University of Calcutta

Lady Brabourne College

B.Sc. Semester 6 Examination 2022

(under cbcs)

Subject: Statistics

Paper: DSE B - Project Work

PROJECT TITLE: STATISTICAL ANALYSIS OF THE USED CARS AND PRICE PREDICTION.

Submitted by :

CU roll number: **193031-11-0178**

CU registration number: **031 - 1214- 0408-19**

ABSTRACT:

In this project, we will be predicting the prices of used cars.

The price of a car depends on a lot of factors like the goodwill of the brand of the car, features of the car, and the mileage it gives and many more. Car price prediction is one of the major research areas in machine learning and statistics. So i want to learn how to train a car price prediction model. In this article, I will take you through how to train a car price prediction model with machine learning and statistical methods using r programming.
. It is based on **finance** and the **marketing** domain. If one ignores the brand of the car, a car manufacturer primarily fixes the price of a car based on the features it can offer a customer. Later, the brand may raise the price depending on its goodwill, but the most important factors are what features a car gives you to add value to your life. The dataset I'm using here to train a car price prediction model was downloaded from Kaggle. It contains data about all the main features that contribute to the price of a car.

Introduction:

This data set contains information about used cars listed on www.cardekho.com.

There is an automobile company XYZ from Japan which aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts.

They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. The company wants to know:

Which variables are significant in predicting the price of a car.
How well those variables describe the price of a car.
Based on various market surveys, the consulting firm has gathered a large data set of different types of cars across the America market.

I as a statistician are required to apply some statistical techniques for the price of cars with the available independent variables. That should help the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels.

This data can be used for a lot of purposes such as price prediction to exemplify the use of linear regression in Machine Learning.

In this section,we will explore the data. First Let's see what colunms we have in the data. we We can observe that data have 301 rows and 9 columns. There are 4 numeric columns and 5 categorical columns. With the first look, we can see that there are no missing values in the data.'selling price and present Price' column/feature are going to be the target column or dependent feature for this project.

The columns in the given dataset is as follows:

1. Car_Name:This column should be filled with the name of the car.
2. Year:This column was filled with the year in which the car was bought
3. Selling_Price:This column should be filled with the price the owner wants to sell the car at.
4. Present_Price:This is the current ex-showroom price of the car.
5. Kms_Driven:This is the distance completed by the car in km
6. Transmission:Defines whether the car is manual or automatic.
7. Fuel_type:type of the car. Here fuel types are petrol,CNG,diesel.
8. Owner: Fuel Defines the number of owners the car has previously had.
9. Seller_Type:Defines whether the seller is a dealer or an individual.
10. In our main data set there is 301 observations and 9 columns.

METHODOLOGY:

Pie slices is used to show the percentage of a particular data from the whole pie. Also, it is useful to interpret visual associations of all the variables' pie chart.

Scatter plot is used to observe the association of each of variables with one another. Covariance matrix is a symmetric matrix that shows covariances of each pair of variables.

Correlation is a statistical tool used to access a possible linear association between variables. It expresses the extend, to which two variables are linearly related to each other. It is used to examine if any of the variables are correlated to each other.

Multiple linear regression is used to estimate the relationship between independent variables and dependent variable.

The equation for the above is: $y= \beta_0+\beta_1x_1+\beta_2x_2 +\beta_3x_3+ \mu$

Where, y is the response variable, x_1, x_2 ,x_3 are explanatory variables,β_0, β_1, β_2 and β_3 are the unknown parameters to be estimated on the basis of sample data. μ is the stochastic disturbance term. The multiple regression model is based on the following assumptions: i)There is a linear relationship between the dependent variables and the independent variables, ii)the independent variables are not too highly correlated with each other, iii)residuals should be normally distributed with a mean of 0 and variance σ.

Simple linear regression is used to model the linear relationship between dependent and independent variable. Consider a model: $y= \beta_0+\beta_1x_1+\mu$

Where y, is response variable, x is the explanatory variable, $\beta0$ and $\beta1$are the unknown parameters to be estimated on the basis of sample data. μ is the stochastic disturbance term. Now following assumptions are made about μ: $E[\mu]=0$, $V[\mu]= \sigma2$, $Cov [\mu_i\mu_j]=0$ for all $i\neq j$, x is non stochastic. The model with all these assumptions is known as classical linear regression model. In addition to all these assumptions if we assume μ_i follows i.i.d normal $(0, \sigma2)$ then the model is known as classical normal linear regression model. Let x_1y_1, x_2y_2……..x_ny_n be a sample of size n from the paired observation xy. The predicted linear regression equation of y on x is given by: $Y=\beta_0+\beta_1x_1$

Where Y is predicted value of y and $\beta0$ and β_1 obtained by minimising the error sum of squares s $=\Sigma^n_{i=1}(yi -Yi)^2$

Polynomial regression- Previously we discussed about transformations on the response and/or the predictor(s) for linearizing a nonlinear relationship. When this fails, we can turn to polynomial regression models that is :

$y = \beta_0+\beta_1x_1+\beta_2x_1^2+…..+\beta_kx_k^k+\varepsilon$ Where, y is response variable, x_1, x_2 and x_3 are explanatory variables, β_0, β_1,β_2 and β_3 are the unknown parameters to be estimated on the basis of sample data. ε is an unobserved random error with mean zero conditioned on a scalar variable x. In this model, for each unit increase in the value of x, the conditional expectation of y increases by $\beta1$ units. The interpretation of parameter β_0 is $\beta_0 = E(y)$ when $x = 0$ and it can be included in the model provided the range of data includes $x = 0$. If $x = 0$ is not included, then β_0

has no interpretation. The model can be written in terms of a matrix X, a response variable y a parameter vector β, and a vector ε.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ 1 & x_3 & x_3^2 & \dots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

A basic assumption in linear regression analysis is that rank of X-matrix is full column rank. In polynomial regression models, as the order increases, the X ' X matrix becomes ill-conditioned. As a result, the (X'X)-1 − may not be accurate, and parameters will be estimated with a considerable error. If values of x lie in a narrow range, then the degree of ill-conditioning increases and multicollinearity in the columns of X matrix enters. For example, if x varies between 2 and 3, then x2varies between 4 and 9. This introduces strong multicollinearity between x and x2. It is expected that all polynomial models should have hierarchy property because only hierarchical models are invariant under linear transformation.

Logistic Regressions is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model. A binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labelled "0" and "1". In the logistic model, the log odds (the logarithm of the odds) for the value labelled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value).

We consider a model with a predictor x1 and one binary (Bernoulli) response Y, which we denote p=P(Y=1). We assume a linear relationship between the predictor variables and the log-odds (also called logit) of the event that Y=1. This linear relationship can be written in the following mathematical form (where Z is the log-odds, b is the base of the logarithm and β_i are parameters of the model)

$Z = \log b(P/1-P) = \beta_0+\beta_1 x_1$

We can recover the odds by exponentiating the log-odds:

$(P/1-P) = b^{\beta 0+ \beta 1 x1}$

The assumptions are: (1) logistic regression does not require a linear relationship between the dependent and independent variables, (2) the error terms (residuals) do not need to be normally distributed, (3) homoscedasticity is not required, and (4) the dependent variable in logistic

To evaluate the performance of a logistic regression model, we must consider few metrics. 1. AIC (Akaike Information Criteria) – AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value. 2. Null

Deviance and Residual Deviance – Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.

Multiple column diagram is use to display more than one data series in multiple vertical columns. A multiple line graph is a line graph that is plotted with two or more lines.

Forecasting is the process of making predictions based on past and present data and most commonly by analysis of trends. Forecasting of future events can be performed on such data which is dependent on Time.

Autoplot automatically create a ggplot (ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.) for time series objects. Autoplot takes an object of type ts or mts and creates a ggplot object.

STACKED BAR DIAGRAM: A stacked bar diagram consists of multiple bar series stacked one upon another. The length of each series is determined by the value in each data point.

BAR DIAGRAM: A bar chart presents categorial data with rectangular bars with heights proportional to the values they represent. It shows comparisons among discrete categories.

PRINCIPAL COMPONENT ANALYSIS:

Principal component analysis (PCA) is a statistical procedure that

a. uses an orthogonal transformation to convert a set of observations of correlated variables into a set of linearly uncorrelated variables (called the principal components)

b. finds direction with maximum variability

c. PCs form a new coordinate system by rotating the original system constructed by X1, X2, …, XP

The PCA is concerned with explaining the variance-covariance (Σ) or correlation (R) structure of **X**= (X1, X2, …, XP)' through a few linear combinations of these variables. Z1=**v1**'**X**=v11X1+v12X2+…+v1pXp

Z2= **v2**'**X**=v21X1+v22X2+…+v2pXps

…

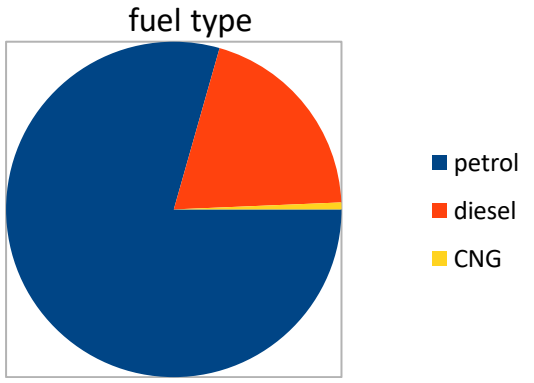
Zp= **v1**'**X**=vp1X1+vp2X2+…+vppXp

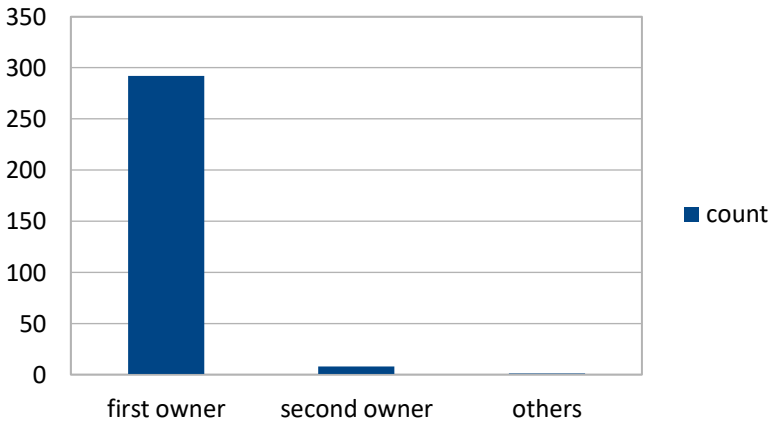
where, vij are the corresponding coefficients (called factor loadings: PCA loadings are the coefficients of the linear combination of the original variables from which the

principal components (PCs) are constructed) for the ith variable in the jth PC in Σ or R and Zis are the Principal Components.

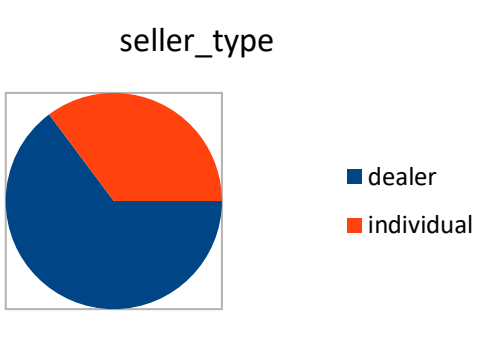
The **scree plot** plots the variances of the PCs along the new feature axes and look for the bend in the plot.

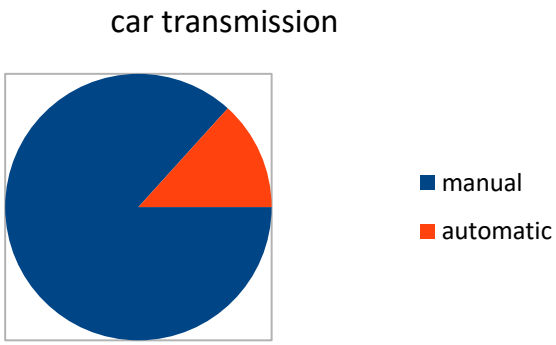
statistical analysis:





here 30.1% diesel car.68.4 %petrol car. 1.5% .CNG car.

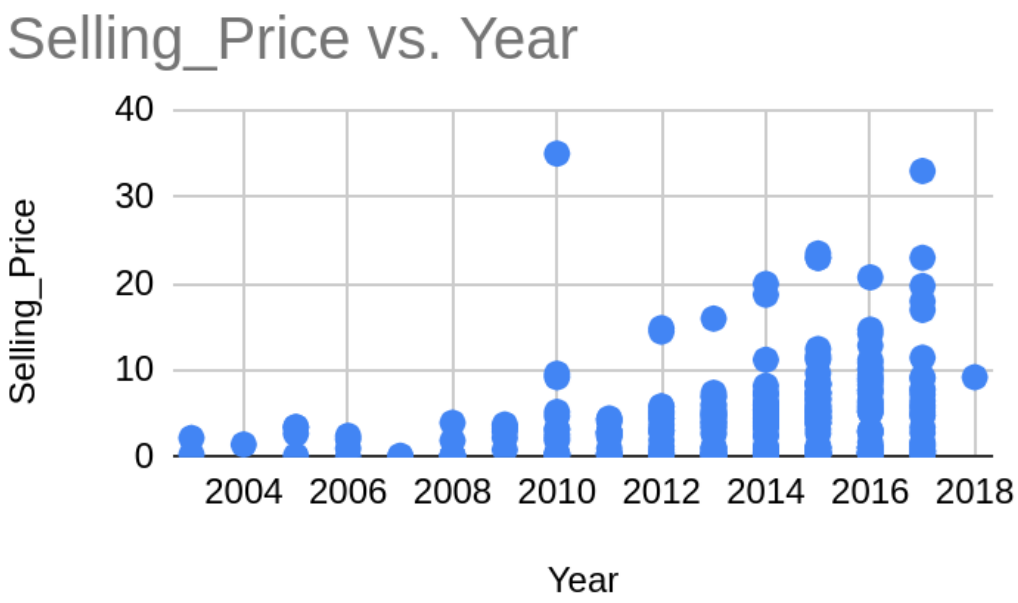




seller type of 62.3%cars are dealer and 37.7% are individual

13.3% car are automatic.86.7% car are

manual



here I plot year vs selling price. We see that as year increase selling price also increase.

summary(car.data)

Car_Name	Year	Selling_Price	Present_Price	Kms_Driven
city : 26	Min. :2003	Min. : 0.100	Min. : 0.320	Min. : 500
corolla altis: 16	1st Qu.:2012	1st Qu.: 0.900	1st Qu.: 1.200	1st Qu.: 15000
verna : 14	Median :2014	Median : 3.600	Median : 6.400	Median : 32000
fortuner : 11	Mean :2014	Mean : 4.661	Mean : 7.628	Mean : 36947
brio : 10	3rd Qu.:2016	3rd Qu.: 6.000	3rd Qu.: 9.900	3rd Qu.: 48767
ciaz : 9	Max. :2018	Max. :35.000	Max. :92.600	Max. :500000

(Other) :215

Fuel_Type	Seller_Type	Transmission	Owner
CNG : 2	Dealer :195	Automatic: 40	Min. :0.00000
Diesel: 60	Individual:106	Manual :261	1st Qu.:0.00000
Petrol:239			Median :0.00000
			Mean :0.04319
			3rd Qu.:0.00000
			Max. :3.00000

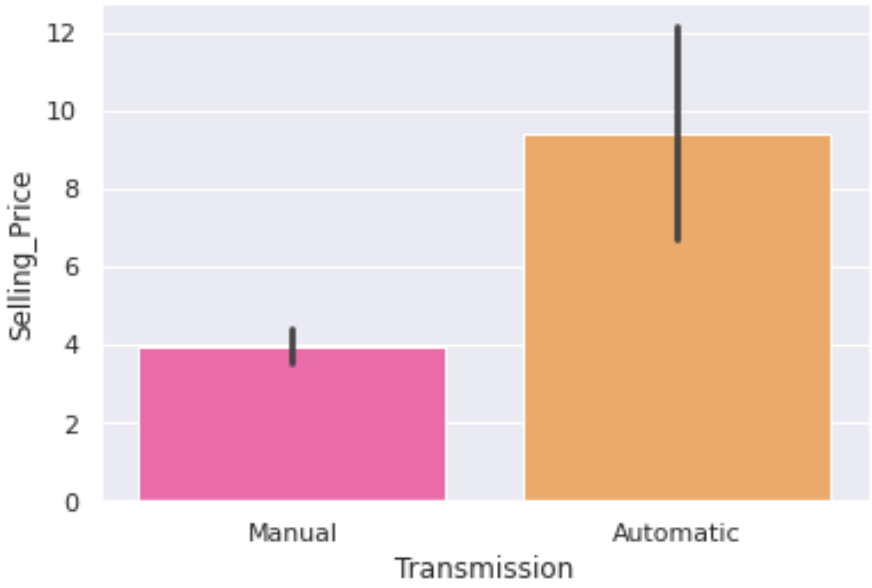
Now we shall show how selling price depend on other factors.

1) Seller_Type:

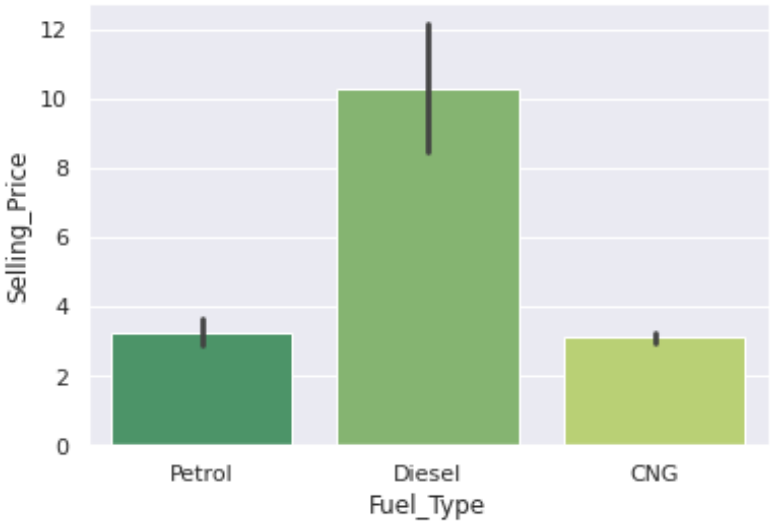
Conclusion:Selling Price of cars seems to have higher prices when sold by Dealers when compared to Individuals

2) Transmission:

It can be observed that Selling Price would be higher for cars that are Automatic



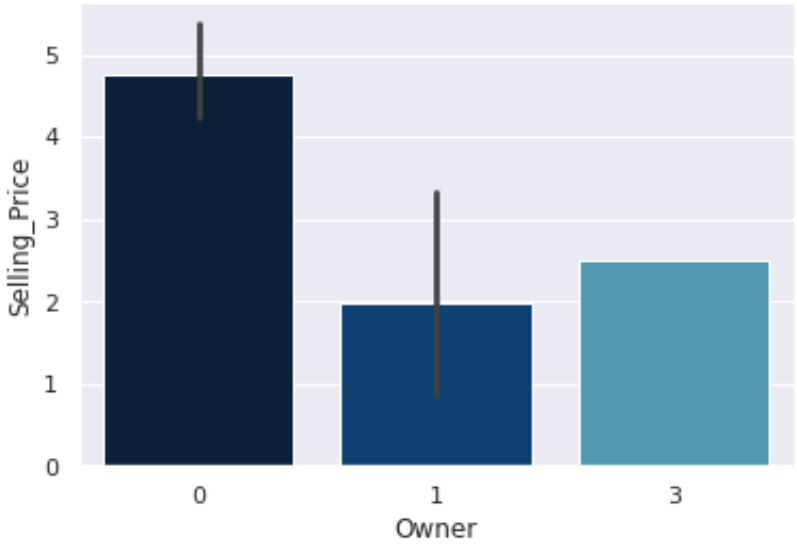
*3)**Fuel_Type***



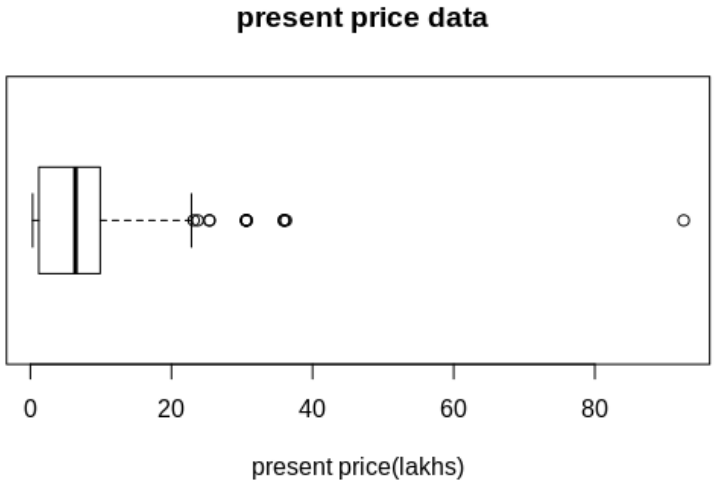
Conclusion:Selling Price of cars with Fuel Type of Diesel is higher than Petrol and CNG.

4)owner:

Selling Price is high with less Owners used Cars

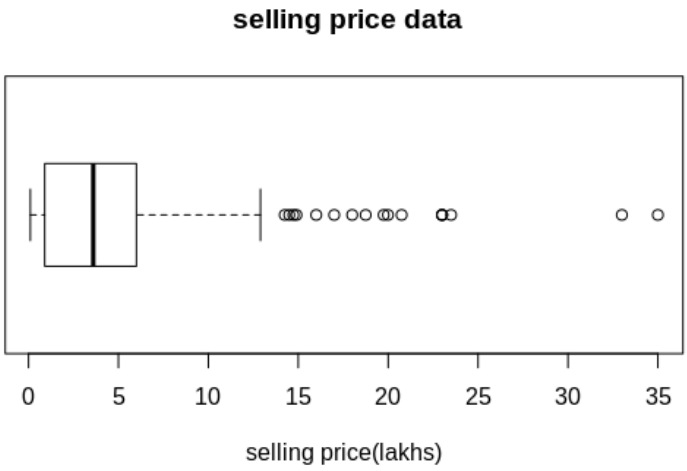


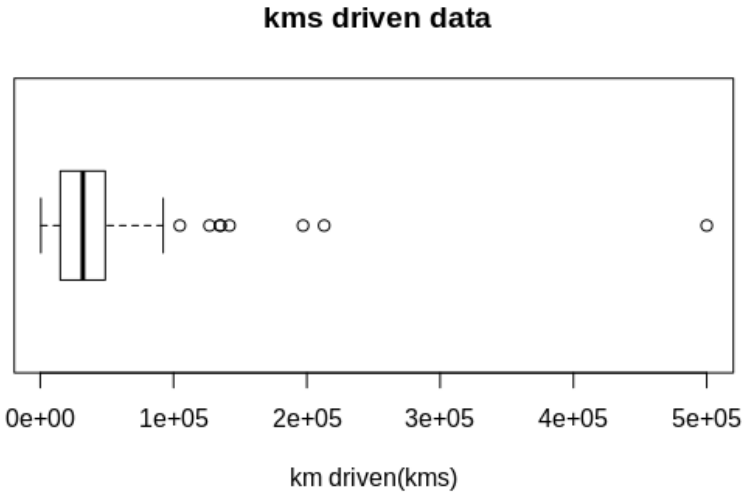
Boxplot:, boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles.] In addition to the box on a box plot, there can be lines (which are called whiskers) extending from the box indicating variability outside the upper and lower quartiles, thus, the plot is also termed as the box-and-whisker plot and the box-and-whisker diagram. Outliers that differ significantly from the rest of the dataset may be plotted as individual points beyond the whiskers on the box-plot.



Most of the cars' present price lies in 1 to 21 lakhs

Most of the cars' selling price lies in 1 to 14 lakhs





CovarianceMatrix(Dispersion Matrix):

call:cov(car.data[, c('Year','Selling_Price','Present_Price','Kms_Driven')])

	Year	Selling_Price	Present_Price	Kms_Driven
Year	8.361085	3.470617	-1.189364	-5.895887e+04
Selling_Price	3.470617	25.834973	38.619337	5.768966e+03
Present_Price	-1.189364	38.619337	74.720731	6.845447e+04
Kms_Driven	-58958.869767	5768.965732	68454.466283	1.512190e+09

Above matrix shows how each of the variables is varying with one another. It is observed that year has negative relationship present price and kms driven respectively that is they are going in opposite direction. And all other variables are having positive relation between them.

Correlation Matrix:correlation between multiple variables:

In this method to compute the correlation between all the variables in the given data frame, we needs to call the cor() function with the entire data frame passed as its parameter to get the correlation between all variables of the given data frame in the R programming language.

Call: cor(car.data[, c('Year','Selling_Price','Present_Price','Kms_Driven')])

	Year	Selling_Price	Present_Price	Kms_Driven
Year	1.00000000	0.23614098	-0.04758421	-0.52434204
Selling_Price	0.23614098	1.00000000	0.87898255	0.02918709
Present_Price	-0.04758421	0.87898255	1.00000000	0.20364703
Kms_Driven	-0.52434204	0.02918709	0.20364703	1.00000000

selling_price **is highly correlated with present_price. selling_Price is positively correlated with**

present price and year , kms driven respectively which implies selling price increases as Present_Price and year ,kms driven increases.

present price is highly correlated with selling price, and has moderately high correlation with kms driven. It is positively correlated with selling price and kms driven and negatively correlated with year.
Kms driven though considered as a general measure of car quality does not have high correlation with any of the other variables. Kms driven is negatively correlated with year. Means if year increases kms driven decreases.

Among the variables, selling price has high correlation with present price, So we construct two simple linear models of selling price as a function of present price and kms driven as a funtion of year.

Simple linear regression of selling price vs present price:

call: x <-car.data$Present_Price

y <-car.data$Selling_Price

Call:

lm(formula = y ~ x, data = car.data)

Coefficients:

(Intercept) **x**

0.7185 **0.5168**

This output indicates that the fitted value is given by
y =.7185 + .5168x

call:

lm(formula = Selling_Price ~ Present_Price, data = car.data)

Residuals:

Min	1Q	Median	3Q	Max
-13.5787	-0.7321	-0.3783	0.8731	13.5560

Coefficients:

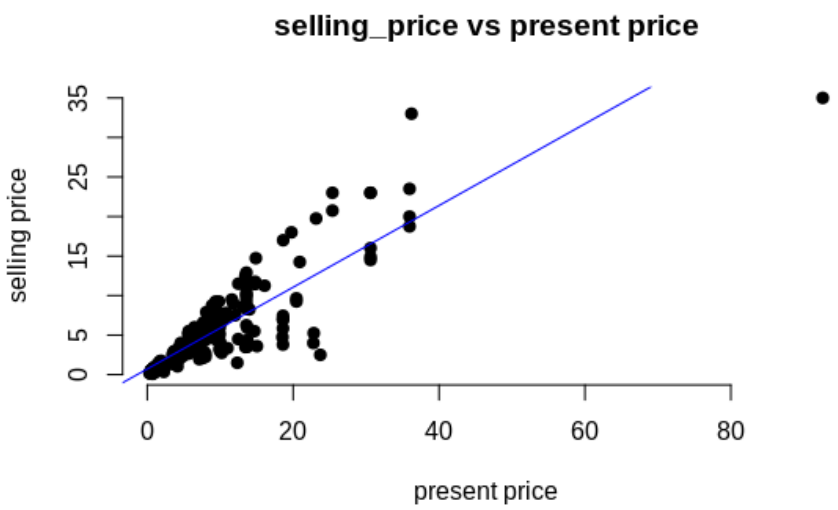
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.71853	0.18677	3.847	0.000146	***
Present_Price	0.51685	0.01622	31.874	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.428 on 299 degrees of freedom

Multiple R-squared: 0.7726, Adjusted R-squared: 0.7718

F-statistic: 1016 on 1 and 299 DF, p-value: < 2.2e-16



conclusion:

According to the model, selling price increases as present price inscrease. The coefficients of the model are very significant to the model. Residual Standard Error is small which implies the actual values of the dataset is close to the predicted values. . Adjusted R^2value is 0.7716 which means

77.16% variation of selling price can be explained by the model. Same for p value: since p-value < alpha=0.05, we reject the null hypothesis at 5% level of significance.

The p-value of the model is < 2.2e-16 so the null hypothesis(there is no relationship between selling price and present price) is rejected, this means that coefficients are significant .

Over all, we can conclude that the model is fitting the data satisfactorily.

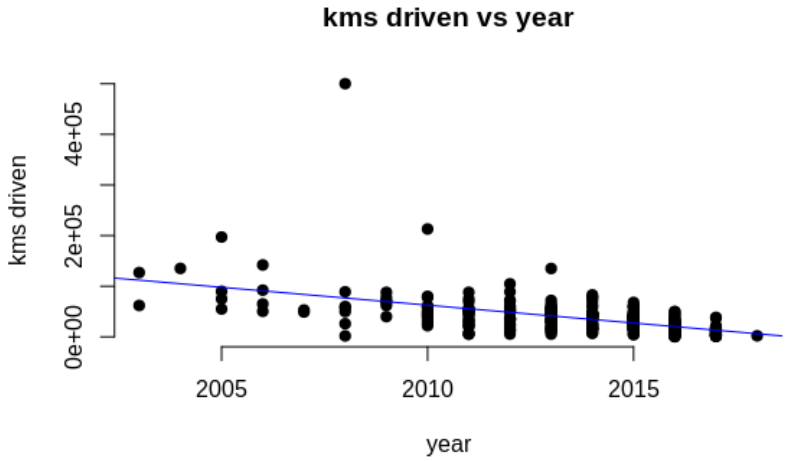
2) Simple linear regression model of kms driven vs year :

call: c <-car.data$Year

d <-car.data$Kms_Driven

Call:

lm(formula = Kms_Driven ~ Year, data = car.data)



Residuals:

Min	1Q	Median	3Q	Max
-74733	-12652	-2169	10677	423367

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	14236207.4	1333535.2	10.68	<2e-16	***
Year	-7051.6	662.3	-10.65	<2e-16	***

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33170 on 299 degrees of freedom

Multiple R-squared: 0.2749, Adjusted R-squared: 0.2725

[plot(c, d, main = "kms driven vs year", xlab = "year", ylab = "kms driven", pch = 19, frame = FALSE)]

[abline(lm(d ~ c, data = car.data), col = "blue")]

under [] it is the code for the above graph.

F-statistic: 113.4 on 1 and 299 DF, p-value: < 2.2e-16

The linear model of kms driven as a function of year is given by: d=14236207.4 -7051.6c. Here kms driven is slowly decrease with year.

We begin by testing whether the explanatory variables collectively have an effect on the response variable, i.e. H_0: $\beta_1 = 0$

The output shows that observed value of statistic is 33170. Since observed value of the test statistic is greater than upper alpha point of t distribution with 299 df, we reject the null hypothesis at 5% level of significance. Hence this indicate that year is not good predictors for kms driven .

In addition, the output also shows that$R^2 = 0.2745$ and R^2adjusted = 0.2725

Same for p value: since p-value < alpha=0.05, we reject the null hypothesis at 5% level of significance.

CONCLUSION: Since null hypothesis is rejected, this means that coefficients are significant but this model is not adequate since here R^2

is .2725 i.e 27.25% of total variabilities captured by this model. Hence this model is not a great fit.

3)Multiple linear model of selling price as a function of present price,year,kms driven:

call:

lm(formula = Selling_Price ~ Year + Present_Price + Kms_Driven,

data = car.data)

Residuals:

Min	1Q	Median	3Q	Max
-12.5282	-0.9751	-0.0672	0.8096	11.6446

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-9.685e+02	9.368e+01	-10.338	<2e-16	***
Year	4. 813e-01	4.649e-02	10.352	<2e-16	***
Present_Price	5.256e-01	1.353e-02	38.859	<2e-16	***
Kms_Driven	-1.214e-06	3.527e-06	-0.344	0.731	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.978 on 297 degrees of freedom

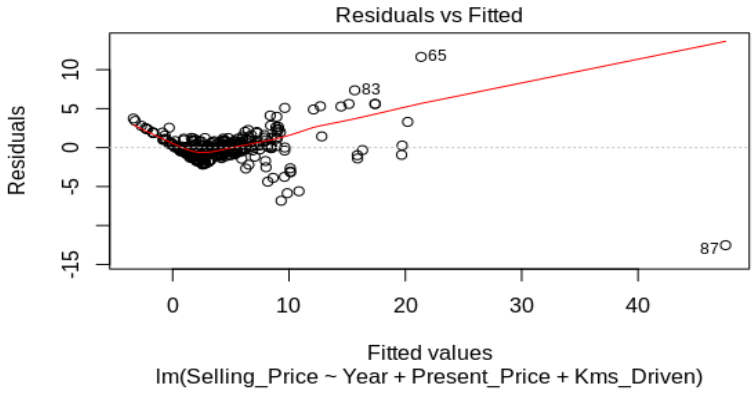
Multiple R-squared: 0.8501, Adjusted R-squared: 0.8486

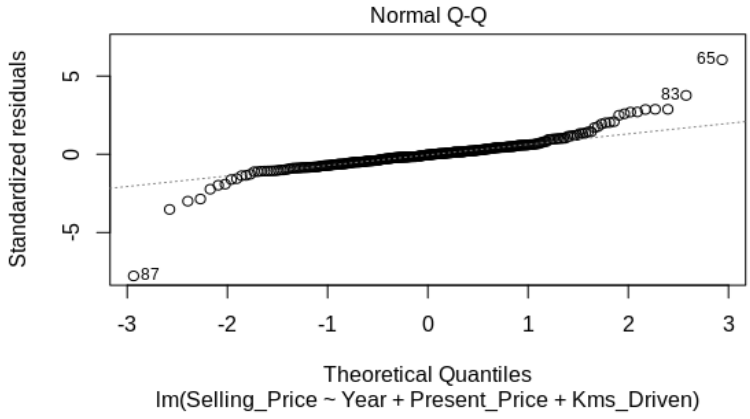
F-statistic: 561.5 on 3 and 297 DF, p-value: < 2.2e-16

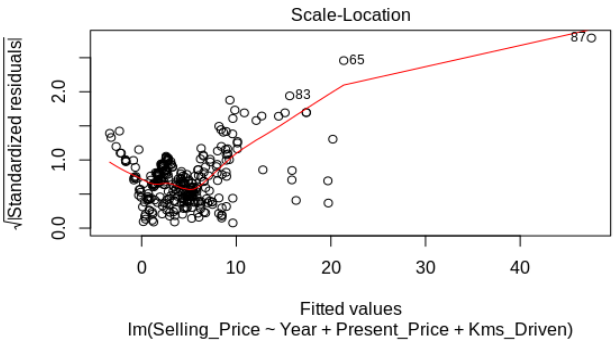
The multiple linear model of selling price as a function of year, present price ,kms driven :

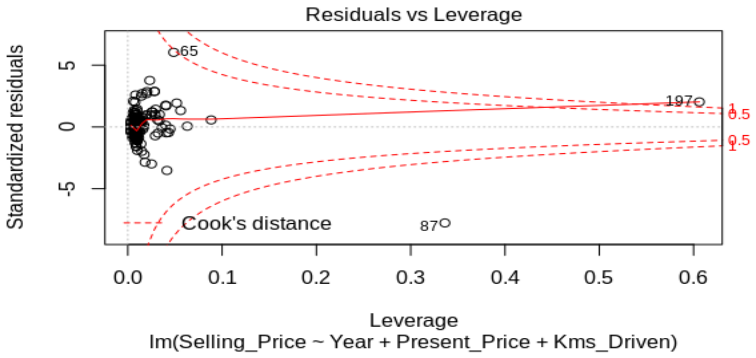
$y = -9.685e+02 + 4.813e-01x + 5.256e-01x_2 - 1.214e-06x_3$

The coefficients of year and present are very significant to the model, the coefficient of intercept has significance in the model but kms driven does not have any significance in the model. Though year has low correlation with selling price, it has high significance in the model.Residual Standard Error is 1.978 on 297 degrees of freedom which is small which implies the actual values of the dataset is close to the predicted values. Adjusted R^2value is 0.8486 which means 84.86% variation of selling price can be explained by the model. The p-value of the model is < 2.2e-16 so the null hypothesis is rejected. Over all, we can conclude that the model is fitting the data **satisfactorily**.









Residual observed vs predicted values are plot against predicted values to check graphically to check hene the residual are following normal distribution.

Now I am interested to predict the value of the selling price on the bases of the above multiple linear regression model. So here I use predict function on r.
call:
Pred_values = predict(model,newdata = car.data)

car.data$pred_sellprice = Pred_values
>View(car.data)
in my data set ,add a extra column named pred_sellprice which contain all predicted value of selling price using the above model. I observed that predicted value is very close to the main value.

n the above multiple linear regression model, the coefficients of kms driven is not significant to the data. So, I remove kms driven from the model and construct another multiple linear model.

4)Multiple linear model of selling price as a function of present price,year:

Call:

lm(formula = Selling_Price ~ Present_Price + Year, data = car.data)

Residuals:

Min	1Q	Median	3Q	Max
-12.4644	-0.9575	-0.0662	0.8236	11.6817

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-985.45942	79.49404	-12.40	<2e-16	***
Present_Price	0.52464	0.01321	39.73	<2e-16	***
Year	0.48972	0.03948	12.41	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.975 on 298 degrees of freedom

Multiple R-squared: 0.8501, Adjusted R-squared: 0.849

F-statistic: 844.7 on 2 and 298 DF, p-value: < 2.2e-16

The model of selling price as a funcion of year,present price $y = 0.52464x_1 + 0.48972x_2$

-985.45942

The coefficients of all the variables are very significant to the model. Residual Standard Error is 1.975 on 298 degrees of freedom which is small which implies the actual values of the dataset is close to the predicted values. Adjusted R^2 value is 0.849 which means 84.9% variation of selling can be explained by the model. The p-value of the model is < 2.2e-16 so the null hypothesis is rejected.

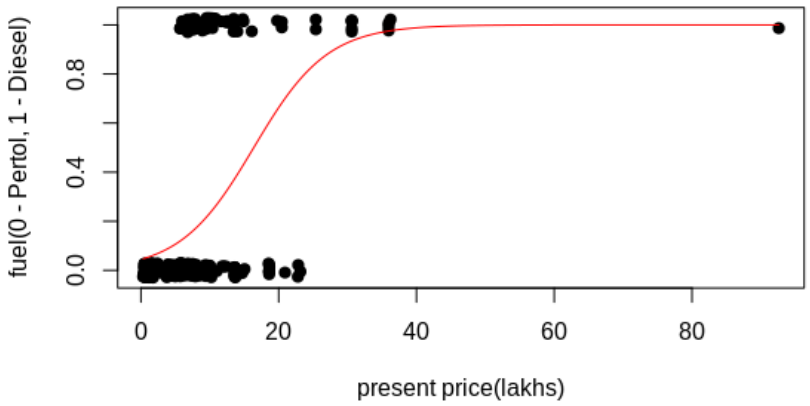
Over all, we can conclude that The model is fitting the data satisfactorily.
The second model explained the variation of selling price more effectively than the first model.

5)Logistic Regression Model of Diesel(1) and Petrol(0) Status as a function of present price:

```
View(car.data.logist)
  fuel <-car.data.logist$Fuel_Type
  presentprice <-car.data.logist$Present_Price
  fuelcode<- ifelse(fuel =="Diesel", 1 ,0)
```

here in r code ,fuelcode implies 1 for Diesel ,0 for petrol.
Call:

glm(formula = fuelcode ~ presentprice,
family = binomial)



Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7405	-0.5723	-0.3483	-0.3177	2.0610

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.06218	0.31936	-9.589	< 2e-16 ***
presentprice	0.18698	0.02845	6.573	4.95e-11 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 299.35 on 297 degrees of freedom [it is a binary logistic model]
Residual deviance: 226.22 on 296 degrees of freedom

AIC: 230.22

Number of Fisher Scoring iterations: 5

They correspond to the following model: presentprice= -3.06218(intercept) +.18698 fuel type(0/1)

The variable, diesel is equal to 1. When the fuel type is equal to petrol it is equal to 0.If I am predicting the present price of petrol cars then the above equation is presentprice= - 3.06218(intercept) +0.18698 0 = -3.06218 . If we are predicting price for diesel cars, we get the following equation : presentprice= -3.06218(intercept) +0.18698*1 =

Present price of diesel car is much higher then the present price of those cars which are driven by petrol. The both p-value of the model is well below from .05 so the null hypothesis is rejected and both the price is statistically significant.

With logistic regression , we estimate the mean of the data , and the variance is derived from the mean. Since we are not estimating the variance from the data (instead just deriving it from the mean) it is possible that the variance is underestimated. Then we have the AIC(Akaike Information criterion) which in this context ,is just the residual deviance adjusted for the number of parameter in the model. AIC is used to compare one model to others.

Over all, we can conclude that The model is fitting the data satisfactorily.

6)Logistic Regression Model of Diesel(1) and Petrol(0) Status as a function of selling price:

Call:

glm(formula = fuelcode ~ sellprice, family = binomial)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6193	-0.5627	-0.3278	-0.2825	2.1625

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.28578	0.34435	-9.542	< 2e-16	***
sellprice	0.33839	0.05022	6.738	1.61e-11	***

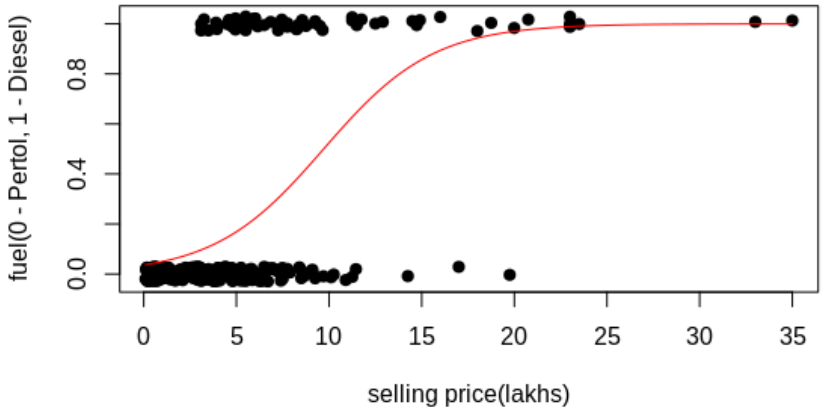
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 299.35 on 297 degrees of freedom

Residual deviance: 211.31 on 296 degrees of freedom

AIC: 215.31



Number of Fisher Scoring iterations: 5

They correspond to the following model: sellprice= -3.28578 (intercept) +.18698 fuel type(0/1)

The variable, diesel is equal to 1. When the fuel type is equal to petrol it is equal to 0.If I am predicting the selling price of petrol cars then the above equation is sellprice= -3.28578 (intercept) + 0.33839*0

=0.33839 If we are predi cting price for diesel cars, we get the following equation : sellprice= -3.28578 (intercept) +0.33839 *1 = -2.94739

Present price of diesel car is much higher then the present price of those cars which are driven by petrol.The both p-value of the model is well below from .05 so the null hypothesis is rejected and both the price is statistically significant.

7)Logistic regression model of Desiel(1) and petrol(0) status as a function of selling price, present price, kms driven:

Call:

glm(formula = fuelcode ~ car.data.logist$Present_Price + car.data.logist$Selling_Price +

car.data.logist$Kms_Driven, family = binomial, data = car.data.logist)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6464	-0.5521	-0.2934	-0.2254	2.1252

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.989e+00	4.657e-01	-8.566	< 2e-16	***
car.data.logist$Present_Price	-1.835e-02	5.881e-02	-0.312	0.7550	
car.data.logist$Selling_Price	3.852e-01	9.211e-02	4.182	2.89e-05	***
car.data.logist$Kms_Driven	1.481e-05	6.140e-06	2.412	0.0159	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 299.35 on 297 degrees of freedom

Residual deviance: 200.43 on 294 degrees of freedom

AIC: 208.43

Number of Fisher Scoring iterations: 6

Among the above 3 models, the first logistic regression model of Status(petrol 0 & desiel:1) as a function of present price has the highest AIC value which implies the first model fits the data less satisfactorily.

Among the above 3 models, the first logistic regression model of Status(automatic:0 & manual:1) as a function of present price has the lowest AIC value which implies the first model fits the data more satisfactorily.

The higher no of deviance indicates bad fit of the data. In first model, the null deviance is 168.09 with 178 degrees of freedom and the residual deviance reduces by (168.09-65.429)=102.661 loss of one degree of freedom. In the second model, the null deviance is 165.682 with 172 degrees of freedom and the residual deviance reduces by (165.682-66.434)=99.248 with loss of one degree of freedom. In third model, the null deviance is 165.682 with 172 degrees of freedom and the residual deviance reduces by (165.682-57.289)=108.393 with loss of six degrees of freedom. In all the models, addition of independent variables decreased the deviance significantly.

8)**Logistic Regression Model of seond owner(1) and first owner(0) Status as a function of selling price:**

Call:glm(formula = ownercode ~ car.data.logist$Selling_Price, family = binomial, data = car.data.logist)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.3811	-0.3367	-0.2290	-0.1598	2.8247

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.5592	0.4398	-5.819	5.92e-09	***
car.data.logist$Selling_Price	-0.2689	0.1477	-1.820	0.0687	.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 87.551 on 297 degrees of freedom
Residual deviance: 82.286 on 296 degrees of freedom

AIC: 86.286

Number of Fisher Scoring iterations: 7

They correspond to the following model: sellprice= -2.5592 (intercept) -0.2689 owner type(0/1)

The variable,second owner is equal to 1. When the owner type is first owner(in our data it is indicate by 0) it is equal to 0.If I am predicting the selling price of first owner cars then the above equation is :selling priec -2.5592 (intercept) -0.2689 *0= -2.5592If we are predicting price for second owner cars, we get the following equation : sellprice= -2.5592 (intercept) -0.2689*1= -2.8281

selling price of first owner car is much higher then the selling price of second owner cars .The both p-value of the model is well below from .05 so the null hypothesis is rejected and both the price is statistically significant.

Logistic Regression Model of Manual(1) and Autometic(0) Status as a function of present price:

all:
glm(formula = transmissioncode ~ presentprice, family = binomial)
Deviance Residuals:

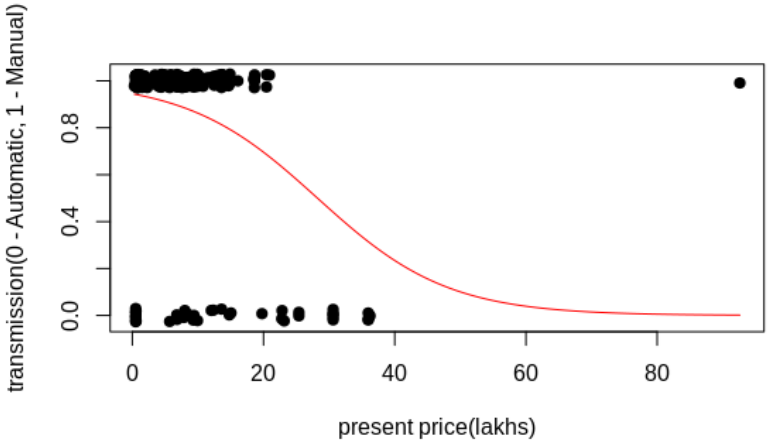
Min	1Q	Median	3Q	Max
-2.3869	0.3506	0.4196	0.5117	3.5999

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.84020	0.29127	9.751	< 2e-16	***
presentprice	-0.10063	0.02129	-4.726	2.29e-06	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)



Null deviance: 231.27 on 297 degrees of freedom

Residual deviance: 204.15 on 296 degrees of freedom

AIC: 208.15

Number of Fisher Scoring iterations: 5

They correspond to the following model: presentprice= 2.84020(intercept) +. -0.10063transmission type(0/1)

The variable,manual is equal to 1. When the transmission type is equal to automatic it is equal to 0.If I am predicting the present price of automatic cars then the above equation is presentprice= 2.84020(intercept) +. -0.10063 *0= 2.84020 If we are predicting price for manual cars, we get the following equation : presentprice= 2.84020(intercept) +. -0.10063*1 = 2.70063

Present price of automatic car is much higher then the present price of manual cars .The both p-value of the model is well below from .05 so the null hypothesis is rejected and both the price is statistically significant.

9)Logistic Regression Model of Manual(1) and Autometic(0) Status as a function of selling price:

call:glm(formula = transmissioncode ~ sellprice, family = binomial)

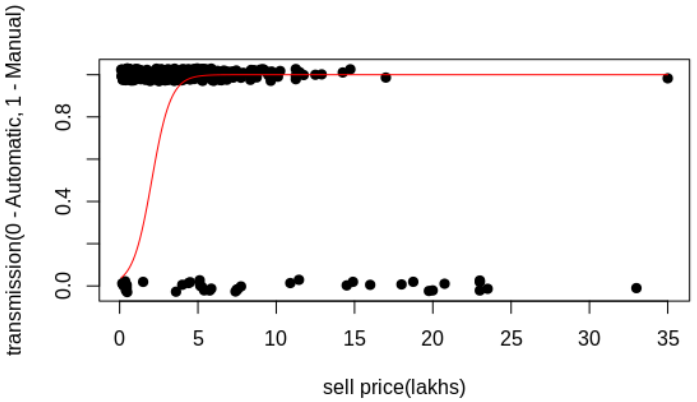
Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.4093	0.3444	0.4111	0.5015	2.4087

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.87374	0.28577	10.056	< 2e-16	***
sellprice	-0.16338	0.03214	-5.083	3.72e-07	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



(Dispersion parameter for binomial family taken to be 1)

Null deviance: 231.27 on 297 degrees of freedom
Residual deviance: 199.83 on 296 degrees of freedom

AIC: 203.83

Number of Fisher Scoring iterations: 5

They correspond to the following model: sellprice= 2.87374 (intercept) -0.16338transmission type(0/1)

The variable,manual is equal to 1. When the transmission type is equal to automatic it is equal to 0.If I am predicting the selling price of automatic cars then the above equation is selling price=2.87374 (intercept) -0.16338 *0= 2.87374 If we are predicting price for manual cars, we get the following equation : sellprice= 2.87374 (intercept) -0.16338*1= 2.15064

selling price of automatic car is much higher then the selling price of manual cars .The both p-value of the model is well below from .05 so the null hypothesis is rejected and both the price is statistically significant.

10)Logistic Regression Model of Manual(1) and Autometic(0) Status as a function of selling price,present price,kms driven:

glm(formula = transmissioncode ~ car.data.logist$Selling_Price + car.data.logist$Present_Price

+ car.data.logist$Kms_Driven, family = binomial, data = car.data.logist)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5409 0.3134 0.4041 0.4957 2.2358

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.242e+00	3.447e-01	9.405	< 2e-16	***
car.data.logist$Selling_Price	-1.941e-01	7.265e-02	-2.672	0.00754	**
car.data.logist$Present_Price	1.982e-02	4.586e-02	0.432	0.66552	
car.data.logist$Kms_Driven	-8.935e-06	4.210e-06	-2.122	0.03380	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 231.27 on 297 degrees of freedom
Residual deviance: 194.80 on 294 degrees of freedom

AIC: 202.8

Number of Fisher Scoring iterations: 5

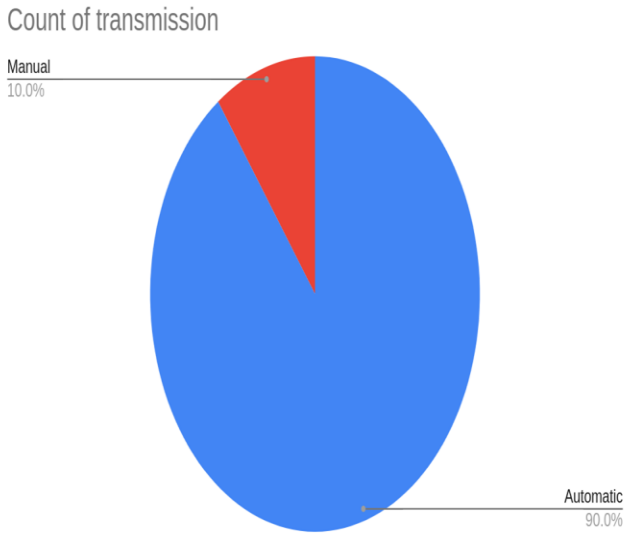
With logistic regression , we estimate the mean of the data , and the variance is derived from the mean. Since we are not estimating the variance from the data (instead just deriving it from the mean) it is possible that the variance is underestimated. Then we have the AIC(Akaike Information criterion) which in this context ,is just the residual deviance adjusted for the number of parameter in the model. AIC is used to compare one model to others.

Among the above 3 models, the third logistic regression model of Status(automati:0 & manual :1) as a function of selling price, present price, kms driven has the lowest AIC value which implies the third model fits the data more satisfactorily.

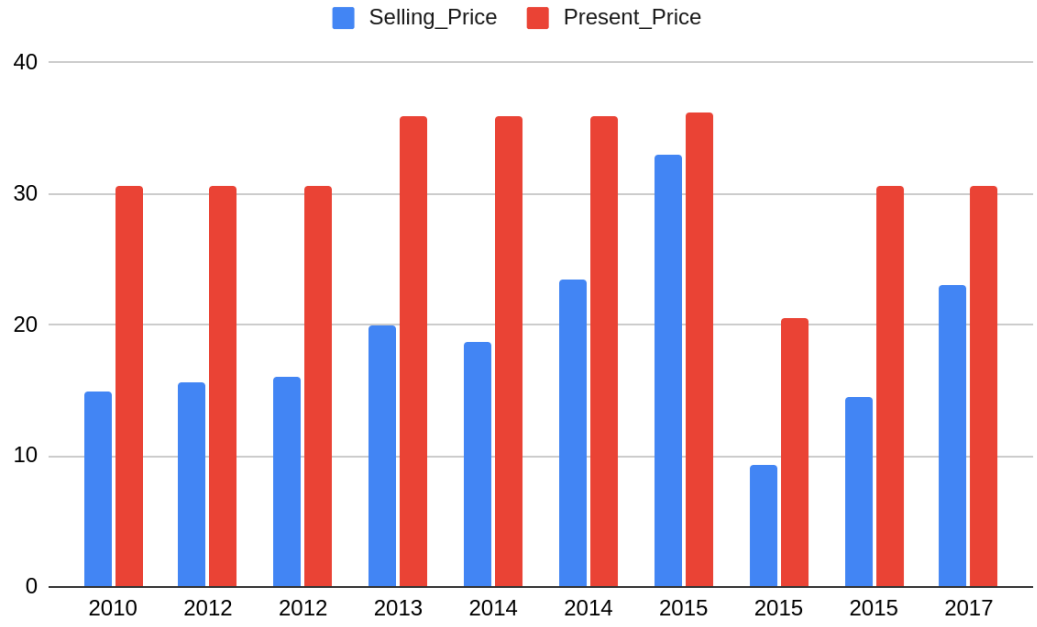
The higher no of deviance indicates bad fit of the data. In third model, the null deviance is 231.27 on 297 degrees of freedom and the residual deviance reduces by (231.27 -194.80 =36.47 loss of three degree of freedom. In the second model, the null deviance is231.27 on 297 degrees of freedom and the residual deviance reduces by (231.27 -199.83)=31.44 with loss of one degree of freedom. Infirst model, the null deviance is 231.27 on 297 degrees of freedom and the residual deviance reduces by (231.27-204.15)= 27.12 with loss of one degrees of freedom. In all the models, addition of independent variables decreased the deviance significantly.

Now I shall do the model of some specific cars. From our main data set I extract different cars data.

1)**Fortuner car**: I have fortuner car data from 2010 to 2017.Most of the fortuner are automatic. The fuel type is diesel.



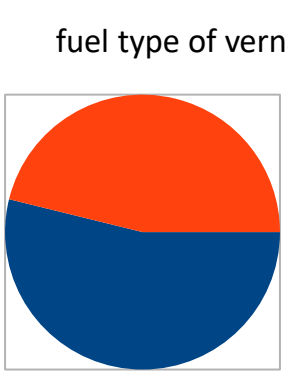
here we see the yearwise price variation:

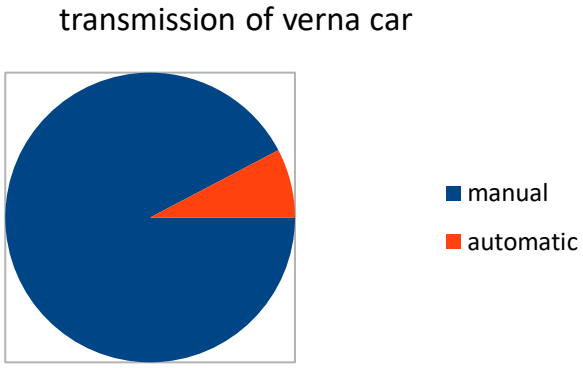


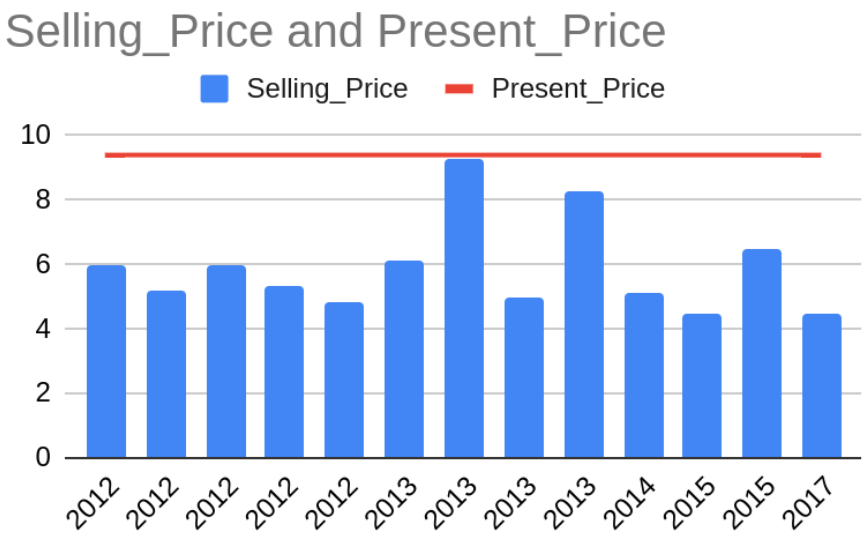
fortnuer cars are high present price as most of the car of this model are automatic. Almost every year present price is above 30 lakhs. And selling price is around 20 lakhs.

Varna car: the fuel of all verna car are diesel and petrol both. most of the verna cars are driven by human meas cars are manual not automatic. we have 2012 to 2017 verna car data.

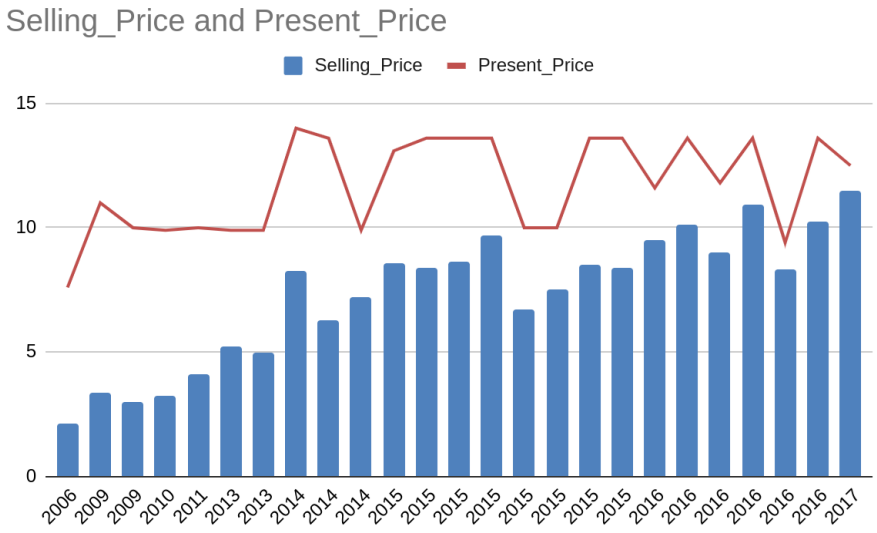
Here year wise present price is constant. Selling price is different. Fuel of verna car are 53.8% petrol and 46.2% diesel .







city car: we have 2006 to 2017 city car data
. Below the graph is 2006 to 2017 selling price and present price variation of city car.



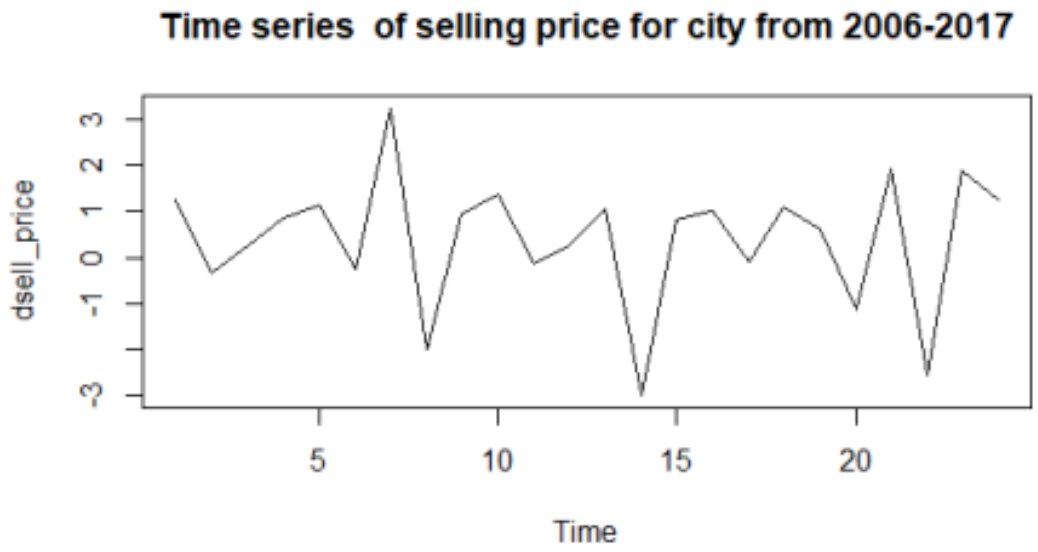
city car price forcast using time series analysis:

Augmented Dickey-Fuller Test
data: car.data. ...car.data$Selling_Price
Dickey-Fuller = -1.6897, Lag order = 2, p-value = 0.6906
alternative hypothesis: stationary

From the ADF test we get p value greater than 0.05, which means our data is not stationary. Since we need stationary data to apply the ARIMA process, we shall convert the non-stationary data into stationary data through differencing.
After differencing the result we got:

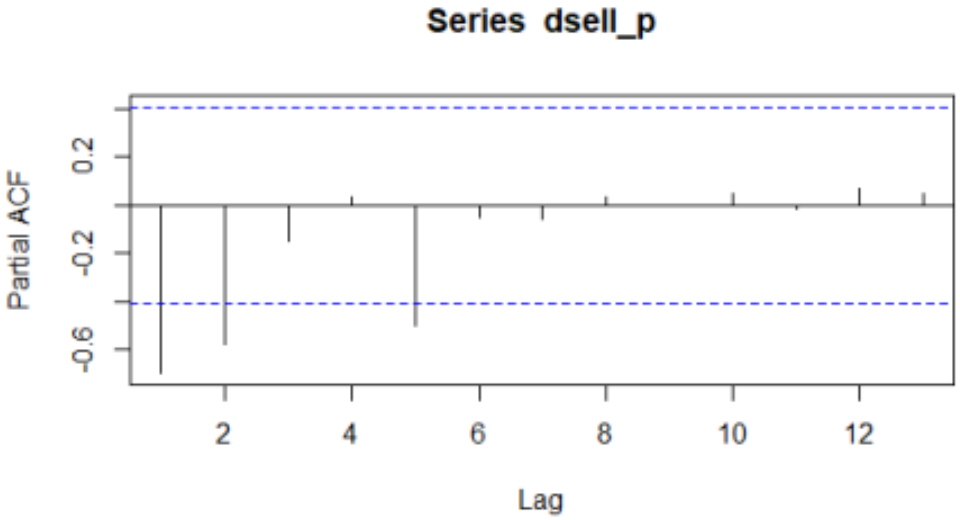
Augmented Dickey-Fuller Test
data: dsell_price
Dickey-Fuller = -2.7343, Lag order = 2, p-value = 0.2926
alternative hypothesis: stationary
From the ADF test again we get p value greater than 0.05. Again we apply ADF test to convert the non-stationary data into stationary data through differencing.
Augmented Dickey-Fuller Test

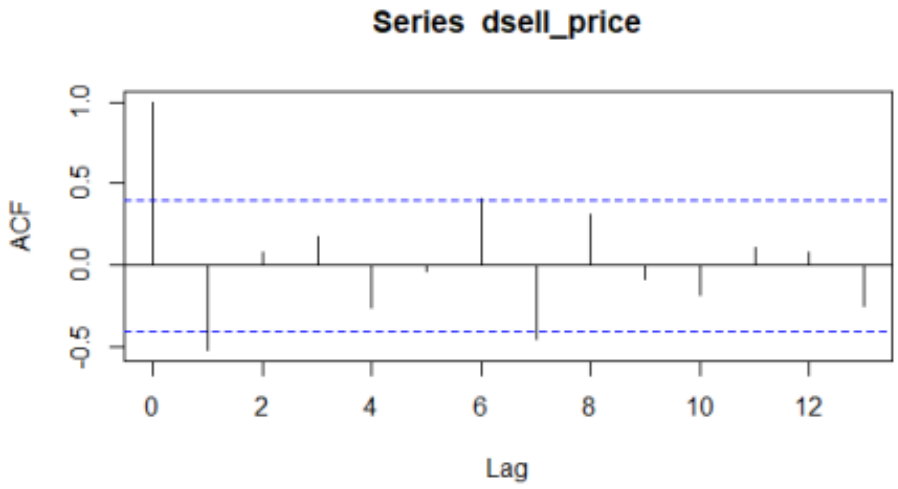
data: dsell_p
Dickey-Fuller = -4.0633, Lag order = 2, p-value = 0.02105
alternative hypothesis:stationary
now the p value becomes less than .05. so the dara is stationary.



stationary time series

P**lotting ACF & PACF:the** ACF & PACF diagosis is employed over time series to determine the order in which we are going to create our model using ARIMA model. A time series is stationary when its mean, variance , and autocorrelation remain constant over time.





BEST FITTED ARIMA MODEL FOR THE ABOVE FORECAST: ARIMA (2,0,1)

Series: dsell_p

ARIMA(2,0,1) with zero mean

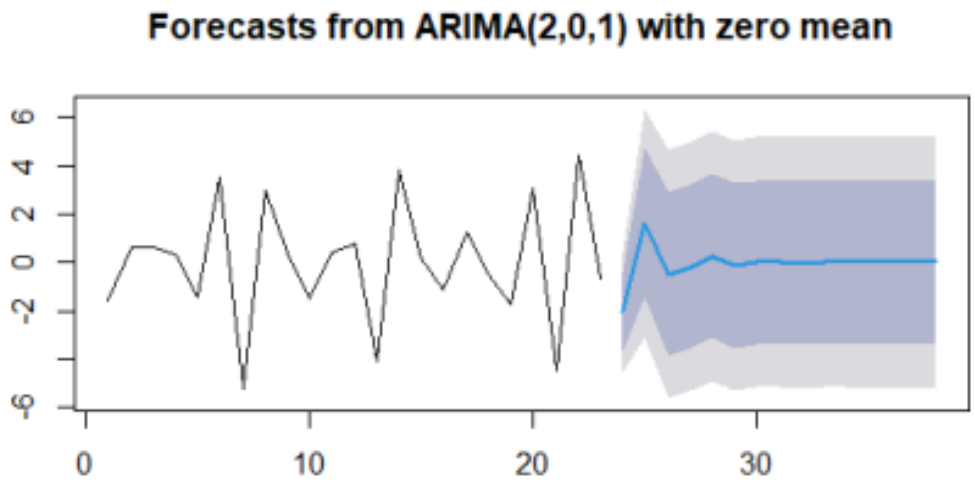
Coefficients:

ar1	ar2	ma1
-0.6930	-0.3133	-0.9296
s.e. 0.2321	0.2375	0.2577

sigma^2 = 1.602: log likelihood = -38.35

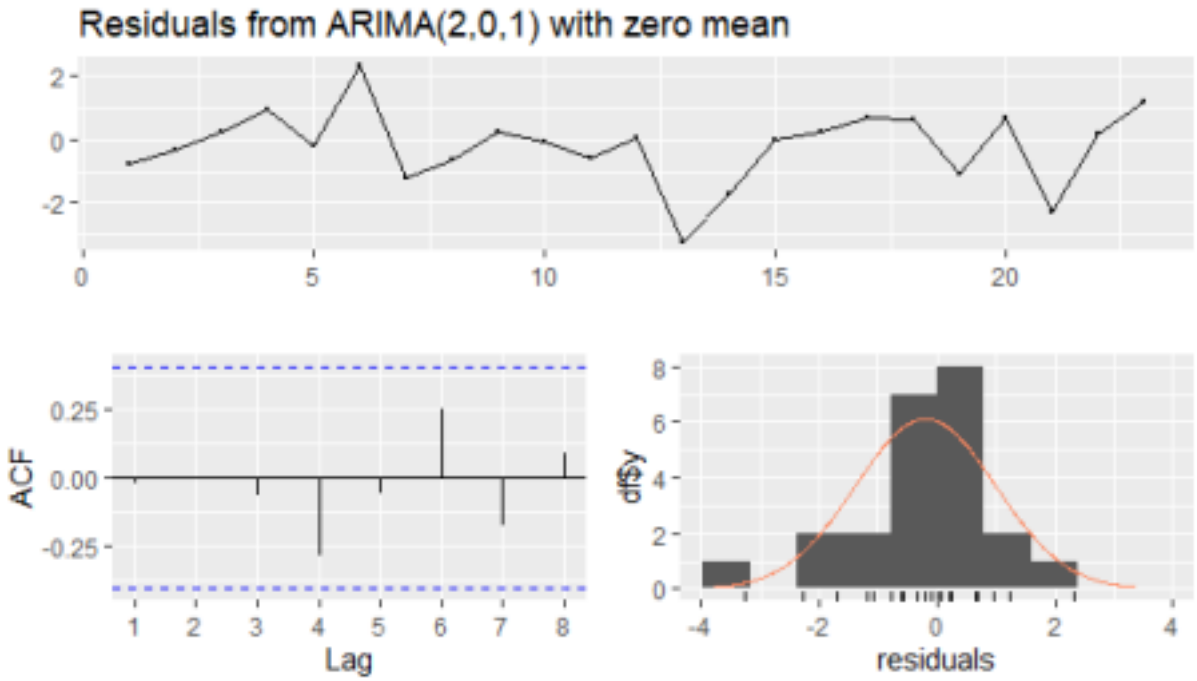
AIC=84.69 AICc=86.91 BIC=89.23

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
24	-2.068169e+00	-3.694542	-0.4417955	-4.555492	0.4191545
25	1.636940e+00	-1.456100	4.7299794	-3.093456	6.3673356
26	-4.866125e-01	-3.847108	2.8738827	-5.626046	4.6528214
27	-1.755296e-01	-3.537130	3.1860707	-5.316654	4.9655944
28	2.740809e-01	-3.105865	3.6540270	-4.895101	5.4432624
29	-1.349648e-01	-3.525741	3.2558114	-5.320709	5.0507799
30	7.680198e-03	-3.383981	3.3993419	-5.179419	5.1947791
31	3.695509e-02	-3.354849	3.4287594	-5.150362	5.2242722
32	-2.801724e-02	-3.420131	3.3640968	-5.215808	5.1597736
33	7.840913e-03	-3.384345	3.4000265	-5.180059	5.1957411
34	3.342337e-03	-3.388843	3.3955280	-5.184558	5.1912427
35	-4.772549e-03	-3.396964	3.3874189	-5.192682	5.1831367



The table and the graph give the 80% and 95% intervals of the point forecast values of selling price for the upcoming years.Here, some of the forecasted values of the selling price are negative that are practically and statistically insignificant. Henceforth, the minimum possible value for selling price can be 0 which is again very impossible but can be considered to be in the said range. So, the forecasted values of the lower boundaries of the intervals can be considered to range from the minimum value 0.

To check the accuracy of our proposed forecasted model, we check the residuals plot.



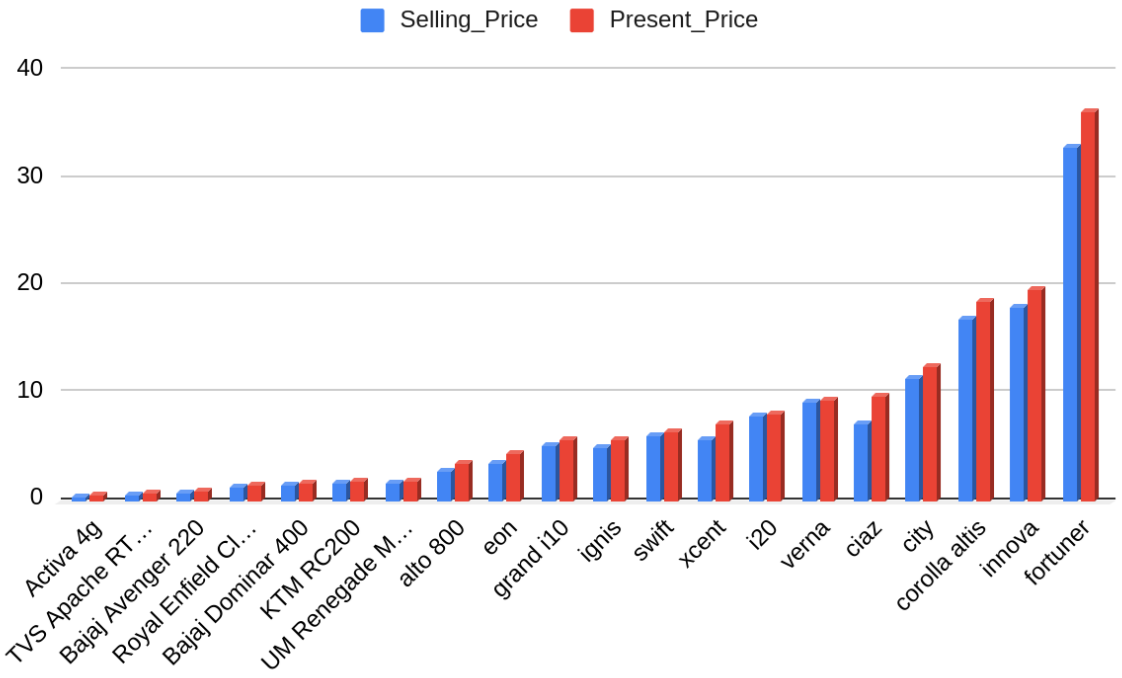
Ljung-Box test

data: Residuals from ARIMA(2,0,1) with zero mean
Q* = 4.8234, df = 3, p-value = 0.1852

Model df: 3. Total lags used: 6

The residuals plot gives a clear interpretation of the fitted ARIMA model. All the lags are within the threshold and the errors follow normal distribution.

• The p-value in Ljung-Box test is 0.1852>0.05, which implies the residuals are not autocorrelated, thus giving a good fit.



the above graph is from all car data of 2017 .We want to see which car price is low or which one is high. Activa 4g, TVS Apahe , bajaj avenger are low price car. Corolla altis, innova, fortuner are high price car.

Conclusion:

I have here analyzed the factors that may have influenced the selling price significantly.

Present price is widely used as a measurement of overall used cars selling price.

I have given here an analysis of the components that have significant relationships with selling price. In this project, I used graphical representation to observe the trend of some factor to give a I have observed that Selling Price of cars seems to have higher prices when sold by Dealers when compared to Individuals .It can be observed that Selling Price would be higher for cars that are Automatic .Selling Price of cars with Fuel Type of Diesel is higher than Petrol and CNG. Selling Price is high with less Owners used Cars. n insight of what makes a car model more valuable in market. In this project I used linear models of selling price as response variable to observe the dependence of selling price on other variables such as present price, kms driven, year. Among the simple linear regression models, the model with present price as response variable explained the variation of selling price more effectively as we know that present price is a key measurement in computing deciding selling price. From the multiple linear regression models of selling price, we can conclude that all the mentioned components
have contribution in explaining the variation of selling price.
I have used binomial logistic regression models with desiel (1) and petrol(0) status of selling price as two categories of response variable and compare between different models. Analyzing the models, I have come to the conclusion that the model with price as explanatory variable has fit the data more effectively.
I have used binomial logistic regression models with second owner (1) and first owner(0) status of selling price as two categories of response variable and compare between different models. Analyzing the models, I have come to the conclusion that the model with price as explanatory variable has fit the data more effectively.
I have used binomial logistic regression models with manual (1) and automatic(0) status of selling price and present price as two categories of response variable and compare between different models.Analyzing the models, I have come to the conclusion that the model with price as explanatory variable has fit the data more effectively.
Then I model some specific cars. I found most of the fortuner are automatic. The fuel type is diesel. fortnuer cars are high present price as most of the car of this model are automatic. Almost every year present price is above 30 lakhs. And selling price is around 20 lakhs.the fuel of all verna car are diesel and petrol both. most of the verna cars are driven by human meas cars are manual not automatic. we have 2012 to 2017 verna car data. Here year wise present price is constant. Selling price is different. Fuel of verna car are 53.8% petrol and 46.2% diesel . we have 2006 to 2017 city car data. I forcast the upcoming years selling price.

REFERENCES:

• Fundamentals of Statistics VOL. I & II (Goon, Gupta, Dasgupta)

• Applied Multivariate Statistical Analysis 6th edition (Johnson and Wichern)

• The Analysis of Time Series (C. Chatfield)

• A Little Book of R for Time Series (Avril Coghlan)

• Introduction to Linear Regression Analysis 5th edition (Douglas C. Montgomery)

methods of multivariate analysis, 3rd edition rencher and christensen

https://www.tutorialspoint.com/r/r_linear_regression.htm

http://r-statistics.co/Complete-Ggplot2-Tutorial-Part1-With-R-Code.html

https://www.scribbr.com/statistics/linear-regression-in-r/

http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/

https://www.geeksforgeeks.org/time-series-analysis-in-r/

https://www.google.com/search?q=multiple+linear+regression+in+r&oq=mu&aqs=chrome.3.69i57j35i39l2j0i67l2j69i61l2j69i60.4456j0j4&sourceid=chrome&ie=UTF-8

https://www.statmethods.net/advstats/timeseries.html

http://r-statistics.co/Outlier-Treatment-With-R.html

httphttps://www.statology.org/glm-fit-fitted-probabilities-numerically-0-or-1-occurred/
://r-statistics.co/Logistic-Regression-With-R.html

ACKNOWLEDGEMENTS:

I would like to express my sincere gratitude to several individuals and organizations for supporting me throughout my final semester project. First, I wish to express my sincere gratitude to my supervisor, Professor Dr. Moutushi Chatterjee, for her enthusiasm, patience, insightful comments, helpful information, practical advice and unceasing ideas that have helped me tremendously at all times in my research and doing of this project. Her immense knowledge, profound experience and professional expertise in statistics has enabled me to complete this research successfully. Second, I wish to thank other faculty members of our department, Dr. Bratati Chakraborty and Prof. Ipshita Samanta who were very helpful as well. Without their support and guidance, this project would not have been possible. I could not have imagined having better supervisors in my study.

I also wish to express my sincere thanks to the University of Calcutta for giving us the chance for doing project. Last but by no means least, I am also grateful to all my professors

Thanks for all your encouragement!

Code:

```
x<-c(car.data$Present_Price)

boxplot(x,horizontal = TRUE,xlab =  "present price(lakhs)", main=" present price data")
```

```
y<-c(car.data$Selling_Price)
z<-c(car.data$Kms_Driven)
stripchart(x, method = "jitter", pch = 19, add = TRUE, col = "blue")
boxplot(y,horizontal = TRUE,xlab =  "selling price(lakhs)", main=" selling price")

boxplot(z,horizontal = TRUE,xlab =  "km driven(kms)", main=" kms driven data”)

View(car.data)
colnames(car.data)
unique(car.data$Fuel_Type)
unique(car.data$year)
unique(car.data$Year)
unique(car.data$Seller_Type)
unique(car.data$Transmission)
unique(car.data$Owner)
summary(car.data$Selling_Price)
summary(car.data$Present_Price)
linear_model <- lm(Selling_Price ~ Present_Price ,data = car.data)
summary(linear_model)
abline(.71853 , .51685)
abline(.71853,.51685)
abline(lm(Selling_Price ~ Present_Price ,data = car.data))
a <-car.data$Selling_Price
a
b <-car.data$Present_Price
abline(lm(a ~ b))
plot(b,a, main = "Main title",
xlab = "X axis title", ylab = "Y axis title",
pch = 19, frame = FALSE)
abline(lm(a ~ b, data = car.data), col = "blue")
```

```
abline(lm(a ~ b, data = car.data), col = "blue")
>
plot(x, y, main = "Main title",
xlab = "X axis title", ylab = "Y axis title",
pch = 19, frame = FALSE)
c <-car.data$Year
plot(c,a, main = "Main title",
xlab = "X axis title", ylab = "Y axis title",
pch = 19, frame = FALSE)
abline(lm(a ~ c, data = car.data), col = "blue")
linear_model <- lm(Selling_Price ~ year ,data = car.data)
linear_model <- lm(Selling_Price ~ Year ,data = car.data)
summary(linear_model)
source('~/ac lm.R')
fit<-lm(Kms_Driven ~ Year,data = car.data)
summary(car.data)
summary(fit)
plot(c, d, main = "kms driven vs year",
xlab = "year", ylab = "kms driven",
pch = 19, frame = FALSE)
c <-car.data$Year
d <-car.data$Kms_Driven
plot(c, d, main = "kms driven vs year",
xlab = "year", ylab = "kms driven",
pch = 19, frame = FALSE)
abline(lm(d ~ c, data = car.data), col = "blue")
model <-lm(Presemt_Price ~ Kms_Driven, data=car.data)
model <-lm(Present_Price ~ Kms_Driven, data=car.data)
summary(model)
model <-lm(Present_Price ~ Year, data=car.data)
```

```
summary(model)
model <-lm(Year ~ Selling_Price, data=car.data)
summary(model)
model <- lm(Selling_Price ~ Year + Present_Price + Kms_Driven, data = car.data)
summary(model)
model <- lm(Selling_Price ~ Year + Present_Price , data = car.data)
summary(model)
plot(model <- lm(Selling_Price ~ Year + Present_Price + Kms_Driven, data = car.data)
source('~/ac lm.R')
plot(lm(Selling_Price ~ Year + Present_Price + Kms_Driven, data =car.data))
avPlot(lm(Selling_Price ~ Year + Present_Price + Kms_Driven, data =car.data))
model<-lm(Selling_Price ~ Year + Present_Price + Kms_Driven, data =car.data)
summary(model)
predict()
Pred_values = predict(model,newdata = test.ds)
Pred_values = predict(model,newdata = test_ds)
Pred_values = predict(model,newdata = car.data)
car.data$pred_sellprice = Pred_values
View(car.data)
model<-lm(Selling_Price ~ Year + Present_Price)
model<-lm(Selling_Price ~ Year + Present_Price
source('~/ac lm.R')
View()
view(car.data)
View(car.data)
model <- lm(Selling_Price ~ Year + Present_Price
View(car.data)
model <- lm(Selling_Price ~ Year + Present_Price)
car.data <- read.csv("~/Downloads/car data.csv")
View(car.data)
```

```
model <- lm(Selling_Price ~ Year + Present_Price)

car.data$Selling_Price

model <- lm(Selling_Price ~ Year + Present_Price)

model <- lm

model <-lm(Selling_Price ~ Present_Price + Year)

model <-lm(a ~ b ,c data=car.data)

model <-lm(Selling_Price ~ Present_Price + Year,data=car.data)

summary(model)

car.data.logist <- read.csv("~/Downloads/car data logist.csv")

View(car.data.logist)

fuel <-car.data.logist$Fuel_Type

presentprice <-car.data.logist$Present_Price

fuelcode<- ifelse(fuel =="Diesel", 1 ,0)

plot(presentprice, jitter(fuelcode, .15), pch = 19,xlab = "present price(lakhs)", ylab = "fuel(0 - Pertol, 1 - Diesel)")

model <- glm(fuelcode~presentprice, binomial)

summary(model)

xv<- seq(min(presentprice),max(presentprice),.01)

yv<-predict(model,list(presentprice=xv),type = "response")

lines(xv,yv, col = "red")

logi.hist.plot(presentprice,fuelcode,boxp=FALSE,type="count",col="gray",xlab = "size")

source('~/ac lm.R')

View(car.data.logist)

fuel <-car.data.logist$Fuel_Type

sellprice<-car.data$Selling_Price

sellprice<-car.data.logist$Selling_Price

fuelcode<- ifelse(fuel =="Diesel", 1 ,0)

plot(sellprice, jitter(fuelcode, .15), pch = 19,xlab = "selling price(lakhs)", ylab = "fuel(0 - Pertol, 1 - Diesel)")

model <- glm(fuelcode~sellprice, binomial)
```

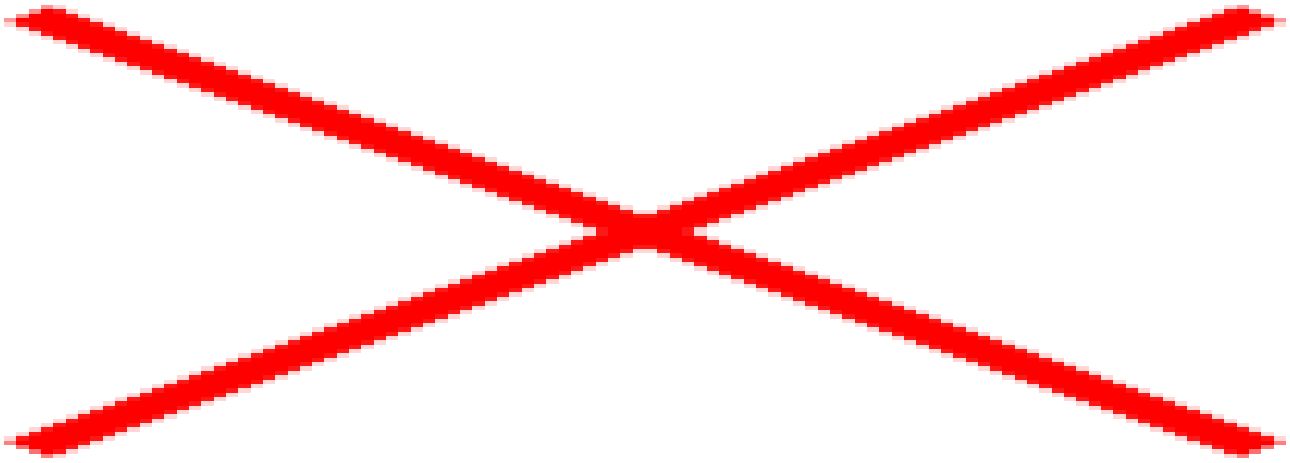
```
xv<- seq(min(sellprice),max(sellprice),.01)

yv<-predict(model,list(sellprice=xv),type = "response")

lines(xv,yv, col = "red")

summary(model)

presentprice <-car.data.logist$Present_Price

transmission<-car.data.logist$Transmission

transmissioncode<-ifelse(transmission =="Manual", 1 ,0)

plot(presentprice, jitter(transmissioncode, .15), pch = 19,xlab = "present price(lakhs)", ylab = "transmission(0 - Automatic, 1 - Manual)")

model <- glm(fuelcode~presentprice, binomial)

model<-glm(transmissioncode~presentprice, binomial)

summary(model)

xv<- seq(min(presentprice),max(presentprice),.01)

yv<-predict(model,list(presentprice=xv),type = "response")

lines(xv,yv, col = "red")

kmdriven<-car.data.logist$Kms_Driven

transmissioncode<-ifelse(transmission =="Manual", 1 ,0)

> plot(kmdriven, jitter(transmissioncode, .15), pch = 19,xlab = "kms driven", ylab = "transmission(0 - Automatic, 1 - Manual)")

plot(kmdriven, jitter(transmissioncode, .15), pch = 19,xlab = "kms driven", ylab = "transmission(0 - Automatic, 1 - Manual)")

xv<- seq(min(kmdriven),max(kmdriven),.01) yv<-predict(model,list(kmdriven=xv),type = "response")

lines(xv,yv, col = "red")

xv<- seq(min(kmdriven),max(kmdriven),.01)

yv<-predict(model,list(kmdriven=xv),type = "response")

lines(xv,yv, col = "red")

xv<- seq(min(kmdriven),max(kmdriven),.02)

        yv<-predict(model,list(kmdriven=xv),type = "response")

    lines(xv,yv, col = "red")

xv<- seq(min(kmdriven),max(kmdriven),.03)                                                    yv<-predict(model,list(kmdriven=xv),type = "response")
```

```
lines(xv,yv, col = "red")

sellprice<-car.data.logist$Selling_Price

sellertype<-car.data.logist$Seller_Type

sellertypecode<-ifelse(sellertype =="Dealer", 1 ,0)

plot(sellprice, jitter(sellertypecode, .15), pch = 19,xlab = "selling price(lakhs)", ylab = "seller type(0 - Pertol, 1 - Diesel)")

model <- glm(sellertypecode~sellprice, binomial)

model <- glm(sellertypecode~sellprice, binomial)

model1 <- glm(sellertypecode~sellprice, binomial)

fit<-glm(sellertypecode~sellprice,binomial)

plot(sellprice, jitter(sellertypecode, .15), pch = 19,xlab = "selling price(lakhs)", ylab = "seller type(0 - individual, 1 - Delear)")

xv<- seq(min(selltprice),max(sellprice),.01)

yv<-predict(model,list(sellprice=xv),type = "response")

lines(xv,yv, col = "red")

xv<- seq(min(sellprice),max(sellprice),.01)

sellertypecode<-ifelse(sellertype =="Dealer", 1 ,0)

xv<- seq(min(sellprice),max(sellprice),.01)

yv<-predict(model,list(sellprice=xv),type = "response")

lines(xv,yv, col = "red")

model <-glm(sellertypecode~sellprice,binomial)

source('~/ac lm.R')

> lines(xv,yv, col = "red")

View(car.data.logist)

sellertype<-car.data.logist$Seller_Type

sellertypecode<-ifelse(sellertype =="Dealer", 1 ,0)

plot(sellprice, jitter(sellertypecode, .15), pch = 19,xlab = "selling price(lakhs)", ylab = "seller type(0 - individual, 1 - Delear)")

model<- glm(sellertypecode~ car.data.logist$Selling_Price,family =  binomial)

transmission<-car.data.logist$Transmissiontransmissioncode<-ifelse(transmission =="Manual", 1 ,0)
```

```
plot(sellprice, jitter(transmissioncode, .15), pch = 19,xlab = "sell price(lakhs)", ylab = "transmission(0 - Automatic, 1 - Manual)")

model <- glm(transmissioncode~sellprice, binomial)

summary(model)

xv<- seq(min(sellprice),max(sellprice),.01) yv<-predict(model,list(sellprice=xv),type = "response")

lines(xv,yv, col = "red")

model<-glm(transmissioncode~ car.data.logist$Selling_Price + car.data.logist$Present_Price + car.data.logist$Kms_Driven, family = binomial,data = car.data.logist)

summary(model)

ownercode<-ifelse(Owner =="1", 1 ,0)

owner<-car.data.logist$Owner

ownercode<-ifelse(Owner =="1", 1 ,0)

ownercode<-ifelse(owner =="1", 1 ,0)

plot(sellprice, jitter(sellertypecode, .15), pch = 19,xlab = "selling price(lakhs)", ylab = "owner(0 -1st owner,1 -2nd owner)")

plot(sellprice, jitter(sellertypecode, .15), pch = 19,xlab = "selling price(lakhs)", ylab = "owner(0 - owner,1 -2nd owner)")

model<-glm(ownercode~car.data.logist$Selling_Price,family = binomial,data=car.data.logist)

summary(model)

xv<- seq(min(car.data.logist$Selling_Price),max(car.data.logist$Selling_Price),.01)

yv<-predict(model,list(car.data.logist$Selling_Price=xv),type = "response")

>   lines(xv,yv, col = "red")

yv<-predict(model,list(car.data.logist$Selling_Price=xv),type = "response")

xv<- seq(min(car.data.logist$Selling_Price),max(car.data.logist$Selling_Price),.01)

yv<-predict(model,list(car.data.logist$Selling_Price=xv),type = "response")

#time series

adf.test(car.data.city...car.data$Selling_Price)

                dsell_price= diff(car.data.city...car.data$Selling_Price)
                View(dsell_price)
                adf.test(dsell_price)
                acf(dsell_price)

                pacf(dsell_price)
```

```
arimafit <- auto.arima(dsell_price)
arimafit
forecast_model <- forecast(arimafit, h = 15)
forecast_model
plot(forecast_model)

checkresiduals(forecast_model)
```

APPENDIX:



STUDENT'S DECLARATION

I hereby declare that the project work with title "**STATISTICAL ANALYSIS OF THE USED CARS AND PRICE PREDICTION**" submitted by me for the partial fulfillment of the degree of B.Sc. Honours in Statistics under the University of Calcutta is my original work and has not been submitted earlier to any University or Institution for the fulfillment of requirement of any course of study. I also declare that no chapter in this manuscript in whole or in part has been incorporated in this report from any earlier work done by others or by me. However, extracts of my literature which has been used for this report has been duly acknowledged providing details of such literature in the references.

Signature: Chetana Das

Name : Chetana Das

Address: Sulantu, Parulia, Purba Bardhaman, 713513

University Roll No. : 193031-11-0178

University Registration No. : 031-1214-0408-19

Place: Lady Brabourne College, Kolkata

Date: 4.7.22

Supervisor's Signature : Moutushi Chatterjee

Name : Dr. Moutushi Chatterjee

Designation : Assistant Professor

Name of the college : Lady Brabourne College, Kolkata

Date: 4/7/22