Analysis and prediction of prices of used cars

Submitted as a requirement of the course
DS203 Programming for Data Science
Centre for Machine Intelligence and Data Sciences
Indian Institute of Technology Bombay 2022

Mayank Gupta
210101002
Electrical Engineering
Indian Institute of Technology Bombay

Harsh Amit Shah
210100063
Electrical Engineering
Indian Institute of Technology Bombay

Mokashi Harish Mahesh
210070053
Electrical Engineering
Indian Institute of Technology Bombay

***Abstract*—People face difficulties in deciding which used car should they purchase. It is difficult to analyse the long-term costs of the vehicle. Different cars are suitable for different people and the perfect choice for a particular region can be made by viewing which car performs better in a particular region. Our work is based on Exploratory Data Analysis and decision tree regressor. The analysis helps people in choosing cars which are appropriate for their own use, economical and sustainable for the environment.**

Keywords—Price prediction, Fuel analysis, Company Popularity, Location

I. INTRODUCTION

With the increase in the difficulty in economic conditions, the sales of second and third-hand cars have dramatically increased, leading to an increased necessity for the analysis of various factors responsible for determining the perfect car to purchase. In the financial year 2021-22, about 4.4 million used vehicles were sold in India. Consumers thus require a reliable source of analysis of the statistics which is available. The use of Data Science and machine learning is comparatively limited in this field. There are many cars available with each one having a different age, mileage, engine power, kilometres run and most importantly price.

We attempt to analyse the data to help consumers understand different factors which should be considered while purchasing a used car. A consumer might suppose that a cheap car would be more economical, but it is very crucial for one to understand the subtle fact that the car might cost a lot in the future as that car might have a lower mileage, might require more expenditure on spare parts depending on the age, etc. We have used factors such as engine power and type of fuel to determine the type of car required for a particular region. The environmental damage caused by a particular car has also been considered to help the consumer understand the difference in environmental harm caused by different cars. People often only consider the outer beauty of vehicles like the interior style, maximum speed of vehicle, paint colour, entertainment system etc, but it is equally important to consider other factors which can significantly influence other important factors like durability and safety. Information about all such factors is mostly unavailable to the buyers, which often leads them into making incorrect decisions by not taking into consideration all the factors.

We have primarily developed the following things:

- Analysed which is the best car company based on all the different factors by extracting the company from the car name.
- We have analysed all different situations that the buyer might be in like location, and total kilometres the user aims to drive the vehicle over its duration to find the perfect car for the user.
- We have considered pollution factors in different locations, depending on the type of fuel so that the user can buy the most environment-friendly car.
- Analysed the cost of cars using the cost of fuels.
- We have used Standard Gradient Descent, Linear Regressor, Bayesian Ridge, Lasso Lars, ARD Regression, Passive Aggressive Regressor, Theil Sen Regressor etc. to determine the perfect optimiser for predicting the prices. Principle component analysis (PCA) was also performed for a better understanding of the data.

In the next few sections, we have explained the dataset along with its important highlights, related work, the environment, and modules used for developing the analysis along with the Machine learning and neural network environment developed for determining the price of the used cars.

II. RELATED WORK

Surprisingly, work on estimating the price of used cars is very recent but also very sparse. Analysis of used cars gives us a lot of insight into what kind of cars do people prefer whether it is in diesel, petrol or manual, automatic. Some of work on prediction of used cars is listed below:

Listiani showed that the regression mode build using support vector machines (SVM) can estimate the residual price of leased cars with higher accuracy than simple multiple regression or multivariate regression. SVM is Predicting the Price of Used Cars using Machine Learning Techniques better able to deal with very high dimensional data (number of features used to predict the price) and can avoid both over-fitting and underfitting.
(https://www.ifis.uni-luebeck.de/~moeller/publist-sts-pw-and-m/source/papers/2009/list09.pdf)
Gonggi proposed a new model based on artificial neural networks to forecast the residual value of private used cars. The main features used in this study were: mileage, manufacturer and estimate useful life. The model was optimised to handle nonlinear relationships which cannot be done with simple linear regression methods. It was found that this model was reasonably accurate in predicting the residual value of used cars.(https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5721273&tag=1)

III. DATASETS

The dataset we have used is from the Kaggle, which is an online community platform for data scientists and machine learning enthusiasts. It is a great platform which allows users to collaborate with other users, find and publish datasets, compete and cooperate with other data scientists to solve data science challenges.

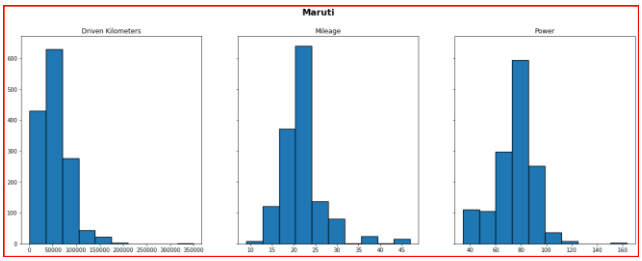
The dataset briefly covers information about cars further branched throughout various independent and interrelated variables such as Brand, Location, Year, Distance covered by car, Fuel type, Transmission, owner type, Mileage, Engine, Power, Seats, new prices, prices when respective car is sold. Majority of the columns we have used take the integer/ floating point values. The columns might be categorical as well for example in the used dataset columns like *Fuel type*, *Transmission*, *Year* list the discrete values out of which some are encoded as values 0,1,2,3 for smooth and efficient predictive analysis.

To make data analysis easier, columns *Engine*, *Power* have been converted to float after certain pre-processing cleaning operations. After similar cleaning operation on *Mileage* variable, all values in column are converted into similar units. Further interesting predictable variables are also added which include Company Name, Pollution factor for clear and better intuitive use of data analysis.

IV. DATA ANALYSIS

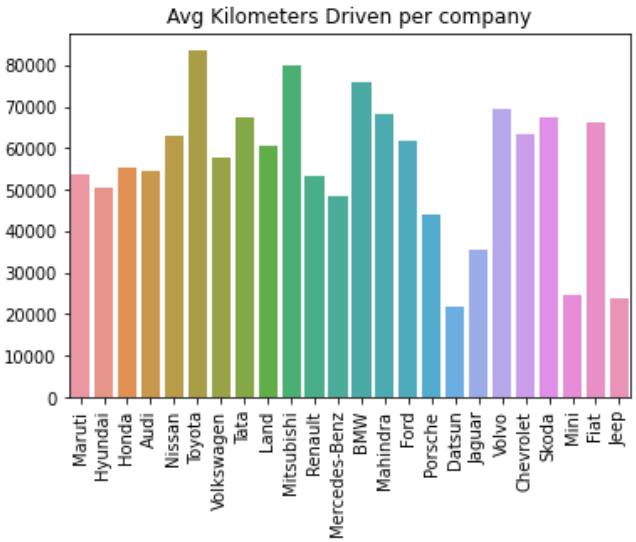
Visual analysis of the data helped in finding columns which are redundant or not very useful for analysis. Few rows which should have only been numeric had their units assigned in the same field which was removed and then converted to numeric data type. The null values were also handled. For obtaining the dataset for the model the new price column was dropped as it had many null values, and that column isn't very helpful in predicting the value of the car's price. The data types, number of nulls in each column, number of unique companies etc. were obtained for a better understanding of the data. The price of the cars was listed in different units and all that was converted to a uniform unit for better processing of the data.

The concentration of the number of cars for each of the properties like mileage, engine power, kilometres driven, age of the car etc. was expressed using bar graphs. The number of cars for each company in the dataset was plotted using a bar graph, and the companies which have a smaller number of cars were dropped as the less number indicates that those companies are now outdated. Analysis of the number of cars for each company for the three properties namely Kilometres driven, Mileage and Power was performed by plotting histograms for each company. Below is an example of this analysis for Maruti Suzuki:

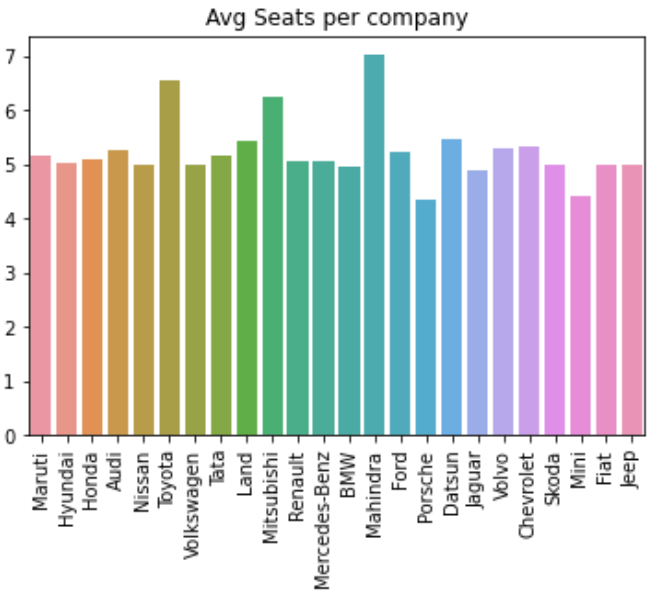


This graph shows a lower mileage, power, and kilometre for Maruti cars. This seems appropriate as Maruti cars being comparatively cheaper will have lower power and mileage and in turn, lower kilometres driven.

The number of average kilometres, mileage, engine power, seats etc. for cars of each company were expressed using a bar graph.

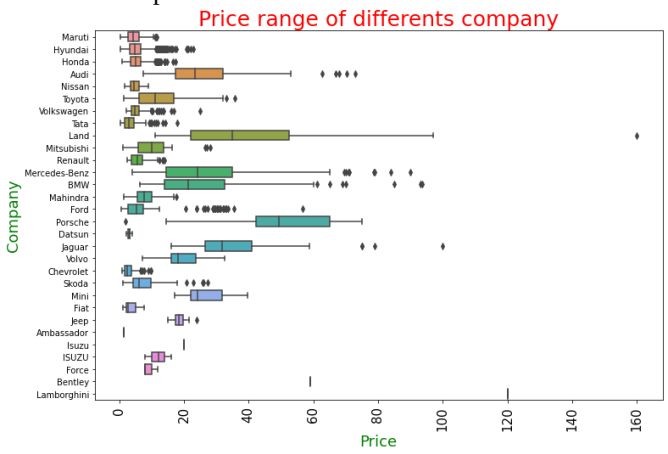


We can infer from the graph that Toyota cars are the most driven cars, which is in alignment with the real-life observations too as we see many Toyota cars running as taxis which cover a lot of distance. Toyota is followed by Mitsubishi which is a racing car company.



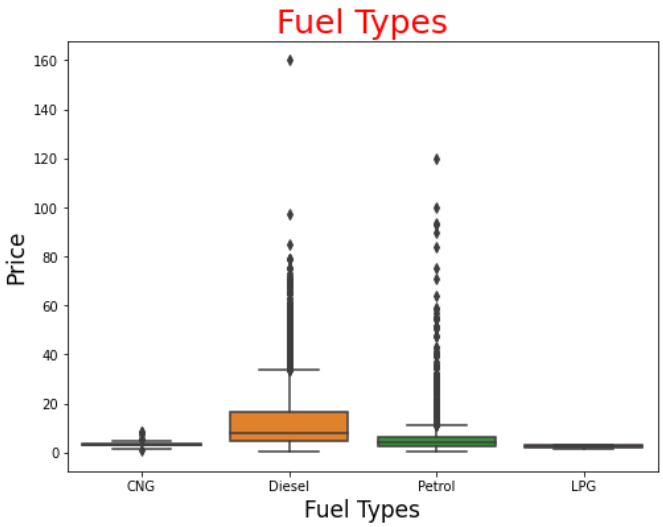
The above graph expresses the average number of seats for each company. Mahindra being a primary SUV manufacturer tops the graph, followed by Toyota.

People might also need the range of prices that a particular car company has to offer. This is expressed using a combined boxplot as follows:



Land rover has the largest price range, followed by Porsche and Audi. The cheaper car companies like Maruti and Hyundai have a smaller range.

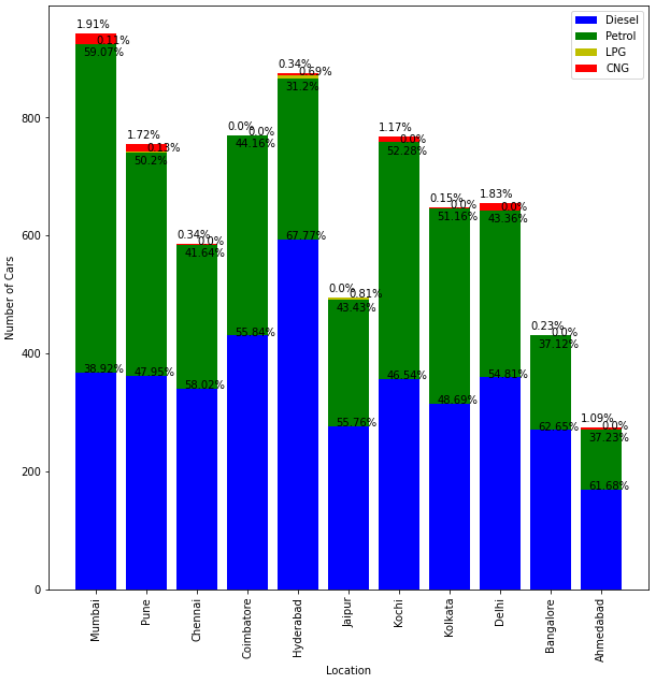
It is crucial to also determine the preferred fuel type used by people. This is expressed using a box plot.



Diesel is the most commonly used fuel followed by Petrol. LPG has a negligible number of users which can be observed in real life too.

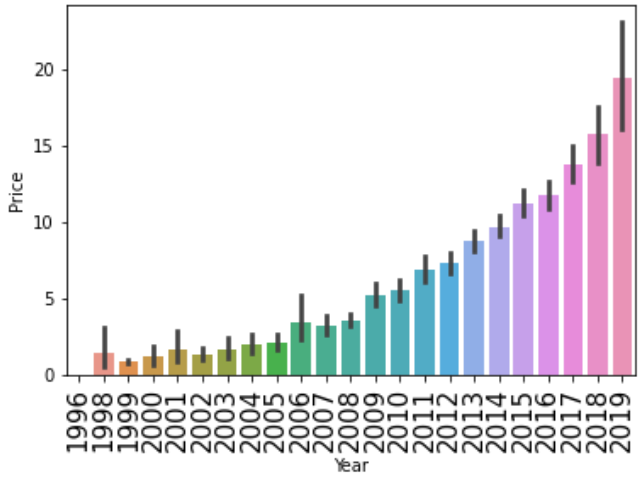
The availability of cars in nearby regions is also an important factor which is considered by buyers. The statistics for this were expressed using a bar plot.
The price of cars in different cities and regions like North, South, East and West were also plotted, indicating that metro cities have higher prices.

The distribution of the number of Petrol, Diesel, LPG and CNG cars for each location was expressed in a stacked bar graph. This was expressed as:



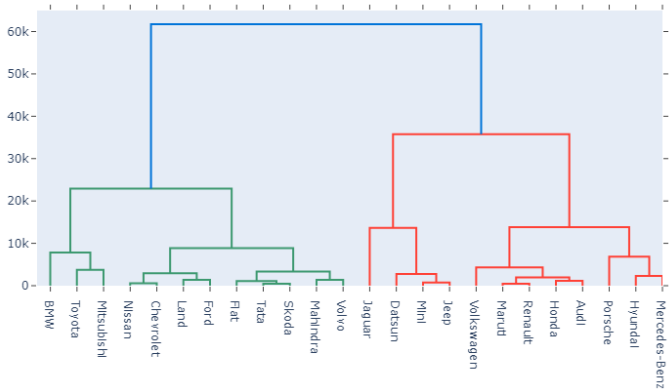
From the graph, it can be seen that in all cities the number of cars which are resold are usually petrol and diesel. CNG and LPG cars don't mostly enter the resale market, as it can be seen from the above graph.

The age of the car a person is purchasing is an important factor which helps us to determine the quality of the car. This is expressed using a bar plot as:



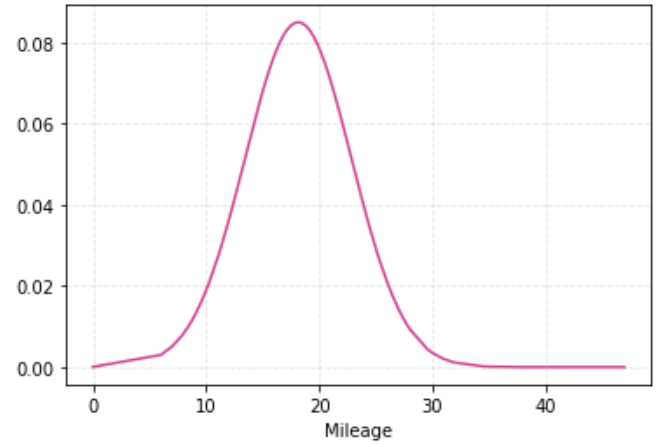
Evidently, it can be seen that the number of cars in the newer years is more compared to the older years, as people usually sell their cars after a few years of usage. People don't keep a car with them for a long period of time.

A dendrogram was made to express the correlation between any two companies. The dendrogram is very useful for people to find substitutes for a car. This was expressed as:



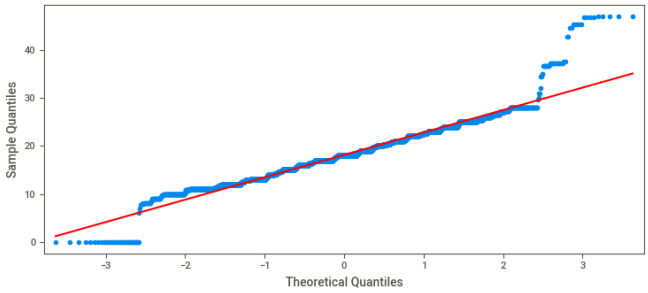
It can be seen that companies like land rover and ford offer similar cars. Mitsubishi and Toyota are similar as we have deduced earlier that both have cars which have a high kilometre drive, similar number of average number of seats etc.

The variables which have a distribution similar to a gaussian distribution were determined using their PDF as:



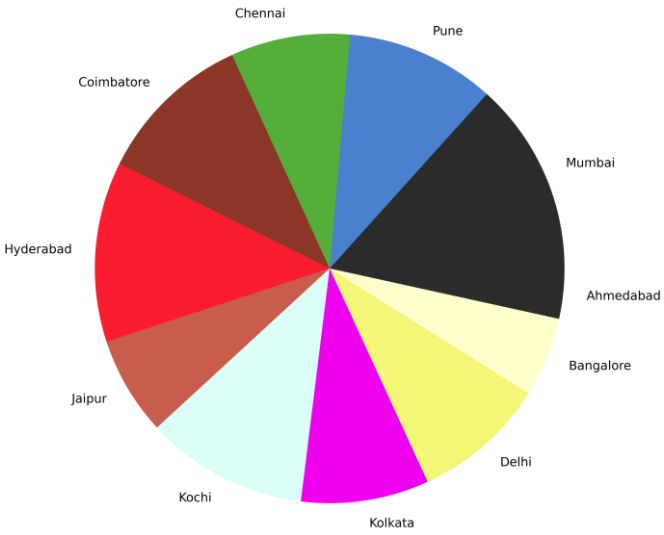
Four variables were found as Gaussian by the above method. These are mileage, power, price, and engine.

The quantile-quantile plot for each gaussian distribution were made. The qq plot for mileage is:

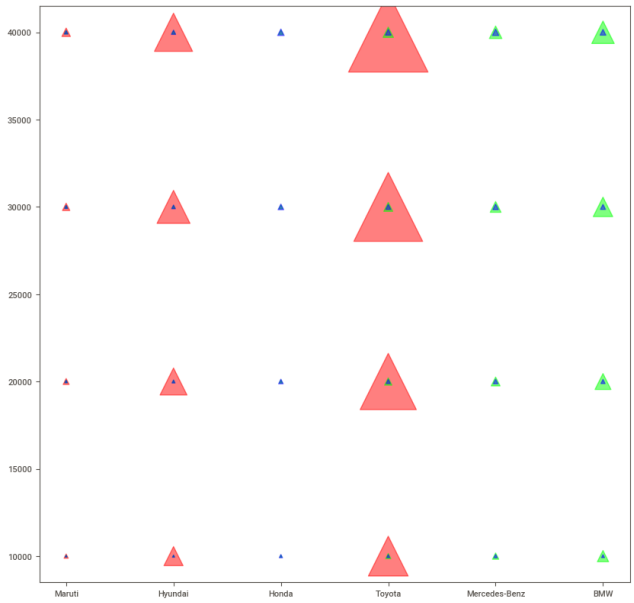


Since this overlaps with the red line, this is a Gaussian distribution.

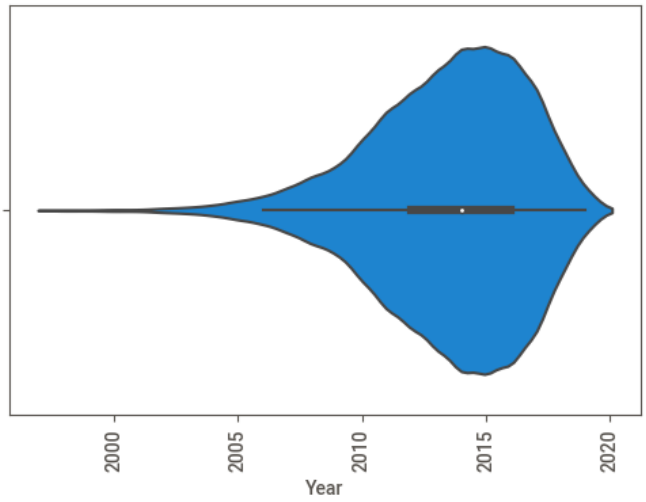
The average price of the cars for each location was expressed using a pie chart to show the costliest location to purchase used cars.



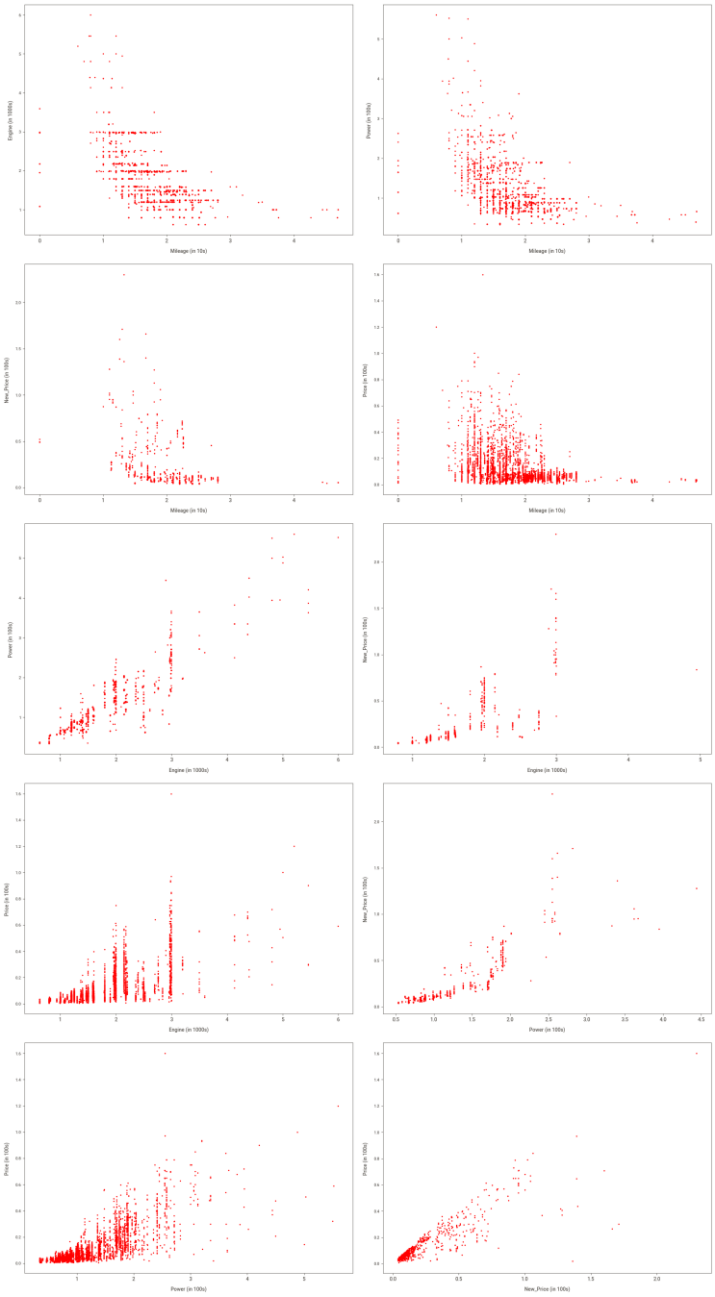
The effective cost of using a car of a specific company was expressed using a scatter plot. This is expressed as:



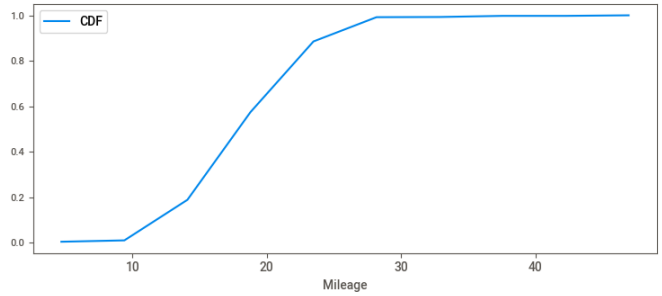
Violin plots were made to express the spread of each numeric variable. This can be seen for the year as:



Scatter plots were used to show the distribution of continuous variables.

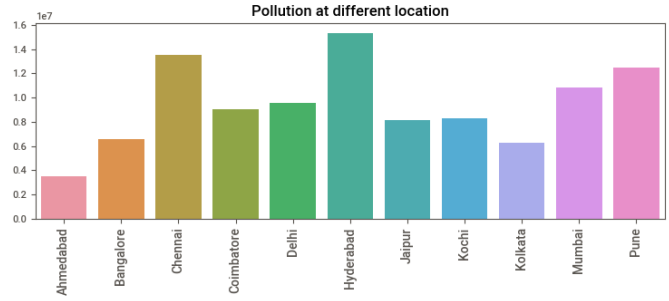


The CDF and PDF for all continuous variables was plotted for better understanding of the data.



Sweetviz is used to perform a complete exhaustive analysis of data, giving output in an html file. This can be seen as:

Major component of pollution is done by cars and different fuels used by them. In our data set we have been provided with number of kilometres driven, mileage and type of gas used by the car. We try to analyse the amount of pollution done by the cars in different regions using a term Pollution Factor. Pollution Factor is calculated as the total amount of CO_2 produced in kilogram by a car. As we know how much litre of gas is used by the car by kilometres driven and mileage, amount of CO_2 produced by one litre of the specific gas used by the car. This gives us an insight of how much pollution is happening in different locations. In later plots we have tried to average the Pollution Factor by number of years in use and number of cars in different locations



V. Results & Inference

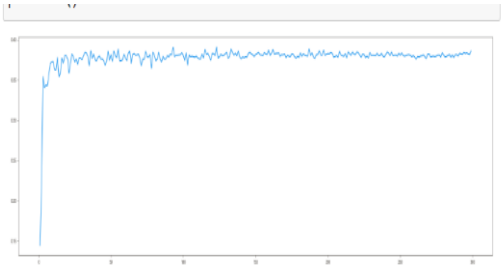
Predictive analysis typically involves selection criteria based on type of target variable, performing model training using most suitable training methods, concluding best model depending upon parameters RMSE error, R2 score.

A. Model training and selection –

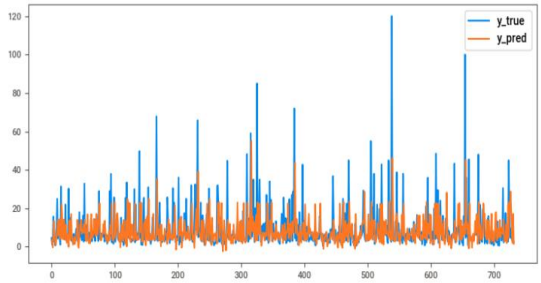
- First of all its very important to choose which factors contribute to most in determining prices of used cars and so they will be most contributing in prediction of target variable which is price. So such variables are first segregated using VIF(Variance Inflation Factor) from the model training dataset **datam_withoutprice** . Variables with VIF>10 are removed.
- Decision tree regression is done in different formats of training dataset-'Original dataset', 'Normalized dataset with one bit hot encoding of categorical variables', 'Decision tree with PCA', 'D-tree regression with hyper parameter tuning'.

	Model_Name	RMSE	R-Squared
0	Dtree_notOneHot_NotNormalized_noValidation	7.967193	0.267873
1	Dtree_OneHot_Normalized_noValidation	0.077796	0.270839
2	Dtree_WithPCA	0.065348	0.447389
3	DTree_withHyperParTuning	0.064103	0.475238

- The above chart shows RMSE error and R2 score of trained models. Its clear that the model with Hyperparameter tuning can be considered as best model in decision tree regression method as due to minimum RMSE and maximum R2 score.
- Model parameters –
  RMSE = 0.064103
  R2 SCORE = 0.475238
- Further Random Forest Regression is performed with hyperparameter tuning in 300 iterations which give best R2 score of 0.38 which is not good compared to that of previous. Its graph is shown below.



- Next training involves use of deep learning network with keras-optimizer technique of 'Adam' optimizer, Relu activation and batch size of 32 in just 5 epochs. The values of predicted values match with that of true values of prices very well as illustrated from below
  graph



B. Inferences of predictive analysis

Predicting prices of used cars in India from given dataset complex as it is bit tough to think and narrow down to more factors from available dataset resulting in RFR trained model with less R2 score. But below are some of the meaningful factors that have attributed D-tree trained model with much better 0.48 R2 score-

- Kilometres driven significantly contributes to prices & rightly so the model indicates a positive correlation i.e. cars with a higher distance travelled would be predicted to have higher duration with better working condition and so the higher prices

- Another significant factor which has contributed to an increase in prices is fuel type since consumption of particular type of fuel may cause reduction in no. of its resources which increases average price of car(current price + increased price of fuel) & its contribution to prices is well justified considering the assumption that people haven't adopted to new electric type vehicles, new fuel types.
- Mileage, Engine and power also has a significant weight in the determination of prices. Very rightly an increase in any of the mentioned parameters will lead to an increase in value of other parameter and also ultimately to prices of used cars.

The other categorical variables such as owner type and no. of seats also contribute in certain proportion in evaluation of prices.

ACKNOWLEDGMENT

We would like to thank our professors, Prof. Amit Sethi,Prof. Manjesh K. Hanawal, Prof. Sunita Sarawagi and Prof. S.Sudarshan for teaching us this course, without which we would not have been able to complete this project. We would also like to thank all the TA's involved with this course, who helped us a lot in doing the various assignments, and completing this course.

REFERENCES

[1] Dataset used: https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction

[2] https://www.kaggle.com/code/ipshitagh/eda-and-visualizations

[3] https://www.kaggle.com/code/karanchinchpure/predict-price-of-used-cars-regression-problem

[4] https://www.geeksforgeeks.org/?newui

[5] https://www.tensorflow.org/api_docs/python/tf/

[6] Dataset used for Prediction

https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction

[7] https://www.kaggle.com/code/ipshitagh/eda-and-visualizations

[8] https://www.kaggle.com/code/karanchinchpure/predict-price-of-used-cars-regression-problem