# 

# 

# Car Price Prediction Project

# 1. Introduction

Regression can be defined as a statistical technique that allows us to study and identify how one or more features relate to one another. Regression analysis is aided by a technique that adapts to assist you to discover which aspects of your data are essential, as well as which aspects are irrelevant.

There are thousands of used cars that are being sold every year in the UK. The used cars are usually posted on platforms like Gumtree, Facebook marketplace, and other platforms. When a new person decides to sell his used vehicle, the biggest confusion occurs on pricing. Unknowingly people ask for overpricing for their vehicles or value under-price their vehicles in most cases. To avoid this confusion based on all the ads that were posted we can build a regression model that automatically predicts the price of the vehicle based on certain features. This would be helpful for the new sellers in the market and also will help buyers by avoiding them paying overprice for a vehicle. Having to check the used car price frequently because of market influences, making it a difficult task to see if a price on the used car is a good deal. To help with purchasing decisions, this project aims to build models that can correctly predict the cost of the used specific car on its features. We perform and evaluate different learning approaches on a data set that comprises the selling prices. It's a manufacturer-set price with government taxes as an extra cost. So, people ordering a new vehicle can feel confident about their investment in it. Used car sales are rising because new car prices are getting higher, and customers can't afford new cars because they are lacking in funds This used purchase price prediction system is necessary to give an accurate assessment of the value of the car using its multiple features. Although there are many sites that offer the service, they may not have the best method of predicting results. Also, various designs and systems may aid in predicting the true market value of a used car. In order to ensure that you get the proper return on your investment, it is imperative to get a handle on market values when buying and selling.

There are lots of different creating and models in every major American city. We found that Random Forest produces better results. Regression with a linear component also provided excellent results with the benefit of significantly decreased costs. For the regression analysis, I will be using multiple techniques of regression models that are available in Machine Learning and Deep Learning. Techniques such as Linear Regression, Lasso Regression, Ridge Regression, Support Vector Regression, Decision Tree Regression etc., are the models that we will be using as part of this project.

I will be using Python language and Jupyter IDE for the whole project. To successfully completing the project I will be using Python libraries such as NumPy, Pandas, Scikit Learn, etc.

## 

## 1.1. Dataset

The data is readily available on the Kaggle platform for the use of research purposes. I will be using this data to perform the regression analysis on the data. This data was scraped from the listings of the used cars. There are different cars from different manufacturers are listed in the dataset. In total there are close to 100k data points in the data.

This data will be further divided into train, validation, and test sets in the project for evaluation and building robust models. These are the following features of the dataset

Model: The Model says what is the model of the Car

Year : This feature tells about the Manufacturing year of the car

Price : This feature tells about the price of the cars

Transmission: Transmission rates of the car

Mileage : The Mileage feature shows how much mileages can give any car.

fuelType: which type of fuel used by the cars

tax : how much tax the buyer has to pay on the car

mpg : This feature tells about the mpg of the cars

enginesize : This feature tells about the Size of the engines of the cars

## 1.2. Project Aim

The goal of this project is to develop a classifier that can predict the price of the cars. The outcome will be assessed using alternative machine learning models. Because the data contains both numerical and text elements, one model has been applied to the textual information and the other to the numeric data. The Machine learning models will take the features as an input and gives the predicted price as an output

## 1.3.Research Questions

## A research topic is a topic that a study tries to find answers for. This refers to something in the study that's addressed by examining the data and telling the storey it reveals. The research objective is typically formulated in a way that highlights a variety of aspects, such as the study's research population and factors, and also the study's main purpose. Research is commonly centred around scientific research. It's not surprising that researchers frequently revisit and revamp their research questions: Research questions tend to be evolving rather than static. Researchers must reassess and adjust questions as they conduct literatures and build a framework for the study.

* How does traditional classification algorithms such as SVM and decision trees perform for regression tasks compared to the regular regression algorithms such as linear regression?
* Regression models in Machine Learning which of these models are capable of accurately predicting the prices of the vehicles?

## 1.4. Objectives

* Clean the data by removing the outliers and handling the missing values.
* Make interpretations from the data visualization during the data analysis.
* Use feature engineering to extract new useful features from the data available.
* Use Regression Analysis to accurately predict the prices of the cars to help the customers correctly price their vehicles on selling platforms.

## 1.5. Tools

There are several tools employed as phase of this project to achieve its main objective. a few essential tools utilized during this Endeavour include:

NumPy: NumPy is a Library in python designed to help with array management. It has features for linear algebra, Fourier transforms, and matrices, as well.

Pandas: Pandas is a Python-based library that works with data analysis and manipulation. The focus is on the operations and data structures required to manipulate tables and time - series data.

Scikit-learn: Scikit-learn is a library for Python that helps with machine learning. It is open source and available to anyone. It includes several algorithms, such as support vector machines, among others.

Matplotlib: Matplotlib is a tremendous 2D plotting library in Python, perfect for visualising array data. Matplotlib is a library built on the concept of NumPy arrays, and it is made to work with the other components of the Scipy stack. Matplotlib has numerous plots, including lines, bars, scatter plots, histograms, and more.

Seaborn: Seaborn is an example of a Python library that works with the matplotlib data visualisation framework and integrates with pandas data structures. Seaborn is Seaborn's central visualisation system, which is crucial in helping the exploration of data. See how the distribution is univariate and bivariate.

## 1.6.Ethical, Social, and Legal Issues

## 

## 1.6.1. Ethical Issues

Ethical issues are those issues that let us understand what is fundamentally right and wrong. The ethical issues that are possible in this case of the project are as follows:

Machine Ethics & Algorithmic Biases:

This issue under the ethical OS toolkit helps us understand right and wrong while building the machine learning algorithms. In our project, we use Machine Learning for training the regression models and the model might behave biased due to the presence of outliers in the data. This needs to be checked and the outliers need to be removed from the data to avoid such issues of bias in the models.

## 1.6.2. Legal Issues

Legal issues are those issues that let us understand what is legally right and wrong. The legal issues that are possible in this case of the project are as follows:

The predictions that are being passed from the algorithms to the users should be accurate and should not have a drastic difference in the original value of the cars being sold and the predicted value. Any misclassification of prices could misguide the people to lose their money or face loss. This could be a reason for the legal issues. To avoid this issue the algorithms should be properly trained on the data and should cross-check the performance before production and the predictions should be monitored while the algorithm is in production.

## 1.6.3. Social Issues

Social issues are the issues that harm society. The social issues that are possible as an outcome of this project are as follows:

Issues related to unemployment as a result of automation can be a possible social issue that might occur as a result of this project.

## 2. Methodology

## 3.1. Installing set-up

I used Python 3.7, which I downloaded from the official Python website, for my project. I installed Python on my machine by setting it up on my hard drive and adding Python to my path. Additionally, I've installed Anaconda on my computer to get everything ready for running Python via Jupyter. After that, I was able to execute this project by using command prompt to install a few libraries. Installing the libraries required the following commands:

pip install pandas

pip install numpy

pip install scipy

pip install scikit learn

pip install matplotlib

pip install seaborn

pip install collections

pip install math

pip install collection

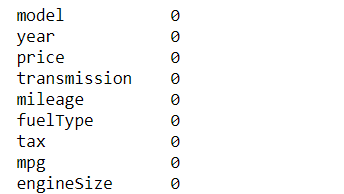
pip install warnings

## 3.2 Data Exploration

The data for this project is available at Kaggle [- https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes](-%20https:/www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes). The dataset includes 10,668 records with 9 variables combination data kinds make up the information. The following is a basic explanation of the variables:

|  |  |  |  |
| --- | --- | --- | --- |
| S. No | Variable | Datatype | Description |
| 1. | model | Object | Model of the cars |
| 2. | year | Int | Manufacturing year of the cars |
| 3. | price | Int | Price of the cars |
| 4. | transmission | Object | Transmission rates of the car |
| 5. | mileage | int | Mileage of the cars |
| 6. | fueltype | Object | Type of fuel used by the cars |
| 7. | tax | int | How much tax buyer has to pay on the cars |
| 8. | mpg | float | Mpg of the cars |
| 9. | enginesize | float | Engine sizes of the cars |

The dataset is further explored to identify Exploratory data Analysis and Null values.



There is no missing values in the dataset. Here In the given dataset there is some columns of Object type. So In the further steps we have to convert the categorical variables into numerical ones.

## 3.3 Exploratory Data Visualization

Exploratory Data Visualization (EDA) is very important part of the data science pipeline or any data science project. Exploratory Data Analysis is a vital method that involves conducting initial investigations on data in order to identify trends, identify discrepancies, evaluate assumptions, and verify conclusions using summary statistics and data visualizations. Exploratory Data Analysis (EDA) is a computational data analysis technique focused on John Tukey's pioneering work. EDA offers a basis for a wide variety of data analytic activities and addresses the diverse types of data and architecture encountered by applied researchers. EDA's fundamental conceptual and computational tools include the use of graphics and interactive data visualisation, a focus on model creation, diagnosis, and evaluation, addressing fundamental measurement issues associated with various distributions.

Although these methods serve as a foundation for all research, EDA places a high value on data - based learning from data to enhance standard hypothesis testing procedures that may neglect critical unanticipated aspects of data and their effect on modelling and estimation. The EDA, it is claimed, is critical both in the early stages of science, where hypotheses and model development must be well-informed.

## 

## 3.3.1. Univariate Analysis

The technique of univariate analysis is for comparing and analyzing the relationship between a single feature and response variable. The prefix "uni" highlights the analysis only covering a single variable and its impact on a parameter.

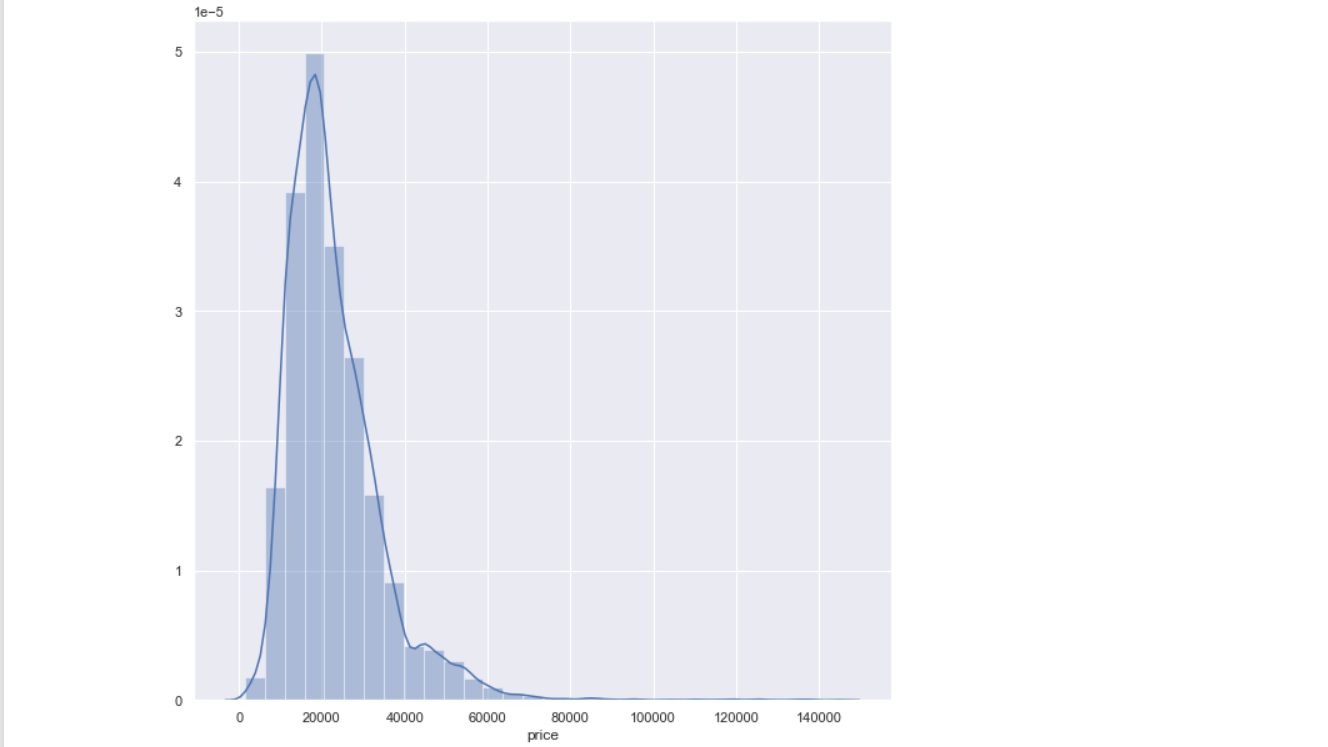
For Example, the study can focus on a variable such as "gender," "height," or "weight."

However, only one variable is examined each time.

## 

## 3.3.1.1. Price

Price is dependent variable in our dataset which is numerical feature.



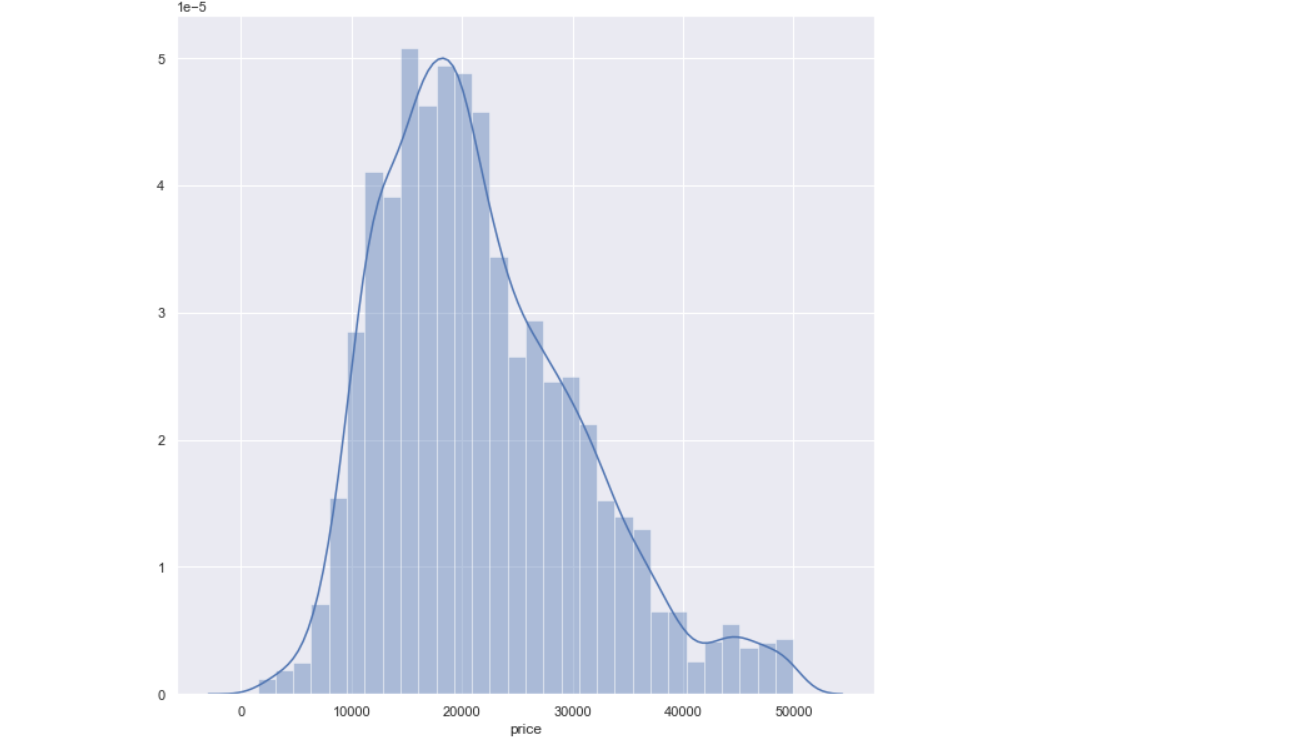
The above plots show that most of the data points, i.e., 75% of the total data points are prices less than 27990 and the least priced car is 1490 and the item that is priced the highest is 1,45000.00. It can be observed that the distribution of the price is skewed towards left side where the maximum prices of cars are under 27990.

The following table shows the description of the Price variable:

Table 1 - Price variable description

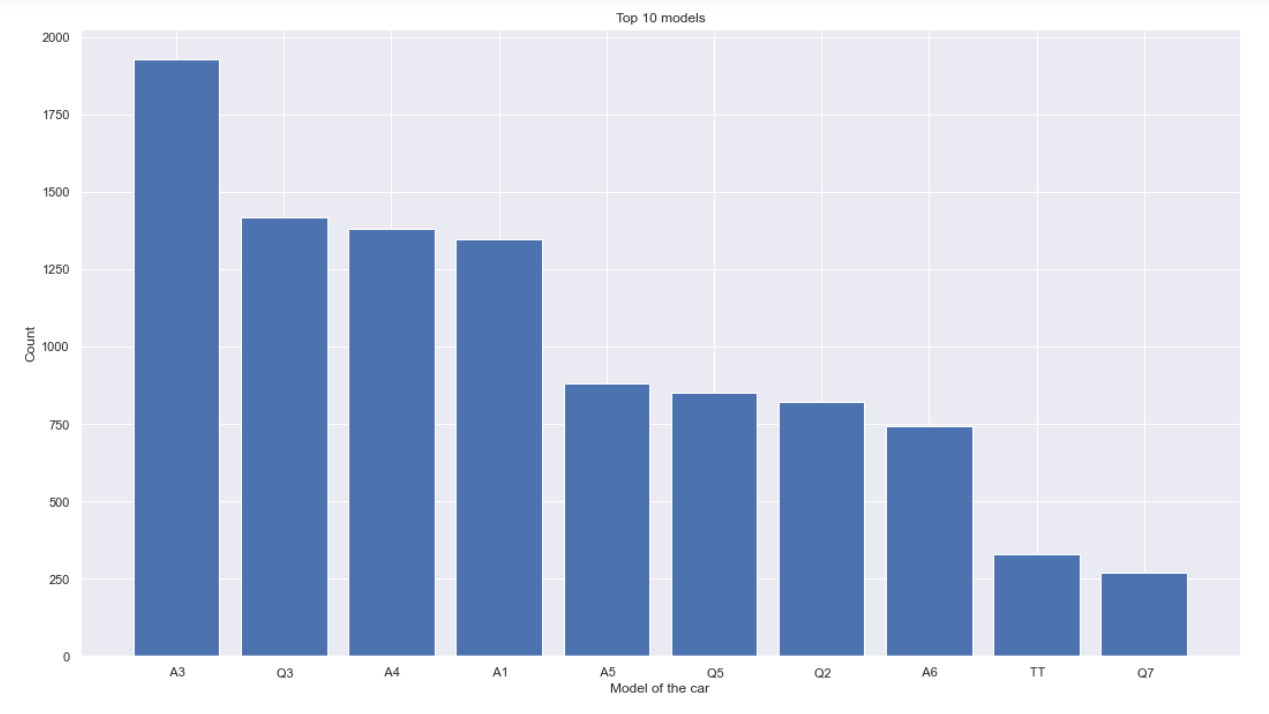
|  |  |
| --- | --- |
| Count | 10668.000 |
| Mean | 22896.6850 |
| Standard Deviation | 11714.841888 |
| min | 1490.0000 |
| 25% | 15130.7500 |
| 50% | 20200.000 |
| 75% | 27990.000 |
| max | 145000.000 |

There is some outliers in the dataset, That’s why the distplot contains some skewness. In our dataset there are 319 data points are outliers. So we will remove these outliers. After removing the outliers our dataset contains 10342 data points and the distplot will be:

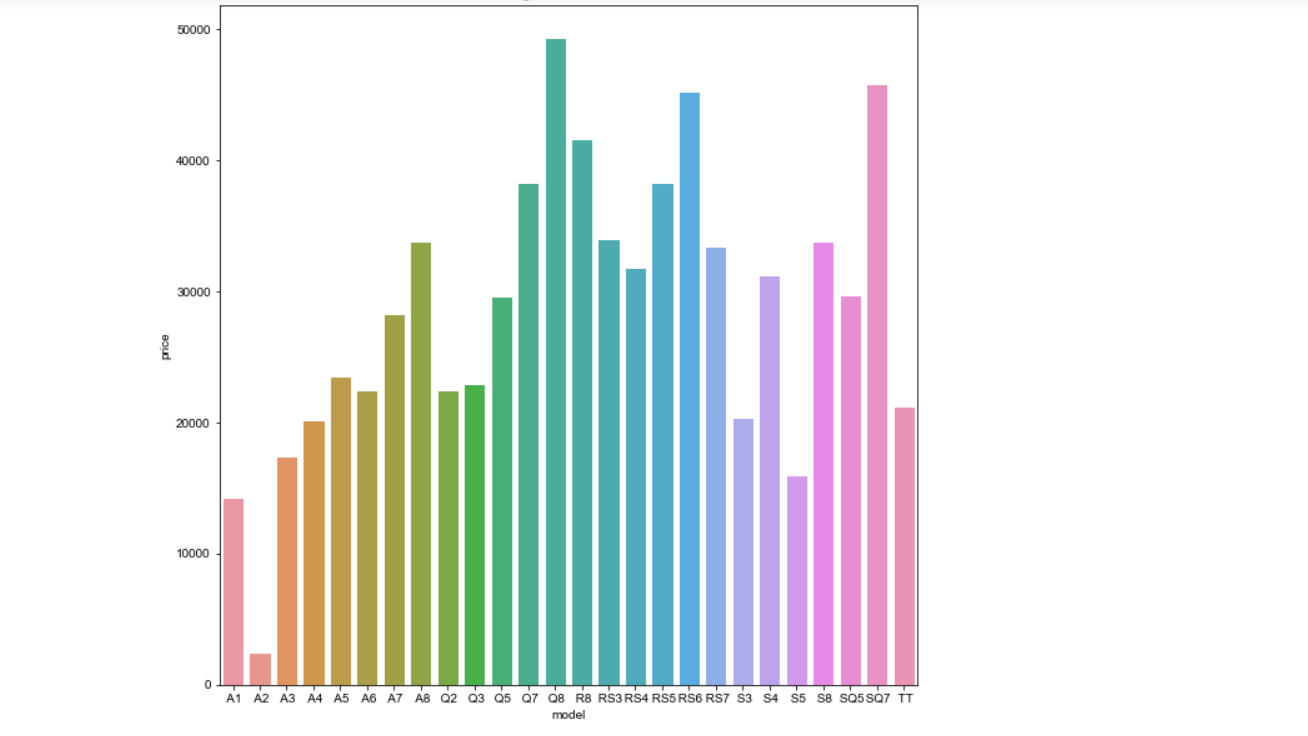


## 3.3.1.2. Model

Model variable in the data describes the Model of the Cars. Similar to the Transmission variable, Model variable is also a categorical variable. There are 26 unique models in the data and the top 10 models in the data are:

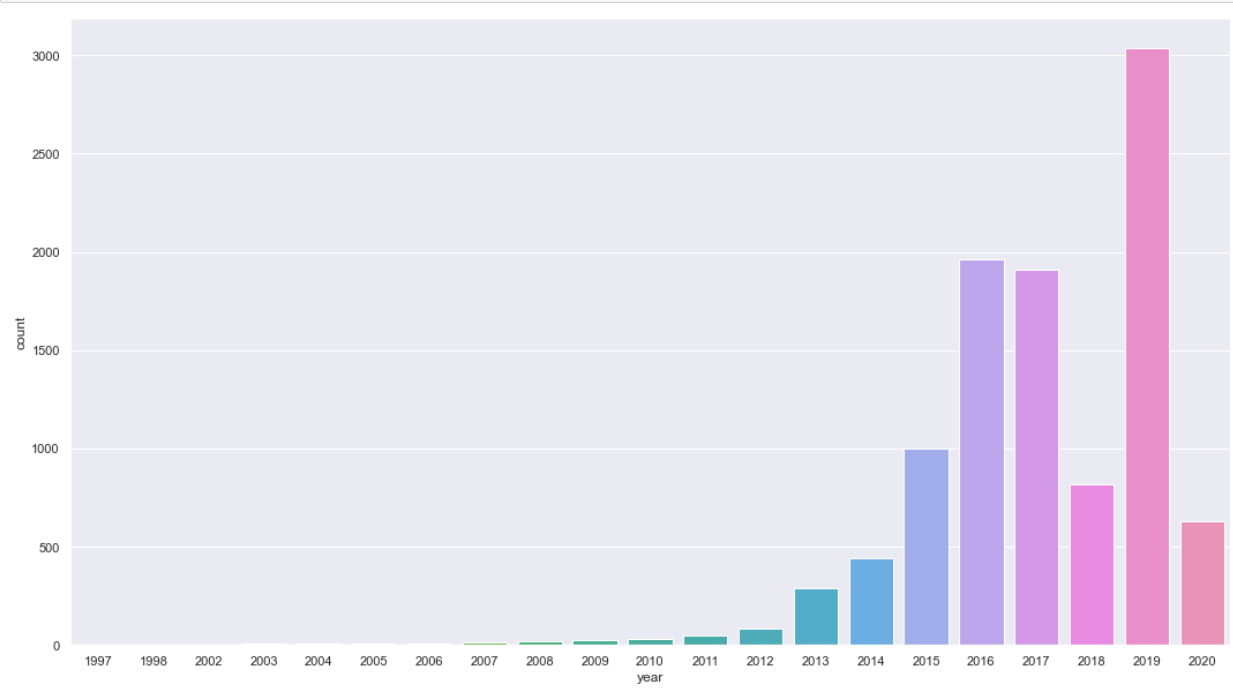


Now, In the below graph we will check the Models of the cars according to the price. The Q8 Model of the car has highest price.



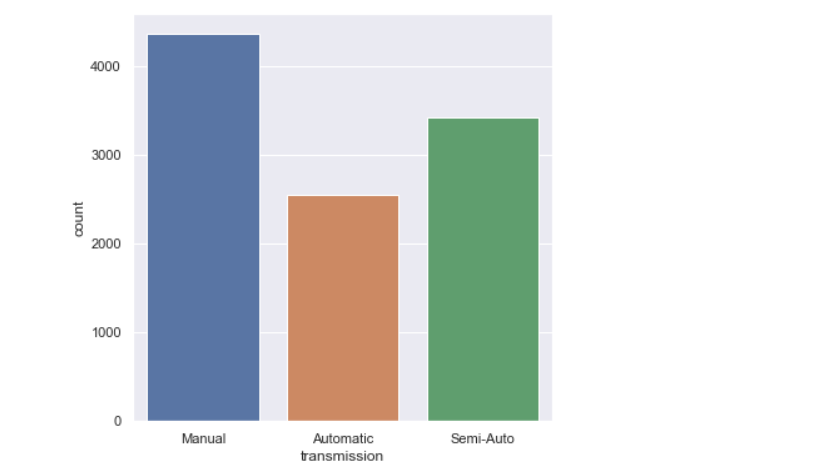
## 3.3.1.3. Year

Year variable in the data describes about the manufacturing year of the cars. Year variable is the int variable. In the below graph we can see that In 2019 is the manufacturing year of 3039 cars and 1997 and 1998 is the manufacturing year of only one car.



## 3.3.1.4 Transmission

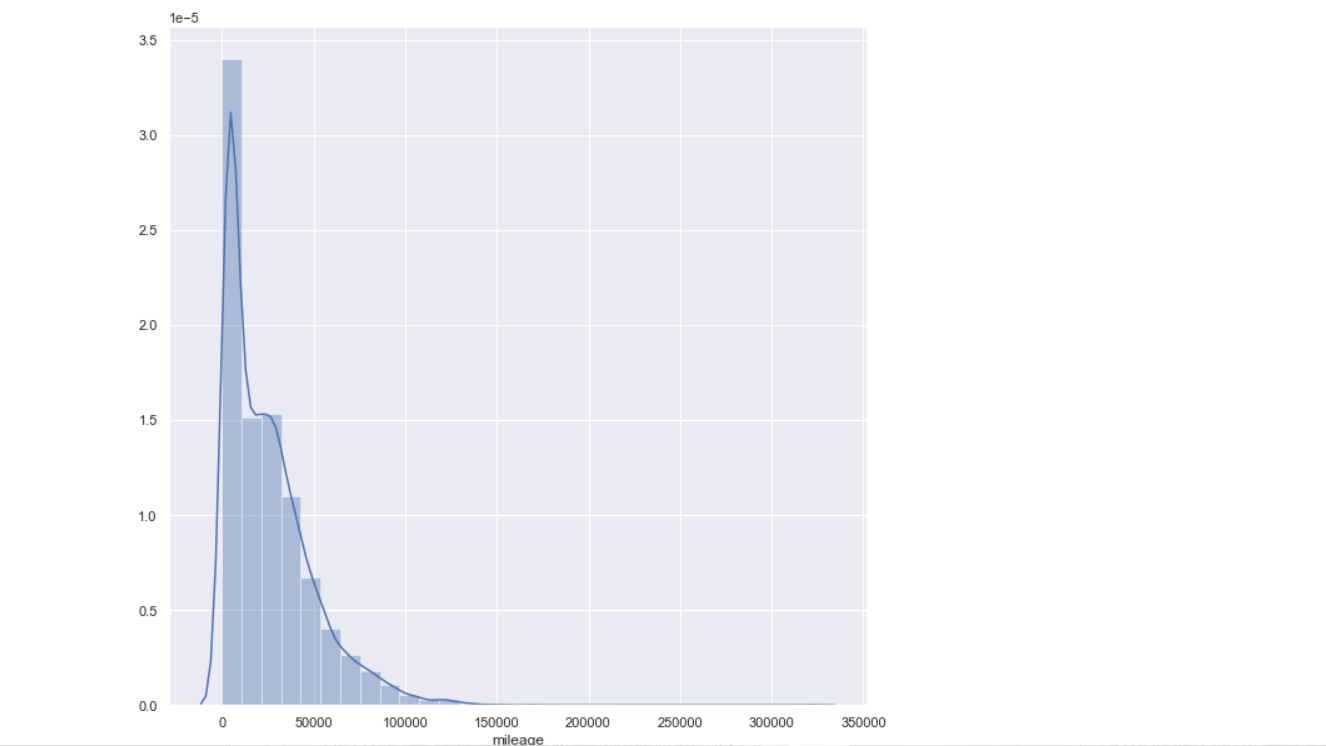
Transmission variable in the data describes about the type of transmission of the cars. This is a categorical feature and there are 3 categories in this feature.



|  |  |
| --- | --- |
| Manual | 4369 |
| Semi-Auto | 3426 |
| Automatic | 2547 |

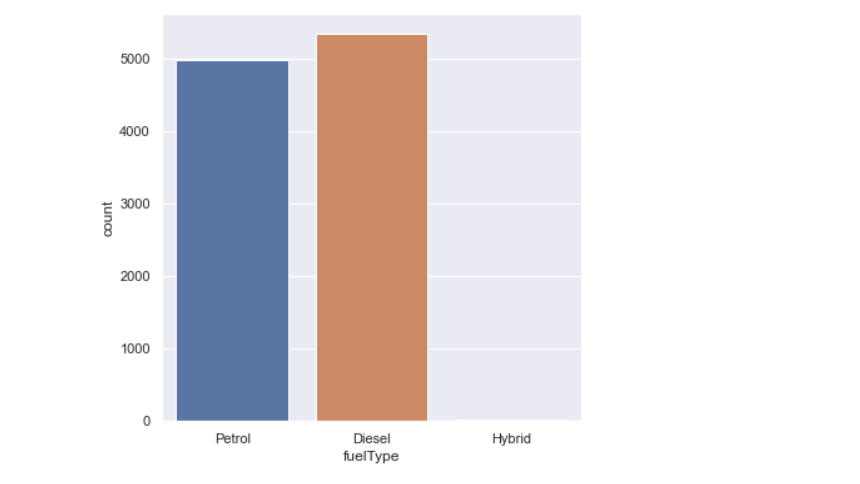
## 3.3.1.5. Mileage

Mileage is a variable in the data describes about the mileage of the cars. This is a integer variable. The minimum mileage of the cars is 5 and the maximum mileage of the cars is 323000 and the 50% cars have the mileage of 19823.



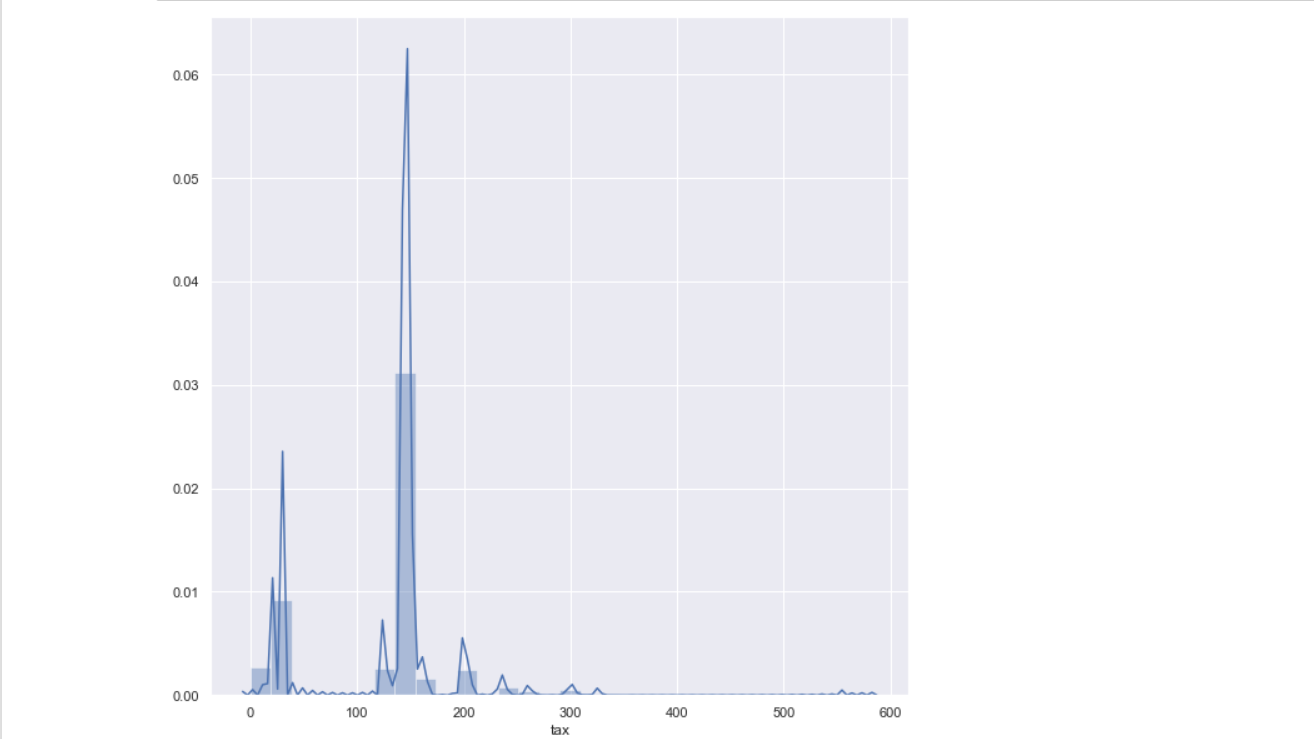
## 3.3.1.6. Fuel Type

Fuel Type is a variable in the data that describes about the Fueltype of the cars. This is a categorical variable. There are 3 categories in this feature petrol, diesel , Hybrid. There are 5345 diesel cars , 4970 petrol and 27 hybrid cars.



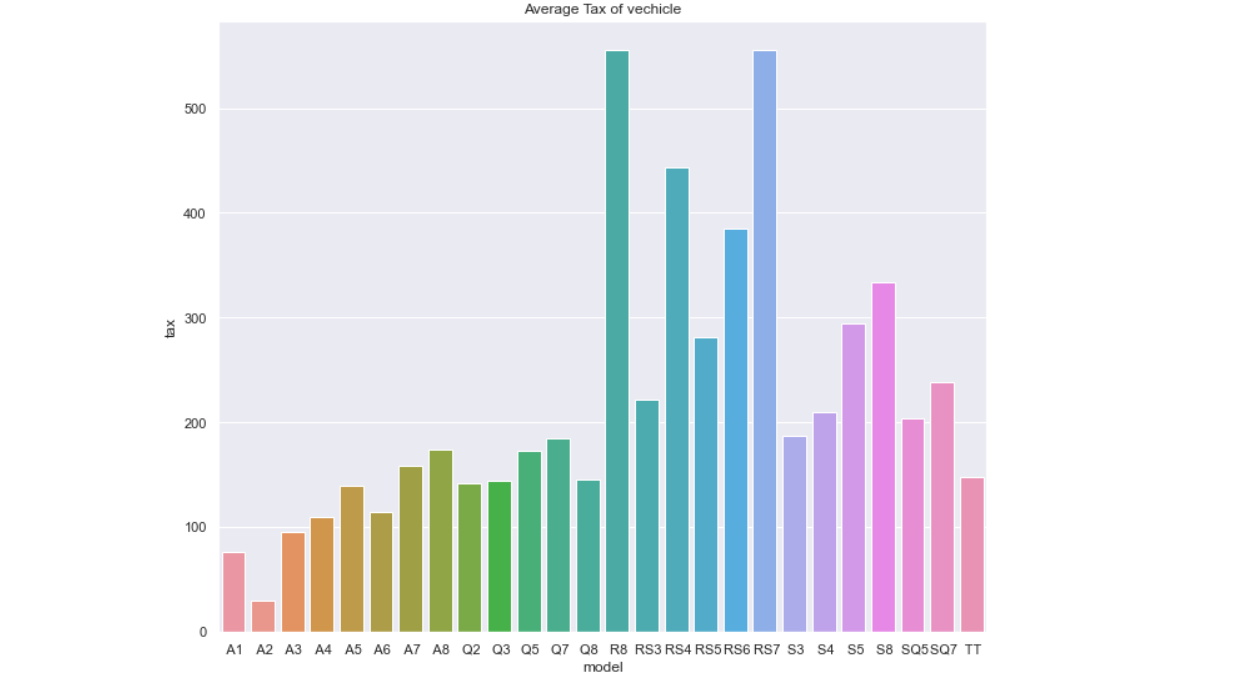
## 3.3.1.7 Tax

Tax is a variable that describes about the tax on the Cars. This is a integer variable. The Maximum value of Tax is 580.00 and the minimum value of tax is 0.

The description of the tax variable is :

|  |  |
| --- | --- |
| Count | 10342.00 |
| mean | 125.00 |
| Standard deviation | 66.9700 |
| min | 0.00000 |
| 25% | 125.000 |
| 50% | 145.000 |
| 75% | 145.000 |
| max | 580.000 |

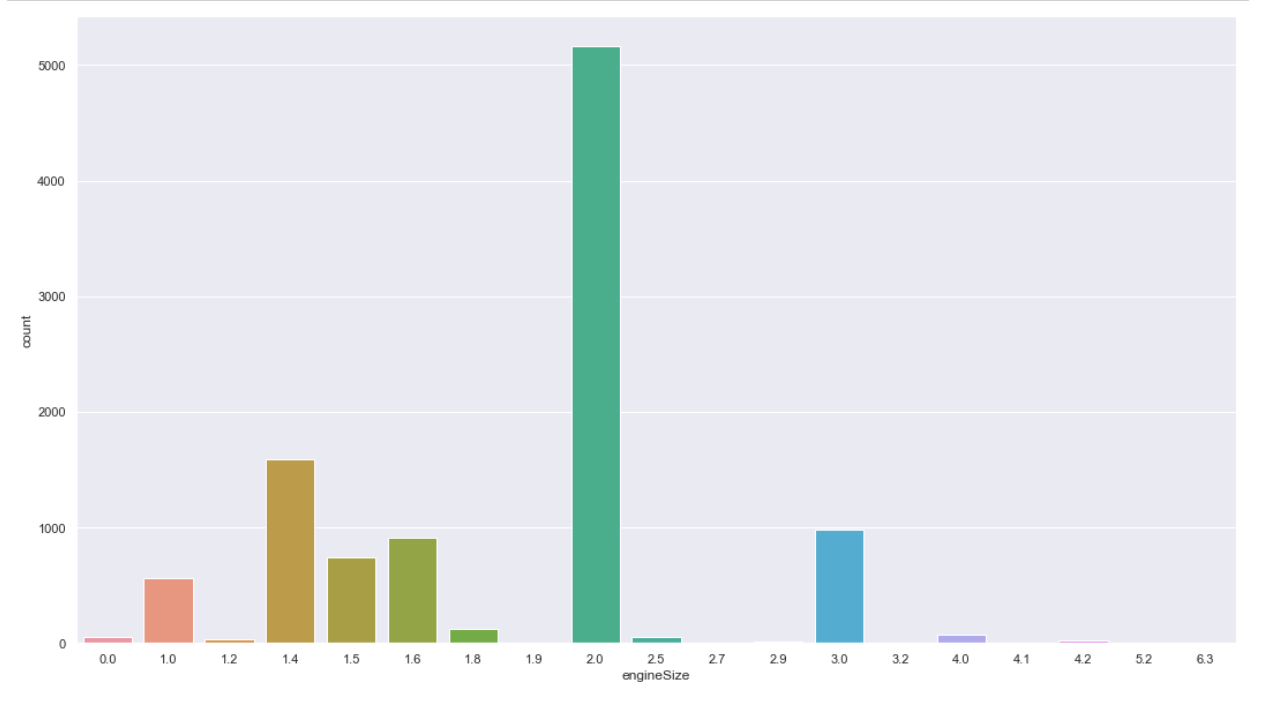
Then i have plotted the tax variable with model. Means how much tax the buyer has to pay on Each model.



On R8 Model and RS7 model the buyer has to pay the maximum tax.

## 3.3.1.8 Engine Size

Engine Size is a feature that gives the information about the size of cars of the Engines. This is integer column.



## 3.4 Data Cleaning

Data cleaning is a process that takes in flawed, inconsistent, irrelevant, overlapping, or flawed data, and transforms it into clean data, which is ready for a team to run numbers on. Data that is irrelevant or potentially problematic for research purposes tends to slow down the work and produce bad results. There are several different approaches to cleaning the data depending on the answers desired and the storage method being used. Data cleaning isn't just about removing information; it's about being able to remove information without destroying a dataset's value. And finally, the ultimate objective of data cleaning is to provide a standardised and uniform data set to make it easy for data analytics and data dashboards to get and use the appropriate data each time a query is made.

Data cleansing, unlike data removal, is used to improve data by eliminating errors, standardising data sets, and cleaning up extraneous data points.

**Steps involved in Data Cleaning:**

3.4.1.Handling missing data

This is the main step of data cleaning . there is no missing values in our dataset. So there is no need to apply this step.

## 3.4.2. Encoding of Categorical Features

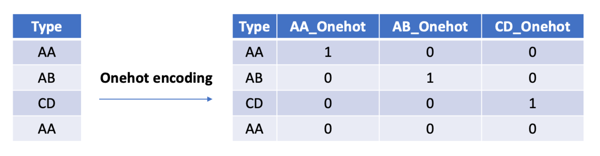
A category variable does have two or more categories. A categorical variable can be either nominal or ordinal. Categories of a nominal variable do not have any built-in ranking. Gender is often a categorical data that is simply composed of two categories with no natural order. The ordinal variable exhibits a well-defined order.

Many machine learning algorithms can't understand the categorical data on their own. On the other hand, Decision trees can learn straight from the data itself. They require all parameters to be numeric because of this the categorical data must be transformed into numbers.

Few types of categorical variable encoding are:

**1) One hot encoding**: One approach to converting data in order to provide it to an algorithm, which improves the prediction results. One-hot conversions split the categorical values into binary columns and apply a 1 or 0 designation to each column. A binary vector represents each integer value.

Example: We can see in the figure in the first-row red is present, so In the first row only red column is 1. Others will be 0.

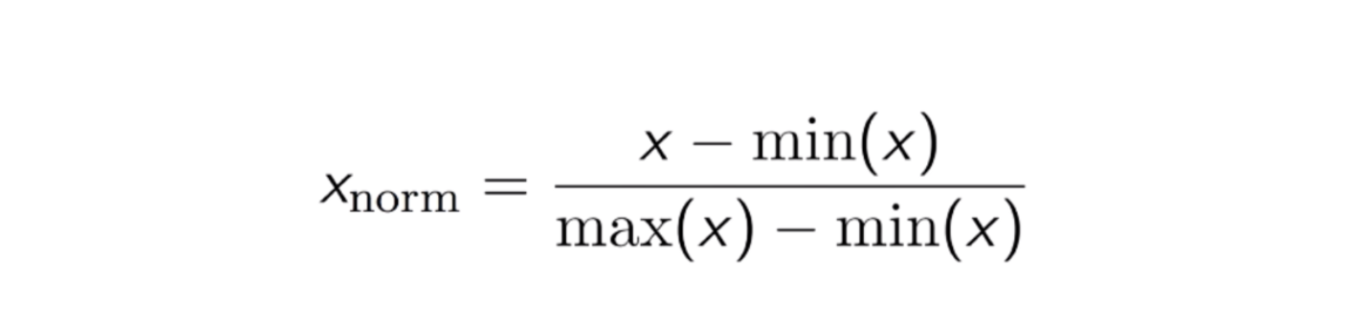


2. **Integer Encoding / Label Encoding**: In Label Encoding simply we will give the numbers to each category.

In this project I have done Label encoding for converting the categorical features into numerical features. Model, Transmission, and fuelType are the categorical features in our dataset.

## 3.4.3 Feature Scaling

Feature Scaling is also very important step of Data Cleaning if our feature lies on different –different scale. It is better to apply feature Scaling. A common practice when preparing data for machine learning is normalization. To normalize, it helps to keep everything on the same scale, to keep different value ranges constant. To make the best use of machine learning, datasets do not have to be normalized.



In this project i have applied feature Scaling on all the features except dependent variable (price). It will convert all the features in between 0 to 1 range.

## 3.5 Algorithm and Techniques

For this project i have used Machine Learning Algorithms like Decision Tree Regressor, Random Forest Regressor , Support Vector Machine Regressor and Linear Regression Algorithms etc.

Machine Learning

Machine learning can be addressed as a subset of artificial intelligence that allows machines to learn and develop automatically without any supervision of the humans. Machine learning is concerned with the development of computer programmes capable of having access to data and use it to train themselves. The learning phase starts with insights or facts, such as instructions to search for correlations in data and make more informed potential judgments based on the examples provided. The aim of the machine learning is to allow computers to train on the data automatically and change their activities as per the requirements without any human involvement or assistance.

Machine learning allows the processing of enormous amounts of data. Although it usually produces more reliable and timely outcomes for identifying lucrative opportunities or risky threats, correctly training it may take additional time and effort. Combining machine learning and artificial intelligence with variety of technologies have the potential to render it much more efficient at analysing vast amounts of data. Within the realm of machine learning, there are several algorithms that are published every day and they are organised based on whether they focus on supervised or unsupervised learning or feature similarities. Machine learning algorithms can be classified as two types. Namely, supervised machine learning algorithms and unsupervised machine learning algorithms.

### 3.5.1. Linear Regression

The first machine learning model that is trained on our data is a linear regression model. Linear regression is one of the most used regression algorithms in Machine Learning. A linear regression model assumes that the relationship between two variables is linear and tries to fit a linear equation on the data. Out of these two variables, one of them is the independent variable and the other is a dependent variable.

Example:

Let us consider 50 students in a class and we are trying to predict the weights of the students based on their heights.

Here, the independent variable is the heights of the students and the dependent variable is the weights of the students. A linear regression model was introduced for this task. Here the linear regression assumes that there is a linear relationship between the heights of the students and weights of the students.

Chart, scatter chart

Description automatically generated

Figure - Linear regression illustration (Agarwal, 2018)

Each input value or column in the linear equation is assigned one scale element, referred to as a coefficient, and denoted by the capital Greek letter Beta. Additionally, an additional coefficient is applied, which provides the line with an additional degree of independence (e.g., the ability to go up and down in a two-dimensional plot), which is often referred to as the intercept or bias coefficient. For instance, in a simple linear regression, the model will take the following form:

**y = mx + b**

where,

x is the independent variable.

y is the dependent variable.

The slope of the line is m, and b is the y-intercept.

If we have more than one input (x) in higher dimensions, the line is referred to as a plane or hyper-plane. Thus, the representation is the equation's structure changes with coefficients' values.

When the coefficient of the equation becomes zero, it essentially eliminates the independent variable's impact on the formula and therefore on the model's prediction. This is significant when considering regularisation techniques, which alter the learning algorithm in order to minimize the difficulty of regression models by imposing a constraint on the absolute size of the coefficients, effectively dropping some to 0.

For this project, we have implemented the linear regression on the data as our initial model. The whole data was split into training and test sets with 70% being the training sets and 30% being the test set. The linear regression model was initially trained on the training data and then it was implemented on the test set. The linear regression was then implemented on the test set to predict the prices of the cars.

### 3.5.2. Ridge Regression

The next machine learning regression model that I have trained on the data is a Ridge regression model. Ridge regression model is an extension to the linear regression model that adds the penalty to the loss function such that the models have small coefficients and behave simply. To be precise, a ridge regression model is a regularization of the linear regression with L2 penalty. Without taking such insignificant variables into consideration, the input coefficients shrink to better reflect the coefficients in the model.

These regression machine learning models primarily derive their results by using the regression equation, which is known as the following:

**Y = XB + e**

Where

x is the independent variable.

y is the dependent variable.

e represents the residual errors.

A hyper parameter named lambda (λ) is used to monitor the penalty's weighting in the loss equation. A default value of 1.0 totally weights the penalty, whereas a value of 0 completely ignores it.

**ridge loss = loss + ( λ \* L2)**

For this project, we have implemented the ridge regression on the data as our second model. Similar to the linear regression, the whole data was split into training and test sets with 70% being the training sets and 30% being the test set. The ridge regression model was initially trained on the training data and then it was implemented on the test set.

## 3.5.3 Lasso Regression

The Lasso is a technique that is used for regularisation. Its application includes using it over regression methods to produce more accurate predictions. Shrinkage occurs in this model. Data shrinkage occurs as values approach a central point. The lasso encourages models that are simple and sparse (i.e. models with fewer parameters). To simplify, this type of regression is helpful when your models have high levels of multicollinearity, when you want to do automated model selection like variable selection/parameter elimination, or when you have multicollinearity in your model.

Lasso Regression uses L1 regularisation method in Lasso Regression. Feature selection is performed on our behalf with greater number of features present because the algorithm can handle it. These regression machine learning models primarily derive their results by using the regression equation, which is known as the following:

**Y = XB + e**

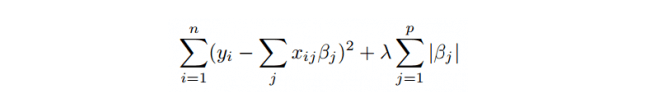
Where

x is the independent variable.

y is the dependent variable.

e represents the residual errors.

A hyper parameter named lambda (λ) is used to monitor the penalty's weighting in the loss equation.



For this project, we have implemented the lasso regression on the data as our third model. Similar to the linear regression, the whole data was split into training and test sets with 70% being the training sets and 30% being the test set. The lasso regression model was initially trained on the training data and then it was implemented on the test set.

## 3.5.4 Random Forest Regressor

Supervised learning algorithms, like random forest, apply ensemble methods to both classification and regression problems. At training time, the algorithm constructs a host of decision trees and uses each one to forecast a different conclusion.

Random forest assumes the "wisdom of crowds," where large numbers of models that are uncorrelated operate as a committee. In comparison to any individual constituent model, this approach is assumed to perform better.

One reason is that trees help each other avoid their own errors. The trees within one random forest are not influenced by each other. The random forest algorithm is essentially a more efficient classifier, which relies on the outcomes of numerous classification models to extrapolate the best possible results.

## 3.5.5 Decision Tree Regressor

Decision tree builds models, which have a tree structure. While progressively developing a decision tree, it starts to break down a dataset in to the progressively smaller subsets. A tree with leaf nodes and decision nodes is the end result. A decision node, such as Outlook, will have multiple paths, with each one representing an attribute's attribute values.

A tree's topmost decision node is the node which is responsible for identifying the root node, the best predictor. When presented with both numerical and categorical data, decision trees can do the heavy lifting.

## 3.5.6 Support Vector Machine Regressor

One of the most common and frequently used classification algorithms is SVMs, also known as Support Vector Machines.. This algorithm, which handles non-linearity, results in a robust prediction model. The algorithm of Support Vector Regression helps predict discrete values and is utilised in a supervised learning context. The method used by Support Vector Regression is the same as the SVMs'.

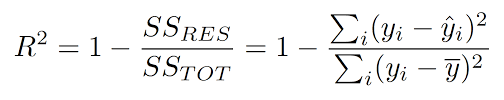
SVR is centred around finding the best-fitting line. The line of best fit in SVR is the hyperplane with the most points.SVR does not make any attempts to minimise error like the other regression models. Instead, it attempts to fit the most ideal line within a range. A hyperplane is the imaginary plane, and the distance between it and the boundary line is the hyperplane's threshold value. SVR's scaling issues are evident in its fit time complexity, which is more than quadratic, making it difficult to work with datasets larger than about 10,000 samples.

Linear SVR or Stochastic gradient descent Regressor is used to handle large datasets. The linear kernel gives a faster implementation of SVR, but does not consider it. The Support Vector Regression model is created with only a portion of the training data, as it uses the cost function to exclude those samples that have results that are close to their target.

## 3.6. Performance Evaluation Metrics

While it is important to validate an evaluation of a model's skill, using metrics is necessary to evaluate the performance of a model. Metrics are employed to determine the most suitable problem-solving method. These metrics are what matter most.

Regression models explain variance in dependent variables (such as expected output or cost) in terms of the independent variables (such as a country's risk-free borrowing rate). R² is a statistic that shows the extent to which the dependent variable is explained by independent variables. In other words, r-squared shows how the data fits the model of regression (the goodness of fit).

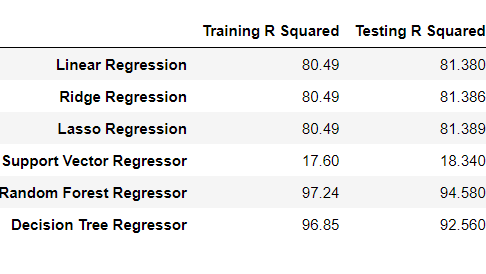


If R2  = 0 The model will be worst.

If R2 = 1 The model is very good

I have calculated the R2 value of each machine learning model for training and testing data.

The values are:



R2  value for Random Forest Algorithm is highest. That’s why Random Forest Algorithm gives accurate results as compare to other models.

## Conclusion

In this paper, I have worked with different models of Machine Learning to predict the price of the cars from the given data to identify the better performing models.

Initially, the exploratory data analysis was performed to understand the insights in the data. In Exploratory data Analysis I have performed various data visualization to understand the data. Further, feature engineering was performed. In feature engineering firstly i have checked the null values . there is no null values in the dataset . Then i have applied Label encoding for converting the categorical variables into numerical ones. Then i have applied feature scaling for normalizing purpose. Then I have divided the data into training and testing data. Then we will give this training data to the Various Machine learning Algorithms (Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor etc)

To evaluate these models, I have used R2 as the metric. By analyzing the performances of the models, it has been observed that the Random Forest has recorded the Highest R2  Values.

To answer the research questions posed, after performing the analysis on the data with different models with all the evidence we can conclude that the Decision Tree , Random Forest Algorithms working better than Linear, Ridge and Lasso. It has been observed that Random Forest gives highest R2  Value..