Exploratory Data Analysis (EDA),Bivariate Analysis, Multiple Linear Regression Using Used Cars Dataset

STATISTICS-II PROJECT

Prepared by
Sandip Kumar Saha
(C21012)

Praxis Business School (PGPDS-21-JAN-KOL)

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to the Praxis Business School for providing me the opportunity to do an elementary level project of Statistics using Python and SPSS.First, I wish to express my special thanks to Pro **Sayantani Roy Choudhury**who has been very helpful throughout the project and for her enthusiasm,patience,helpful comments,practical suggestions and unceasing ideas that have helped me tremendously throughout the project.Her immense knowledge, profound experience and professional expertise in Data Quality Control has enabled me to complete this research successfully. Without her support and guidance, this project would not have been possible. I could not have imagined having a better supervisor in my study.

Finally, last but by no means least; also to everyone in the PGPDS-21-JAN batch. It was a great sharing experience with all of you during these three months.

Thanks for all the Encouragement!
Sandip K. Saha

Business Problem

Predicting the price of a car is an important and interesting problem at this juncture and this is also relevant for a developing country like India. In many developed countries, it is common to lease a car rather than buying it outright. A lease is a binding contract between a buyer and a seller(or a third party –usually a bank, insurance firm or other financial institutions) in which the buyer must pay fixed instalments for a predefined number of months/years to the seller/financer.After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e. its expected resale value. Thus, it is of commercial interest to seller/financers to be able to predict the salvage value (residual value) of cars with accuracy.Now two things can occur from the business perspective of used cars:

- If the residual value is under-estimated by the seller/financer at the beginning, the instalments will be higher for the clients who will certainly then opt for another seller/financer.
- If the residual value is overestimated, the instalments will be lower for the clients but then the seller/financer may have much difficulty at selling these high-priced used cars at this over-estimated residual value.

Thus, we can see that estimating the price of used cars is of very high commercial importance as well.Manufacturers' from Germany made a loss of 1 billion Euros in their USA market because of mis-calculating the residual value of leased cars.

The Client:To be able to predict used cars market value can help both buyers and sellers.

Used car sellers (dealers): They are one of the biggest target groups that can be interested in the results of this study. If used car sellers better understand what makes a car desirable, what the important features are for a used car, then they may consider this knowledge and offer a better service.

Online pricing services: There are websites that offer an estimated value of a car. They may have a good prediction model. However, having a second model may help them to give a better prediction to their users. Therefore, the model developed in this study may help online web services that tell a used car's market value.

Individuals: There are lots of individuals who are interested in the used car market at some point in their life because they wanted to sell their car or buy a used car. In this process, it's a big corner to pay too much or sell less than its market value.

Data Source:

https://www.dataminingbook.com/content/datasets-download-r-and-python-editions

Initial Data Description:

The file ToyotaCorolla.csv contains data on used cars(ToyotaCorolla) on sale during late summer of 2004 in the Netherlands.It has 1436 records containing details on 38 attributes.The goal is to predict the price of a used Toyota Corolla based on its specification.

Dataset Information:

The data description for all the features used are described below:

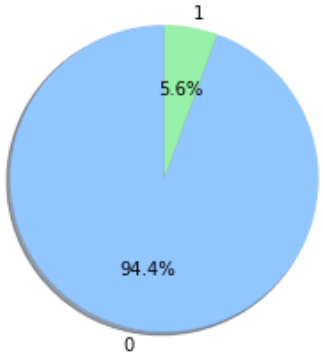
- ID:The ID number created while preparing the data
- Model: Particular Toyota model or variant of the same model
- Price: price of the manufactured car
- Age_08_04: age of the car as per August 2004
- Mfg_Month: Month of Manufacturing

- Mfg_Year: Year of Manufacturing
- KM: Number of kilometers the car has been driven
- Fuel_type: Whether the car is of Petrol or Diesel type
- HP: Horsepower is the metric used to indicate the power produced by a car's engine
- Color: colors offered by the Manufacturer
- Met_color: whether the car appears to be that of a polished metal
- Automatic: automatic car is an automobile with an automatic transmission that doesn't require a driver to shift gears manually
- CC: refers to cubic centimetres, or the metric measurement of engine capacity
- Doors: Number of doors in the car
- Cylinders: Number of cylinders in the car (A cylinder is a vital part of the engine. It's a chamber where fuel is combusted and power is generated)
- Gears: Number of gears in the car (As you accelerate, the drive gear will apply power to the wheels and progressively shift into higher 'gears' as the engine RPM reaches the desired level)
- Quaterly_Tax: Tax expenses for quarterly basis (euro)
- Weight: weight of the car (kg)
- Mfr_Guarantee: if the car has guarantee or not
- BOVAG_Guarantee: happens to be a kind of guarantee
- Guarantee_Period: years of guarantee
- ABS: ABS or an Anti-Lock Braking System is a piece of safety equipment that prevents the wheels of a vehicle from locking up under emergency
- Airbag_1: airbags in the front
- Airbag_2: airbags available in the backend
- Airco: if the car has air condition feature
- Automatic_airco: if the car has automatic air condition facility

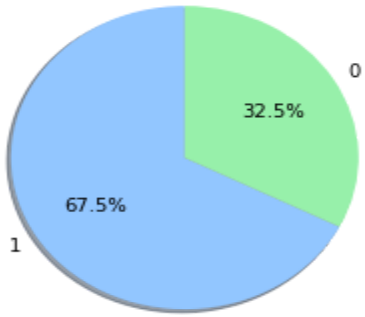
- BoardComputer: board computer continuously informs the driver by specific displays on the instrument panel about the current or average fuel consumption
- CD_player: if the car supports CD player option
- Central_lock: allow the driver or front passenger to simultaneously lock or unlock all the doors of an automobile
- Powered_windows: The windows on cars which can be opened or shut with the help of buttons
- Power_Steering: power-assisted steering is a system for steering that uses power from the engine so that it is easier for the driver to steer the vehicle
- Radio: it is the brains and command center for a car's audio system
- Mist Lamps: Fog lights are designed to be used at low speed in fog, heavy mist, snow and other poor-visibility situations
- Sport_Model: A sports car is a car designed with an emphasis on dynamic performance, such as handling, acceleration, top speed, or thrill of driving
- Backseat_Divider: a removable seat designed to hold a small child safely while riding in an automobile
- Metallic_Rim: The rim is the "outer edge of a wheel, holding the tire"
- Radio_cassette: an analog magnetic tape recording format for audio recording and playback
- Parking_Assistant: Park Assist is an automated parking aid that helps drivers park with greater precision
- Tow_Bar: a metal bar on the back of a vehicle that is used for towing something

Exploratory Data Analysis:

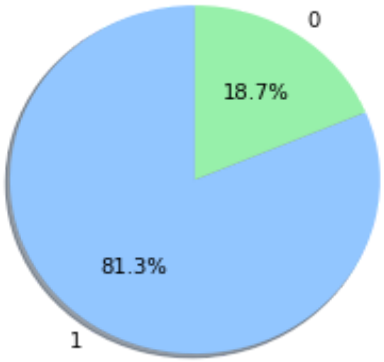
% of automobile with an automatic transmission



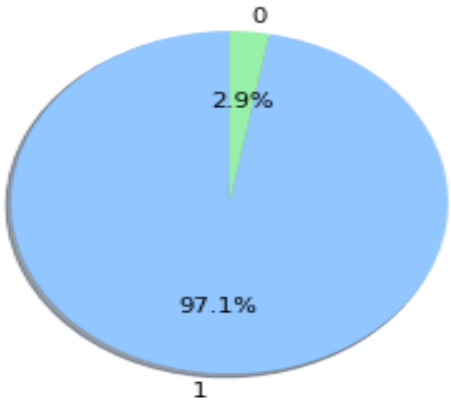
% of the car appears to be that of a polished metal



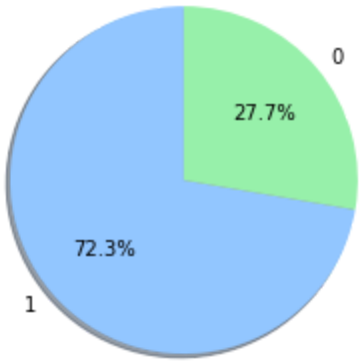
% of Anti-Lock Braking System



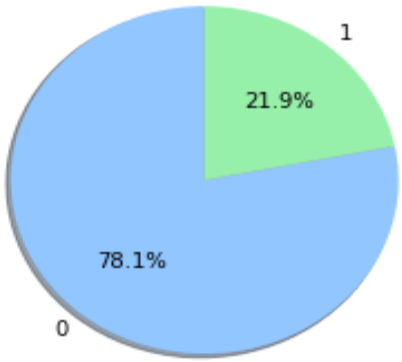
% of airbags in the front



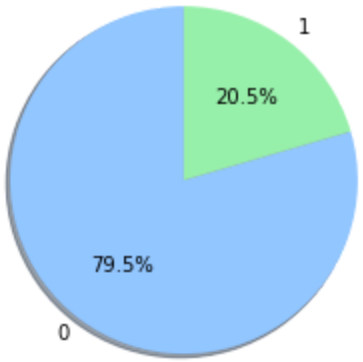
% of airbags in the backend



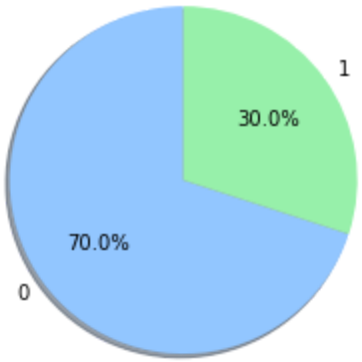
% of car supports CD player option



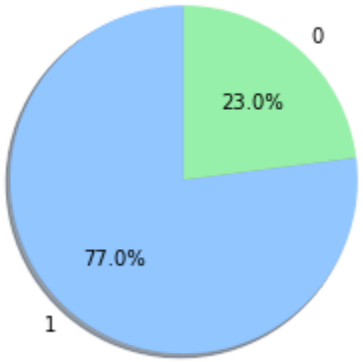
% of cars has Metallic Rim



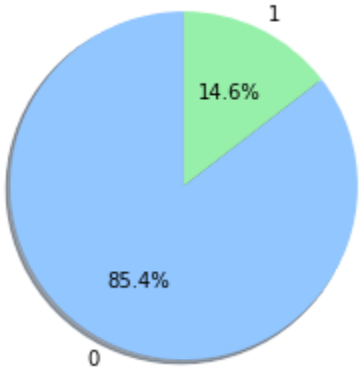
% of cars which has design like sports model

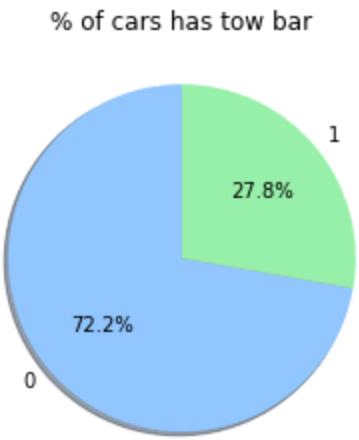


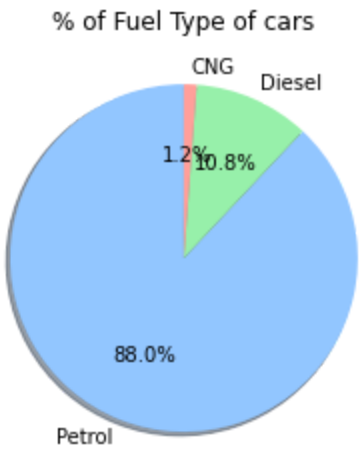
% of cars backseat divider

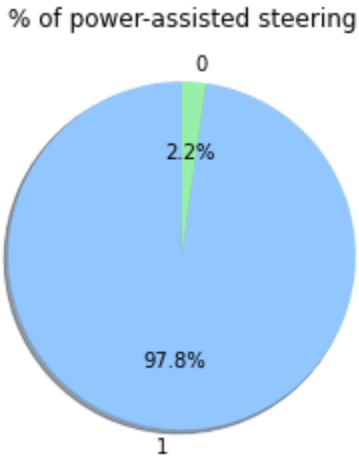


% of cars has Radio cassette



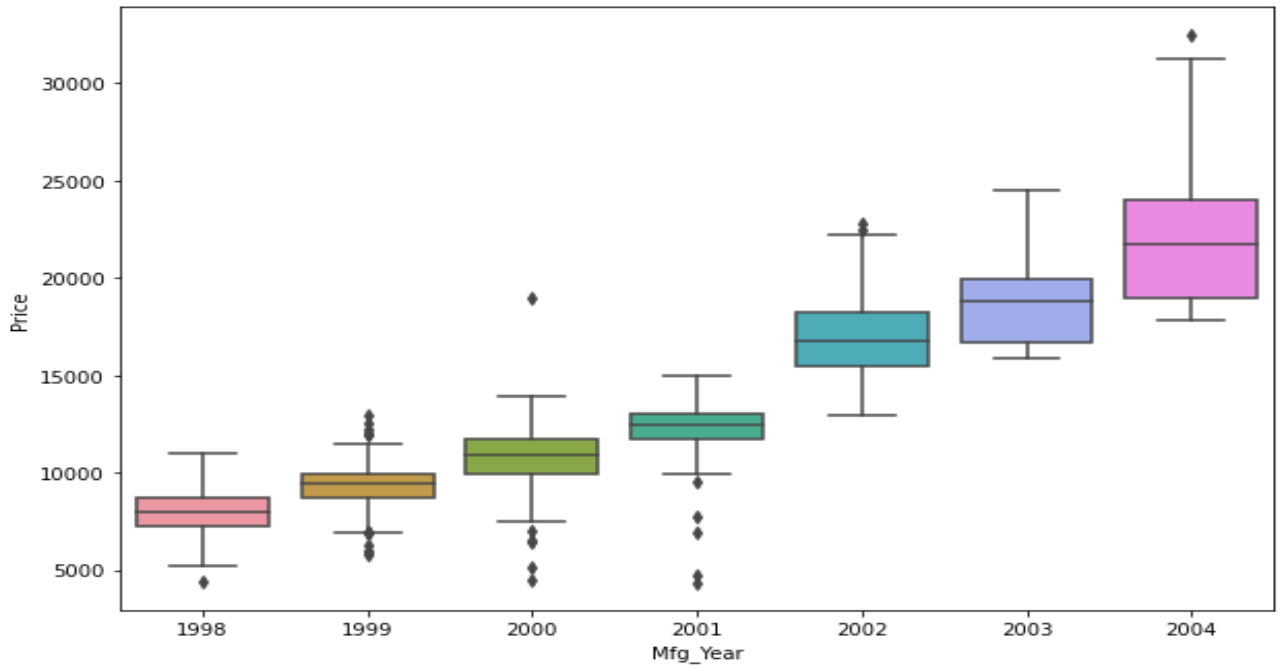




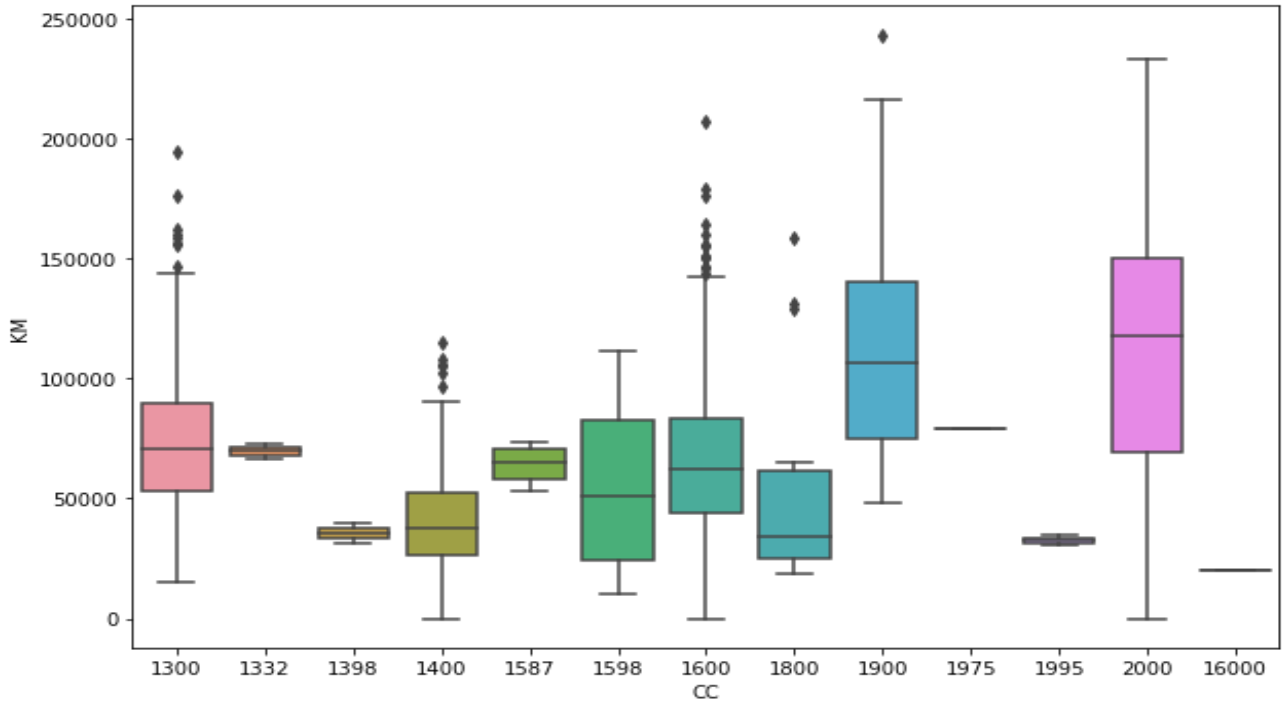


Relationship between Manufacturing Year and Price:

We can see that the mean price of these cars is increasing from the manufacturing year 1998 to 2004.The Highest average price of the car is $ 20000.

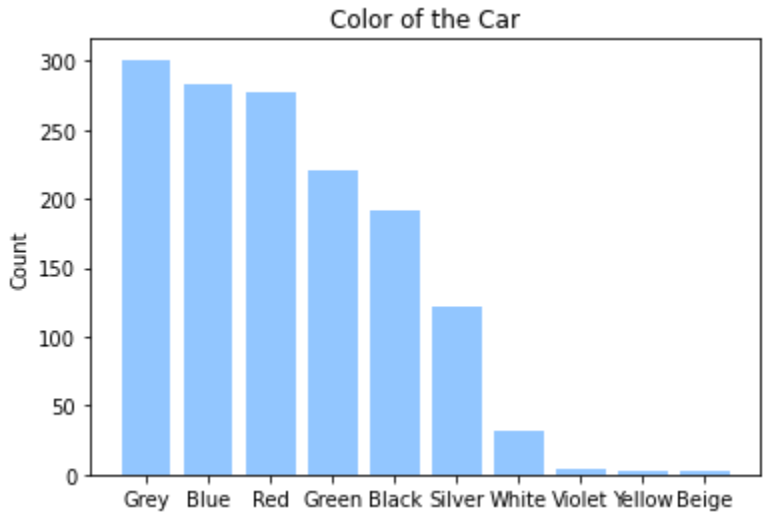


Relationship between the Engine CC and total distance travelled by the car:



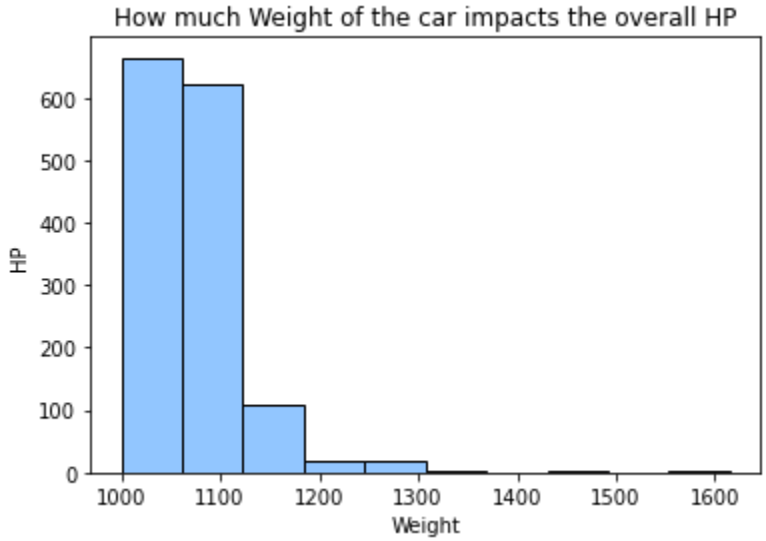
We can see that 2000 CC cars have an average travel time of more than 100000 KM which is the highest among all.

Color Distribution of Cars:



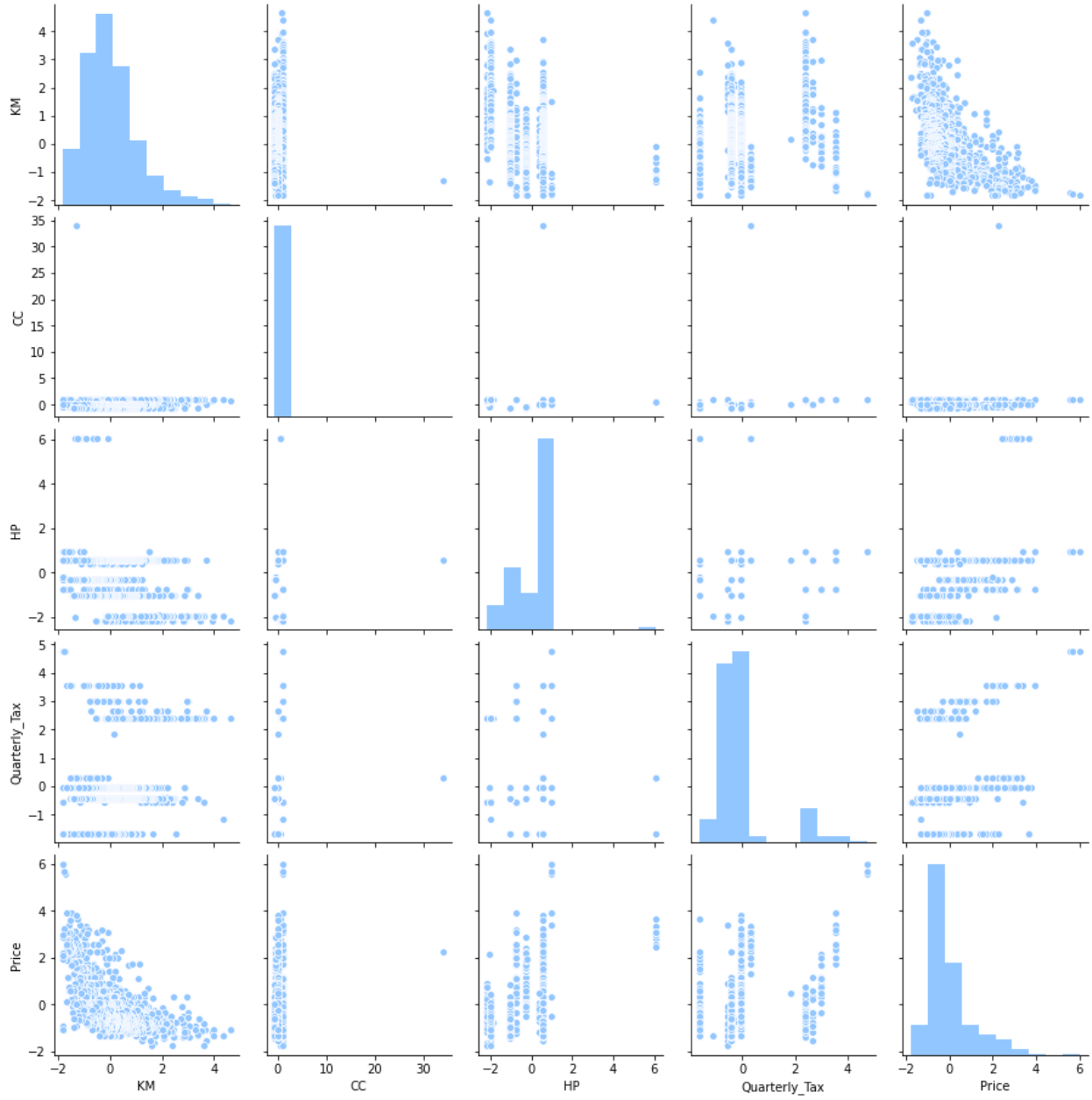
We conclude that Grey color cars have the highest priority among the users followed by Blue and Red etc.

Relationship between Weight and HP:



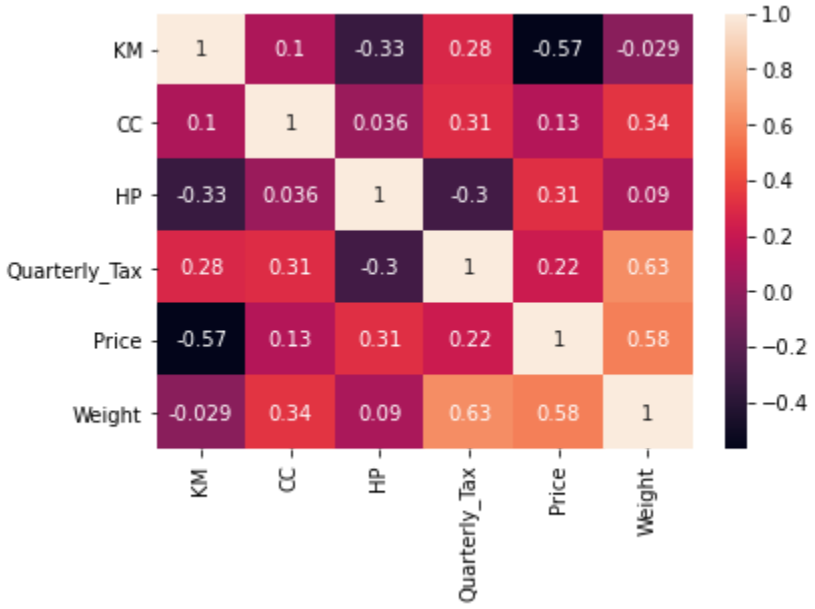
HP and Weight have an inverse relationship which is sort of obvious but the data also proves it.

Distribution of Histograms of some important features



We have drawn the histograms for the most important continuous features like-KM,CC,HP,Quarterly_Tax and Price.The data looks like it is normally distributed but the visual aid is not sufficient so i conducted Shapiro-Wilk test to check the Normality and it passed the Alternate Hypothesis of p<alpha(0.05)

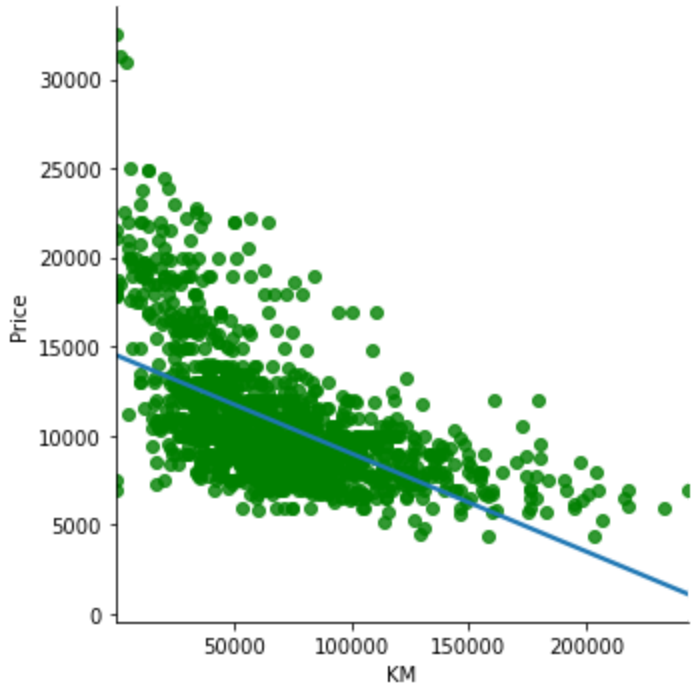
Correlation between some important features:



We can see that Quarterly Tax and Weight is highly correlated but there is no causation here.

Bivariate Analysis:

We see that the price of the cars are falling down as more distance is driven by the Car.



Output of Bivariate Analysis:

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.325
Model:	OLS	Adj. R-squared:	0.324
Method:	Least Squares	F-statistic:	690.0
Date:	Sat, 14 Aug 2021	Prob (F-statistic):	1.76e-124
Time:	11:32:43	Log-Likelihood:	-13525.
No. Observations:	1436	AIC:	2.705e+04
Df Residuals:	1434	BIC:	2.706e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.451e+04	163.915	88.510	0.000	1.42e+04	1.48e+04
KM	-0.0551	0.002	-26.268	0.000	-0.059	-0.051

Omnibus:	390.716	Durbin-Watson:	0.386
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1115.783
Skew:	1.388	Prob(JB):	5.14e-243
Kurtosis:	6.308	Cond. No.	1.63e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.63e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Check Multicollinearity effect before model building(VIF):

KM	2.065644
HP	5.841397
Automatic	1.084151
CC	8.508782
Doors	1.303334
Quarterly_Tax	4.038076
Weight	3.521738
Mfr_Guarantee	1.160153
BOVAG_Guarantee	1.325926
Guarantee_Period	1.305012
Airco	1.670606
Automatic_airco	1.533340
Boardcomputer	2.395371
CD_Player	1.479365
Powered_Windows	1.551147
Sport_Model	1.217501
Metallic_Rim	1.193388
Fuel_Type_1	17.546820
Fuel_Type_CNG	2.294240

Variation Inflation factors which are greater than 5 are discarded (Thumb rule). VIF is a measure of Multicollinearity.The variance inflation factor is the ratio of the variance of estimating some parameter in a model that includes multiple other terms by the variance of a model constructed using only one term. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis.

Final OLS Result for the Train data:

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.811
Model:	OLS	Adj. R-squared:	0.808
Method:	Least Squares	F-statistic:	264.5
Date:	Thu, 12 Aug 2021	Prob (F-statistic):	0.00
Time:	15:55:23	Log-Likelihood:	-597.01
No. Observations:	1005	AIC:	1228.
Df Residuals:	988	BIC:	1312.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0072	0.014	-0.516	0.606	-0.035	0.020
KM	-0.3810	0.017	-22.474	0.000	-0.414	-0.348
Automatic	-0.0049	0.014	-0.353	0.724	-0.032	0.022
Doors	0.0351	0.015	2.279	0.023	0.005	0.065
Quarterly_Tax	0.1094	0.021	5.229	0.000	0.068	0.150
Weight	0.2553	0.022	11.766	0.000	0.213	0.298
Mfr_Guarantee	0.0351	0.015	2.353	0.019	0.006	0.064
BOVAG_Guarantee	0.0341	0.016	2.176	0.030	0.003	0.065
Guarantee_Period	0.1561	0.015	10.139	0.000	0.126	0.186
Airco	0.0726	0.018	4.134	0.000	0.038	0.107
Automatic_airco	0.2447	0.016	14.952	0.000	0.213	0.277
Boardcomputer	0.2161	0.017	12.399	0.000	0.182	0.250
CD_Player	0.0993	0.016	6.046	0.000	0.067	0.132
Powered_Windows	0.0622	0.017	3.608	0.000	0.028	0.096
Sport_Model	0.0664	0.015	4.419	0.000	0.037	0.096
Metallic_Rim	0.0370	0.015	2.463	0.014	0.008	0.067
Fuel_Type_CNG	-0.0327	0.016	-2.075	0.038	-0.064	-0.002

Omnibus:	92.134	Durbin-Watson:	2.016
Prob(Omnibus):	0.000	Jarque-Bera (JB):	441.909
Skew:	0.262	Prob(JB):	1.10e-96
Kurtosis:	6.206	Cond. No.	3.53

Final OLS Result on the Test Data:

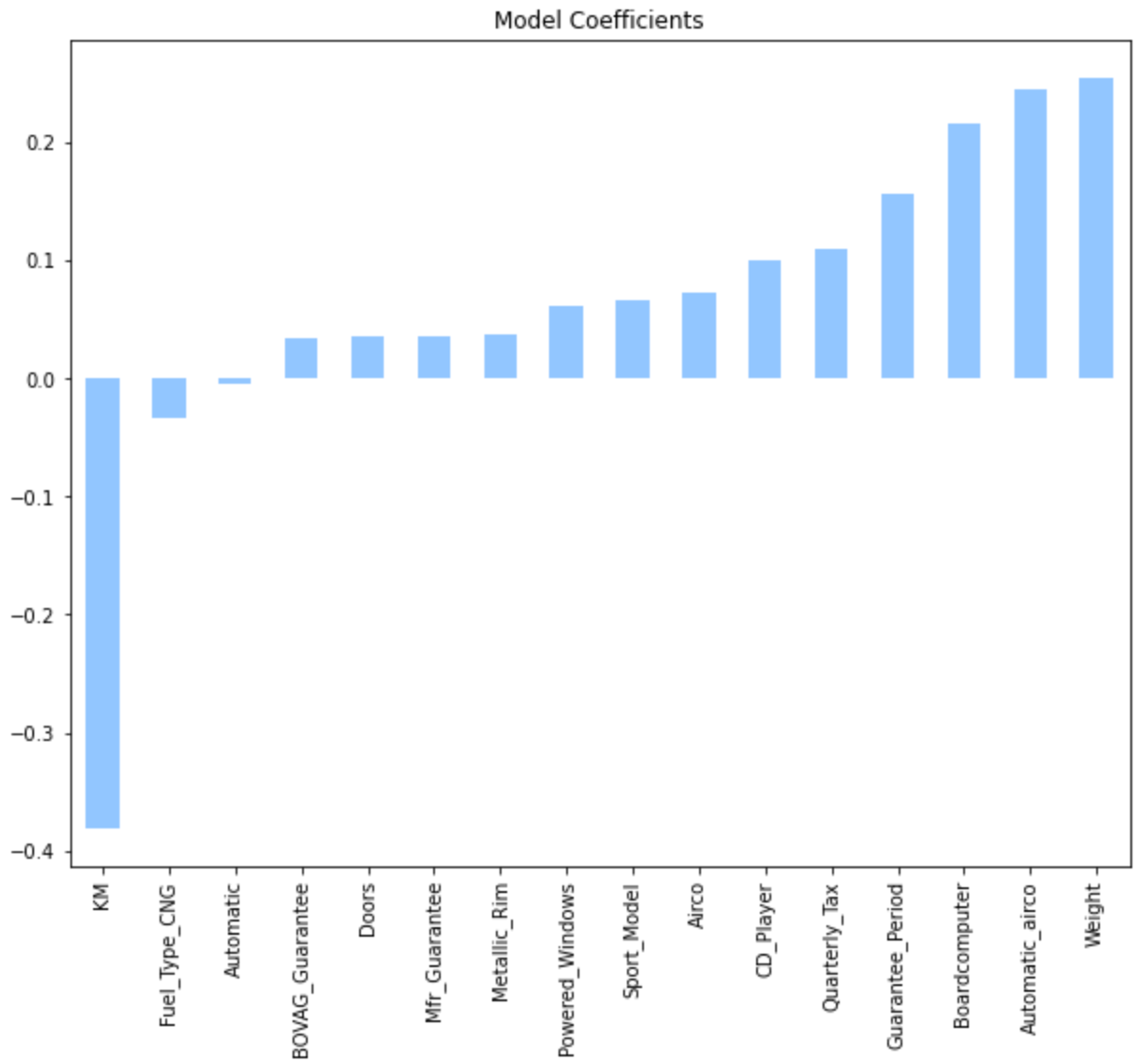
OLS Regression Results

Dep. Variable:	Price	R-squared:	0.841
Model:	OLS	Adj. R-squared:	0.835
Method:	Least Squares	F-statistic:	137.3
Date:	Thu, 12 Aug 2021	Prob (F-statistic):	2.92e-154
Time:	15:56:02	Log-Likelihood:	-206.52
No. Observations:	431	AIC:	447.0
Df Residuals:	414	BIC:	516.2
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0173	0.019	0.894	0.372	-0.021	0.055
KM	-0.3331	0.024	-13.972	0.000	-0.380	-0.286
Automatic	-0.0128	0.022	-0.581	0.561	-0.056	0.031
Doors	-0.0324	0.021	-1.571	0.117	-0.073	0.008
Quarterly_Tax	-0.0020	0.031	-0.065	0.948	-0.062	0.058
Weight	0.4003	0.038	10.527	0.000	0.326	0.475
Mfr_Guarantee	0.0660	0.021	3.200	0.001	0.025	0.107
BOVAG_Guarantee	0.0298	0.021	1.416	0.158	-0.012	0.071
Guarantee_Period	0.1509	0.021	7.347	0.000	0.110	0.191
Airco	0.1206	0.025	4.845	0.000	0.072	0.169
Automatic_airco	0.1739	0.025	6.963	0.000	0.125	0.223
Boardcomputer	0.2658	0.026	10.421	0.000	0.216	0.316
CD_Player	0.0538	0.023	2.316	0.021	0.008	0.099
Powered_Windows	0.0470	0.024	1.942	0.053	-0.001	0.094
Sport_Model	0.0463	0.021	2.243	0.025	0.006	0.087
Metallic_Rim	0.0236	0.021	1.121	0.263	-0.018	0.065
Fuel_Type_CNG	0.0158	0.018	0.881	0.379	-0.019	0.051

Omnibus:	24.180	Durbin-Watson:	2.132
Prob(Omnibus):	0.000	Jarque-Bera (JB):	41.294
Skew:	0.376	Prob(JB):	1.08e-09
Kurtosis:	4.317	Cond. No.	4.21

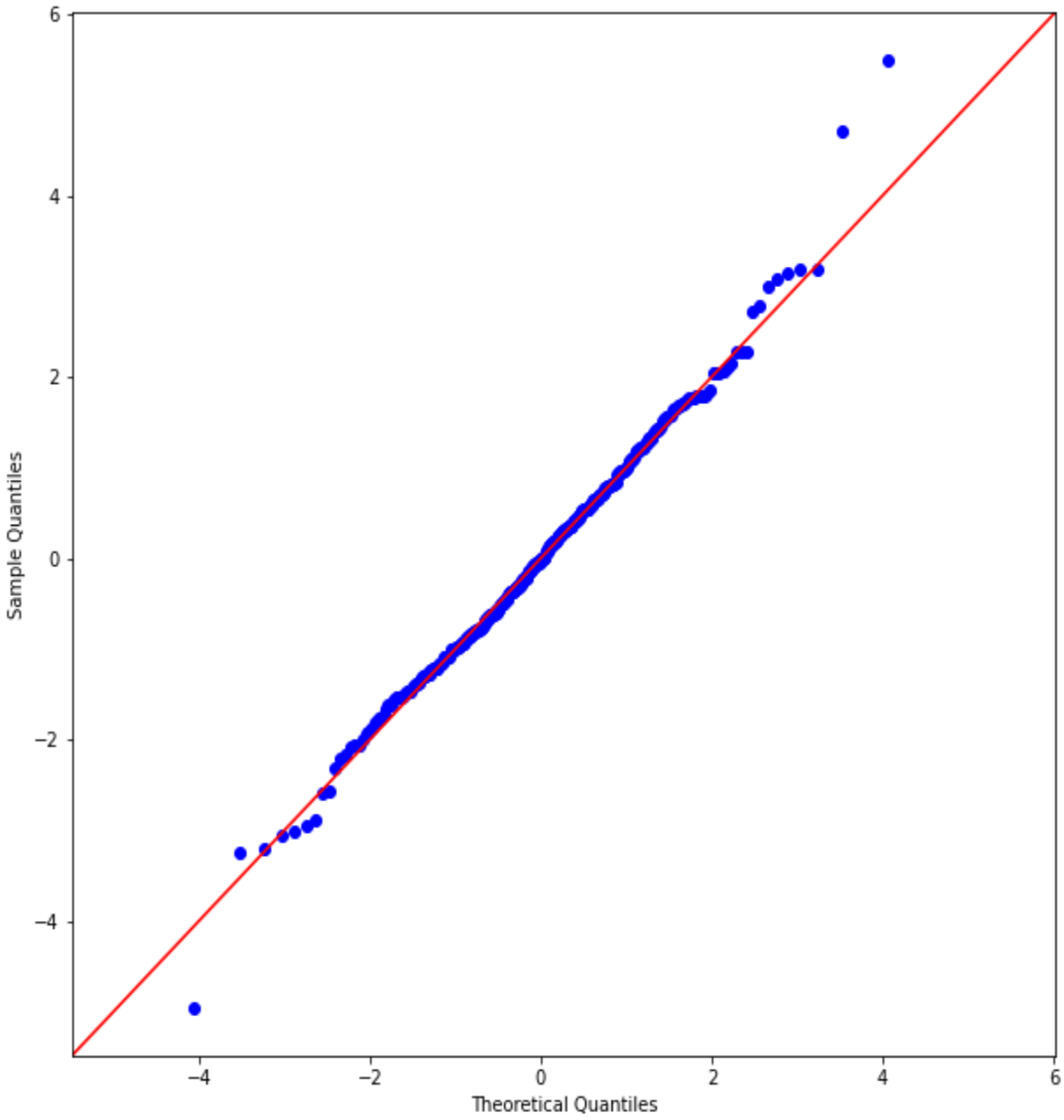
Importance of Model features are shown in a graph:



Model Analysis:

- variables showing +ve effect on regression model are Automatic_airco,Automatic,BOVAG_Guarantee,Powered_Windows,Sport_Model,Metallic_Rim, CD_Player,Airco,Mfg_Guarantee,Doors,Guarantee_Period,HP
- Higher the value of beta coefficient higher is the impact
- car features like the above mentioned plays an important role in the sales of used car
- car price has a -ve effect on Fuel_Type,Mfg_month etc keeping all other variables constant
- Some variables which hardly affect model prediction for car price determination are CC,KM,Weight,Quaterly_Tax,Mfg_Year etc.

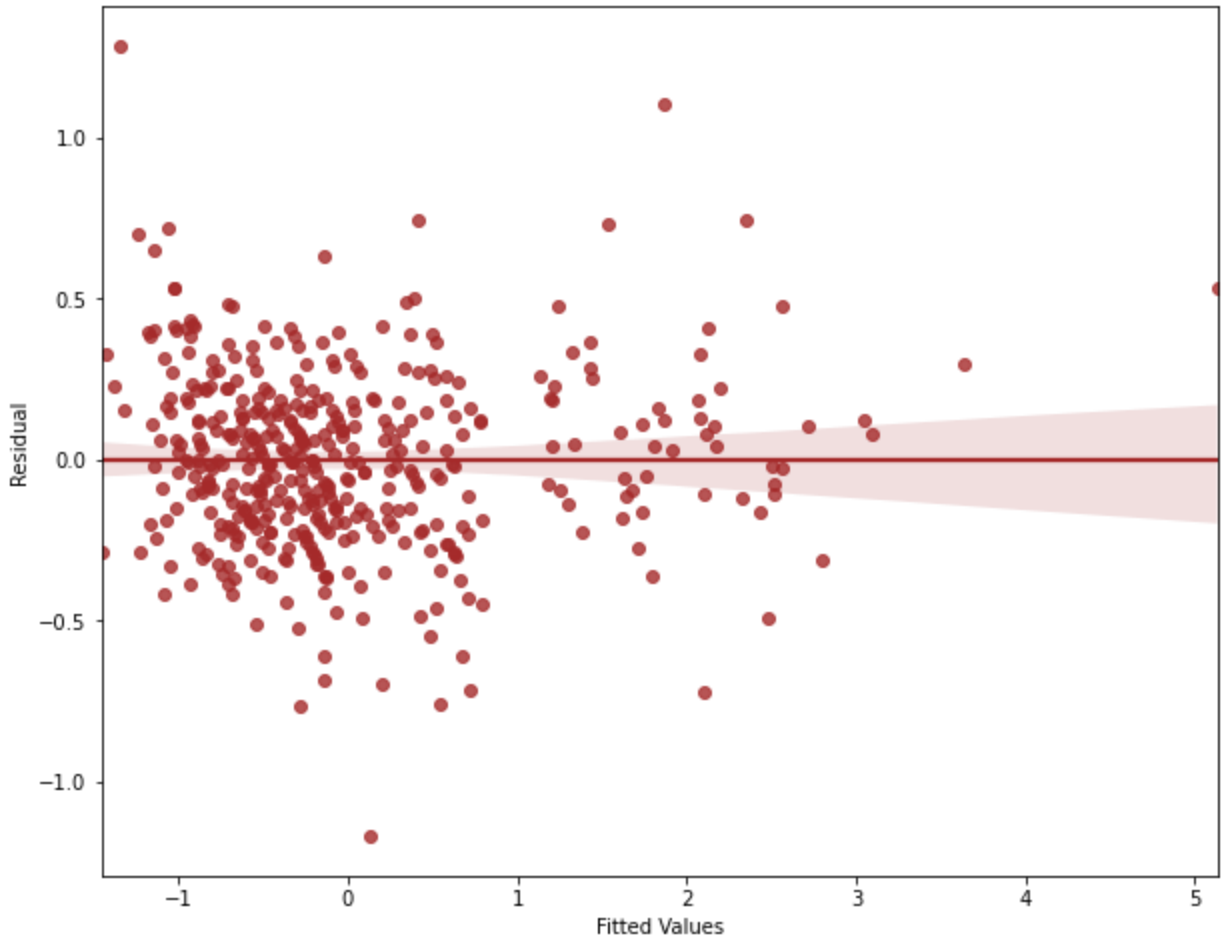
Checking Normality of Residuals:



Although for OLS estimation we do not need to assume that the residuals follow Normal Distribution,it is an important assumption for the Hypothesis Test.The Q-Q plot compares the cumulative distribution function of two probability distributions against each other.

- The diagonal line is the cumulative distribution of a normal distribution, whereas the dots represent the cumulative distribution of residuals.
- Since the dots are close to the diagonal line we can conclude that the residuals follow an approximate Normal distribution.

To check the Homoscedasticity of Residuals:

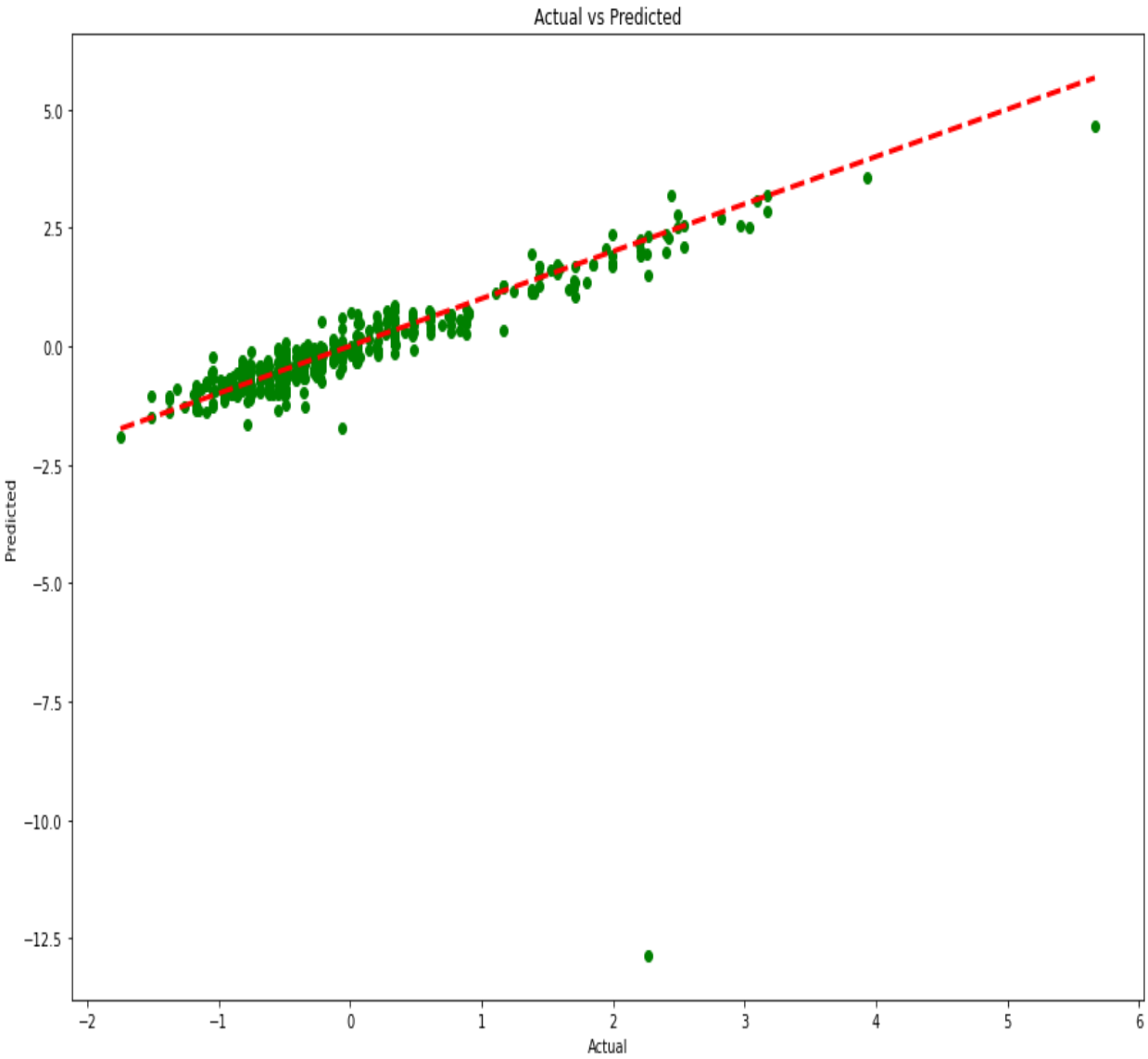


An important assumption of the regression model is that the residuals have constant variance across different values of the explanatory variable(X). I.e. The variance of residuals is assumed to be independent of variable X. Failure to meet this assumption will result in unreliability of the hypothesis tests.

- The plot above is between standardized predicted values versus the standardized residuals.
- If there is heteroscedasticity (non-constant variance of residuals) then we can expect a funnel type shape in the residual plot which is not the case.
- A funnel shape indicates that the variance of residuals depends on the value of independent variable X.

Scatter plot between Actual vs Predicted Price:

Below Graph shows the best fit line between the predicted value and actual value using the OLS regression.It looks like there are very few outliers but for the time we can ignore this.



References:

- **https://towardsdatascience.com/predicting-used-car-prices-with-machine-learning-techniques-8a9d8313952**
- **https://www.mordorintelligence.com/industry-reports/india-used-car-market**