

USED CARS PRICE PREDICTION



Submitted by:

SACHIN GUPTA

**ACKNOWLEDGMENT**

Working on this project has an incredible experience that will have an impact on my career.

It is pleasant gratification to present micro credit defaulter project.

I have completed this project by taking the help from Google, Bing and You tube.

**INTRODUCTION**

A car is a wheeled motor vehicle used for transportation. Most definitions of cars say that they run primarily on roads, seat one-to-eight people, have four wheels and mainly transport people rather than goods.

Cars came into global use during the 20th century, and developed economies depend on them.

Predicting the price of used cars in both an important and interesting problem.

Sometimes selling your used car becomes crucial as we are not able to identify its fair price accurately. The depreciation of a car depends on a variety of factors so the car owner needs to be aware of the worth of their vehicle. With the rapid expansion of Machine Learning, this problem can also be solved by minimizing human efforts and time.

Let’s see an end-to-end solution for a similar problem.

* Business Problem Framing

Steps Involved

* Importing the required packages into our python environment.
* Importing the house price data and do some EDA on it.
* Data Visualization on the house price data.
* Feature Selection & Data Split.

Conceptual Background of the Domain Problem

Used Cars price prediction can help the developer determine the selling price of a used cars and can help the customer to arrange the right time to purchase a cars. There are three factors that influence the price of a cars which include cars years, price and variant.

* Review of Literature

I have used visualization tool such as Cat Plot to understand the data in a better way.

I have also used label encoder technique and convert all the data into numerical form to do the data analysis in an easier way.

I used describe method for five-point summary analysis and also found the number of rows and columns in dataset.

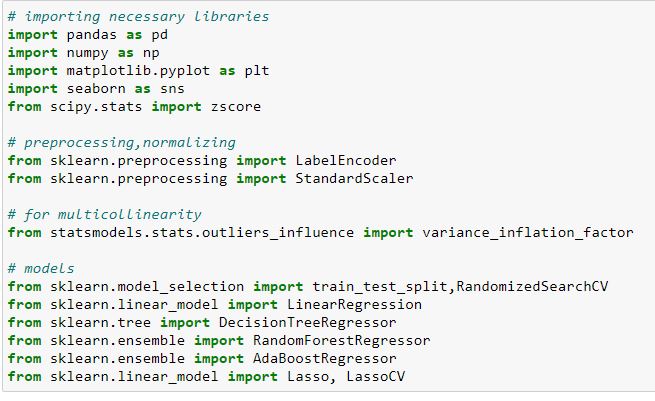
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

In this particular project I need to understand the importance of cars and how much easy for a human to buy a used cars for his family . I have done the exploratory data analysis process and try to figure out the used cars Price Prediction in a better way.

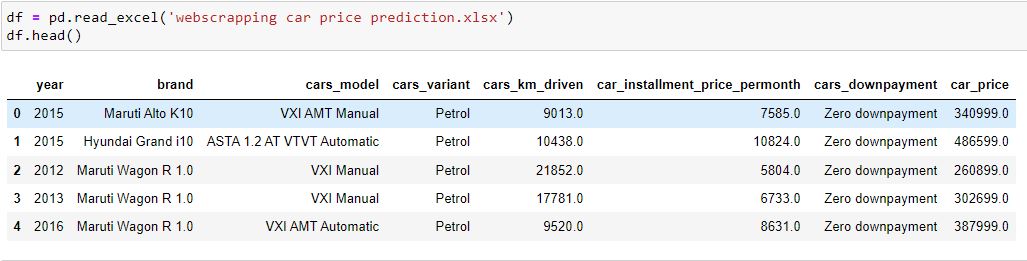
**Importing Libraries:**

*I am importing all the library which I required for EDA, visualization, prediction and finding all matrices. The reason of doing this is that it become easier to use all the import statement at one go and we do not require to import the statement again at each point***.**



* Data Sources and their formats

,Now I am going to upload or read the files/data-sets using pandas. For this I used read csv method.



There are train and test 2 separate dataset. I have used head function to get the brief information about first 5 rows of both the dataset.

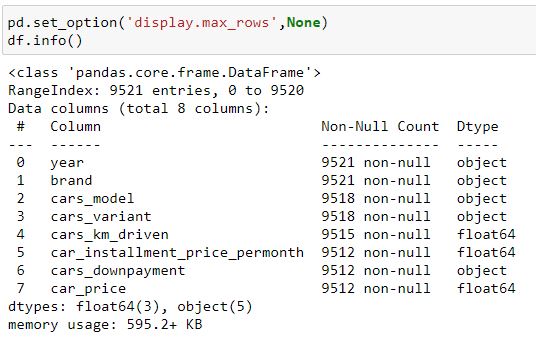


Total datasets contain 9521 rows and 8 columns and each of the features is extremely important in this dataset.

* Data Preprocessing Done

*In the data pre-processing stage, I have found out if there is any missing data in dataset, for a particular column if there are any outliers present and how to handle the outliers. I have also found the total shape of the data set. I have also found out the dataset description using describe method. So, in this pre-processing process I have mainly cleansed the data and prepared the right set of data for further processing. Below are the steps for data Pre-processing in detailed manner*.

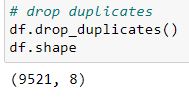
Finding Null Values



*There are missing values in the dataset which i will fixed it soon.*

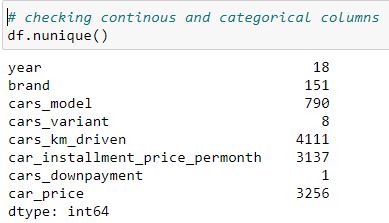
*It is a mixed dataset as 3 columns are of float type and 5 are object type.*

Checking The Duplicates



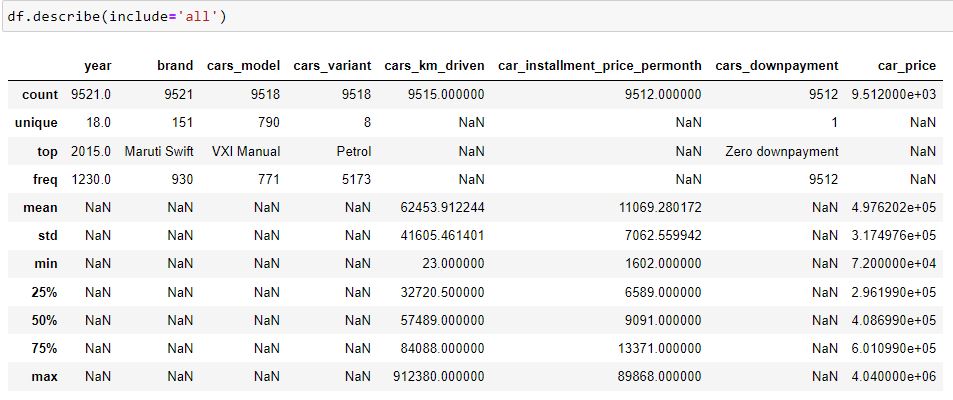
There are no duplicates in the dataset.

Checking continuous and categorical columns



There are 3 categorical columns and 5 continuous columns.

Using Describe



Above statistics data show that there are few outliers in this dataset.

There is also difference between mean and 50% value in some of the columns which used to get fix for better prediction.

Also, number of rows in each column are not same, means there are null values in the data set.

Also, the mean and 50% value of most of the column are not same and the STD and mean have a major difference to each other.

•Visualizations

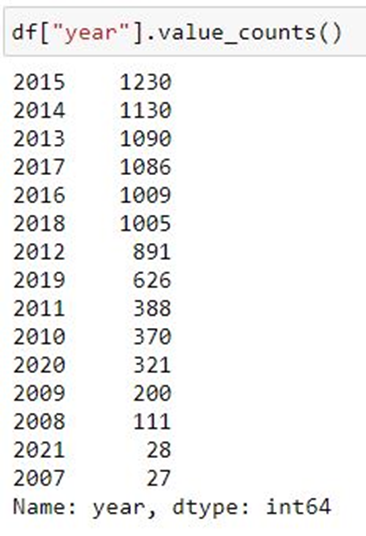
I start off with catplot () function and use x argument to specify the axis we want to show the categories.

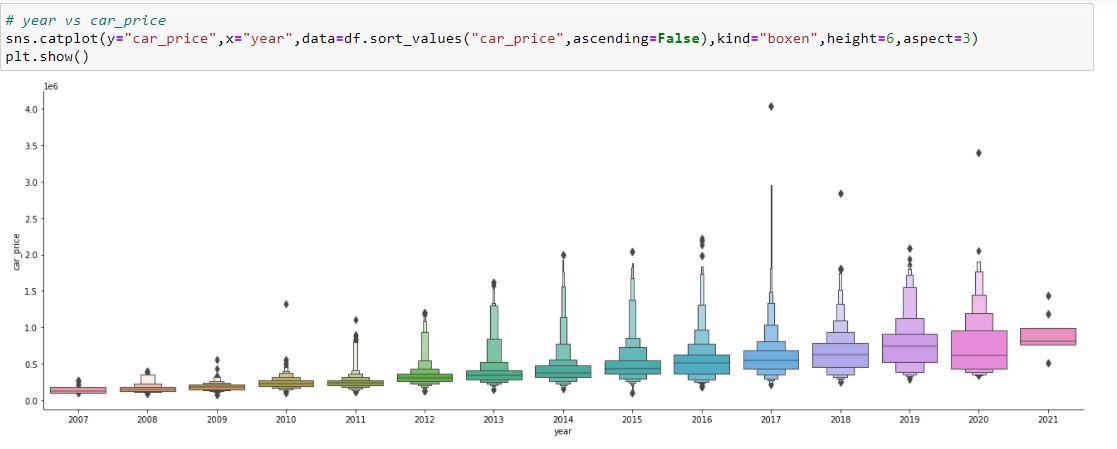
Another popular choice for plotting categorical data is a bar plot. In the count plot example, our plot only needed a single variable.

In the bar plot, we often use one categorical variable and one quantitative.

Description catplot is a complementary function to interact plot () that is designed for plotting interactions when both predictor and moderator (s) are categorical (or, in python terms, factors)..

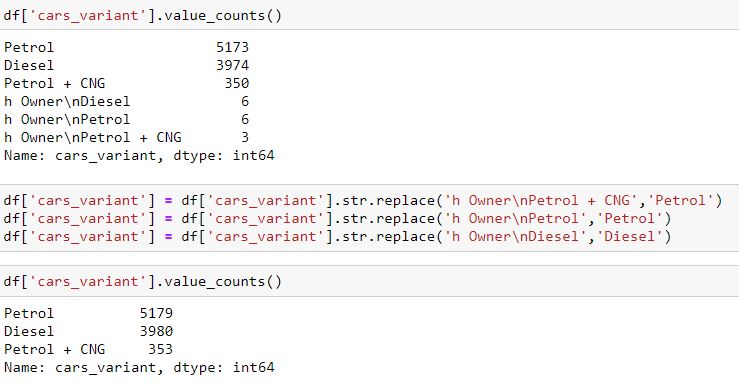
1. Year

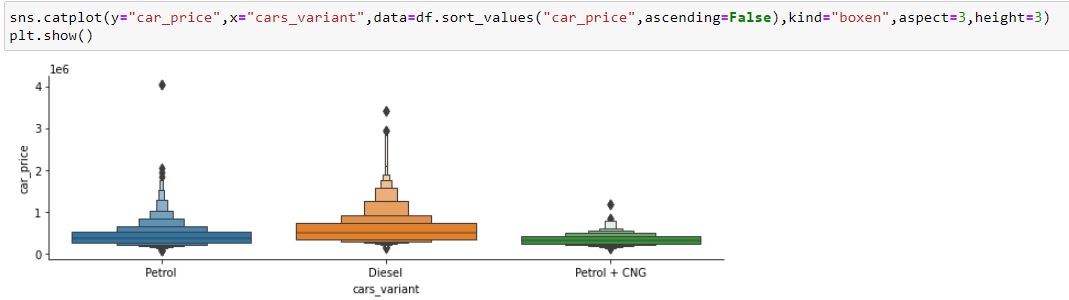




From the above cat plot i can say that 2021 car brand have highest car price and remaining years cars values are less than 2021.

b)cars\_variant

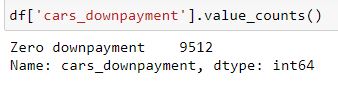


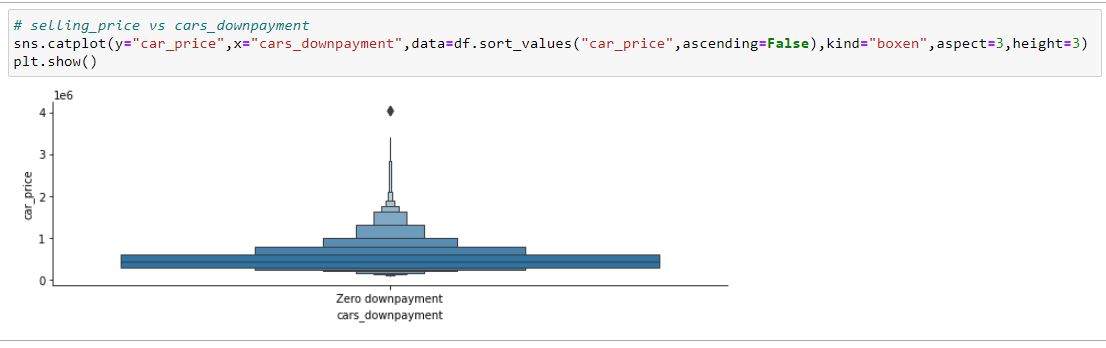


In above cat plot I can say that diesel used car prices are more as compare to petrol and Cng.

In India diesel car are more preferable to purchase for an individual.

c) cars\_downpayment



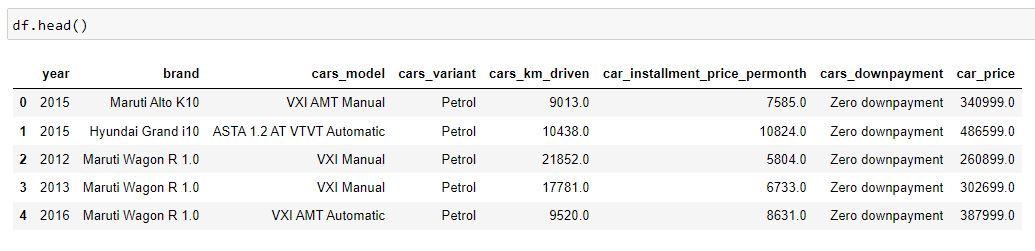


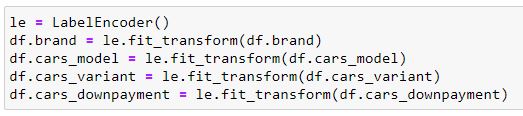
From the above cat plot I can analyse that most of the used cars are purchased in zero downpayment.

**Label Encoding**

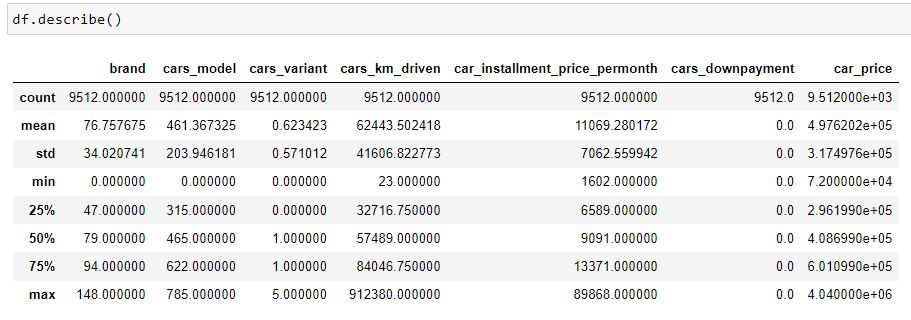
*This approach is very simple and it involves converting each value in a column to a number. Hereby, I would focus on Label-Encoder. So it is used as a part of SciKit-learn library (one of the most widely used Python library) and are used to convert text or categorical data into numerical data which the model expects and perform better with.*

*After Applying Label Encoder I have converted all the strings columns into integer.*





**Describe**



After using label encoding in this dataset I use describe function to understand the dataset perfectly.

**Scatter Plot**

**Scatter plots are the graphs that present the relationship between two variables in a data-set .It represents data points on a two-dimensional plane or on a Cartesian system.**

**The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.**

**These plots are often called scatter graphs or scatter diagrams.**

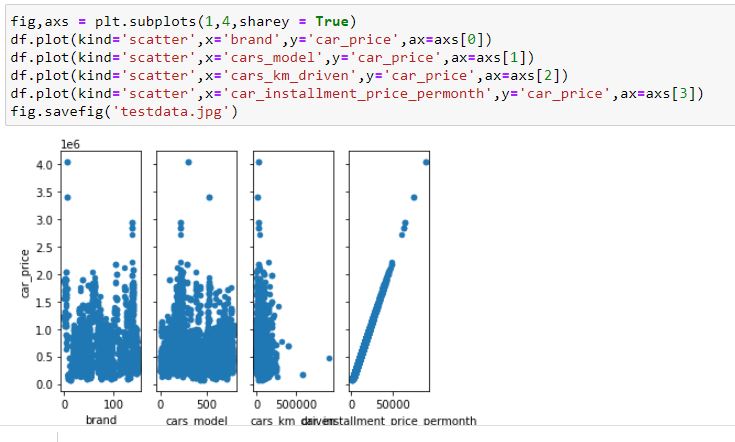
**Scatter plots are used in either of the following situations:**

**a) When I have paired numerical data,**

**b) When there are multiple values of the dependent variable for a unique value of an independent variable,**

**c) In determining the relationship between variables in some scenarios, such as identifying potential root causes of problems,**

**checking whether two products that appear to be related both occur with the exact cause and so on.**



**By analysing scatter plot I can determine that continuous columns has important relationship in used cars and prices.**

**Histogram plot**

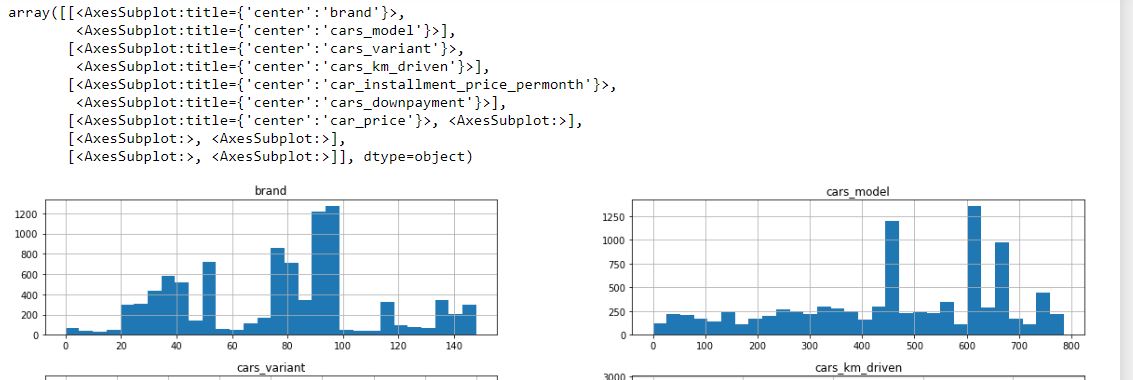
**Histograms group the data in bins and is the fastest way to get idea about the distribution of each attribute in dataset.**

**The following are some of the characteristics of histograms:**

**a) That provides us a count of the number of observations in each bin created for visualization.**

**b) From the shape of the bin, I can easily observe the distribution i.e. weather it is Gaussian, skewed or exponential.**

**Histograms also help us to see possible outliers**



he above output shows that it created the histogram for each attribute in the dataset.

From this, I can observe that perhaps car\_price, cars\_downpyment, car\_installment\_price\_permonth,cars\_km\_driven and cars\_variant attribute may have exponential distribution while brand and cars\_model have Gaussian distribution.

**Box Plot**

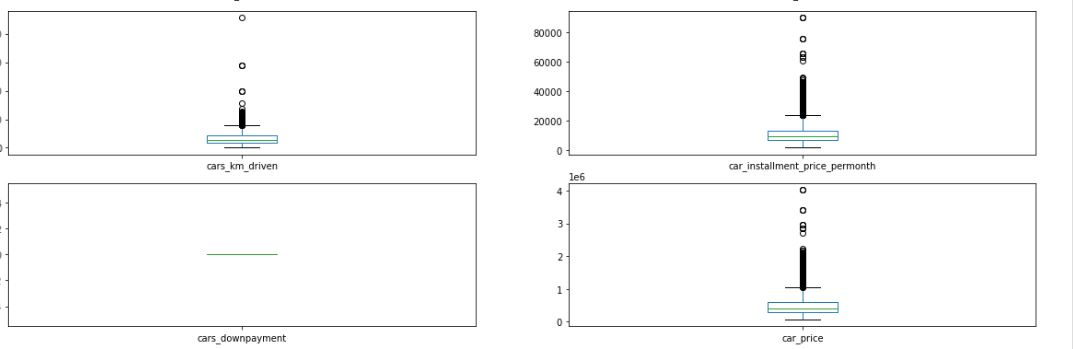
#### In the box plot, a box is created from the first quartile to the third quartile, a vertical line is also there which goes through the box at the median.

#### Here x-axis denotes the data to be plotted while the y-axis shows the frequency distribution.

#### Boxplots are also very beneficial when large numbers of data sets are involved or compared.

#### In simple words, I can define the box plot in terms of descriptive statistics related concepts.

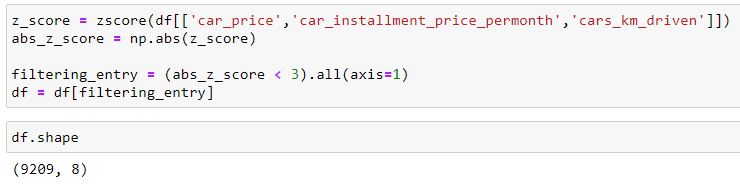
#### That means box or whiskers plot is a method used for depicting groups of numerical data through their quartiles graphically.

****

#### From the above box plot I can see that some columns have outliers which I will remove soon.

**Z Score**

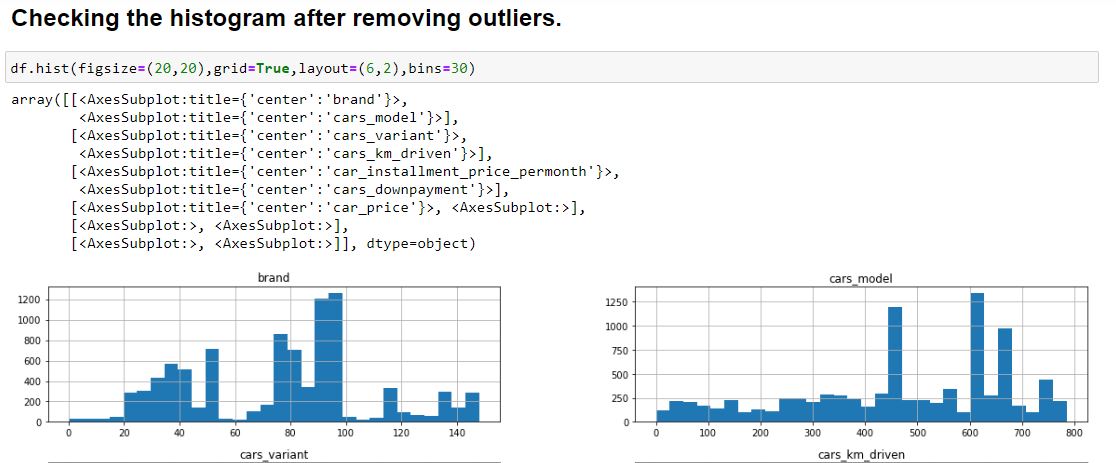
#### The z score is also called the standard score represents the number of standard deviations with which the value of an observation point or data differ than the mean value of what is observed.



#### I am using zscore to remove it by replacing the outliers with the continuous columns.

#### And from shape I got 9209 rows which shows major of the outliers are cleaned.

**Checking the histogram after removing outliers.**



#### After Removing Skewness of most of the columns now it is perfectly distributed for Model Building

**Model Building**

#### A machine learning model is built by learning and generalizing from training data, then applying that acquired knowledge to new data it has never seen before to make predictions and full fill its purpose. Lack of data will prevent you from building the model. Regression is a supervised machine learning technique which is used to predict continuous values.

#### To find out the relationship between all the input variable I have used correlation function and find out whether there is a positive/negative relationship between a pair of variables. It shows the correlation between different features and label.

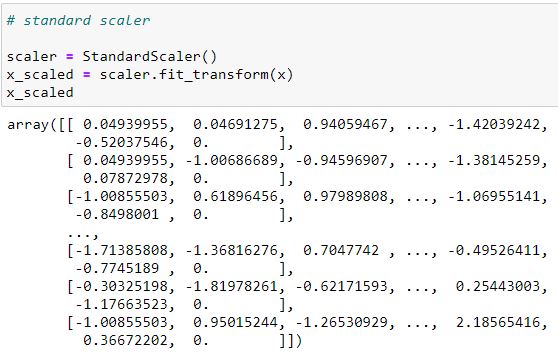
Data Inputs- Logic- Output Relationships

To find out the relationship between all the input variable I have used correlation function and find out whether there is a positive/negative relationship between a pair of variables. It shows the correlation between different features and label.

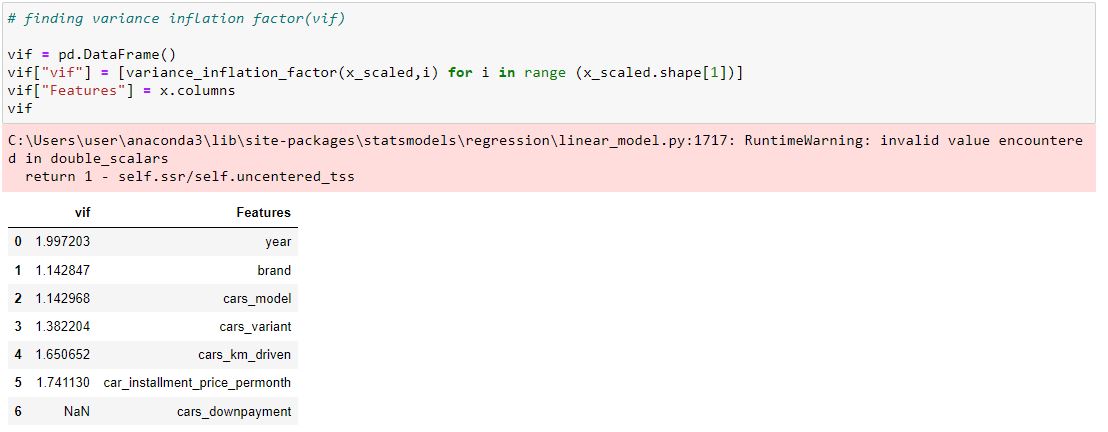
#### Splitting the dataset into the features and label.

**Standard Scaler**

#### The scikit library provides various types of scalers above I will used the Standard Scaler to scale the data.

 The scikit library provides various types of scalers above I will used the StandardScaler to scale the data**.**

**VIF - variance inflation factor**



By using Variance Inflation Factor(VIF) I can see that there is a strong relationship among many Features as values are greater than 5 which shows there is multi-collinearity exists.

* Hardware and Software Requirements and Tools Used

Hardware and Software Requirements and Tools Used:

For this particular dataset the Hardware is used Windows as operating system, and the software used are mainly Jupyter notebook for model building and various internal packages that are defined in the anaconda/jupyter notebook.