Exercise 1

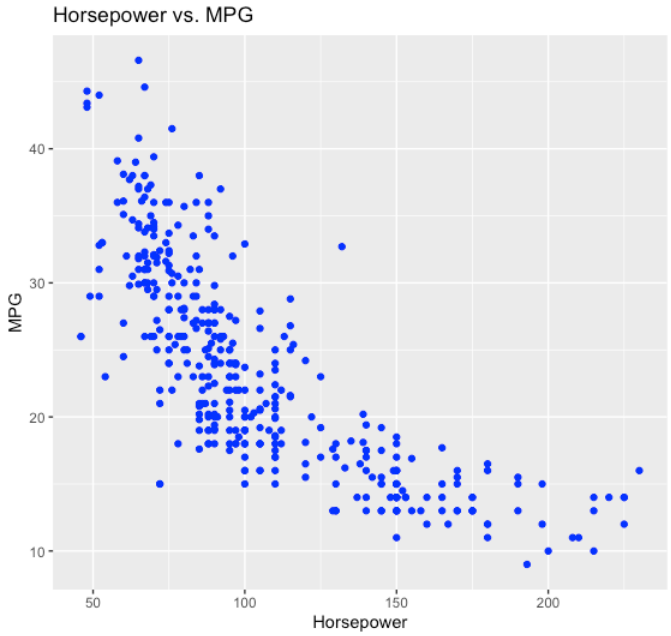
Name: Collin Real (yhi267)

Consider a well-known auto dataset (Auto.txt available on Canvas), which consists of 392 observations and two variables: mpg: miles per gallon and horsepower: Engine displacement (engine horsepower). As in those previous analyses in Chapter 2, we take **mpg** as the dependent/outcome variable and **horsepower** as the predictor variable.

a) Read data into R using the read.table function in R. For instance, I read the data from my computer using

```
Auto <- read.table("C://Users/DTY670/Desktop/STA6543/Exercise/Exercise1/Auto.txt",
header = T)  #You need to change the highlighted path.
```

b) Draw a scatterplot to check the relationship between horsepower (x-axis) and mpg (y-axis) and interpret the relationship between mgp and horsepower.



There is a negative correlation between horsepower and MPG: as horsepower goes up, MPG goes down. More horsepower = less fuel efficiency.

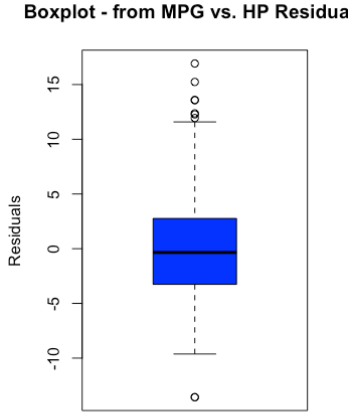
c) Write down the least square regression equation and circle the results from your outputs.

MPG = -0.1578*horsepower + 39.9359

```
Call:
lm(formula = mpg ~ horsepower, data = auto_df)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

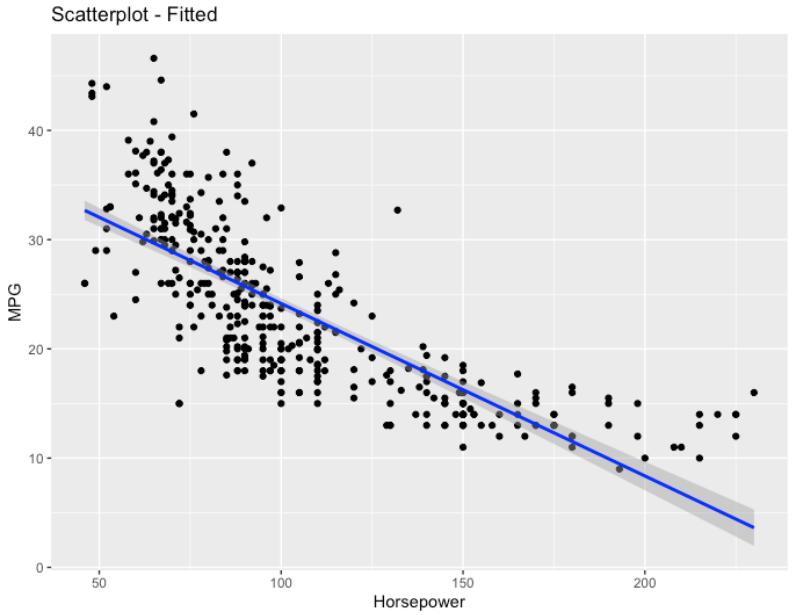
d) Find proportion of the variation that can be explained by the least squares regression line (i.e., R^2).
The proportion of variation that can be explained is .6059. The adjusted R-squared is .6049.

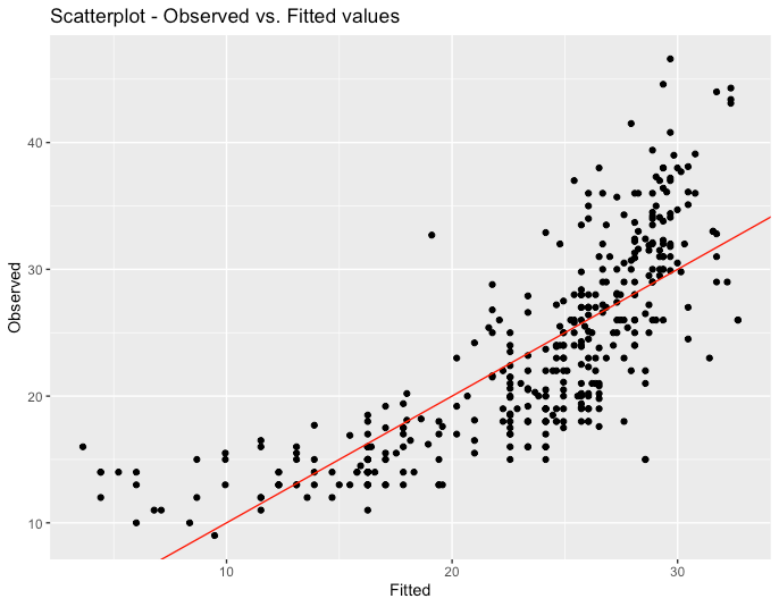
e) Draw the boxplot for the residuals from the linear regression model between mpg and horsepower to check if the data contain any potential outliers?



The data does contain a few potential outliers.

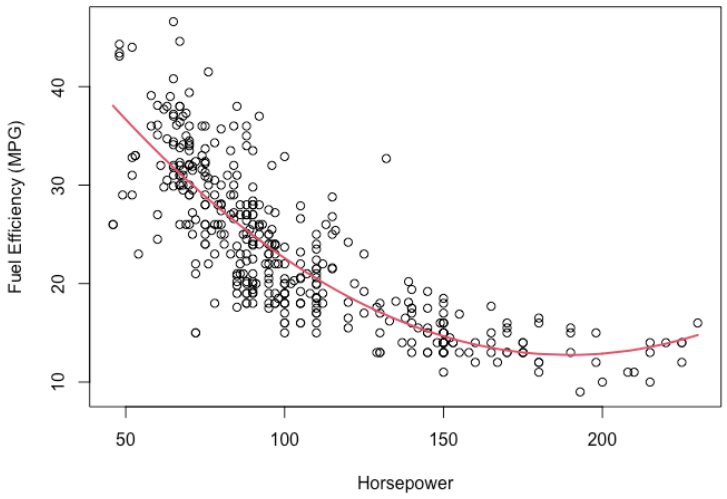
f) Fit a single linear model and conduct 10-fold CV to estimate the error. In addition, draw the scatter plot with the fitted line and the scatter plot between the observed and fitted values.

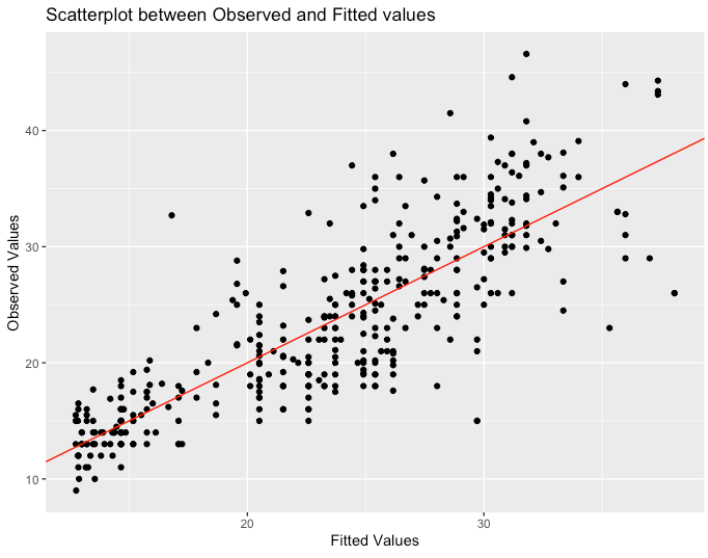




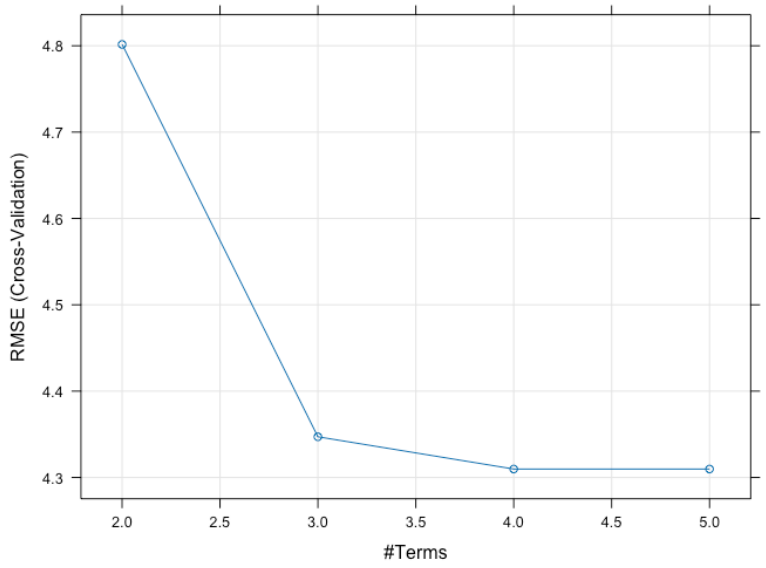
g) Fit a quadratic model and conduct 10-fold CV to estimate the error and draw the scatter plot with the fitted line and the scatter plot between the observed and fitted values. (Hint, you need to sort your data in an ascending order, such that

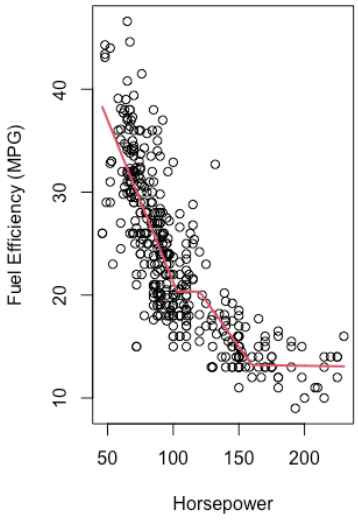
```
#Create the squared term of horsepower, called horsepower2
Auto$horsepower2 = Auto$horsepower^2
#sort the data in an ascending order
Auto = Auto[order(Auto[,3],decreasing=FALSE),]
```

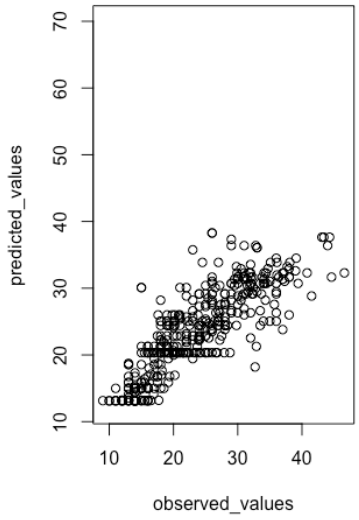




h) Fit a mars model with optimal tuning parameters that you choose and conduct 10-fold CV to estimate the error and draw the scatter plot with the fitted line and the scatter plot between the observed and fitted values.







i) Compare the three fitted models that obtained in g), h) and i) and suggest which model should be preferred according to your criteria, such as R^2 or root mean square error (RMSE), or others.

Since the mars model has the lowest Mean Absolute Error (i.e. the difference between predicted and actual values), the highest R-squared (i.e. the proportion of variance that is explained by the dependent variable), and the lowest RMSE, I choose it as the best fit model.