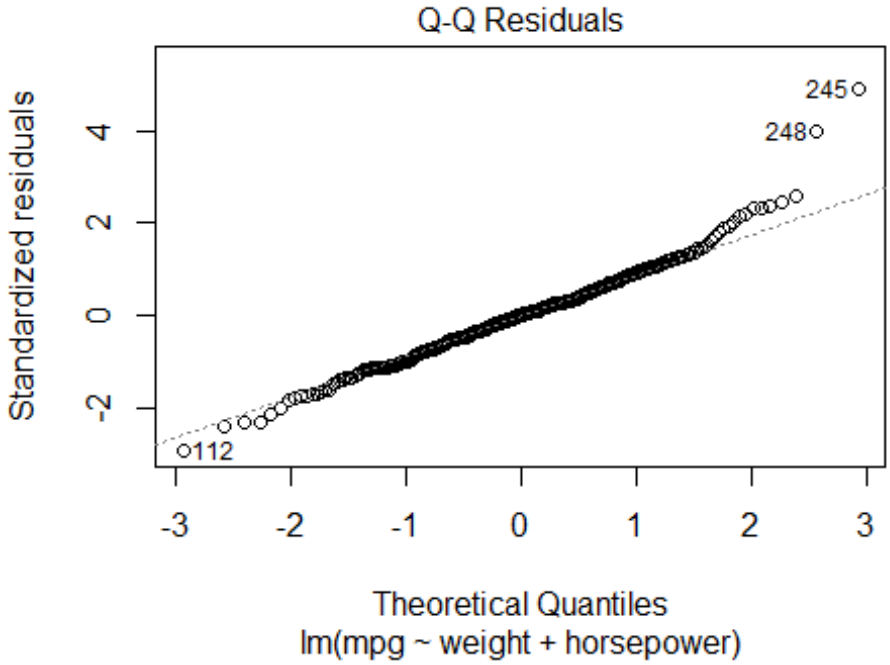
The step function determined that weight and horsepower were the best variables to use in order to create the necessary linear regression model, as shown by the R findings above. The step function chooses a set of variables that produce the lowest AIC statistic by using the Akaike Information Criterion (AIC) as a criterion. As a result, we decide to make weight and horsepower the model's final variables.

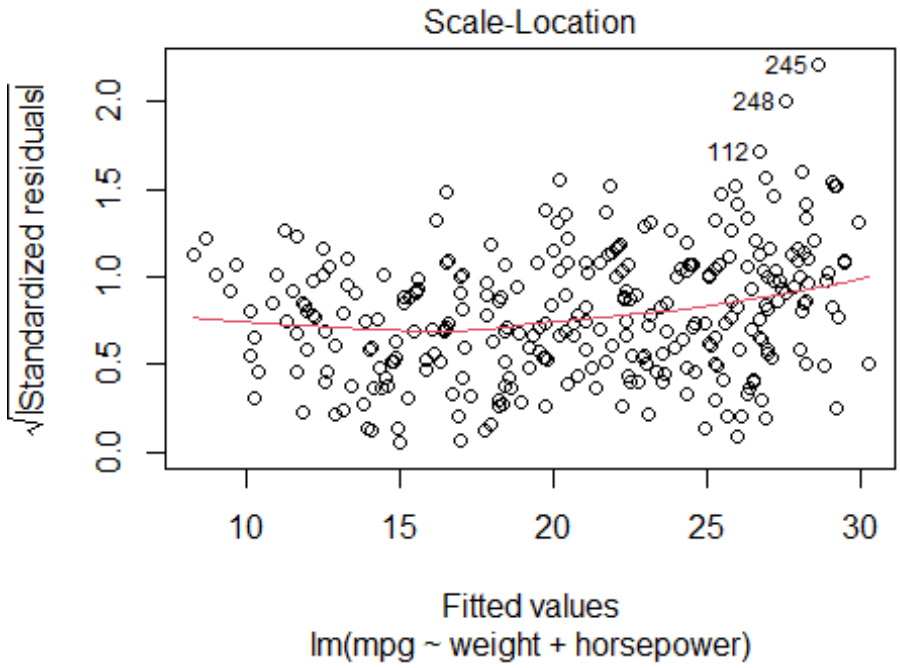
```
final_model <- lm(mpg ~ weight + horsepower, data = Auto_mpg_data_train)
summary(final_model)
```

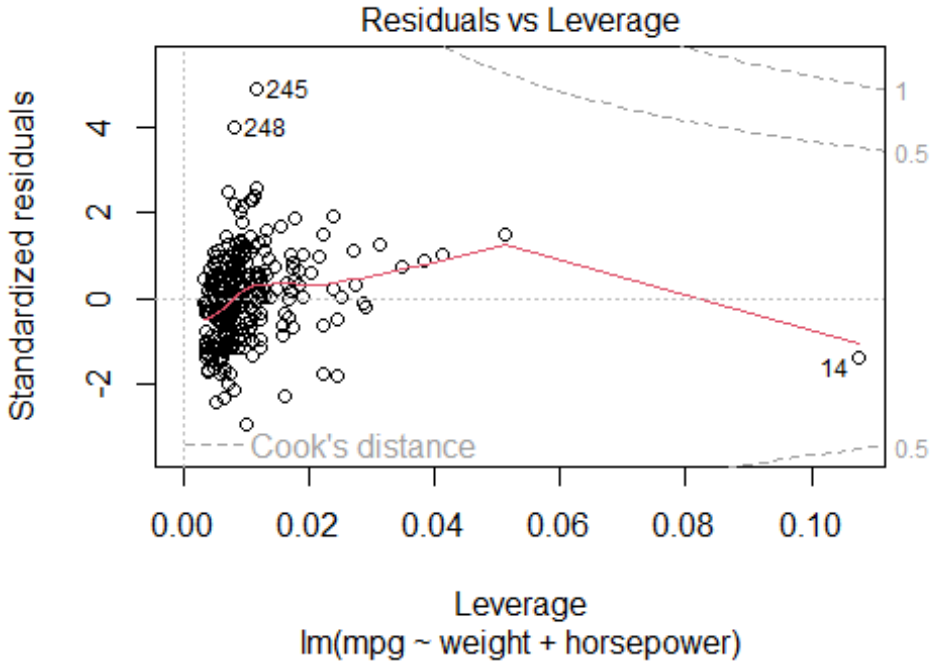
```
##
## Call:
## lm(formula = mpg ~ weight + horsepower, data = Auto_mpg_data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7069 -1.8380  0.0207  1.6877 14.5038
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.2587429  0.6420610  62.702  < 2e-16 ***
## weight      -0.0052041  0.0003808 -13.666  < 2e-16 ***
## horsepower  -0.0277594  0.0083560  -3.322  0.00101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.982 on 297 degrees of freedom
## Multiple R-squared:  0.7796, Adjusted R-squared:  0.7781
## F-statistic: 525.2 on 2 and 297 AUTO_MPG_DATA,  p-value: < 2.2e-16
```

```
plot(final_model)
```

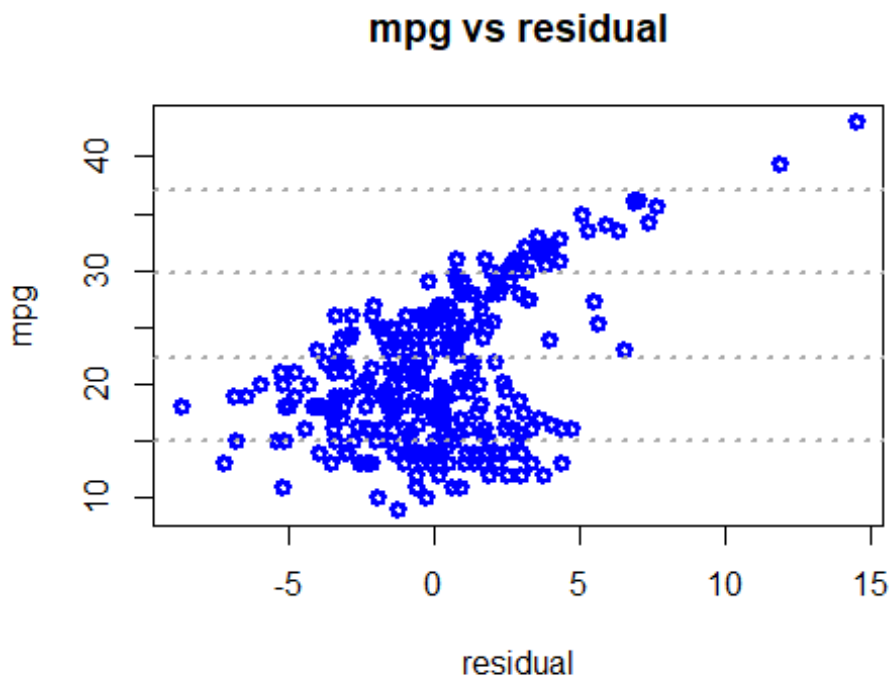

Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ weight + horsepower)
245
248
112
-10
-5
0
5
10
15
10
15
20
25
30


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(mpg ~ weight + horsepower)
245
248
112
-2
0
2
4
-3
-2
-1
0
1
2
3

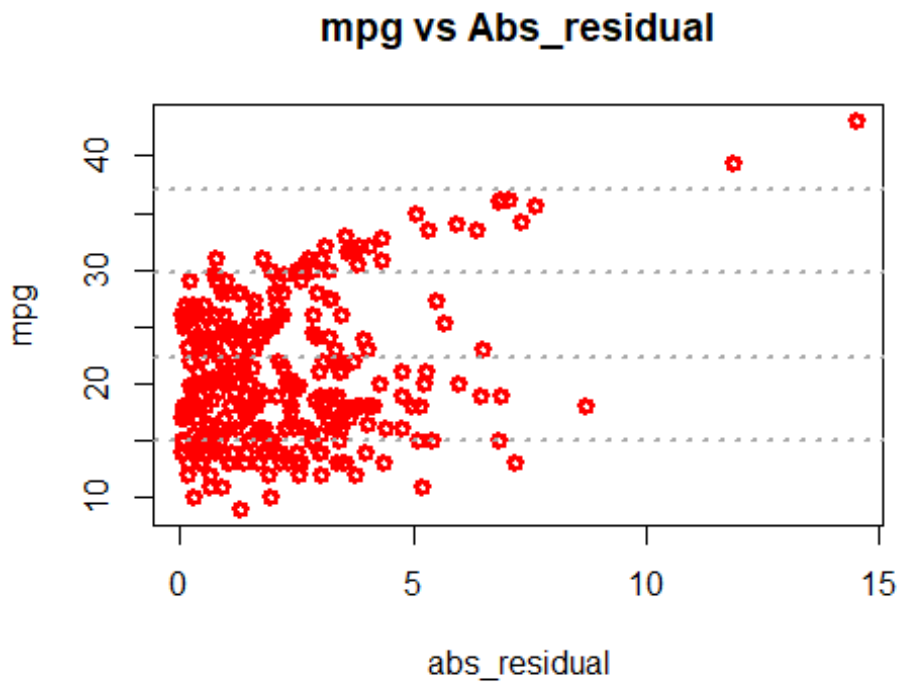





```r
#residuals vs. the predictor variable
residual <- final_model$residuals
plot(Auto_mpg_data_train$mpg~residual,lwd=3, col="blue",main="mpg vs residual", xlab="res
idual",ylab = "mpg")
grid(NA, 5, lwd = 2,col = "darkgray")
```

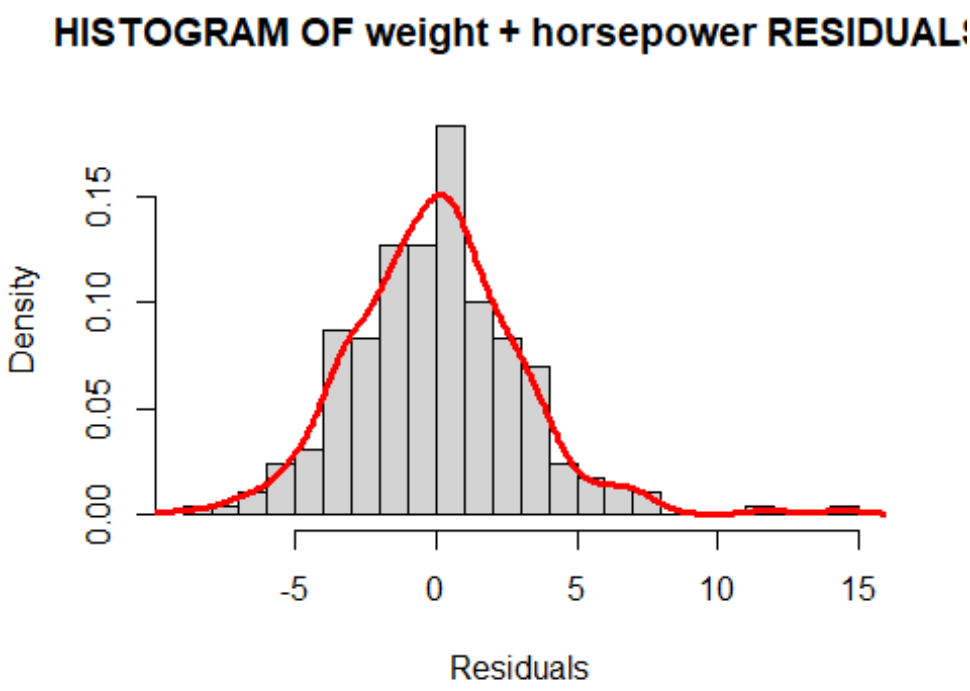



```r
#absolute value of the residuals vs. the predictor variable
abs_residual <- abs(residual)
plot(Auto_mpg_data_train$mpg~abs_residual,lwd=3, col="red",main="mpg vs Abs_residual", xl
ab="abs_residual",ylab = "mpg")
grid(NA, 5, lwd = 2,col = "darkgray")
```




```r
#histogram of the residuals
hist(residual,prob=T,breaks=20,main="HISTOGRAM OF weight + horsepower RESIDUALS",xlab="Re
```

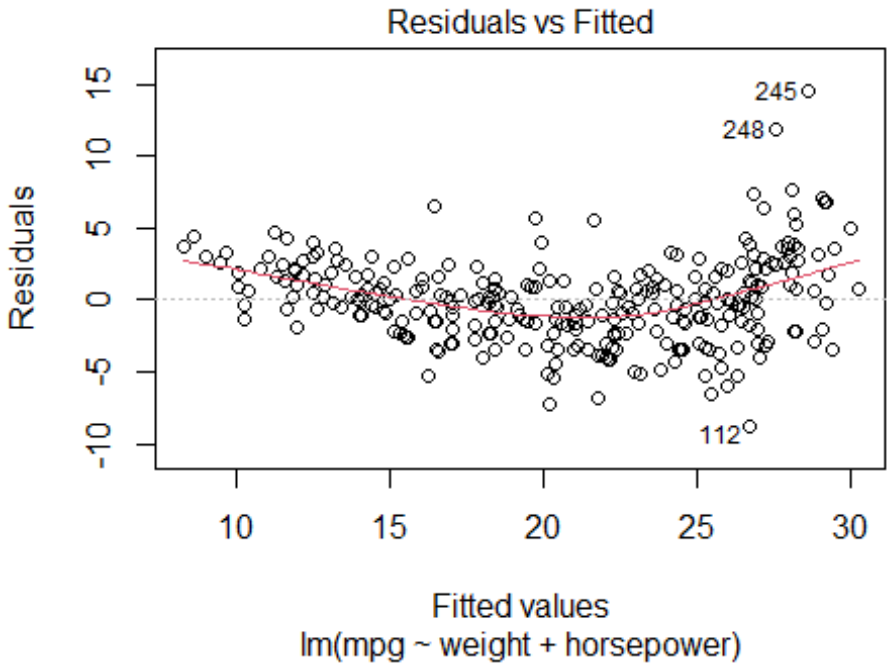
```
siduals")
lines(density(residual),col="red",lwd=3)
```

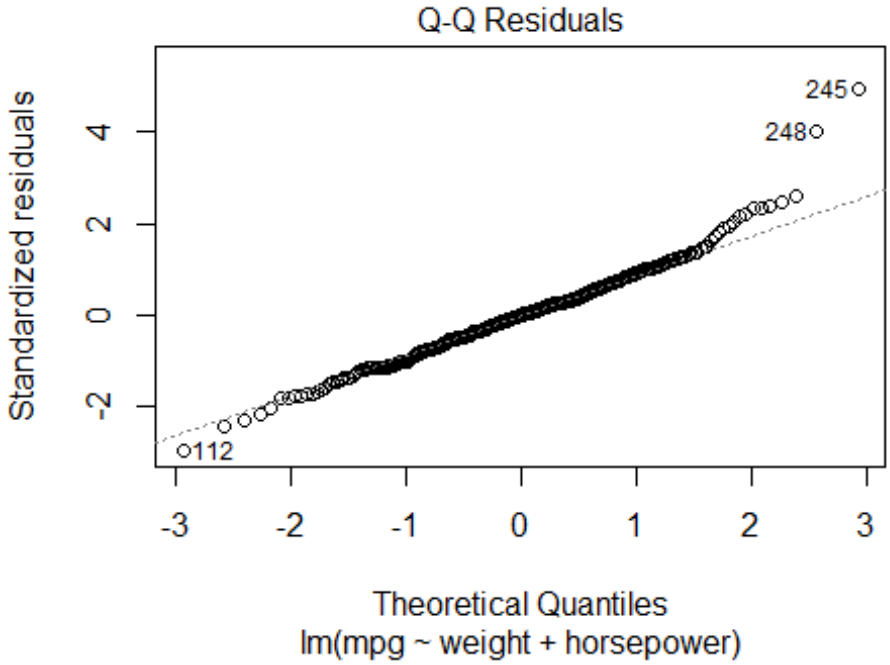


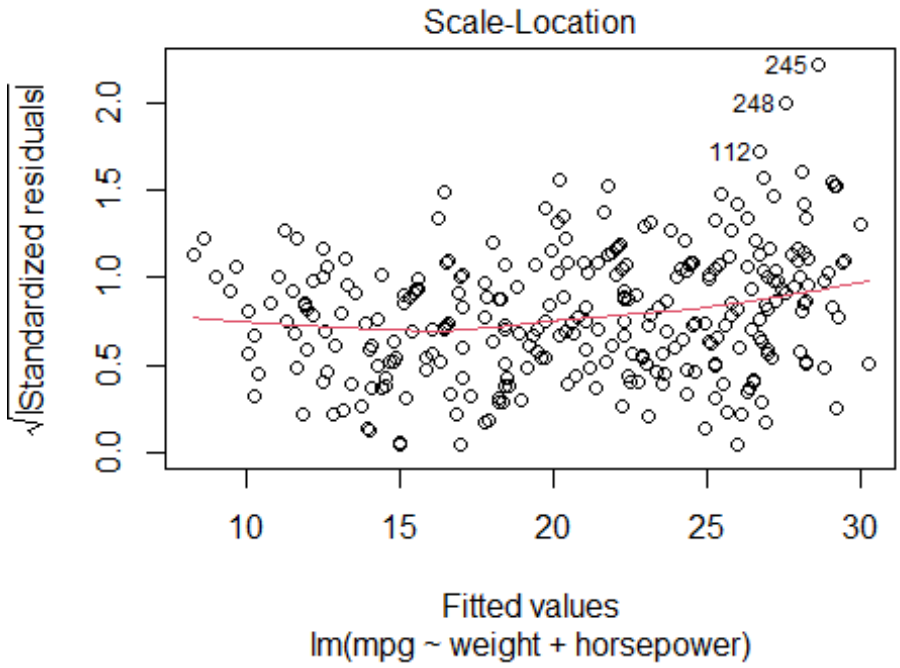

```
gvlma(final_model)

##
## Call:
## lm(formula = mpg ~ weight + horsepower, data = Auto_mpg_data_train)
##
## Coefficients:
## (Intercept)       weight   horsepower
##   40.258743    -0.005204    -0.027759
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = final_model)
##
##                      Value   p-value                   Decision
## Global Stat        150.774 0.000e+00 Assumptions NOT satisfied!
## Skewness            16.365 5.225e-05 Assumptions NOT satisfied!
## Kurtosis            59.193 1.432e-14 Assumptions NOT satisfied!
## Link Function       65.957 4.441e-16 Assumptions NOT satisfied!
## Heteroscedasticity   9.259 2.343e-03 Assumptions NOT satisfied!
```

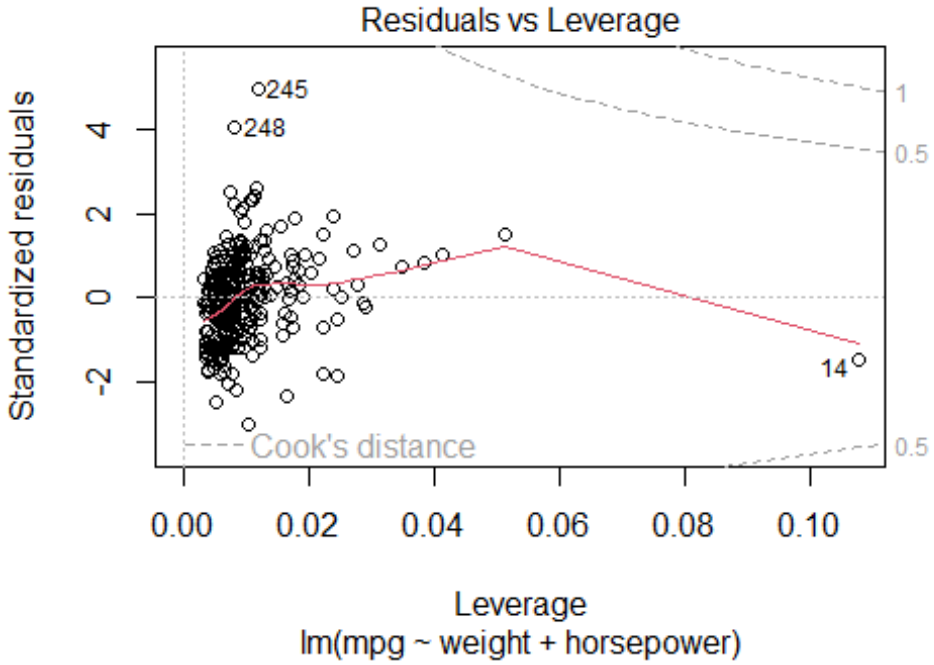
It's probable that data outliers are affecting the model's quality because none of the assumptions are met. In order to remedy this, we can continue by eliminating the anomalies from the data and reevaluating the model to look for any possible enhancements.

```r
final_model2 <- lm(mpg ~ weight + horsepower, data = Auto_mpg_data_train[-c(112,245,248),
])
summary(final_model2)

##
## Call:
## lm(formula = mpg ~ weight + horsepower, data = Auto_mpg_data_train[-c(112,
##     245, 248), ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7327 -1.7911  0.0047  1.6783 14.5074
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.2859314  0.6408855  62.860  < 2e-16 ***
## weight      -0.0052394  0.0003785 -13.844  < 2e-16 ***
## horsepower  -0.0269430  0.0083047  -3.244  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.96 on 294 degrees of freedom
## Multiple R-squared:  0.7824, Adjusted R-squared:  0.7809
## F-statistic: 528.6 on 2 and 294 AUTO_MPG_DATA,  p-value: < 2.2e-16

plot(final_model2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ weight + horsepower)
245
248
112


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(mpg ~ weight + horsepower)
245
248
112





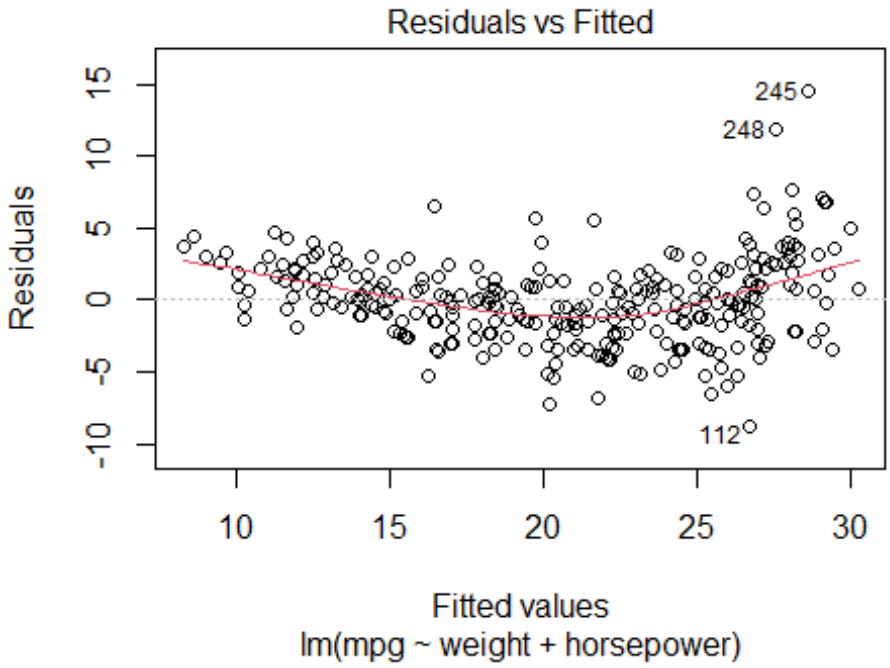

```
# Global Validation of Linear Models Assumptions
gvlma(final_model2)

##
## Call:
## lm(formula = mpg ~ weight + horsepower, data = Auto_mpg_data_train[-c(112,
##     245, 248), ])
##
## Coefficients:
## (Intercept)       weight   horsepower
```

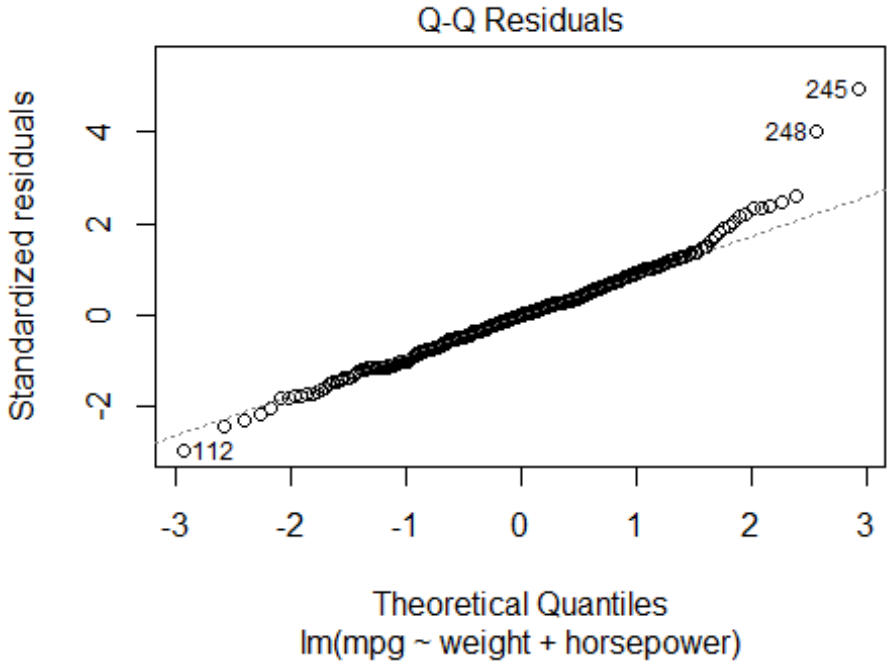
```
##   40.285931    -0.005239    -0.026943
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = final_model2)
##
##                     Value   p-value                   Decision
## Global Stat        158.37 0.000e+00 Assumptions NOT satisfied!
## Skewness            18.69 1.537e-05 Assumptions NOT satisfied!
## Kurtosis            64.02 1.221e-15 Assumptions NOT satisfied!
## Link Function       65.56 5.551e-16 Assumptions NOT satisfied!
## Heteroscedasticity  10.09 1.490e-03 Assumptions NOT satisfied!
```

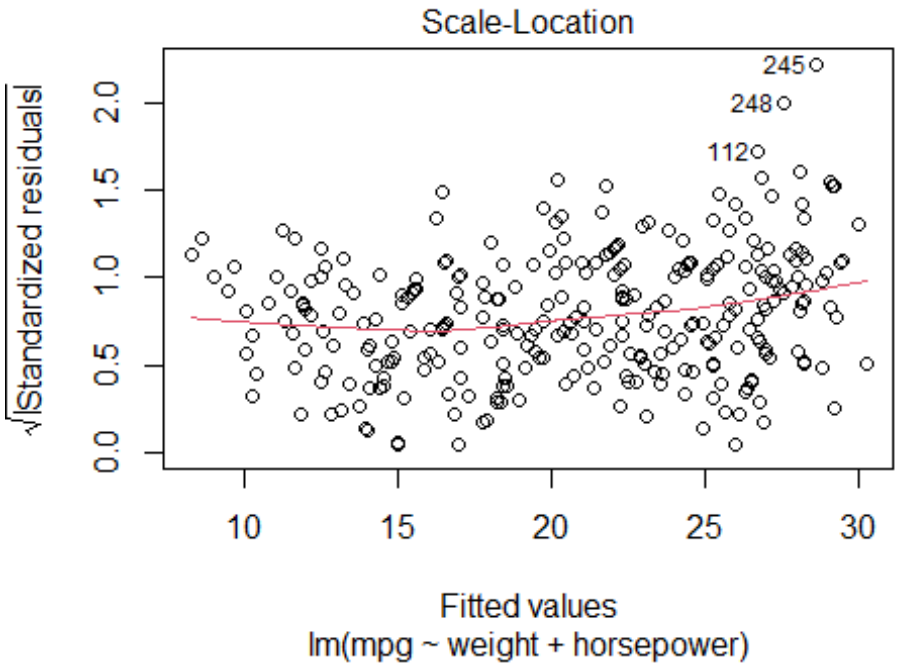
```
final_model2 <- lm(mpg ~ weight + horsepower, data = Auto_mpg_data_train[-c(112,245,248),
])
summary(final_model2)
```

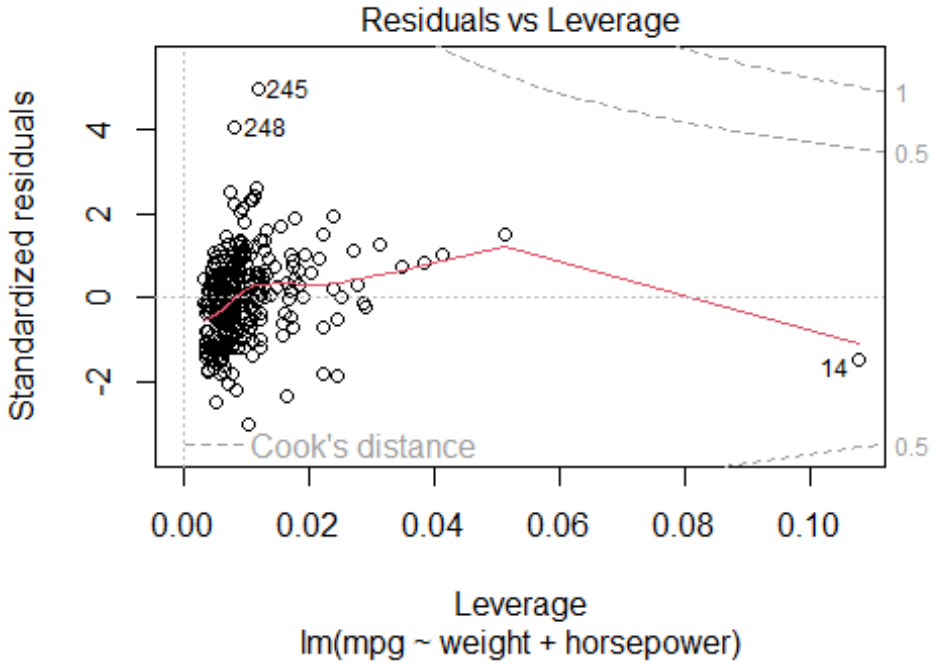
```
##
## Call:
## lm(formula = mpg ~ weight + horsepower, data = Auto_mpg_data_train[-c(112,
##     245, 248), ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7327 -1.7911  0.0047  1.6783 14.5074
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.2859314  0.6408855  62.860  < 2e-16 ***
## weight      -0.0052394  0.0003785 -13.844  < 2e-16 ***
## horsepower  -0.0269430  0.0083047  -3.244  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.96 on 294 degrees of freedom
## Multiple R-squared:  0.7824, Adjusted R-squared:  0.7809
## F-statistic: 528.6 on 2 and 294 AUTO_MPG_DATA,  p-value: < 2.2e-16
```

```
plot(final_model2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ weight + horsepower)
245
248
112


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(mpg ~ weight + horsepower)
245
248
112






```
# Make predictions and compute the R2, RMSE and MAE
predict_final <- final_model2 %>% predict(Auto_mpg_data_test)
data.frame( R2 = R2(predict_final, Auto_mpg_data_test$mpg),
            RMSE = RMSE(predict_final, Auto_mpg_data_test$mpg),
            MAE = MAE(predict_final, Auto_mpg_data_test$mpg))

##           R2    RMSE      MAE
## 1 0.5427488 8.0093 6.881424
```

```r
predictions_error <- RMSE(predict_final, Auto_mpg_data_test$mpg)/mean(Auto_mpg_data_test$
mpg)
predictions_error
```

```
## [1] 0.2507164
```

```r
compare_final <- as.data.frame(cbind(Auto_mpg_data_test$mpg,predict_final),row=FALSE)
names(compare_final) <- c("observed","predict_final")
head(compare_final)
```

```
##   observed predict_final
## 1     34.5      27.13531
## 2     31.8      27.95114
## 3     37.3      27.26704
## 4     28.4      23.87199
## 5     28.8      23.59136
## 6     26.8      23.04123
```

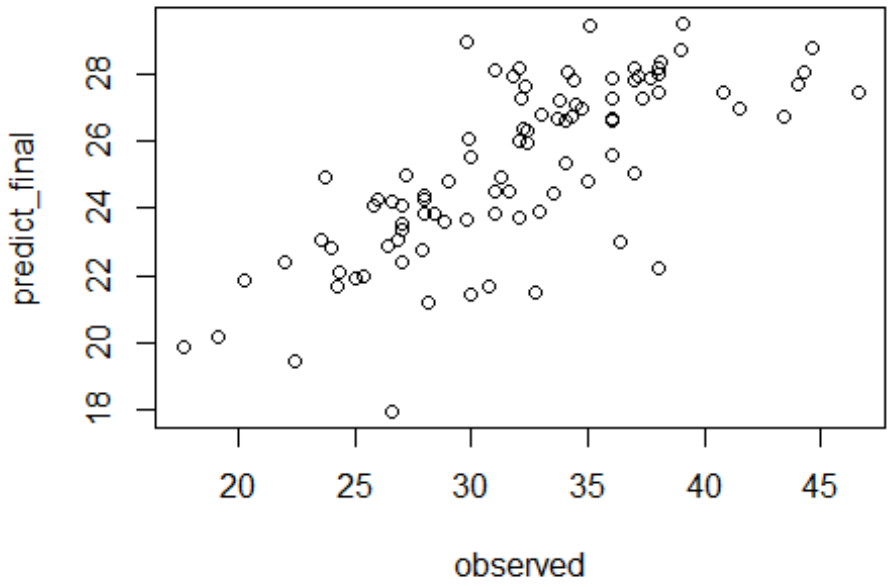
```r
cor(compare_final)
```

```
##                observed predict_final
## observed      1.0000000     0.7367148
## predict_final 0.7367148     1.0000000
```

```r
summary(compare_final)
```

```
##     observed     predict_final
##  Min.   :17.60   Min.   :17.94
##  1st Qu.:27.73   1st Qu.:23.49
##  Median :32.05   Median :25.21
##  Mean   :31.95   Mean   :25.21
##  3rd Qu.:36.00   3rd Qu.:27.46
##  Max.   :46.60   Max.   :29.53
```

```r
plot(compare_final)
```



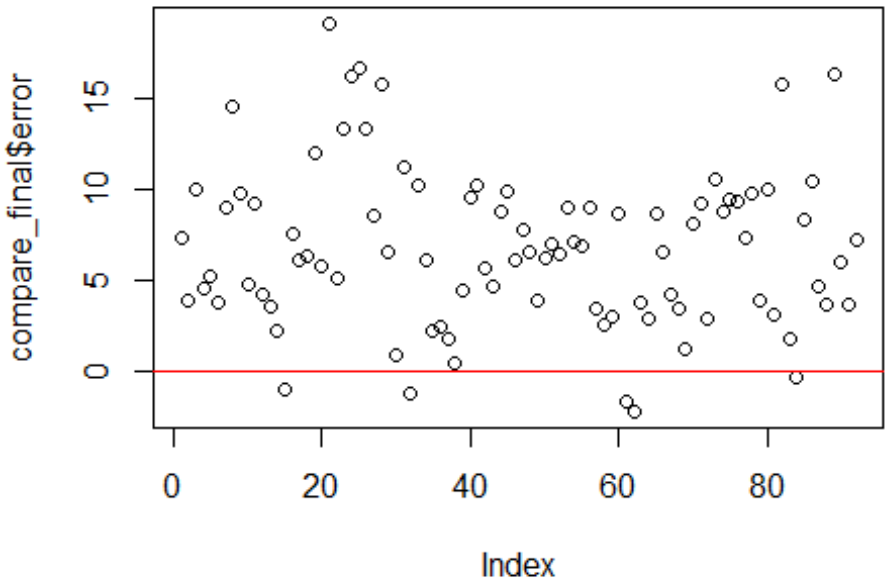

```
compare_final$error = compare_final$observed - compare_final$predict_final
head(compare_final)
```

```
##   observed predict_final     error
## 1     34.5      27.13531  7.364688
## 2     31.8      27.95114  3.848857
## 3     37.3      27.26704 10.032958
## 4     28.4      23.87199  4.528012
## 5     28.8      23.59136  5.208635
## 6     26.8      23.04123  3.758767
```

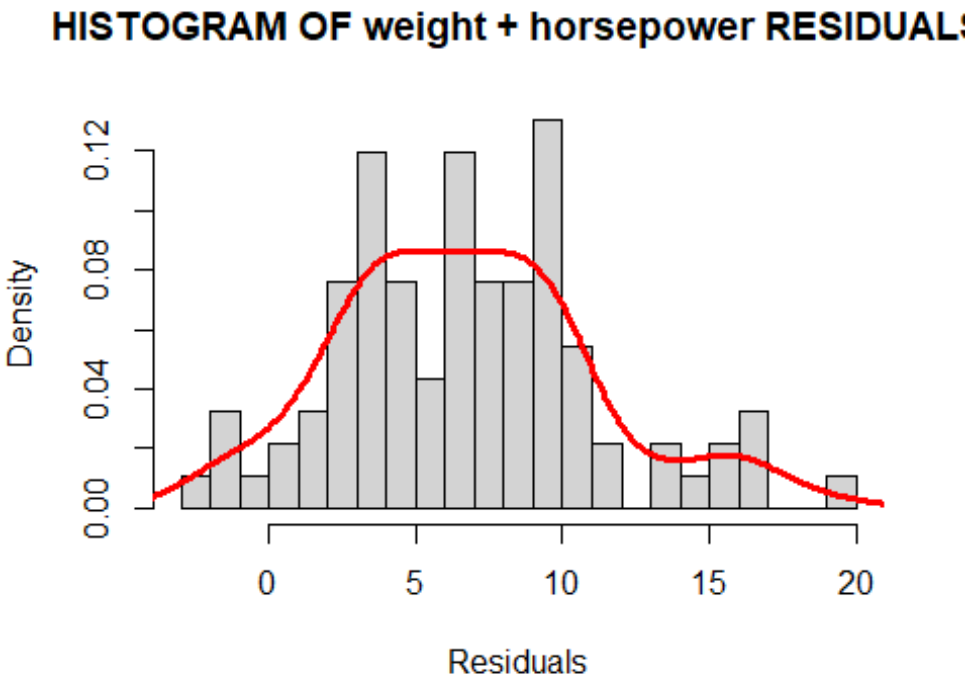
```
summary(compare_final)
```

```
##     observed     predict_final       error        
##  Min.   :17.60   Min.   :17.94   Min.   :-2.241  
##  1st Qu.:27.73   1st Qu.:23.49   1st Qu.: 3.712  
##  Median :32.05   Median :25.21   Median : 6.471  
##  Mean   :31.95   Mean   :25.21   Mean   : 6.738  
##  3rd Qu.:36.00   3rd Qu.:27.46   3rd Qu.: 9.255  
##  Max.   :46.60   Max.   :29.53   Max.   :19.120
```

```
# Residuals plot
plot(compare_final$error)
abline(h = 0, col= 'red')
```




```
#histogram of the residuals
hist(compare_final$error,prob=T,breaks=20,main="HISTOGRAM OF weight + horsepower RESIDUAL
S",xlab="Residuals")
lines(density(compare_final$error),col="red",lwd=3)
```



Therefore, it appears that most of the time, our regression model predicts more than the actual number.
Therefore, mpg = 40.2859314 - 0.0052394 * weight - 0.0269430 * horsepower is the final formula.