

**CAR PRICE PREDICTION**

Submitted by:

Ankita Prashant Chaudhari

**ACKNOWLEDGMENT**

[1] NATIONAL TRANSPORT AUTHORITY. 2014. Available from: http://nta.gov.mu/English/Statistics/Pages/Archive.aspx [Accessed 15 January 2014].

[2] MOTORS MEGA. 2014. Available from: http://motors.mega.mu/news/2013/12/17/auto-market-8-decrease-sales-newcars/ [Accessed 17 January 2014].

[3] LISTIANI, M., 2009. Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Thesis (MSc). Hamburg University of Technology.

[4] RICHARDSON, M., 2009. Determinants of Used Car Resale Value. Thesis (BSc). The Colorado College.

[5] WU, J. D., HSU, C. C. AND CHEN, H. C., 2009. An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications. Vol. 36, Issue 4, pp. 7809-7817.

[6] DU, J., XIE, L. AND SCHROEDER S., 2009. Practice Prize Paper - PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation and Genetic Algorithms to Used-Vehicle Distribution. Marketing Science, Vol. 28, Issue 4, pp. 637-644

**INTRODUCTION**

**Business Problem Framing**

You are required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the business strategy etc. to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

**Conceptual Background of the Domain Problem**

No matter what the problem is, but data science has the same framework to follow for every problem. These are the steps to follow for solving the problem from the higher viewpoint:

* EDA(Exploratory Data Analysis) and Data Cleaning
* Feature Engineering
* Feature Selection
* Modeling
* Model Evaluation
* Conclusion of the problem

**Review of Literature**

The first paper is Predicting the price of Used Car Using Machine Learning Techniques.In this paper, they investigate the application of supervised machine learning techniques to predict the price of used cars. The predictions are based on historical data collected from daily newspapers. Different techniques like multiple linear regression analysis, k-nearest neighbours, naïve bayes and decision trees have been used to make the predictions. The Second paper is Car Price Prediction Using Machine Learning Techniques. Considerable number of distinct attributes are examined for the reliable and accurate prediction. To build a model for predicting the price of used cars in Bosnia and Herzegovina, they have applied three machine learning techniques (Artificial Neural Network, Support Vector Machine and Random Forest). The Third paper is Price Evaluation model in second hand car system based on BP neural networks. In this paper, the price evaluation model based on big data analysis is proposed, which takes advantage of widely circulated vehicle data and a large number of vehicle transaction data to analyze the price data for each type of vehicles by using the optimized BP neural network algorithm. It aims to establish a second-hand car price evaluation model to get the price that best matches the car

**Motivation for the Problem Undertaken**

Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices.

**Analytical Problem Framing**

**Mathematical/ Analytical Modeling of the Problem**

**1.What is Statistical Modeling and How is it Used?**

Statistical modelingis the process of applying statistical analysis to a dataset. A statistical model is a mathematical representation (or mathematical model) of observed data.

When [data analysts](https://www.northeastern.edu/graduate/blog/what-does-a-data-analyst-do/) apply various statistical models to the data they are investigating, they are able to understand and interpret the information more strategically. Rather than sifting through the raw data, this practice allows them to identify relationships between variables, [make predictions](https://www.northeastern.edu/graduate/blog/predictive-analytics/) about future sets of data, and visualize that data so that non-analysts and stakeholders can consume and leverage it.

**Important Statistical Techniques in Data Analysis**

Before any statistical model can be created, an analyst needs to collect or fetch the data cars on a database, clouds, social media, or within a plain excel file. To do this, analysts must also have a solid grasp of data structure and management, including how and where data is stored, fetched, and maintained. Those working in this field should thus share a passion for facts and data, and understand the basics of data manipulation, as well.

Once it comes time to analyze the data, there are an array of statistical models analysts may choose to utilize. According to Mello, most common techniques will fall into the following two groups:

* Supervised learning, including regression and classification models.
* Unsupervised learning, including clustering algorithms and association rules.

### Regression Models

Data analysts use **regression models** to examine relationships between variables. Regression models are often used by organizations to determine which independent variables hold the most influence over dependent variables—information that can be leveraged to make essential [business decisions](https://www.northeastern.edu/graduate/blog/data-driven-decision-making/).

“The most traditional regression models that have been used for a long time are logistic regression, linear regression, and polynomial regression,” Mello says. “These are the most common.”

Other examples of regression models can include stepwise regression, ridge regression, lasso regression, and elastic net regression.

### Classification Models

Classification is a process in which an algorithm is used to analyze an existing data set of known points. The understanding achieved through that analysis is then leveraged as a means of appropriately classifying the data. Classification is a form of machine learning that can be particularly helpful in analyzing very large, complex sets of data to help make more accurate predictions.

“Classification models are a form of supervised machine learning which is often used when the analyst needs to understand how they got to a certain point,” Mello says. “They give you more than just an output; [they give you] more information that you can use to explain the results of the prediction to your boss or stakeholder.”

Some of the most common classification models include decision trees, random forests, nearest neighbor, and  Naive Bayes.

There are also the neural networking models that are more used in AI. “These are very powerful models, and they can make accurate predictions very well,” Mello says, “but you typically cannot explain what is happening behind the scenes.”

**Types of the Model:**

Here it is the regression modeling technique use because in this dataset **selling\_price** is target variable and in this column all continuous values are present so use regression models.

**Problem Definition:**

With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model.

**Data Analysis:**

The first and for most step involves importing necessary libraries and packages and loading the dataset as a pandas dataframe. Therefore, we load dataset for further analysis.

**Introduction**

Determining whether the listed price of a car is a challenging task, due to the many factors that drive a used vehicle’s price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models . We will compare the performance of various machine learning algorithms like Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor and choose the best out of it. Depending on various parameters we will determine the price of the car. Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value because of which it will be possible to predict the actual price a car rather than the price range of a car. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user’s inputs.

**Data Loading and Visualisations:**

The first and foremost step involves importing necessary libraries and packages and loading the dataset as a pandas dataframe. Data visualization is the graphical representation of information and data. By using [visual elements like charts, graphs, and maps](https://www.tableau.com/learn/articles/data-visualization/glossary), data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Our eyes are [drawn to colors and patterns](https://www.tableau.com/learn/whitepapers/tableau-visual-guidebook). We can quickly identify red from blue, square from circle. Our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we [quickly see trends and outliers](https://www.tableau.com/reports/business-intelligence-trends). If we can see something, we internalize it quickly. It’s storytelling with a purpose. If you’ve ever stared at a massive spreadsheet of data and couldn’t see a trend, you know how much more effective a visualization can be.

**Importing libraries**

We will start by importing the libraries we will require for performing EDA. These include NumPy, Pandas, Matplotlib, and Seaborn.

### Reading data:

We will now read the data from a CSV file into a Pandas DataFrame in this there are two dataset first is train dataset and second is test dataset.

Install This project requires anaconda python, because below libraries already available.

• Numpy

• Matplotlib

• Seaborn

• Skit-learn

• Pandas

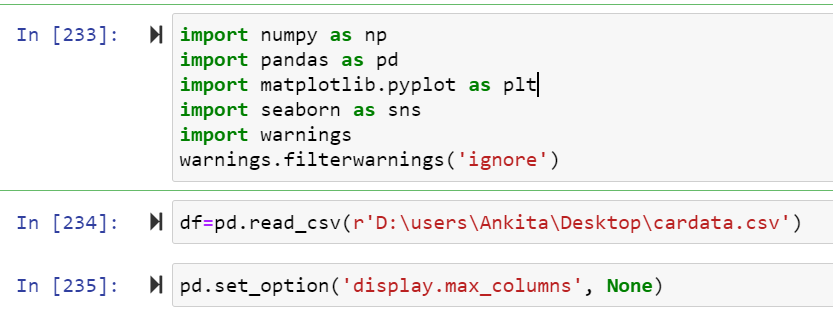
Also need to have software, Install, run and execute a Jupyter notebook. .

**EDA**:

There are no shortcuts in a machine learning project lifecycle. We can’t simply skip to the model building stage after gathering the data. We need to plan our approach in a structured manner and the exploratory data analytics (EDA) stage plays a huge part in that. I can say this with the benefit of hindsight having personally gone through this situation plenty of times. In my early days in this field, I couldn’t wait to dive into machine learning algorithms but that often left my end result hanging in the balance. I discovered, through personal experience and the advice of my mentors, the importance of spending time exploring and understanding my data.

## The Importance of Exploratory Data Analysis (EDA): There are no shortcuts in a machine learning project lifecycle. We can’t simply skip to the model building stage after gathering the data. We need to plan our approach in a structured manner and the exploratory data analytics (EDA) stage plays a huge part in that. I can say this with the benefit of hindsight having personally gone through this situation plenty of times. In my early days in this field, I couldn’t wait to dive into machine learning algorithms but that often left my end result hanging in the balance. I discovered, through personal experience and the advice of my mentors, the importance of spending time exploring and understanding my data.

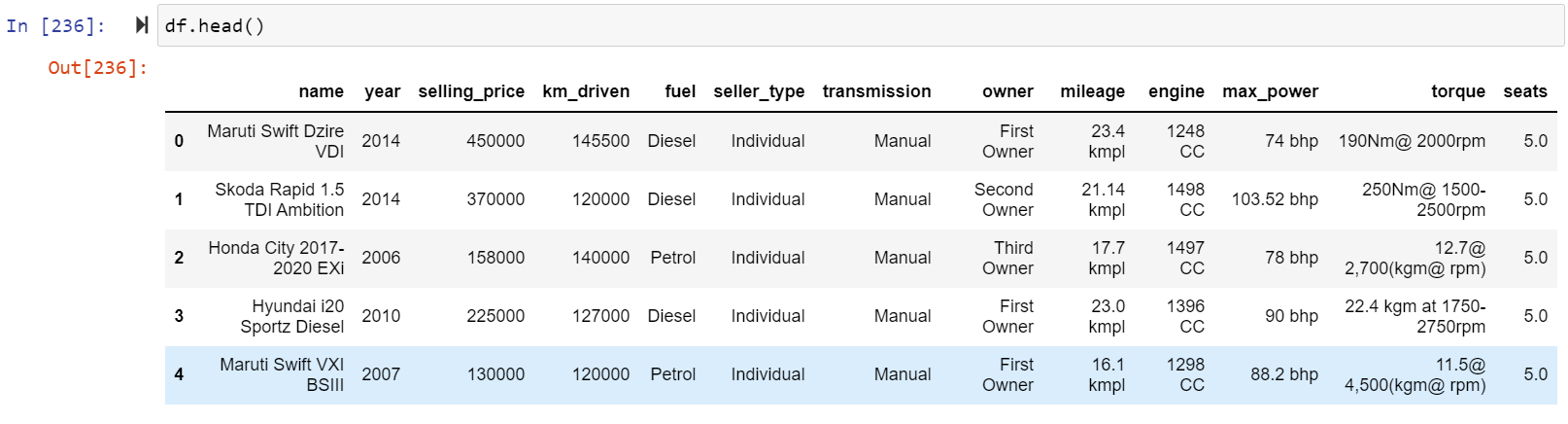
**Loading dataset :**



Here importing all necessary library and also load car dataset.

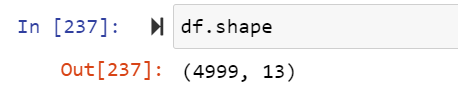
**Data Sources and their formats**

Then need to read and load car dataset. And Then this below code isusefor display all the columns in dataset. Let us have a look at how our dataset. The structure and details of the data are given below in this there are dataset in which no of row and columns are present:



Here selling\_price is our Target Variable in car dataset and in car dataset selling\_price are target variable and there was continues numeric values are

so this is **Regression Problem** so Need use Regression Algorithm.



Check the how much Columns and rows present in dataset using df.shape() is use to check the rows and columns count in this above dataset there are 5000 rows and 13 columns are present in car dataset.

### The algorithms we are going to cover are:

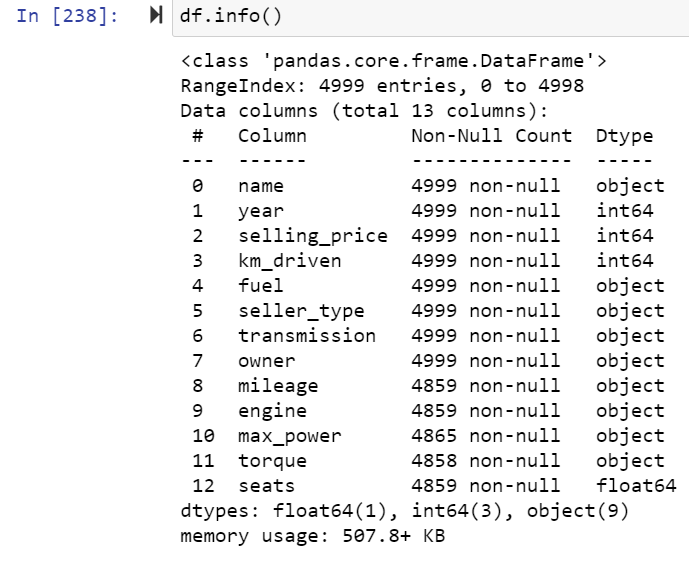
1. Linear Regression

2. Decision TreeRegressor

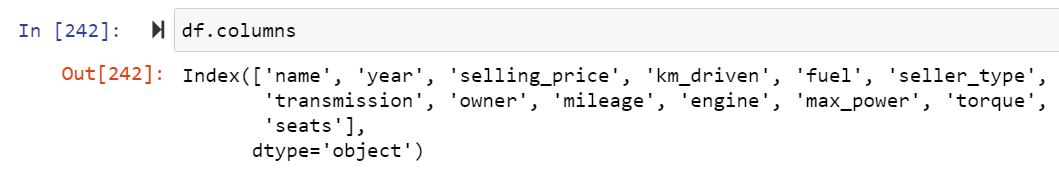
3. Support Vector Regression

4. Lasso Regression

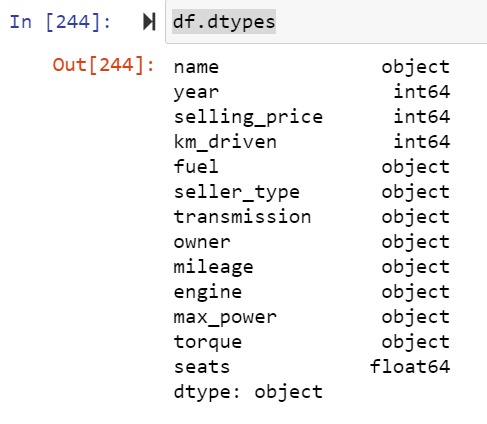
5. Random Forest Regressor



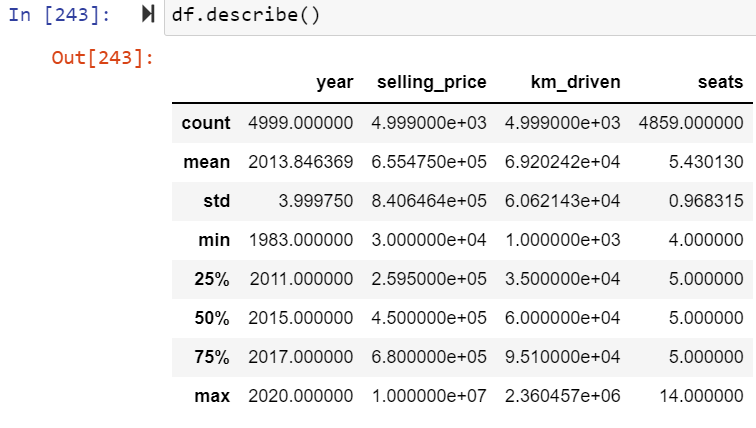
In above using df.info() is used to get the information all the columns in dataset and its data types. In above dataset there are 13 columns in dataset.



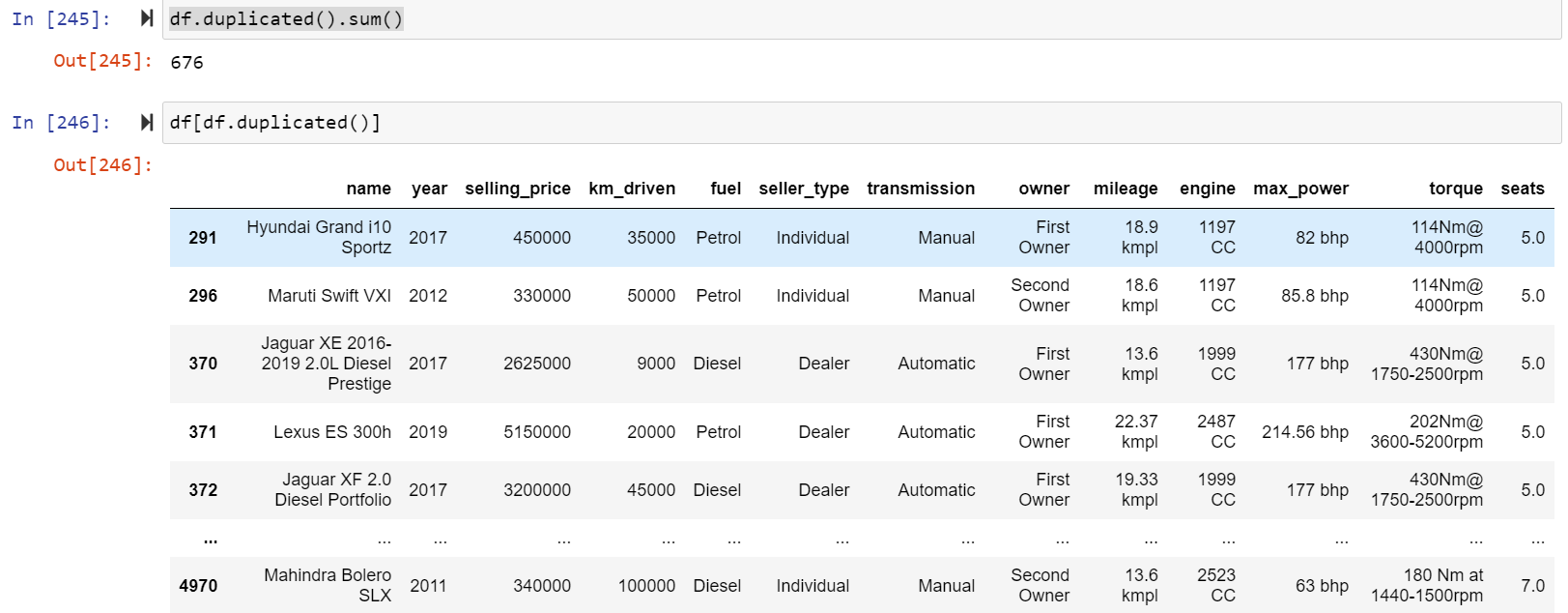
In above df.colums is use to get all the column names in car dataset.

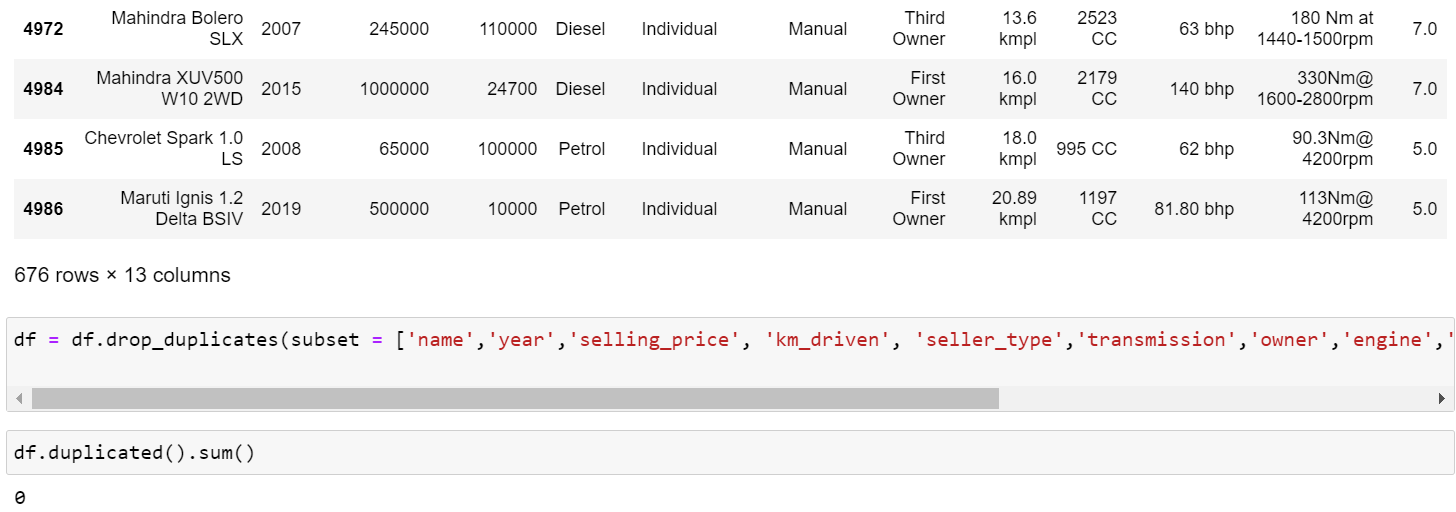


In above code df.dtypes is used to check the data types of all the column in car dataset.

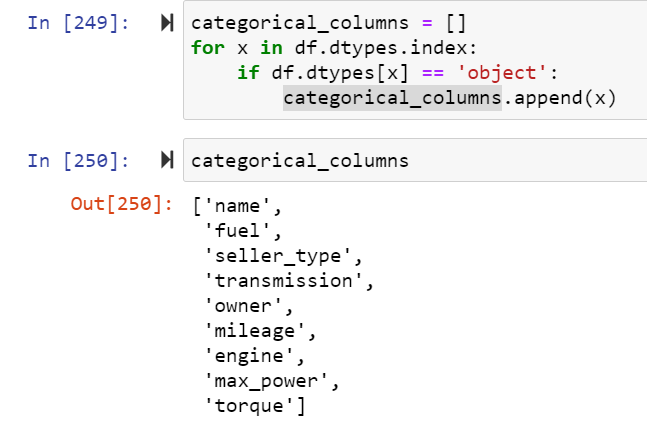


Above describe the information of car dataset using df.describe().

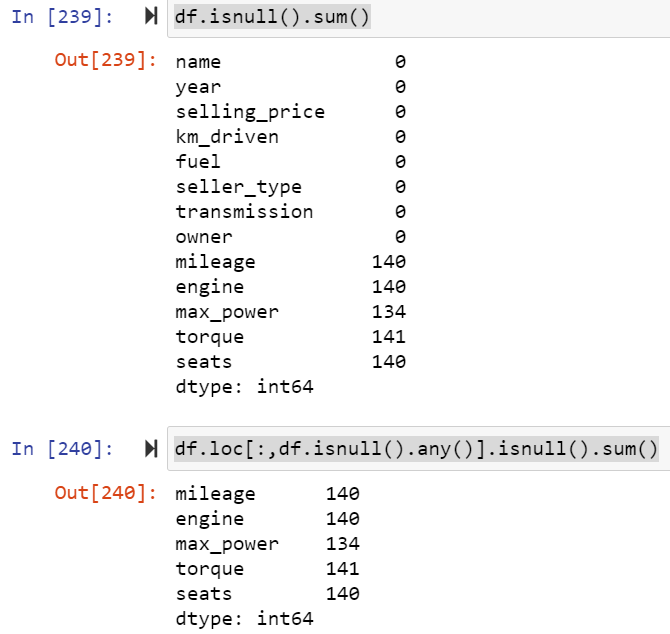




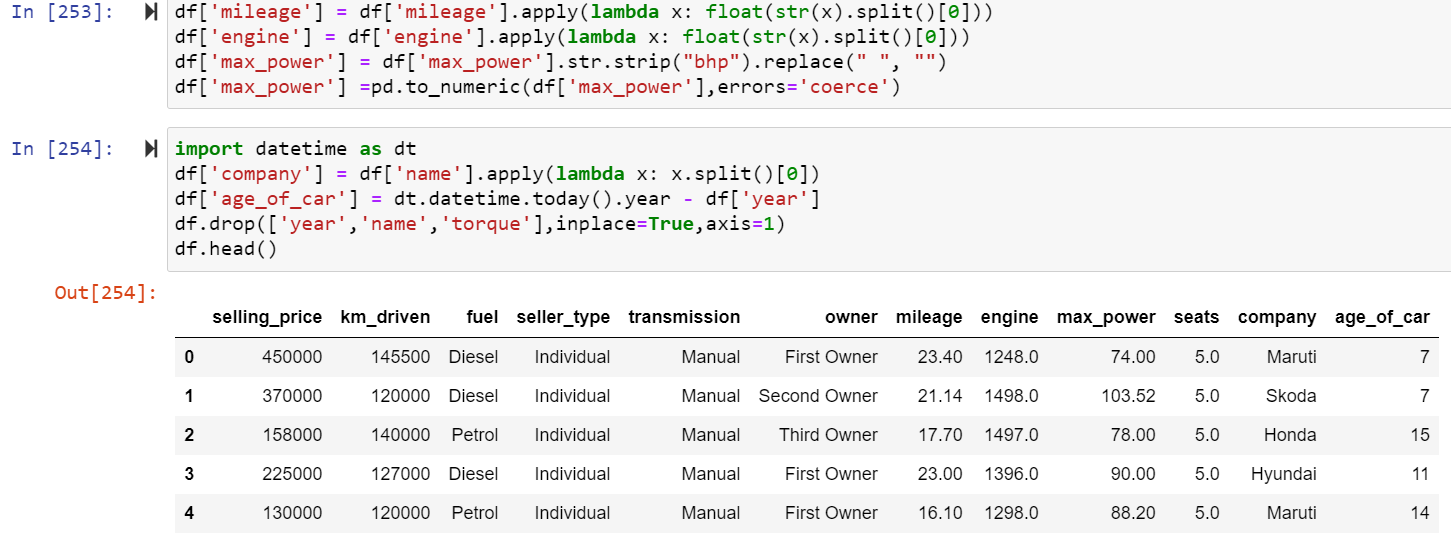
Above check the duplicates values of car dataset there was no of duplicates values are present in dataset so drop that duplicate values from dataset.



Here above are list of all columns in car dataset display names of column in which data types are object.



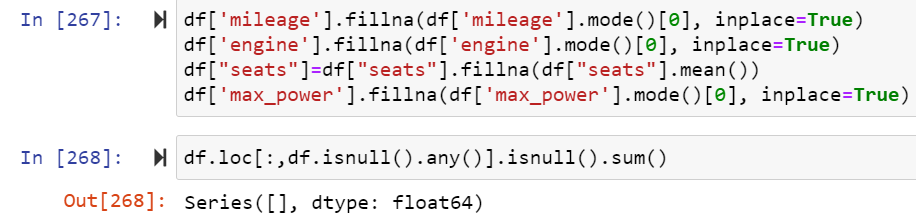
Check the how much null values present in dataset using df.isnull() is use to check the null values in dataset in this dataset there are no of null values are present in car dataset.



Here above ‘milenge’, ‘engine’, ‘max\_power’ columns are categorical and numerical values are present so split that values and convert it into the numeric and also drop year,name and torque column because it is not much important.

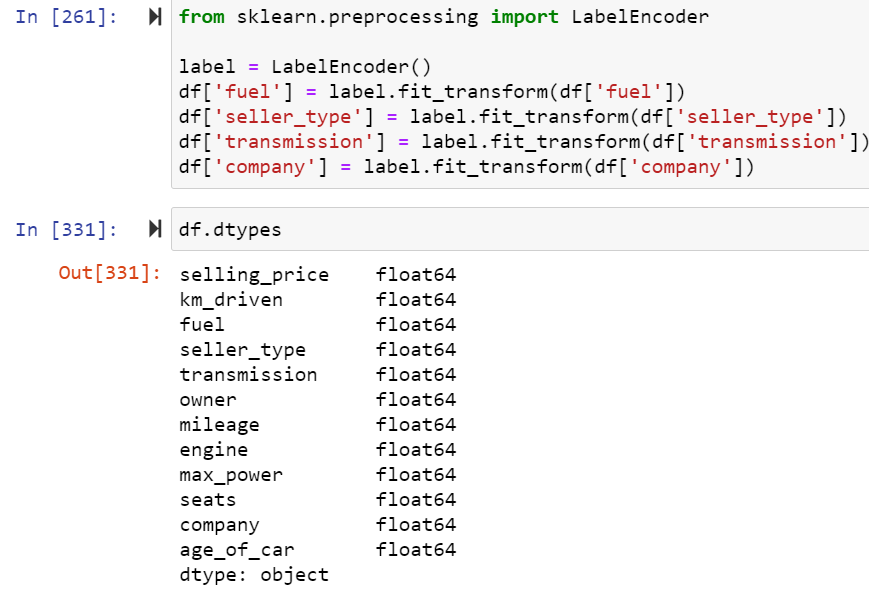
**Data Preprocessing Done**

In order to get a better understanding of the data, we plotted a histogram of the data. We noticed that the dataset had many outliers, primarily due to large price sensitivity of used cars. Typically, models that are the latest year and have low mileage sell for a premium, however, there were many data points that did not conform to this. This is because accident history and condition can have a significant effect on the car’s price. Since we did not have access to vehicle history and condition, we pruned our dataset to three standard deviations around the mean in order to remove outliers.

****

Fill all null values in milenge, engine, seats and max\_power columns in car dataset.

**Data Inputs- Logic- Output Relationships**

 Here use label encoder to replace the categorical values into numeric in car dataset .

**State the set of assumptions (if any) related to the problem under consideration**

The price of a new car in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But, due to the increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. Existing System includes a process where a seller decides a price randomly and buyer has no idea about the car and it’s value in the present day scenario. In fact, seller also has no idea about the car’s existing value or the price he should be selling the car at. To overcome this problem we have developed a model which will be highly effective. Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value. Because of which it will be possible to predict the actual price a car rather than the price range of a car. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user’s inputs.

**Hardware and Software Requirements and Tools Used**

**Hardware requirements**

Operating system- Windows 7,8,10

Processor- dual core 2.4 GHz (i5 or i7 series Intel processor or equivalent AMD)

RAM-4GB

200 GB HDD.

Intel 1.66 GHz Processor Pentium 4

**Software Requirements**

Python Pycharm PIP 2.7

* 1. Coding Language: Python3, Python
  2. Coding software : Anaconda, Jupyter Notebook

1. Microsoft Office Word.
2. Snipping Tools (For Screenshots).
3. Microsoft Excel

Jupyter Notebook Chrome

**Non Functional Requirements:**

1: Platform Independent: The application would be platform independent if all the requirements are installed in the device.

2: Performance: The application should have better accuracy and should provide the information in less time.

3: Capacity: The capacity of the storage should be high so that large amount of data can be stored in order to train the model.

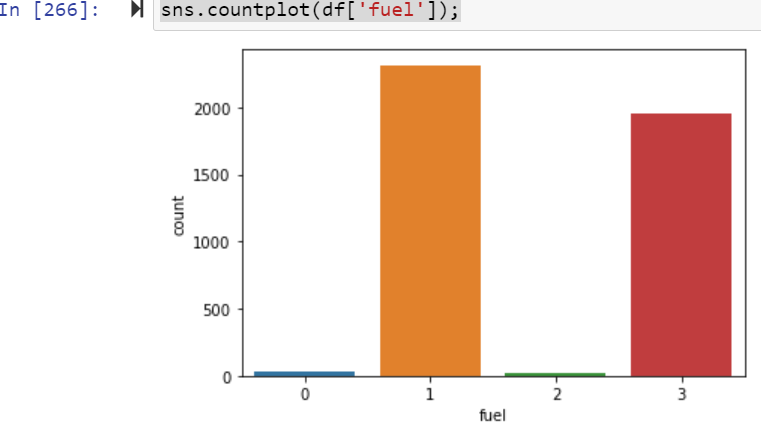
**Graphical representation**

### We will start with Univariate Analysis. We will be using a bar graph for this purpose.

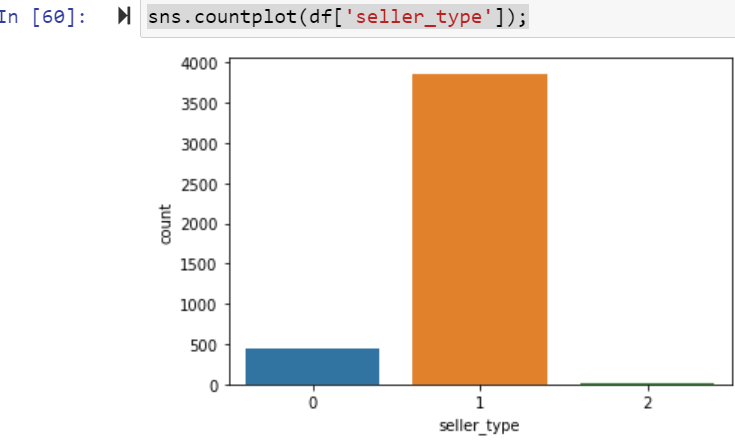
1. Univarient Analysis.
2. Bi Varient Analysis
3. Multivarient Analysis

**Univarient Analysis of Train dataset**:

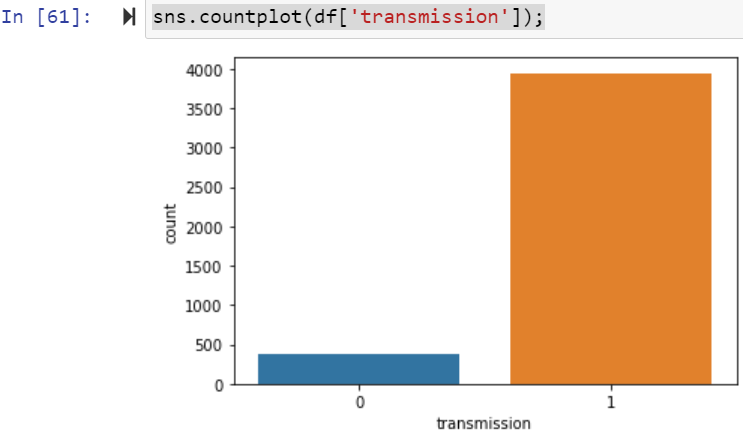
Univariate analysis **explores each variable in a data set**, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable. It describes each variable on its own.

****

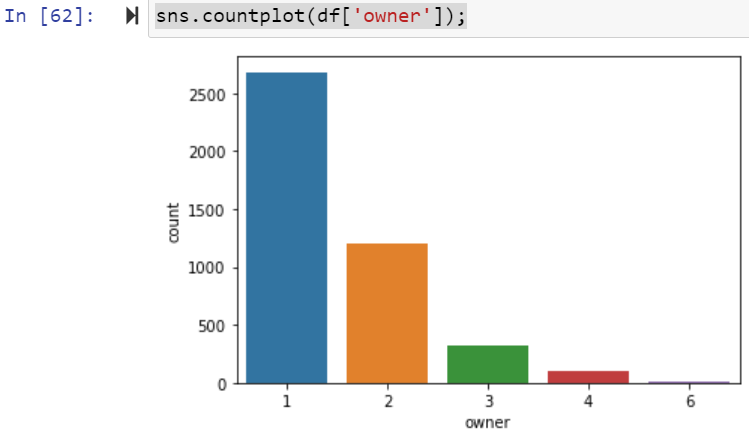
Univarient analysis of ‘fuel’ column in car dataset.

****

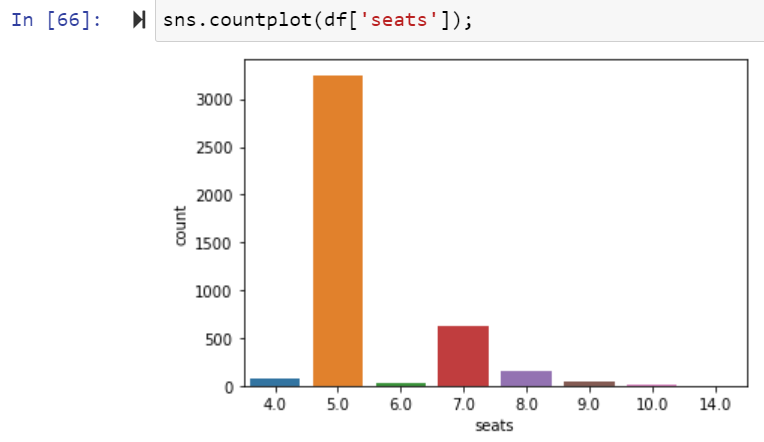
Univarient analysis of ‘seller\_type’ column in car dataset.



Univarient analysis of ‘transmission’ column in car dataset.



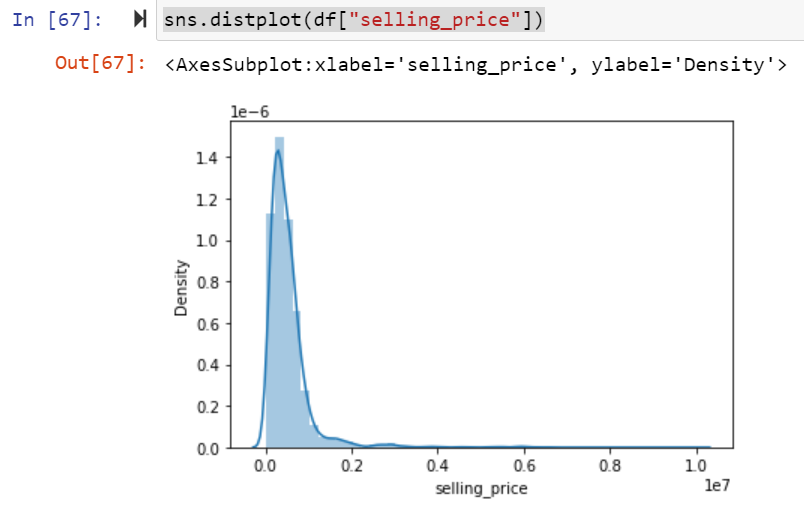
Univarient analysis of ‘owner’ column in car dataset.

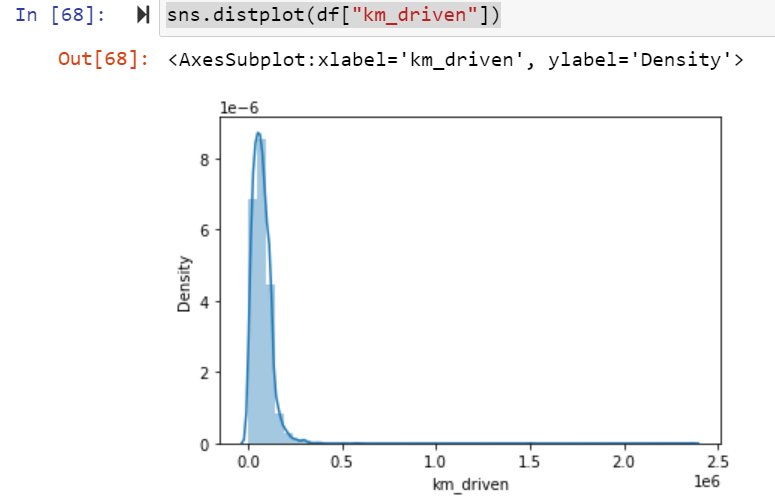


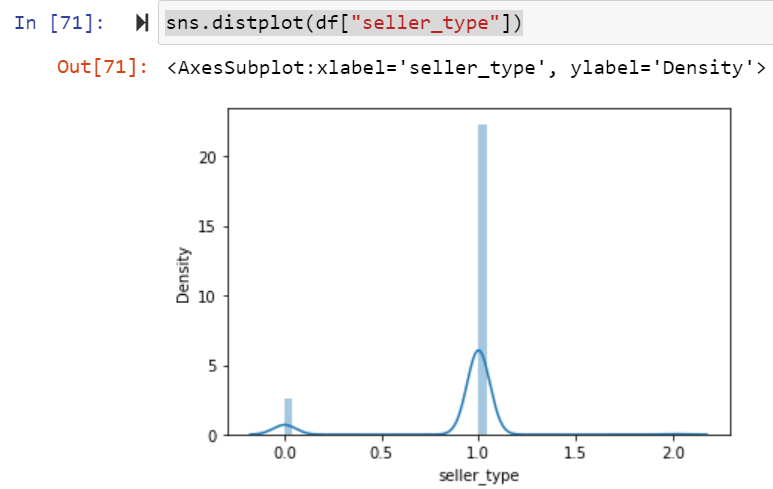
Univarient analysis of ‘seats’ column in car dataset.

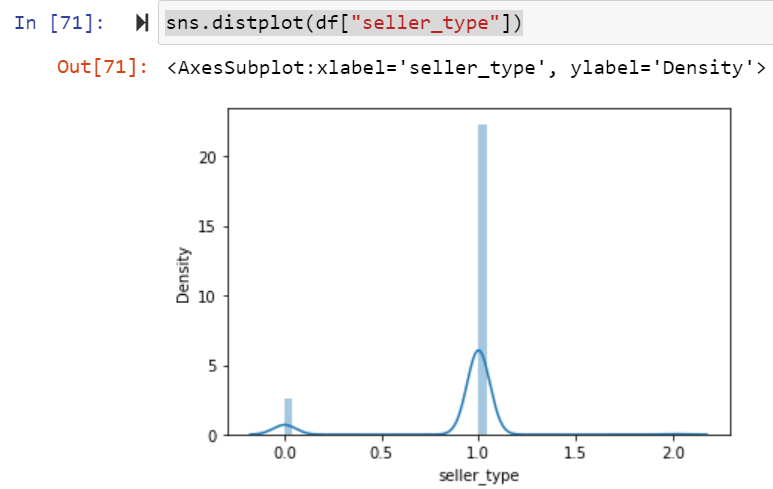
**Checking Distribution of columns in train datset they are normally distributed or not:**

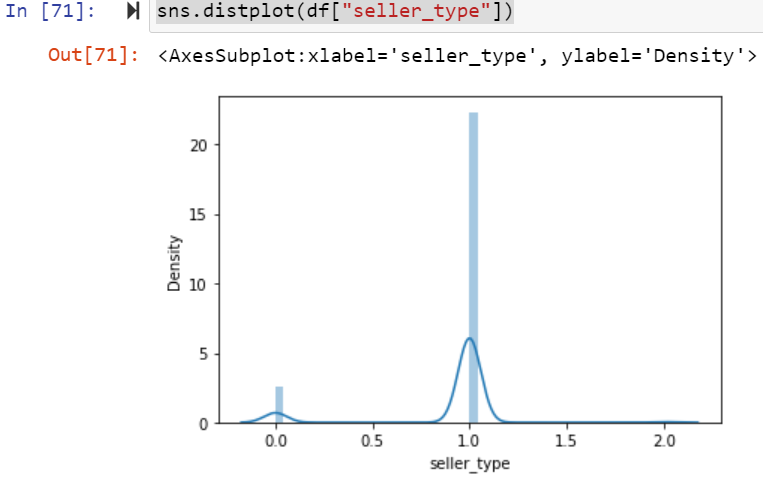
Normally distributed data, there is a constant proportion of data points lying under the curve between the mean and a specific number of standard deviations from the mean. Thus, for a normal distribution, almost all values lie within 3 standard deviations of the mean

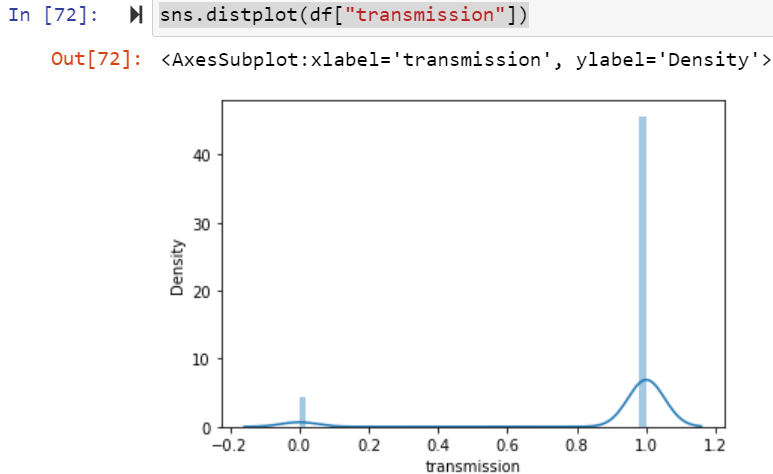


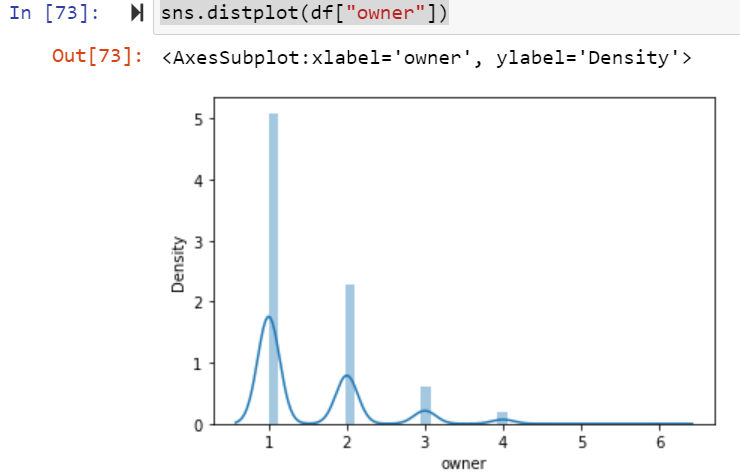


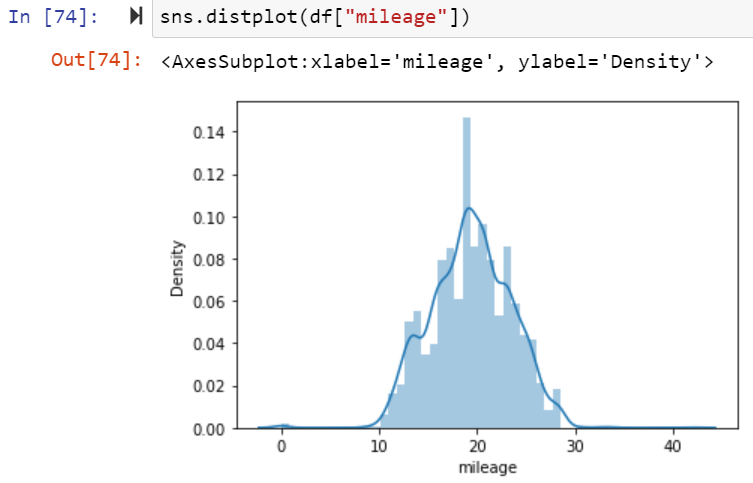
****

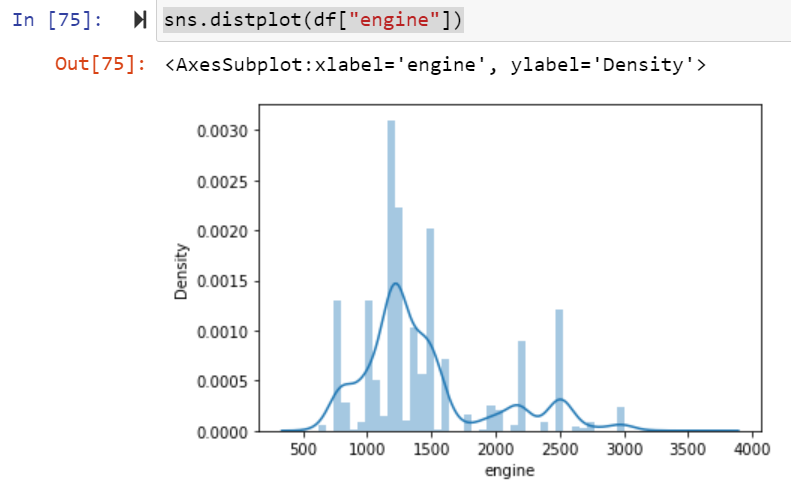
****

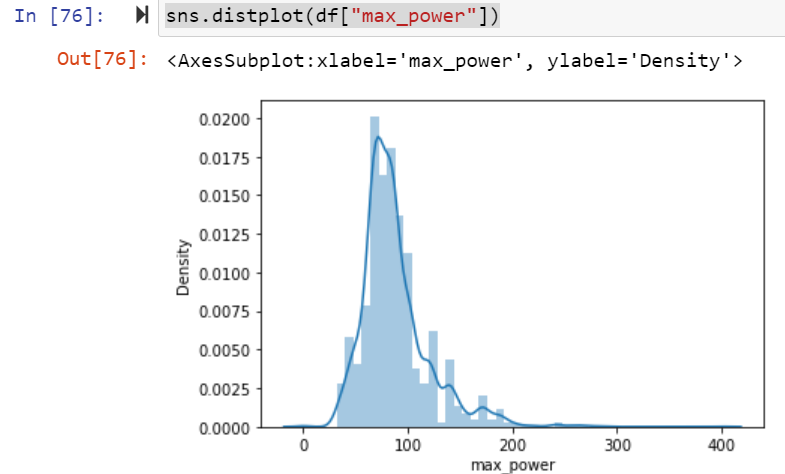
****

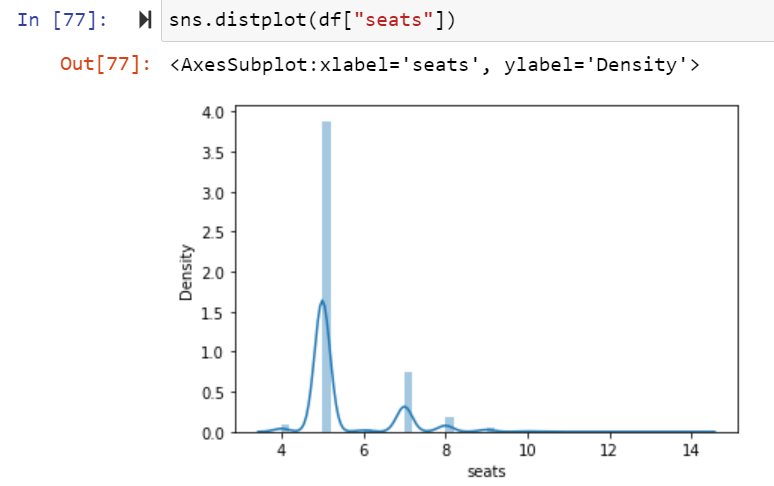
****

****

****



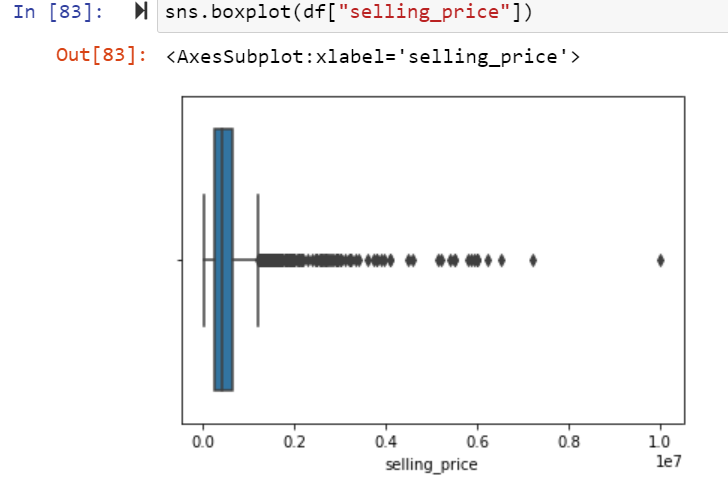


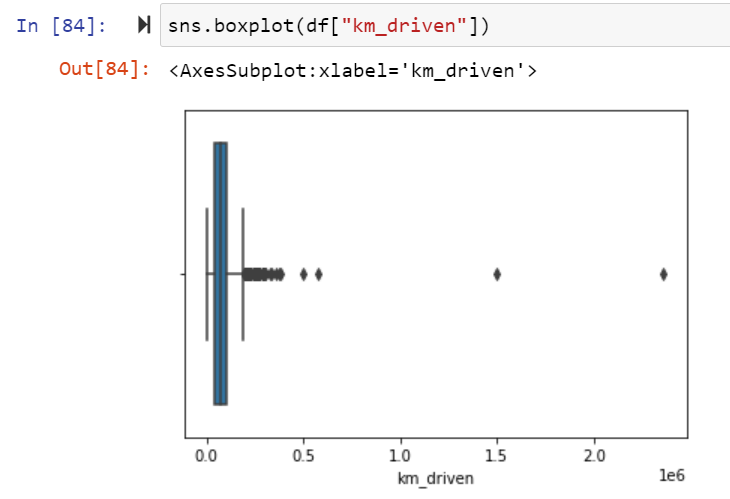


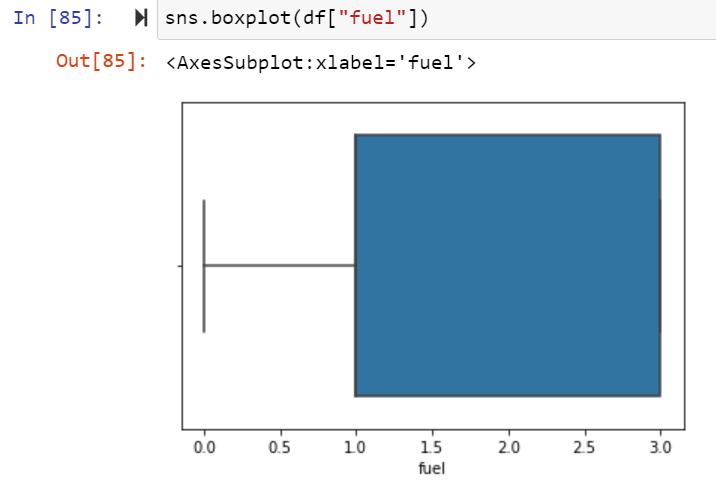
Here some columns are normally distributed and some columns are not normally distributed.

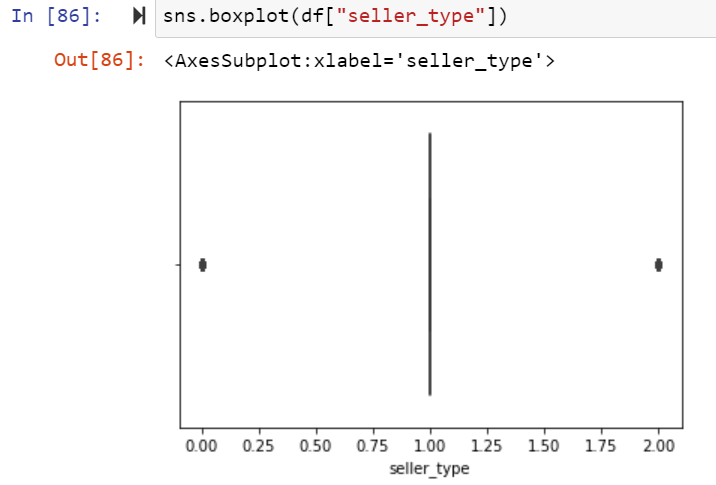
**Checking Outliers:**

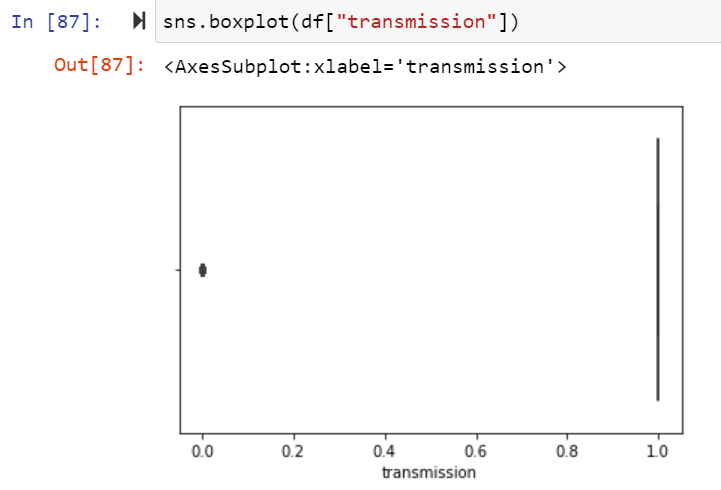
An outlier is an **object(s) that deviates significantly from the rest of the object collection**. It is an abnormal observation during the Data Analysis stage, that data point lies far away from other values. An outlier is an observation that diverges from well-structured data.

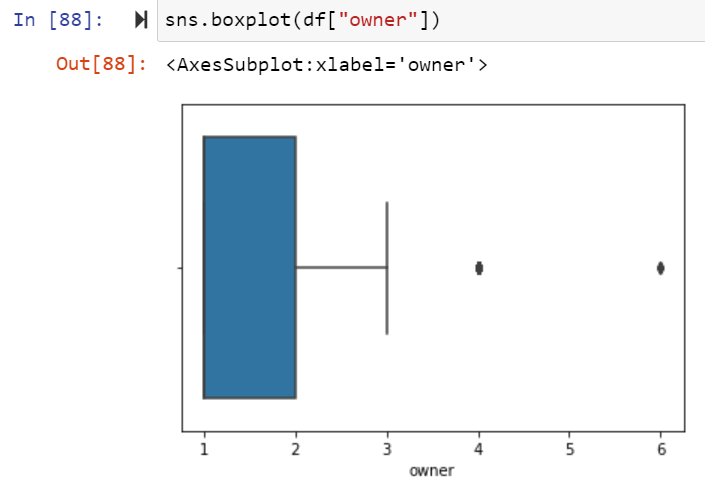
****

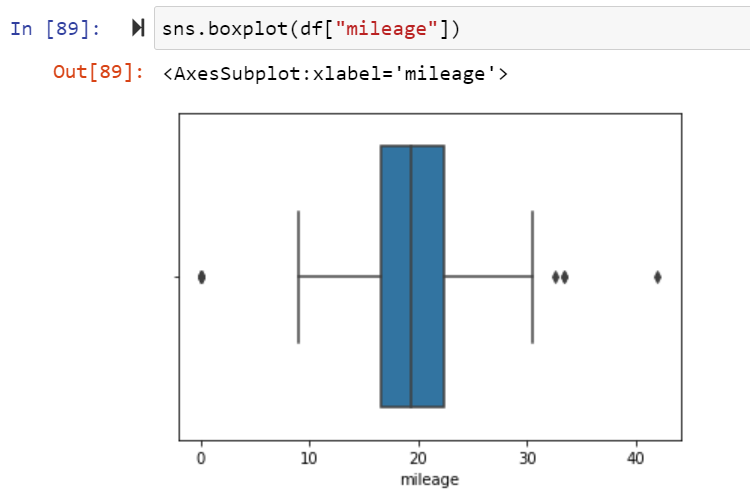
****

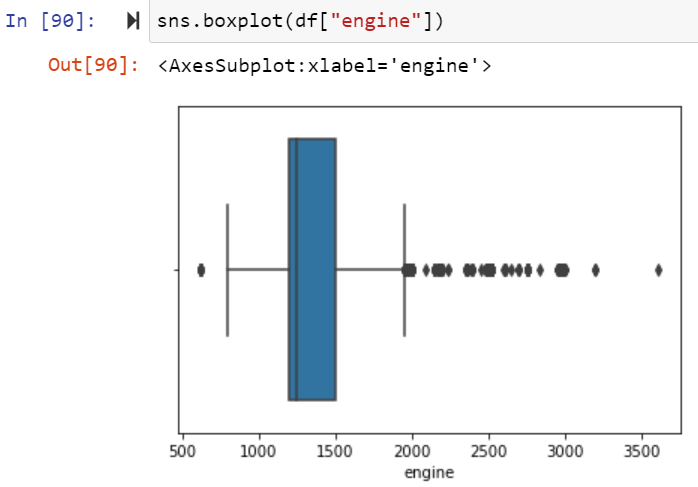


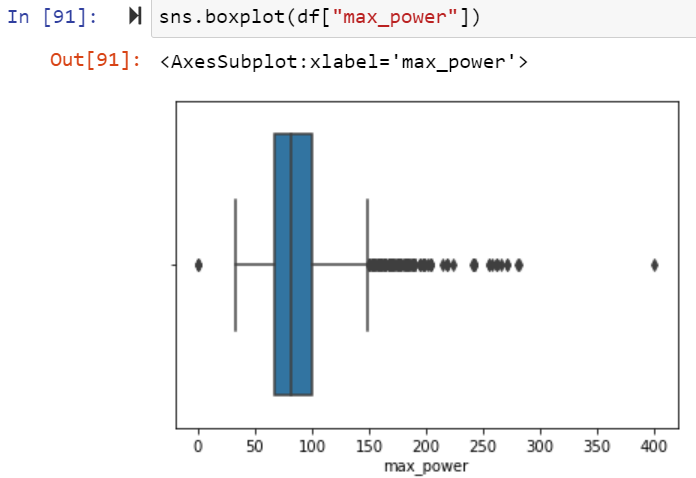


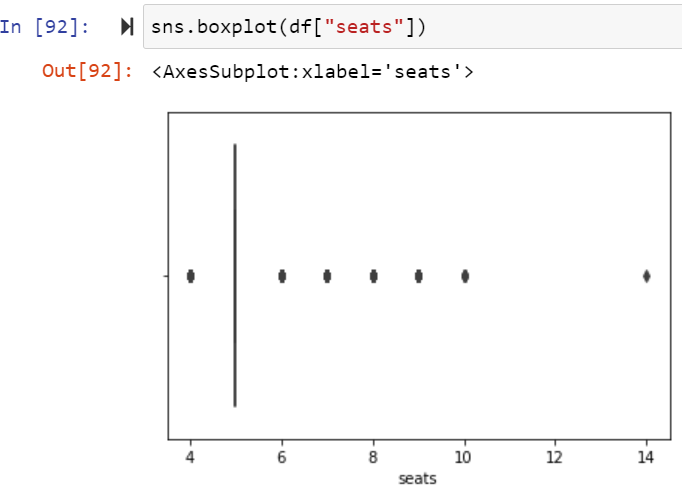




****

****

****

****

Here only ‘fuel’ column outlier is not present but all the remaining columns outliers was present.

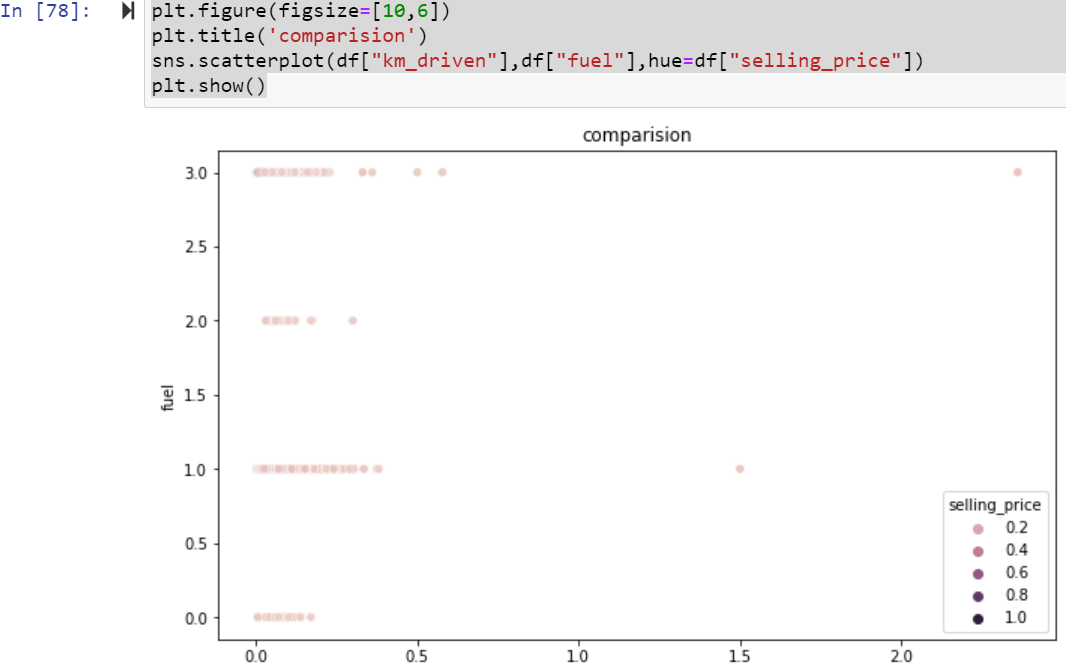
**Bi varient Analysis:**

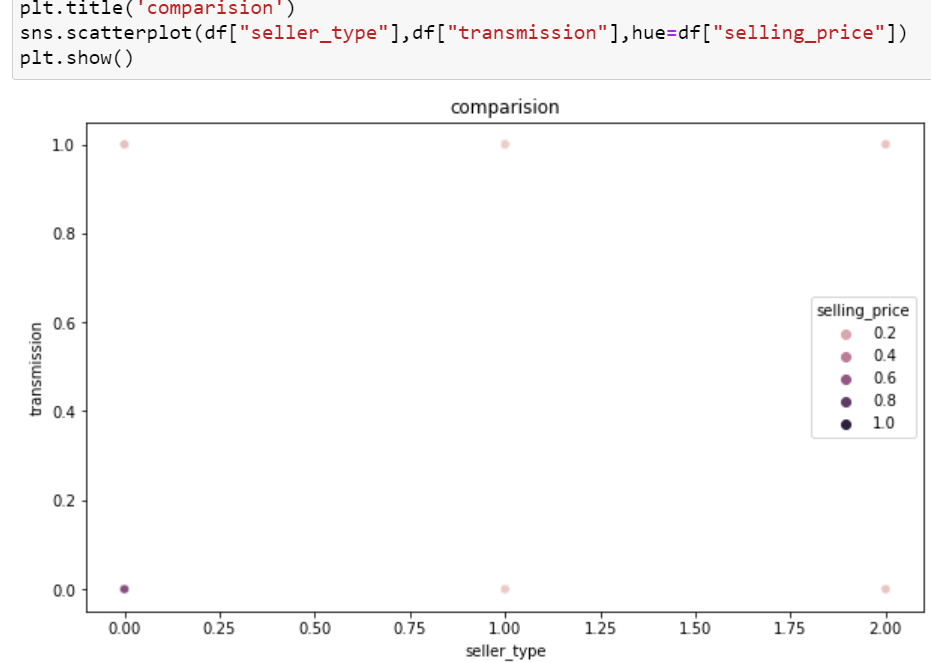
A scatter plot represents individual pieces of data using dots. These plots make it easier to see if two variables are related to each other. The resulting pattern indicates the type (linear or non-linear) and strength of the relationship between two variables.

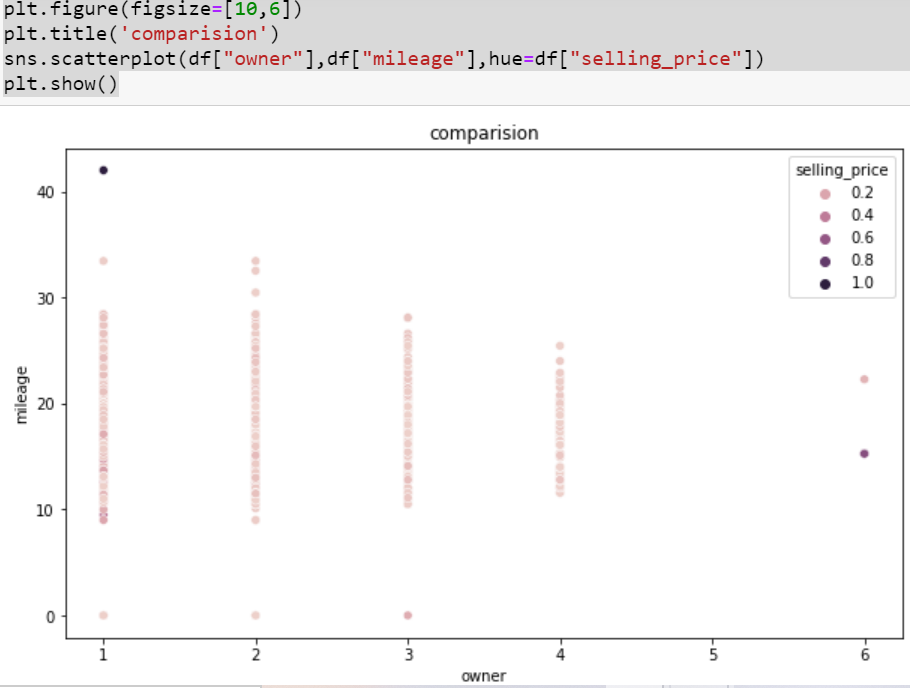
#### Scatter Plot of Car Dataset:

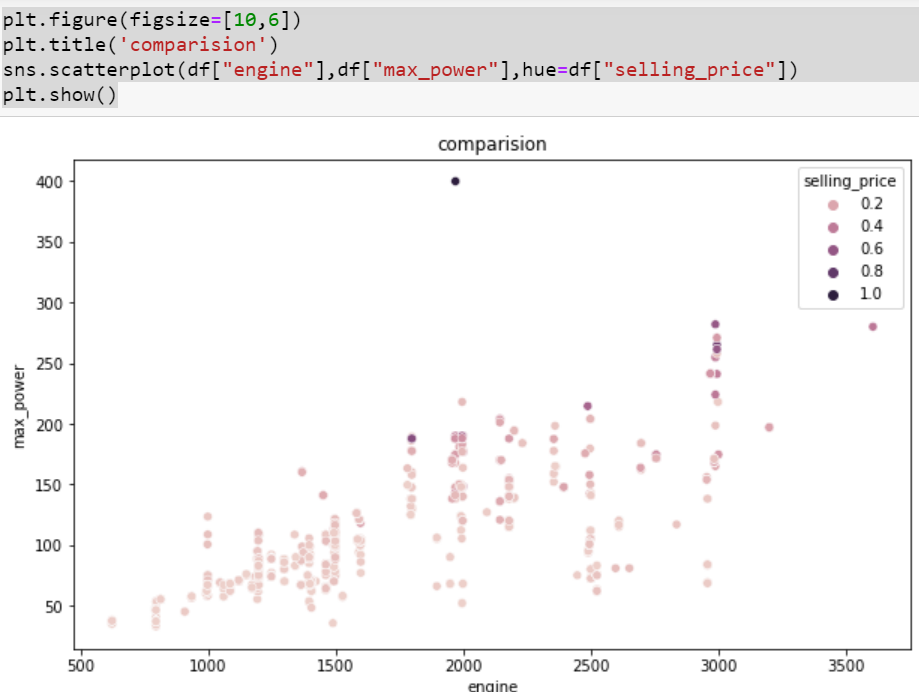
A scatter plot represents individual pieces of data using dots plots make

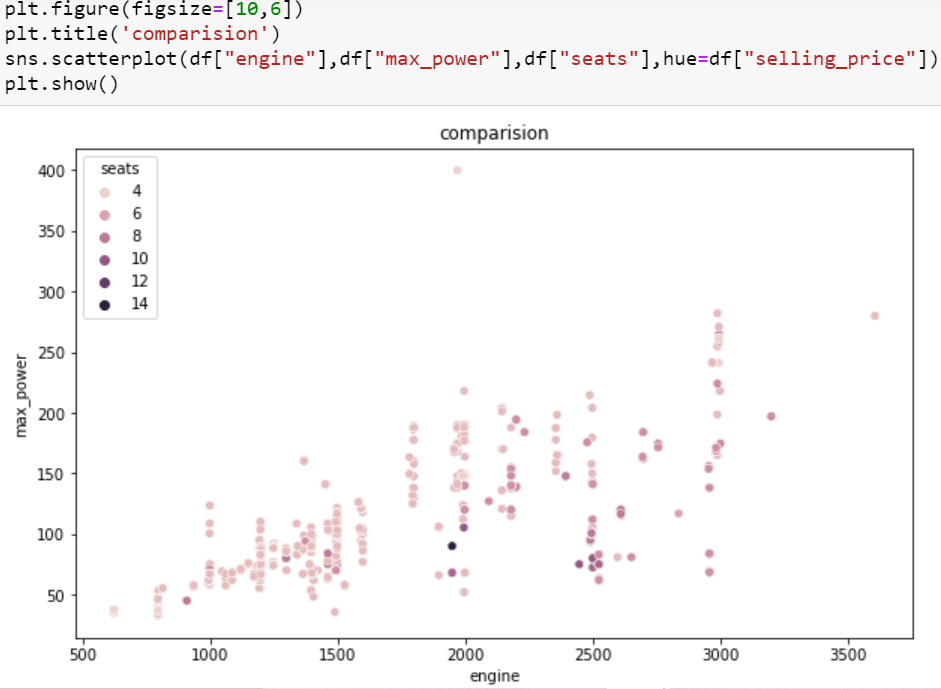
it easierto see if two variables are related to each other. The resulting pattern indicates the type (linear or non-linear) and strength of the relationship between two variables.





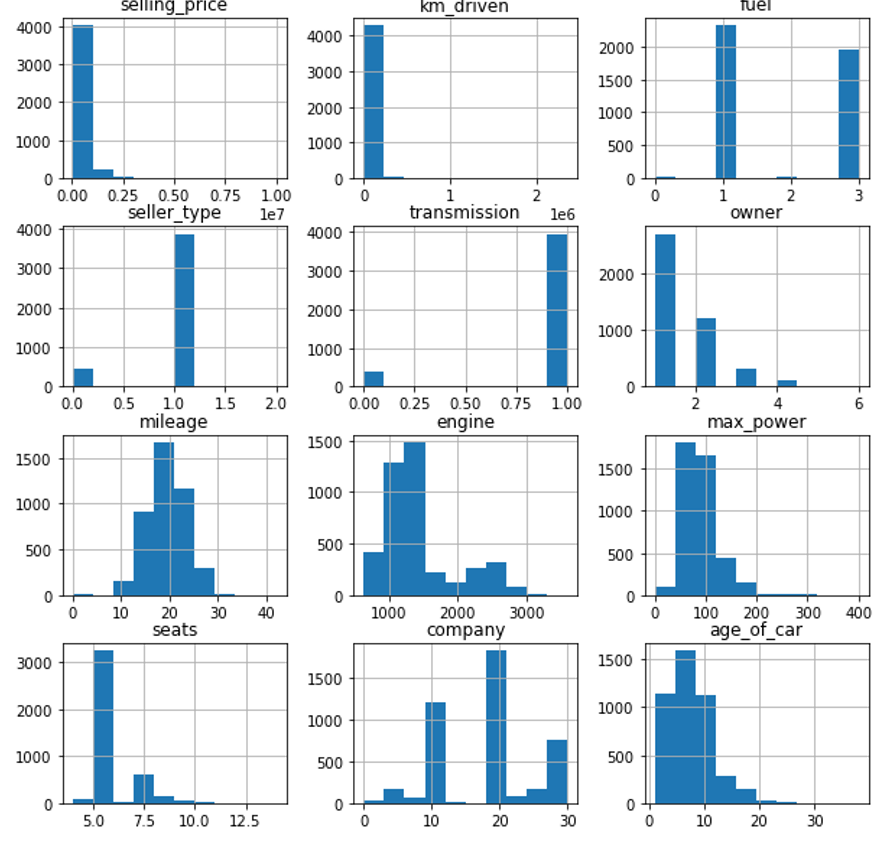
****

****

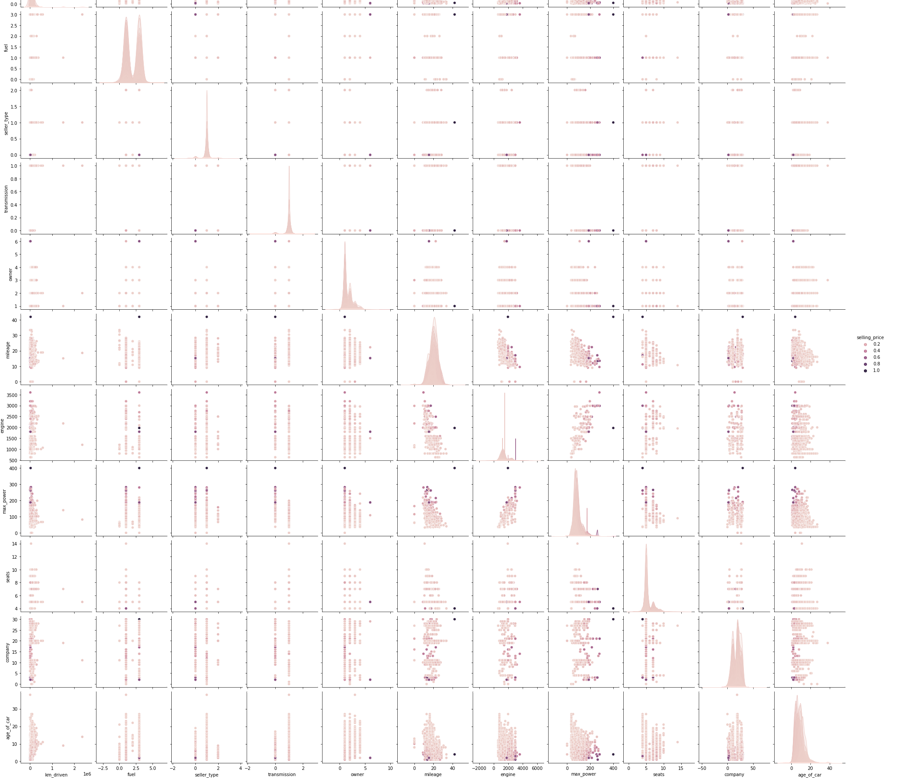
****

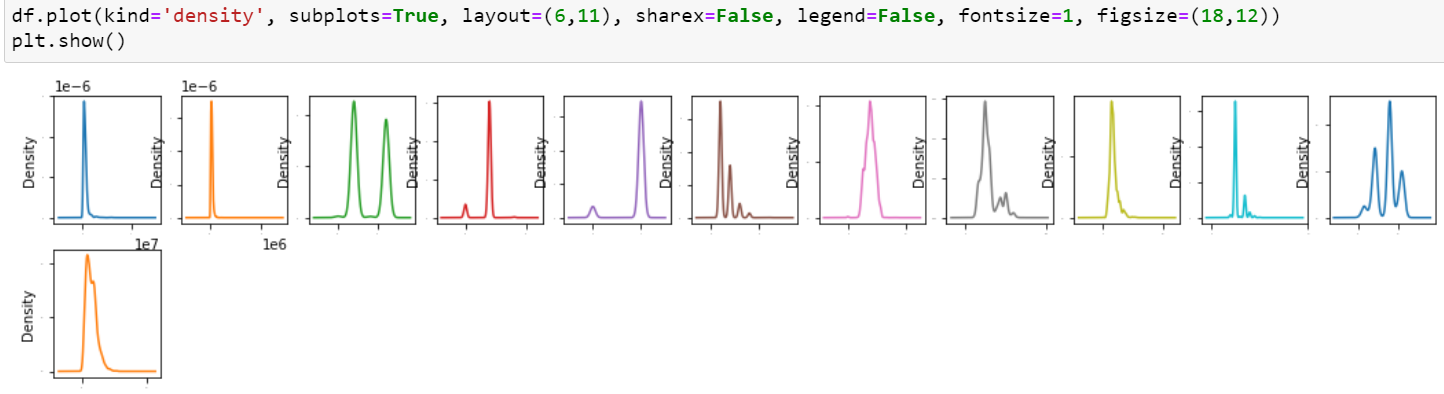
**Histplot Train Dataset**:

A histogram is a classic visualization tool that represents the. A histogram provides **a visual representation of the distribution of a dataset**: location, spread and skewness of the data; it also helps to visualize whether the distribution is symmetric or skewed left or right. ... In brief, a histogram summarizes the distribution properties of a continuous numerical variable.

****

**Multi Varient Analysis of Dataset:**

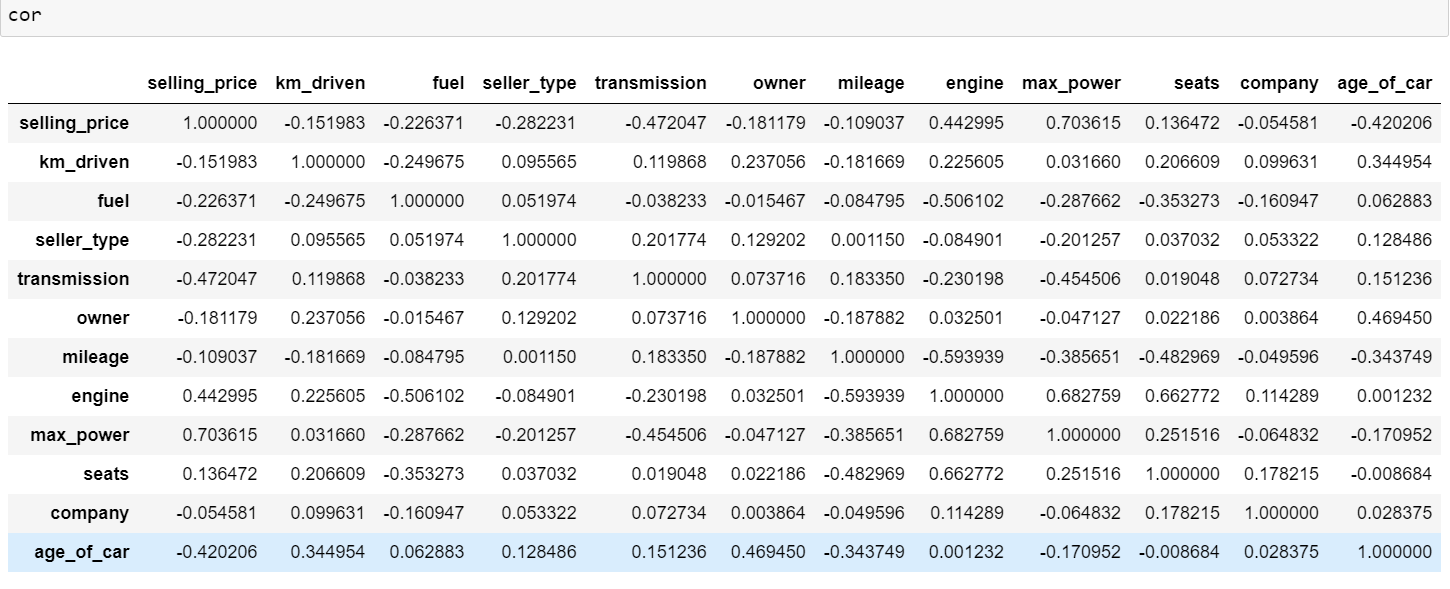
****

****

**Visualizations:**

**Heatmap:**

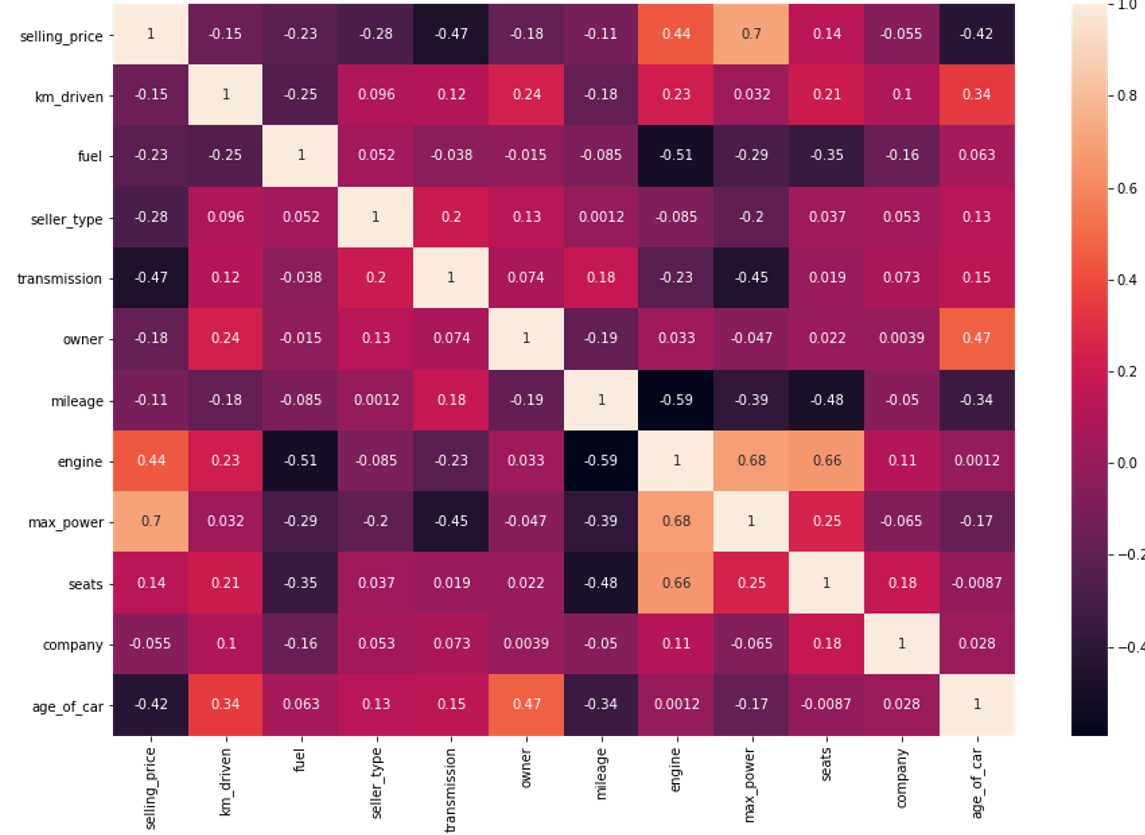
Correlation heatmap is graphical representation of correlation matrix representing correlation between different variables. The value of correlation can take any values from -1 to 1. ... Correlation between two variables can also be determined using scatter plot between these two variables.

****

After that Checking Correlation of all independent columns with Target column.

plt.figure(figsize=(15,10))

sns.heatmap(df.corr(),annot=True)

****

Heat map shows the correlation of every independent variable in dataset with target variable. Here above heatmap the every independent variable check correlation with selling\_price target variable.

We can observe that features are not much correlated in the data. But there is one thing that we can notice is that after log transforming ‘Price’ column, correlation with few features got increased which is a good thing. We will be using log-transformed ‘Price’ to train the model.

plt.figure(figsize=(15,10))

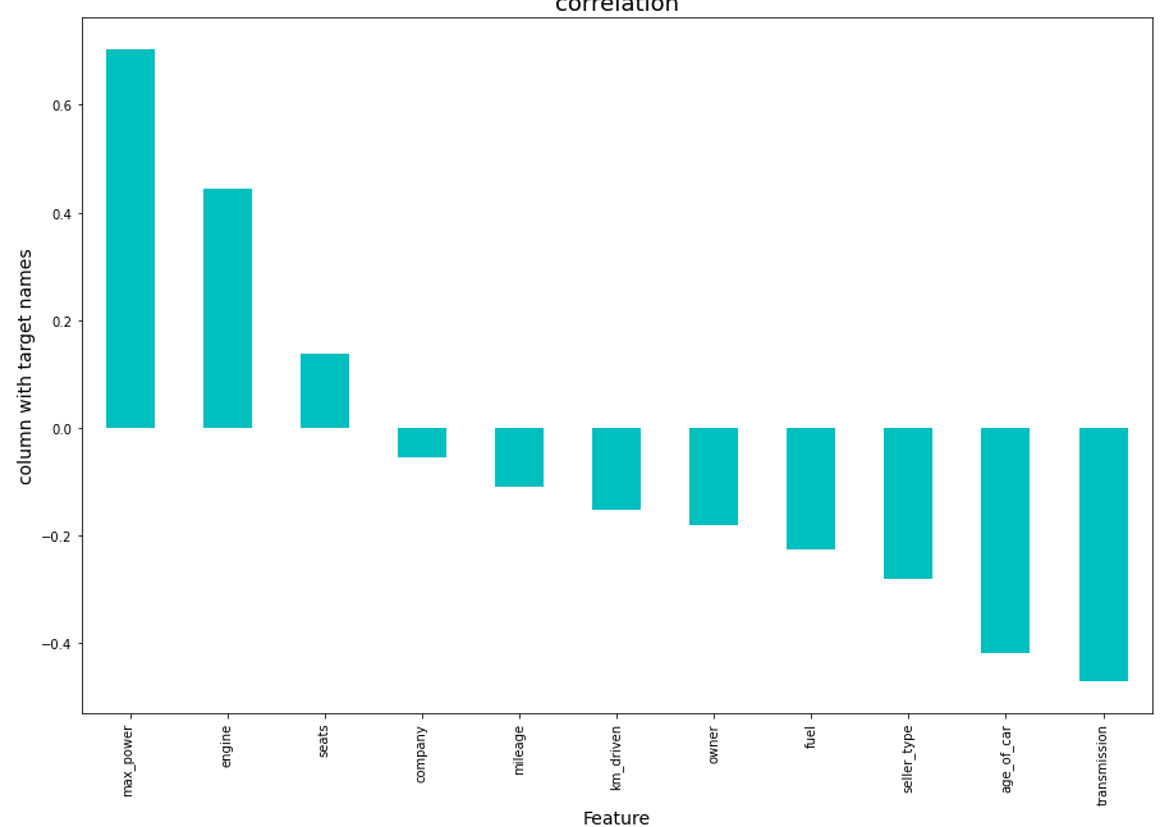
df.corr()['selling\_price'].sort\_values(ascending=False).drop(['selling\_price']).plot(kind='bar',color='c')

plt.xlabel('Feature',fontsize=14)

plt.ylabel('column with target names',fontsize=14)

plt.title('correlation',fontsize=18)

plt.show()



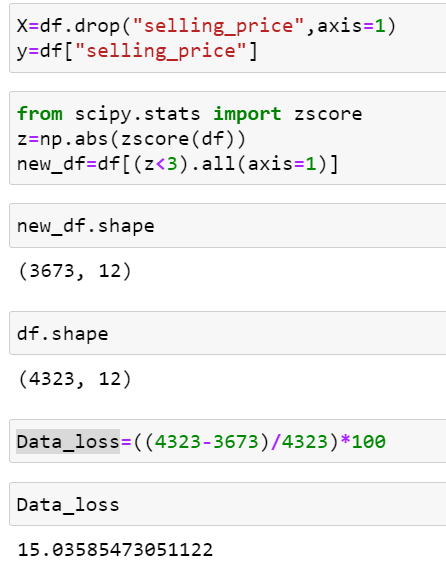
Here we shows the correlation in another way in this above we show some columns are positively and some columns are negatively correlated with target variable.

**State the set of assumptions (if any) related to the problem under consideration:**

In this we analyse the accuracy of predicting house prices when using Multiple linear, Lasso, Ridge, Random Forest regression algorithms. Thus, the purpose of this is to depend the knowledge in regression methods in machine learning. In addition, the given datasets should be processed to enhance performance, which is accomplished by identifying the necessary features by applying one of the selection methods 2 to eliminate the unwanted variables since each car has its unique features that help to estimate its price. These features may or may not be shared with all cars, which means they do not have the same influence on the car pricing resulting in inaccurate output.

**Z-score:**

Take your data point, subtract the mean from the data point, and then divide by your standard deviation. That gives you your Z-score. You can use Z-Score to determine outliers.One of the most commonly used tools in determining outliers is the Z-score. Z-score is just the number of standard deviations away from the mean that a certain data point is.In your future data science life, Z-scores are gonna be a really useful way to think about how usual or how unusual a certain data point is. And that’s going to be really valuable once we start making inferences based on our data. In this story, we will take a deep dive into our notebooks and learn how to detect outliers using Z-Score.

****

After that Cheking Data loss of the dataset after using preprocessing steps here using zscore method checking dataloss here after zscore there 15 percent dataloss.

**IQR:**

The interquartile range rule is useful in detecting the presence of outliers. [Outliers](https://www.thoughtco.com/what-is-an-outlier-3126227) are individual values that fall outside of the overall pattern of a data set. This definition is somewhat vague and subjective, so it is helpful to have a rule to apply when determining whether a data point is truly an outlier—this is where the interquartile range rule comes in.

## What Is the Interquartile Range?

Any set of data can be described by its [five-number summary](https://www.thoughtco.com/what-is-the-five-number-summary-3126237). These five numbers, which give you the information you need to find patterns and outliers, consist of (in ascending order):

* The minimum or lowest value of the dataset
* The first quartile Q1, which represents a quarter of the way through the list of all data
* The [median](https://www.thoughtco.com/what-is-the-median-3126370) of the data set, which represents the midpoint of the whole list of data
* The third quartile Q3, which represents three-quarters of the way through the list of all data
* The maximum or highest value of the data set.

These five numbers tell a person more about their data than looking at the numbers all at once could, or at least make this much easier. For example, the [range](https://www.thoughtco.com/what-is-the-range-in-statistics-3126248), which is the minimum subtracted from the maximum, is one indicator of how spread out the data is in a set (note: the range is highly sensitive to outliers—if an outlier is also a minimum or maximum, the range will not be an accurate representation of the breadth of a data set).

Range would be difficult to extrapolate otherwise. Similar to the range but less sensitive to outliers is the interquartile range. The [interquartile range](https://www.thoughtco.com/what-is-the-interquartile-range-3126245) is calculated in much the same way as the range. All you do to find it is subtract the first quartile from the third quartile:

IQR = Q3 – Q1.

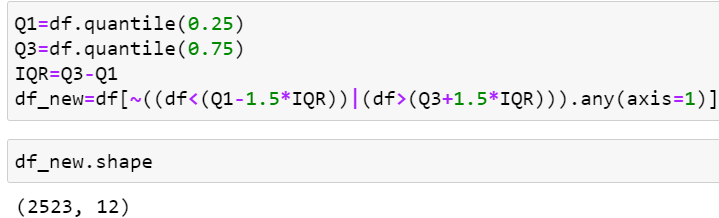
The interquartile range shows how the data is spread about the median. It is less susceptible than the range to outliers and can, therefore, be more helpful.

## Using the Interquartile Rule to Find Outliers

Though it's not often affected much by them, the interquartile range can be used to detect outliers. This is done using these steps:

1. Calculate the interquartile range for the data.
2. Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
3. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
4. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

Remember that the interquartile rule is only a rule of thumb that generally holds but does not apply to every case. In general, you should always follow up your outlier analysis by studying the resulting outliers to see if they make sense. Any potential outlier obtained by the interquartile

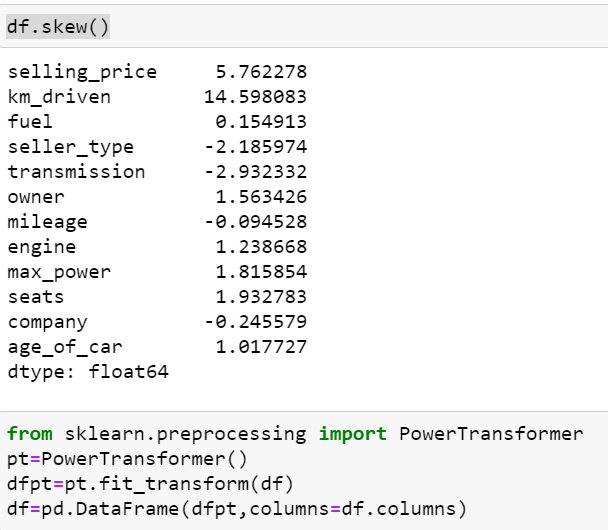


Here After Applying IQR method checking dataloss here using IQR method more dataloss.

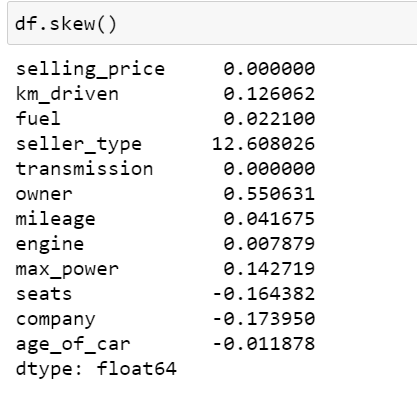
**Skewness:**

The skewness is a measure of symmetry or asymmetry of data distribution, and kurtosis measures whether data is heavy-tailed or light-tailed in a normal distribution. Data can be positive-skewed (data-pushed towards the right side) or negative-skewed (data-pushed towards the left side).

**Checking skewness in dataset:**







To remove skewness use PowerTransformer and log transformer technique in dataset.

**Model/s Development and Evaluation**

**Identification of possible problem-solving approaches (methods)**

**Approach**

Importing the required libraries and reading the dataset.

Merging of the two datasets

* Understanding the dataset

1. Exploratory Data Analysis (EDA) –

* Data Visualization

1. Feature Engineering

* Duplicate value removal
* Missing value imputation
* Encoding of categorical variables
* Dropping of redundant feature columns
* Check for the outliners and removal of outliers.

1. Model Building

* Performing train test split
* Feature Scaling
* Linear Regression Model
* Ridge Regression
* Lasso Regressor
* Gradient Boosting Regressor
* Decision Tree Regressor
* Random Forest Regressor

1. Model Validation

* R2 square error

1. Hypermeter Tuning (GridSearchCV)

For Random Forest Regressor

1. Checking for Feature Importance
2. Creating the final model and making predictions

**Testing of Identified Approaches (Algorithms)**

* LinearRegression, Lasso, Ridge
* SVR
* DecisionTreeRegressor
* RandomForestRegressor
* AdaBoostRegressor
* GradientBoostingRegressor

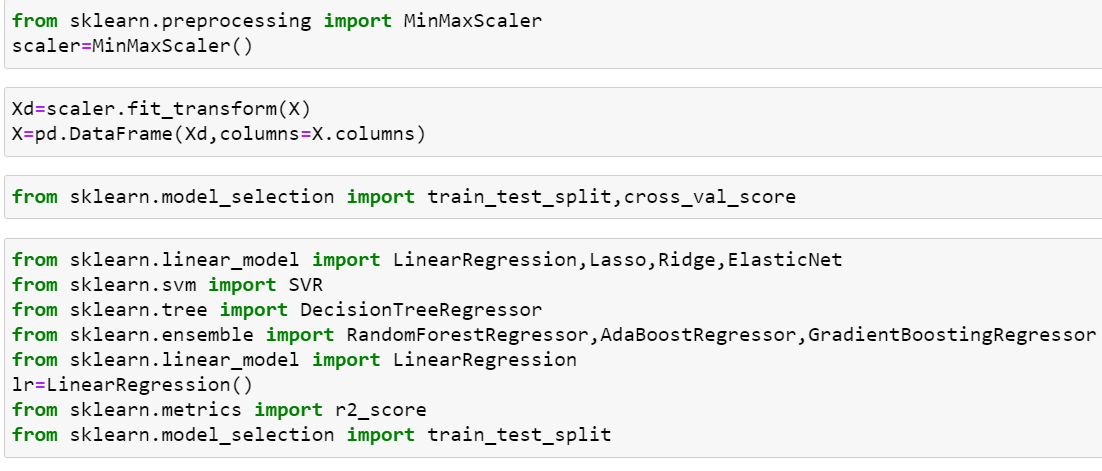
**Building Machine Learning Models:**

Run and Evaluate selected models:

## MinMaxScaler Transform

We can apply the MinMaxScaler to the dataset directly to normalize the input variables.

We will use the default configuration and scale values to the range 0 and 1. First, a MinMaxScaler instance is defined with default hyperparameters. Once defined, we can call the fit\_transform() function and pass it to our dataset to create a transformed version of our dataset.



**Testing of Identified Approaches (Algorithms)**

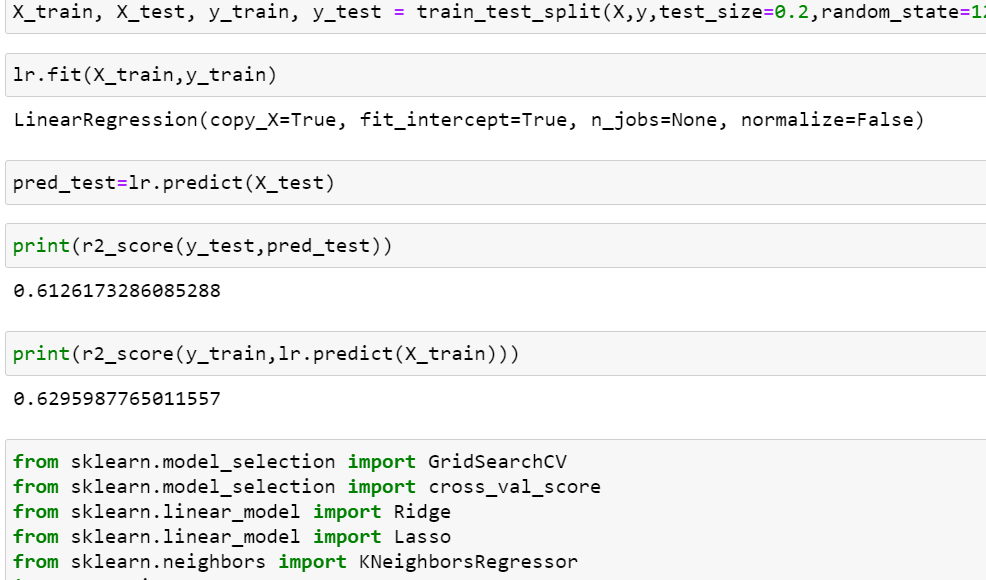
**Lasso Regression:**

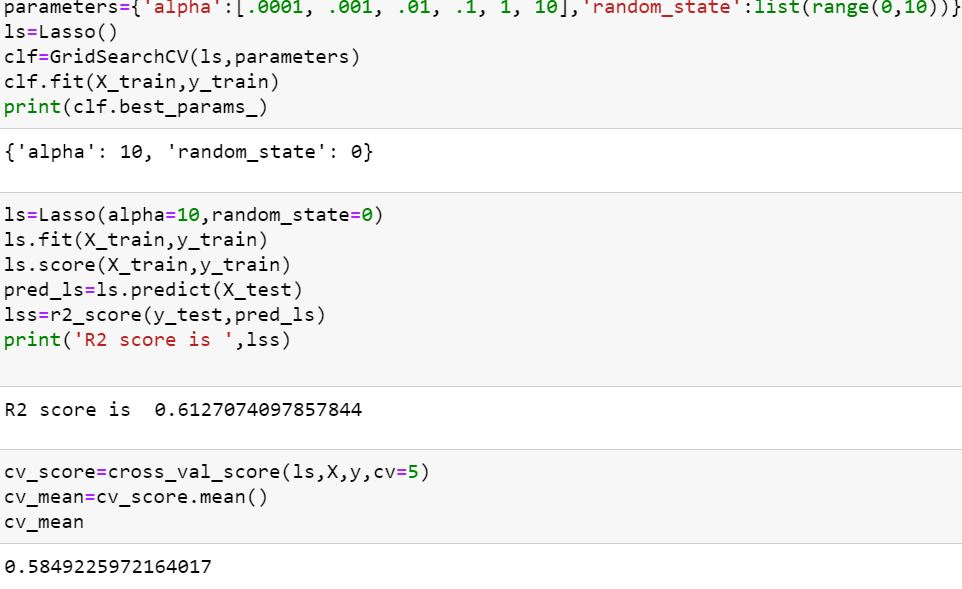
The word “LASSO” stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator. It is a statistical formula for the regularisation of data models and feature selection.LASSO or L1 regularization is a technique that can be used to improve many models, including generalized linear models (GLMs) and Neural networks. LASSO stands for “**least absolute shrinkage and selection operator**.” However, you might wonder if the phrase or the acronym came first.

**What is Lasso Regression**?

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

Lasso Regression uses L1 regularization technique (will be discussed later in this article). It is used when we have more number of features because it automatically performs feature selection.





Here r2 score is 0.61 for dataset and cross validation score is 0.58 in Lasso algorithm.

**Ridge Regression:**

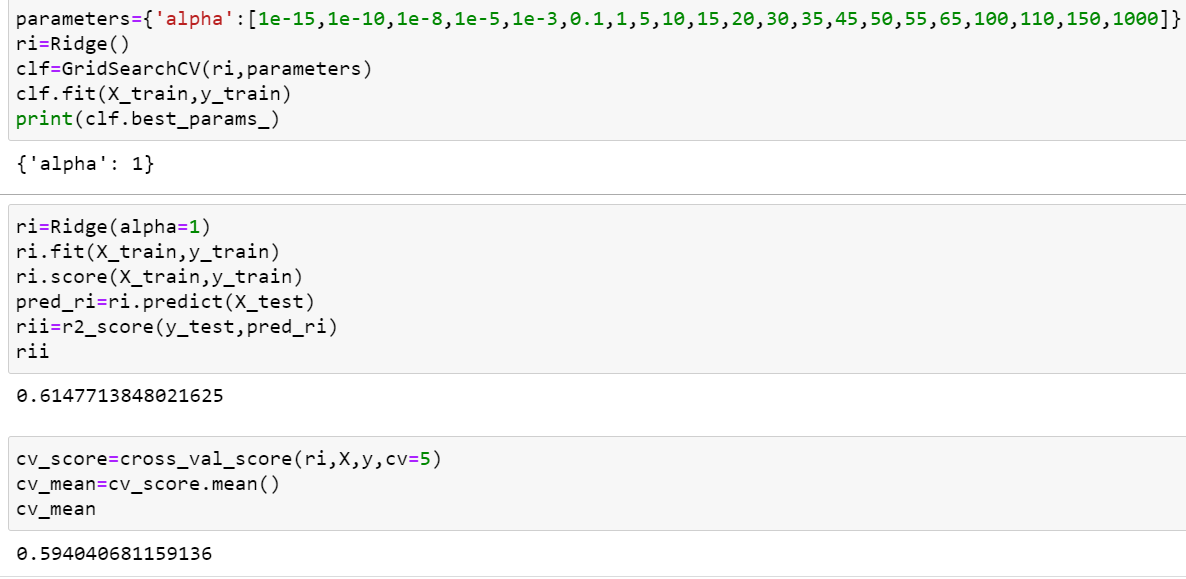
Ridge [regression](https://www.mygreatlearning.com/blog/what-is-regression/) is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

**Ridge Regression Models**

For any type of regression machine learning models, the usual regression equation forms the base which is written as:

**Y = XB + e**

Where Y is the dependent variable, X represents the independent variables, B is the regression coefficients to be estimated, and e represents the errors are residuals. Once we add the lambda function to this equation, the variance that is not evaluated by the general model is considered. After the data is ready and identified to be part of L2 regularization, there are steps that one can undertake.



Here r2 score is 0.61 for dataset and cross validation score is 0.59 in Ridge algorithm.

**RandomForest Regression:**

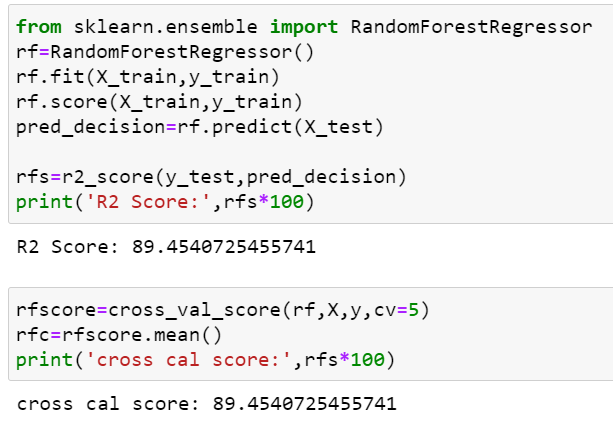
Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Random forest is an ensemble of decision tree algorithms.It is an extension of [bootstrap aggregation (bagging)](https://machinelearningmastery.com/bagging-ensemble-with-python/) of decision trees and can be used for classification and regression problems.

In bagging, a number of decision trees are created where each tree is created from a different bootstrap sample of the training dataset. A [bootstrap sample](https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/) is a sample of the training dataset where a sample may appear more than once in the sample, referred to as **sampling with replacement**.

Bagging is an effective ensemble algorithm as each decision tree is fit on a slightly different training dataset, and in turn, has a slightly different performance. Unlike normal decision tree models, such as [classification and regression trees](https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/) (CART), trees used in the ensemble are unpruned, making them slightly overfit to the training dataset. This is desirable as it helps to make each tree more different and have less correlated predictions or prediction errors.

Predictions from the trees are averaged across all decision trees resulting in better performance than any single tree in the model.



Here r2 score is 0.89 for dataset and cross validation score is 0.89 in Lasso algorithm.

**SVR Regression:**

Support Vector Regression (SVR) uses the same principle as SVM, but for regression problems. Let’s spend a few minutes understanding the idea behind SVR.

Recall the idea of SVR. Here, you give a set of input vectors and defined an output. The SVR then fits a model and tries to learn from those input vectors and finally predicts the response for a given new input vector. While working with time series data like stock prices, you need to determine which will be the "feature vector".

This is because time series data is time-dependent i.e. there will be a lot of past values and you can not take everything as feature vectors. For example, you have a stock price data set that contains the prices of a single stock from the previous six months. Now, based on this data you want to forecast the future price of the stock.

Here, you have to transform the past data to build some feature vectors. There are many ways you can do this i.e. averaging the past one month's prices or the current price of the stock divided by the moving average. This will minimize the input vectors and make it easier for the SVR to fit them. SVR features are unordered x-y pairs, so you can not get a model that considers time order. If you want to maintain the time order, you can build separate SVRs i.e. one for the past 10 days, one for the past 1 month, etc. and then take the average of all the predictions and forecast the values.

#### Advantages and Disadvantages of Support Vector Regression

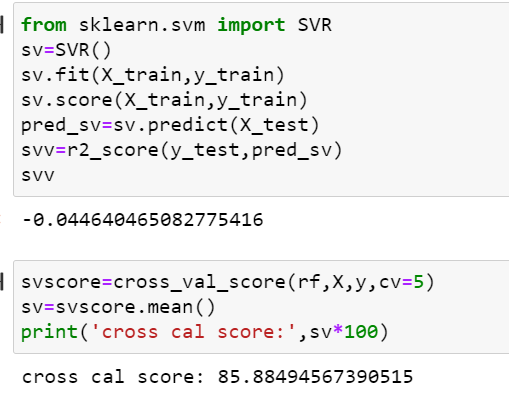
There are some key benefits to choose a support vector machine for regression tasks. There are some drawbacks as well. Let's talk about them-

**The key advantages are-**

* SVM works really well with high-dimensional data. If your data is in higher dimensions, it is wise to use SVR.
* For data with a clear margin of separations, SVM works relatively well.
* When data has more features than the number of observations, SVM is one of the best algorithms to use.
* As a discriminative model, it need not to memorize anything about data. Therefore, it is memory efficient.

**Some drawbacks are-**

* It is a bad option when the data has no clear margin of separation i.e. the target class contains overlapping data points.
* It does not work well with large data sets.
* For being a discriminative model, it separates the data points below and above a hyperplane. So, you will not get any probabilistic explanation of the output.
* It is hard to understand and interpret SVM as its underlying structure is quite complex.

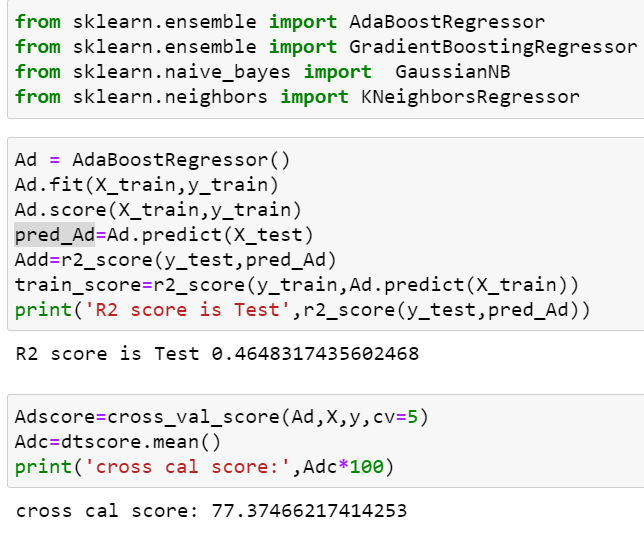


Here r2 score is -0.044 for dataset and cross validation score is 85 in Lasso algorithm.

**AdaBoostRegressor:**

Adaboost stands for Adaptive Boosting and it is widely used ensemble learning algorithm in machine learning. Weak learners are boosted by improving their weights and make them vote in creating a combined final model. In this post, we'll learn how to use [AdaBoostRegressor](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html" \t "_blank) class for the regression problem. AdaboostRegressor starts fitting the regressor with the dataset and adjusts the weights according to error rate. The tutorial covers:

* Preparing data
* Defining the model
* Predicting and checking the accuracy

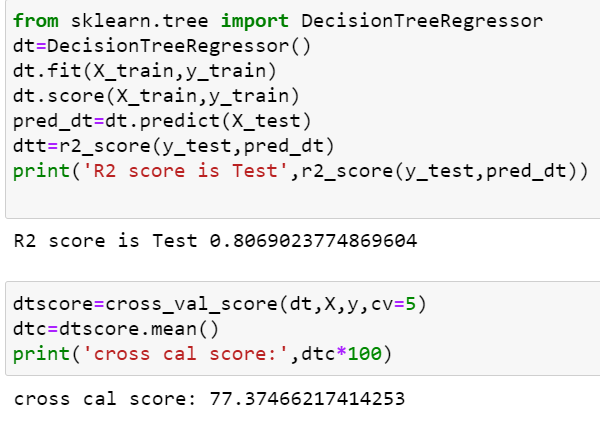


Here r2 score is 0.46 for dataset and cross validation score is 77 in Lasso algorithm.

**DecisionTreeRegressor:**

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

**Decision Tree** is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs, and utility. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.



Here r2 score is 0.80 for dataset and cross validation score is 77 in Lasso algorithm.

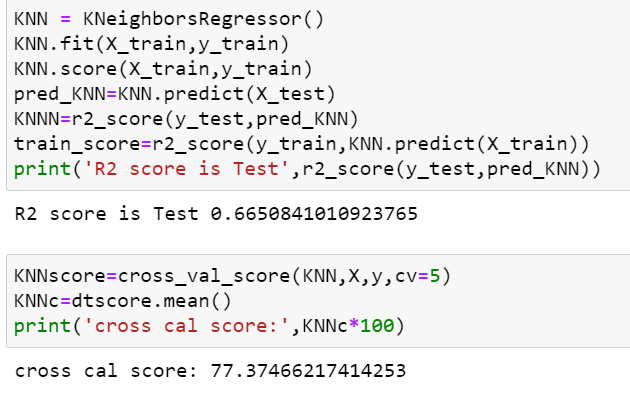
**KNN Regressor:**

As we saw above, KNN algorithm can be used for both classification and regression problems. The KNN algorithm uses '**feature similarity**' to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set.

The kNN algorithm is one of the most famous [machine learning](https://realpython.com/learning-paths/machine-learning-python/) algorithms and an absolute must-have in your machine learning toolbox. Python is the go-to programming language for machine learning, so what better way to discover kNN than with Python’s famous packages [NumPy](https://realpython.com/numpy-tutorial/) and [scikit-learn](https://scikit-learn.org/stable/)!

Below, you’ll explore the kNN algorithm both in theory and in practice. While many tutorials skip the theoretical part and focus only on the use of libraries, you don’t want to be dependent on automated packages for your machine learning. It’s important to learn about the mechanics of machine learning algorithms to understand their potential and limitations.

At the same time, it’s essential to understand how to use an algorithm in practice. With that in mind, in the second part of this tutorial, you’ll focus on the use of kNN in the Python library scikit-learn, with advanced tips for pushing performance to the max.



Here r2 score is 0.66 for dataset and cross validation score is 77 in KNN algorithm.

**Select Best model:**

Here we check r2 score and after that check cross

validation score of all the model the Random forest repressor is the best model because r2 score and cross. validation score is 0.89 So apply hyper parameter tuning on it.

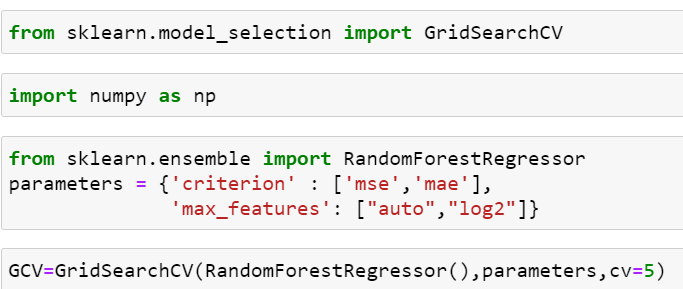
**GridSearchCv**:

GridSearchCV is **a library function that** is a member of sklearn's model\_selection package. It helps to loop through predefined hyperparameters and fit your estimator (model) on your set. So, in the end, you can select the best parameters from the listed hyperparameters.

**Parameterlist:**

There is a list of different machine learning models. They all are different in some way or the other, but what makes them different is nothing but input parameters for the model. These input parameters are named as **Hyperparameters.**These hyperparameters will define the architecture of the model, and the best part about these is that you get a choice to select these for your model. Of course, you must select from a specific list of hyperparameters for a given model as it varies from model to model.

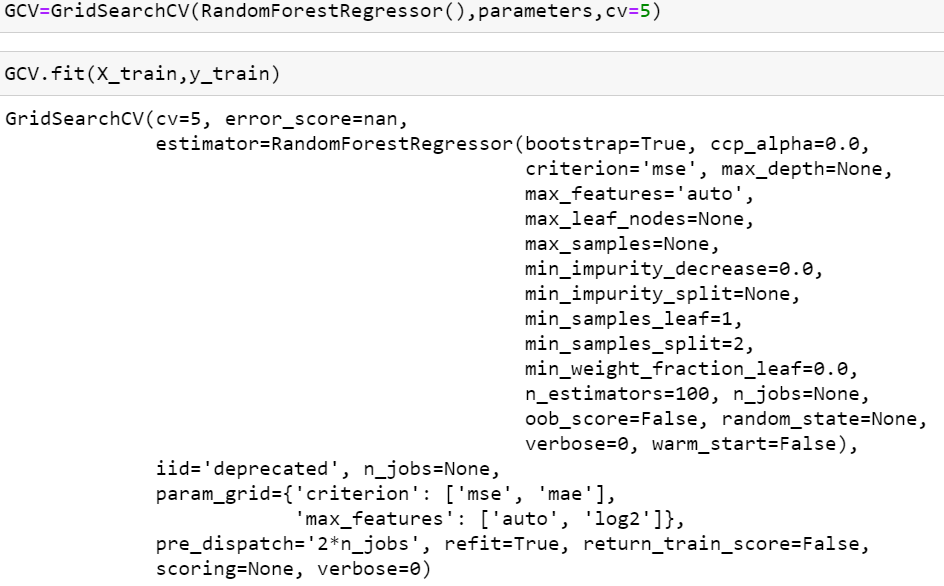
There is a list of different machine learning models. They all are different in some way or the other, but what makes them different is nothing but input parameters for the model. These input parameters are named as **Hyperparameters.**These hyperparameters will define the architecture of the model, and the best part about these is that you get a choice to select these for your model. Of course, you must select from a specific list of hyperparameters for a given model as it varies from model to model.

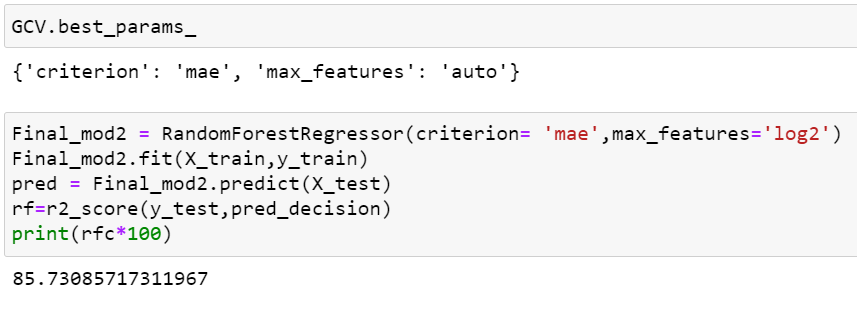


Here above are the parameter list of RandomForestRegressor model.

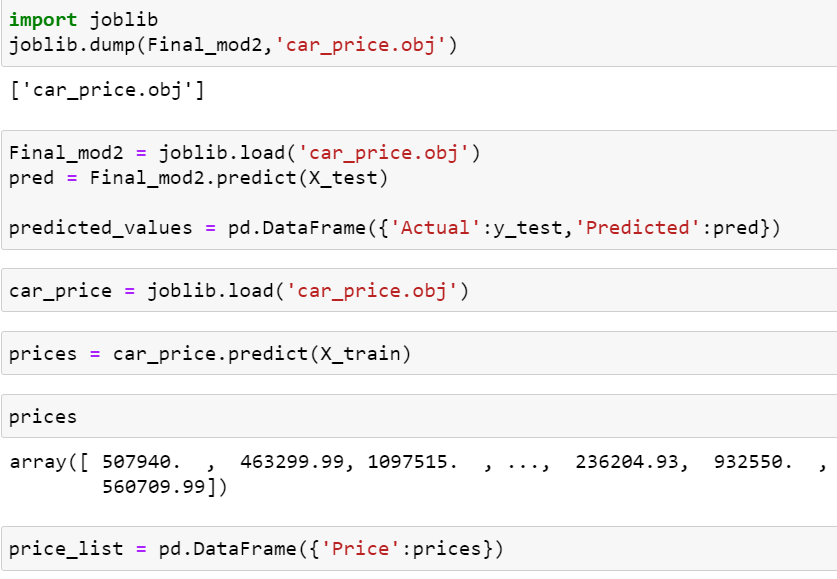
The above code block we have the following parameters  
max\_features: In this maximum features there are two values auto and auto,sqrt.

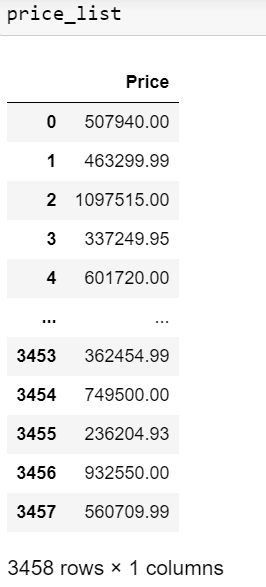
Criterion: In criterion have mse and mae are two parameters.





Here after use different parameter list the best parameter list is select. Above are the best parameter list RandomForestRegressor. Put this parameters into the model so output is finally best score is RandomForestRegressor is 85 so it is the best score.





Finally Load the model and predict the values.

**Identification of possible problem-solving approaches (methods)**

The prices of new cars in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But due to the increased price of new cars and the incapability of customers to buy new cars due to the lack of funds, used cars sales are on a global increase. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used car’s actual market value. It is important to know their actual market value while both buying and selling.

**METHODOLGY**

There are two primary phases in the system:

1. Training phase: The system is trained by using the data in the data set and fits a model (line/curve) based on the algorithm chosen accordingly.

2. Testing phase: the system is provided with the inputs and is tested for its working. The accuracy is checked. And therefore, the data that is used to train the model or test it, has to be appropriate. The system is designed to detect and predict price of used car and hence appropriate algorithms must be used to do the two different tasks. Before the algorithms are selected for further use, different algorithms were compared for its accuracy.

**Objective**

To develop a efficient and effective model which predicts the price of a car according to user’s inputs.

To achieve good accuracy.

**Key Metrics for success in solving problem under consideration**

 I divided out my project into three parts:

1. [**Exploratory Data Analysis**](https://towardsdatascience.com/a-machine-learning-project-predicting-used-car-prices-efbc4d2a4998#c0d7)

## Understanding my Data

## Exploring Categorical Data

## Exploring Numerical Data

## Columns with too many Null Values

## Removing Outliers

## Dropping Remaining Columns

## Visualizing variables and relationships

1. [**Data Modelling**](https://towardsdatascience.com/a-machine-learning-project-predicting-used-car-prices-efbc4d2a4998#207c)

## Dummy Variables

## Scaling the data

## Checking accuracy of the model

1. [**Feature Importance**](https://towardsdatascience.com/a-machine-learning-project-predicting-used-car-prices-efbc4d2a4998#eced)

**Visualizations**

For visualization purpose use heatmap for dataset: data visualization is a method of graphically representing numerical data where the value of each data point is indicated using colors. ... More importantly, heatmaps help to classify the sections that are performing sub-par and need optimization

**Interpretation of the Results:**

House price prediction project there are in car dataset the Selling\_Price are the target variable was present and there was continues values are present so it regression problem need to use regression algorithm.

After that check the shapes of the dataset and check null values are present that that dataset and remove the null values .

Check the data types of the dataset and need to convert categorical columns into numeric.

And also the use zscore and IQR to check the data loss.

And check the skewness and remove the skewness using power transform and log transform method.

**Use scaling technique.**

And use different Regression models and depend on r2 score and cross validation score select Random forest repressor is the best model and apply hyper parameter tuning on it.

After that load that model and predict the values.

**Interpretation of the Results**

The results of our tests were quantified in terms of the R score of our predictions. score is a statistical 2 R 2 measure of how close the data are to the fitted regression line.

|  |  |
| --- | --- |
| **Learning Algorithm** | **R Score on Car Data** |
| Lasso | 0.61 |
| Ridge | 0.61 |
| RandomForestRegressor | 0.89 |
| AdaBoostRegressor | 0.46 |
| KNeighborsRegressor | 0.66 |
| SVR | 0.85 |

**CONCLUSION**

The increased prices of new cars and the financial incapability of the customers to buy them, Car sales are on a global increase. Therefore, there is an urgent need for a Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The proposed system will help to determine the accurate price of car price prediction. This paper compare different algorithms for machine learning : Linear Regression, Lasso Regression and Ridge Regression,SVM,Random Forest Regressor, KNN, DecisionTree Regressor

**Learning Outcomes of the Study in respect of Data Science**

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

In this project drop some columns because not much important.

Remove the null values using mean and median In this many columns null values are present so require time to remove null values of all the columns.

Use label encoder to convert categorical values into numeric.

In this using IQR technique more data loss.

Remove the skewness is difficult because dataset size is big and so much skewness is present so time is require to remove the skew ness.

Use different Regression modelling technique and using r2 score and cross validation score select random forest is the best model and apply hyper parameter tuning on it and load the model predict the values.

**Limitations of this work and Scope for Future Work**

In future this machine learning model may bind with various website which can provide real time data for price prediction. Also we may add large historical data of car price which can help to improve accuracy of the machine learning model. We can build an android app as user interface for interacting with user. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.