Car Price Prediction

using Multiple Linear Regression

MSDA3055-03-S24: Linear Regression and Time series

By: Sai Nithisha Marripelly

Sai Koushik Kadiyala

Kartheek Kamadana

Naaz Nagori

Amanda Mahajan

Predictive Modeling for Car Prices: A Multiple Linear Regression Approach

Introduction

"Driving Innovation: Exploring Car Price Prediction through Multiple Linear Regression Analysis"

In a bid to conquer the American automotive market, Chinese giant Geely Auto has enlisted our consultancy to unravel the factors dictating car prices. Our mission? To decode the intricate relationships between variables and pricing, empowering Geely to make informed strategic decisions in a fiercely competitive landscape.

Research Questions

Which Variables Drive Car Prices in the American Market?
How Effectively Can Multiple Linear Regression Model Car Prices?
Implications for Business Strategy and Market Penetration

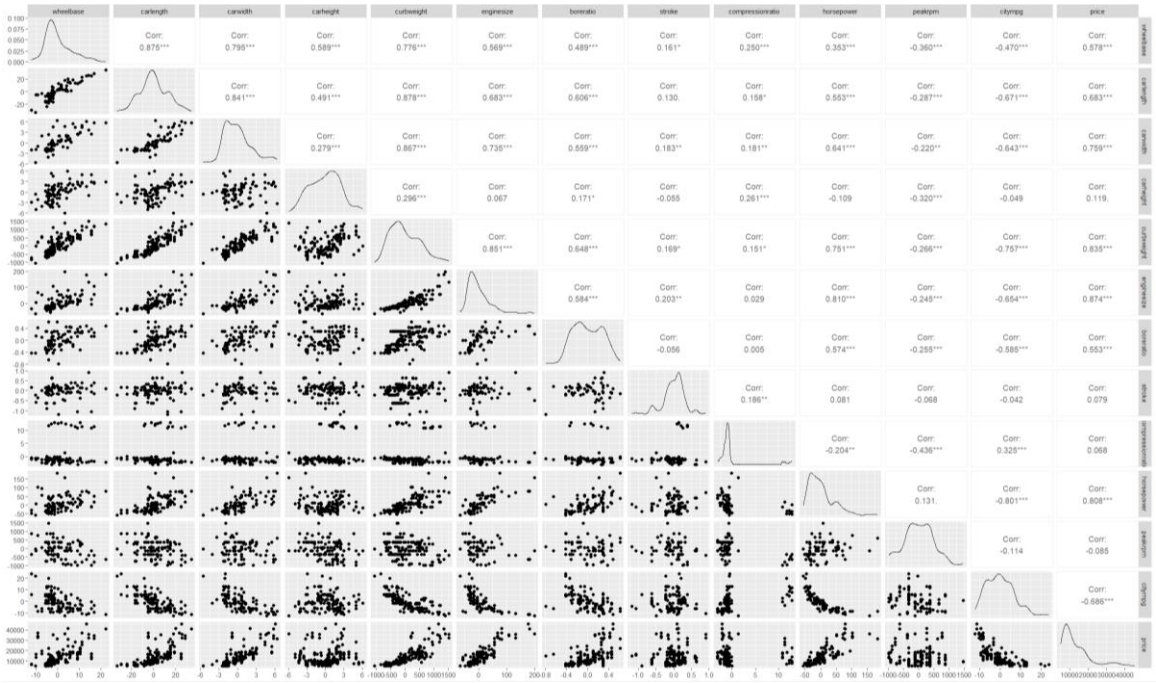
Data Description

The dataset under examination, sourced from Kaggle, consists of 205 rows encompassing a wide array of variables related to vehicle attributes. This comprehensive dataset showcases a mix of both numerical and categorical data, revealing its expansive dimensions and diverse data types pertinent to the prediction of car prices. Below is a glimpse into the dataset description:

1	Car_ID	Unique id of each observation	(Integer)
2	Symboling	Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.	(Categorical)
3	carCompany	Name of car company	(Categorical)
4	fueltype	Car fuel type i.e gas or diesel	(Categorical)
5	aspiration	Aspiration used in a car	(Categorical)
6	doornumber	Number of doors in a car	(Categorical)
7	carbody	body of car	(Categorical)
8	drivewheel	type of drive wheel	(Categorical)
9	enginelocation	Location of car engine	(Categorical)

10	wheelbase	Wheelbase of car	(Numeric)
11	carlength	Length of car	(Numeric)
12	carwidth	Width of car	(Numeric)
13	carheight	height of car	(Numeric)
14	curbweight	The weight of a car without occupants or baggage	(Numeric)
15	enginetype	Type of engine	(Categorical)
16	cylindernumber	cylinder placed in the car	(Categorical)
17	enginesize	Size of car	(Numeric)
18	fuelsystem	Fuel system of car	(Categorical)
19	boreratio	Boreratio of car	(Numeric)
20	stroke	Stroke or volume inside the engine	(Numeric)
	compressionratio	compression ratio of car	(Numeric)
22	horsepower	Horsepower	(Numeric)
23	peakrpm	car peak rpm	(Numeric)
24	citympg	Mileage in city	(Numeric)
25	highwaympg	Mileage on highway	(Numeric)
26	price(Dependent variable)	Price of car	(Numeric)

The below plot shows correlation between Y (Price) and Continuous predictor variables:



Data Cleaning and Data Preparation

Our initial model considered the categorical variable with at most five unique values. However, generating interaction terms with all possible combinations would have resulted in a high number of predictor variables. This could lead to issues with multicollinearity and model complexity. To address these concerns, we strategically grouped the categorical values into three or less concise categories.

Extracting Company Name and Creating New Columns:

- We extracted the company name from the 'CarName' column and created a new 'Company' column to better categorize the data based on the car manufacturer.

Cleaning Company Names and Creating BrandType:

- We cleaned specific company names for consistency and created a new 'BrandType' column to categorize brands into luxury and midrange categories based on industry classifications.

Encoding Categorical Predictors as Factors:

- To facilitate statistical calculations and model building, we encoded categorical predictors such as 'symboling', 'BrandType', 'enginelocation', 'cylindernumber', and 'fuelsystem' as factors.
-

Creating Dummy Variables for Categorical Predictors:

- Using the 'fastDummies' library, we created dummy variables for categorical predictors, transforming them into binary indicators for each category.
-

Centering and Creating Squared Terms:

- We centered continuous variables to have a mean of zero, aiding in the interpretation of interaction terms. Additionally, we created squared terms for continuous variables to capture potential non-linear relationships in the data.
-

Preparing Interaction Terms:

- Interaction terms were generated to capture the combined effects of predictors, enhancing the predictive power of our models.
-

Removing Unnecessary Columns:

Price-related columns were removed from certain datasets to ensure model accuracy and prevent multicollinearity issues.

Dataset Splitting and Model Validation

To assess the generalizability of our modeling approach and ensure the chosen data split was effective, we employed a train-test split strategy. The data was divided into two sets:

Training Set (80%): This larger portion was used to train the models. The training process involves fitting the model parameters to the data to learn the underlying relationships between the predictor variables and the dependent variable.

Test Set (20%): This unseen portion of the data was held out from the training process. It was used to evaluate the performance of the trained models on unfamiliar data.

Training Set Model (fwd_final1):

```
fwd_final1 <- lm(price ~ wheelbase+carlength+carwidth+carheight+curbweight+enginesize+boreratio+stroke+compressionratio+horsepower+peakrpm+citympg+enginelocation_rear+symboling_Safe+BrandType_Midrange+cylindernumber_NormalPerf+fuelsystem_FuelInjected, data = train_data)
```

```
Residual standard error: 2584 on 147 degrees of freedom
Multiple R-squared:  0.911,     Adjusted R-squared:  0.9007
F-statistic: 88.46 on 17 and 147 DF,  p-value: < 2.2e-16
```

Test Set Model (fwd_final2):

```
fwd_final2 <- lm(price ~ wheelbase+carlength+carwidth+carheight+curbweight+enginesize+boreratio+stroke+compressionratio+horsepower+peakrpm+citympg+enginelocation_rear+symboling_Safe+BrandType_Midrange+cylindernumber_NormalPerf+fuelsystem_FuelInjected, data = test_data)
```

```
Residual standard error: 2588 on 22 degrees of freedom
Multiple R-squared:  0.9262,    Adjusted R-squared:  0.8691
F-statistic: 16.24 on 17 and 22 DF,  p-value: 1.039e-08
```

The model achieved an R-squared of 0.911 on the training set and 0.92 on the test set. This indicates good generalizability, as the model can capture the underlying relationship in the unseen test data as well.

Model Selection and Approach

Identified statistically significant individual predictors using regsubsets. This method likely involves iteratively fitting models with different combinations of predictor variables and selecting the model with the best performance based on a specific metric (e.g., adjusted R-squared).

Built a model using forward selection with AIC criteria which included all possible interactions between the significant predictors we identified earlier.

This likely resulted in a large number of predictor variables (i.e. 121), making the model complex and potentially difficult to interpret.

Next, we used the first 50 predictor variables from 121 predictor variables as input to regsubsets.

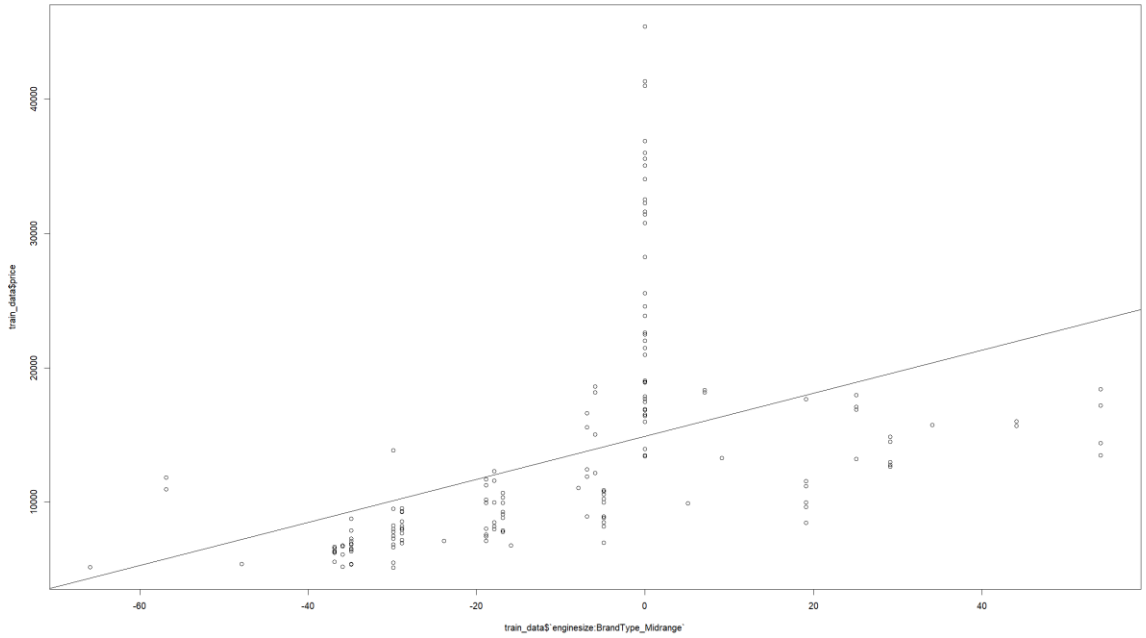
Regsubsets identified models with only 8 informative predictor variables, which was used in the further steps for model building.

fwd.final <- lm(price~ curbweight+enginesize+`enginesize:BrandType_Midrange`+BrandType_Midrange+`curbweight:enginelocation_rear`+`stroke:compressionratio`+`enginesize:compressionratio`+`horsepower:BrandType_Midrange`, data = train_data)

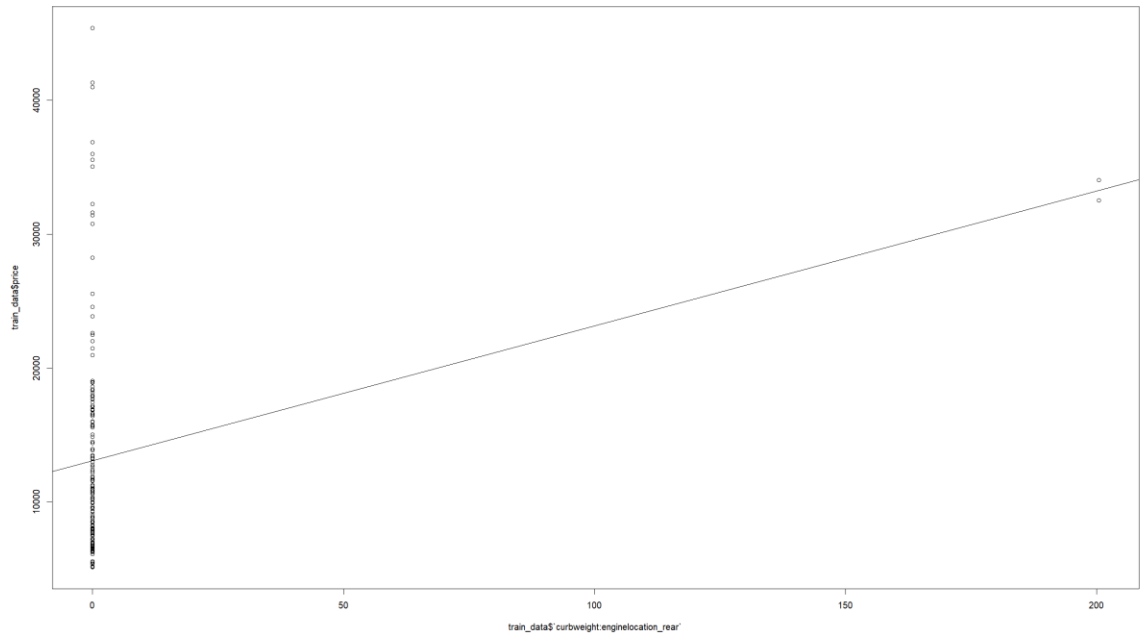
```
Residual standard error: 2151 on 156 degrees of freedom
Multiple R-squared:  0.9345,    Adjusted R-squared:  0.9311
F-statistic: 278.2 on 8 and 156 DF,  p-value: < 2.2e-16
```

Slope plots of Interaction terms in the model

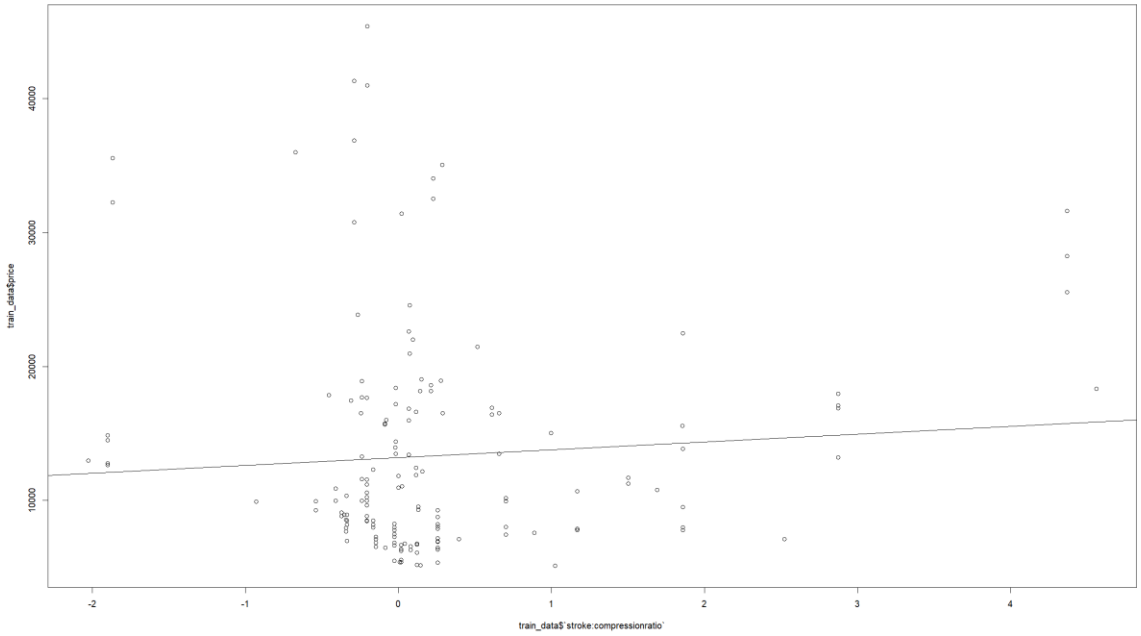
In plots depicting interaction terms versus the response variable with a noticeable slope, a significant relationship can be inferred. The slope indicates the strength and direction of this relationship, highlighting the interaction's influence on the response variable's behavior.



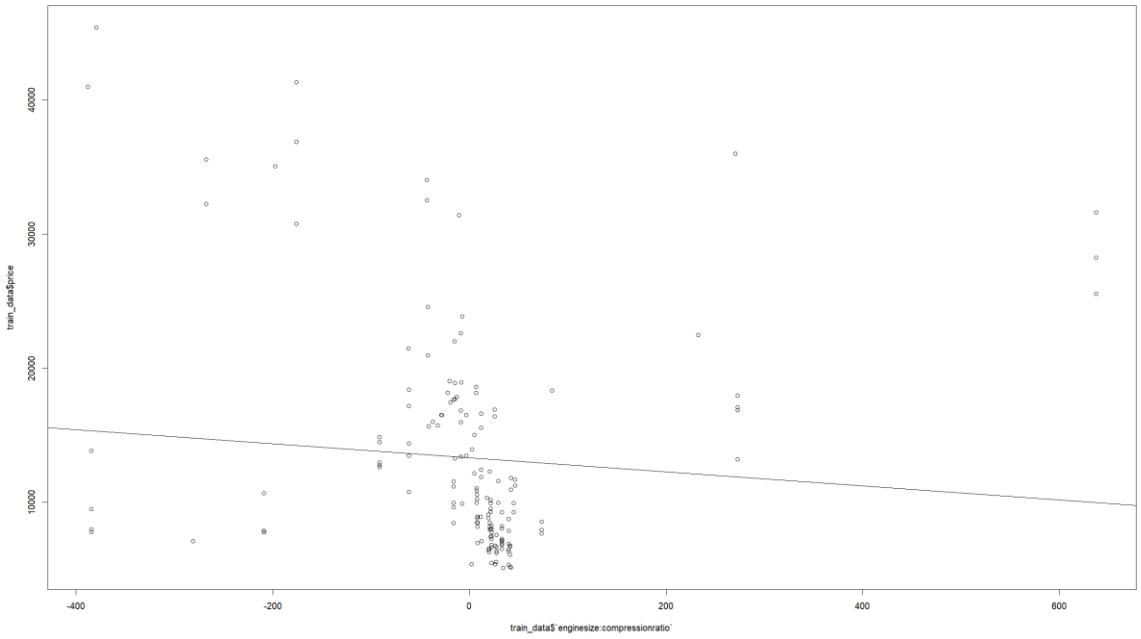
The above plot suggests that enginesize:BrandType_Midrange has significant relation with train_data$Price.



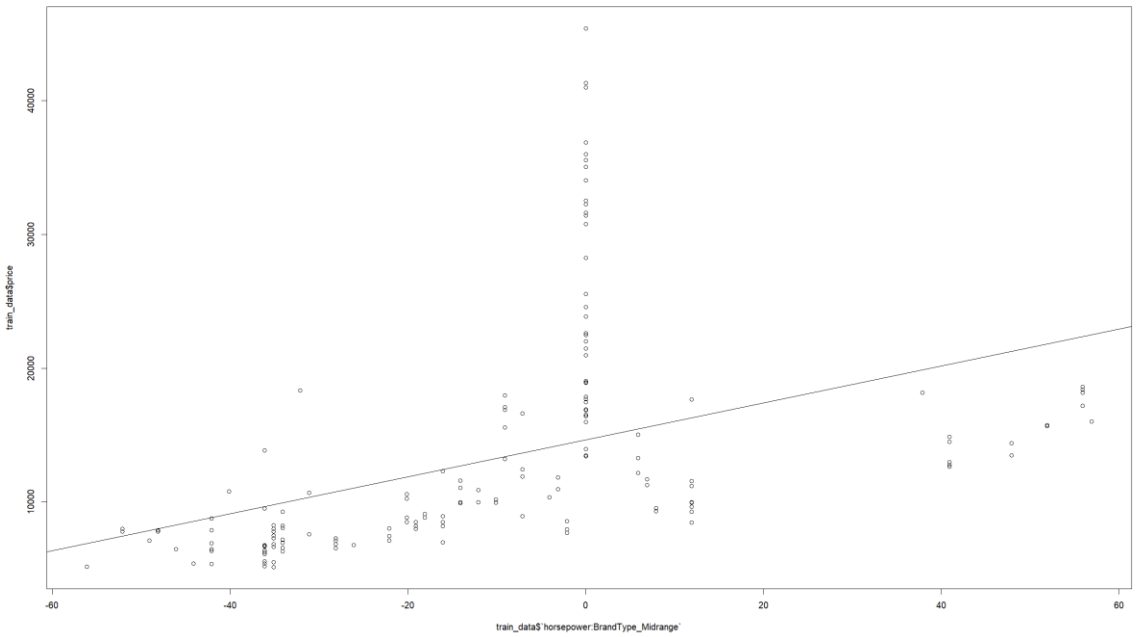
The above plot suggests that curbweight:enginelocation_rear has significant relation with train_data$Price.



The above plot suggests that the influence of the stroke:compressionratio on the train_data$Price is not substantial or conclusive.



The above plot suggests that enginesize:compressionratio has significant relation with train_data$Price.



The above plot suggests that horsepower:BrandType has significant relation with train_data$Price.

We proceeded by incorporating individual predictors involved in the interaction terms from the model, namely enginelocation_rear, Stroke, Compressionratio, and Horsepower.

VIF Calculation

During VIF calculation, we identified aliased coefficients within the model. Similarly, when plotting interaction slopes against the response variable, we observed two identical plots involving interaction terms: enginelocation_rear and curbweight:enginelocation_rear. To address this redundancy, we removed the duplicated interaction term curbweight:enginelocation_rear from the model.
Subsequently, we proceeded to eliminate predictor variables with VIF values exceeding 7.

> **vif_fwd_2**

Curbweight	**enginesize**	**enginesize:BrandType_Midrange**
5.814121	5.008280	2.267670
BrandType_Midrange	**enginelocation_rear**	**stroke:compressionratio**
1.904032	1.146645	6.848867
Stroke	**compressionratio**	**enginesize:compressionratio**
2.007663	5.445955	1.705219

Summary of Model: fwd_final1_3:

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    15669.6091   538.0630  29.122  < 2e-16 ***
curbweight                         6.2586     0.7968   7.854 6.25e-13 ***
enginesize                        92.2712     9.0821  10.160  < 2e-16 ***
`enginesize:BrandType_Midrange`  -77.3064    11.4260  -6.766 2.56e-10 ***
BrandType_Midrange             -4920.7329   573.1139  -8.586 8.94e-15 ***
enginelocation_rear             9970.7878  1711.2096   5.827 3.17e-08 ***
`stroke:compressionratio`       1780.1035   437.3756   4.070 7.47e-05 ***
stroke                          1926.6486   792.6519   2.431 0.016214 *
compressionratio                -332.5377    97.1683  -3.422 0.000794 ***
`enginesize:compressionratio`     -5.9621     1.6414  -3.632 0.000381 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2246 on 155 degrees of freedom
Multiple R-squared:  0.929,     Adjusted R-squared:  0.9249
F-statistic: 225.5 on 9 and 155 DF,  p-value: < 2.2e-16
```

Outliers with respect to Y:

```
> table[table$"abs(t)-qt(1-alpha/(2*n),n-p-1)">0,]
   curbweight enginesize horsepower BrandType_Midrange enginesize:BrandType_Midrange     Y
15   824.4341   82.09268   77.88293                  0                             0 41315
42  1394.4341  199.09268  157.88293                  0                             0 36000
      yhat         e    e_star         r         d abs(t)-qt(1-alpha/(2*n),n-p-1)
15 30174.9 11140.097  4.534570  4.626579  11596.76                        1.145324
42 45878.1 -9878.096 -4.020873 -4.621435 -13049.27                        1.139118
```

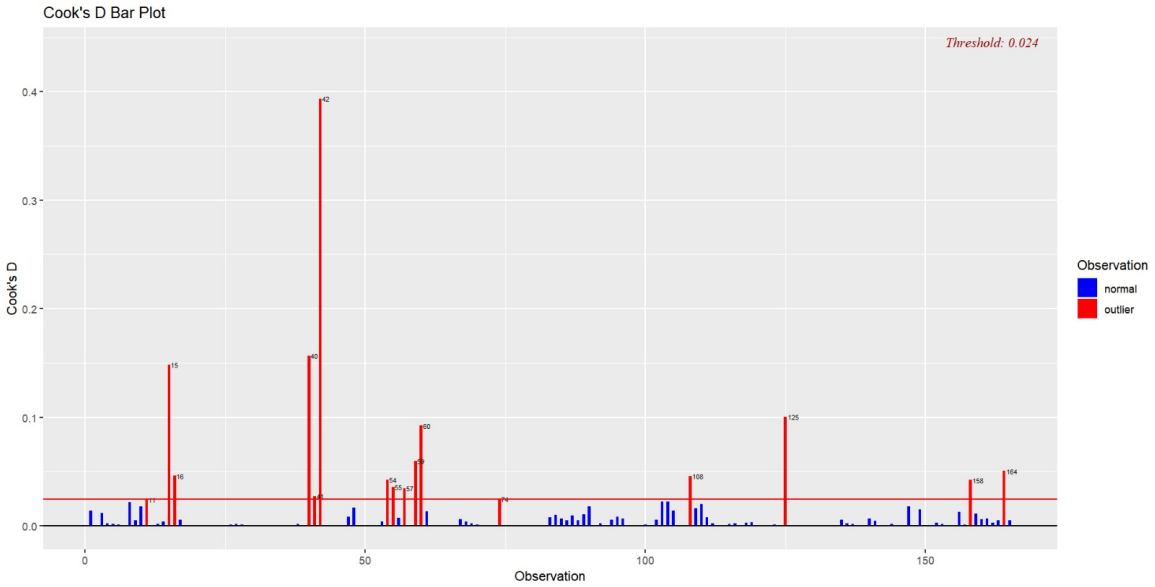
15, 42 are the outliers with respect to Y

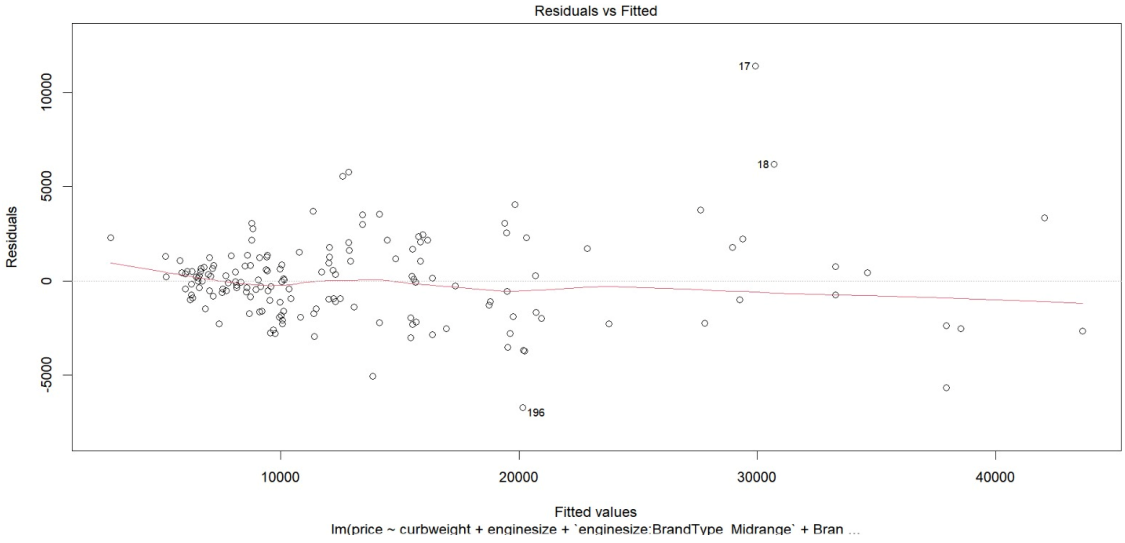
Outliers with respect to X:

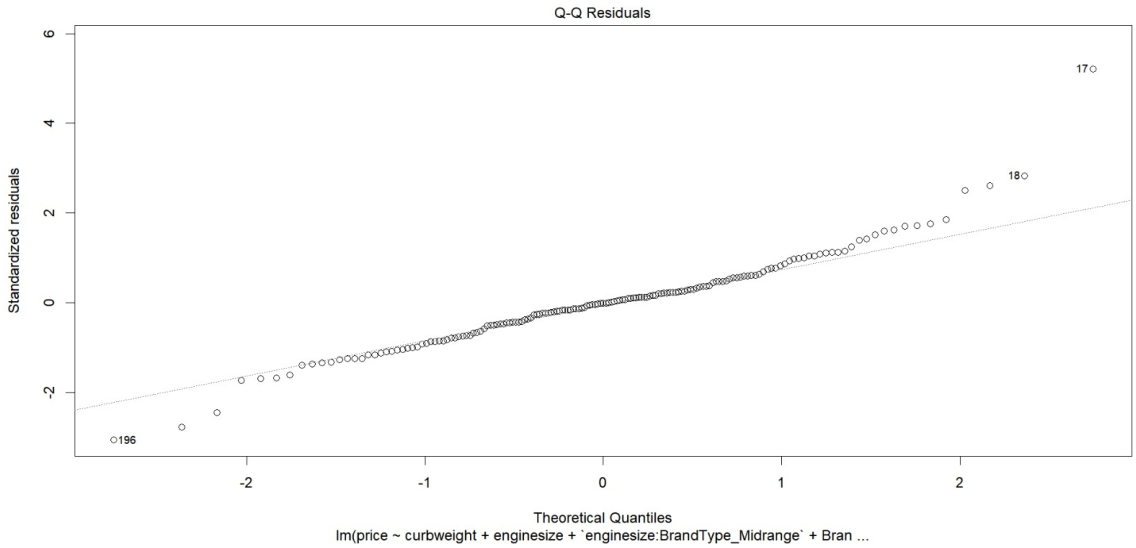
```
> LT[LT$"h_ii-2p/n">0,]
    curbweight enginesize horsepower BrandType_Midrange enginesize:BrandType_Midrange       Y       h_ii    h_ii-2p/n
4    -218.5659  -17.90732  -2.117073                  0                       0.00000 13950.0 0.07652520 9.207908e-05
40   1510.4341  131.09268  71.882927                  0                       0.00000 32250.0 0.10338384 2.695072e-02
41   1510.4341  131.09268  71.882927                  0                       0.00000 35550.0 0.10338384 2.695072e-02
42   1394.4341  199.09268 157.882927                  0                       0.00000 36000.0 0.24301513 1.665820e-01
47   -175.5659  -56.90732  -3.117073                  1                     -56.90732 10945.0 0.09118050 1.474738e-02
48   -175.5659  -56.90732  -3.117073                  1                     -56.90732 11845.0 0.09118050 1.474738e-02
59   1344.4341  181.09268  79.882927                  0                       0.00000 40960.0 0.19675446 1.203213e-01
60   1159.4341  177.09268  79.882927                  0                       0.00000 45400.0 0.19270827 1.162752e-01
103   200.4341   67.09268 102.882927                  0                       0.00000 32528.0 0.08896104 1.252791e-02
104   200.4341   67.09268 102.882927                  0                       0.00000 34028.0 0.08896104 1.252791e-02
105   810.4341   76.09268 183.882927                  0                       0.00000 31400.5 0.23024545 1.538123e-01
125   554.4341  -34.90732 -42.117073                  1                     -34.90732  8778.0 0.12034227 4.390915e-02
```

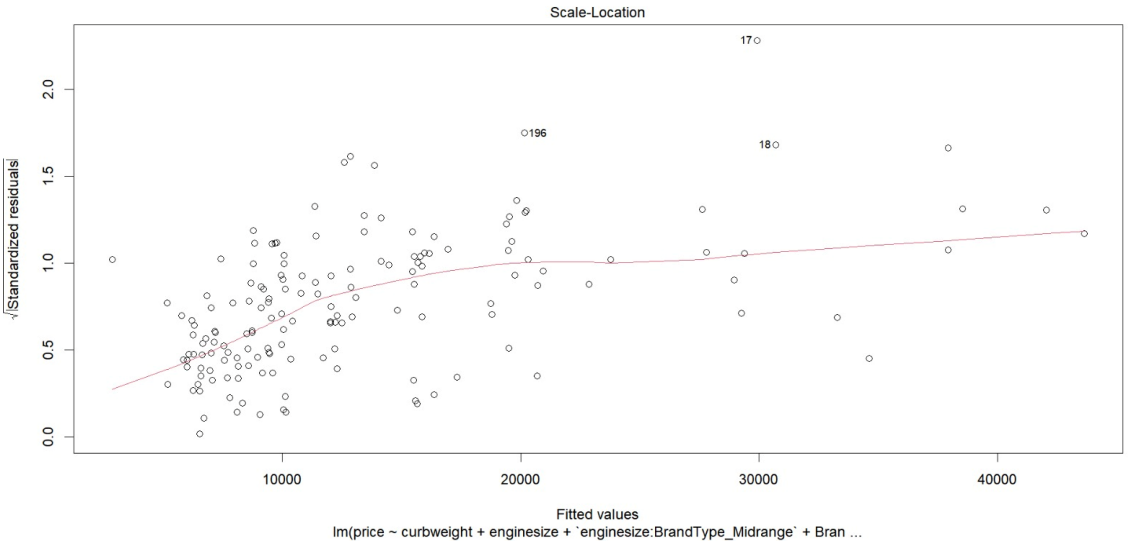
4, 40, 41, 42, 47, 48, 59, 60, 103, 104, 105, 125 are the outliers to X

Diagnostic Plots for Influential Cases:

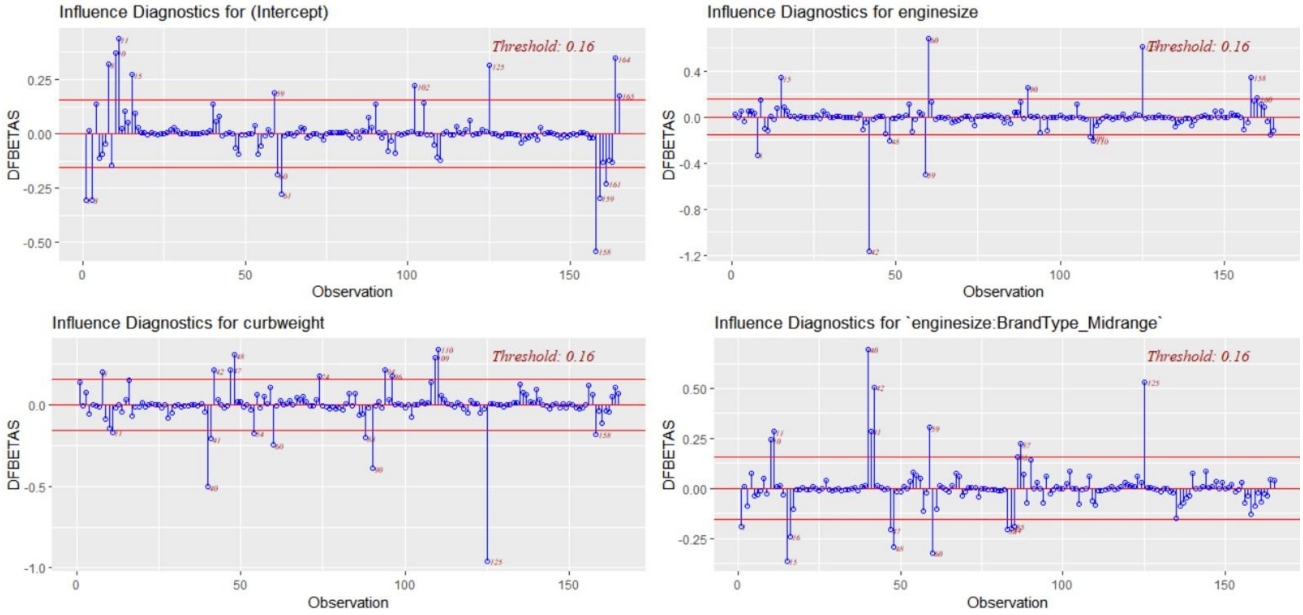




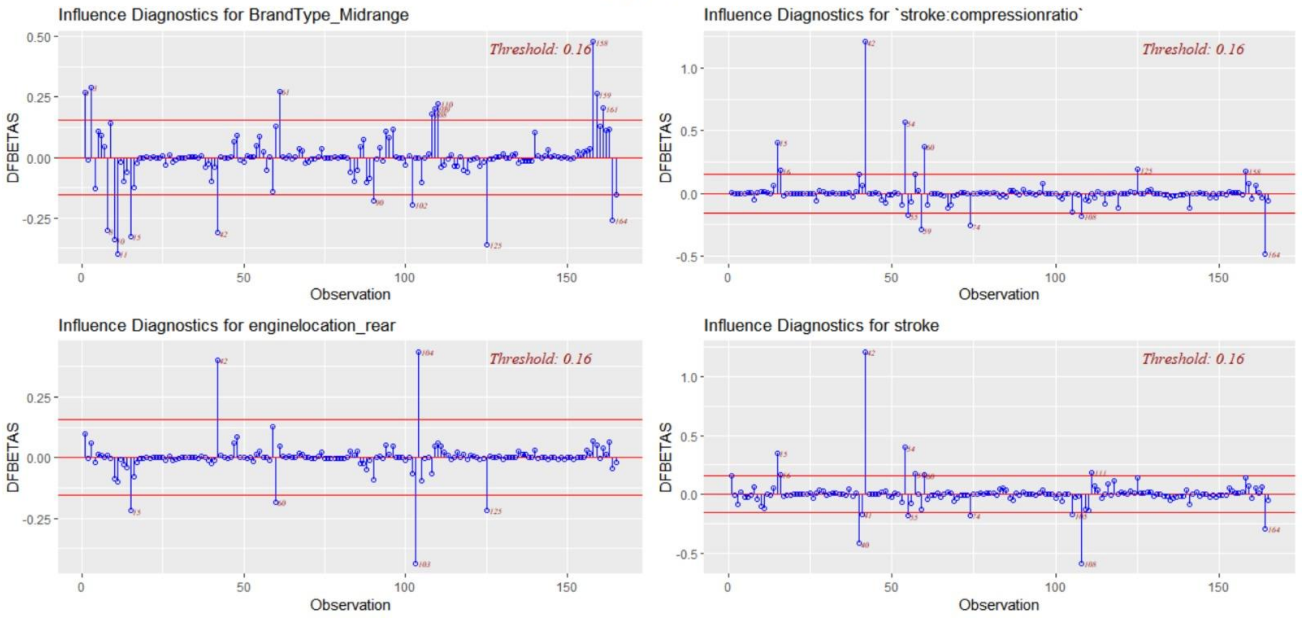

Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(price ~ curbweight + enginesize + `enginesize:BrandType_Midrange` + Bran ...
17
18
196


Scale-Location
√|Standardized residuals|
Fitted values
lm(price ~ curbweight + enginesize + `enginesize:BrandType_Midrange` + Bran ...
17
196
18

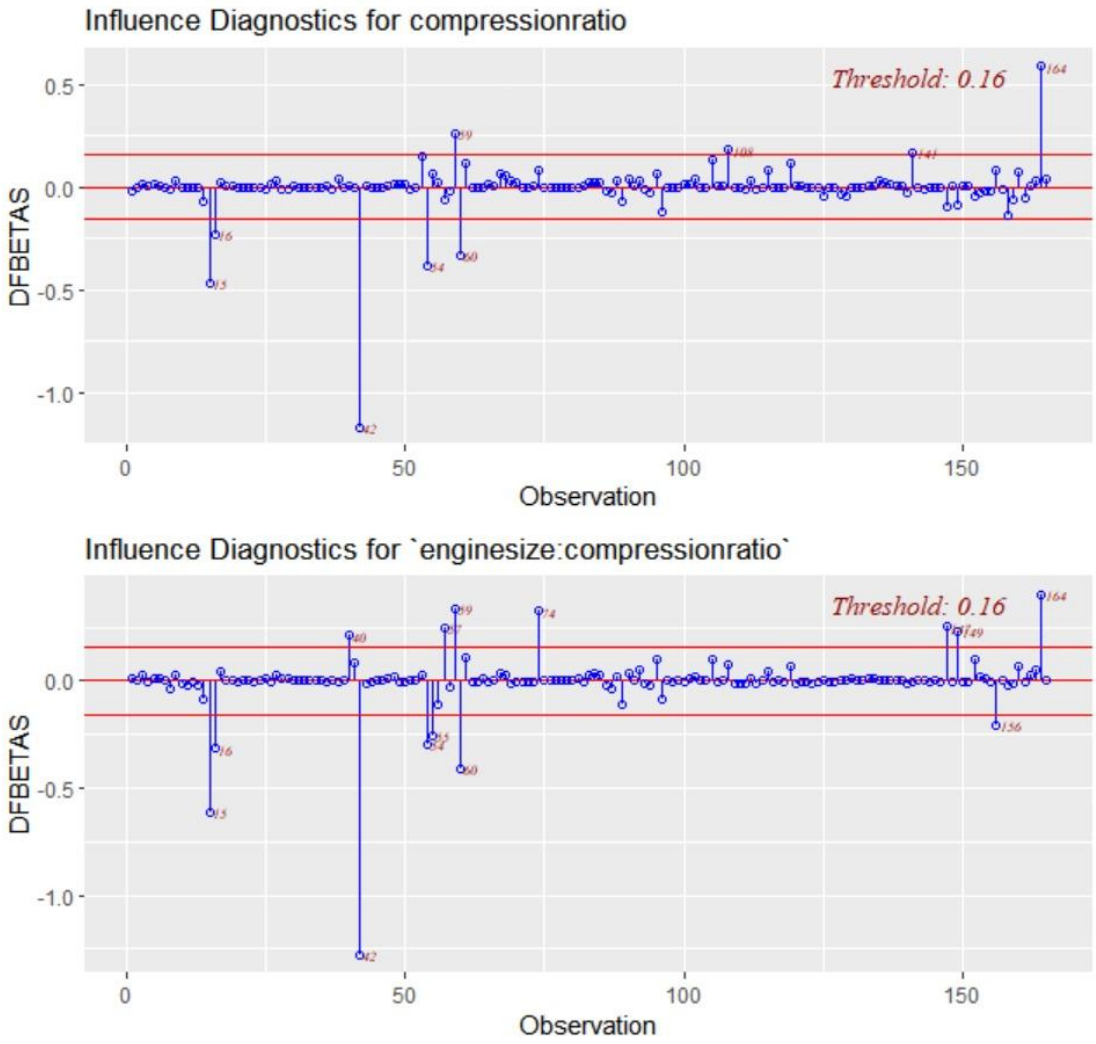
page 1 of 3



page 2 of 3



page 3 of 3



Next, we identified and removed influential outliers, specifically cases 10, 11, 40, 41, 42, 47, 48, and 125, based on their impact on both predictor variables (X) and the response variable (Y). These outliers were identified through thorough analysis of the plots mentioned earlier.

Summary of Final_fwd:

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   14608.8730   542.6713  26.920  < 2e-16 ***
curbweight                        7.9767     0.8861   9.002 1.05e-15 ***
enginesize                      109.2575    10.5418  10.364  < 2e-16 ***
`enginesize:BrandType_Midrange` -118.1932    13.3543  -8.851 2.55e-15 ***
BrandType_Midrange            -3673.0492   581.1987  -6.320 2.95e-09 ***
enginelocation_rear            9803.3674  1613.3535   6.076 1.01e-08 ***
`stroke:compressionratio`       715.9929   494.1924   1.449   0.1495
stroke                         1336.2387   882.1077   1.515   0.1320
compressionratio               -153.7624   106.8649  -1.439   0.1523
`enginesize:compressionratio`    -3.4606     1.8493  -1.871   0.0633 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2036 on 147 degrees of freedom
Multiple R-squared:  0.9364,    Adjusted R-squared:  0.9325
F-statistic: 240.5 on 9 and 147 DF,  p-value: < 2.2e-16
```

Statistical Analysis

F-test for Model:

Null Hypothesis (H0): There is no significant relationship between the predictor variables (such as engine size, brand type, cylinders, horsepower, etc.) and Car Prices.

Alternative Hypothesis (Ha): There is a significant relationship between the predictor variables and Car Price.

Decision rule: If F*>F, conclude alternative hypothesis (Ha), there is a significant relationship between predictor variables and car price. Otherwise, we conclude Null hypothesis(H0)

```
> #f value calculation
> n<-nrow(train_data_filtered)
> p<-10
> f_value<-qf((1-0.05),p-1,n-p)
> f_value
[1] 1.944096
```

Conclusion:

From the above summary F* = 240.5 and F value = 1.94.

Since F* > F, we conclude **alternate hypothesis (Ha) ,** there is a significant relationship between predictor variables and car price.

T-test for Beta values:

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   14608.8730   542.6713  26.920  < 2e-16 ***
curbweight                        7.9767     0.8861   9.002 1.05e-15 ***
enginesize                      109.2575    10.5418  10.364  < 2e-16 ***
`enginesize:BrandType_Midrange` -118.1932    13.3543  -8.851 2.55e-15 ***
BrandType_Midrange            -3673.0492   581.1987  -6.320 2.95e-09 ***
enginelocation_rear            9803.3674  1613.3535   6.076 1.01e-08 ***
`stroke:compressionratio`       715.9929   494.1924   1.449   0.1495
stroke                         1336.2387   882.1077   1.515   0.1320
compressionratio               -153.7624   106.8649  -1.439   0.1523
`enginesize:compressionratio`    -3.4606     1.8493  -1.871   0.0633 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2036 on 147 degrees of freedom
Multiple R-squared:  0.9364,    Adjusted R-squared:  0.9325
F-statistic: 240.5 on 9 and 147 DF,  p-value: < 2.2e-16

> tvalue<- qt(1-0.05/2,147)
> tvalue
[1] 1.976233
```

Null Hypothesis (H0): States that coefficient of Xk is zero

Alternative Hypothesis (Ha): States that coefficient of Xk is not zero

Decision rule: If |t*| > t, conclude alternative hypothesis (Ha). Otherwise, we conclude Null hypothesis (H0)

Conclusion: We opted to include predictor variables with |t*| values greater than the corresponding t-value. However, for variables such as stroke:compressionratio, stroke, compressionratio, and enginesize:compressionratio, the |t*| values fell below the t-value, indicating their insignificance. Consequently, we can accept the **Null hypothesis** for these variables.

Final Models

We arrived at three distinct models, each utilizing the same set of predictor variables but with different transformations applied to the response variable (Y).

Residual Analysis	final_forward0 (Y = price)	final_forward1 (Y = 1/price)	final_forward2 (Y = log10(price))
Normality (r-critical = 0.987**)**	**rEe = 0.951**	**rEe = 0.989**	**rEe = 0.993**
Adjusted r-squared	**0.9326**	**0.8679**	**0.9182**
Homoscedasticity (BP test**)**	**p-value < 2.2e-16**	**p-value = 0.232**	**p-value = 0.2797**

We encountered failures in both the normality test and Breusch-Pagan test for the final_forward0 model. Therefore, we have chosen to proceed with the other two models for further analysis.

Model Validation

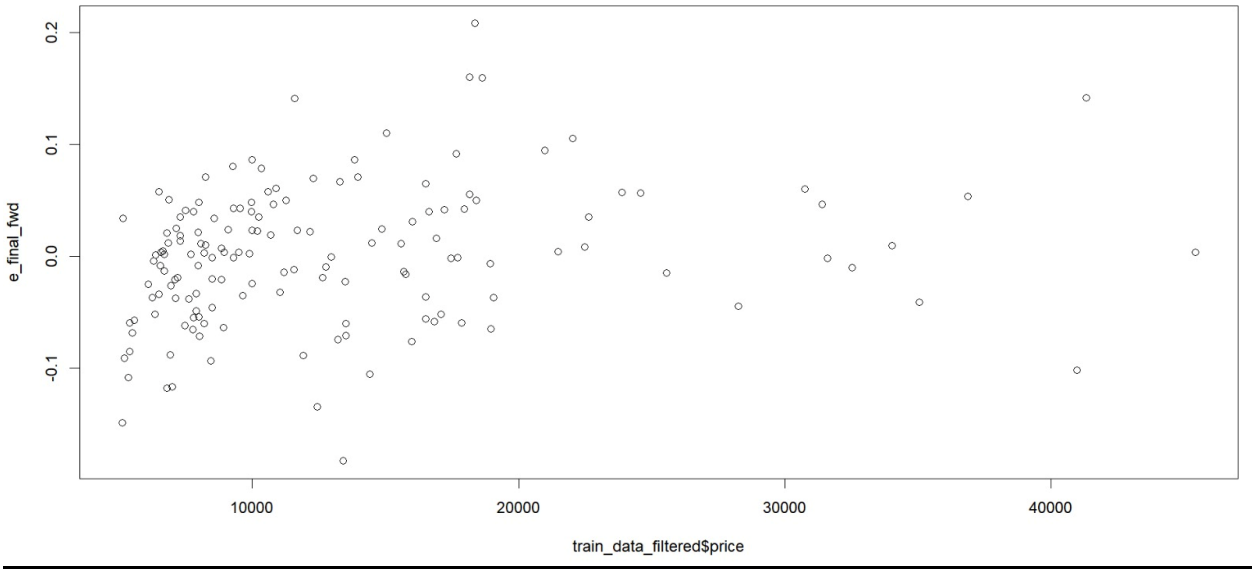
	Train	Test
Model 1	**Adjusted R-squared: 0.8679**	**Adjusted R-squared: 0.9182**
Model 2	**Adjusted R-squared: 0.9182**	**Adjusted R-squared: 0.8397**

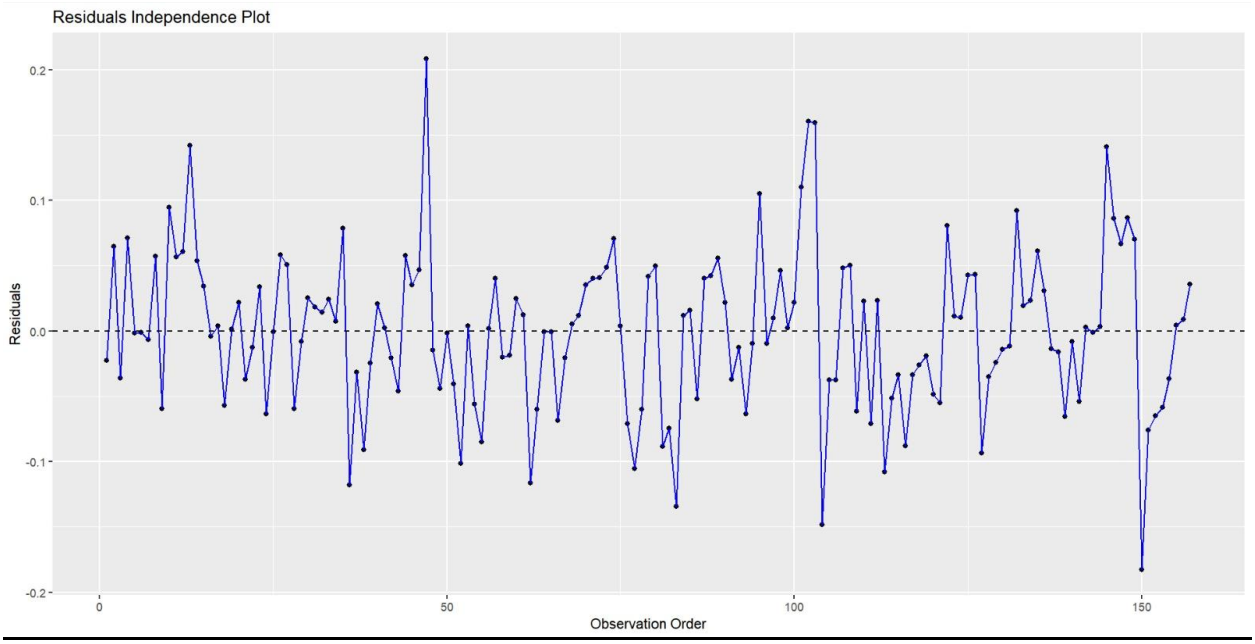
After comparing the difference in adjusted R-squared values between the train and test datasets for both models, we observed a smaller difference in the second model compared to the first. As a result, we have decided to proceed with model 2 due to its more consistent performance across both datasets.

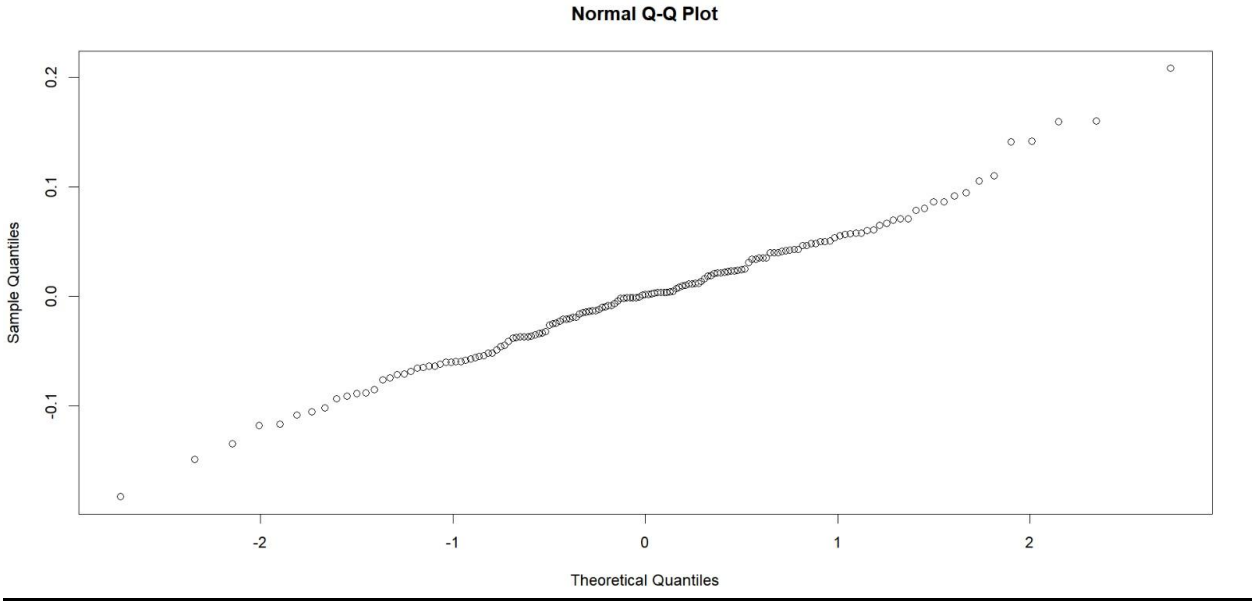
final_forward2<- lm(log10(price)~ curbweight+compressionratio+enginesize:compressionratio+enginesize+enginesize:BrandType_Midrange+BrandType_Midrange+enginelocation_rear, data = train_data_filtered)

```
> summary(final_forward2)

Call:
lm(formula = log10(price) ~ curbweight + compressionratio + `enginesize:compressionratio` +
    enginesize + `enginesize:BrandType_Midrange` + BrandType_Midrange +
    enginelocation_rear, data = train_data_filtered)

Residuals:
      Min        1Q    Median        3Q       Max
-0.182735 -0.037445  0.001717  0.040078  0.208533

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4.152e+00  1.659e-02 250.241  < 2e-16 ***
curbweight                     3.103e-04  2.592e-05  11.974  < 2e-16 ***
compressionratio              -3.498e-04  1.224e-03  -0.286   0.7754
`enginesize:compressionratio` -9.025e-05  3.921e-05  -2.302   0.0227 *
enginesize                     5.999e-04  3.073e-04   1.952   0.0528 .
`enginesize:BrandType_Midrange` -4.733e-04  3.816e-04  -1.240   0.2168
BrandType_Midrange            -1.314e-01  1.767e-02  -7.434 7.66e-12 ***
enginelocation_rear            2.631e-01  4.791e-02   5.492 1.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06242 on 149 degrees of freedom
Multiple R-squared:  0.9219,    Adjusted R-squared:  0.9182
F-statistic: 251.3 on 7 and 149 DF,  p-value: < 2.2e-16
```

Residual Analysis for Model 2







Accuracy:

R-squared -> 0.9219

Adjusted R-squared -> 0.9182

p-value -> 2.2e-16

Final Regression Model

log10(Y) = 4.15 + 0.0003 * curbweight – 0.0003 * compressionratio – 0.00009 * enginesize:compressionratio + 0.0005 * enginesize – 0.0004 * enginesize:BrandType_Midrange – 0.1 * BrandType_Midrange + 0.2 * enginelocation_rear

Conclusion:

In conclusion, our thorough analysis and model selection process led us to identify and refine three models. While the final_forward0 model faced challenges with normality and homoscedasticity tests, the other two models demonstrated acceptable performance across various evaluation metrics.

As a result, we have decided to focus our further analysis and interpretation on these two selected models. This approach ensured that our modeling efforts are anchored in statistically sound and reliable frameworks, providing meaningful insights for decision-making purposes.

We have identified several key variables that significantly impact car prices in the American market. These variables include engine size and location, curb weight, horsepower, compression ratio, and certain brand types.

The insights gained from our analysis have significant implications for business strategy and market penetration. By leveraging the identified influential variables, businesses can tailor their pricing strategies, product offerings, and marketing efforts to better align with consumer preferences and market demands. This strategic alignment can lead to improved market penetration, competitive advantage, and overall business performance in the American automotive market.