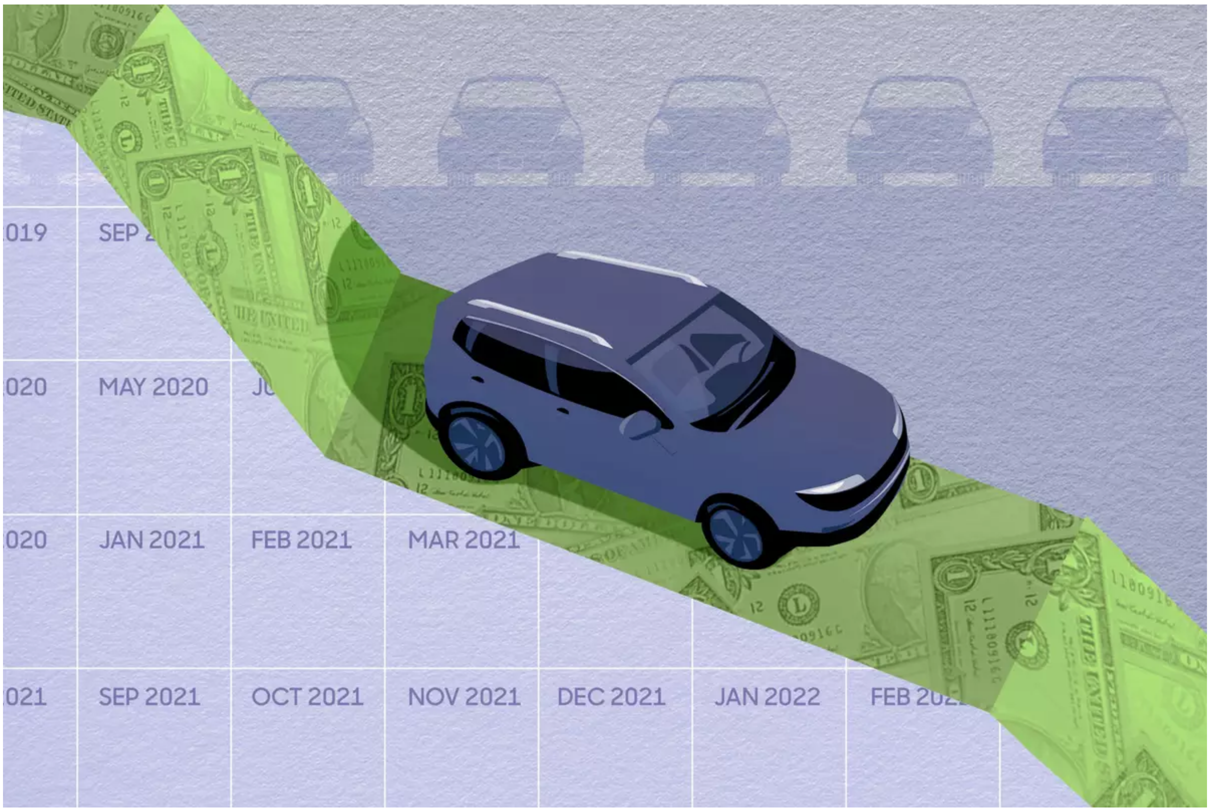
Final Project

Automobile Price Prediction and Clustering



Adityo Das Gupta - 261107470

Instructor: Prof. Juan Camilo Serpa

December, 15th 2022

Contents

I. Introduction

This project utilizes the statistical foundations of predictive data analytics using the R programming language on the Automobile dataset, which consists of data from 1985 Ward's Automotive Yearbook. The project is divided into two phases, where the first phase is concerned with developing a model to accurately predict the price of listed automobiles, while the second phase involves clustering various makes of cars on the basis of their specifications and features.

The final objective of my analysis is to create a recommendation system which would integrate the price forecasts from the first model with the cluster groups from the second model and help customers pick a vehicle of their choice or help car manufacturing companies identify segments for their advertising or production purposes.

While statistical measures lay the foundations for data-driven projects, cars must also be analysed through intuition, influence, and rational. Therefore, alongside statistical knowledge, I will use intuition and creativity along the way to make strategic decisions while building the models that could convey data-driven insights for future automobile productions. This report will help consumers pick an ideal car on the basis of their needs and also manufacturers to target certain segments of the population they want to cater to using the insights provided.

II. Phase 1 - Price Prediction

1. Data Description

In the original dataset, there were 26 attributes including the target variable: 'price' for each combination of car make, model, fuel-type, aspiration and body-style. To improve the efficiency, multiple data pre-processing steps were performed to ensure that we have a comprehensive understanding of the data and that it was concise enough for model building.

i. Data Preprocessing

I converted 'number of doors' and 'number of cylinders' from categorical to numerical columns by simply replacing the values from their word form to integers. This was done to ensure ease in visualising trends with the target variable and outliers. Variable 'normalized losses' was dropped since 20% of the records were missing. Furthermore, variable named 'engine-location' was dropped since it was homogeneous throughout all observations and hence provided no statistical significance to model formulation

ii. Missing Value Imputation

The following variables were found to have missing values and were respectively imputed by;

number of doors: Originally being a categorical variable, missing values were replaced by the mode of the respective body-style the record belonged to.

bore: The bore of a car appears to follow a normal distribution (see Plot 1a) hence missing values were replaced by the mean of the respective body-style the record belonged to.

stroke: The stroke of a car appears to follow a normal (see Plot 1a) hence missing values were replaced by the mean of the respective body-style the record belonged to.

peak-rpm: The peak rpm of a car appears to not follow a normal distribution (see Plot 1a) hence missing values were replaced by the median of the respective body-style the record belonged to.

horsepower: The stroke of a car appears to not follow a normal distribution (see Plot 1a) hence missing values were replaced by the median of the respective body-style the record belonged to.

price: The price of a car appears to follow a normal distribution (see Plot 1a) hence missing values were replaced by the mean of the respective body-style the record belonged to.

Other than observing the histograms, for a variable to follow normal distribution I checked if the mean, median and mode are very close to each other.

iii. Outlier Detection

All numerical variables were plotted against the target variable price using scatterplots (see Plot 2a and Plot 2b). Since there were only a handful of records available (205) combined with not visually observing any outliers, I decided to not drop any records.

iv. Dummy Variable

The categorical variables ‘make’, ’fuel-type’, ‘aspiration’, ‘body-style’, ‘drive-wheels’, ‘engine-type’, ‘fuel-system’ were converted from string to factors since models in R process categorical variables as factors.

2. Model Selection

i. Linearity Test

To observe any linear trends in the data, I performed Tukey test and observed residual plots of all the numerical predictors against the target variables. As seen in (Plot 3a and Plot 3b) and (Table2 and Table3) ‘symboling’, ‘wheel-base’, ‘length’, ‘width’, ‘height’, ‘curb-weight’, ‘horsepower’, ‘city-mpg’, ‘highway-mpg’ were found to be fairly non-linear.

ii. Choice of Model

As majority of the variables were observed to have non-linear relationships with the target variable, I decided to pick a Gradient Boosted Decision Tree as my choice of model since decision trees are known to effectively capture non-linear relationships and boosted models are state of the art, powerful learning models which have shown considerable success recently.

iii. Hyperparameter Tuning

I split the data into train and test sets by an 8:2 ratio, and ran the model on the training set using different combination of parameters (see Table 4), adding to a total of 81 different models using k folds cross validation with k= 4 resulting in 324 fits. I used the gbm function from the gbm package to perform the above tests. The distribution parameter was kept 'gaussian' throughout all the fits since it was a regression problem. The models were compared on the basis of resulting RMSE with the best one having the least error.

3. Model Results, Interpretations and Conclusions

The optimal parameters with the least RMSE were interaction.depth=5, n.trees = 100, shrinkage=0.01 and n.minobsinnode=5. The model evaluation metrics for the train and test dataset can be seen in the table below.

	MSE	**R^2**
Train Set	14983404	92.28%
Test Set	9301877	93.47%

The R^2 of the model was calculated by computing the correlation between the prediction and train/test target variable and then squaring it.

In terms of significance of each predictor, the top 5 predictors were 'make', 'engine-size', 'curb-weight', 'highway-mpg' and 'wheel-base' in that order as seen in (Plot 4a).

Once we had the final model ready, we stored the predictions in the base data set in a column called 'predicted_price'. This column will be used as predictor for the clustering model in the second phase.

III. Phase 2 - Clustering Analysis

1. Data Description

Going by intuition and pragmatism, I decided to select predictors which are realistically considered by consumers while buying a car. Hence, I chose to keep "wheel.base", "length", "width", "height", "curb.weight", "num.of.cylinders", "engine.size", "compression.ratio", "horsepower", "peak.rpm", "city.mpg", "highway.mpg" and "predicted_price" out of the newly updated Automobile data set.

i. Dimensionality Reduction

The characteristics of the different groups/clusters should be concise and objective, restricted to 2-3 features at max while proposing business solutions. With account to the large number of predictors I had, I decided to perform PCA on the dataset to reduce the number of predictor variables. The data was scaled before the PCA was performed.

I plotted the partial variance explained and the cumulative sum of it against the principal components (see Plot 5a). Looking at the plots and the cumulative sum of variance explained I decided to keep 5 principal components which together accounted for **92.21%** of variance in the data.

2. Model Selection

I chose the K-Means algorithm to perform clustering on the data set. To select my optimal k (number of clusters), I plotted an elbow curve which is the within cluster variation against a bunch of different values for k ranging from 1 to 10 (see Plot 6a). I chose k=5 as the optimal number of clusters since the inter cluster variation was fairly low at **895.12** and the model was not too complex with a large number of clusters.

3. Model Results and Interpretations

In terms of the PCA I observed that the first principal component's eigenvector had high magnitudes along curb-weight and highway.mpg which were also a part of the most significant predictors selected by the gradient boosted model. The fact that these variables help explain a lot of the variance in the dataset makes them integral while making predictions of car price. Additionally, it can be observed that 'horsepower', 'number of cylinders', 'curb-weight', 'engine type' and 'width' are mostly correlated with the price since the eigenvectors are close to each other. While 'height' and 'peak-rpm' are not since thety are orthogonal to the price.

The clusters can be visualised by plotting them against the first and second principal component (see Plot 7a). Additionally, I used the 'Plotly' library to visualise the clusters in 3 dimensions for visual validation (see Plot 8a). This gives us a good idea of how the clusters are formed over the data.

For interpreting each cluster, I looked at the value of cluster centres and related them to the magnitude of the eigenvectors of the principal component. For example, the clusters with high centre values for PC1 will tend to have high values for the variables whose eigenvectors have large magnitudes along PC1. Moreover, I plotted box plots for all the predictors cluster-wise for additional insights as seen in (Plot 9a, Plot 9b and Plot 9c).

Following are my inferences made for each cluster

Predictor	**Cluster 1**	**Cluster 2**	**Cluster 3**
wheel.base	average	large	less than large
length	average	large	large
width	below average	large	less than large
height	above average	variable above average	large
curb.weight	average	large	above average
number of cylinders	average	high	average
engine.size	below average	very large	average
compression.ratio	average	low	average
horsepower	below average	high	average
peak.rpm	same across all clusters		
city.mpg	average	very low	average
highway.mpg	above average	very low	average
predicted_price	just above the low	high	average

Predictor	**Cluster 4**	**Cluster 5**
wheel.base	small	small
length	below average	small
width	average	small
height	small	average
curb.weight	average	small
number of cylinders	average	average
engine.size	variable below average	small
compression.ratio	average	high
horsepower	average	low
peak.rpm	same across all clusters	
city.mpg	average	high
highway.mpg	average	high
predicted_price	varies low to average	low

Key Points :

Cluster 2 is observed to have high end cars with the best mechanics and features. Naturally placing them on the higher end of the price spectrum

Cluster 5 has the lower end cars which are low cost, low on the features but has very high mileage on road

Cluster 1, 3 and 4 have cars whose features are relatively around the mean for each of the specific features with little variation. For example, when it comes to mechanics, Cluster 1 has cars which fall below average but the mileage is above average while Cluster 3 falls above average and just below Cluster 2 in terms of mechanics. Cluster 4 varies between low and average with a large interval for the price of the cars.

IV. Business Implications

It's Christmas time and I am planning to gift my father a car out of the new models launching next month. Since he's a car buff and I'm not, I have no idea where to start and don't want to disappoint him by gifting him a car which won't live up to his hopes since he's very particular about owning a car which is dependable with strong mechanics and features. At the same time, I have constraints on my budget and can't go all out spending and need some price estimates on the cars being launched next month.

Luckily, I had this model and figured out group of cars I should look at. Not compromising on the mechanics and features but making sure the cars aren't really expensive, I decided to target the cluster 3 of cars. Following is the list from which I should plan to buy.

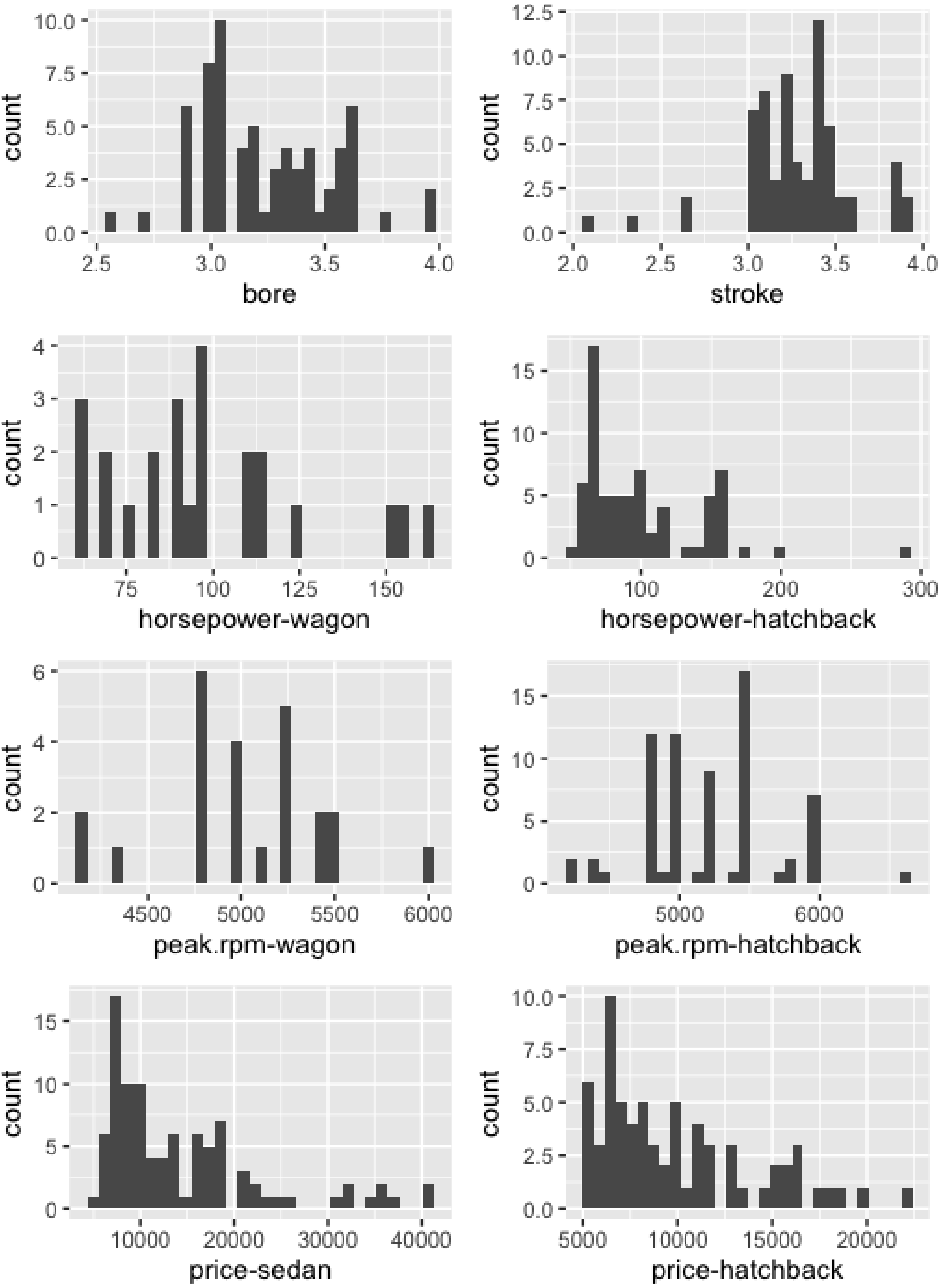
Table 1: Partial subset of Cluster 3

make	fuel.type	aspiration	body.style	predicted_price
mercury	gas	turbo	hatchback	14330.08$
nissan	gas	std	sedan	15213.85$
nissan	gas	std	wagon	15145.55$
nissan	gas	std	sedan	15199.99$
peugot	gas	std	sedan	14731.06$
peugot	diesel	turbo	sedan	14860.21$
peugot	gas	std	wagon	14953.89$
peugot	diesel	turbo	wagon	15560.18$
peugot	gas	std	sedan	14879.55$
peugot	diesel	turbo	sedan	14860.21$
peugot	gas	std	wagon	15011.42$
peugot	diesel	turbo	wagon	15560.18$
peugot	gas	std	sedan	14822.01$

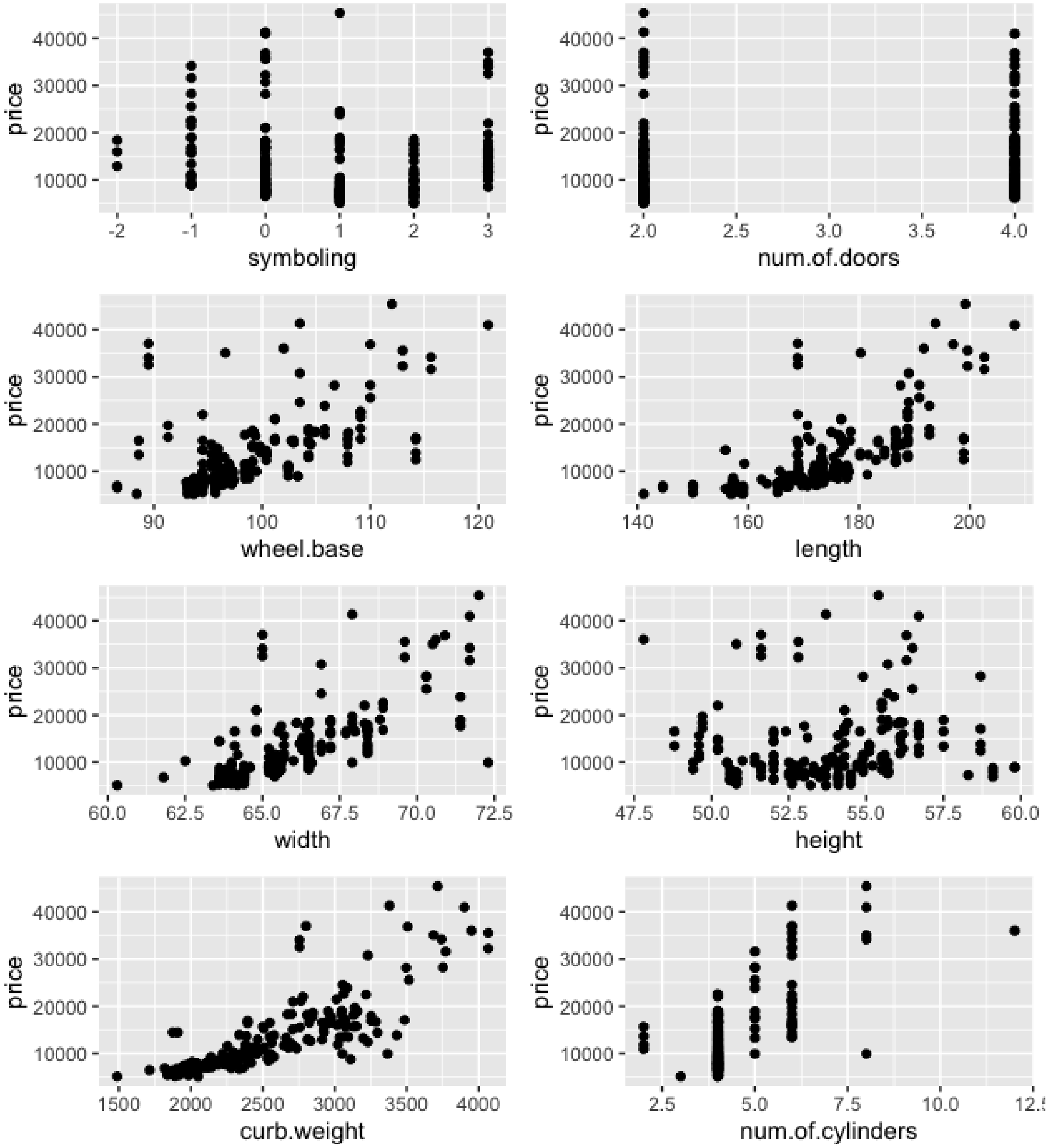
Similar to this scenario stated above, manufacturing companies can estimate a rough price they should launch their car at using my first model. Once they have price, they can use the clustering model to identify the segment it lies in and accordingly get insights about the demographic they should target using information from previous advertising and marketing campaigns.

V. Appendices

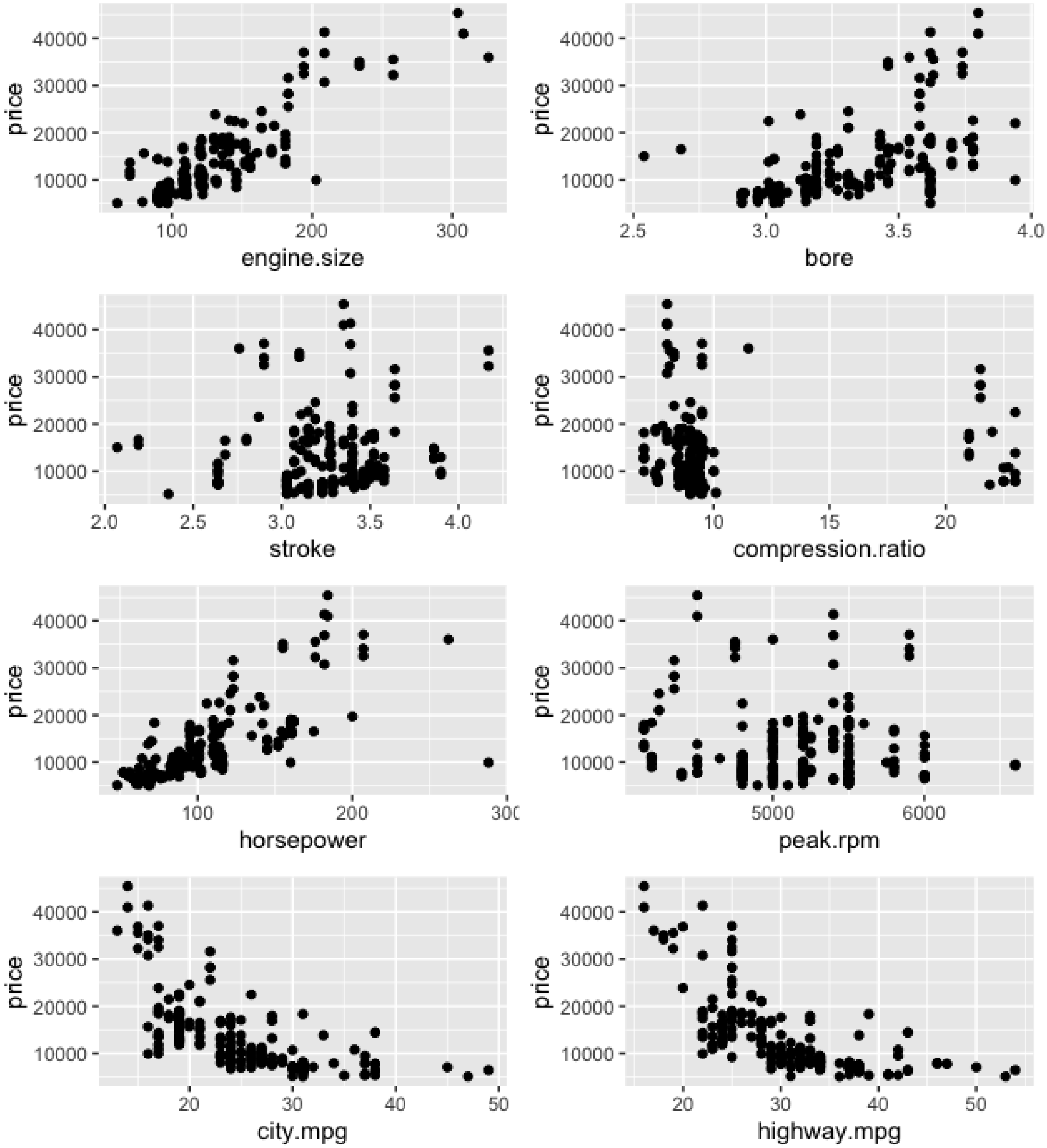
Plot 1. Distribution of variables with missing values



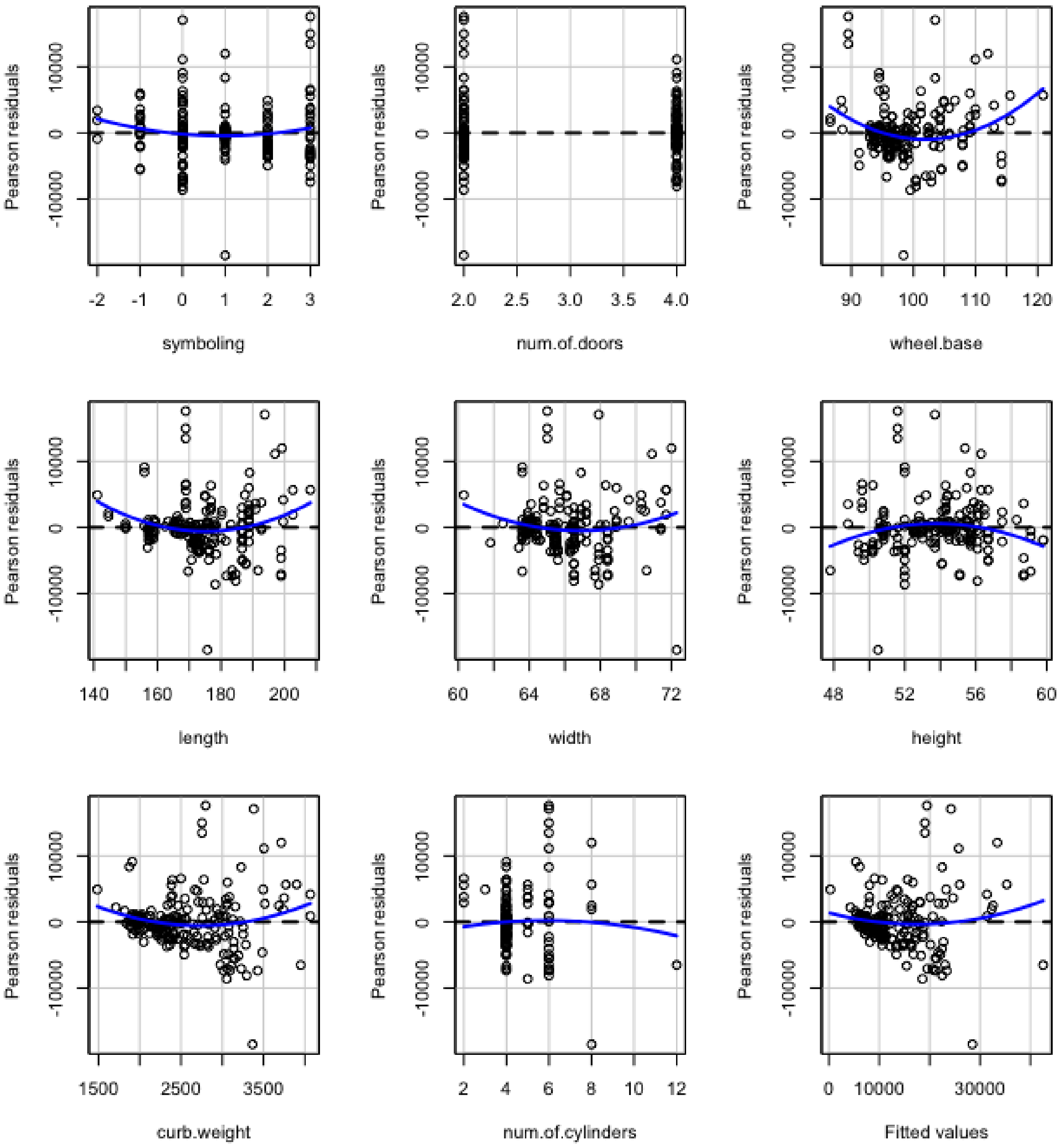
Plot 2a. Scatter Plots against price



Plot 2b. Scatter Plots against price



Plot 3a. Residual Plots



Plot 3b. Residual Plots

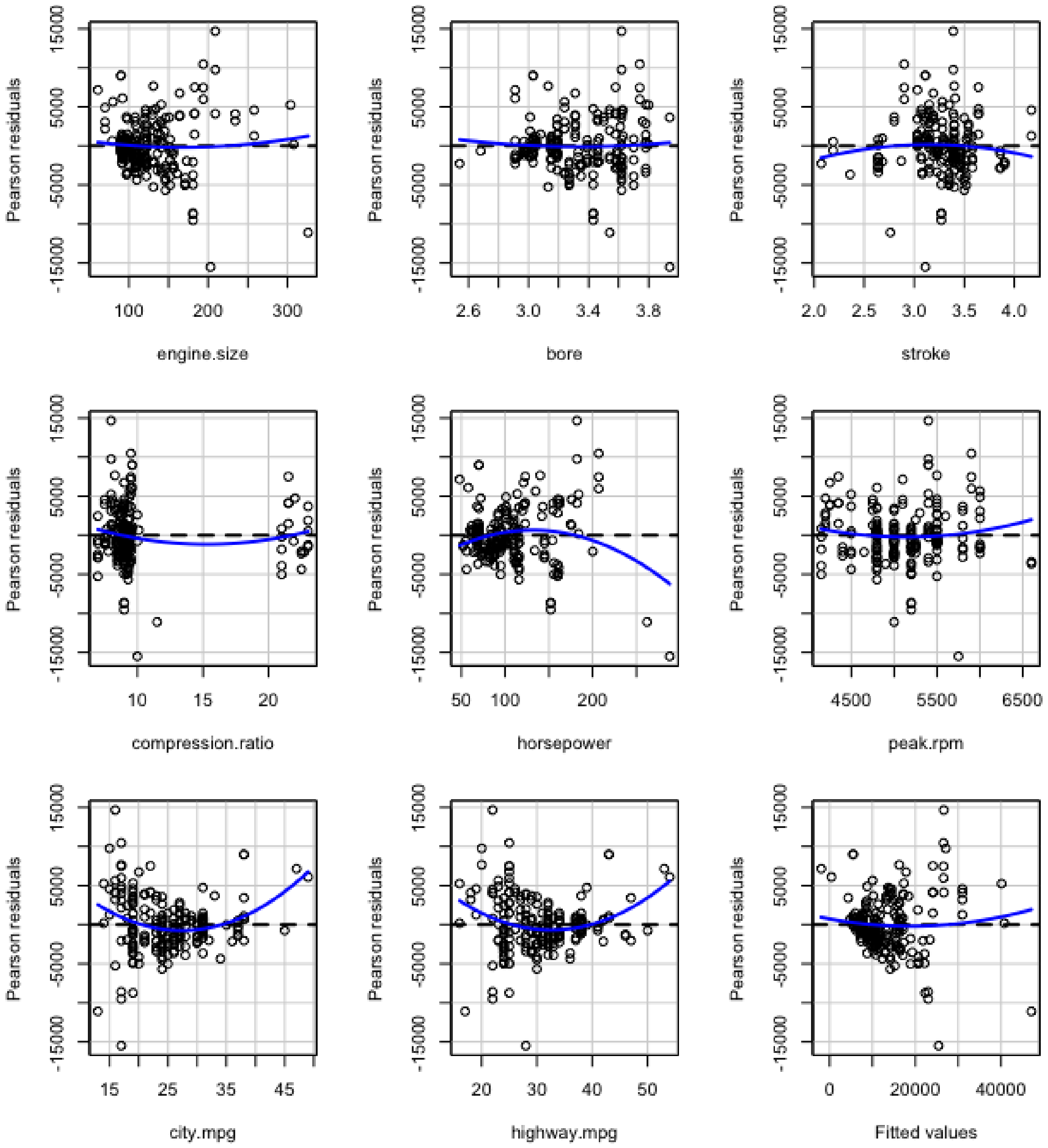


Table 2: Tukey Test

	Test stat	Pr(>|Test stat|)
symboling	1.996	0.047
number of doors	0.161	0.872
wheel.base	4.050	0.0001
length	3.595	0.0004
width	2.291	0.023
height	-2.688	0.008
curb.weight	2.704	0.007
number of cylinders	-0.700	0.485
Tukey test	1.821	0.069

Table 3: Tukey Test

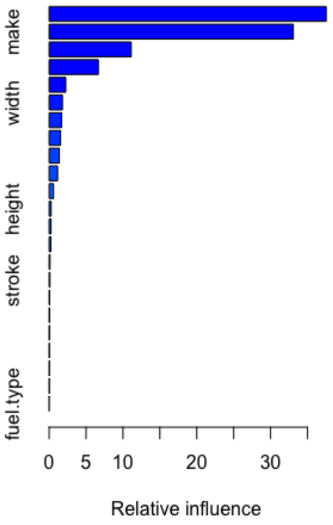
	Test stat	Pr(>|Test stat|)
engine.size	0.870	0.386
bore	0.483	0.629
stroke	-1.129	0.260
compression.ratio	1.145	0.253
horsepower	-3.970	0.0001
peak.rpm	1.471	0.143
city.mpg	5.073	0.00000
highway.mpg	4.759	0.00000
Tukey test	1.309	0.190

Table 4. Hyperparmeter Tuning

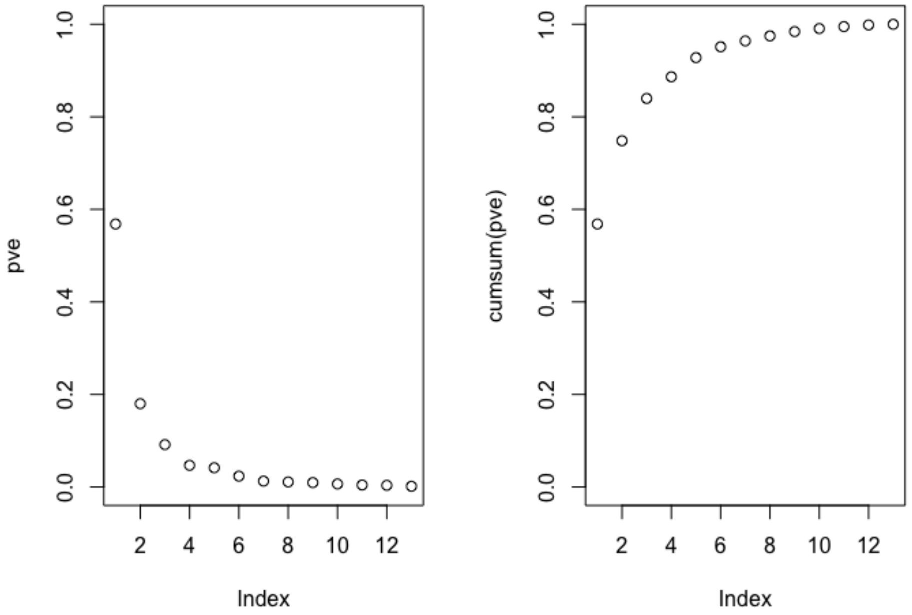
Parameter	Value 1	Value 1	Value 3
interaction.depth	1	3	5
n.trees	25	50	100
shrinkage	0.1	0.01	0.001
n.minobsinnode	5	10	15

.

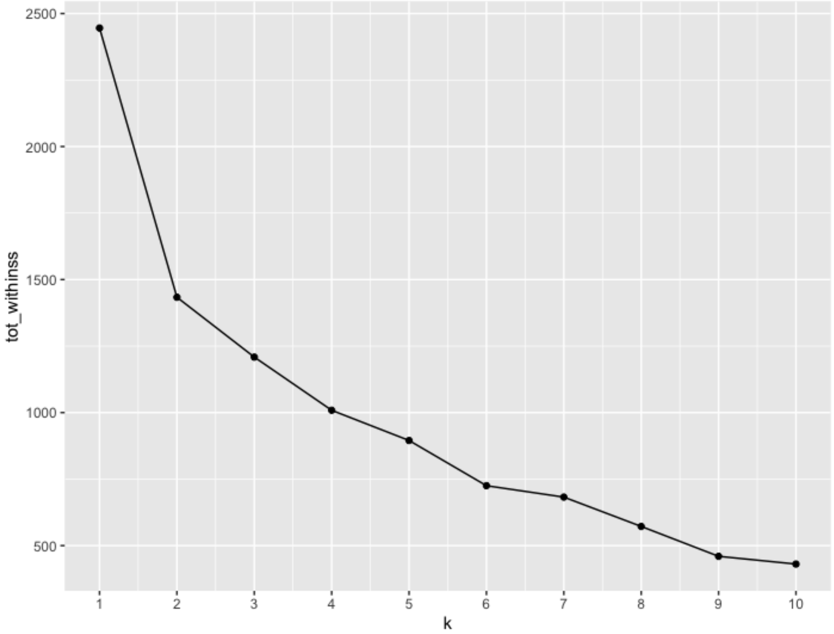
Plot 4a. Predictor Significance



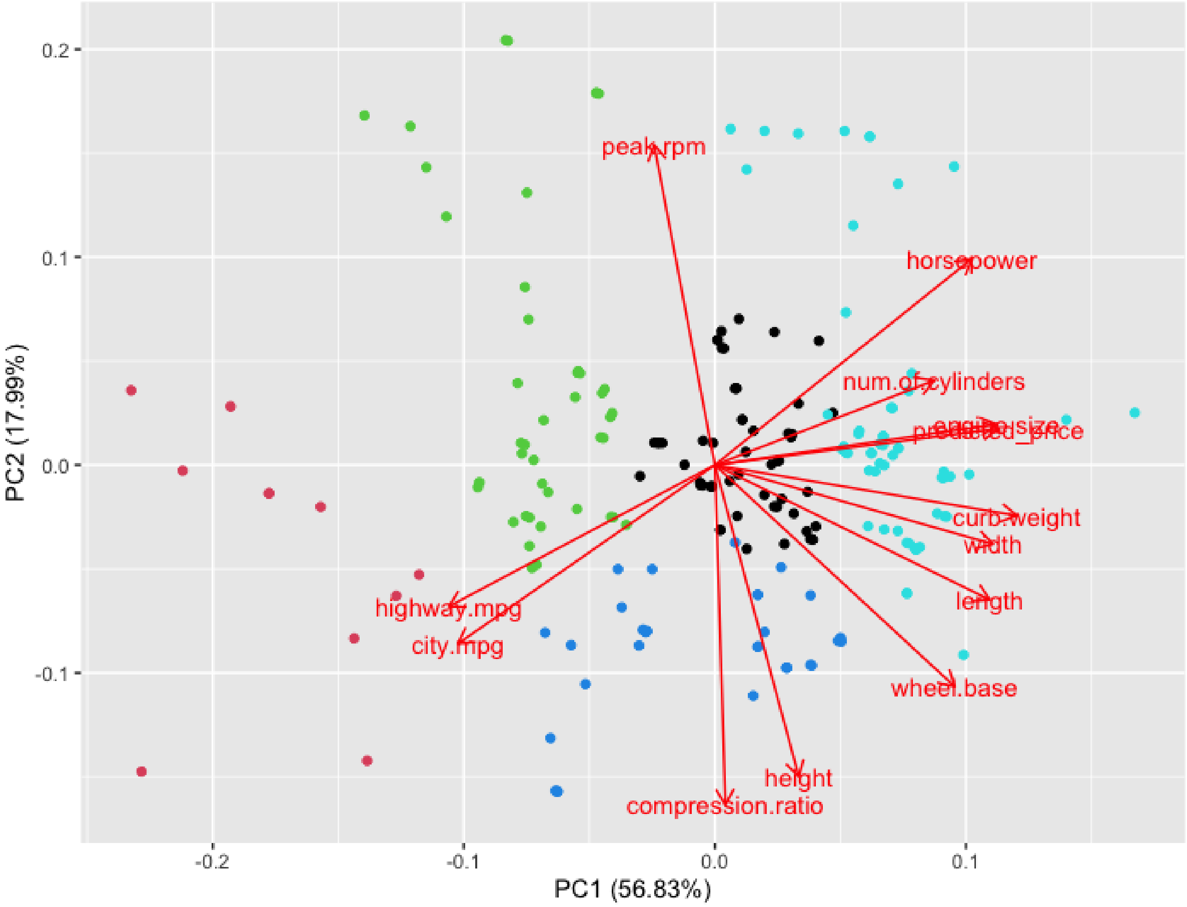
Plot 5a. Scree Plot



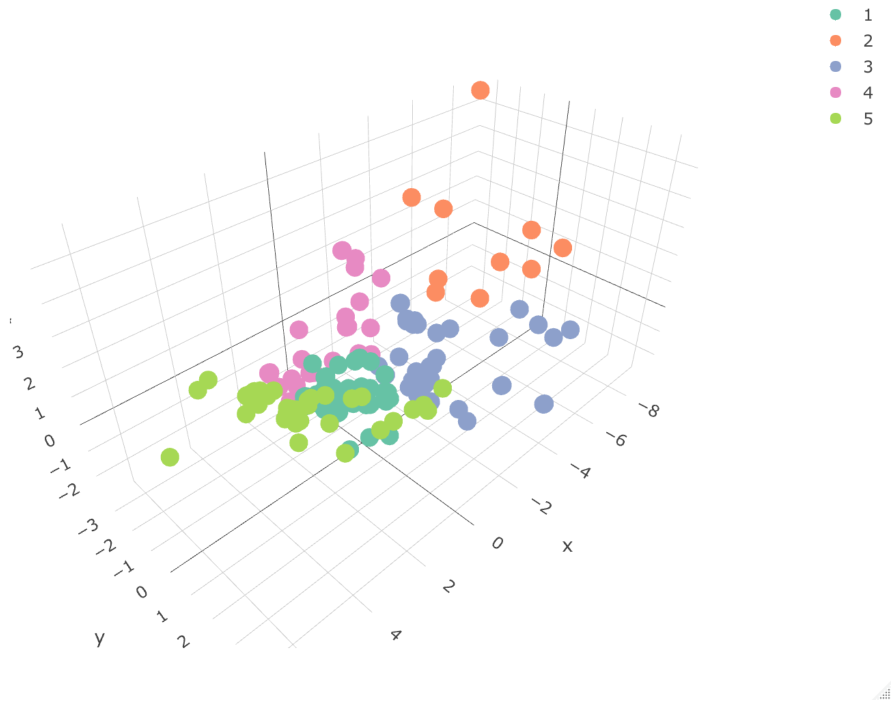
Plot 6a. Elbow Curve



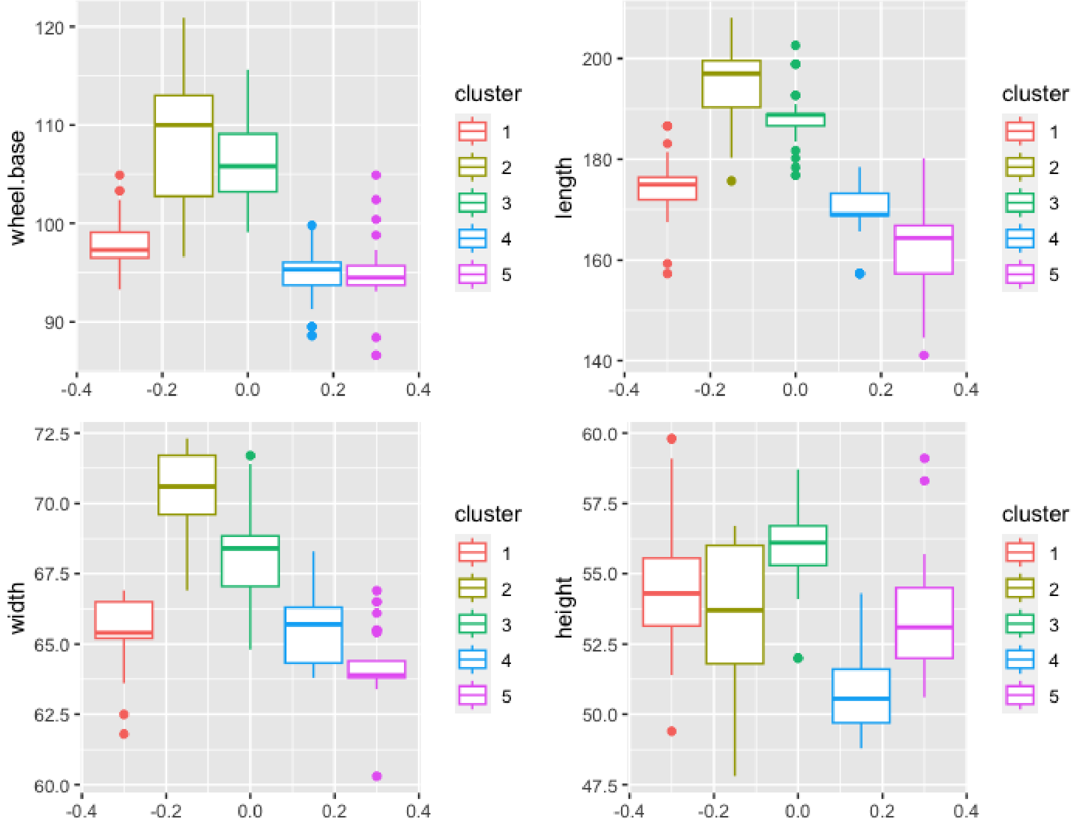
Plot 7a. Clusters



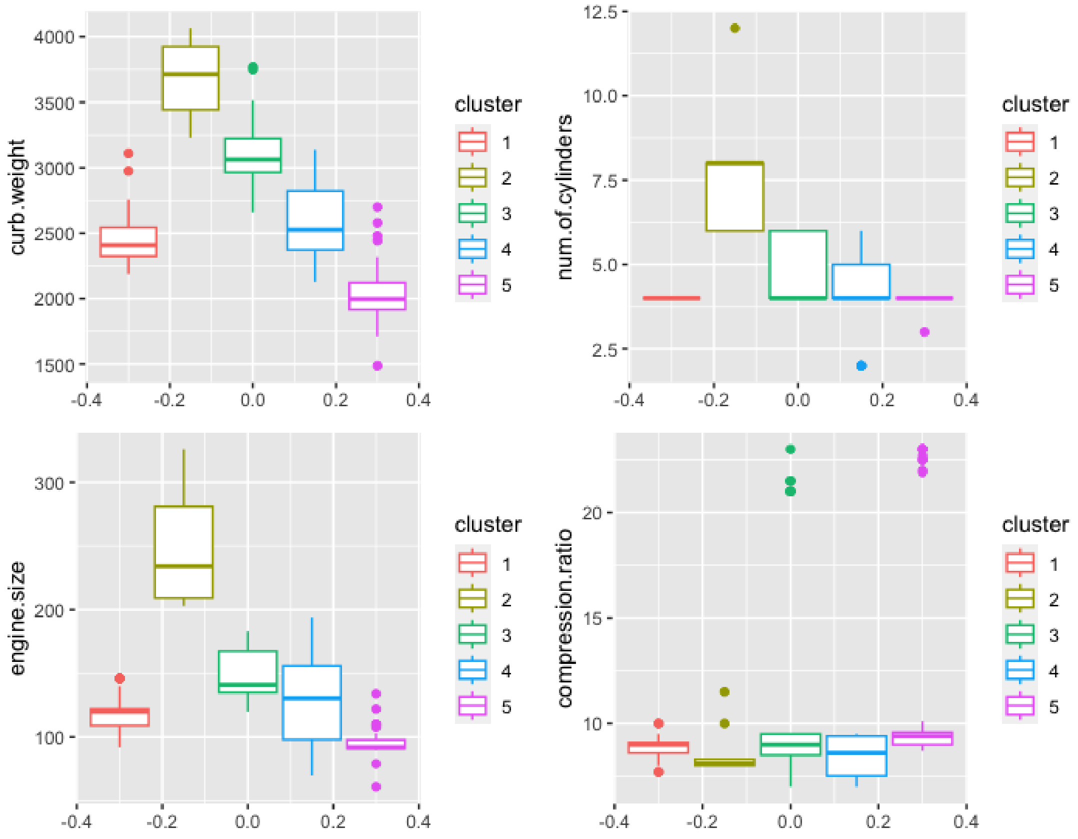
Plot 8a. Clusters across first 3 Principal Components



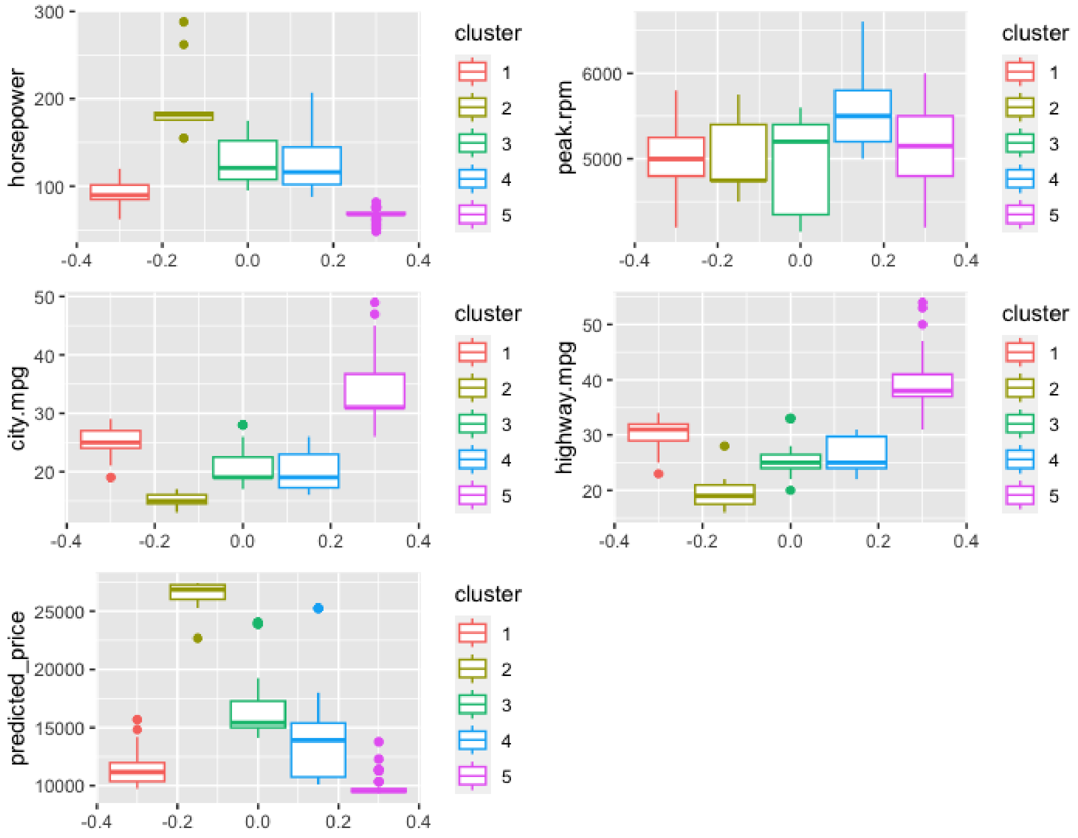
Plot 9a. Boxplot acorss clusters



Plot 9b. Boxplot acorss clusters



Plot 9c. Boxplot acorss clusters


horsepower
peak.rpm
city.mpg
highway.mpg
predicted_price
cluster