Summer 2020 DANA 4810 Group 7 - Project
1) Topic: Study of Used Car Market in India

Full Name of Each Group Member	Student ID of Each Group Member
Chin Wei (David) Shih	#100334127
Zhixuan Zhang (Eric)	#100338057
Chun Ching Look (Cyrus)	#100347726
Simranjit Singh (Simran)	#100348495
Kailash Sukumaran (Kailash)	#100350193
Guneesh Bhatia	#100346420

2) Introduction (Objectives and Motivation)

Buying a used car is an event that a lot of us will have to encounter in our lives. With an abundance of cars in the market, it can be a daunting experience for many people. It is always a gamble when you purchase a used car as many fear the underlying issues that may arise. While we will leave the inspection of vehicles to the experts, there is quite a number of information we can research upon for understanding how the price of a used car is set. We were fortunate enough to obtain a dataset on kaggle that focuses on the main factors that attributes to a car's second hand value. From the dataset, we will explore which factors contribute to the outcome that determines a used car's price. We will first explore the dataset and provide the background of where the information in the dataset is obtained. Then we will use graphs and provide a brief summary on the context of each variable/predictor that we will be using. Finally we will conduct inferential analysis to see if our model is useful in predicting a used car's selling price.

3) Descriptive Analysis

Brief description of the dataset

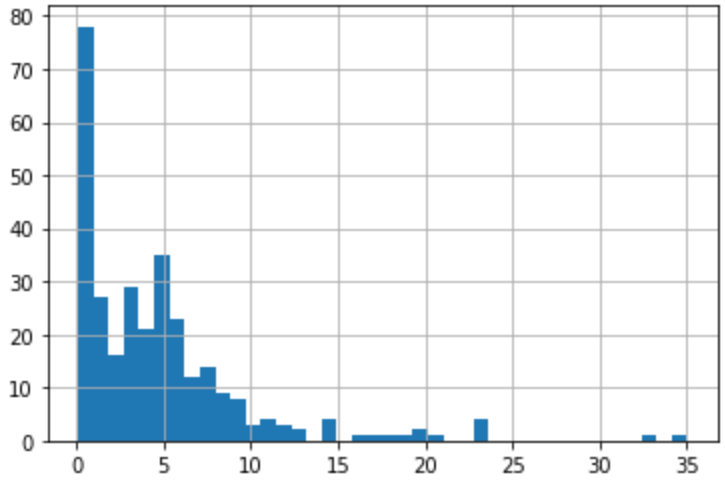
The dataset was found on Kaggle.com. The owners of this dataset collected information of all used cars listed on CarDekho.com - a well known used car search tool in India. The creators of this dataset used data scraping tools to scrape data from CarDekho.com. There were 2 versions of the dataset: a complete list of all 4,340 cars (as of June 2019) and a smaller list of just 300 cars (As of June 2018). For simplicity's sake, we will be using the smaller list for our dataset. The information collected from CarDekho are as follows: (1) Car Name, (2) Year, (3) Selling Price, (4) Present Price (Vehicle's original sticker price), (5), Kilometers Driven, (6) Fuel Type, and (7) Number of Owners.

Variable name	Variable Description	Type	Scale of measurement	Unit of Measurement/List of categories
Year	Year of the car when it was bought	Numerical	Interval	2003-2018
Selling Price	Our output (Y) variable	Numerical	Interval	Lakh, 1 Lakh = 100,000 Indian Rupees

Present Price	Price when car was bought	Numerical	Interval	Lakh, 1 Lakh = 100,000 Indian Rupees
Kms Driven		Numerical	Interval	KMs
Fuel Type	Petrol, Diesel or CNG (Compressed Natural Gas)	Categorical	Nominal	Petrol/Diesel
Seller Type	Dealer vs Owner	Categorical	Nominal	Dealer/Individual
Transmission	Manual vs Automatic	Categorical	Nominal	Manual/Automatic
Owner	0 = second hand car, 1 =third hand car and so on	Categorical	Nominal	0/1/3

(A) Output Variable: Selling Price

Selling Price Histogram: Price in Lakh on Y axis, freq on X axis.



Data is right skewed. More cheaper valued cars in the dataset.

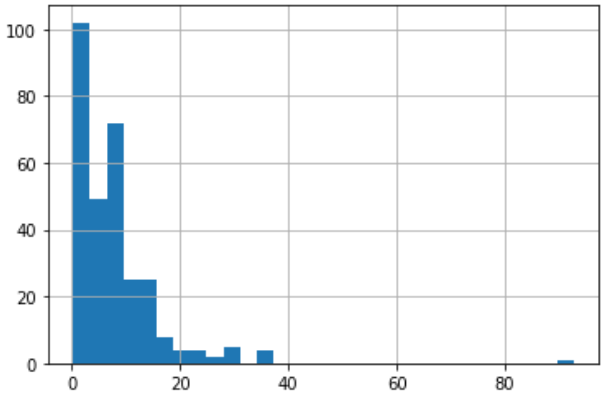
Summary of Selling Price:

count	301.000000
mean	4.661296
std	5.082812
min	0.100000
25%	0.900000
50%	3.600000
75%	6.000000
max	35.000000

(B) Graphs for Numerical Variables:

(1) Present Price

Histogram of Present Price: (Skewed data)
Price in Lakh on Y axis, freq(count) on X axis.



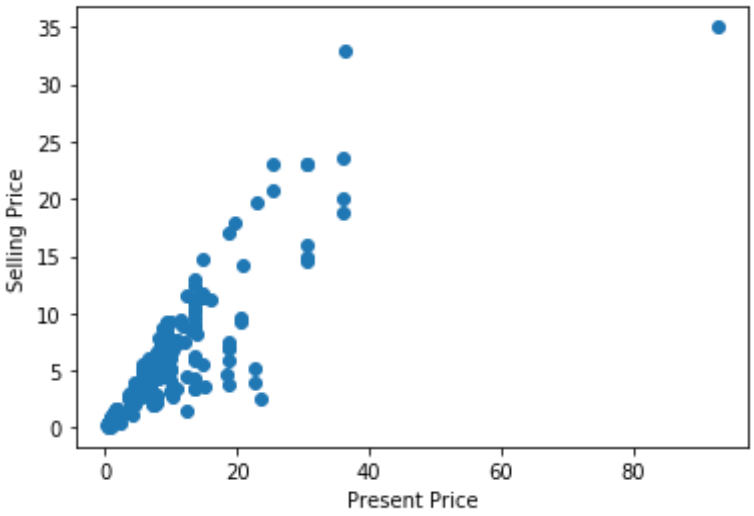
Data is right skewed. More cheaper valued (based on present price) cars in the dataset.

Summary Table of Present Price:

```
count    301.000000
mean       7.628472
std        8.644115
min        0.320000
25%        1.200000
50%        6.400000
75%        9.900000
max       92.600000
Name: Present_Price, dtype: float64
```

For 301 records, the mean value of present price is 7.628472 Lakh with a standard deviation 8.644115 Lakh . The present price ranges from 0.32 Lakh to 92.6 Lakh .

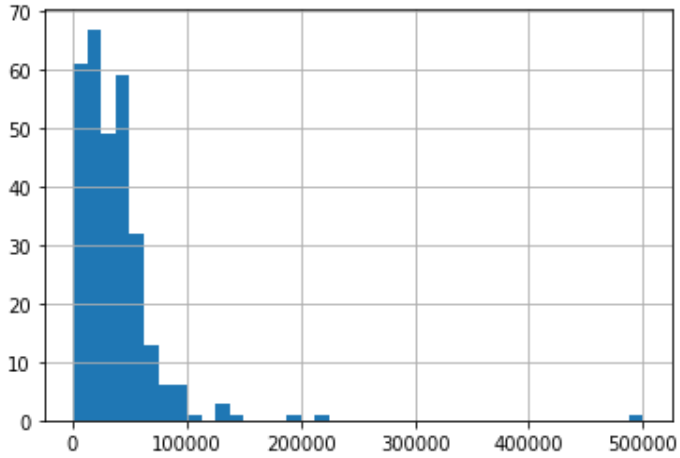
Present Price vs Selling Price:



Linear trend among present price vs selling price. This is logical because more expensive cars will still demand a higher selling price compared to a cheaper car of the same year.

(2) KMs Driven

Histogram Kms Driven: (skewed data)



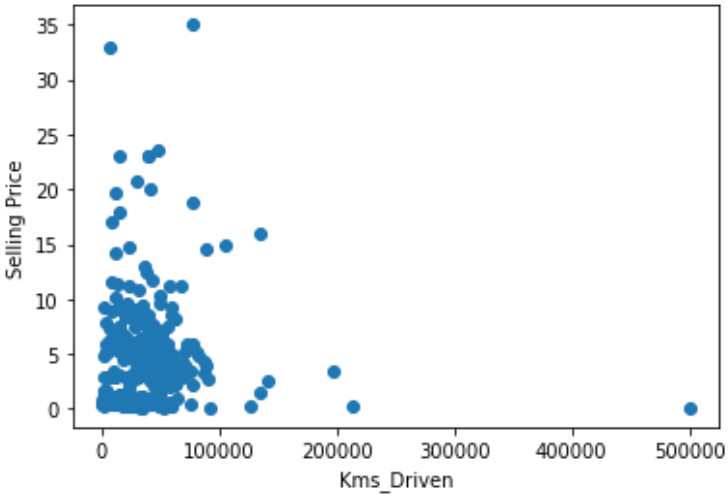
Lots of cars on the market with less than 100,000 kms driven. Right skewed data because of outliers present around 500,000 km.

Summary Table of Kms Driven:

```
count       301.000000
mean      36947.205980
std       38886.883882
min         500.000000
25%       15000.000000
50%       32000.000000
75%       48767.000000
max      500000.000000
Name: Kms_Driven, dtype: float64
```

For 301 records, the mean value of Kms Driven is 36,947.20598 km with a standard deviation 38,886.883882 km. The Kms Driven ranges from 500 km to 500,000 km.

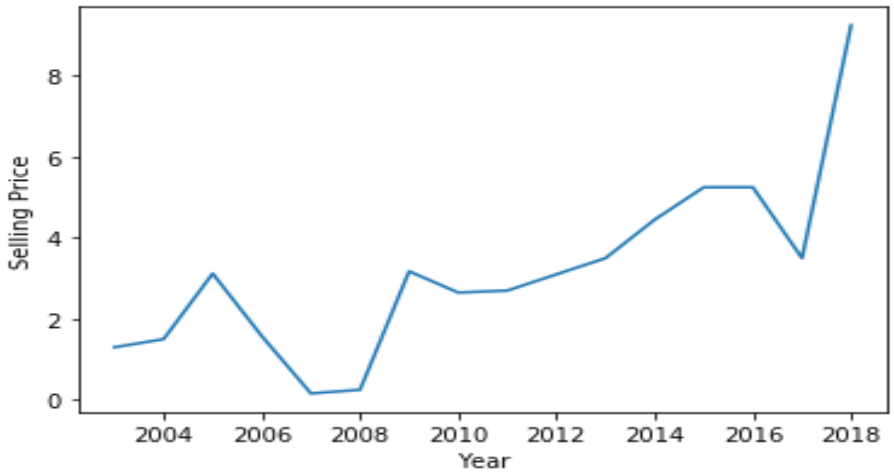
Kms Driven vs Selling price:



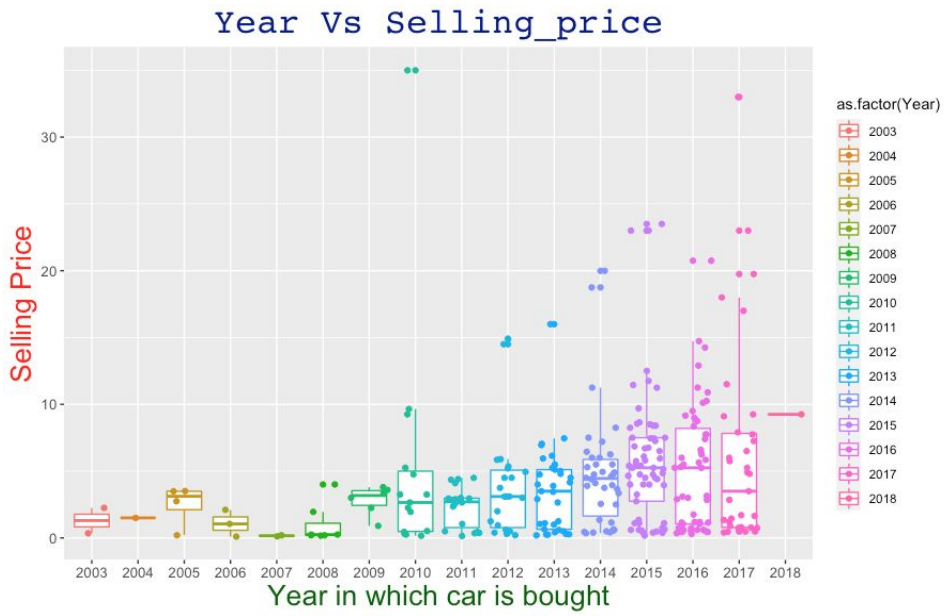
Outliers skews data. Low mileage cars sell at a higher price.

(3) Year

Year vs Selling price:



Newer cars tend to be more expensive than older cars.



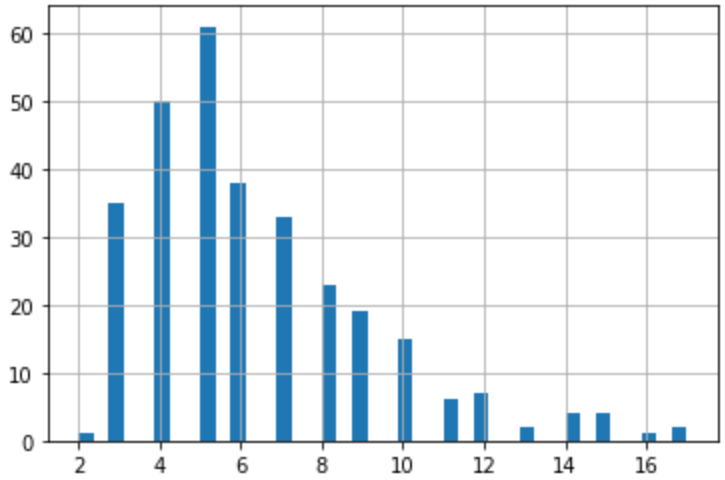
From the graph it's clear that the pattern of the market has changed from primary market to secondary market, since 2010 it's evident that more cars have started to sell in the secondary market. This proof adds up value to our project as the secondary market is equally important as the primary market in the current era.

Summary Table of Year:

```
count     301.000000
mean     2013.627907
std         2.891554
min      2003.000000
25%      2012.000000
50%      2014.000000
75%      2016.000000
max      2018.000000
Name: Year, dtype: float64
```

For 301 records, the mean value of Year is 2013.627907 with a standard deviation 2.891554. The Year ranges from 2003 to 2018.

Converted Year to Number of Years old with year 0 as 2019 (year data set was created)
Histogram of # of year old:(skewed data)



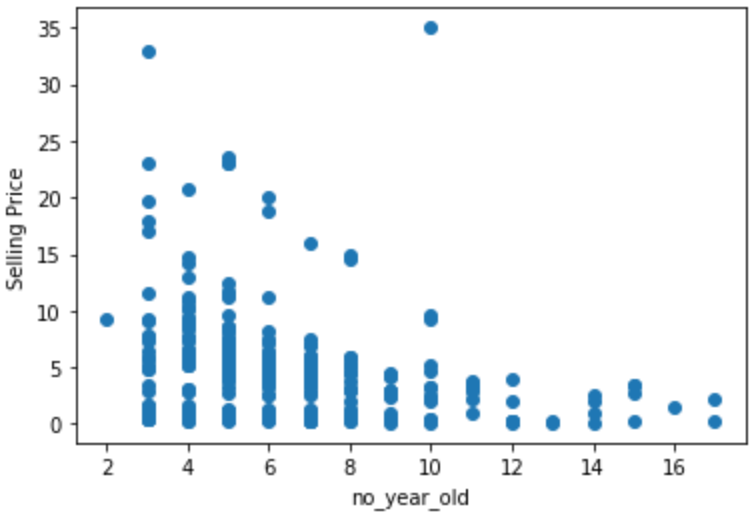
Right skewed data. Lots of cars in the dataset that is 5 years old. Seems to be the sweet spot in the used car market because most of the depreciation of the car is within the first few years.

Summary Table of # of year old:

```
count    301.000000
mean       6.372093
std        2.891554
min        2.000000
25%        4.000000
50%        6.000000
75%        8.000000
max       17.000000
Name: no_year_old, dtype: float64
```

For 301 records, the mean value of # of year old is 6.372093 years old with a standard deviation 2.891554 years old. The # of year old ranges from 2 years old to 17 years old.

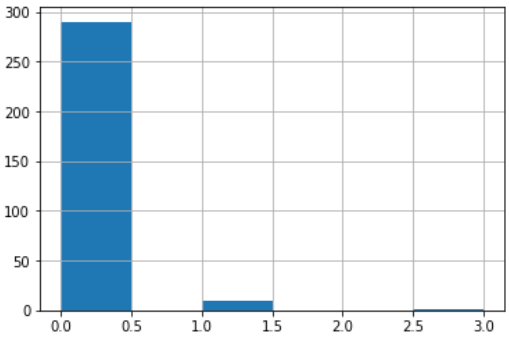
# of year old vs Selling price:



Newer cars have higher selling points (true for most values except outliers).

(4) Number of Owners

Histogram # of owners:



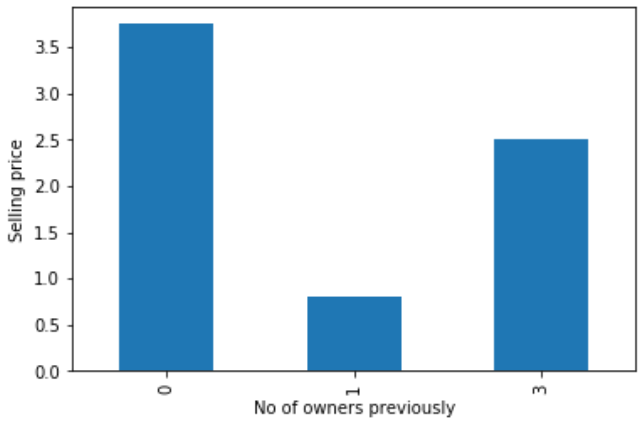
Lots of cars that had only 1 owner (selling as second hand). Few with more than 1 owner.

Summary Table of # of Owners:

```
count    301.000000
mean       0.043189
std        0.247915
min        0.000000
25%        0.000000
50%        0.000000
75%        0.000000
max        3.000000
Name: Owner, dtype: float64
```

For 301 records, the mean value of # of Owners is 0.043189 with a standard deviation 0.247915. The # of Owners ranges from 0 to 3.

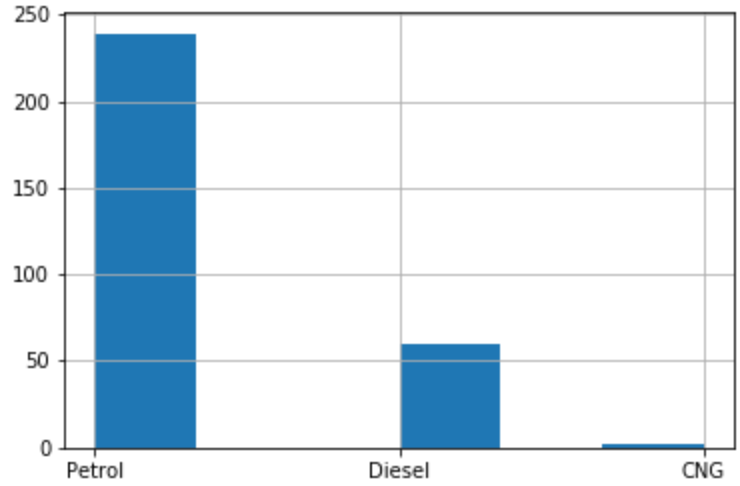
# of owners vs Selling price:



Cars with fewer owners have high selling prices. Cars with more than 2 owners also have high selling prices. One underlying factor could be that these are collector type cars that are hard to find giving them a higher selling price even having exchanging multiple hands.

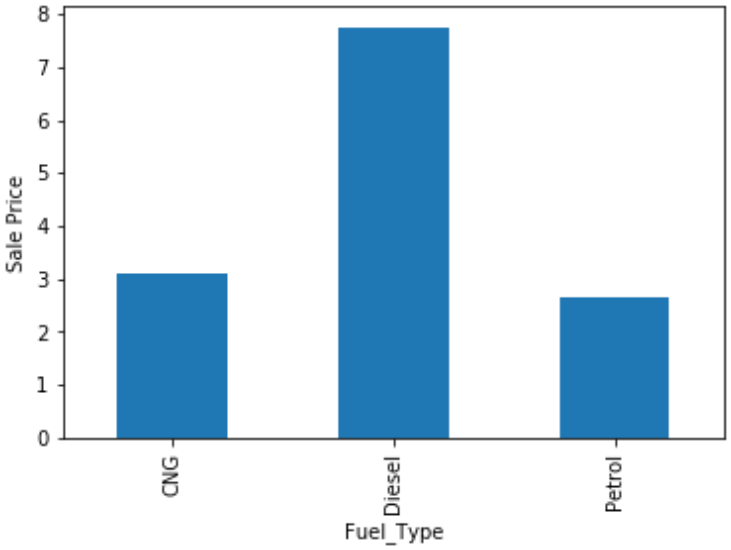
(C)Graphs for Categorical Variables:

(1) Fuel Types
Histogram of Fuel Types.



Lots of cars with petrol as the main fuel type, diesel second. CNG only has 2 rows of data.

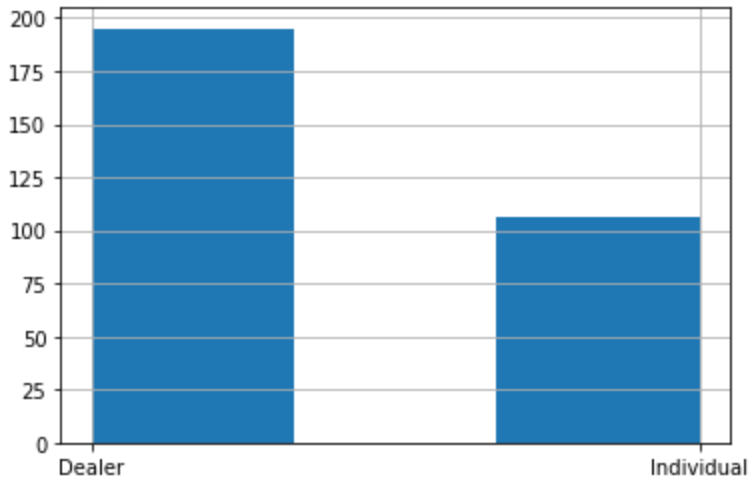
Bar Graph for Fuel type vs Sale price



Diesel has a higher selling point. An underlying variable could be because of the longevity of diesel engines compared to other types giving them a higher sale price.

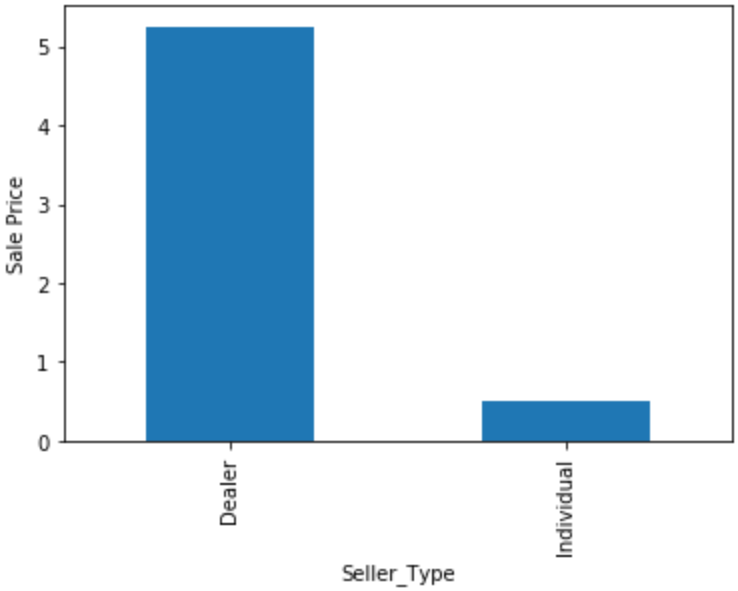
	Frequency:		Relative Frequency:
Petro	239	Petro	0.794020
Diesel	60	Diesel	0.199336
CNG (Compressed natural Gas)	2	CNG	0.006645

(2) Seller Type
Histogram of Seller Type:



Good mix of dealer selling vehicles and individual/private sellers. Dealers have more listed vehicles in the dataset/website used to scrape the data.

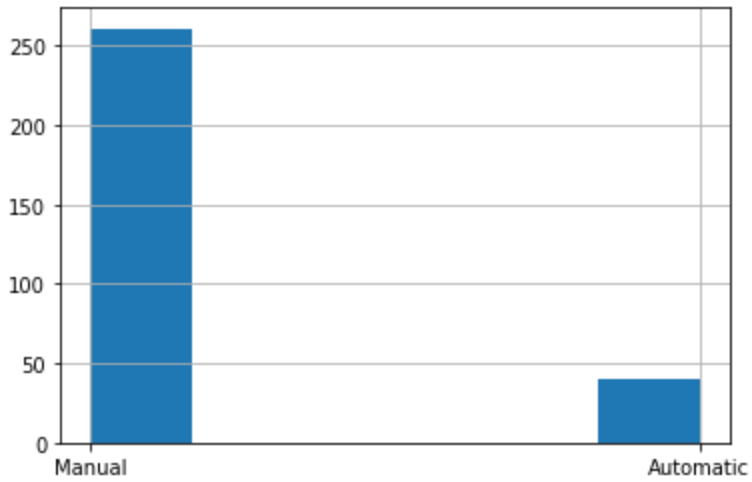
Bar Graph for Seller type vs Sale price:



Dealers sell cars that cost more compared to individuals.

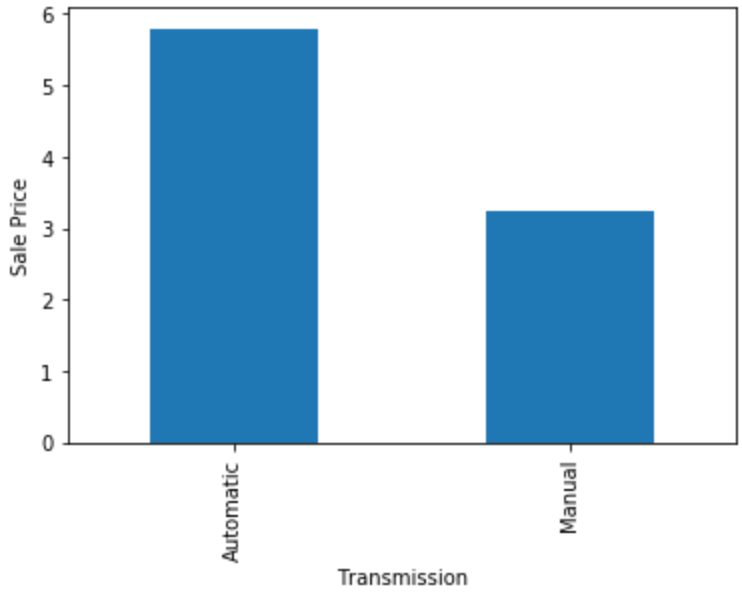
	Frequency:		Relative Frequency:
Dealer	195	Dealer	0.647841
Individual	106	Individual	0.352159

(3) Transmission Type
Histogram for Transmission Type:



More manual vehicles listed. This could be because of geographic demographics as manual cars are more desired in this region than automatic.

Bar Graph for Transmission type vs Sale price:



Although manual cars are more common, automatic cars have a higher sale price. This happens because of a reduced supply (more rare) of automatic vehicles in the region/dataset/website.

	Frequency:		Relative Frequency:
Manual	261	Manual	0.86711
Automatic	40	Automatic	0.13289

4) Inferential Analysis

To select predictor variables for model building, we need to remove unnecessary predictors in the beginning. Unnecessary predictors will add noise when doing estimation of other quantities in this statistical study. We will save much time in doing analysis after removing redundant predictors.

First, we will remove the “Car_Name” column from the dataset because the “Car Name” variable has little impact in analyzing the selling price of used cars. Next, we will remove the 86th row where the “Owner” variable = 3 because there is only one corresponding record. It has very little impact on building a multi linear regression model. Then, we will remove the rows where the “Fuel Type” variable is CNG because there are only two corresponding records.

Second, we will convert categorical predictors into dummy variables. For the “Owner” predictor, it has 2 levels (i.e. “0” and “1”); For the “Fuel Type” predictor, it has 2 levels (i.e. "Diesel", "Petrol")

Third, we will propose a model based on our dataset. We will fit the model with all the predictors in the dataset. The response variable is “Selling Price”. The predictors include 3 numerical predictors (i.e. “Year”, “Present Price”, “Kms Driven”) and 4 categorical predictors (i.e. “Fuel Type”, “Seller Type”, “Transmission”, “Owner”).

Fourth, we will use "olsrr", a R software package, to select suitable predictors. We will use 4 variable screening methods which are backward elimination, forward selection, stepwise regression, and all-possible-regression selection.
For backward elimination, we chose a 0.20 significance level to stay (SLS). We started with all the predictors in the model. After removing the “Owner” predictor, having the highest p-value greater than SLS, no more predictors are removed. All of their p-values are less than SLS.

For forward selection, we chose a 0.20 significance level to enter (SLE). After 6 steps of forward selection, the final model is produced. All predictors satisfy the condition of 0.20 SLE except the “Owner” predictor.

For stepwise regression, we chose a 0.10 significance level to enter (SLE) and a 0.30 significance level to stay (SLS). After 6 steps of stepwise regression, the final model is produced. All predictors satisfy the condition of 0.10 SLE and 0.30 SLS except the “Owner” predictor.

For all-possible-regression selection, we fit the model with all the possible subsets of predictors. Among the 127 possible subsets of predictor, we chose the one having the highest adjusted coefficient of determination (i.e. Adjusted R-Square). The adjusted R-Square of the chosen one is 0.883236993. It includes all predictors except the “Owner” predictor, so the number of predictors is 6. Its Mallow's Cp is 6.687804. Since the Cp is not much bigger than 7 (number of predictors + 1 = 6 + 1), the final model is not a bad fit.

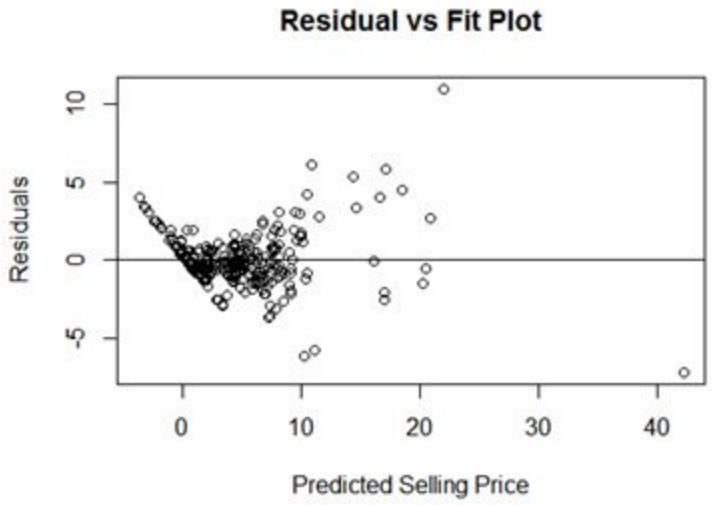
With reference to the above predictor variables selection, we decide to use the following multiple linear regression model. The model equation is as follows:

E(y) = beta0 + (beta1)(x1) + (beta2)(x2) + (beta3)(x3) +(beta4)(x4) + (beta5)(x5) + (beta6)(x6)
y = Selling Price; x1 = Year; x2 = Present Price; x3 = Kms Driven; x4 = Fuel Type;
x5 = Seller Type; x6 = Transmission

The fitted equation is as follows:

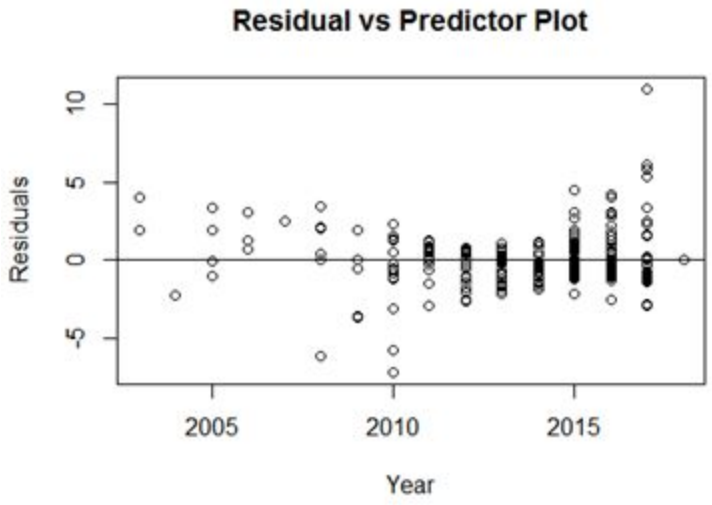
y_hat = -786.5 + 0.3929(x1) + 0.4421(x2) - 0.000006210(x3) - 1.813(x4) - 1.068(x5) - 1.524(x6)

To investigate the observed residuals to see if they support the assumptions, we did a residual analysis for response variable and numerical predictors.

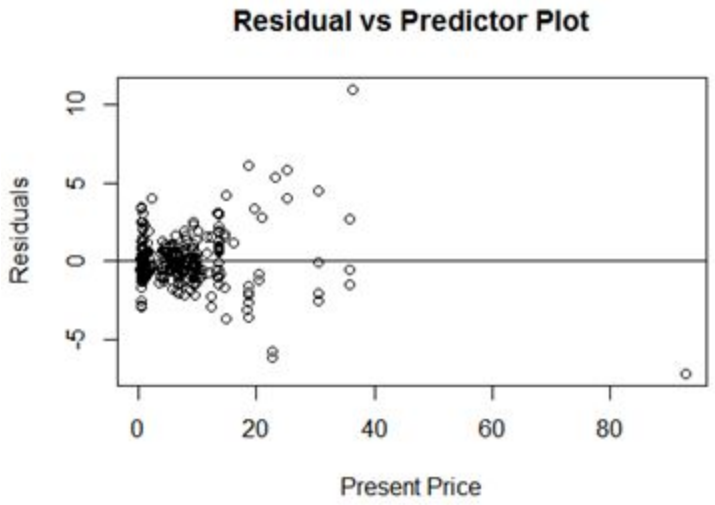


We prepared a Residuals vs. Fit Plot. Our first step is to obtain residuals and predicted value of response variable (selling price). Then, we created a Residuals vs. Fit Plot. As can be seen from the graph, most of the residuals lie randomly around the residual = 0 line. Most (~95%)
residuals within 2s of 0. Linearity assumption is met. The residuals show no obvious pattern. Equal variances assumption is met. There are 2 residuals stand out from the majority of residuals. One is located near the top of y-axis while one is located the right end of x-axis. We cannot prove that there is no outlier.

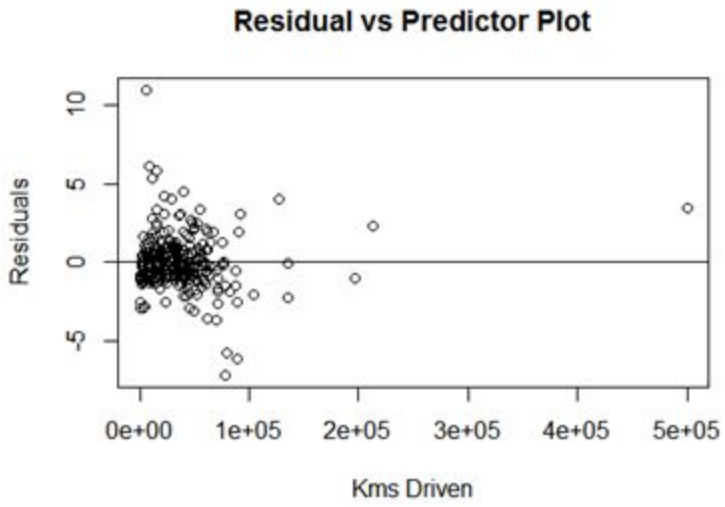
We prepared 3 Residuals vs. Predictor Plots for 3 numerical predictors. They are “Year”, “Present Price”, “Kms Driven”.



For “Year” predictor, most of the residuals lie randomly around the residual = 0 line. Most (~95%) residuals within 2s of 0. Linearity assumption is met. The residuals show no obvious pattern. Equal variances assumption is met. There is 1 residual stands out from the majority of residuals. It is located near the top of y-axis. We cannot prove that there is no outlier.

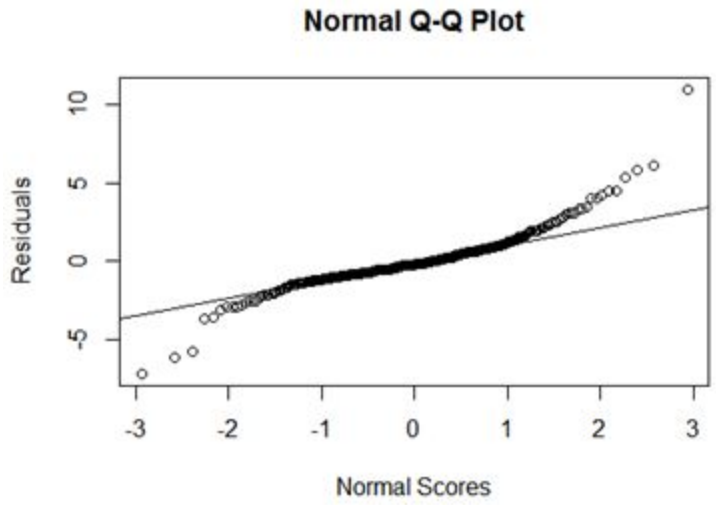


For “Present Price” predictor, most of the residuals lie randomly around the residual = 0 line. Most (~95%) residuals within 2s of 0. Linearity assumption is met. The residuals show no obvious pattern. Equal variances assumption is met. There is 1 residual stands out from the majority of residuals. It is located near the top of y-axis. We cannot prove that there is no outlier.

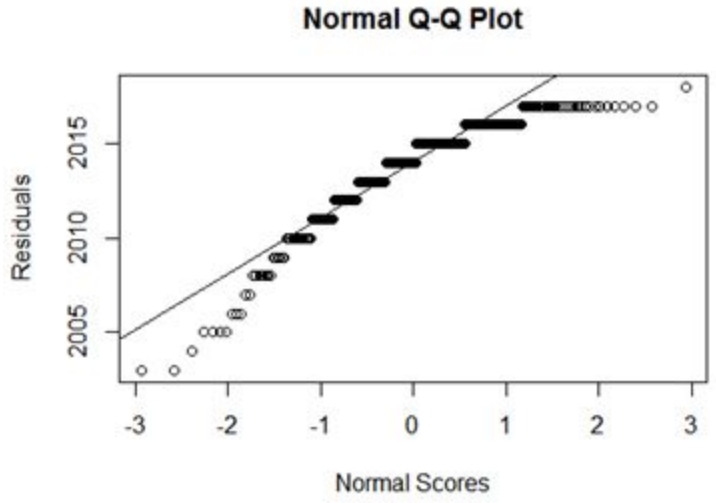


For “Kms Driven” predictor, most of the residuals lie randomly around the residual = 0 line. Most (~95%) residuals within 2s of 0. Linearity assumption is met. The residuals show no obvious pattern. Equal variances assumption is met. There are 2 residuals stand out from the majority of residuals. One is located near the top of y-axis while one is located the right end of x-axis. We cannot prove that there is no outlier.

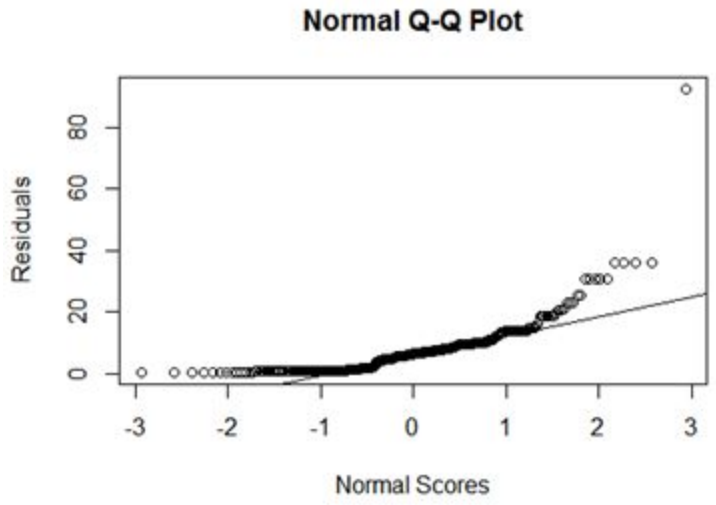
To check the normality assumption for response variable and numerical predictors, we created 4 Normal probability plots.



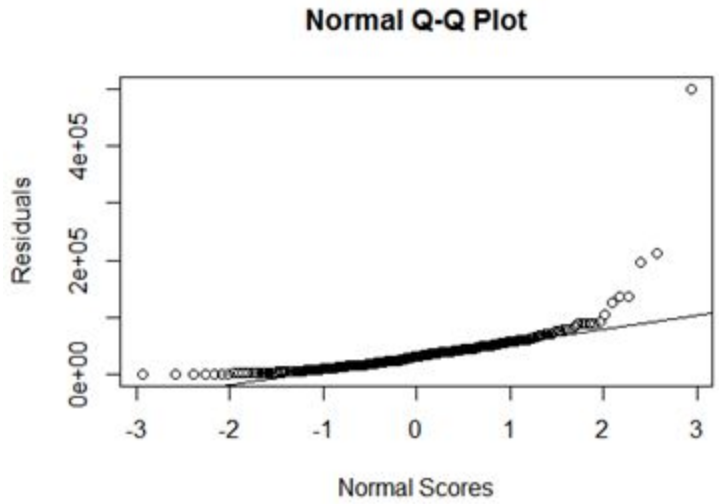
For response variable (selling price), its Normal probability plot shows a linear trend. The normality assumption is met.



For “Year” predictor, its Normal probability plot shows a linear trend. The normality assumption is met.



For “Present Price” predictor, its Normal probability plot shows a linear trend. The normality assumption is met.



For “Kms Driven” predictor, its Normal probability plot shows a linear trend. The normality assumption is met.

Model Testing

We conducted a Partial-F test for comparing our reduced model with the complex model.

Equation of Complex Model:

E(y) = beta0 + (beta1)(x1) + (beta2)(x2) + (beta3)(x3) +(beta4)(x4) + (beta5)(x5) + (beta6)(x6) + (beta7)(x7)

Equation of Reduced Model:

E(y) = beta0 + (beta1)(x1) + (beta2)(x2) + (beta3)(x3) +(beta4)(x4) + (beta5)(x5) + (beta6)(x6)

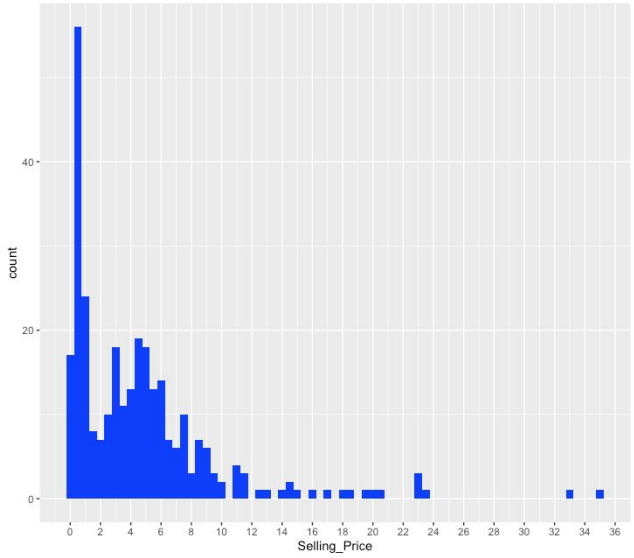
y = Selling Price; x1 = Year; x2 = Present Price; x3 = Kms Driven; x4 = Fuel Type; x5 = Seller Type; x6 = Transmission; x7 = Owner

We fitted both the Reduced and Complex Model in anova() function and observed that the p-value is 0.41, which is greater than the significant level (0.05). This means that a reduced model is sufficient in predicting the selling price of a car.

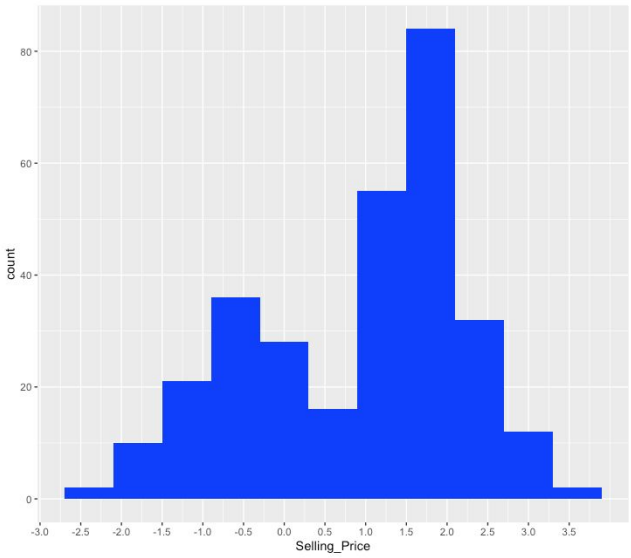
Variable Transformation

We have done natural log transformation for our response variable “Selling Price” because there was strong skewness in this response variable.

Before Transformation

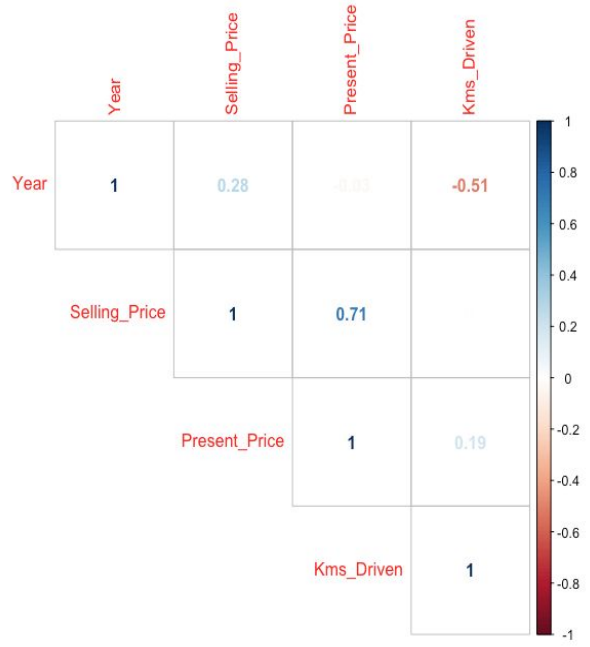


After transformation



Multicollinearity

We checked multicollinearity within all the numerical variables. It is clear from the correlation plot that there is NO strong correlation among the predictors.



Further Estimation and Prediction

To make estimation and predictions, we split our dataset into “training set” and “test set”. We used 80% of our data (238 rows) for training the regression model and 20% (60 rows) for testing the regression model.
RMSE (Root Mean Square Error) of the training set = 0.40
RMSE of the test set = 0.49

We can see that RMSE for both the train and test data is quite low, which indicates that our regression model is highly efficient in predicting the selling price of a car.

Interaction Term

Based on our knowledge and background information on the used car market, it is unnecessary to add interaction terms. When considering interaction among predictors, there is NO strong interaction among the predictors. We need not to include any interaction terms between both categorical and numerical predictors.

5) Discussion & Conclusion & Limitations

Limitations

Even though the sample size (n) is about 300, there is still sampling error. It exists because we are getting samples in our studies. We can only claim the sampling error should be small when the sample size is so large. Moreover, all the samples included in the data set is for one company and one company alone (Cardekho). Other brands of used cars not sold by Cardekho will not be in this dataset. We can say that the dataset has selection bias. It is barely impossible to have 100% match between the target population (i.e. All used cars from the Indian market) and the sampling frame (i.e. All used car data from Cardekho). Under this condition, we can only choose a better sampling frame and clearly the current one using in this project is the best.

All variable selection methods (backward elimination, forward selection, stepwise regression, and all possible regression selection) do have some drawbacks. It is always dangerous to use the selected model as the final model for doing prediction and making inferences. Thus, we used all 4 variable selection methods together to get a more comprehensive guide.

Conclusion

As demonstrated from the graphs in the report above, we conclude the following details:

1. Many cars in the market are driven less than 10,000kms. Due to the presence of outliers ,even the low mileage cars are sold at higher prices.
2. New cars are more expensive than the older ones. Further, the market trend has slightly changed and more cars are being sold since 2010.
3. Cars with few owners have high selling price.However surprisingly ,the selling price of the cars with multiple owners is more than cars having single owner because these cars might be a collector type cars which are hard to get.
4. Automatic cars have higher selling price than manual cars because of the demand and supply of the type of cars in that region,even when the manual cars are more common and preferred in that geographical location.
5. Stepwise regression analysis was done to select the most significant variables in the dataset.Among the 127 subsets of the predictors, six predictors with highest adjusted coefficient of determination were chosen. Post feature selection, residual analysis for every predictor was found to be following the underlying assumptions.
6. Partial F test was conducted to test the utility of the model and the reduced model was found to be more significant in predicting the price of the vehicle than the complex model.
7. RMSE for both the training and the testing set were found to be low,indicating the efficiency of the model for predicting the price of a vehicle.