New Automobile Pricing

Minjiao Yang
Jan 5, 2020

Summary

The foreign car company wants to learn the factors that can affect car prices in the US. They obtain 215 records of 22 different brands of cars sold in the US with detailed information of their engine, figure and manufacturer's suggested retail price. In total 26 variables in the dataset, and with 6 duplicated records, 50 missing values, 4 mis-entered values, and 7 abnormal price values. After removing all of these data, we build the model on the remaining 145 data points. And by analyzing the residual of the model and carefully select the best model based on the AIC, BIC, RMSE and R^2 criteria, we obtain our final model with the lowest AIC, BIC, RMSE value and highest R^2 value. All variables in the model are significant and the model satisfies the linear regression assumption. So we find out the car price depends on CarCategory, Carbody, Enginelocation, Curbweight enginetype, fuelsystem and peakrpm.. And we can calculate the estimated car price though our final model. And from our model, we can see that Luxury cars are significantly expensive than in middle and economy cars. Convertible car is the most expensive ones, while hatchback is the cheapest one. Also, the engine located at the rear is significantly more expensive than in the front. The car with ohcv engine type and mpfi fuel system are the most expensive. Moreover, as the curbweight and stroke increase, car prices increase. Car prices increase. car price decrease with peakrpm increasing.

Table of Content

1. Introduction

2. Data Summary

3. EDA
    1. EDA for Response Variable
    2. EDA for Categorical Predictors
    3. EDA for Numerical Predictors

4. Modeling
    1. Methodology for Model Selection
    2. Model Selection

5. Conclusion

6. Further Discussion

7. Appendix
    1) Output Summary for full model (model 1), model 2 and model 3
    2) Residual Analysis of Model 2
    3) Boxcox plot of Car price (model 2)
    4) Output summary of model 4 and model 5
    5) Residual Analysis for Model 5
    6) Plots of Predicted Price vs. Factors

8. R code

1 Introduction

A foreign automobile company aspires to enter the US market by setting up their manufacturing unit here and producing cars locally to give competition to their US and European counterparts. They have contracted an automobile consulting company to learn the factors that have an effect on car prices. They want to know what factors affect car prices in the US market, and which variables are significant in predicting the price of the car.

2 Data summary

The dataset obtained by automobile consulting company is for 215 automobiles of different types currently sold in the US. For each automobile, detailed information of their engine, figure and manufacturer's suggested retail price were shown in the dataset. The 26 variables present in the original data-set are summarized in Table 1.

Variable Name	Description	Type	Levels
Car_ID	ID of each obs	Int	
Symboling	Insurance risk rating; higher is more risky	Continous	
CarName	Name of car model	Name	
fueltype	Car fuel type	Factor	gas, diesel
aspiration	Aspiration used in the car	Factor	std, turbo
doornumber	Number of doors on the car	Factor	four, two
carbody	Body of the car	Factor	convertible, hardtop, hatchback, sedan, wagon
drivewheel	Type of drive wheel	Factor	4wd, fwd, rwd
enginelocation	Location of the car engine	Factor	front, rear
wheelbase	The distance between front and rear axis of the car	Continous	
carlength	Length of the car	Continuous	
carwidth	Width of the car	Continuous	
carheight	Height of the car	Continuous	
curbweight	Weight of the car without occupants or baggage	Continuous	
enginetype	Type of engine	Factor	dohc, I, ohc, ohcf, ohcv, rotor
cylindernumber	Number of engine cylinders, more cylinders more power generated in less time	Factor	eight, five, four, six, three, two, twelve
enginesize	Engine capacity /how much power the car produce	Continuous	
fuelsystem	Fuel system of the car	Factor	1bbl, 2bbl, 4bbl, idi, mfi,mpfi, spdi, spfi
boreratio	Bore-to-stroke ratio of the cylinder	Continuous	
stroke	Storke length inside a cylinder	Continuous	
compressionratio	Volume ratio with the piston out:in	Continuous	
horsepower	Engine power	Continuous	
peakrpm	How many times engine's crank shaft make one full roation per minute	Continuous	
citympg	Fuel mileage per gallon in city driving	Continuous	
highwaympg	Fuel mileage per gallon in highway driving	Continuous	
price	Suggested retail price of the car	Continous	
215 observations (6 duplicates, 50 missing, 4 mis-entered, 7 abnormal price values)			

Table 1. Original Car Price Data Summary

Before conducting the analyses, the variable “Car_ID” are proved to be redundant and was thus discarded. Also, 41 records with a missing value of “Engine location”, 10 with the missing value of “Horsepower” and 4 with mis-entered value of “Engine type” were omitted as well as 6 abnormal values of car price and 6 duplicate records. In addition, there are some categorical variables with only 1 instance, such as 3 cylindernumber and 12 cylindernumber. The same problem can be found in “Fuelsysem” variable. So, we will remove those with number of instances less than 2 as they are not too useful for use

and can be considered as a noise. In total 70 records were removed from dataset. The variables that were retained were re-coded, the entires of "CarName" variable were categorized into 22 different brand and an additional variable "CarCategory" was created according to the mean prices of the brands with three levels "economy" for mean price less than 10k, "middle" for mean price between 10k and 20k, and "luxury" for mean price greater than 20k. The transformed data is given in Table 2.

Variable Name	Type	Levels
Symboling	Continous	
CarCategory	Factor	0 (middle), 1 (luxury), 2 (economy)
fueltype	Factor	0 (gas), 1(diesel)
aspiration	Factor	0 (std), 1 (turbo)
doornumber	Factor	0 (four), 1 (two)
carbody	Factor	0 (convertible), 1(hardtop), 2 (hatchback), 3 (sedan), 4 (wagon)
drivewheel	Factor	0 (4wd), 1 (fwd), 2 (rwd)
enginelocation	Factor	0 (front), 1 (rear)
wheelbase	Continous	
carlength	Continuous	
carwidth	Continuous	
carheight	Continuous	
curbweight	Continuous	
enginetype	Factor	0 (dohc), 1 (I), 2 (ohc), 3 (ohcf), 4 (ohcv), 5 (rotor)
cylindernumber	Factor	0 (two), 1 (four), 2 (five), 3 (six), 4 (eight)
enginesize	Continuous	
fuelsystem	Factor	0 (1bbl), 1 (2bbl), 2 (4bbl), 3 (idi), 4 (mpfi), 5 (spdi)
boreratio	Continuous	
stroke	Continuous	
compressionratio	Continuous	
horsepower	Continuous	
peakrpm	Continuous	
citympg	Continuous	
highwaympg	Continuous	
price	Continous	
145 observations		

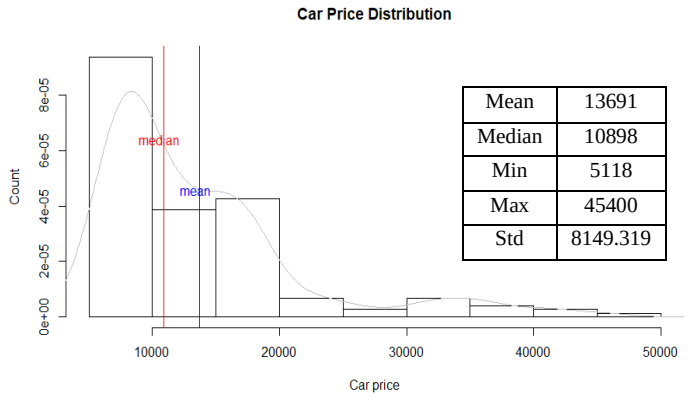
Table 2. Transformed Car Price Data Summary

3 Exploratory Data Analysis

Before modeling the data, we first visualize the response variable and its relationship with each predictor.

3.1 EDA for Response Variable

Since we're trying to predict car prices with historical data, a simple histogram plot of car price distribution is shown in figure 1.



Mean	13691
Median	10898
Min	5118
Max	45400
Std	8149.319

Figure 1. Car Price Distribution

We can see the plot is right-skewed with most car prices are below 15k. There is a significant difference between the mean and median of the price distribution. In addition, the data points are far spread out from the mean, which indicates a high variance in the car prices.

3.2 EDA for Categorical Predictors

Next, we will visualize the Categorical predictors and numerical predictors individually.

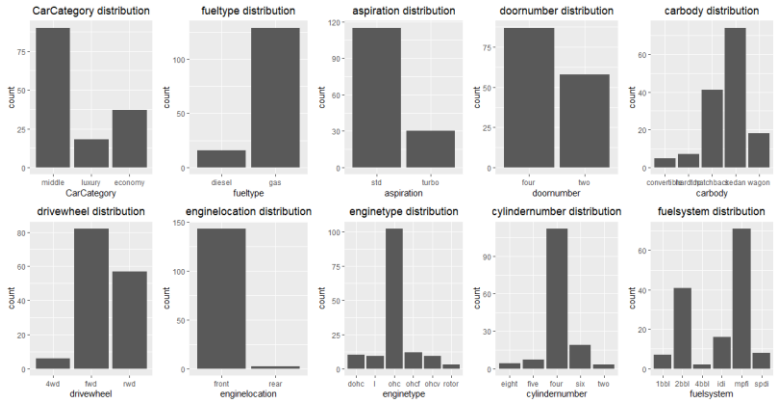


Figure 2. Categorical Predictors Distribution

In Figure 2, a barplot was performed to show the distribution of each Categorical Predictors. We can see middle type cars ($10k~$20k) are the most popular in the US, especially sedan and hatchback. Compare to diesel, gas is still the major fueltype for US car. std aspiration is more popular than turbo. And, people seem to prefer four-door car with fwd drivewheel. In addition, the number of the car with front-engine location, ohc engine type and four cylinder cars are sold more than others. And for fuelsystem, people prefer 2bbi and mpfi. To sum up, the most popular car in US is the sedan around 10k~20k (middle) with four door, fwd drivewheel, gas fueltype, mpfi fuelsystem, std aspiration, front ohc engine and four cylinder.

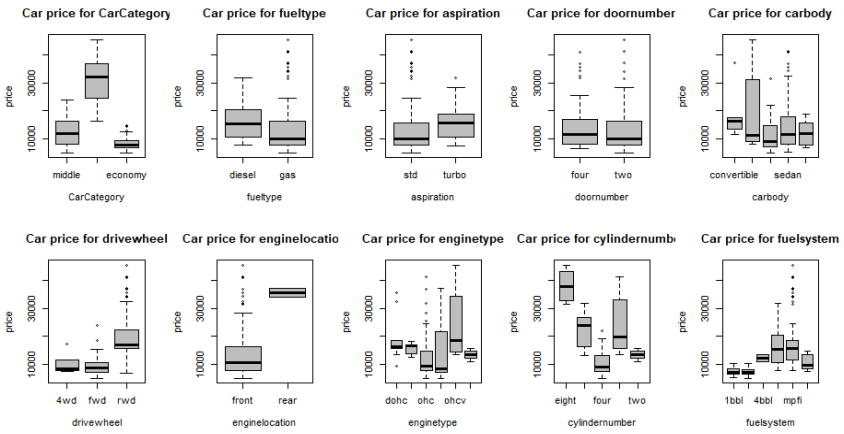


Figure 3. Categorical Predictors vs. Car Price

In Figure 3, a boxplot was performed to show the relationship between Categorical Predictors and Car Price. Start from CarCategory, the price of luxury cars (>20k) are significantly higher than the other two, Diesel seems significantly expensive than gas, turbo seems significantly expensive than std. Four-door and two-door cars are approximate at the same price. Convertible car is sold more expensive, while the hardtop car has highest price range. Rwd cars are significantly more expensive, and the rear engine car is obviously sold for a higher price. The car price for ohcv engine is higher than others, and it has the highest price range. Eight cylinder car is significantly the most expensive car, while six cylinder has the highest price range. Price for mpfi fuelsystem car is higher, and with few outliers at higher price. So from Figure 3, seems that CarCategory, Fueltype,aspiration, carbody, drivewheel, enginelocation, enginetype and clindernumber will affect car price.

3.3 EDA for Numerical Predictors

Then, we move on to Numerical predictors. In Figures 5 and 6, we made scatterplots for Numerical predictors to investigate their relationship with Car price.

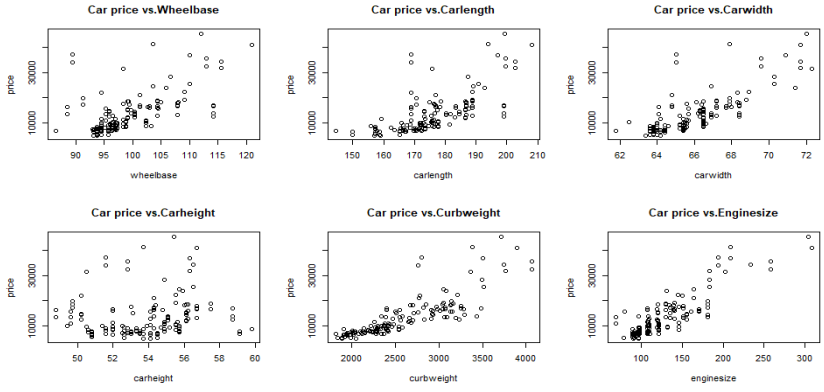


Figure 5. Numerical Predictors vs. Car Price

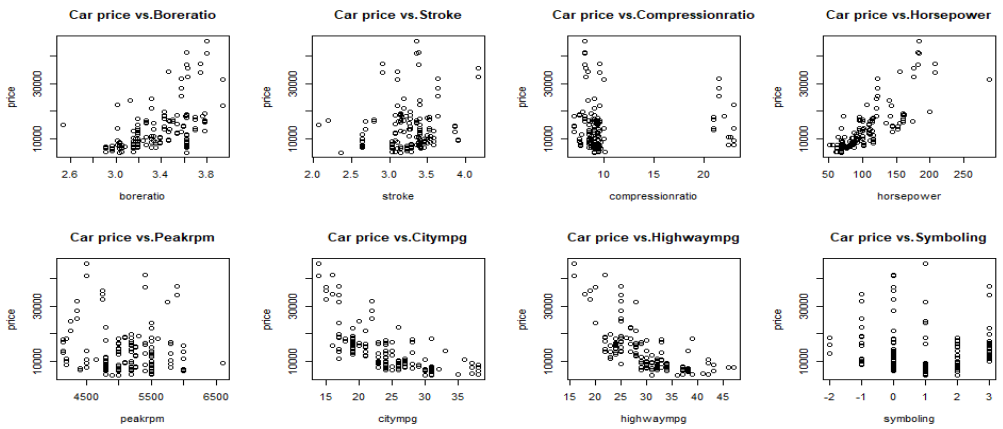


Figure 6. Numerical Predictors vs. Car Price

We can see carwidth, car length and curbweight seems to have a positive correlation with car price, carheight doesn't show any significant trend with price. Enginesize, boreratio, horsepower, and wheelbase seem to have significant positive correlation with price, citympg and high waympg seem to have a significant negative correlation with price.

For further investigation of the correlation between numerical predictors and car price, Pearson correlation test was performed, and the result was shown in figure 7.

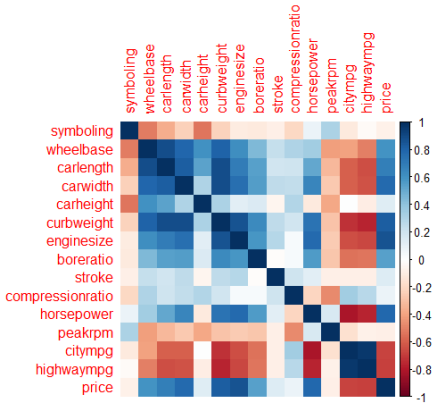


Figure 7. Correlation plot of Numerical Predictors and Car Price

We can see that wheelbase, carlength,carwidth,curbweight,enginesize, horsepower,citympg and highwaympg are highly correlated with car price. So I believe these numerical predictors will affect car prices.

4 Car Price Analysis

4.1 Methodology for Model Selection

To prove my conjecture, we model the linear regression using price as the response variable. But linear regression is a parametric model that follows some assumptions. Linear regression that doesn't follow the assumptions may be misleading. So when we build our linear model, we need to verify that:

- The relationship between predictors and response is linear, which is graphically investigated by residual plots. And if there's a pattern in the residual vs.fitted value plots, the model does not meet the linearity assumption.
- The residuals of the linear regression model follow the normal distribution. This assumption is important since the violation would result in inaccurate estimation and inference. This is checked by Sphiro-Wilknormality test, with null hypothesis that residuals are normally distributed.
- Heterocedasitcity means that the variances of the error terms are non-constant. One can identify non-constant variances in the errors from the presence of a funnel shape in the residual plot.
- Multicollinearity indicate that there is a correlation between the predictors. This can be checked by measuring the cariance inflation factor (VIF) with a VIF value exceeds 5 or 10 indicates collinearity.

Then for model selection, we based on the following criterions.

- All variables in the model are significant at the 0.05 level. Otherwise, the model is reduced by eliminating insignificant variables.
- Model with lower RMSE (root mean square error) and higher R^2 is preferred. RMSE is a measure of how speard out these reiduals are, and R^2 represents the proportion of the variance for a dependent variable that's explained by variables in the regression model. Both are used for measuring the goodness-of-fit of the model.
- The model with lower Akaike formation criterion (AIC) or Schwardz-Bayesian information Criterion (BIC) is selected because AIC and BIC consider both goodness-of-fit of the model and parsimony by :

$$AIC = -2nlog\hat{L} + 2p$$
$$BIC = -2nlog\hat{L} + plogn$$

Where p is the number of parameters in the model, n is the sample size, and $\hat{L}$ is the likelihood of the model with estimated parameter.

4.2 Model Selection

We start with building the full model by including all predictors. And we use backward selection to select the model (model 2). Then, we eliminate all the variables that are not significant in Model 2 and obtain a simpler model (Model 3) shown in table 2.

Model	Predictors	RMSE	R^2	AIC	BIC
1	All	2068	0.9526	2660	2779
2	CarCategory, aspiration, carbody, enginelocation, carwidth, curbweight, enginetype, cylindernumber, enginesize, fuelsystem, boreratio, stroke, compressionratio, peakrpm	2015	0.9507	2646	2735
3	CarCategory, aspiration, carbody, enginelocation, carwidth, curbweight, enginetype, cylindernumber, enginesize, fuelsystem, stroke, compressionratio, peakrpm	2034	0.949	2647	2734

Table 3. Model Comparison

Among the two models, model 2 is preferred with RMSE (2015), AIC (2646) and higher R-square(0.9507) value. Then we need to verify if our model 2 satisfies the assumption of linear regression.

In the residual plot (shown in Appendix Figure 11), there is a pattern in the data with the residuals has become more negative as the fitted values increase before increased again. The pattern indicates that our model may not be linear enough. Then we run the Saphrio-Wilk normality test to check the normality of

residual (shown in Appendix Table 4). With p-value < 0.05, we reject the null hypothesis that the residuals follow a normal distribution. Next, we check the Autocorrelation of standard errors by running the Durbin-Wattson Test (shown in Appendix Table 4). Then for the Heteroscedasticity, from our residual plot, we can see observe that on lower fitted values, the residuals are concentrated around the value of 0. As the fitted value increases, the residuals are also got bigger. So, the heteroscedasticity is present in model 2. In the end, we check for multicollinearity. when measuring the VIF, we got a warning message that there're aliased coefficients in the model. Therefore, there's collinearity exists in the model.

Since our model 2 does not meet the assumptions of linear regression, we will try to fix them. To make the model more linear, we can transform some of the variables. We first will take out the variables that have a strong correlation with other variables. Based on the correlation plot (Figure 7), engine size is highly correlated with other variables. Moreover, the residual plot of model 2 indicating that our model is not linear, this suggesting a transformation is needed for our variables. Boxcox plot (Figure 12 in Appendix) shows we need log transformation on our variables. After running our improved model and use backward selection to select the variable, we obtain our model (model 4) for Car price. Again, we discard the non-significant term and obtain a better model model 5, shown in Table 5.

Model	Predictors	RMSE	R^2	AIC	BIC
4	CarCategory, carbody, enginelocation, carwidth, curbweight, enginetype, cylindernumber, fuelsystem, stroke, compressionratio, peakrpm	0.05847	0.9403	-387	-306
5	CarCategory, carbody, enginelocation, carwidth, curbweight, enginetype, fuelsystem, stroke, compressionratio, peakrpm	0.05898	0.9378	-386	-314
final model	CarCategory, carbody, enginelocation, curbweight, enginetype, fuelsystem, stroke, peakrpm	0.06037	0.9337	-381	-316

Table 5. Model Comparison

Although, model 4 have some advantage in RMSE (smaller), R-square (larger), AIC (larger). But It contains significant term "cylindernumber" and cylindernumber is liaised with another variable in the model. Then we analyze the residual of model 5 and discover that in the VIF value of compressionratio is too big (see figure 16 in Appendix), which indicate that compressionratio is strongly correlated with other variable in the model. So, we must eliminate it. After we run the improved model, we discard the non-significant term we obtain our final model:

$$\log_{10}(Car\ price)_{ijklm} = \beta_0 + \beta_1 \times I(CarCategory_i) + \beta_2 \times I(carbody_j) + \beta_3 \times I(enginelocation_k) + \beta_4 \times curbweight + \beta_5 \times I(enginetype_l) + \beta_6 \times I(fuelsystem_m) + \beta_7 \times stroke + \beta_8 \times peakrpm + \varepsilon_{ijklm}$$

Where, $i = 0,1,2;\ j = 0,1,2,3,4; k = 1,2;\ l = 1,2,3,4,5,6;\ m = 1,2,3,4,5,6$. Baseline is Carcategory=middle, carbody=convertible, enginelocation=front, enginetype=dohc,, and fuelsystem=1bbl. Output summary is shown in figure 9.

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.145e+00  1.417e-01  22.198  < 2e-16 ***
factor(CarCategory)luxury  1.297e-01  2.188e-02   5.931 2.79e-08 ***
factor(CarCategory)economy -3.992e-02  1.929e-02  -2.070 0.040520 *
factor(carbody)hardtop     -8.683e-02  3.764e-02  -2.307 0.022736 *
factor(carbody)hatchback   -8.711e-02  3.247e-02  -2.683 0.008298 **
factor(carbody)sedan       -6.786e-02  3.167e-02  -2.143 0.034060 *
factor(carbody)wagon       -1.213e-01  3.474e-02  -3.491 0.000668 ***
factor(enginelocation)rear  1.835e-01  5.999e-02   3.058 0.002729 **
curbweight                  3.323e-04  2.167e-05  15.338  < 2e-16 ***
factor(enginetype)l        -3.273e-02  3.207e-02  -1.021 0.309463
factor(enginetype)ohc       5.572e-02  2.338e-02   2.383 0.018681 *
factor(enginetype)ohcf      4.037e-02  3.814e-02   1.059 0.291873
factor(enginetype)ohcv      3.174e-03  2.922e-02   0.109 0.913690
factor(enginetype)rotor     1.461e-01  6.570e-02   2.223 0.028005 *
factor(fuelsystem)2bbl     -2.559e-02  3.157e-02  -0.811 0.419104
factor(fuelsystem)4bbl     -3.020e-02  8.019e-02  -0.377 0.707092
factor(fuelsystem)idi       3.581e-02  3.926e-02   0.912 0.363456
factor(fuelsystem)mpfi      3.803e-02  3.144e-02   1.210 0.228691
factor(fuelsystem)spdi      3.072e-02  3.306e-02   0.929 0.354649
stroke                     -4.738e-02  2.291e-02  -2.068 0.040672 *
peakrpm                     4.605e-05  1.662e-05   2.770 0.006460 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06037 on 124 degrees of freedom
Multiple R-squared:  0.9337,    Adjusted R-squared:  0.9231
F-statistic: 87.37 on 20 and 124 DF,  p-value: < 2.2e-16
```

Figure 17. Output Summary for final model

In the residual plot, there's no obvious pattern and heteroscedasticity is not presented in our final model. Also, points in the qqnorm plot fall in the straight line. Then for Saphrio-Wilk normality test (see table 5 in Appendix). the p-value is greater than 0.05, and we conclude that the residuals follow the normal distribution. In the end, VIF values for variables indicating that there's no collinearity between variables.

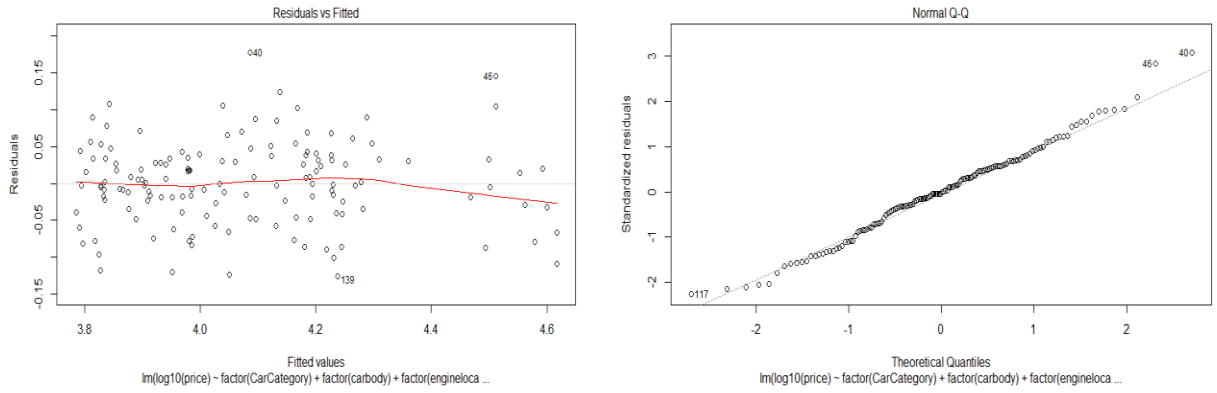


Figure 18. Residual Plots for final model

```
                            GVIF Df GVIF^(1/(2*Df))
factor(CarCategory)     5.536159  2        1.533918
factor(carbody)         2.329323  4        1.111485
factor(enginelocation)  1.947702  1        1.395601
curbweight              4.917982  1        2.217652
factor(enginetype)     24.987406  5        1.379660
factor(fuelsystem)     26.467874  5        1.387624
stroke                  2.167652  1        1.472295
peakrpm                 2.556491  1        1.598903
```

Figure 19. VIF for final model

5 Conclusion

In conclusion, we find out the car price depends on CarCategory, Carbody, Enginelocation, Curbweight enginetype, fuelsystem, and peakrpm. And we can use our final model to get estimated car price (since the response variable is log-transformed, we need to transform it back by taking the anti-log base to obtain the real price). Luxury cars are significantly expensive than in middle and economy cars. Convertible car is the most expensive ones, while hatchback is the cheapest one. Also, the engine located at the rear is significantly more expensive than in the front. The car with ohcv engine type and mpfi fuel system are the most expensive. Moreover, as the curbweight and stroke increase, car prices increase. car price decrease with peakrpm increasing.

6 Further Discussion

Although those missing, duplicate, mis-entered data take up 32% of the original dataset. But they occur at random, so analyzing the data without those observations is legitimate.

When we use training data for the final model and test the model with test data, we obtain RMSE of 3630 for test data and 1847 for trained data, and we conclude that our model seems over fit and we can try to adjust the proportion training- test dataset and see if the model overfit problem is improved.

7 Appendix

1) Output Summary for full model (model 1), model 2 and model 3

```
Coefficients: (2 not defined because of singularities)
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.051e+04  2.052e+04  -0.512 0.609692
symboling          2.199e+02  2.569e+02   0.856 0.393909
fueltypegas       -1.904e+04  8.239e+03  -2.311 0.022753 *
aspirationturbo    1.477e+03  1.054e+03   1.401 0.164016
doornumbertwo     -6.746e+01  7.773e+02  -0.087 0.930996
carbodyhardtop    -3.137e+03  1.445e+03  -2.171 0.032154 *
carbodyhatchback  -3.885e+03  1.301e+03  -2.986 0.003508 **
carbodysedan      -2.592e+03  1.490e+03  -1.739 0.084893 .
carbodywagon      -3.370e+03  1.643e+03  -2.050 0.042800 *
drivewheelfwd     -9.913e+02  1.191e+03  -0.832 0.407111
drivewheelrwd     -9.664e+02  1.464e+03  -0.660 0.510749
enginelocationrear 8.369e+03  2.701e+03   3.098 0.002494 **
wheelbase          1.891e+01  1.074e+02   0.176 0.860561
carlength         -2.861e+01  5.977e+01  -0.479 0.633168
carwidth           7.704e+02  2.816e+02   2.736 0.007289 **
carheight         -1.291e+02  1.666e+02  -0.775 0.440009
curbweight         5.632e+00  2.437e+00   2.311 0.022771 *
enginetypel       -1.255e+03  1.680e+03  -0.747 0.456445
enginetypeohc      2.897e+03  1.029e+03   2.815 0.005821 **
enginetypeohcf     1.483e+03  1.967e+03   0.754 0.452703
enginetypeohcv    -3.378e+03  1.462e+03  -2.310 0.022825 *
enginetyperotor    4.102e+03  4.305e+03   0.953 0.342741
cylindernumberfive -6.800e+03  2.967e+03  -2.292 0.023902 *
cylindernumberfour -5.419e+03  3.215e+03  -1.685 0.094868 .
cylindernumbersix  -3.376e+03  2.083e+03  -1.621 0.108008
cylindernumbertwo          NA         NA      NA       NA
enginesize         9.101e+01  2.844e+01   3.200 0.001814 **
fuelsystem2bbl    -9.119e+02  1.168e+03  -0.781 0.436634
fuelsystem4bbl    -3.302e+03  2.882e+03  -1.146 0.254473
fuelsystemidi             NA         NA      NA       NA
fuelsystemmpfi    -7.540e+02  1.219e+03  -0.619 0.537560
fuelsystemspdi    -3.118e+03  1.516e+03  -2.057 0.042154 *
boreratio         -1.583e+03  1.891e+03  -0.837 0.404376
stroke            -3.673e+03  1.041e+03  -3.528 0.000620 ***
compressionratio  -1.470e+03  6.038e+02  -2.435 0.016558 *
horsepower        -2.087e+01  2.326e+01  -0.897 0.371666
peakrpm            2.719e+00  8.001e-01   3.398 0.000956 ***
citympg            5.995e+01  1.823e+02   0.329 0.742870
highwaympg         6.973e+00  1.546e+02   0.045 0.964114
CarCategoryluxury  5.349e+03  9.966e+02   5.367 4.74e-07 ***
CarCategoryeconomy -1.538e+03  8.294e+02  -1.854 0.066516 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2068 on 106 degrees of freedom
Multiple R-squared:  0.9526,    Adjusted R-squared:  0.9356
F-statistic: 56.05 on 38 and 106 DF,  p-value: < 2.2e-16
```

Figure 8. Output summary of Model 1

```
Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -2.909e+04  1.412e+04  -2.061 0.041525 *
aspirationturbo     1.158e+03  7.769e+02   1.490 0.138852
carbodyhardtop     -3.509e+03  1.301e+03  -2.696 0.008061 **
carbodyhatchback   -4.204e+03  1.177e+03  -3.572 0.000518 ***
carbodysedan       -3.272e+03  1.127e+03  -2.903 0.004424 **
carbodywagon       -4.144e+03  1.239e+03  -3.344 0.001111 **
enginelocationrear  6.769e+03  2.402e+03   2.818 0.005679 **
carwidth            6.675e+02  2.431e+02   2.746 0.006995 **
curbweight          3.790e+00  1.652e+00   2.295 0.023552 *
enginetypel        -6.205e+02  1.450e+03  -0.428 0.669508
enginetypeohc       2.621e+03  9.501e+02   2.758 0.006751 **
enginetypeohcf      1.748e+03  1.681e+03   1.040 0.300393
enginetypeohcv     -2.978e+03  1.185e+03  -2.513 0.013347 *
enginetyperotor     5.937e+03  3.835e+03   1.548 0.124371
cylindernumberfive -4.782e+03  2.312e+03  -2.069 0.040805 *
cylindernumberfour -3.464e+03  2.589e+03  -1.338 0.183612
cylindernumbersix  -2.359e+03  1.820e+03  -1.297 0.197374
cylindernumbertwo          NA         NA      NA       NA
enginesize          9.702e+01  2.486e+01   3.902 0.000160 ***
fuelsystem2bbl     -4.696e+02  1.075e+03  -0.437 0.663112
fuelsystem4bbl     -2.490e+03  2.712e+03  -0.918 0.360467
fuelsystemidi       1.733e+04  7.350e+03   2.358 0.020060 *
fuelsystemmpfi     -5.551e+02  1.083e+03  -0.513 0.609210
fuelsystemspdi     -2.735e+03  1.281e+03  -2.135 0.034836 *
boreratio          -2.640e+03  1.486e+03  -1.776 0.078302 .
stroke             -3.458e+03  9.206e+02  -3.756 0.000271 ***
compressionratio   -1.251e+03  5.171e+02  -2.419 0.017098 *
peakrpm             2.380e+00  6.218e-01   3.827 0.000211 ***
CarCategoryluxury   5.783e+03  9.050e+02   6.390 3.58e-09 ***
CarCategoryeconomy -1.155e+03  7.025e+02  -1.644 0.102873
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2015 on 116 degrees of freedom
Multiple R-squared:  0.9507,    Adjusted R-squared:  0.9388
F-statistic: 79.94 on 28 and 116 DF,  p-value: < 2.2e-16
```

Figure 9. Output summary of Model 2

```
Coefficients: (1 not defined because of singularities)
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                -3.073e+04  1.421e+04  -2.162 0.032647 *
factor(CarCategory)luxury   5.603e+03  9.075e+02   6.174 9.93e-09 ***
factor(CarCategory)economy -8.022e+02  6.800e+02  -1.180 0.240491
factor(aspiration)turbo     1.062e+03  7.821e+02   1.358 0.177002
factor(carbody)hardtop     -3.491e+03  1.313e+03  -2.658 0.008951 **
factor(carbody)hatchback   -4.167e+03  1.188e+03  -3.509 0.000640 ***
factor(carbody)sedan       -3.211e+03  1.137e+03  -2.824 0.005576 **
factor(carbody)wagon       -4.035e+03  1.249e+03  -3.231 0.001603 **
factor(enginelocation)rear  7.514e+03  2.387e+03   3.148 0.002085 **
carwidth                    6.809e+02  2.452e+02   2.777 0.006391 **
curbweight                  3.213e+00  1.634e+00   1.966 0.051648 .
factor(enginetype)l        -9.502e+02  1.451e+03  -0.655 0.513918
factor(enginetype)ohc       2.410e+03  9.513e+02   2.534 0.012605 *
factor(enginetype)ohcf      2.741e+01  1.386e+03   0.020 0.984258
factor(enginetype)ohcv     -3.375e+03  1.174e+03  -2.874 0.004813 **
factor(enginetype)rotor     2.928e+03  3.473e+03   0.843 0.400822
factor(cylindernumber)five -6.306e+03  2.166e+03  -2.912 0.004306 **
factor(cylindernumber)four -5.986e+03  2.185e+03  -2.739 0.007119 **
factor(cylindernumber)six  -3.400e+03  1.738e+03  -1.956 0.052850 .
factor(cylindernumber)two          NA         NA      NA       NA
enginesize                  8.145e+01  2.348e+01   3.469 0.000732 ***
factor(fuelsystem)2bbl     -7.073e+02  1.077e+03  -0.657 0.512544
factor(fuelsystem)4bbl     -2.971e+03  2.723e+03  -1.091 0.277458
factor(fuelsystem)idi       1.819e+04  7.401e+03   2.458 0.015427 *
factor(fuelsystem)mpfi     -8.130e+02  1.083e+03  -0.751 0.454358
factor(fuelsystem)spdi     -2.900e+03  1.289e+03  -2.250 0.026331 *
stroke                     -3.661e+03  9.219e+02  -3.971 0.000124 ***
compressionratio           -1.315e+03  5.205e+02  -2.526 0.012866 *
peakrpm                     2.237e+00  6.223e-01   3.595 0.000476 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2034 on 117 degrees of freedom
Multiple R-squared:  0.9494,    Adjusted R-squared:  0.9377
F-statistic: 81.29 on 27 and 117 DF,  p-value: < 2.2e-16
```

Figure 10. Output summary of Model 3

2) Residual Analysis of Model 2

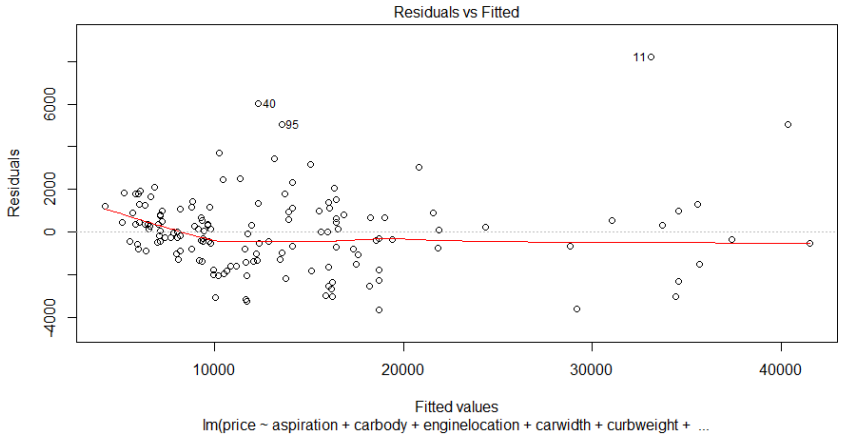


Figure 11. Residual vs. Fitted Plot for Model 2

SHAPIRO-WILK NORMALITY TEST	
W=0.94343	p-value=1.342e-05

Table 4. Shaprio-Wilk Normality Test for Residual of Model 2

3) Boxcox plot of Car price (model 2)

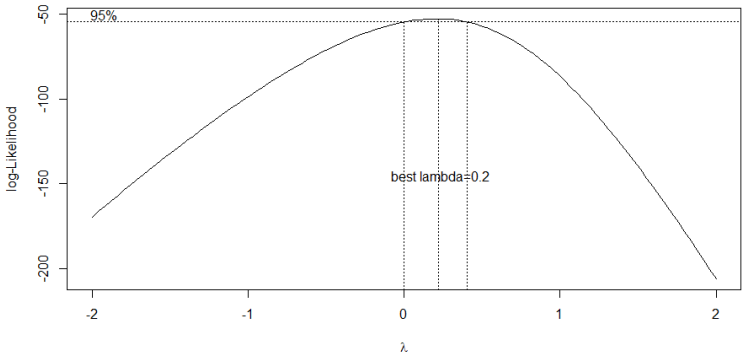


Figure 12. boxcox plot of Car price (model 2)

4) Output summary of model 4 and model 5

```
Coefficients: (1 not defined because of singularities)
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.866e+00  4.069e-01   7.044 1.30e-10 ***
factor(CarCategory)luxury   1.034e-01  2.416e-02   4.278 3.82e-05 ***
factor(CarCategory)economy -4.064e-02  1.952e-02  -2.082 0.039485 *
factor(carbody)hardtop     -9.514e-02  3.702e-02  -2.570 0.011397 *
factor(carbody)hatchback   -1.169e-01  3.401e-02  -3.436 0.000814 ***
factor(carbody)sedan       -9.082e-02  3.245e-02  -2.799 0.005985 **
factor(carbody)wagon       -1.327e-01  3.465e-02  -3.829 0.000207 ***
factor(enginelocation)rear  1.964e-01  6.401e-02   3.069 0.002660 **
carwidth                    1.420e-02  6.913e-03   2.054 0.042153 *
curbweight                  2.548e-04  3.456e-05   7.374 2.42e-11 ***
factor(enginetype)l        -5.893e-02  3.762e-02  -1.566 0.119911
factor(enginetype)ohc       5.150e-02  2.632e-02   1.956 0.052769 .
factor(enginetype)ohcf      1.813e-02  3.968e-02   0.457 0.648563
factor(enginetype)ohcv     -4.444e-02  3.273e-02  -1.358 0.177148
factor(enginetype)rotor     7.382e-02  8.059e-02   0.916 0.361465
factor(cylindernumber)five -9.876e-02  5.332e-02  -1.852 0.066503 .
factor(cylindernumber)four -9.209e-02  5.522e-02  -1.668 0.097995 .
factor(cylindernumber)six  -3.464e-02  4.506e-02  -0.769 0.443639
factor(cylindernumber)two          NA         NA      NA       NA
factor(fuelsystem)2bbl     -3.620e-02  3.093e-02  -1.170 0.244278
factor(fuelsystem)4bbl     -5.085e-02  7.819e-02  -0.650 0.516774
factor(fuelsystem)idi       4.705e-01  1.718e-01   2.740 0.007099 **
factor(fuelsystem)mpfi      2.649e-02  3.111e-02   0.851 0.396228
factor(fuelsystem)spdi     -7.795e-03  3.661e-02  -0.213 0.831778
stroke                     -6.849e-02  2.311e-02  -2.964 0.003668 **
compressionratio           -3.210e-02  1.234e-02  -2.602 0.010446 *
peakrpm                     4.985e-05  1.706e-05   2.922 0.004158 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05847 on 119 degrees of freedom
Multiple R-squared:  0.9403,    Adjusted R-squared:  0.9278
F-statistic: 75.03 on 25 and 119 DF,  p-value: < 2.2e-16
```

Figure 13. Output summary of Model 4

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.689e+00  3.374e-01   7.970 9.53e-13 ***
factor(CarCategory)luxury   1.252e-01  2.144e-02   5.842 4.38e-08 ***
factor(CarCategory)economy -4.320e-02  1.952e-02  -2.213 0.028733 *
factor(carbody)hardtop     -9.199e-02  3.722e-02  -2.471 0.014839 *
factor(carbody)hatchback   -1.009e-01  3.244e-02  -3.109 0.002334 **
factor(carbody)sedan       -7.789e-02  3.136e-02  -2.484 0.014362 *
factor(carbody)wagon       -1.216e-01  3.402e-02  -3.574 0.000505 ***
factor(enginelocation)rear  2.260e-01  6.122e-02   3.692 0.000334 ***
carwidth                    1.365e-02  6.031e-03   2.263 0.025383 *
curbweight                  2.699e-04  3.186e-05   8.469 6.56e-14 ***
factor(enginetype)l        -7.667e-02  3.531e-02  -2.171 0.031853 *
factor(enginetype)ohc       3.001e-02  2.467e-02   1.217 0.226113
factor(enginetype)ohcf      8.831e-03  3.959e-02   0.223 0.823878
factor(enginetype)ohcv     -1.467e-02  2.930e-02  -0.501 0.617349
factor(enginetype)rotor     1.383e-01  6.424e-02   2.152 0.033347 *
factor(fuelsystem)2bbl     -3.693e-02  3.114e-02  -1.186 0.237885
factor(fuelsystem)4bbl     -5.080e-02  7.868e-02  -0.646 0.519714
factor(fuelsystem)idi       3.333e-01  1.584e-01   2.105 0.037345 *
factor(fuelsystem)mpfi      2.477e-02  3.110e-02   0.796 0.427304
factor(fuelsystem)spdi     -2.514e-04  3.643e-02  -0.007 0.994506
stroke                     -5.775e-02  2.268e-02  -2.546 0.012132 *
compressionratio           -2.293e-02  1.147e-02  -1.998 0.047888 *
peakrpm                     4.586e-05  1.674e-05   2.740 0.007070 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05898 on 122 degrees of freedom
Multiple R-squared:  0.9378,    Adjusted R-squared:  0.9266
F-statistic: 83.59 on 22 and 122 DF,  p-value: < 2.2e-16
```

Figure 14. Output summary of Model 5

5) Residual Analysis for Model 5

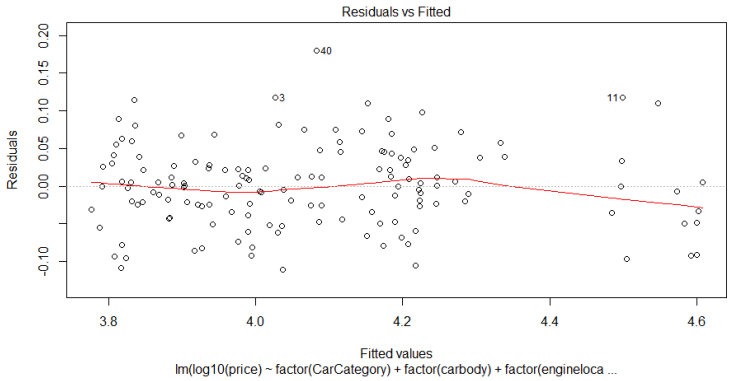


Table 15. Shaprio-Wilk Normality Test for Residual of Model 5

SHAPIRO-WILK NORMALITY TEST	
W=0.98932	p-value=0.3347

Table 6. Shaprio-Wilk Normality Test for Residual of Model 5

	GVIF	Df	GVIF^(1/(2*Df))
factor(CarCategory)	5.979902	2	1.563772
factor(carbody)	2.612500	4	1.127540
factor(enginelocation)	2.125514	1	1.457914
carwidth	6.661035	1	2.580898
curbweight	11.145880	1	3.338545
factor(enginetype)	35.701114	5	1.429777
factor(fuelsystem)	1656.208939	5	2.098512
stroke	2.226295	1	1.492077
compressionratio	95.676860	1	9.781455
peakrpm	2.716178	1	1.648083

Figure 16. VIF for Model 5

6) Plots of Predicted car Price vs. Factors

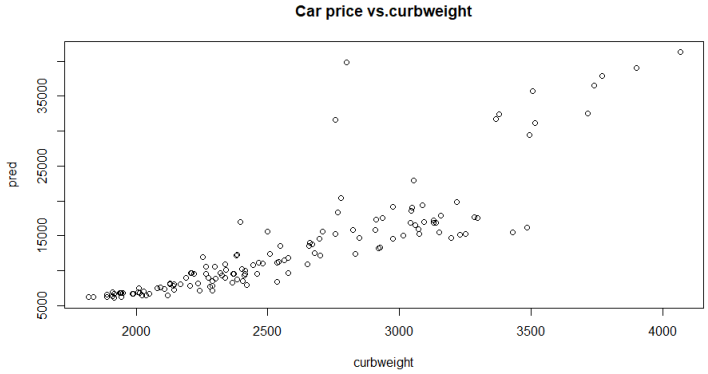


Figure 20. Predicted price vs. curbweight

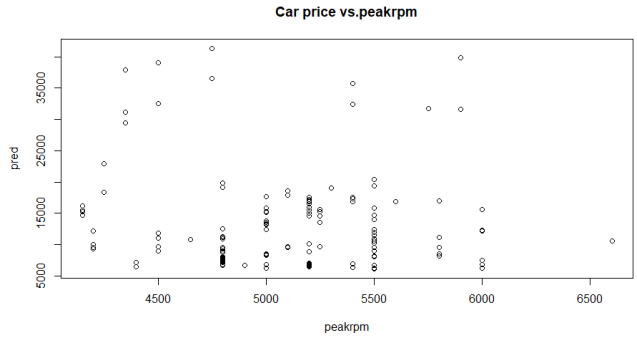


Figure 21. Predicted price vs.peakrpm

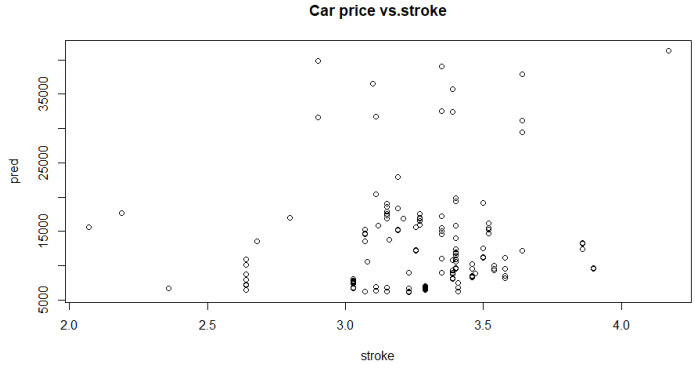


Figure 22. Predicted price vs.stroke

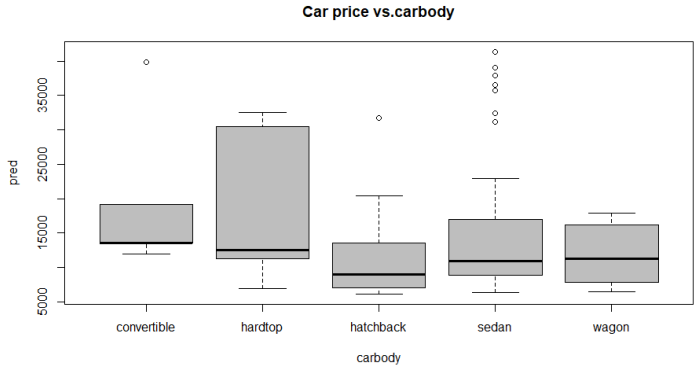


Figure 23. Predicted price vs.carbody

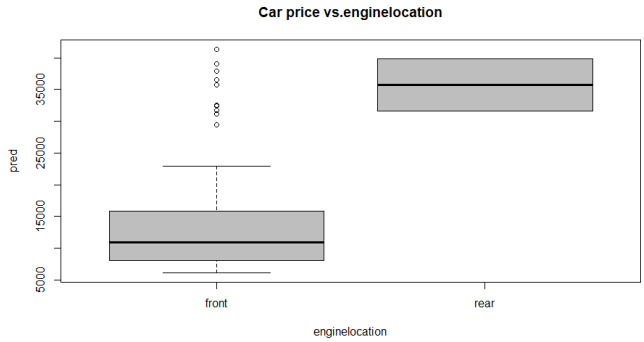


Figure 24. Predicted price vs. enginelocation

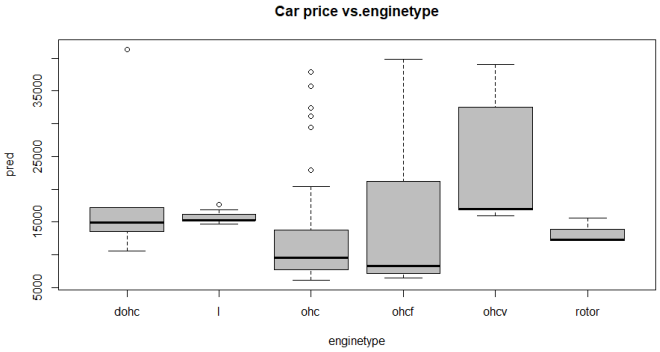


Figure 25. Predicted price vs. enginetype

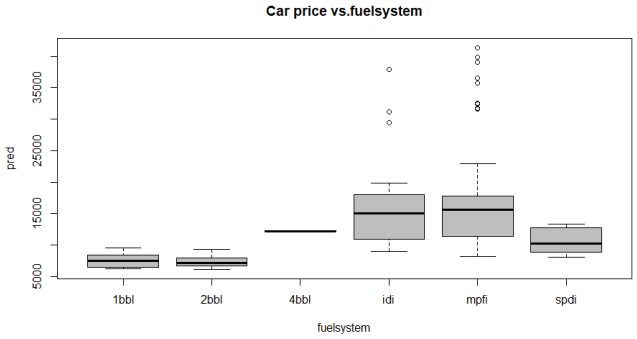


Figure 26. Predicted price vs. fuelsystem

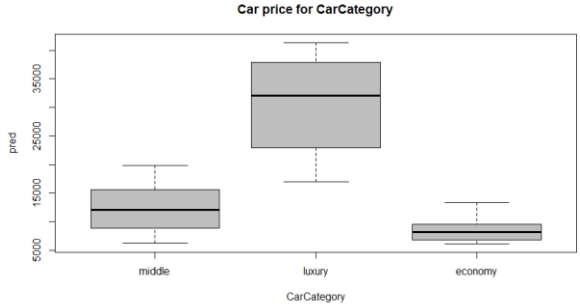


Figure 27. Predicted price vs. carCategory

8 R Code

```
##### Import Data ####

library(ggplot2)

library(dbplyr)

library(readr) # importing csv files

library(MASS) # Boxcox function

library(car) # qqPlot function

library(moments) # skeweness and kurtosis functions

library(stringr)

library(corrplot)

library(tidyverse)

library(knitr)

library(gridExtra)

library(Metrics)

Car<- read.csv("C:/Users/chloe/Desktop/R project/test folder/QEM Aug 2019/QEM JAN 2019/CarPrice.csv")


###### Data Cleaning #####

summary(Car)

Cdata=Car[Car$price>1000 & Car$price<100000,] #Remove Outlier in response

Cdata<- Cdata[!(is.na(Cdata$enginelocation) | Cdata$enginelocation==""), ] #Remove missing values enginetype

Cdata<-na.omit(Cdata) #Remove missing values in horsepower

Cdata<-Cdata[!(is.na(Cdata$enginetype) | Cdata$enginetype=="FALSE"),] # remove mis-entered values

remove<-c(206:214)

Cdata = Cdata[!Cdata$car_ID%in%remove, ] #remove duplicates

r1<-c("mfi","spfi")

Cdata = Cdata[!Cdata$fuelsystem%in%r1, ] #remove 1 observation

r2<-c("three","twelve")

Cdata = Cdata[!Cdata$cylindernumber%in%r2, ]
```

```
Cdata = Cdata[,-1] #Car_ID
summary(Cdata)



###### Replace Value of CarName #####
alfaromero<-Cdata %>%
  filter(str_detect(CarName, "alfa-romero"))
mean(alfaromero$price) #14997.5
audi<-Cdata %>%
  filter(str_detect(CarName, "audi"))
mean(audi$price) #17631.25
bmw<-Cdata %>%
  filter(str_detect(CarName, "bmw"))
mean(bmw$price) #26203.33
dodge<-Cdata %>%
  filter(str_detect(CarName, "dodge"))
mean(dodge$price) #9272
honda <-Cdata %>%
  filter(str_detect(CarName, "honda"))
mean(honda$price) #8288.5
isuzu <-Cdata %>%
  filter(str_detect(CarName, "isuzu"))
mean(isuzu$price) #15233.38
jaguar <-Cdata %>%
  filter(str_detect(CarName, "jaguar"))
mean(jaguar$price) #34600
mazda <-Cdata %>%
  filter(str_detect(CarName, "mazda"))
mean(mazda$price) #11210.33
buick <-Cdata %>%
```

```
  filter(str_detect(CarName, "buick"))
mean(buick$price) #34312
mitsubishi<-Cdata %>%
  filter(str_detect(CarName, "mitsubishi"))
mean(mitsubishi$price) #9956
nissan<-Cdata %>%
  filter(str_detect(CarName, "nissan"))
mean(nissan$price) #11188.29
peugeot<-Cdata %>%
  filter(str_detect(CarName, "peugeot"))
mean(peugeot$price) #15877.78
plymouth<-Cdata %>%
  filter(str_detect(CarName, "plymouth"))
mean(plymouth$price) #7285.5
porsche<-Cdata %>%
  filter(str_detect(CarName, "porsche"))
mean(porsche$price) #31118.62
renault<-Cdata %>%
  filter(str_detect(CarName, "renault"))
mean(renault$price) #9595
saab<-Cdata %>%
  filter(str_detect(CarName, "saab"))
mean(saab$price) #14638
subaru<-Cdata %>%
  filter(str_detect(CarName, "subaru"))
mean(subaru$price) #8140.333
toyota<-Cdata %>%
  filter(str_detect(CarName, "toyo"))
mean(toyota$price) #10110.42
volkswagen<-Cdata %>%
```

```
  filter(str_detect(CarName, "volk"))
mean(volkswagen$price) #10466
volvo<-Cdata %>%
  filter(str_detect(CarName, "volvo"))
mean(volvo$price) #17760.71
Cdata$CarName<- gsub("alfa.*","alfa-romero", Cdata$CarName)
Cdata$CarName<- gsub("audi.*","audi", Cdata$CarName)
Cdata$CarName<- gsub("bmw.*","bmw", Cdata$CarName)
Cdata$CarName<- gsub("buick.*","buick", Cdata$CarName)
Cdata$CarName<- gsub("chevrolet.*","chevrolet", Cdata$CarName)
Cdata$CarName<- gsub("dodge.*","dodge", Cdata$CarName)
Cdata$CarName<- gsub("honda.*","honda", Cdata$CarName)
Cdata$CarName<- gsub("isuzu.*","isuzu", Cdata$CarName)
Cdata$CarName<- gsub("jaguar.*","jaguar", Cdata$CarName)
Cdata$CarName<- gsub("mazda.*","mazda", Cdata$CarName)
Cdata$CarName<- gsub("mercury.*","mercury", Cdata$CarName)
Cdata$CarName<- gsub("mitsubishi.*","mitsubishi", Cdata$CarName)
Cdata$CarName<- gsub("nissan.*","nissan", Cdata$CarName)
Cdata$CarName<- gsub("peugeot.*","peugeot", Cdata$CarName)
Cdata$CarName<- gsub("plymouth.*","plymouth", Cdata$CarName)
Cdata$CarName<- gsub("porsche.*","porsche", Cdata$CarName)
Cdata$CarName<- gsub("renault.*","renault", Cdata$CarName)
Cdata$CarName<- gsub("saab.*","saab", Cdata$CarName)
Cdata$CarName<- gsub("subaru.*","subaru", Cdata$CarName)
Cdata$CarName<- gsub("toyo.*","toyota", Cdata$CarName)
Cdata$CarName<- gsub("volk.*","volkswagen", Cdata$CarName)
Cdata$CarName<- gsub("volvo.*","volvo", Cdata$CarName)
e<-c("chevrolet","dodge","honda","mitsubishi","plymouth","renault","subaru")
m<-c("alfa-
romero","audi","mazda","isuzu","mercury","nissan","peugeot","saab","volkswagen","volvo","toyota")
```

```
l<-c("bmw","buick","jaguar","porsche")
Cdata$CarName[Cdata$CarName%in%e]<-"economy"
Cdata$CarName[Cdata$CarName%in%m]<-"middle"
Cdata$CarName[Cdata$CarName%in%l]<-"luxury"
Cdata$CarCategory=Cdata$CarName
Cdata=Cdata[,-2]
summary(Cdata)

###### Recode variables#####

Cdata<-Cdata %>% mutate_at(c("CarCategory","fueltype","aspiration","doornumber",
                        "carbody","drivewheel","enginelocation","enginetype",
                        "cylindernumber","fuelsystem"), as_factor)

###### EDA #####
hist(Cdata$price,prob=TRUE,main = "Car Price Distribution",xlab="Car price",ylab="Count")#highly screwed positive data
mx=mean(Cdata$price)
lines(density(Cdata$price),col="grey")
abline(v = mx, col = "blue", lwd = 1)
md=median(Cdata$price)
abline(v = md, col = "red", lwd = 1)
text(locator(), labels = c("median", "mean"),col=c("red","blue"))
summary(Cdata$price)



p1<-ggplot(data.frame(Cdata$CarCategory), aes(x=Cdata$CarCategory)) +
  geom_bar()
g1=print(p1 + labs(y="count", x = "CarCategory")
     +ggtitle("CarCategory distribution")
```

```
    +theme(plot.title = element_text(hjust = 0.5)))
p2<-ggplot(data.frame(Cdata$fueltype), aes(x=Cdata$fueltype)) +
  geom_bar()
g2=print(p2 + labs(y="count", x = "fueltype")
    +ggtitle("fueltype distribution")
    +theme(plot.title = element_text(hjust = 0.5)))
p3<-ggplot(data.frame(Cdata$aspiration), aes(x=Cdata$aspiration)) +
  geom_bar()
g3=print(p3 + labs(y="count", x = "aspiration")
    +ggtitle("aspiration distribution")
    +theme(plot.title = element_text(hjust = 0.5)))
p4<-ggplot(data.frame(Cdata$doornumber), aes(x=Cdata$doornumber)) +
  geom_bar()
g4=print(p4 + labs(y="count", x = "doornumber")
    +ggtitle("doornumber distribution")
    +theme(plot.title = element_text(hjust = 0.5)))
p5<-ggplot(data.frame(Cdata$carbody), aes(x=Cdata$carbody)) +
  geom_bar()
g5=print(p5 + labs(y="count", x = "carbody")
    +ggtitle("carbody distribution")
    +theme(plot.title = element_text(hjust = 0.5)))
p6<-ggplot(data.frame(Cdata$drivewheel), aes(x=Cdata$drivewheel)) +
  geom_bar()
g6=print(p6 + labs(y="count", x = "drivewheel")
    +ggtitle("drivewheel distribution")
    +theme(plot.title = element_text(hjust = 0.5)))
p7<-ggplot(data.frame(Cdata$enginelocation), aes(x=Cdata$enginelocation)) +
  geom_bar()
g7=print(p7 + labs(y="count", x = "enginelocation")
    +ggtitle("enginelocation distribution")
```

```
    +theme(plot.title = element_text(hjust = 0.5)))
p8<-ggplot(data.frame(Cdata$enginetype), aes(x=Cdata$enginetype)) +
  geom_bar()
g8=print(p8 + labs(y="count", x = "enginetype")
    +ggtitle("enginetype distribution")
    +theme(plot.title = element_text(hjust = 0.5)))
p9<-ggplot(data.frame(Cdata$cylindernumber), aes(x=Cdata$cylindernumber)) +
  geom_bar()
g9=print(p9 + labs(y="count", x = "cylindernumber")
    +ggtitle("cylindernumber distribution")
    +theme(plot.title = element_text(hjust = 0.5)))
p10<-ggplot(data.frame(Cdata$fuelsystem), aes(x=Cdata$fuelsystem)) +
  geom_bar()
g10=print(p10 + labs(y="count", x = "fuelsystem")
    +ggtitle("fuelsystem distribution")
    +theme(plot.title = element_text(hjust = 0.5)))
grid.arrange(g1, g2,g3,g4,g5,g6,g7,g8,g9,g10, ncol=5)

plot(price~factor(CarCategory),Cdata,main="Car price for CarCategory",
    xlab="CarCategory",col="grey")
plot(price~factor(fueltype),Cdata,main="Car price for fueltype",
    xlab="fueltype",col="grey")
plot(price~factor(aspiration),Cdata,main="Car price for aspiration",
    xlab="aspiration",col="grey")
plot(price~factor(doornumber),Cdata,main="Car price for doornumber",
    xlab="doornumber",col="grey")
plot(price~factor(carbody),Cdata,main="Car price for carbody",
    xlab="carbody",col="grey")
plot(price~factor(drivewheel),Cdata,main="Car price for drivewheel",
    xlab="drivewheel",col="grey")
```

```
plot(price~factor(enginelocation),Cdata,main="Car price for enginelocation",
    xlab="enginelocation",col="grey")
plot(price~factor(enginetype),Cdata,main="Car price for enginetype",
    xlab="enginetype",col="grey")
plot(price~factor(cylindernumber),Cdata,main="Car price for cylindernumber",
    xlab="cylindernumber",col="grey")
plot(price~factor(fuelsystem),Cdata,main="Car price for fuelsystem",
    xlab="fuelsystem",col="grey")

par(mfrow=c(1,1))
plot(price~wheelbase,Cdata,main="Car price vs.Wheelbase")
plot(price~carlength,Cdata,main="Car price vs.Carlength")
plot(price~carwidth,Cdata,main="Car price vs.Carwidth")
plot(price~carheight,Cdata,main="Car price vs.Carheight")
plot(price~curbweight,Cdata,main="Car price vs.Curbweight")
plot(price~enginesize,Cdata,main="Car price vs.Enginesize")
plot(price~boreratio,Cdata,main="Car price vs.Boreratio")
plot(price~stroke,Cdata,main="Car price vs.Stroke")
plot(price~compressionratio,Cdata,main="Car price vs.Compressionratio")
plot(price~horsepower,Cdata,main="Car price vs.Horsepower")
plot(price~peakrpm,Cdata,main="Car price vs.Peakrpm")
plot(price~citympg,Cdata,main="Car price vs.Citympg")
plot(price~highwaympg,Cdata,main="Car price vs.Highwaympg")
plot(price~symboling,Cdata,main="Car price vs.Symboling")
Ccon=Cdata[,c(1,8:12,15,17:24)]
c=cor(Ccon,method="pearson")
corrplot(c, method="color")
######### Modeling #######
f<-lm(price~.,data=Cdata)
s=summary(f)
```

```
AIC(f)
BIC(f)
step.model <- stepAIC(f, direction = "backward",
                      trace = TRUE)
summary(step.model)
anova(step.model)
f2=lm(price~factor(CarCategory)+factor(aspiration)
    +factor(carbody)+factor(enginelocation)+carwidth
    +curbweight+factor(enginetype)+factor(cylindernumber)
    +enginesize+factor(fuelsystem)+stroke+compressionratio+peakrpm,Cdata)
summary(f2)
AIC(f2)
BIC(f2)
plot(step.model)
shapiro.test(step.model$residuals)
durbinWatsonTest(step.model)
vif(step.model)
bc=boxcox(step.model)
bc$x[which(bc$y==max(bc$y))]
text(locator(), labels ="best lambda=0.2",col="black")


f3=lm(log10(price)~factor(CarCategory)+factor(aspiration)
     +factor(carbody)+factor(enginelocation)+carwidth
     +curbweight+factor(enginetype)+factor(cylindernumber)
     +boreratio+factor(fuelsystem)+stroke+compressionratio+peakrpm,Cdata)

step<- stepAIC(f3, direction = "backward",
                      trace = TRUE)
summary(step)
```

```
Anova(step)
AIC(step)
BIC(step)


f4=lm(log10(price)~factor(CarCategory)+factor(carbody)+factor(enginelocation)+carwidth
    +curbweight+factor(enginetype)+factor(fuelsystem)+stroke+compressionratio+peakrpm,Cdata)
summary(f4)
Anova(f4)
AIC(f4)
BIC(f4)
plot(f4)
shapiro.test(f4$residuals)
vif(f4)

f5=lm(log10(price)~factor(CarCategory)+factor(carbody)+factor(enginelocation)
    +curbweight+factor(enginetype)+factor(fuelsystem)+stroke+peakrpm,Cdata)
summary(f5)
Anova(f5)
AIC(f5)
BIC(f5)
plot(f5)
shapiro.test(f5$residuals)
vif(f5)

Cdata$pred=(10)^predict(f5)
plot(pred~peakrpm,Cdata,main="Car price vs.peakrpm")
plot(pred~stroke,Cdata,main="Car price vs.stroke")
plot(pred~curbweight,Cdata,main="Car price vs.curbweight")
plot(pred~factor(carbody),Cdata,main="Car price vs.carbody",xlab="carbody",col="grey")
```

```
plot(pred~factor(enginelocation),Cdata,main="Car price
vs.enginelocation",xlab="enginelocation",col="grey")

plot(pred~factor(enginetype),Cdata,main="Car price vs.enginetype",xlab="enginetype",col="grey")

plot(pred~factor(fuelsystem),Cdata,main="Car price vs.fuelsystem",xlab="fuelsystem",col="grey")

plot(pred~factor(CarCategory),Cdata,main="Car price for CarCategory",

    xlab="CarCategory",col="grey")


#err <- 10^(f5$fitted.values) - Cdata$price

#err2 <- err^2

#rmse <- sqrt(mean(err2))


################## Cross Validation

set.seed(123)

samplesize<-round(0.7 * nrow(Cdata),0)

index<-sample(seq_len(nrow(Cdata)),size=samplesize)

data_train<-Cdata[index,]

data_test<-Cdata[-index,]

f5=lm(log10(price)~factor(CarCategory)+factor(carbody)+factor(enginelocation)

     +curbweight+factor(enginetype)+factor(fuelsystem)+stroke+peakrpm,data_train)

summary(f5)

pred5<-predict(f5,newdata=data_test %>% select(-price))

rmse(10^(f5$fitted.values),data_train$price) #train rmse 1847.

rmse(10^(pred5),data_test$price) #test rmse 3630.1,since test data is larger than traning data, then our
model is overfit the trianing datasetst$price)
```