FINAL PROJECT REPORT

ON

PRICE AND SYMBOLING PREDICTION MODEL

Harsh Hareshkumar Shukla
Khushru Irani
Nisarg Patel
Rama Mani Deepika Maram
Trupti Kirve
Vivek Bhavsar

TABLE OF CONTENTS

ABSTRACT

The market for automobiles today has increased tremendously and so has the price range of the automobiles having similar specifications and features. We picked our dataset from ***UCI Machine Learning Library*** resource and performed analysis on this dataset. The dataset consist of 205 instances and 26 variables. Our project work, moves around these variables to identify how they influence the price of the vehicle. The data was collected as a part of 1985 Ward's Automotive Yearbook , Personal auto manuals of Insurance Service Office and Insurance Collision Report by the Insurance Institute for Highway Safety . Ward is an American organization which has covered the automotive industry for about 80 years.

Our study helps us figure out a pricing model and various correlations between the features to help us predict the price in the best way. We use Linear Regression Model to predict the price of the automobile. We also analyzed different aspect of the selected dataset and came up with a Decision Tree Model to predict the symboling feature in the dataset. The symboling feature explains the risk associated with the automobile. This feature will help buyers to make decisions based on the risk associated on the automobile.

Various exploratory analysis techniques are used before devising the machine learning models. The techniques includer Principal Component Analysis, Factor Analysis, Canonical Correlation Analysis, Correspondence Analysis. All these techniques create a subset of dataset. These techniques explain the proportion of variance and explain the relationship between the variables.

The results of the Linear Regression Model when performed on a set of variables devised from Principal Component Analysis shows an *R squared value of 0.94*. The variables under consideration are horsepower, no. of cylinders, fuel-type, aspiration and width.

The results of Decision Tree are helpful in order to determine and build a predictive model for symboling variable to determine the risk level of considering insurance of a vehicle based upon key determining points such as engine, brand, horsepower and all to check sustainability of the model with decent amount of accuracy in training and testing set with less amount of complexity and depth. The model uses only five key variables for prediction which makes it faster.

INTRODUCTION

As discussed the automobile industry has increased rapidly over few decades including car price, the dataset we worked on was from 1985 which has quite important variables that were retrieved from Ward's Automotive yearbook from 1985, Personal auto manuals of Insurance Service Office and Insurance Collision Report by the Insurance Institute for Highway Safety. Ward is an American organization which has covered the automotive industry for about 80 years. We have selected our dataset from UCI Machine Learning Library Resources and when we were doing initial stage we learned the dataset is made of 205 observations and 26 different variable from ordinal, categorical to continuous. We have performed various analysis on data to predict the price and risk of the insurance since they both seemed dominant variables to predict based upon domain knowledge also we dived dipper to look for different aspects of the dataset.

Data Preprocessing:

Data Preprocessing a technique which involves transforming the raw data into an understandable format. Real-world data is often incomplete, noisy, inconsistent. In-order to get quality results to we need to feed in quality data. To achieve this we performed data preprocessing.

- ***Cleaning and filling in Missing values:*** We found that there are few missing values across the data.
  - ***Price:*** We started with the Dependent variable Price having four missing values. Out of 205 instances in the dataset, there were four observations that had missing values for Price, we consider to remove them as Price of the automobile may depend on different factors like brand, mileage etc, which doesn't make sense taking the average***.***
  - ***Normalized_losses:*** There were 37 missing values. Filled in these missing values using SPSS. The data transformation technique used to fill in missing values was by the mean of nearby points.
  - ***No_of_doors:*** There was one missing value. We filled it manually by considering the values of make, aspiration, engine_loc.
  - ***Bore and stroke:*** There were four missing values each. Research as per the make, body_style, wheel_drive gave a direction to fill in the values manually.
  - ***Horsepower and peak_rpm:*** There were two missing values in each. Considering the make of the car we filled in values manually.

- ***Identifying Outliers:*** An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Box-plot technique is used to identify the outliers.

- ***Box-plot of Price vs Make:*** was plotted to identify the outliers in the price: We found that certain automobiles belonging to certain brands have high price. One automobile belonging to dodge, 3 of Honda, 2 mitsubishi, 1 of plymouth and 4 of toyota. Due to the limitation of data, we cannot just smooth out these values as we know, few brands have automobiles which are expensive. Refer appendix for the plot.

  A similar analysis was performed to identify outliers in curb_weight, engine_Size, horsepower and peak_rpm.
  Again, due to limited availability of the data we cannot just smooth out these values. The data points may be actual and possible inflection points. Please see Appendix B for reference.

- ***Distributions:***

- The first histogram below depicts information on the distribution of price variable. Moreover, it is not normally distributed but it is skewed to the right.
- The second histogram below depicts information on the distribution of Highway Mpg variable. Moreover, the histogram is slightly right skewed and not normally distributed.
- The third histogram below depicts information on the distribution of City Mpg variable. Moreover, the histogram is slightly skewed to the right and not normally distributed.
- The fourth histogram below depicts information on the distribution of Horsepower variable. Moreover, the histogram is skewed to the right and not normally distributed. Please see Appendix C for reference.

- ***Correlations:*** To identify the relationships between numeric variables we plotted a correlation matrix. The correlation matrix shows that there are strong relationships between variables.

- The selected dataset has few instances (205) and 26 variables. However, due to the strong correlations among the variables we were able to choose these data set. Having such strong relationships among the variable can lead to multicollinearity issue.
- To overcome this issue, we identified the variables with show high correlations and decided to input only one variable in the model when designing the Linear Regression Model.
- A Principal Component Analysis was also performed to identify the important components in the dataset and work with those as final variables for the Linear Regression Model.
- The purpose was to remove the redundant information from the dataset and also achieve parsimony. Please see Appendix D for reference.
- We have also checked spearman's correlation for decision tree building and results were positive since there wasn't correlation among categorical variables.

LITERATURE REVIEW

- ***Related Work:***

Predicting the price of the Automobile has been an interesting topic for research in Machine Learning. We found many research papers on similar topic. One such research was carried out for predicting the price of used automobiles in Mauritius. They used multiple linear regression, decision trees and k-nearest neighbors, in order to predict the prices. The comparison of the prediction results from these techniques showed that the prices from these methods are closely comparable. However, according to there research decision tree was unable to classify and predict numeric values. The research also concluded that the limited number of instances in data set do not offer high prediction accuracies.

- ***Method Selection:***

To perform the analysis of predicting the price of the automobile, the data is collected from UCI Machine Learning Repository. Multiple sources are used for data collection like the 1985 Ward's Automotive Yearbook, Personal auto manuals of Insurance Service Office and Insurance Collision Report by the Insurance Institute for Highway Safety. See Appendix A for reference.

METHODS

- ***Principal Component Analysis:***

PCA was performed on the dataset to identify the most significant variables that are affecting the price variable. Moreover, we have highly correlated data so PCA will work well on the numeric variables also the variables which are highly correlated with other variables and variables that have weak or zero correlation are removed. Also the variables with zero variance are eliminated from the dataset. However, PCA concluded that total 14 components could explain 100% proportion of variance.

Furthermore, when generated using the Kaiser-Mayer[1] scree plot we determined that 78.2% proportion of variance can be explained with the help of first three components.

Table: Principal Component Analysis with Varimax Rotation for Automobiles

	PC1	**PC2**	**PC3**
Component 1			
Wheel_base	0.657	0.630	
Length	0.813	0.469	
Width	0.818	0.381	0.122
curb_weight	0.913	0.320	
engine_size	0.863	0.116	0.237
bore	0.690	0.195	-0.179
horsepower	0.901	-0.249	
city_mpg	-0.912	0.217	0.126
highway_mpg	-0.924	0.132	0.123
price	0.871	0.102	0.139
Component 2			
Height	0.137	0.726	-0.398
Compression_ratio	-0.140	0.715	0.413
peak_rpm		-0.724	
Component 3			
Stroke	0.106		0.834

Table: Components Explained

Component	Generalized Description	Component Loading	Percent of Variance Explained
1	Component 1	7.124	50.9%
2	Component 2	2.618	18.7%
3	Component 3	1.210	8.6%

Formulae for Each component:

Component 1 = *0.657 wheel_base + 0.813 length + 0.818 width + 0.913 curb_weight + 0.863 engine_size + 0.690 bore + 0.901 horsepower - 0.912 city_mpg - 0.924 highway_mpg + 0.871 price*
Component 2 = *0.726 height + 0.715 compression_ratio - 0.724 peak_rpm*
Component 3 = *0.834 stroke*

- ***Principal Component Regression:***

We have extended the Principal Component Analysis and performing Principal Component Regression. Assumption of **PCR** is that the directions in which the predictors show the most variation are the exact directions associated with the response variable. The Principal Component Regression was performed with the help of pcr and caret packages. Based on the Validation Plot we see that the RMSE is round 4200. Hence, we see that it is greater than that of Linear Regression model which is why we think this is not a best model for our investigation.

- ***Principal Component Analysis for Dummy Variables using PCA mix data:***

Since the data consisted of many categorical variables, it made sense to explore how the PCA results would be by converting Categorical data to Dummy variables and then performing Principal Component Analysis. This was possible by exploring the package PCA Mixdata.

- This package divides the variables into Quantitative and Qualitative and the performs Principal Component Analysis
- We analysed the Coefficients and Squared loading from the results. Refer Appendix C.
- The observation was that the 100% variance is Explained by 62 dimensions which was a lot as each of the dimension explained 1- 2% of variance
- For which, we consider *19 dimensions/components* making sure all the important variables are taken into consideration and explains *74.99949%* of the total variance
- The sum of rotation variance is 39.5578 which is considerably less
- This is not a suitable method for this dataset as it requires many dimensions to achieve a threshold

variance or to explain the important variables

Therefore, our investigation towards PCA was exploring the components by fitting a *Principal Regression Model*, we find that the Root Mean square Error is high than compared to that of the *Linear Regression* Root Mean Square Error. Also from the results we observe that using the dummy variables might not be not an efficient choice

.

- ***Common Factor Analysis:***

To perform the Common Factor Analysis on the dataset, Initially PCA was performed and the result showed that total 14 components/ factors could explain 100% proportion of variance. However, CFA concluded that with the help of only first three components 71.2% proportion of variance can be explained.

	Factor1	**Factor2**	**Factor3**
Factor 1			
length	0.681	0.649	
width	0.729	0.493	
curb_ weight	0.867	0.407	
engine_ size	0.919		
Bore	0.611		
horsepower	0.934		
city_ mpg	-0.830		0.483
highway_ mpg	-0.835		
price	0.893		
Factor 2			
wheel_ base	0.506	0.739	
height		0.675	
Factor 3			

compression_ ratio			0.637
Peak _rpm			-0.509

Table: Components Explained

Factors	**Generalized Description**	**Component Loading**	**Percent of Variance Explained**
1	Factor 1	6.308	45.1%
2	Factor 2	2.211	15.8%
3	Factor 3	1.447	10.3%

Formulae for Each Factor:

Factor 1 = *0.681 length + 0.729 width + 0.867 curb_weight + 0.919 engine_size + 0.611 bore + 0.934 horsepower - 0.830 city_mpg - 0.835 + 0.893 highway_mpg price*
Factor 2 = *0.739 wheel_base + 0.675 height*
Factor 3 = *0.637 compression_ratio - 0.509 peak_rpm*

Furthermore, Factor analysis gives 71.2% proportion of variance with first three factors while PCA gives 78.2% proportion of variance. This is because PCA uses all unique, error and shared variance while Factor Analysis only uses the shared variance. Please see Appendix F for reference.

- ***Canonical Correlation Analysis:***

Canonical correlation analysis(CCA) is a method used to identify and measure the relationship among two sets of variables. Canonical correlation analysis determines a set of canonical variates, orthogonal linear combinations of the variables within each set that explains the variability both within and between the sets.

The selected dataset does not directly fit for CCA. To come up with sets of Independent Variables(IV) and Dependent Variables (DV), certain amount of exploratory analysis was required. The best approach to carry out this was to conduct a Principal Component Analysis or Common Factor Analysis. From these two techniques.

We can devise the significant components which explains maximum variance. For the selected dataset I performed CFA and identified the first two factors.

Factor 1 consist of certain set of variables which I marked as Y variables and factor consist of set of variables which I marked as X variables. In case of cross - loadings, as per the higher weights the variables were assigned the variables to X or Y set.

A Canonical Correlation Analysis was performed on the above two factors. To identify the significance of the variates a Wilk's Lambda test was performed. The wilk's L test shows that all the variates are significant and the p-value is < 0.05 for all the variates. The first variate as per the p-value tells that all are significant, the second says all below it are significant.
Here, we choose only two variates as we choose only 2 factors from Common Factor Analysis to proceed with CCA.

Wilk's L	F-test	df	P-values
0.11	14.3	40	0.00
0.4	7.56	27	0.00
0.65	5.3	16	0.00
0.92	2.47	7	0.02

Fig: Wilk's Lambda Test (shows significance)

The two canonical correlation values for the first two variates were calculated using the cor function and for variate 1 the value is 0.8555 and 0.6244.

To understand the relationship and the weights of each variable in the variate we identified the standardized coefficients.

The covariates for Factor 1 are as follows:

CV1 = *-0.1066price -0.427wheel_base - 0.3404length + 0.2683width - 0.8385curb_weight + 0.0733engine_size - 0.127bore + 0.4913horsepower – 0.8489city_mpg + 0.067highwaympg*
CV2 = *0.045price + 1.2850wheel_base + 0.6405length – 0.7960 width– 1.4545curb_weight - 0.4048engine_size – 0.1800bore + 0.5695horsepower – 0.3044city_mpg – 0.5747highwaympg*

The covariates for Factor 2 are as follows:

CV1 = *0.3023 symboling– 0.3681height – 0.4352compression_ratio + 0.3043peak_rpm*
CV2 = *-0.3181 symboling + 0.6701 height – 0.6016compression_ratio + 0.4005peak_rpm*

The purpose of CCA is to show how the variables are related to each other. Hence, covariate 1 shows high dependency on variables like price, bore,curb_weight and city_mpg. Covariate 2 shows high relationship with compression_ratio.

- ***Correspondence Analysis:***

Correspondence Analysis is a visual method where we can have a contingency table with rows and columns, such that the positions of the row and column points are consistent with their associations in the table. To get a final global outlook of the categorical variables/factors in our data set.

Using tableau created contingency table for the first pair of variables body_styles and drive wheels. It was done by simply counting the number of combinations that exists for each of the possible levels of the categorical variables for the pair.

	Body Style				
Drive Wheels	convertible	hardtop	hatchback	sedan	wagon
4wd			1	3	4
fwd	1	1	49	55	12
rwd	5	7	18	36	9

The correspondence analysis of 2 categorical variables(body_styles vs drive wheels) is as seen in the plot. Also if we try and analyse the same, by drawing a line from the origin to the “fwd” level of Drive Wheels feature in our model, we can infer that more hatchbacks and sedan correspond to fwd drive type of vehicles.

More analysis regarding other pairs of categorical variables is on-going and we are looking into Multi Correspondence Analysis to include more than pairs of variables for the same. Please see Appendix G for reference.

- ***Decision Trees:***

Decision tree is used for predicting the risk of auto insurance but here we have used Symboling variable to conduct different aspect for the data. Moreover, the data was divided into training(66%) and testing(34%) sets also multiple cases for parent node and child node were performed to come up with the optimal decision tree.
Note: to select the % for training and testing we have run various different combination from 50% division to 90 and 10% division which can be seen in line graph in appendix F and out of all we have found 66% training and 34% testing are most accurate and with least gap between training and testing accuracy consequently we have decided to go with that.

In here to build up the binary decision tree classifier we have used Crt method moreover in order to measure impurity in nodes I have decided to use Gini index which usually yield the purest quality of nodes for decent model building techniques.

Furthermore, out of all cases that was performed for parent and child the Np=20 and Nc=10 was optimal amongst all as it gave high accuracy in both training and testing with less complexity and optimal depth too. However, the importance of the independent variables in descending order were wheel_base, width and make. See Appendix F for reference.

Rules of decision tree:

- *If number_of_doors == two and city_mpg_improvement<=21.5 then symbolling = 3*
- *If number_of_doors == two and city_mpg > 21.5 and bore_improvement<= 3.41 then symbolling = 1*
- *If number_of_doors == two and city_mpg > 21.5 and bore_improvement > 3.41 then symbolling = 2*
- *If number_of_doors == four and make_improvement == “ ’bmw’, ’Chevrolet’, ‘Honda’, ‘isuzu’, ‘jaguar’, ‘mazda’, ‘nissan’, ‘peugot’, ‘renault’, ‘subaru’, ‘toyota’, ‘volkswagen’” then symbolling = 0*
- *If number_of_doors == four and make_improvement == “ ’audi’,’dodge’,’mercedes-benz’,’mitsubishi’,’plymouth’,’saab’, ‘volvo’ ” normalized_losses_improvement <= 123.5 then symbolling = -1*
- *If number_of_doors == four and make_improvement == “ ’audi’,’dodge’,’mercedes-benz’,’mitsubishi’,’plymouth’,’saab’, ‘volvo’ ” normalized_losses_improvement <= 123.5 then*

symbolling = 1

- ***Linear Regression:***

In Linear regression given an input **x∈R**, where x1, . . . , xm represent predictors (also independent variables), we find a prediction yˆ∈R for the price of the automobile **y** ∈ **R** using a linear regression model. We evaluate this model base on R -squared value. First we used the label encoder to obtain the dummy variables for categorical data after that we tried to remove the multicollinearity among the variable so for that we check for correlation among variables and find the VIF for the variables against the target variable i.e. price. After doing that process we removed the unnecessary variables with high VIF and high correlation amongst each other and finally we reduce the predictors from 25 to 5.

Now to build the model we have used k fold cross validation method. Now when our model is ready we will check for p-values, r-squared, Adj r-squared and F-values.

We have tried different splits of Test and Train data such as 50-50, 60-40, but 80-20 seems to be having good accuracy and other estimates comparatively.

Out[31]:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.947
Model:	OLS	**Adj. R-squared:**	0.945
Method:	Least Squares	**F-statistic:**	575.6
Date:	Wed, 06 Jun 2018	**Prob (F-statistic):**	4.14e-121
Time:	11:32:18	**Log-Likelihood:**	-1929.0
No. Observations:	201	**AIC:**	3870.
Df Residuals:	195	**BIC:**	3890.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
horsepower	183.3848	7.786	23.553	0.000	168.029	198.740
make	-195.3183	41.298	-4.729	0.000	-276.767	-113.869
wheel_base	364.2305	68.673	5.304	0.000	228.794	499.667
fuel_type	-8019.5219	994.867	-8.061	0.000	-9981.603	-6057.440
aspiration	-3837.4299	780.579	-4.916	0.000	-5376.891	-2297.969
width	-476.2200	106.804	-4.459	0.000	-686.860	-265.580

Omnibus:	34.774	**Durbin-Watson:**	0.960
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	88.844
Skew:	0.740	**Prob(JB):**	5.10e-20
Kurtosis:	5.902	**Cond. No.**	689.

Table: Evaluation of Linear Regression Model & Final variables used in the model

From the table we can interpret that ***r-squared value 0.94 and F-statistic 575.6***. At last p-value for all the variable is less than 0.0005. RMSE for this model is little high than we expect.we are getting 81.27% accuracy score and our cross predicted accuracy is 73.89%.

- ***Neural Network:***

After doing decision tree we decided to perform neural network analysis with symbolling variable to check for better accuracy, though it is part of future work.

For Neural network analysis we have used multilayer perceptron and standardized rescaling of Covariates With hold out partitioning technique (70% training, 30% testing). Moreover, scaled conjugate gradient for optimization algorithm were used. We kept maximum of 15 minutes for training time to reduce high processing time also the minimum Relative change in training and testing was 0.0001. Furthermore, factors were used for one of the layers of the multilayer analysis. however, it was not feasible and time consuming, hence we concluded to go with categorical variable and continuous variables for various layers.

In nutshell, neural analysis gave ***66.4%*** accuracy in training and ***63%*** accuracy in testing which proves our partitioning technique achieve minimum distance between accuracies. See appendix G for reference.

DISCUSSION AND FUTURE WORK

- ***Conclusion:***

Comparing Principal Component Regression to Linear Regression we notice that the Root-Mean Square Error is high for PCR model than the linear regression model. Moreover, when comparing PCA and Factor analysis we conclude that PCA explains more proportion of variance than Factor analysis.

The linear Regression Model used shows a high R-squared value meaning it explains high variance in the data and hence we can suggest this model to predict the price of the automobile. Also, the model involves few features this helps achieving parsimony.

The Decision tree gives accuracy for training and testing set of data for symboling variable which helps in determining the level of risk in insurance. Furthermore, neural network was also performed though decision tree gave better accuracy with less complexity. In nutshell, we concluded to keep decision tree as our final prediction model for symboling variable.

- ***Limitations and Future work:***

The dataset used for this project has very less number of observations though we used it because when checked for collinearity it was concluded that there is high correlation among the variables. Moreover, different analysis methods were performed with their constraints on variables types. Furthermore, other limitation is that the dataset used was relatively old and had no latest information on the current automobile market as well insurance strategies.

In addition to that, for future research latest data for current automobile industry should be collected with sufficient observations for good prediction on price of different cars in the current market since the automobile is rapidly changing industry. In addition, we need to include the driver ticket or model general tickets to estimate the parameter risk.

In future, we are planning to implement Neural network analysis, cluster analysis and K nearest neighbor analysis to predict possible risk level.

REFERENCES

[1] Uci Machine Learning Repository - Automobile Data Set
https://archive.ics.uci.edu/ml/datasets/automobile

[2] Principal Components Regression, Pt.1: The Standard Method
Nina Zumel - http://www.win-vector.com/blog/2016/05/pcr_part1_xonly/

[3] Decision Tree
https://en.wikipedia.org/wiki/Decision_tree

[4] Multiple Linear Regression
https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python

[5] Author Age Prediction from Text using Linear Regression
https://homes.cs.washington.edu/~nasmith/papers/nguyen+smith+rose.latech11.pdf

[6] Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic Chris Lee-Bruce Hellinga-Frank Saccomanno - Transportation Research Record: Journal of the Transportation Research Board - 2003
https://trrjournalonline.trb.org/doi/abs/10.3141/1840-08

[7] The Prediction of Car Ownership. (m. j. h. mogridge journal of transport economics and policy vol. 1, no. 1 (jan., 1967), pp. 52-74 published by: university of bath)
https://www.jstor.org/stable/20052037?seq=1#page_scan_tab_contents

[8] New Car Sales and Used Car Stocks: A Model Of the Automobile Market
James Berkovec -https://www.jstor.org/stable/2555410?seq=1#page_scan_tab_contents

[9] Price Adjustment in an Automobile Insurance Market: A Test Of the Sheshinski-weiss Model
https://www.jstor.org/stable/135732?seq=1#page_scan_tab_contents

[10] Price Adjustment in an Automobile Insurance Market: A Test Of the Sheshinski-weiss Model Bev Dahlby - https://www.jstor.org/stable/135732?seq=1#page_scan_tab_contents

[11] Principal Component Regression
https://en.wikipedia.org/wiki/Principal_component_regression

[12] Deep Learning Model For Car Price Prediction Using Tensorflow
http://androidkt.com/car-price-prediction/

[13] Ibm Knowledge Center
https://www.ibm.com/support/knowledgecenter/en/SSLVMB_sub/statistics_mainhelp_ddita/spss/neural_network/idh_idd_mlp_output.html

APPENDIX

APPENDIX A: Attribute Description

ATTRIBUTE DESCRIPTION					
#	**ATTRIBUTES**	**TYPE**	**DESCRIPTION**	**VALUES**	**RANGE**
1	Symboling	Ordinal	Risk factor associated with the car	integer	-3 to 3
2	Normalized Losses	Interval	Average loss per car year on year	continuous	65 to 256
3	Make	Categorical	The brand of the car.	22 Levels	NA
4	Fuel_type	Categorical	Determines whether the vehicle is Gas or Diesel type.	2 Levels	NA
5	Aspiration	Categorical	The combustion of the engine	2 Levels	NA
6	Number_of_doors	Categorical	The no. of doors for a vehicle..	2 Levels	NA
7	Body_style	Categorical	Exterior car shape and body style	5 Levels	NA
8	Drive_wheel	Categorical	Type of wheels of the automobile	3 Levels	NA
9	Engine_location	Categorical	The location of the engine	2 Levels	NA
10	Wheel_base	Interval	The distance from the centers of the front wheel and rear wheel.	continuous	86.6 to 120.9
11	Length	Interval	The length of the wheel-base	continuous	141.1 to 208.1
12	Width	Interval	The Width of the wheel-base	continuous	60.3 to 72.3
13	Height	Interval	The height of the vehicle	continuous	47.8 to 59.8
14	Curb_weight	Interval	Engine specification	continuous	1488 to 4066
15	Engine_type	Categorical	The type of the engine.	7 Levels	NA
16	Number_of_cylinders	Categorical	The no. of cylinders	7 Levels	NA
17	Engine_size	Interval	The size of the engine.	continuous	61 to 326
18	Fuel_system	Categorical	Fuel injection technique of the vehicle	8 Levels	NA
19	Bore	Interval	Size in diameter of the cylinder	continuous	2.54 to 3.94
20	Stroke	Interval	How far the piston travels inside the cylinder	continuous	2.07 to 4.17
21	Compression_ratio	Interval	Specification for many combustion engines	continuous	7 to 23
22	Horsepower	Interval	Power of the Engine	continuous	48 to 288
23	Peak_rpm	Interval	Revolutions per minute	continuous	4150 to 6600
24	City_mpg	Interval	Mileage of the car on city roads	continuous	13 to 49
25	Higway_mpg	Interval	Mileage of the car on Highways	continuous	16 to 54

26	Price	Interval	The price of the vehicle	continuous	5118 to 45400

APPENDIX B: Visualizations

- **EDA Using Boxplots**

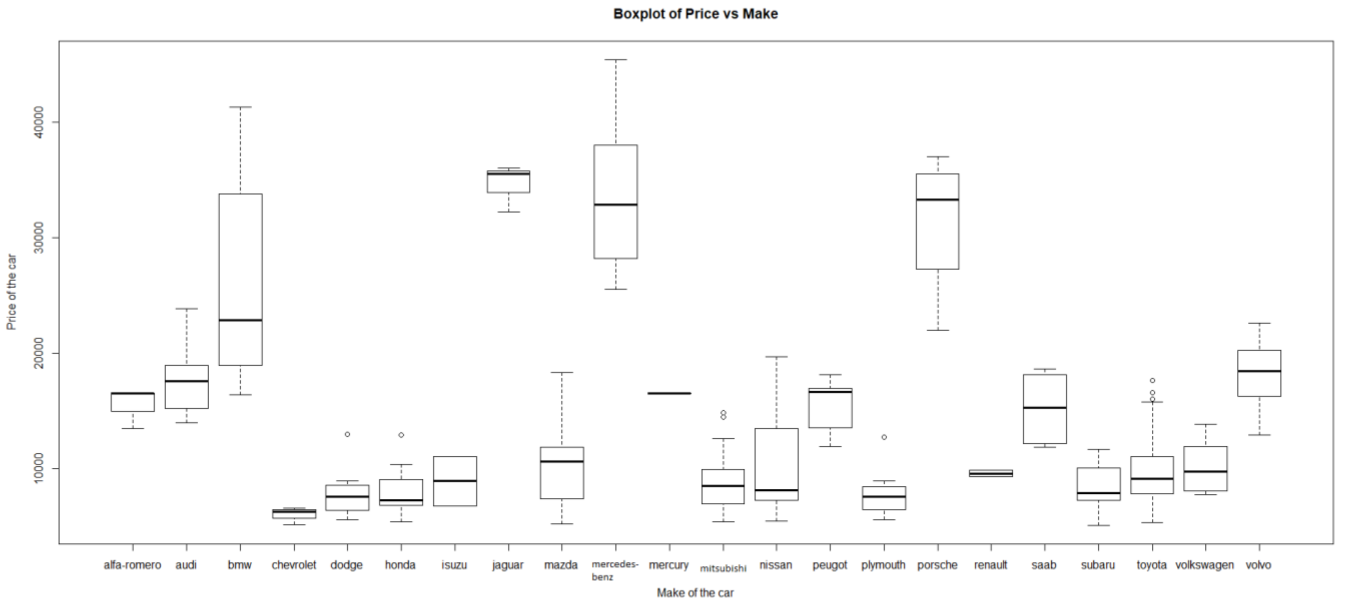


Fig: Box-Plot of Price vs Make of the car

We can identify that there are few outliers for certain make of the automobile like dodge, honda, mitsubishi, plymouth and toyota.

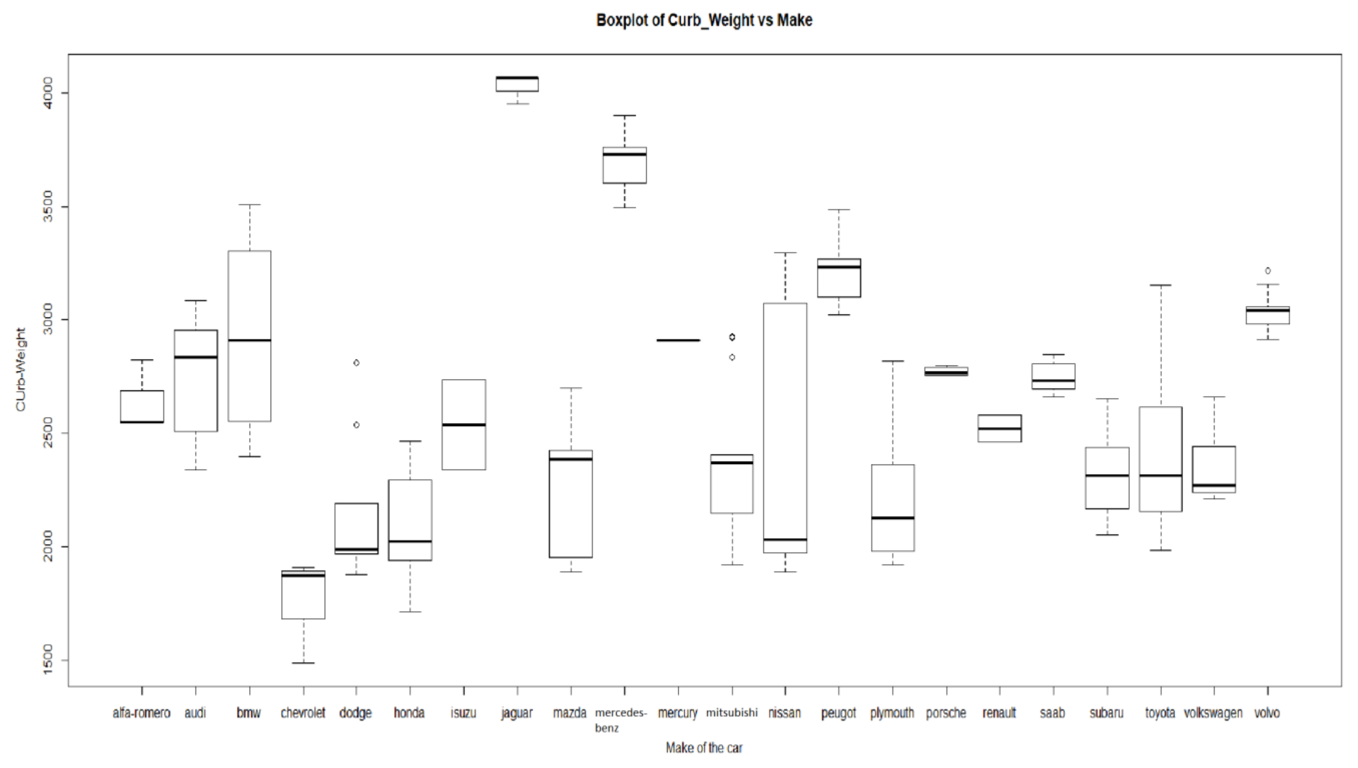


Fig: Box-Plot of Price vs Curb-Weight

There are certain outliers for the make dodge, mitsubishi and volvo.

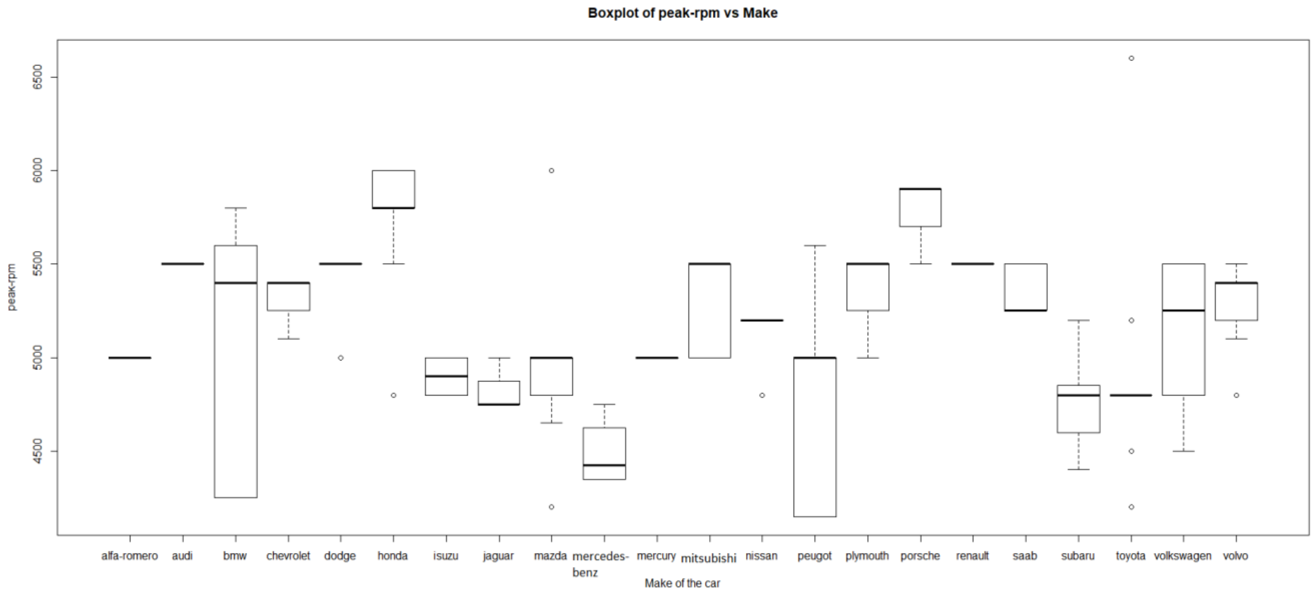


Fig: Box-Plot of Price vs Peak-rpm

- **Histograms**

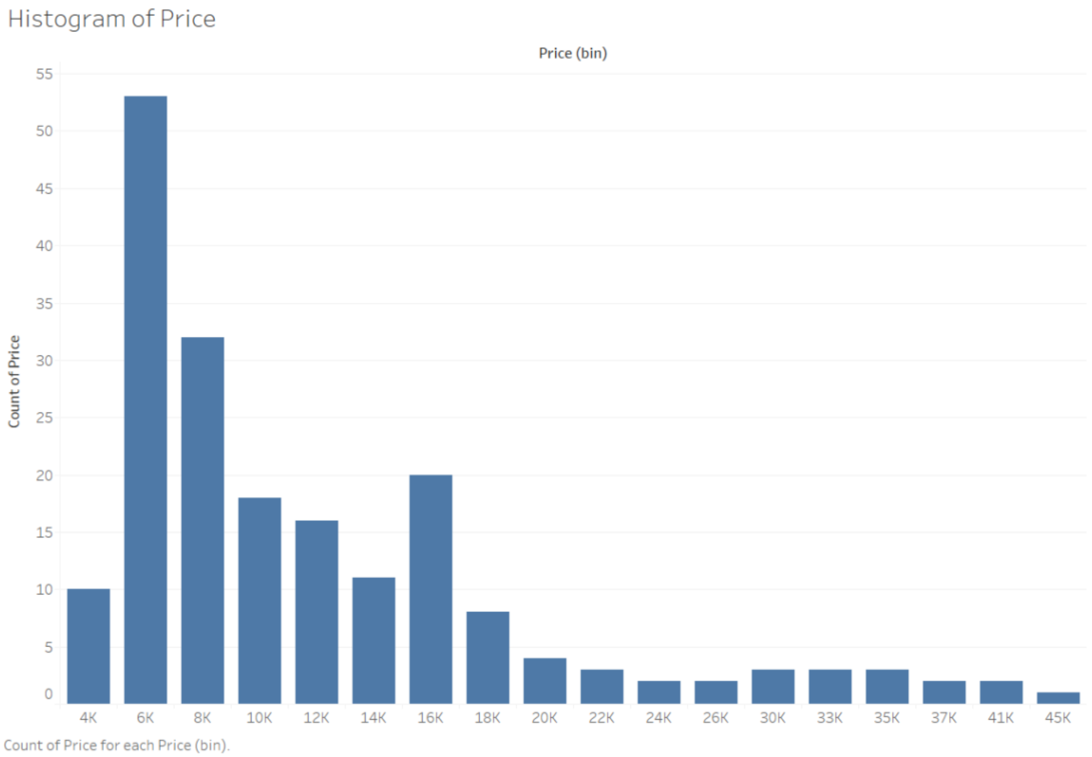


Fig : Histogram of Price

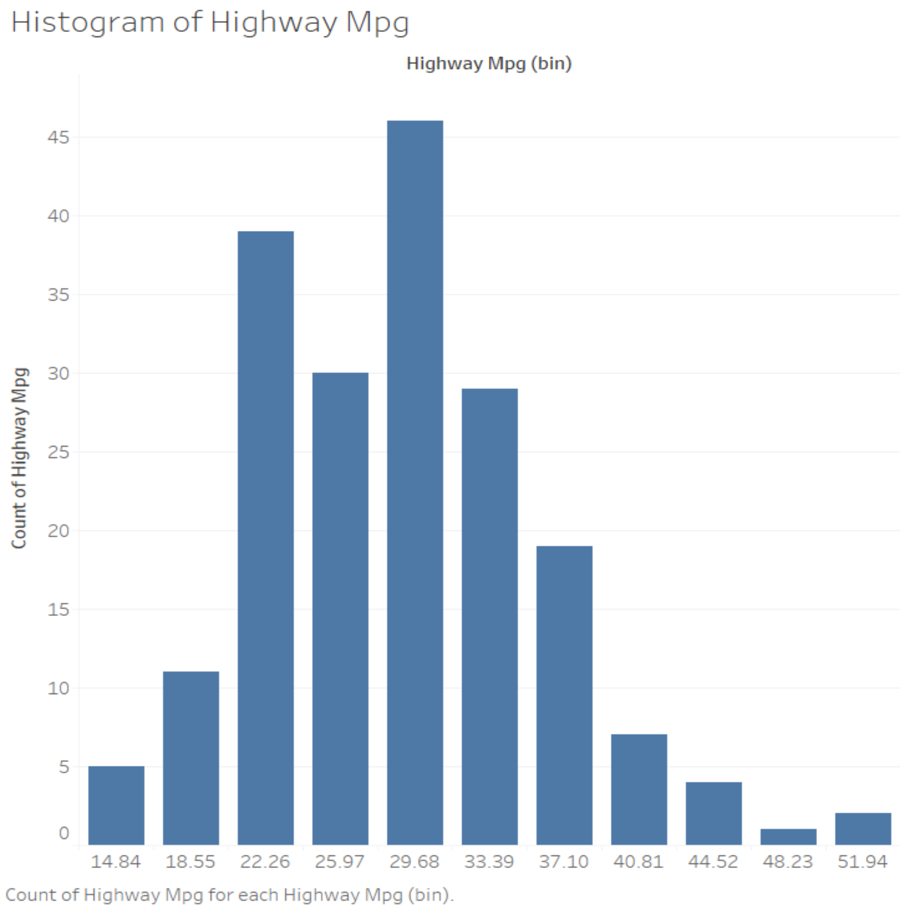


Fig : Histogram of Highway Mpg

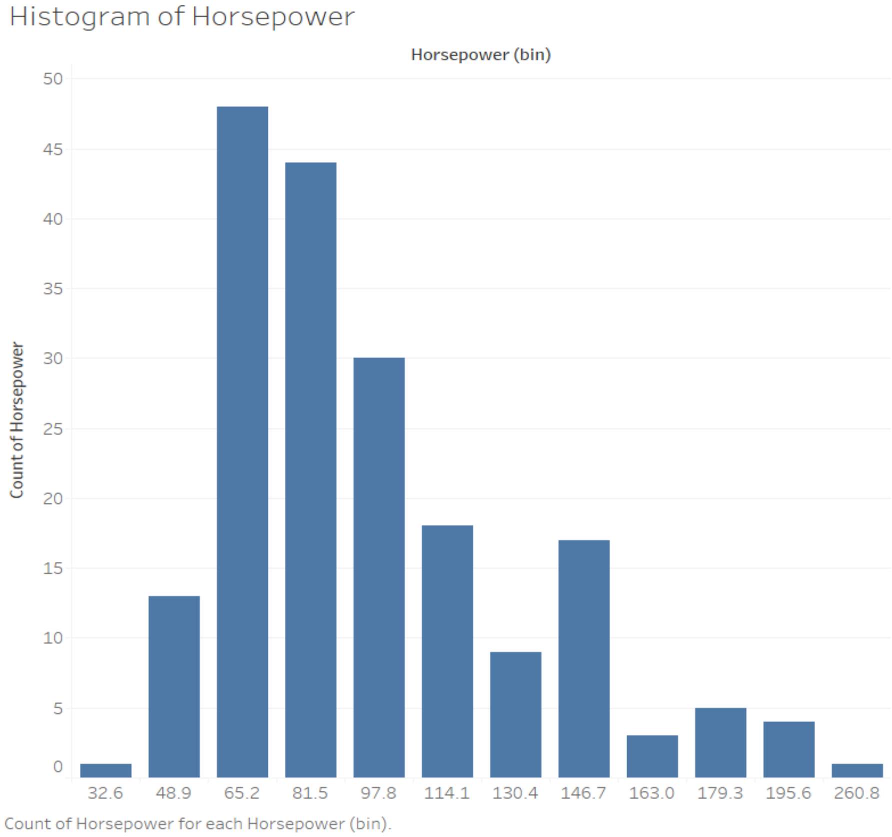


Fig : Histogram of Horsepower

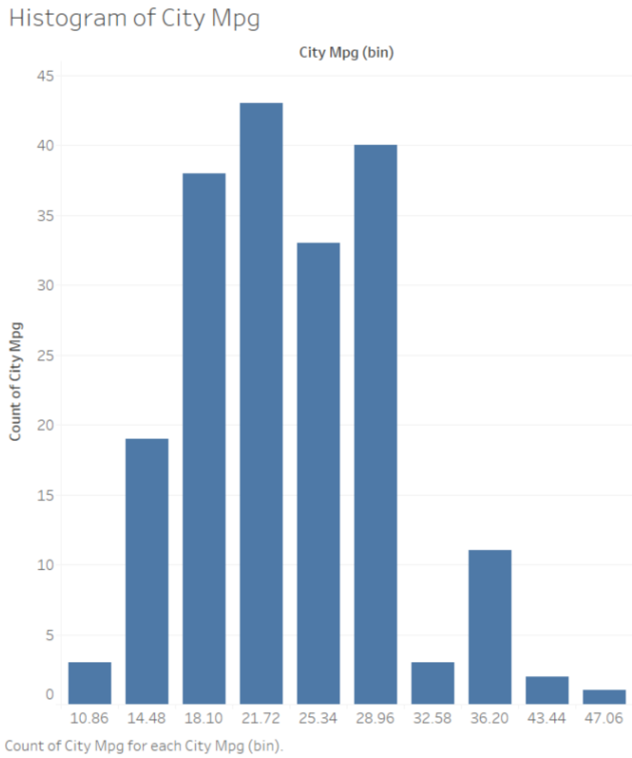


Fig : Histogram of City Mpg

- **Correlation Plots:**

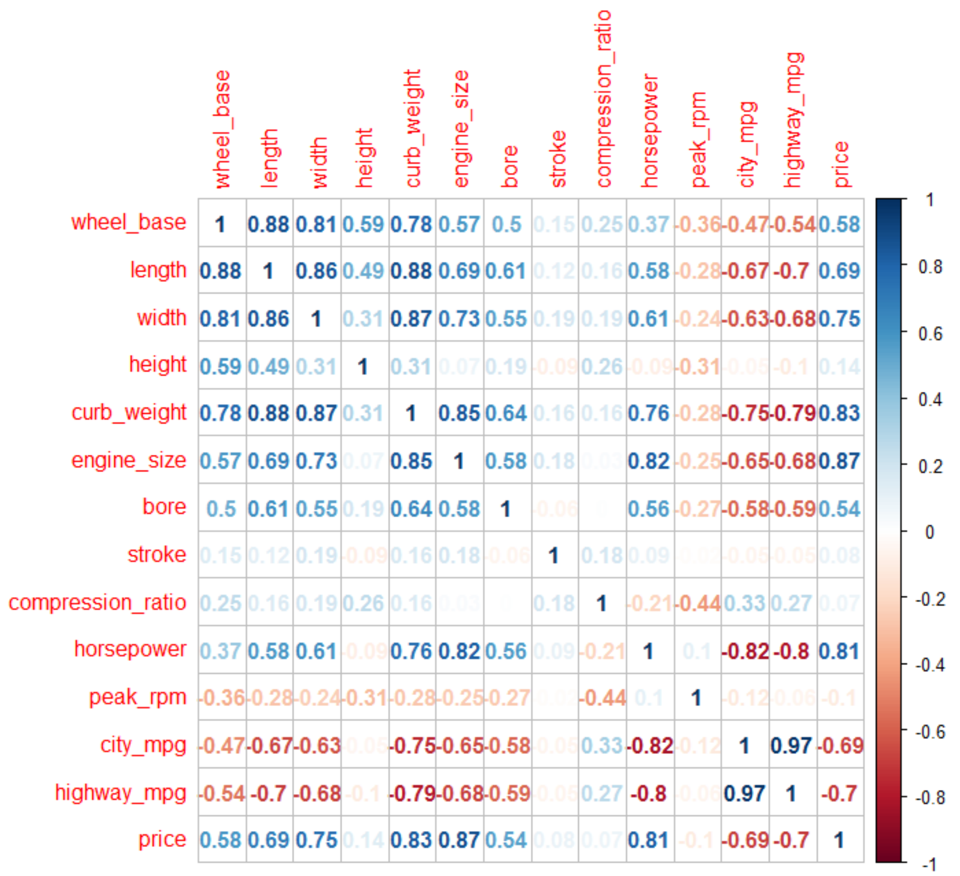


Fig: Correlation plot

A correlation value of and above 0.7 shows strong relationship. The information provided by these variables is redundant. As we can visualize from our plot that variables like length,width,curb_weight, wheel_base, engine_size, horsepower, city_mpg, highway_mpg, price all show strong relationship.

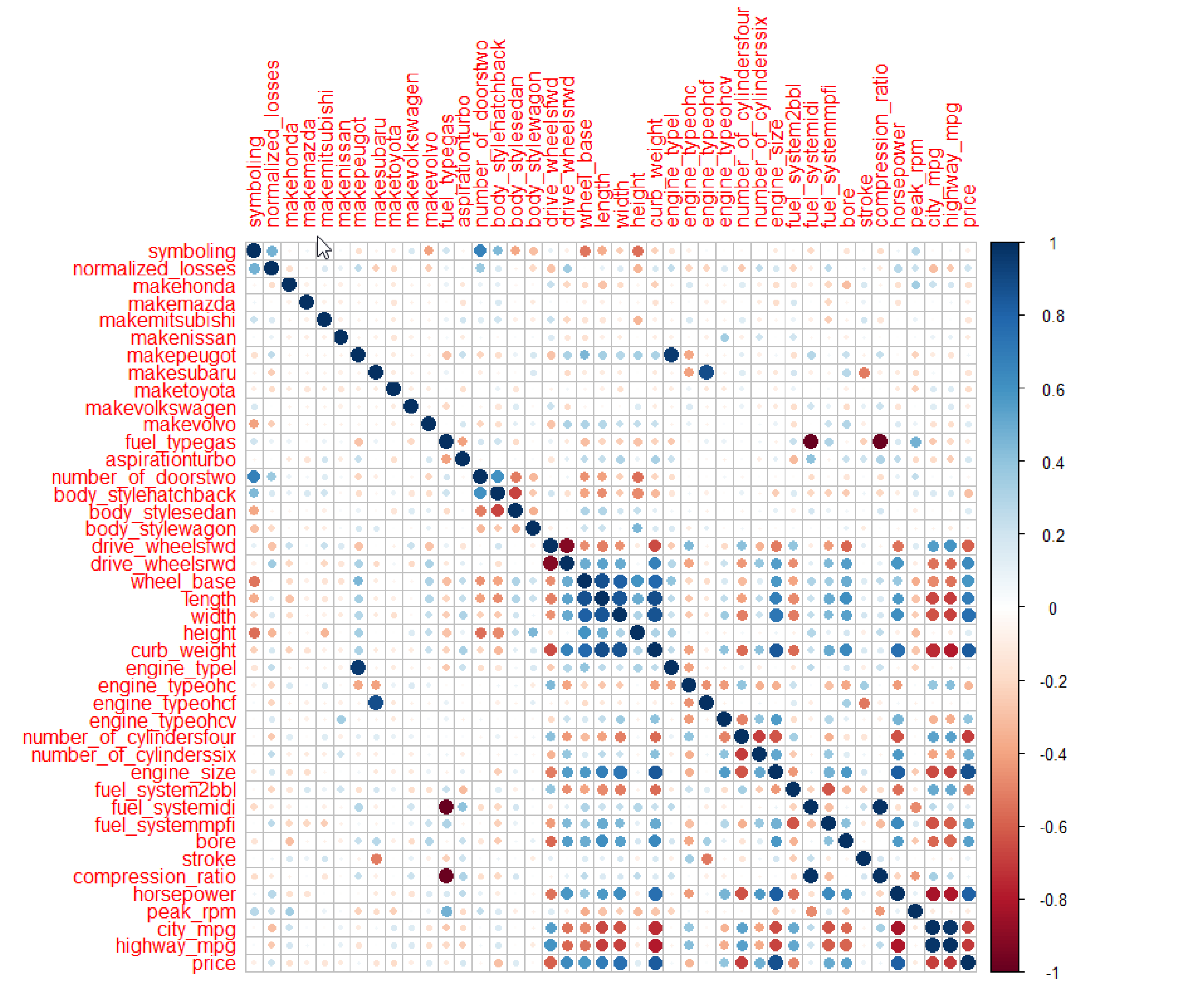


Fig: Correlation(Categorical variables converted as Dummy variables)

APPENDIX C: Principal Component Analysis

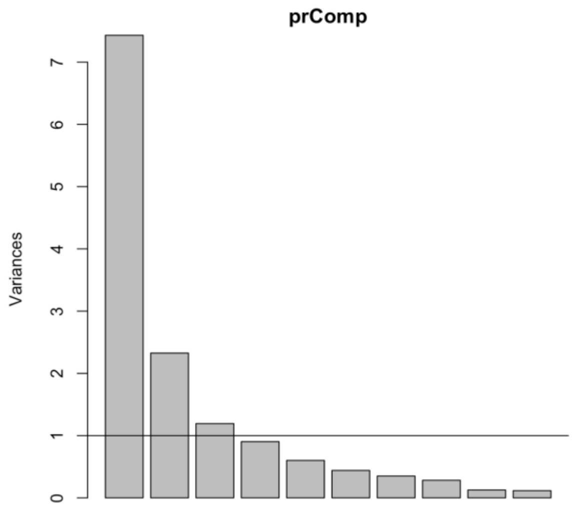


Fig: Scree Plot to determine the number of components to be considered

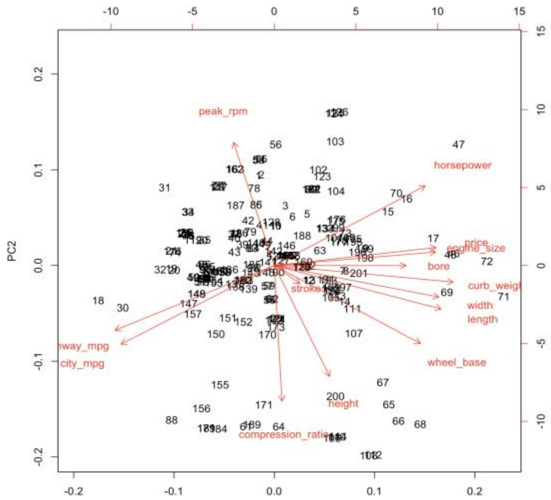


Fig: Biplot of the components

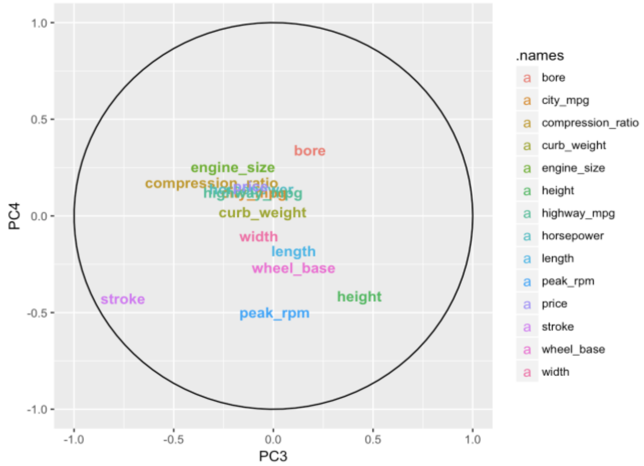


Fig: PCA Plot

- ***Principal Component Regression***:

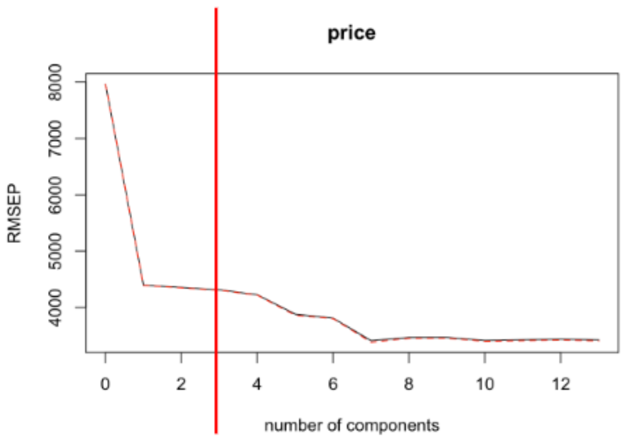


Fig: Validation Plot showing RMSE

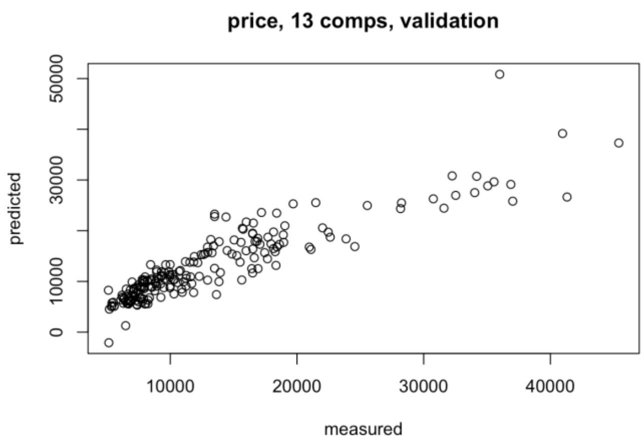


Fig: Prediction Plot

- ***Principal Component Analysis for Dummy Variables using PCA mixdata:***

	Eigenvalue	**Proportion**	**Cumulative**
dim 1	10.308834760	15.859745785	15.85975
dim 2	5.686936489	8.749133060	24.60888
dim 3	3.676441100	5.656063231	30.26494
dim 4	3.084942545	4.746065453	35.01101
dim 5	2.955436702	4.546825695	39.55783
dim 6	2.419595964	3.722455330	43.28029
dim 7	2.315772933	3.562727590	46.84302
dim 8	2.213097733	3.404765744	50.24778
dim 9	1.947845221	2.996684955	53.24447
dim 10	1.924883101	2.961358617	56.20583
dim 11	1.795204816	2.761853564	58.96768
dim 12	1.778403098	2.736004767	61.70368
dim 13	1.660582577	2.554742426	64.25843
dim 14	1.520040737	2.338524211	66.59695
dim 15	1.426346235	2.194378823	68.79133
dim 16	1.368078225	2.104735731	70.89606
dim 17	1.352963962	2.081483018	72.97755
dim 18	1.314263477	2.021943811	74.99949
dim 19	1.257467740	1.934565754	76.93406
dim 20	1.154774206	1.776575701	78.71063

dim 21	1.120348670	1.723613338	80.43425
dim 22	1.027941144	1.581447914	82.01569
dim 23	0.980981224	1.509201883	83.52490
dim 24	0.935356279	1.439009659	84.96391
dim 25	0.872122566	1.341727024	86.30563
dim 26	0.818146272	1.258686573	87.56432
dim 27	0.767782073	1.181203190	88.74552
dim 28	0.723314651	1.112791771	89.85831
dim 29	0.645351254	0.992848083	90.85116
dim 30	0.627829751	0.965891925	91.81705
dim 31	0.558570561	0.859339325	92.67639
dim 32	0.488582652	0.751665619	93.42806
dim 33	0.468480017	0.720738488	94.14880
dim 34	0.434913764	0.669098099	94.81790
dim 35	0.360315194	0.554331067	95.37223
dim 36	0.334270030	0.514261585	95.88649
dim 37	0.319455069	0.491469336	96.37796
dim 38	0.273117595	0.420180916	96.79814
dim 39	0.242223047	0.372650842	97.17079
dim 40	0.212881217	0.327509564	97.49830
dim 41	0.192247172	0.295764881	97.79406
dim 42	0.174822846	0.268958224	98.06302

dim 43	0.165306798	0.254318151	98.31734
dim 44	0.158235637	0.243439441	98.56078
dim 45	0.138900506	0.213693086	98.77447
dim 46	0.120835311	0.185900479	98.96037
dim 47	0.108245941	0.166532217	99.12691
dim 48	0.098589610	0.151676322	99.27858
dim 49	0.088856054	0.136701622	99.41528
dim 50	0.066923388	0.102959058	99.51824
dim 51	0.055703158	0.085697166	99.60394
dim 52	0.051978293	0.079966605	99.68391
dim 53	0.049951953	0.076849158	99.76076
dim 54	0.040610809	0.062478168	99.82323
dim 55	0.037096460	0.057071476	99.88031
dim 56	0.024070539	0.037031599	99.91734
dim 57	0.018502219	0.028464952	99.94580
dim 58	0.013671268	0.021032720	99.96683
dim 59	0.011851046	0.018232378	99.98507
dim 60	0.007266625	0.011179424	99.99625
dim 61	0.002439715	0.003753407	100.00000

APPENDIX D: Common Factor Analysis

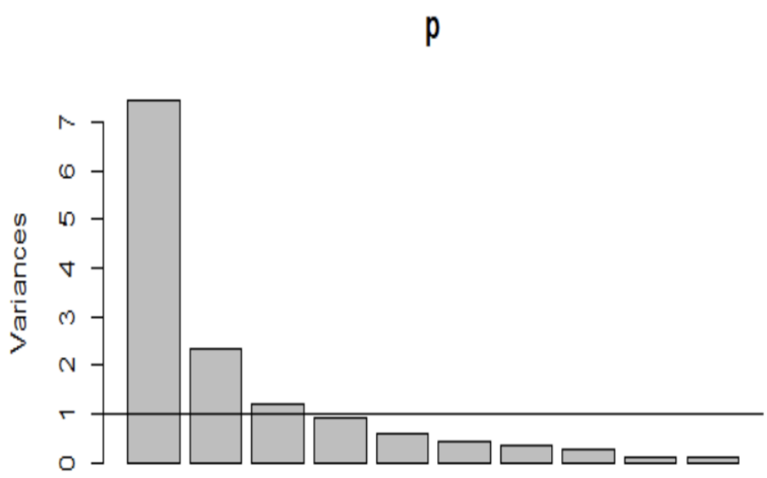
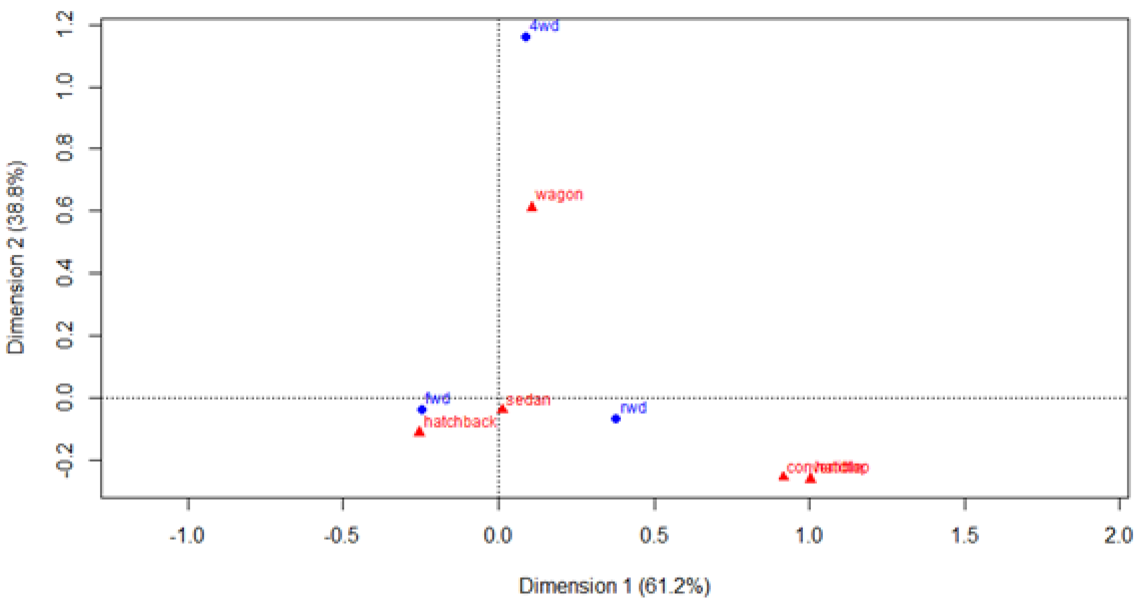
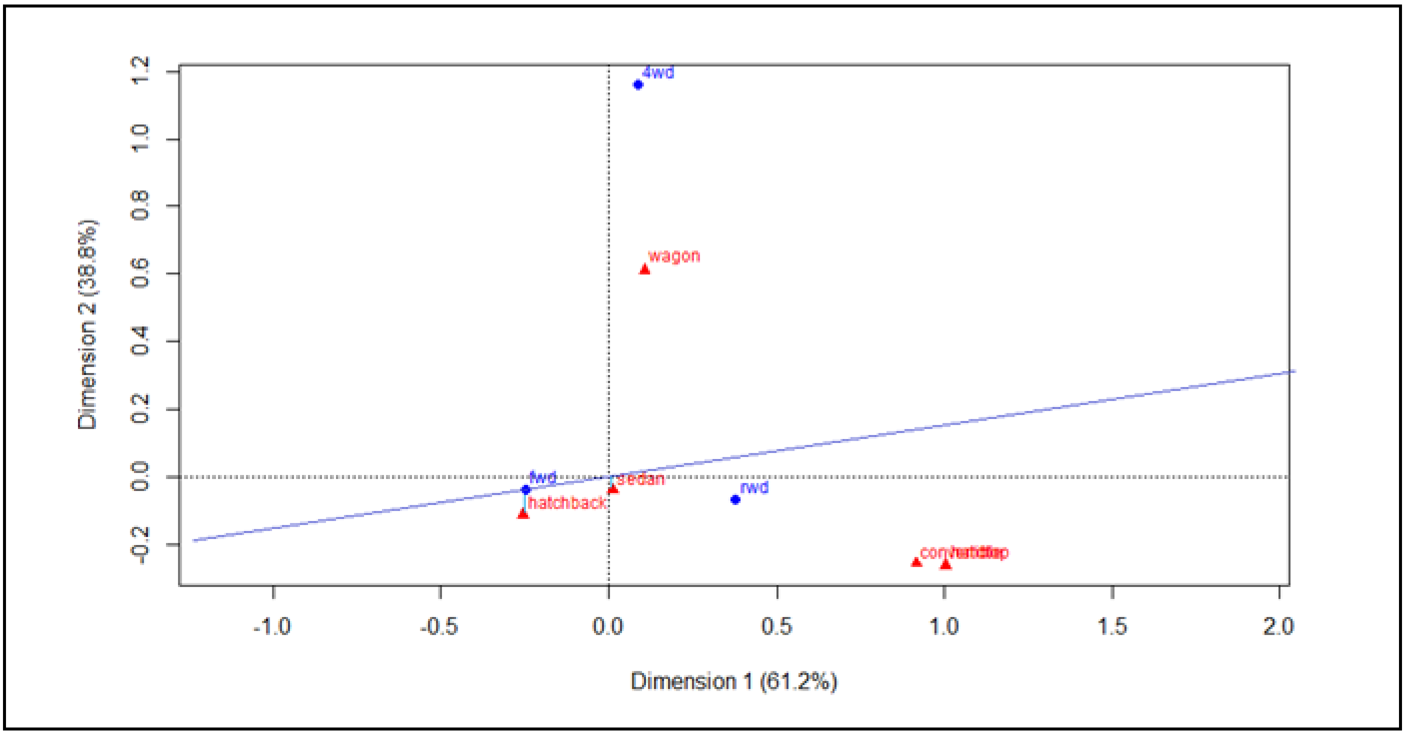


Fig: Scree Plot to determine the number of factors to be considered

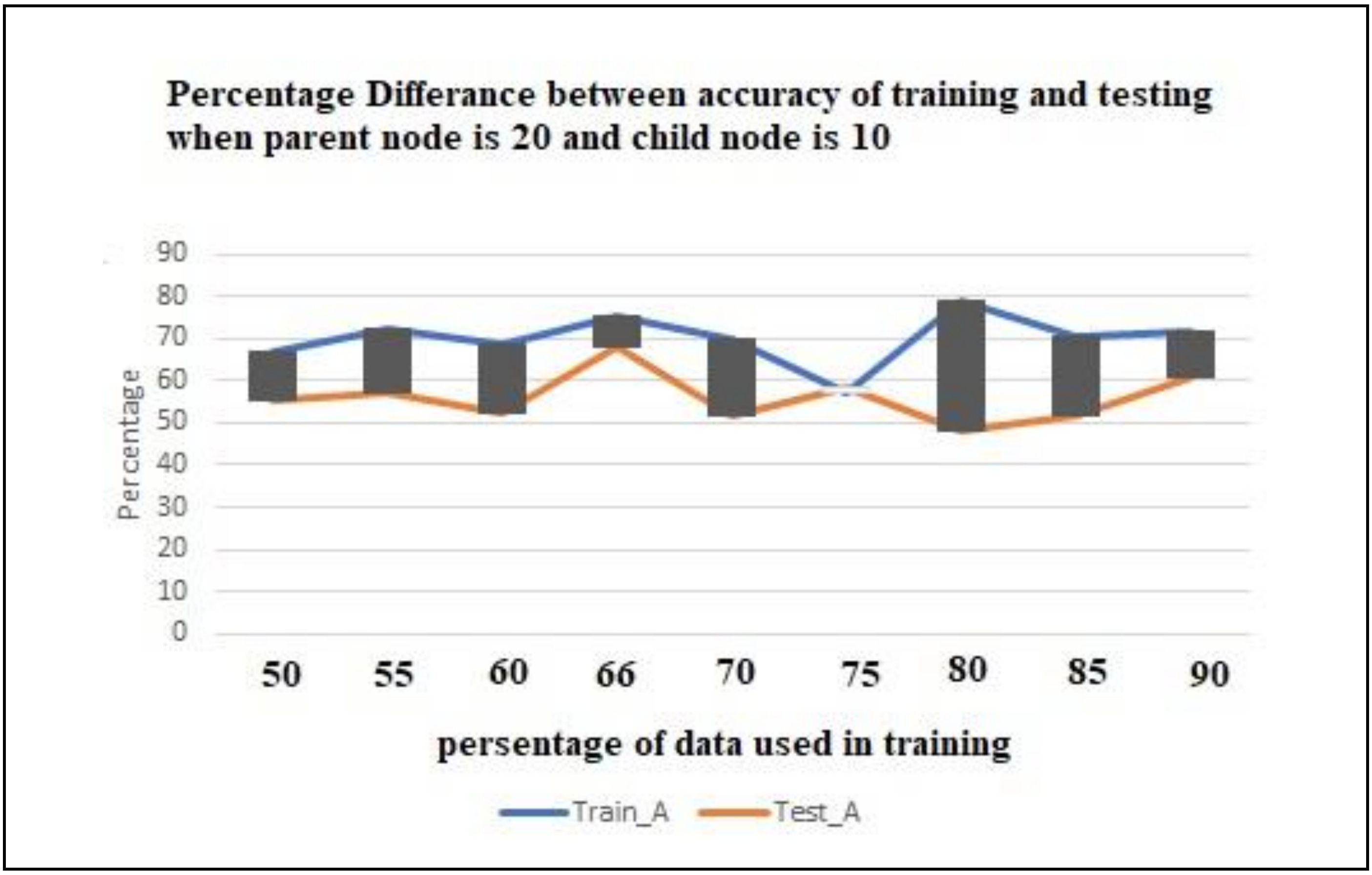
APPENDIX E: Correspondence Analysis





APPENDIX F: Decision Tree

- ***Various percentage ratio tried to see which one fits best:***



- ***Number of cases for parent and child performed:***

Parent Node	Child Node	Training Accuracy	Testing Accuracy	Terminal Node	Depth
50	25	63.2%	47.7%	4	3
40	20	61.5%	63.6%	4	3
30	15	66.7%	50%	4	2
20	10	75.6%	68.2%	6	3
15	8	73.2%	60.6%	7	3

- ***Classification table for final decision tree***

Classification

Sample	Observed	Predicted						
		-2	-1	0	1	2	3	Percent Correct
Training	-2	0	1	0	0	0	0	0.0%
	-1	0	8	4	1	0	0	61.5%
	0	0	0	42	1	1	1	93.3%
	1	0	0	7	30	0	2	76.9%
	2	0	1	3	9	9	0	40.9%
	3	0	0	0	2	0	13	86.7%
	Overall Percentage	0.0%	7.4%	41.5%	31.9%	7.4%	11.9%	75.6%
Test	-2	0	2	0	0	0	0	0.0%
	-1	0	6	3	0	0	0	66.7%
	0	0	0	16	2	0	2	80.0%
	1	0	0	1	10	0	2	76.9%
	2	0	2	2	2	3	1	30.0%
	3	0	0	0	2	0	10	83.3%
	Overall Percentage	0.0%	15.2%	33.3%	24.2%	4.5%	22.7%	68.2%

Growing Method: CRT
Dependent Variable: symbolling

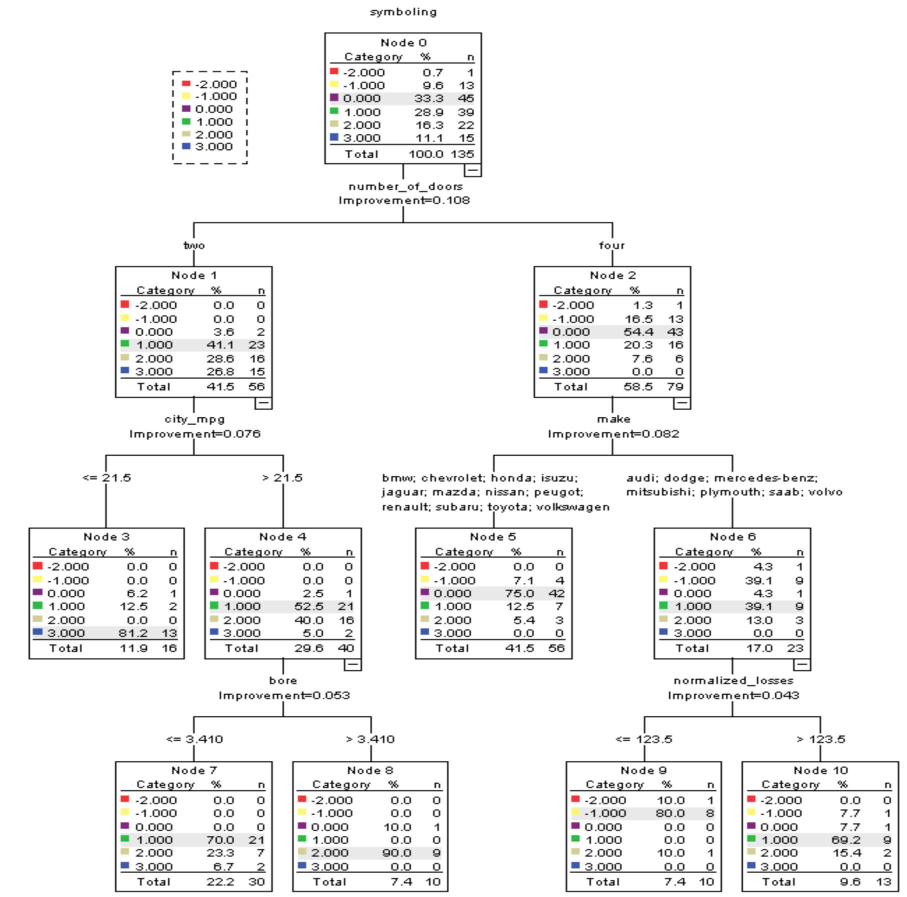
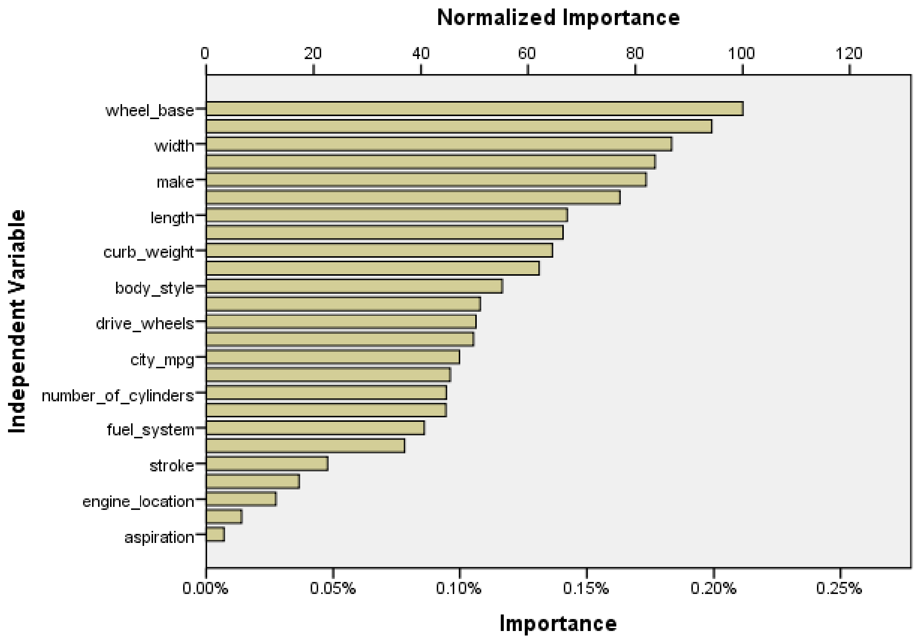


Fig: Final decision tree



Growing Method:CRT

Dependent Variable:symboling

APPENDIX F: Linear Regression

```
KFold(n_splits=2, random_state=1, shuffle=False)

KFold(n_splits=2, random_state=1, shuffle=False)
```

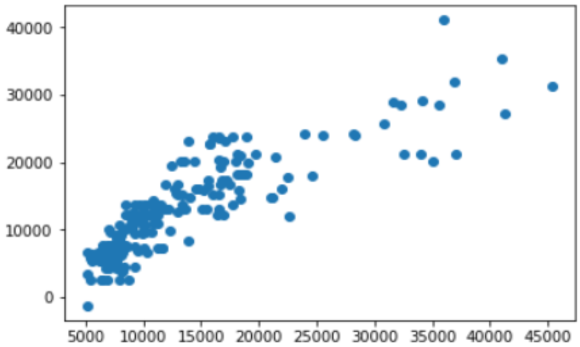
```
from sklearn.cross_validation import cross_val_score, cross_val_predict
from sklearn import metrics

scores = cross_val_score(model5, X5, Y5, cv=3)
print ("Cross-validated scores:", scores)
```

```
Cross-validated scores: [ 0.84661308  0.70591048  0.30844597]
```

```
predictions = cross_val_predict(model5, X5, Y5, cv=3)
plt.scatter(Y5, predictions)
accuracy = metrics.r2_score(Y, predictions)
print ("Cross-Predicted Accuracy:", accuracy)
```

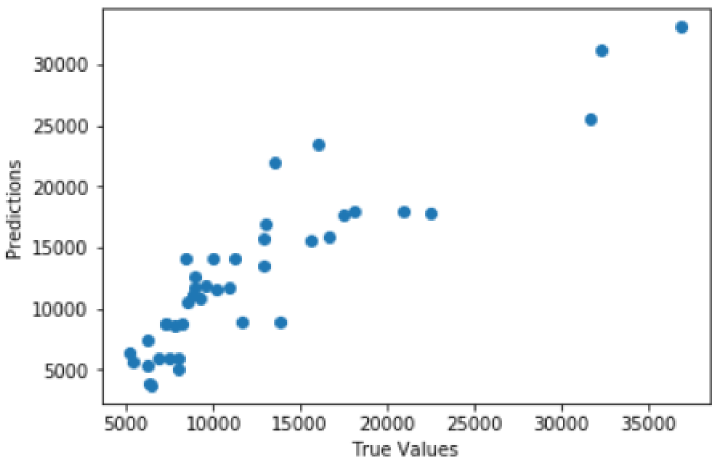
```
Cross-Predicted Accuracy: 0.738970407777
```



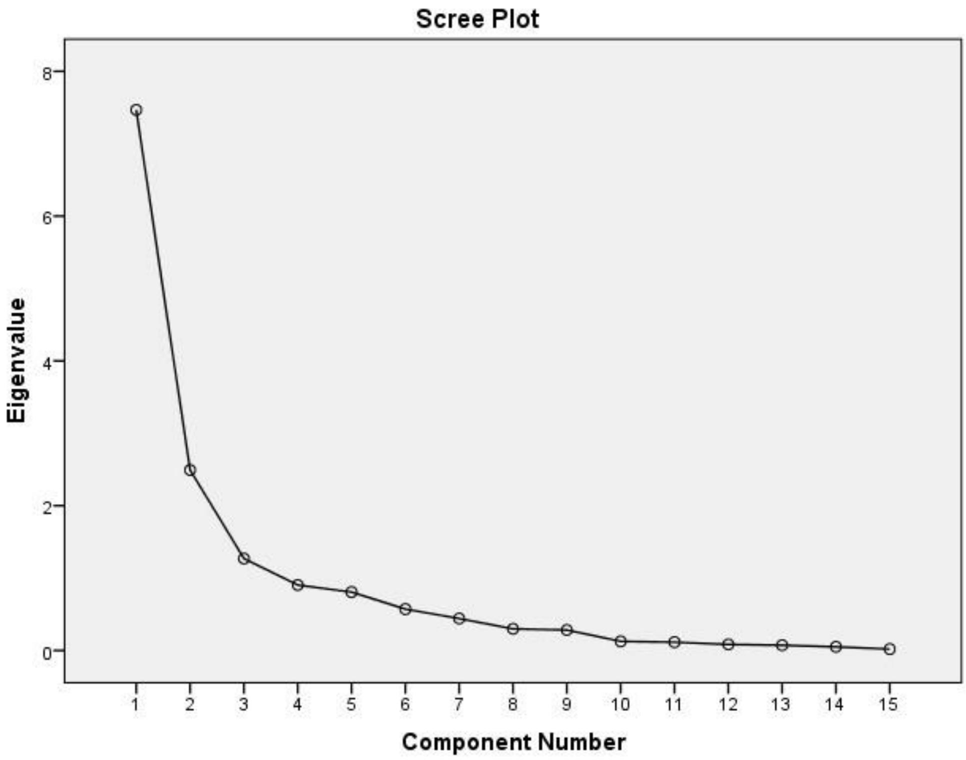

```
-49075.5568335
[  1.32933339e+02  -1.88655060e+02   2.80605139e+02  -1.49402543e+03
  -4.15183230e+03  -3.26194673e+03   1.28674788e+03   1.06349140e+04
   8.39433392e+02   4.08168496e+00   4.00304981e+02]
3150.22586666
```

```
plt.scatter(Y5_test,Y5_pred)
plt.xlabel("True Values")
plt.ylabel("Predictions")
print ("Accuracy Score:" , model5.score(X5_test, Y5_test))
```

```
Accuracy Score: 0.812741333264
```



APPENDIX G: Neural Networks



Component Matrixa

	Component		
	1	2	3
bore	.707	.047	-.239
stroke	.139	.065	.749
compression_ratio	.036	.713	.449
horsepower	.820	-.428	.036
peak_rpm	-.210	-.692	-.074
city_mpg	-.833	.420	.156
highway_mpg	-.864	.340	.156
price	.875	-.093	.100
wheel_base	.784	.447	-.047
length	.898	.255	-.068
width	.888	.175	.110
height	.285	.675	-.413
curb_weight	.965	.098	.068
engine_size	.872	-.065	.172
normalized_losses	.204	-.500	.401

Extraction Method: Principal Component Analysis.
a. 3 components extracted.

Network Information

Input Layer	Factors	1	make
		2	fuel_type
		3	aspiration
		4	number_of_doors
		5	body_style
		6	drive_wheels
		7	engine_location
		8	engine_type
		9	number_of_cylinders
		10	fuel_system
	Covariates	1	normalized_losses
		2	wheel_base
		3	length
		4	width
		5	height
		6	curb_weight
		7	engine_size
		8	bore
		9	stroke

		10	compression_ratio
		11	horsepower
		12	peak_rpm
		13	city_mpg
		14	highway_mpg
		15	price
	Number of Units[a]		73
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		4
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	symboling
	Number of Units		6
	Activation Function		Softmax
	Error Function		Cross-entropy

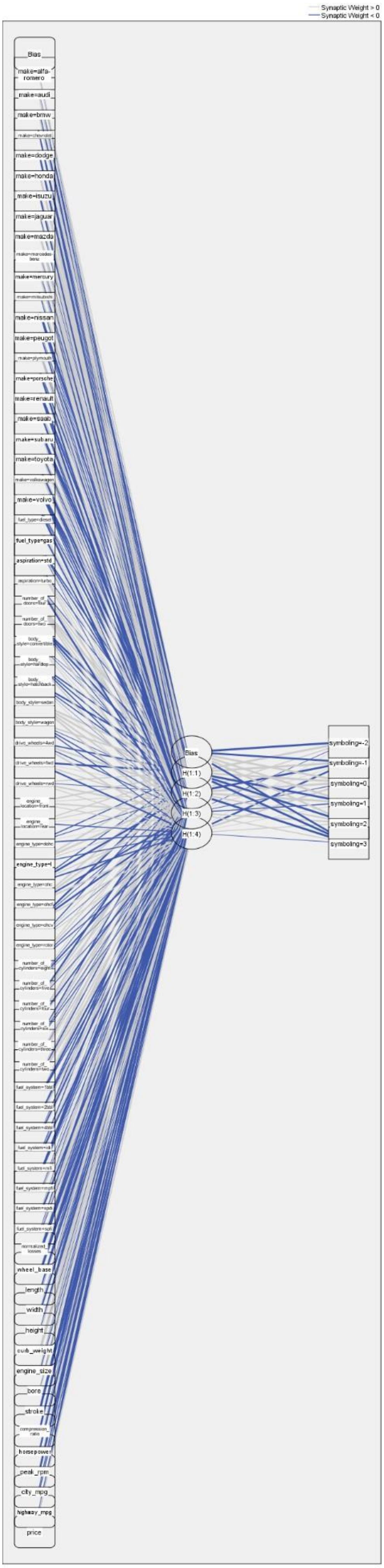
a. Excluding the bias unit

Classification

Sample		Observed		Predicted					
		-2	-1	0	1	2	3	Percent Correct	
Training		-2	0	0	1	0	0	0	0.0%
	-1	0	14	2	0	0	0	87.5%	
	0	0	7	38	2	0	0	80.9%	
	1	0	1	4	27	0	5	73.0%	
	2	0	0	3	14	0	9	0.0%	
	3	0	0	0	1	0	18	94.7%	

	Overall Percent			0.0%	15.1%	32.9%	30.1%	0.0%	21.9%	66.4%
Testing		-2	0	1	1	0	0	0	0.0%	
	-1	0	4	1	1	0	0	66.7%		
	0	0	0	12	4	0	1	70.6%		
	1	0	1	1	12	1	0	80.0%		
	2	0	0	2	4	0	0	0.0%		
	3	0	0	0	0	2	6	75.0%		
	Overall Percent			0.0%	11.1%	31.5%	38.9%	5.6%	13.0%	63.0%

Dependent Variable: symboling



Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Softmax