Huyen Truong – BS22DSY032

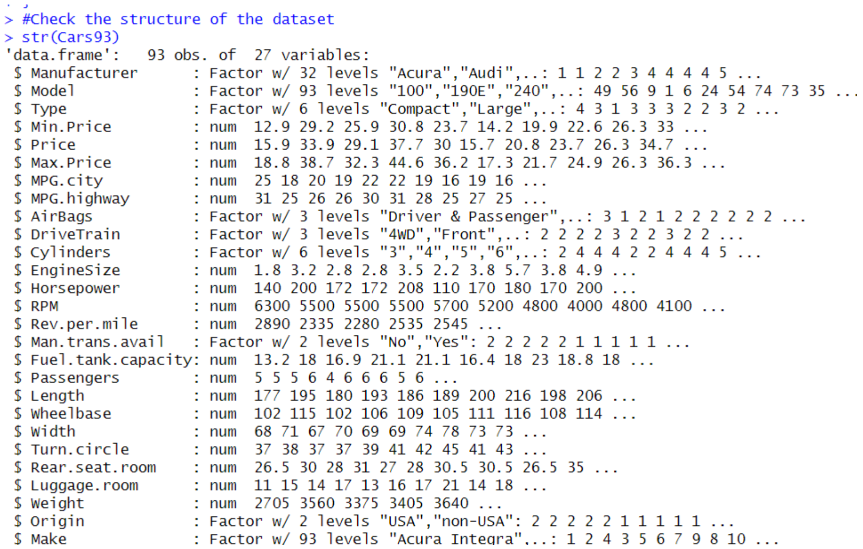
ASSIGNMENT

[PROGRAMMING FOR ANALYTICS]

R Code Part Explanations:

***Note**: The comprehensive R code used in this analysis can be found in the accompanying file 'R code part _ Assignment.R'. Within the file, concise explanations are provided for each step, elucidating the methodology employed. Additionally, brief interpretations of the results obtained are included, offering insights into the outcomes derived from the analysis.*

Check the structure of the dataset:



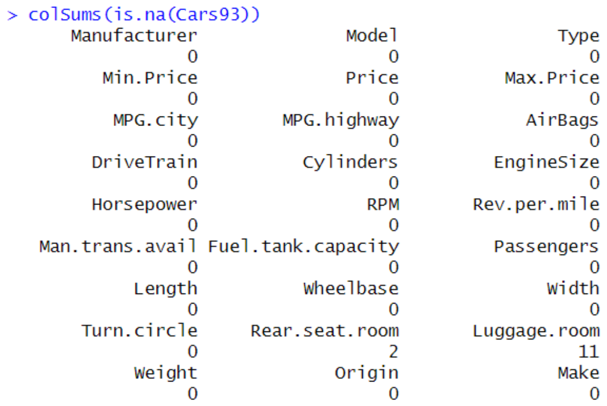
```
> #Check the structure of the dataset
> str(Cars93)
'data.frame':	93 obs. of  27 variables:
 $ Manufacturer      : Factor w/ 32 levels "Acura","Audi",..: 1 1 2 2 3 4 4 4 4 5 ...
 $ Model             : Factor w/ 93 levels "100","190E","240",..: 49 56 9 1 6 24 54 74 73 35 ...
 $ Type              : Factor w/ 6 levels "Compact","Large",..: 4 3 1 3 3 3 2 2 3 2 ...
 $ Min.Price         : num  12.9 29.2 25.9 30.8 23.7 14.2 19.9 22.6 26.3 33 ...
 $ Price             : num  15.9 33.9 29.1 37.7 30 15.7 20.8 23.7 26.3 34.7 ...
 $ Max.Price         : num  18.8 38.7 32.3 44.6 36.2 17.3 21.7 24.9 26.3 36.3 ...
 $ MPG.city          : num  25 18 20 19 22 22 19 16 19 16 ...
 $ MPG.highway       : num  31 25 26 26 30 31 28 25 27 25 ...
 $ AirBags           : Factor w/ 3 levels "Driver & Passenger",..: 3 1 2 1 2 2 2 2 2 2 ...
 $ DriveTrain        : Factor w/ 3 levels "4WD","Front",..: 2 2 2 2 3 2 2 3 2 2 ...
 $ Cylinders         : Factor w/ 6 levels "3","4","5","6",..: 2 4 4 4 2 2 4 4 4 5 ...
 $ EngineSize        : num  1.8 3.2 2.8 2.8 3.5 2.2 3.8 5.7 3.8 4.9 ...
 $ Horsepower        : num  140 200 172 172 208 110 170 180 170 200 ...
 $ RPM               : num  6300 5500 5500 5500 5700 5200 4800 4000 4800 4100 ...
 $ Rev.per.mile      : num  2890 2335 2280 2535 2545 ...
 $ Man.trans.avail   : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 1 1 1 1 ...
 $ Fuel.tank.capacity: num  13.2 18 16.9 21.1 21.1 16.4 18 23 18.8 18 ...
 $ Passengers        : num  5 5 5 6 4 6 6 6 5 6 ...
 $ Length            : num  177 195 180 193 186 189 200 216 198 206 ...
 $ Wheelbase         : num  102 115 102 106 109 105 111 116 108 114 ...
 $ Width             : num  68 71 67 70 69 69 74 78 73 73 ...
 $ Turn.circle       : num  37 38 37 37 39 41 42 45 41 43 ...
 $ Rear.seat.room    : num  26.5 30 28 31 27 28 30.5 30.5 26.5 35 ...
 $ Luggage.room      : num  11 15 14 17 13 16 17 21 14 18 ...
 $ Weight            : num  2705 3560 3375 3405 3640 ...
 $ Origin            : Factor w/ 2 levels "USA","non-USA": 2 2 2 2 2 1 1 1 1 1 ...
 $ Make              : Factor w/ 93 levels "Acura Integra",..: 1 2 4 3 5 6 7 9 8 10 ...
```

The dataset comprises both numeric data and non-numeric variables.

Approach: Multiple regression on numeric data

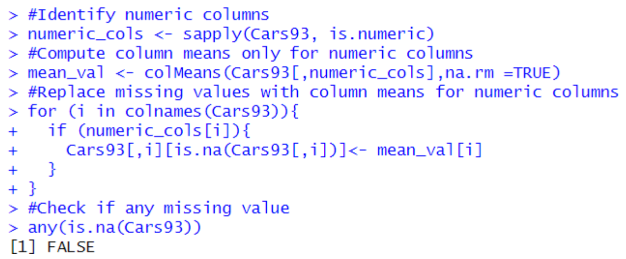
I. Dealing with the missing values.

Check which columns have missing values:



```
> colSums(is.na(Cars93))
     Manufacturer             Model              Type
                0                 0                 0
        Min.Price             Price         Max.Price
                0                 0                 0
         MPG.city       MPG.highway           AirBags
                0                 0                 0
       DriveTrain         Cylinders        EngineSize
                0                 0                 0
       Horsepower               RPM      Rev.per.mile
                0                 0                 0
Man.trans.avail Fuel.tank.capacity        Passengers
                0                 0                 0
           Length         Wheelbase             Width
                0                 0                 0
      Turn.circle    Rear.seat.room      Luggage.room
                0                 2                11
           Weight            Origin              Make
                0                 0                 0
```

There are 2 columns having missing values that are ‘Rear.seat.room’ and ‘Luggage.room’ within the dataset. Then we replaced these missing values with the respective column mean values. A check was conducted to ensure the absence of any further missing values in the dataset.

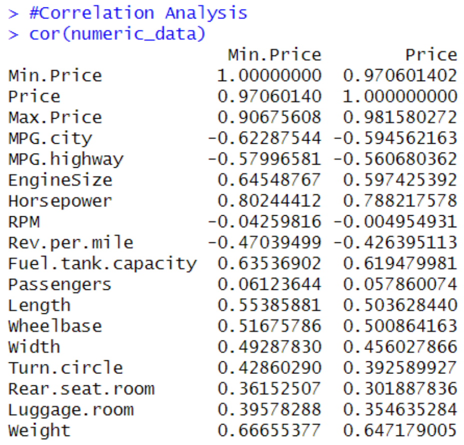


```
> #Identify numeric columns
> numeric_cols <- sapply(Cars93, is.numeric)
> #Compute column means only for numeric columns
> mean_val <- colMeans(Cars93[,numeric_cols],na.rm =TRUE)
> #Replace missing values with column means for numeric columns
> for (i in colnames(Cars93)){
+   if (numeric_cols[i]){
+     Cars93[,i][is.na(Cars93[,i])]<- mean_val[i]
+   }
+ }
> #Check if any missing value
> any(is.na(Cars93))
[1] FALSE
```

II. Perform Multiple Regression on numeric variables.

We have the new dataset ‘numeric_data’ that contains 18 numeric variables.

1. Correlation Analysis



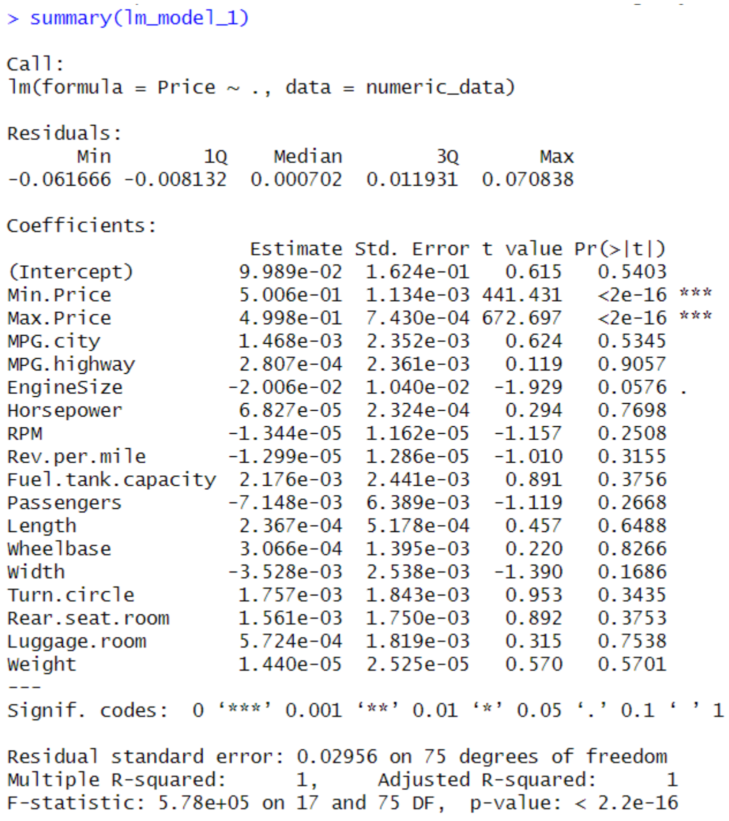
```
> #Correlation Analysis
> cor(numeric_data)
                     Min.Price        Price
Min.Price           1.00000000  0.970601402
Price               0.97060140  1.000000000
Max.Price           0.90675608  0.981580272
MPG.city           -0.62287544 -0.594562163
MPG.highway        -0.57996581 -0.560680362
EngineSize          0.64548767  0.597425392
Horsepower          0.80244412  0.788217578
RPM                -0.04259816 -0.004954931
Rev.per.mile       -0.47039499 -0.426395113
Fuel.tank.capacity  0.63536902  0.619479981
Passengers          0.06123644  0.057860074
Length              0.55385881  0.503628440
Wheelbase           0.51675786  0.500864163
Width               0.49287830  0.456027866
Turn.circle         0.42860290  0.392589927
Rear.seat.room      0.36152507  0.301887836
Luggage.room        0.39578288  0.354635284
Weight              0.66655377  0.647179005
```

<u>Interpretation:</u>

- High Correlation between ‘Price’ and ‘Min.Price’, ‘Max.Price’, ‘Horsepower’

- High Correlation between ('Min.Price' & 'Max.Price'), ('MPG.highway' &'MPG.city'), etc. - Multicollinearity

2. Fitting Multiple Linear Regression on numeric data



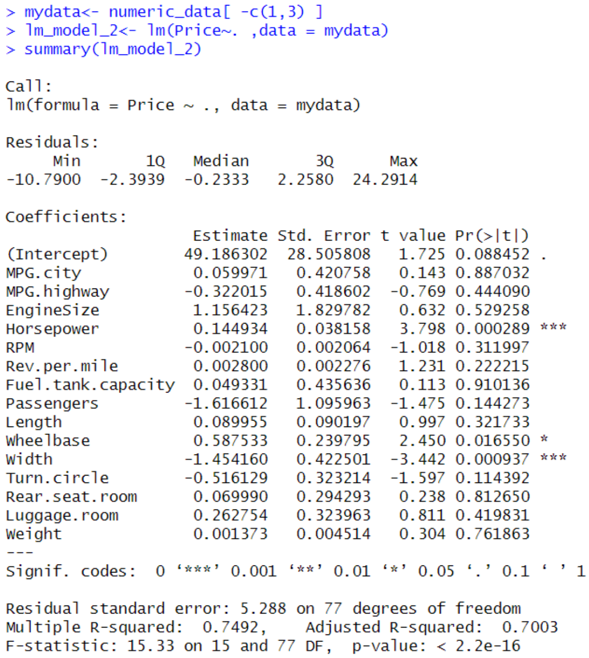
```
> summary(lm_model_1)

Call:
lm(formula = Price ~ ., data = numeric_data)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.061666 -0.008132  0.000702  0.011931  0.070838 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        9.989e-02  1.624e-01   0.615   0.5403    
Min.Price          5.006e-01  1.134e-03 441.431   <2e-16 ***
Max.Price          4.998e-01  7.430e-04 672.697   <2e-16 ***
MPG.city           1.468e-03  2.352e-03   0.624   0.5345    
MPG.highway        2.807e-04  2.361e-03   0.119   0.9057    
EngineSize        -2.006e-02  1.040e-02  -1.929   0.0576 .  
Horsepower         6.827e-05  2.324e-04   0.294   0.7698    
RPM               -1.344e-05  1.162e-05  -1.157   0.2508    
Rev.per.mile      -1.299e-05  1.286e-05  -1.010   0.3155    
Fuel.tank.capacity 2.176e-03  2.441e-03   0.891   0.3756    
Passengers        -7.148e-03  6.389e-03  -1.119   0.2668    
Length             2.367e-04  5.178e-04   0.457   0.6488    
Wheelbase          3.066e-04  1.395e-03   0.220   0.8266    
Width             -3.528e-03  2.538e-03  -1.390   0.1686    
Turn.circle        1.757e-03  1.843e-03   0.953   0.3435    
Rear.seat.room     1.561e-03  1.750e-03   0.892   0.3753    
Luggage.room       5.724e-04  1.819e-03   0.315   0.7538    
Weight             1.440e-05  2.525e-05   0.570   0.5701    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02956 on 75 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 5.78e+05 on 17 and 75 DF,  p-value: < 2.2e-16
```

Interpretation:

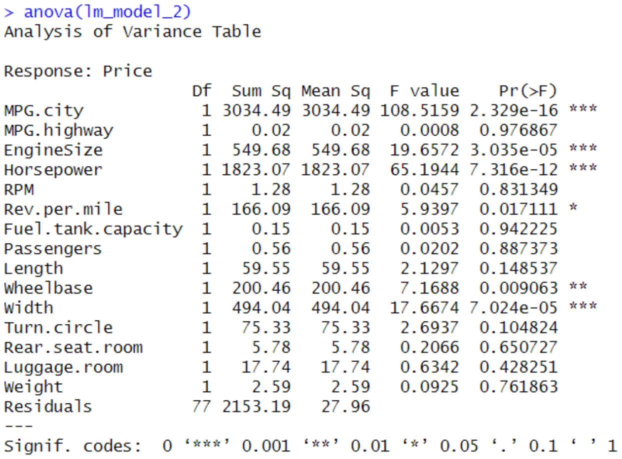
- Among the predictors, 'Min.Price' and 'Max.Price' (p_value<0.05) have a significant positive association with Price.
- We also consider 'EngineSize'.
- p_value associated with the F-statistic < 2.2e-16
  - ➔ The model is statistically significant at a very high level of confidence.
- R-squared and Adjusted R-squared = 1 → Possibility of overfitting
  - ➔ Need of further validation

Upon careful consideration, we have decided to exclude the variables 'Min.Price' and 'Max.Price' from our linear model. Since 'Price' inherently encapsulates the range between 'Min.Price' and 'Max.Price' ('Price' is the average of 'Min.Price' and 'Max.Price'), including both variables simultaneously could potentially lead to multicollinearity issues within our model.

After excluding variables 'Min.Price' and 'Max.Price', a new linear model, denoted as 'lm_model_2', was fitted using the updated dataset 'mydata':



```
> mydata<- numeric_data[ -c(1,3) ]
> lm_model_2<- lm(Price~. ,data = mydata)
> summary(lm_model_2)

Call:
lm(formula = Price ~ ., data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.7900  -2.3939  -0.2333   2.2580  24.2914 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       49.186302  28.505808   1.725 0.088452 .  
MPG.city           0.059971   0.420758   0.143 0.887032    
MPG.highway       -0.322015   0.418602  -0.769 0.444090    
EngineSize         1.156423   1.829782   0.632 0.529258    
Horsepower         0.144934   0.038158   3.798 0.000289 ***
RPM               -0.002100   0.002064  -1.018 0.311997    
Rev.per.mile       0.002800   0.002276   1.231 0.222215    
Fuel.tank.capacity 0.049331   0.435636   0.113 0.910136    
Passengers        -1.616612   1.095963  -1.475 0.144273    
Length             0.089955   0.090197   0.997 0.321733    
Wheelbase          0.587533   0.239795   2.450 0.016550 *  
Width             -1.454160   0.422501  -3.442 0.000937 ***
Turn.circle       -0.516129   0.323214  -1.597 0.114392    
Rear.seat.room     0.069990   0.294293   0.238 0.812650    
Luggage.room       0.262754   0.323963   0.811 0.419831    
Weight             0.001373   0.004514   0.304 0.761863    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.288 on 77 degrees of freedom
Multiple R-squared:  0.7492,	Adjusted R-squared:  0.7003 
F-statistic: 15.33 on 15 and 77 DF,  p-value: < 2.2e-16
```



```
> anova(lm_model_2)
Analysis of Variance Table

Response: Price
                   Df  Sum Sq Mean Sq  F value    Pr(>F)    
MPG.city            1 3034.49 3034.49 108.5159 2.329e-16 ***
MPG.highway         1    0.02    0.02   0.0008  0.976867    
EngineSize          1  549.68  549.68  19.6572 3.035e-05 ***
Horsepower          1 1823.07 1823.07  65.1944 7.316e-12 ***
RPM                 1    1.28    1.28   0.0457  0.831349    
Rev.per.mile        1  166.09  166.09   5.9397  0.017111 *  
Fuel.tank.capacity  1    0.15    0.15   0.0053  0.942225    
Passengers          1    0.56    0.56   0.0202  0.887373    
Length              1   59.55   59.55   2.1297  0.148537    
Wheelbase           1  200.46  200.46   7.1688  0.009063 ** 
Width               1  494.04  494.04  17.6674 7.024e-05 ***
Turn.circle         1   75.33   75.33   2.6937  0.104824    
Rear.seat.room      1    5.78    5.78   0.2066  0.650727    
Luggage.room        1   17.74   17.74   0.6342  0.428251    
Weight              1    2.59    2.59   0.0925  0.761863    
Residuals          77 2153.19   27.96                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

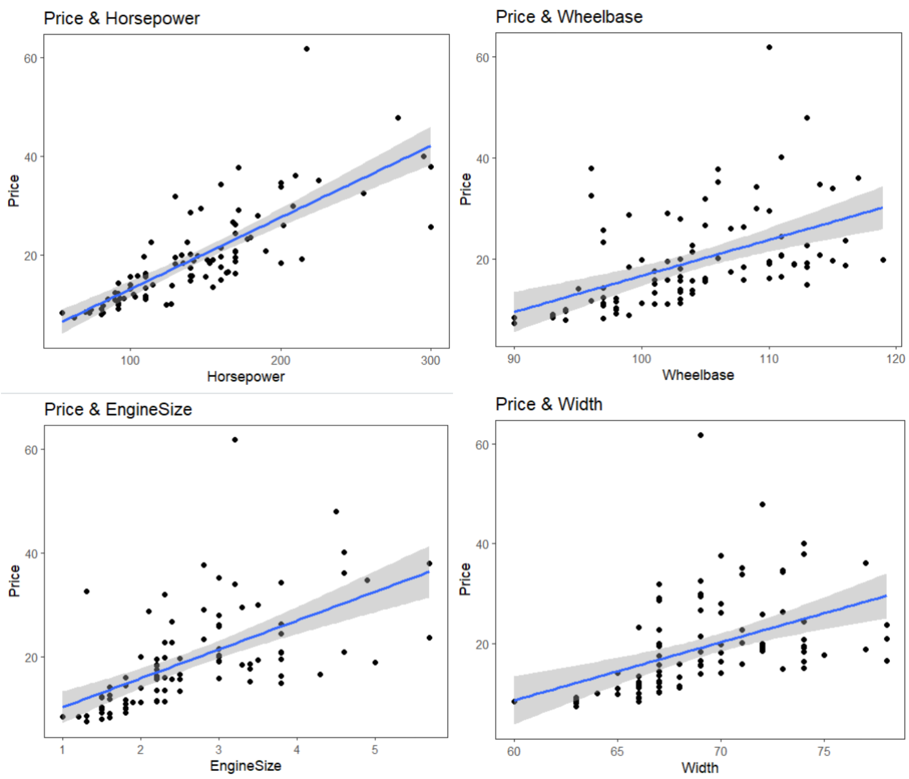
<u>Interpretation:</u>

- With p_value< 0.05, ‘Horsepower’ and ‘Wheelbase’ have a significant positive association with ‘Price’, while ‘Width’ demonstrates a significant negative association.
- We consider ‘EngineSize’ variable because of the large coefficient and having F_value high and p_value <0.05 in ANOVA test.
- Multiple R-squared value of 0.7492 and an Adjusted R-squared value of 0.7003 indicates a reasonably good fit of the model to the data.

3. Using ggplot to plot the necessary graphs.

Investigating variables with potential impact on ‘Price’ as indicated by the summary statistics and ANOVA results.

Scatterplot between ‘Price’ and ‘Horsepower’, ‘Wheelbase’, ‘EngineSize’ and ‘Width’ respectively:

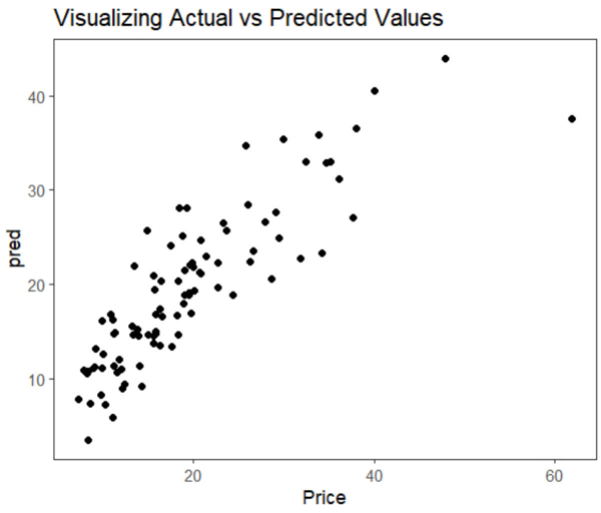


<u>Interpertation:</u>

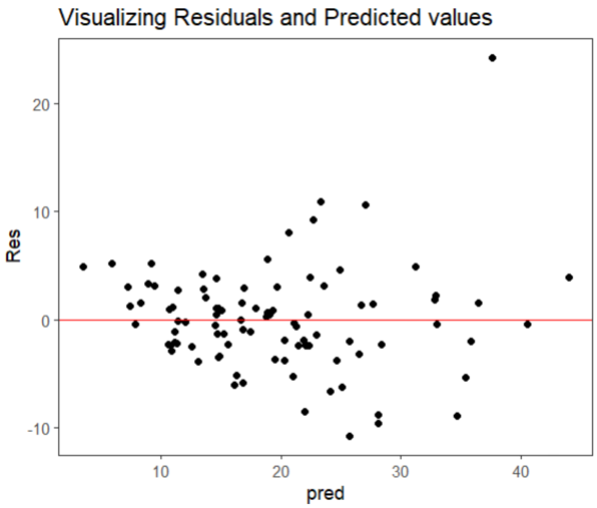
- After plotting scatterplots between ‘Price’ and ‘Horsepower’, ‘EngineSize’, and ‘Wheelbase’, it is evident that there appears to be a positive linear relationship between Price and each of these variables: ‘Horsepower’, ‘Wheelbase’ and ‘EngineSize’.
- The linear model coefficient for Width is negative (-1.454160) from summary(lm_model_2), seemingly contradicting the positive trend observed in the scatterplot.

  One explanation could be confounding variables. Wider cars might often be luxury models with higher prices, even if width itself doesn't directly increase the price. These other factors could be masking the true (potentially negative) effect of width on price. Further analysis to account for these confounding variables might be necessary to understand the true influence of ‘Width’ on ‘Price’.

Visualizing Actual and Predicted Values:



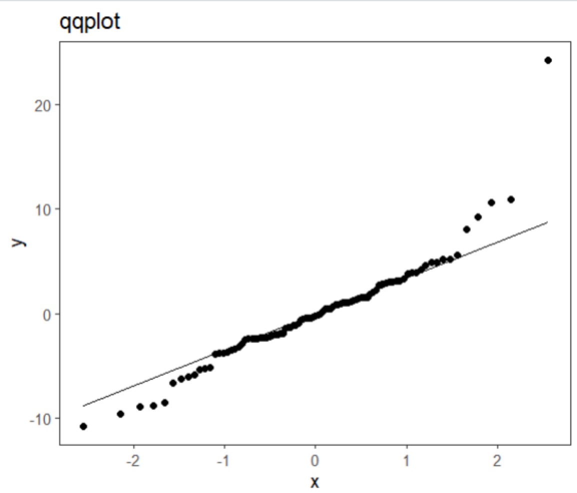
Visualizing the relationship between Residuals and Predictor



<u>Interpretation:</u>

From the graph, we observe that the points are randomly distributed above and below the reference line. This suggests that the model captures the linear relationship between the predictors and the dependent variable ‘Price’.

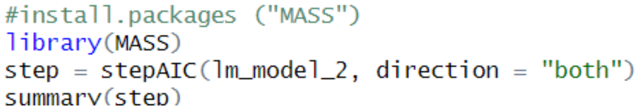
Checking whether the distribution of the Residuals is bell shaped:



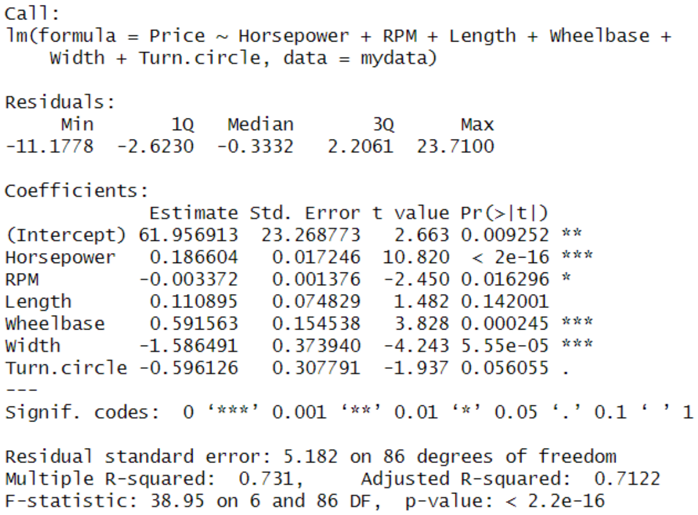
Interpretation:

The qqplot of the residuals from the linear model suggests that the central part of the error distribution aligns with a normal distribution. However, there are deviations in the tails, indicating the presence of potential outliers or a non-normal error distribution.

4. Finding out the most influential explanatory variables using stepAIC()



```
#install.packages ("MASS")
library(MASS)
step = stepAIC(lm_model_2, direction = "both")
summary(step)
```



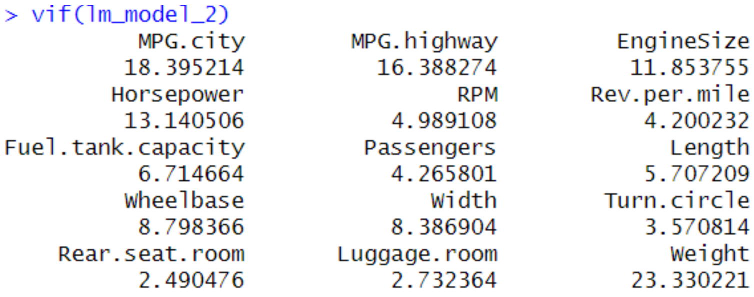
```
Call:
lm(formula = Price ~ Horsepower + RPM + Length + Wheelbase + 
    Width + Turn.circle, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.1778  -2.6230  -0.3332   2.2061  23.7100 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 61.956913  23.268773   2.663 0.009252 ** 
Horsepower   0.186604   0.017246  10.820  < 2e-16 ***
RPM         -0.003372   0.001376  -2.450 0.016296 *  
Length       0.110895   0.074829   1.482 0.142001    
Wheelbase    0.591563   0.154538   3.828 0.000245 ***
Width       -1.586491   0.373940  -4.243 5.55e-05 ***
Turn.circle -0.596126   0.307791  -1.937 0.056055 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.182 on 86 degrees of freedom
Multiple R-squared:  0.731,	Adjusted R-squared:  0.7122 
F-statistic: 38.95 on 6 and 86 DF,  p-value: < 2.2e-16
```

Interpretation:

From the summary, the p-values associated with the variables ‘Horsepower’, ‘Wheelbase’, ‘Width’, and ‘RPM’ are below 0.05. This indicates that these variables have a statistically significant impact on the ‘Price’.

5. Measure for Multicollinearity using Variance Inflation Factor (VIF) to measure the correlation (linear association) between each x variable with other x's.

Criteria: VIF>5 indicates multicollinearity



```
> vif(lm_model_2)
         MPG.city       MPG.highway        EngineSize 
        18.395214         16.388274         11.853755 
       Horsepower               RPM      Rev.per.mile 
        13.140506          4.989108          4.200232 
Fuel.tank.capacity        Passengers            Length 
         6.714664          4.265801          5.707209 
        Wheelbase             Width       Turn.circle 
         8.798366          8.386904          3.570814 
   Rear.seat.room      Luggage.room            Weight 
         2.490476          2.732364         23.330221 
```

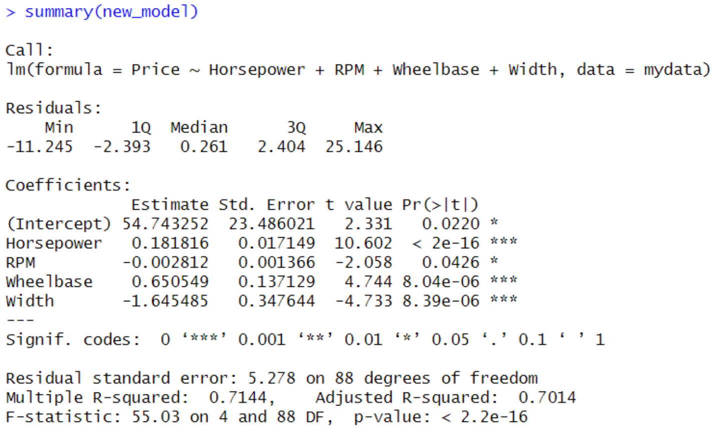
Interpretation:

VIF > 5 indicates multi-collinearity. Hence, multi-collinearity exists between 'MPG.city', 'MGP.highway', 'EngineSize', 'Horsepower', 'Fuel.tank.capacity', 'Wheelbase', 'Width' and 'Weight'.

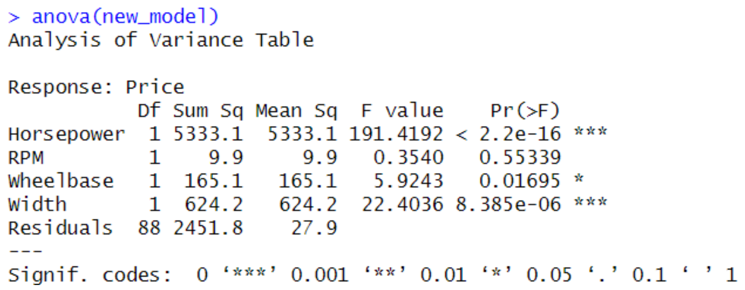
6. Tackling Multicollinearity by Removing highly correlated variable – Stepwise Regression

After doing Stepwise Regression, we have a new model:

Price ~ Horsepower + RPM + Wheelbase + Width



```
> summary(new_model)

Call:
lm(formula = Price ~ Horsepower + RPM + Wheelbase + Width, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.245  -2.393   0.261   2.404  25.146 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 54.743252  23.486021   2.331   0.0220 *  
Horsepower   0.181816   0.017149  10.602  < 2e-16 ***
RPM         -0.002812   0.001366  -2.058   0.0426 *  
Wheelbase    0.650549   0.137129   4.744 8.04e-06 ***
Width       -1.645485   0.347644  -4.733 8.39e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.278 on 88 degrees of freedom
Multiple R-squared:  0.7144,	Adjusted R-squared:  0.7014 
F-statistic: 55.03 on 4 and 88 DF,  p-value: < 2.2e-16
```



```
> anova(new_model)
Analysis of Variance Table

Response: Price
           Df Sum Sq Mean Sq  F value    Pr(>F)    
Horsepower  1 5333.1  5333.1 191.4192 < 2.2e-16 ***
RPM         1    9.9     9.9   0.3540   0.55339    
Wheelbase   1  165.1   165.1   5.9243   0.01695 *  
Width       1  624.2   624.2  22.4036 8.385e-06 ***
Residuals  88 2451.8    27.9                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation:

- ‘Horsepower’ has the highest F-statistic and a very low p-value in both analyses, making it the most influential variable.
- ‘Wheelbase’ and ‘Width’ also have significant p-values in the summary and high F-statistics with low p-values in the ANOVA table, suggesting they are influential.
- ‘RPM’ has a significant p-value in the summary but a low F-statistic and high p-value in the ANOVA. Its influence might be weaker compared to others.

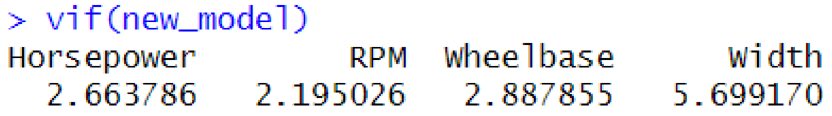
New model:

y^hat = 0.182*Horsepower + 0.650*Wheelbase - 1.645*Width -0.003*RPM + 54.743

Based on the analysis:

- ‘Horsepower’, ‘Wheelbase’, and ‘Width’ are the most influential variables affecting on ‘Price’ in this model.
- The model has a good fit, explaining over 70% of the price variance, and is statistically significant.

Checking for multicollinearity in the new model:



```
> vif(new_model)
Horsepower        RPM  Wheelbase      Width 
  2.663786   2.195026   2.887855   5.699170 
```

Interpretation:

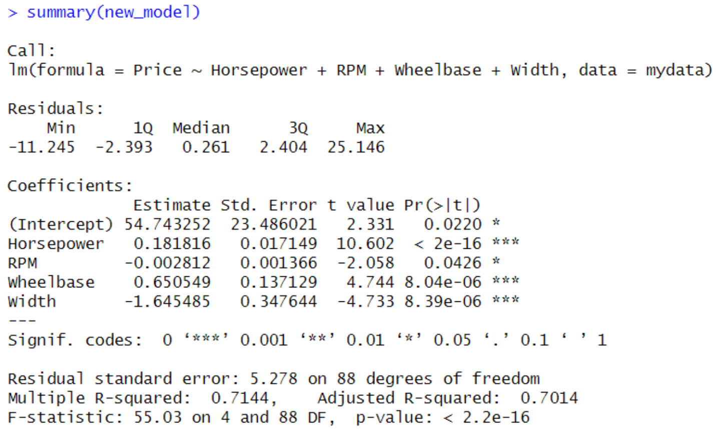
All variables except ‘Width’ have VIF values below 5, indicating no significant multicollinearity concerns. Although ‘Width’ has a VIF value of 5.699, falling between the acceptable range of 5 to 10, it can be deemed acceptable in this context.

7. **Performing model utility test**

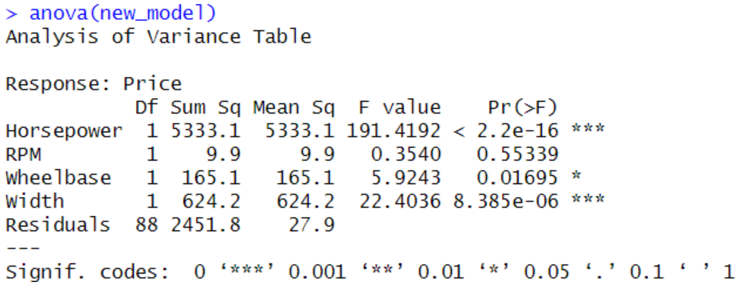
Defining the null hypothesis and alternative hypothesis:

H0: β1=β2=⋯= 0

H1=At least one β≠0 whilst p<0.05



```
> summary(new_model)

Call:
lm(formula = Price ~ Horsepower + RPM + Wheelbase + Width, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.245  -2.393   0.261   2.404  25.146 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 54.743252  23.486021   2.331   0.0220 *  
Horsepower   0.181816   0.017149  10.602  < 2e-16 ***
RPM         -0.002812   0.001366  -2.058   0.0426 *  
Wheelbase    0.650549   0.137129   4.744 8.04e-06 ***
Width       -1.645485   0.347644  -4.733 8.39e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.278 on 88 degrees of freedom
Multiple R-squared:  0.7144,    Adjusted R-squared:  0.7014 
F-statistic: 55.03 on 4 and 88 DF,  p-value: < 2.2e-16
```



```
> anova(new_model)
Analysis of Variance Table

Response: Price
             Df Sum Sq Mean Sq  F value    Pr(>F)    
Horsepower    1 5333.1  5333.1 191.4192 < 2.2e-16 ***
RPM           1    9.9     9.9   0.3540   0.55339    
Wheelbase     1  165.1   165.1   5.9243   0.01695 *  
Width         1  624.2   624.2  22.4036 8.385e-06 ***
Residuals    88 2451.8    27.9                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation:

- The results of the test indicate that the p-value is less than 0.05, suggesting that variables such as ‘Horsepower’, ‘Width’, ‘Wheelbase’, and ‘RPM’ have a statistically significant impact on the Price.
- The overall F-statistic (55.03) and its p-value (< 2.2e-16) indicate a statistically significant model.

Therefore, we **reject** the null hypothesis in favor of the alternative hypothesis, concluding that at least one predictor variable in the model has a significant effect on the Price.

8. The 99% confidence interval of the slopes

Intercept	[-7.09 116.58]
Horsepower	[0.136 0.226]
RPM	[-0.0064 0.00078]
Wheelbase	[0.29 1.01]
Width	[-2.56 -0.73]

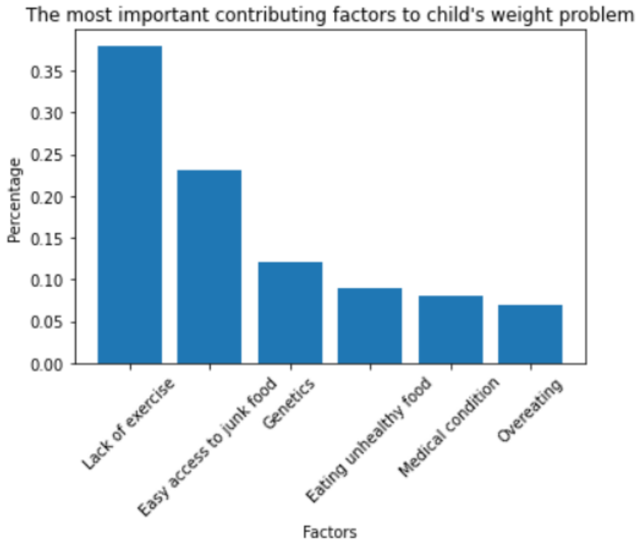
Python code explanations:

***Note**: The comprehensive Python code used in this analysis can be found in the accompanying file ‘Python Part _ Assignment R’. Within the file, brief explanations are provided for each step, elucidating the methodology employed.*

For **questions 1,2,3**, I wrote the functions following requirements and test functions to ensure that the code works well.

Question 4:

a. Construct a bar chart for the data:



b. It would be reasonable that:
"Lack of exercise" with "Overeating" might be grouped into a single category named 'Unhealthy Lifestyle Habits' or 'Behavioral Factors'. Both factors contribute to an imbalance between calorie intake and expenditure, leading to weight gain.

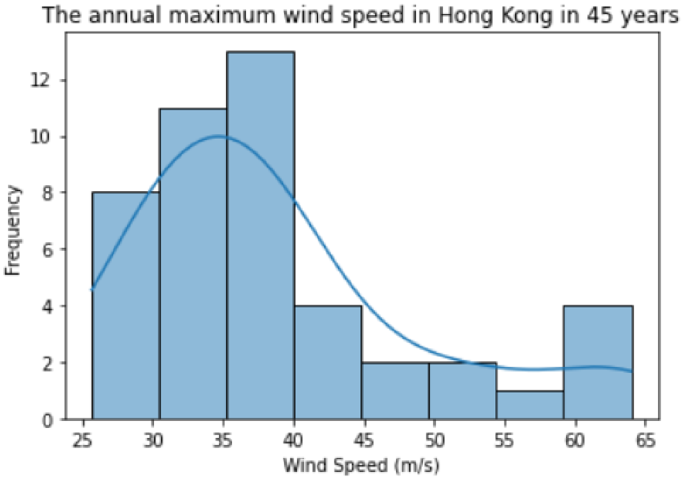
and "Easy access to junk food" and "Eating unhealthy food" can be aggregated into the category 'Poor Dietary Choices' or 'Unhealthy Eating Habits'. Both factors involve consuming high-calorie, low-nutrient foods, which can contribute to weight gain.

While we keep 'Genetics' and 'Medical condition' separate.

OR we have another way to combine some of those factors into a single category:

'Lack of Exercise', 'Easy access to junk food', 'Eating unhealthy food' and 'Overeating might be combined into single category named 'Lifestyle choice ' because they all relate to lifestyle choices.

Question 5:



From the above histograms, we can see that:

- The histogram appears to be positively skewed, as the tail extends to the right.
- The histogram is unimodal since there is only one peak.

In conclusion, the histogram for the annual maximum wind speed data is positively skewed and unimodal.