**Automobile data**

Automobile data is a multivariate dataset that contains records of 205 cars and features 26 attributes of the cars available. This dataset is available on [UCI Machine learning repository](https://archive.ics.uci.edu/ml/datasets/Automobile) as well as [Kaggle.com](https://www.kaggle.com/toramky/automobile-dataset)

**Overview of Dataset**

The dataset contains 206 observations and 26 feature variables. The feature contains Technical, actuarial as well as Insurance aspects of the cars

|  |  |  |
| --- | --- | --- |
| Feature Variable | Data Type | Description |
| Symboling | Integer | Actuarial aspect of car, symboling decides the level of risk. |
| normalized.losses | Character | Insurance aspect of car |
| Make | Character | It tells us about the brand name of the car |
| fuel.type | Character | The fuel which the car uses gasoline or diesel |
| Aspiration | Character | Whether the car is Turbocharged or standard |
| num.of.doors | Character | Number of doors the car has |
| body.style | Character | Whether the car is hatchback sedan or other type |
| drive.wheels | Character | Whether the car is front wheel drive, rear wheel drive or 4- wheel drive |
| engine.location | Character | Whether the engine is in front or at the back. |
| wheel.base | Numeric | Wheel base of the vehicle |
| length | numeric | Length of the vehicle |
| Width | numeric | Width of the vehicle |
| Height | numeric | Height of the vehicle |
| curb.weight | integer | Curb weight of the vehicle |
| engine.type | character | Engine type of the vehicle w.r.t the camshaft position |
| num.of.cylinders | character | Number of cylinders in the engine |
| engine.size | numeric | Size of the engine |
| fuel.system | character | Fuel injection system of the vehicle |
| Bore | character | Bore length of the engine |
| Stroke | character | Stroke length of the engine |
| compression.ratio | numeric | Compression ratio of the engine |
| Horsepower | character | Horsepower of the engine |
| peak.rpm | character | Peak RPM of the engine |
| city.mpg | Integer | Miles per gallon when driven in City |
| highway.mpg | Integer | Miles Per gallon when driven at highway |
| Price | Character | Price of the vehicle |

**Exploratory Data analysis**

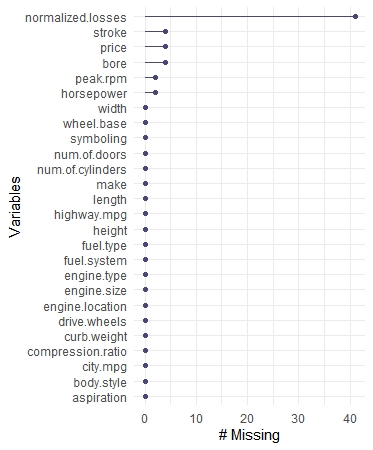
We will start our EDA by first converting the variables into the proper datatype. Our Initial data looks like something below

On checking we see that variables normalized.losses , horsepower, peak.rpm, bore, stroke, price have incorrect data type, hence we will convert it to the correct data type. The functioned used for this is as.numeric.

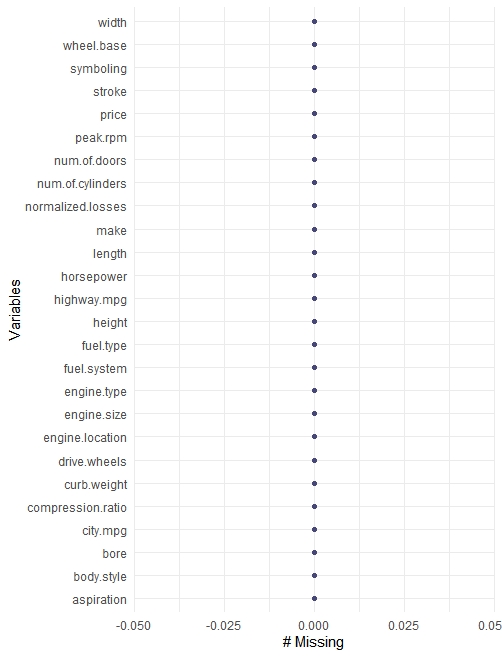
On being converted to the required datatytpe it was found that there were few missing values in these variables.

**Missing Value analysis**

Once the data types are converted to the required there were few missing that got inducted in our dataset. Missing values can be visualised using the R library **naniar** and using the function **gg\_miss\_var.**



The above visualisation shows that the variable normalized losses has the highest number of the missing values followed by stroke, price, bore, peak.rpm, horsepower. Hence used **naniar** library and **gg\_miss\_var** **function** to remove the missing values.

****

The above graph shows that there are no more any missing values in our dataset. For the missing value imputation median was used as mean can alter the outliers.

**Creating dummy Variables**

For our categorical variables, we will create the dummy variables. We will create dummy variable for the aspiration, fuel.type, num.of.doors, engine.location, drive.wheels, engine.type, num.of.cylinders, fuel.system, body.style, make.

**Treating column Symboling**

Symboling is a parameter set up by the Actuaries in order to determine the risk level of any vehicle. New features were created based on the below levels.

Symboling <= -1 low risk

Symboling = 0,1 medium risk

Symboling > 1 high risk

Now that we have treated symboling column we will now eliminate all the un necessary variables. As we had created the dummy variables so in order to avoid the multicollinearity we will have to create (n-1) variables hence we will eliminate the unnecessary variables.

After removing the variable our dataset will have 205 observations and 60 variables. We will now build model on top of it.

**Model Building**

The initial dataset is split into training set and test set with a ratio of 80/20. For the evaluation R-squared value is taken into consideration and for the model building and model selection, statistical significance of the variables and variance inflation factor was taken a criterion.

Our first model has been named as regressor\_linear. Our model should not contain many variables hence stepwise regression is used to get the significant variables.

6 iterations were run and our 6th model had all the significant variables with 3, 2- or 1-star rating. Our final model has following variables.

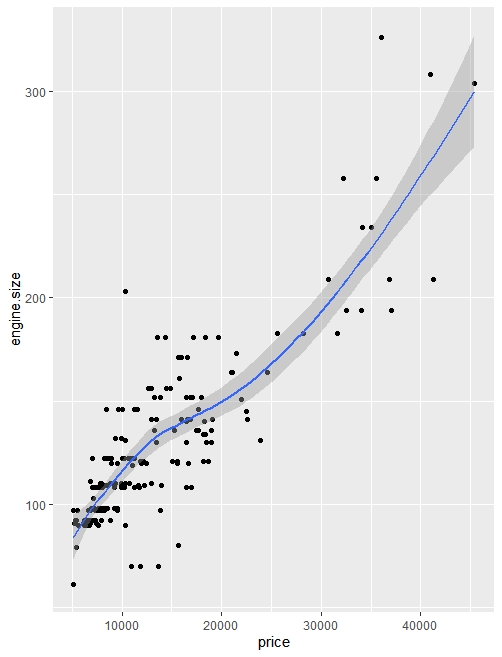
Price, normalized.losses, width, engine.size , stroke peak.rpm aspiration\_turbo engine.location\_front, engine.type\_dohcv , engine.type\_ohcv , num.of.cylinders\_five num.of.cylinders\_four, num.of.cylinders\_six, body.style\_sedan , make\_alfa-romero, make\_bmw, make\_isuzu, make\_jaguar make\_mercedes-benz, engine.type\_ohcf

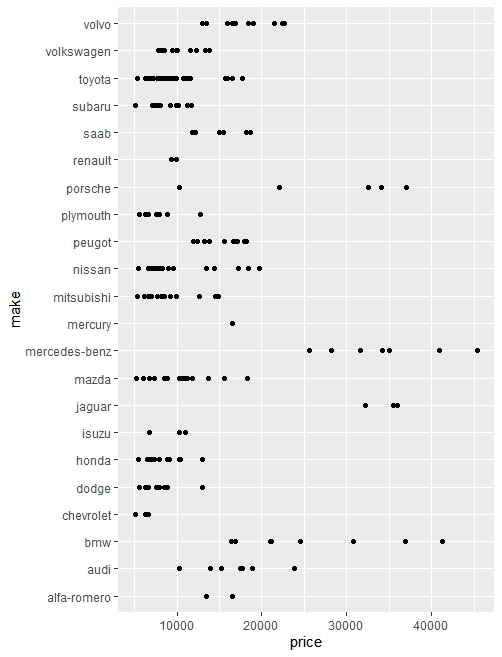
**Model Details and selection**

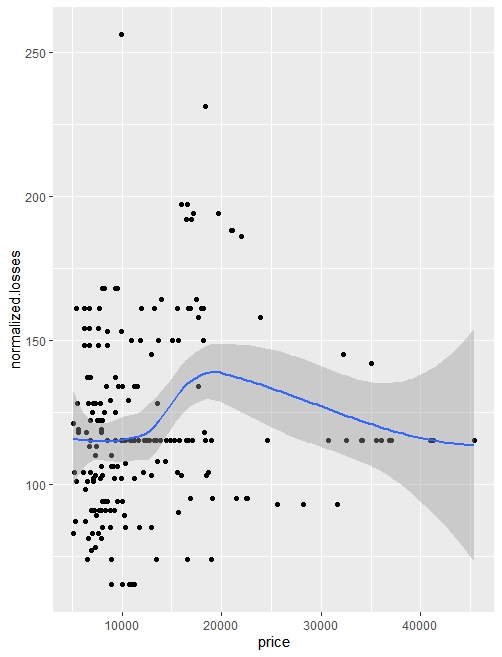
Our final model regressor\_linear\_6 had adjusted r square of 0.951. and had 20 variables. Hence model 6 was used for making the predictions. From the below table we can see that model\_6 has lowest R\_squared value and hence we will select this model to make predictions

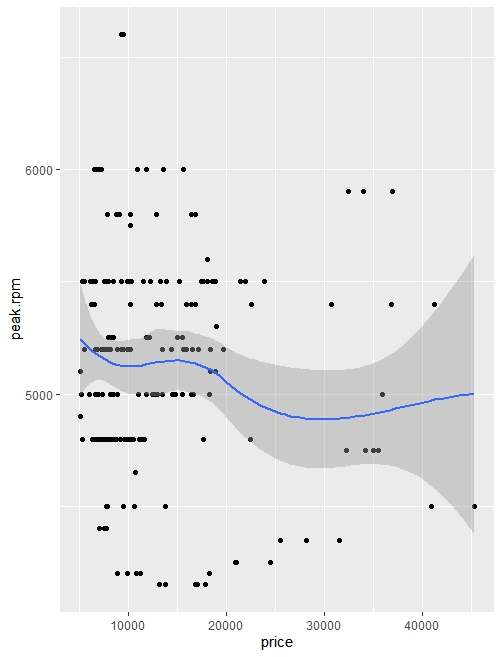
|  |  |
| --- | --- |
| R\_square | model\_num |
| 0.9146680 | Model\_6 |
| 0.9194104 | Model\_5 |
| 0.9225940 | Model\_4 |
| 0.9268626 | Model\_3 |
| 0.9270909 | Model 1 |
| 0.9290358 | Model\_2 |

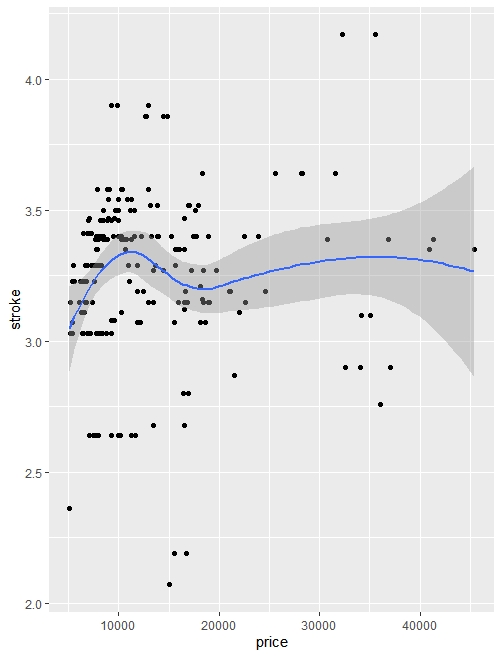
**Important Visualisations.**

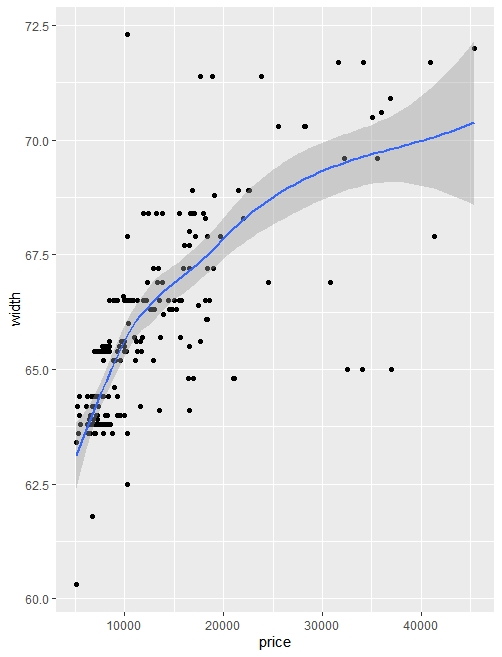












**FINAL INFERENCE**

So, for final inference we can conclude that the following variables are good predictors for the price of cars normalized.losses, width, engine.size, stroke, peak.rpm, aspiration\_turbo, engine.location\_front engine.type\_dohcv, engine.type\_ohcv, num.of.cylinders\_five, num.of.cylinders\_four, num.of.cylinders\_six, body.style\_sedan, make\_alfa-romero, make\_bmw, make\_isuzu, make\_jaguar, make\_mercedes-benz, engine.type\_ohcf

**Making Sense out of data.**

For communicating with our stakeholder, we can convey that as per our given dataset Price of any vehicle is dependent, Body type (sedans are more expensive than hatchback), Width (some vehicles like SUV’s are wider than the usual sedans or hatchbacks and hence has higher occupancy making them expensive) and brand name.

Let’s take a look at why the variable did not make it till the last iteration of our model. The length of any vehicle is largely dependent cabin, engine location and engine type. For our final model the variables which provide similar data as of length are Engine location, Body type sedan, and number of cylinders.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*The End\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*