Logo, company name

Description automatically generated

**Second Hand Cars Re-selling Price Analysis**

**Advance Diploma in Data Science.**

**Machine Learning 02.**

**Module leader: Mr. Chameera De Silva**

ADDS212P

16th January 2023

Inaam Muaz - COADDS212P-013

[**1. Introduction**](#_heading=h.rwjpsnoxo0fb) **3**

[1.1. About the Dataset](#_heading=h.nrw1fdjeyns7) 3

[**2. Implementation**](#_heading=h.6ioqbwxxywkd) **3**

[2.1. Libraries](#_heading=h.dhqlv7fnw4gb) 3

[2.2. Loading the Raw Data](#_heading=h.qrvfy71yzf6) 4

[2.3. Data Processing](#_heading=h.40p4dxm7zhg1) 4

[2.3.1. Dealing with Null and Missing Data](#_heading=h.ehikvpnz6iz9) 4

[2.3.2. Dealing with variables of interest.](#_heading=h.ib67n2u5l82c) 5

[2.3.3. Dealing with outliers](#_heading=h.iwy3w3y0k4n4) 6

[2.4. Checking the OLS assumptions.](#_heading=h.iymhw8e9a1sa) 9

[2.4.1. Linearity](#_heading=h.zc2cgrw569py) 9

[2.4.2. Checking for Multicollinearity](#_heading=h.nkvyqn1um819) 10

[Creating dummies](#_heading=h.kwa4anl6ofn6) 11

[2.5. Downloading Preprocessed Data](#_heading=h.52817qsocf6f) 12

[2.6. Linear Regression Model](#_heading=h.85h3ml23m7p4) 12

[2.6.1. Scaling the dataset.](#_heading=h.6t25ogoluwnx) 12

[2.6.2. Train Test Split](#_heading=h.c6p39xkd32na) 12

[2.6.3. Creating the Regression.](#_heading=h.pbx85n3q7wyx) 13

[2.6.4. Checking the weights and bias.](#_heading=h.54h126j6cr4t) 14

[**3. Conclusion**](#_heading=h.ka88inmu3ggz) **19**

# Introduction

Owning a car has always been a milestone for any individual in their life. Due to the current economic condition, the majority of the population would opt for a used car over a brand new car. This fuels the argument and what variable factors into the price of a second hand car. Will these factors be significant when it comes to predicting the second har car reselling price. We will be using a dataset consisting of prices of second hand cars.

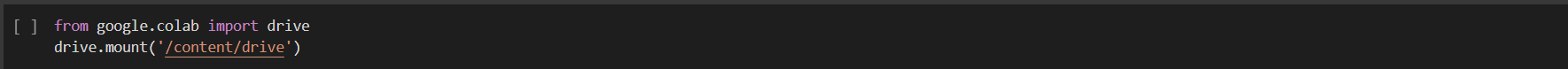
## 1.1. About the Dataset

Our data set is quite large, with nearly 4500 rows. We are supposed to analyze the price of a used car here. Now, the price (target variable) of cars is determined by the data variables listed below.

* Brand : The BMW car will be more expensive than the Toyota.
* Mileage : The greater the mileage the more expensive the car.
* EngineV : The greater the engine volume the more expensive the car. As sports cars are more expensive than the family car.
* Year : The older the car, the cheaper its price.
* EngineType : Whether it is petrol, diesel or gas
* Registration : Whether the car has been registered or not.
* Year : Year of the car
* Model : Model of the car

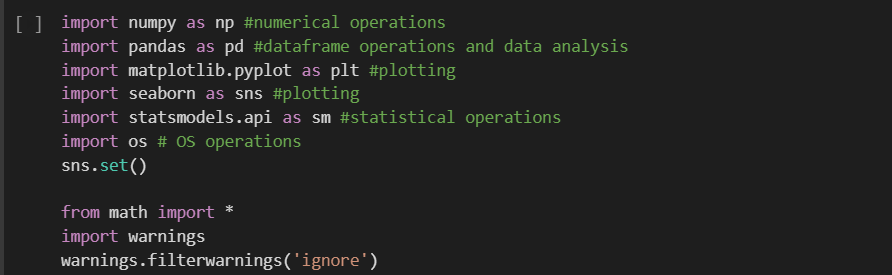
# Implementation

We will be using python as our programming language for this project. Our coding platform will be google collabs. Our dataset was in my google drive. We mounted our drive to Google Collabs using the below code.



## 2.1. Libraries

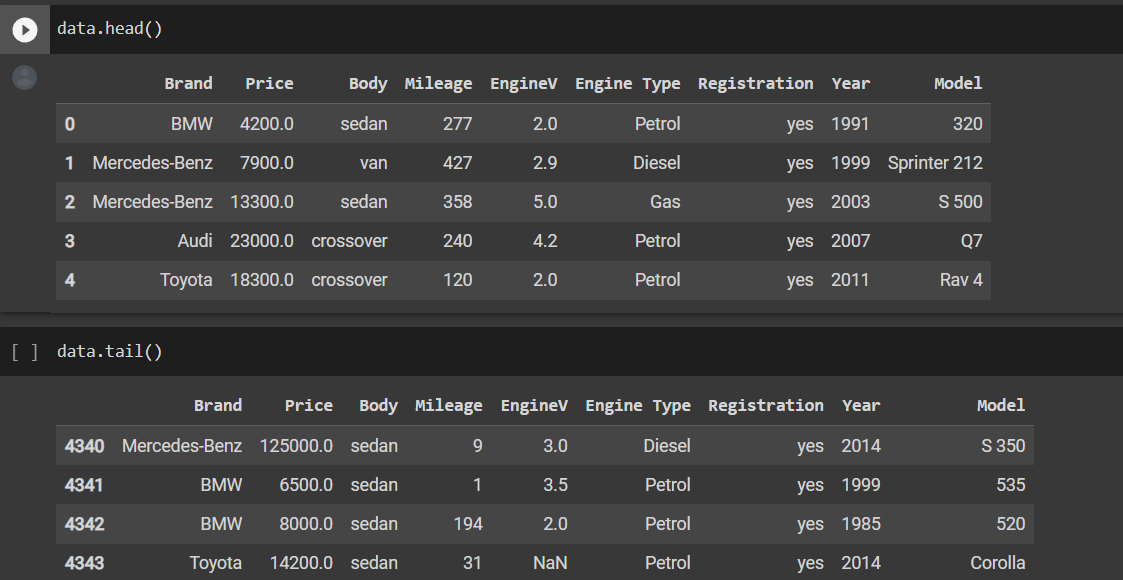
The below libraries will be required for this project.



## 2.2. Loading the Raw Data

We loaded the csv using *“pandas.read\_csv”.*

We ran the following codes to get a feeling of the dataset.

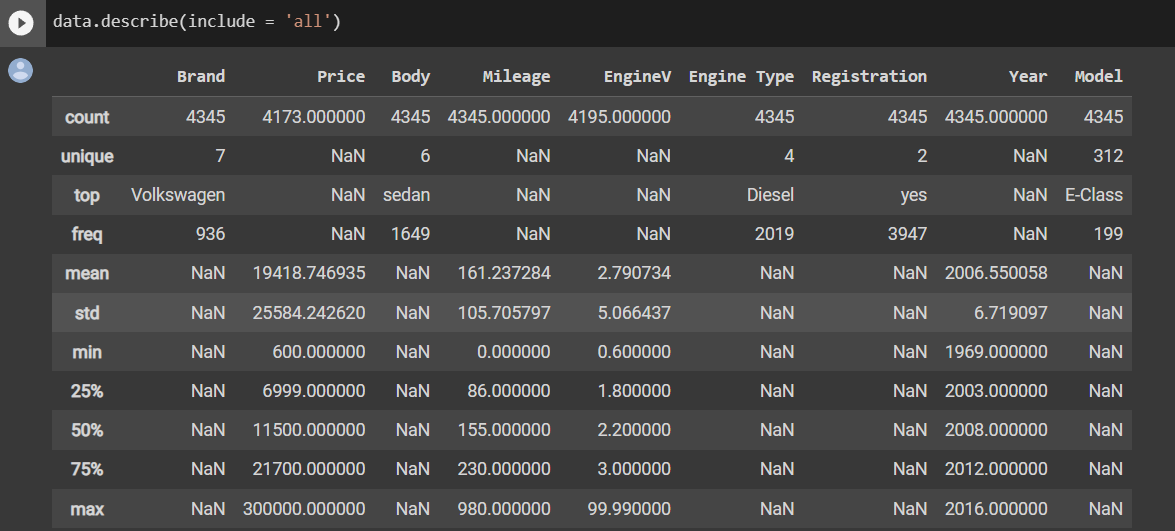


## 2.3. Data Processing

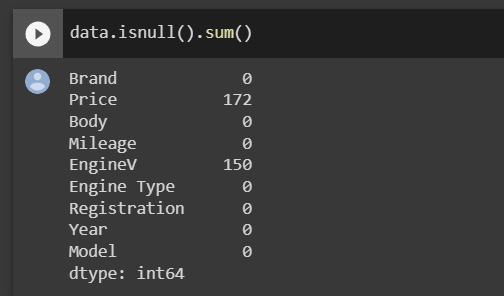
Data preprocessing, which is a crucial phase in Machine Learning, can be defined as the altering or dropping of data before usage in order to ensure or increase performance.

### 2.3.1. Dealing with Null and Missing Data

From the descriptive statistics from the below code we can infer that there are some missing values in our data. We also can see that there are outliers in the variables, we will be treating this in the latter part of the project.

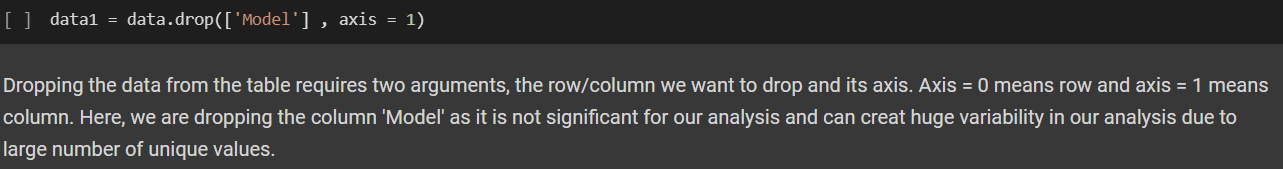


We checked the columns with null values and treated them.

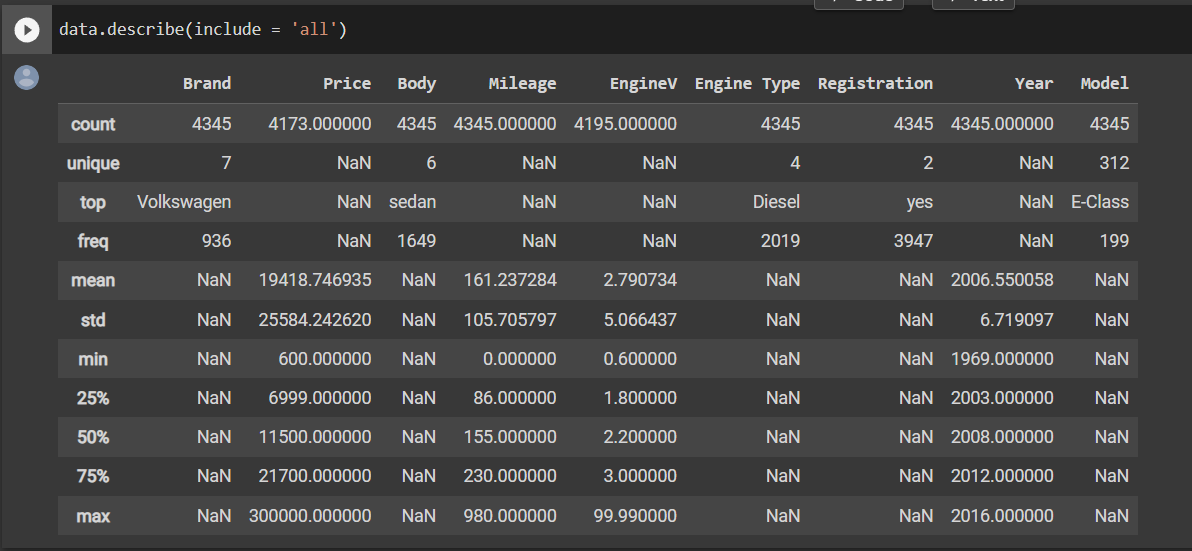


Before dealing with the issue, we need to understand the variables with interest.

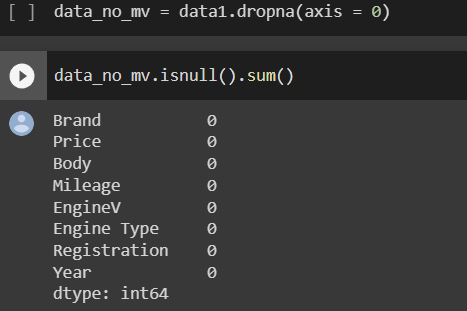
### 2.3.2. Dealing with variables of interest.



After removing the *model* variable, let’s run descriptive statistics on the remaining variables.



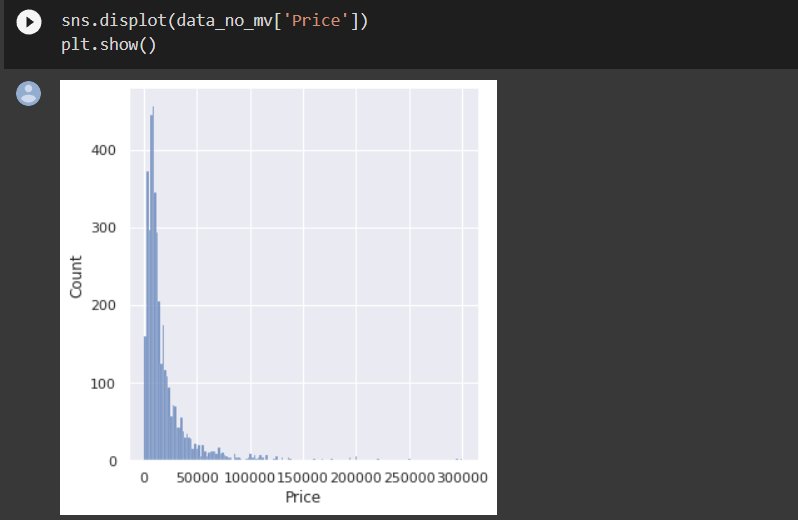
We now have two columns with null values, and we need to effectively handle them. We will remove the rows with missing values in these columns since the percentage of missing values is less than 5% of the total number of rows. The below code will be used to do this.



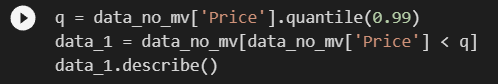
Now, our data set does not have any null values.

### 2.3.3. Dealing with outliers

The *Price, EngineV* and *Year* are variables with outliers. We can plot a distribution plot to get a sense of the outliers and the shape of the data.



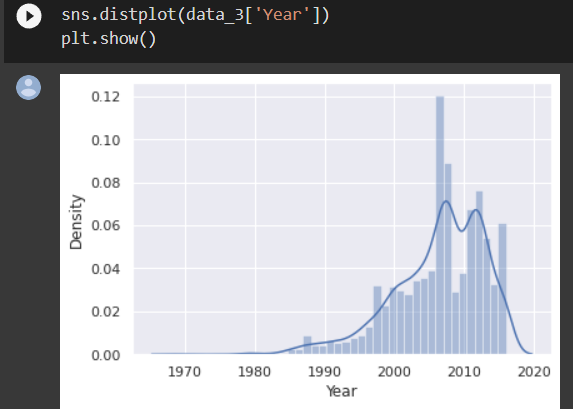
We can use the quantile method to deal with this. We will be keeping the 99 percentile values of the prices from our data. Let's see what happens.

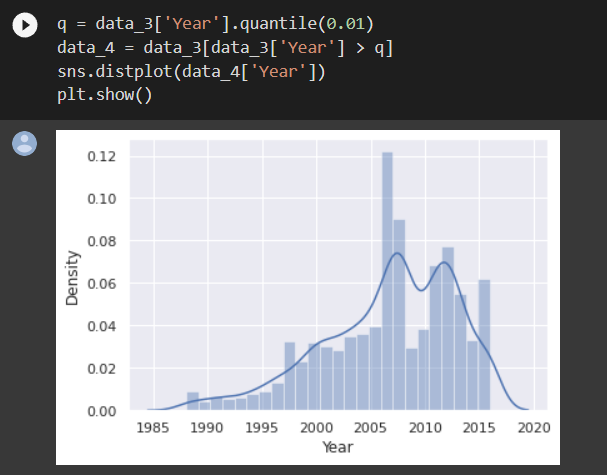


Let’s re-plot the price column.

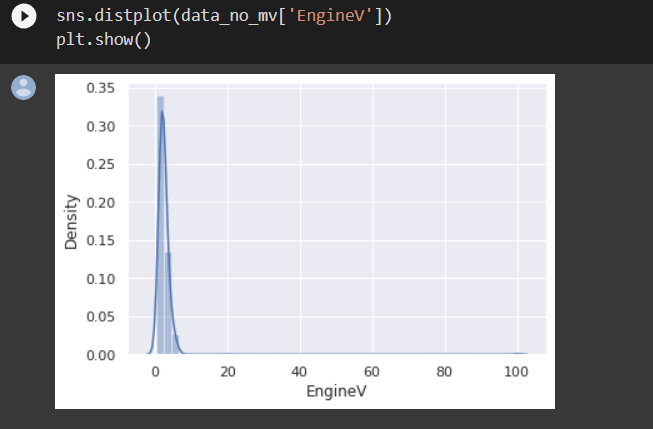


We treated the *Year* and *Mileage* variable again using the same quantile method. For the *year* variable, we had to change the code as the distribution was negatively skewed.



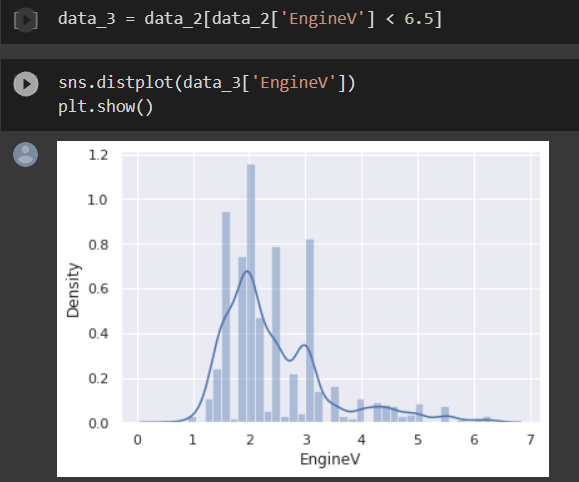


We sensed something unusual when plotting the *EngineV* column.



The "EngineV" column has a lot of extreme outliers. What causes this?

Manually inspecting the data reveals that the column contains some values equal to 99.99. This is so that we can usually fill up the blank cells with these values. But generally speaking (This is where domain expertise is required), a car's engine volume cannot be greater than 6.5 or lower than 0.6. So now we also have to deal with this error. Let’s do that.



Let’s save the cleaned and processed to a new variable and reset the index.



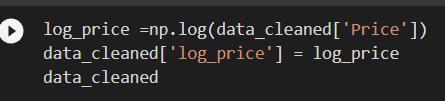
## 2.4. Checking the OLS assumptions.

### 2.4.1. Linearity

We will be using the multi variable regression, so we have to check OLS assumptions. One of them is to look for linearity. Let’s do that.



From the scatter plots we can sport patterns between the dependent and the independent variables of our data but these are not the linear ones. To make our data best fit for the linear regression model, we have to transform our data to get a linear relationship between the dependent and the independent variables.



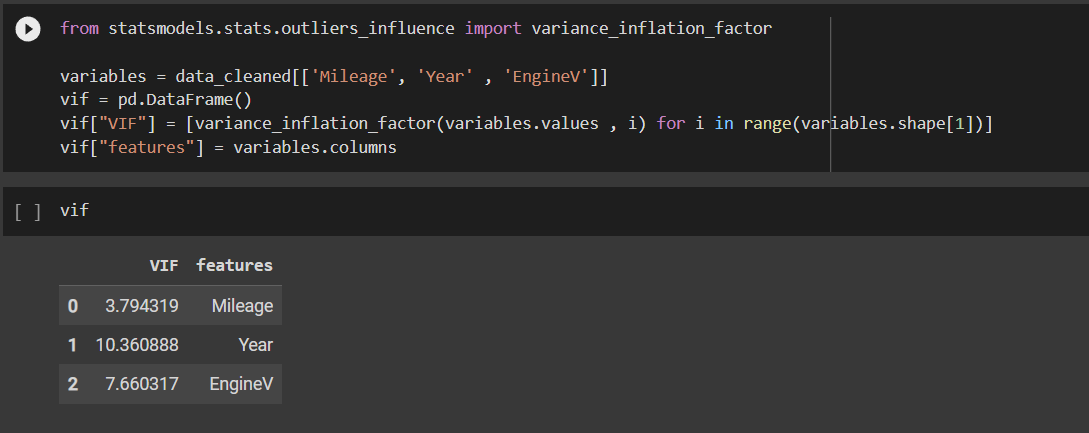
Let’s replot and see the change.



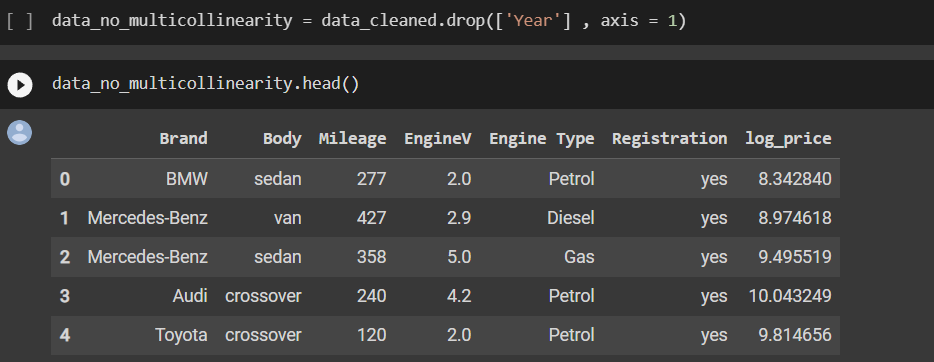
Now, the pattern is much more linear. This will improve our model.

### 2.4.2. Checking for Multicollinearity

The OLS assumption of no multicollinearity says that there should be no linear relationship between the independent variables. We can treat this using Variable interest factor. Let’s do that.

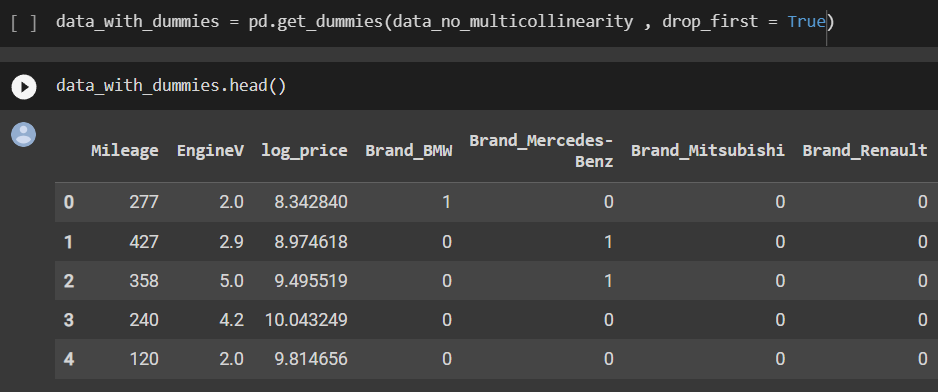


Let’s remove *Year* as it’s VIF is nearing 10.

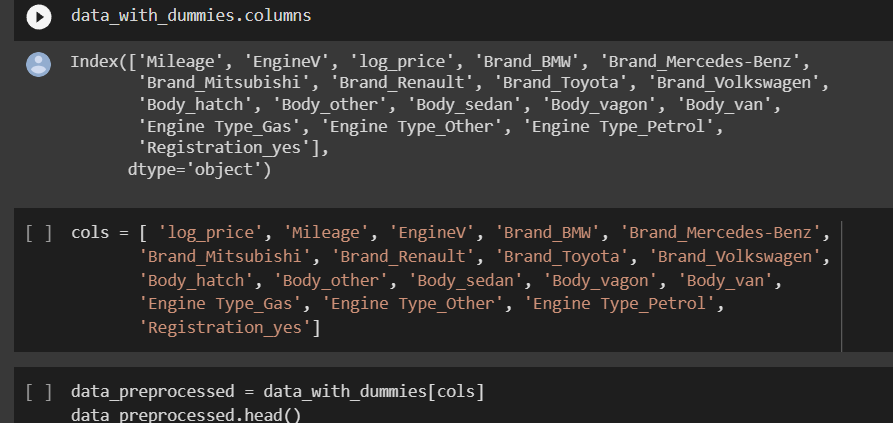


## Creating dummies

Creating dummies will solve the problem of categorical variables in our dataframe. Let’s do this. This will increase the dimensionality of the dataframe.

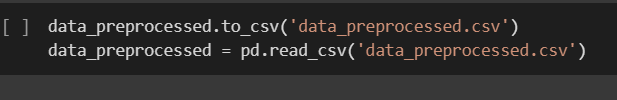


Optional step is to tidy our dataframe by reordering our columns.



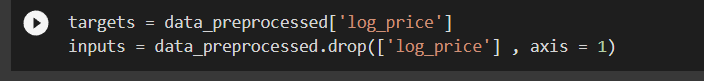
## 2.5. Downloading Preprocessed Data

As our dataset has been cleaned we can download this and reimport.



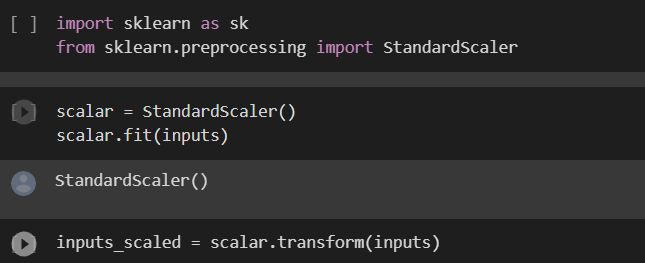
## 2.6. Linear Regression Model

Let’s slice our dataset into dependent and independent variables.



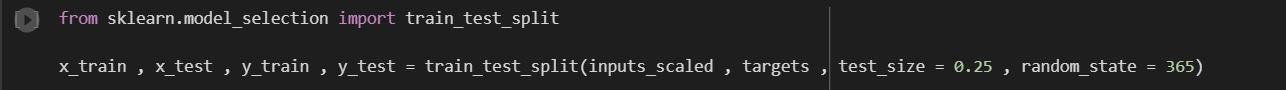
### 2.6.1. Scaling the dataset.

This step can be done in the preprocessing part also. Scaling the variables will improve the performance of the model and work well with distance calculating algorithms.



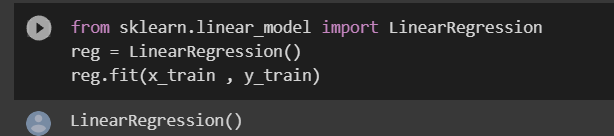
### 2.6.2. Train Test Split

Train test split is a model validation process that allows you to simulate how your model would perform with new data. Let’s do that. It can be easily done using the method train\_test\_split from sklearn.



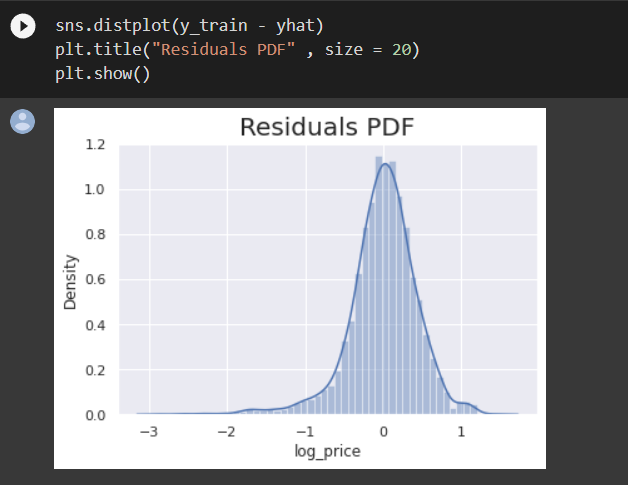
### 2.6.3. Creating the Regression.

Now let’s fit the model.



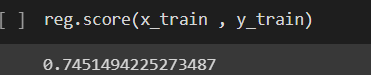
Let’s predict and plot the prediction





We can see our residuals are normally distributed. This indicates normality.

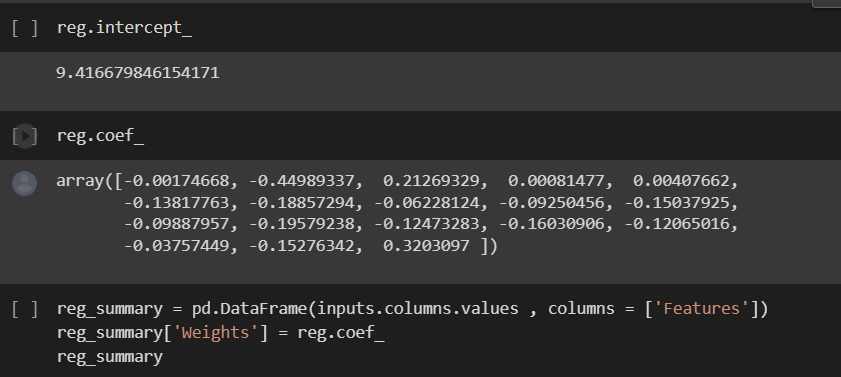
Let’s check the accuracy of our model.

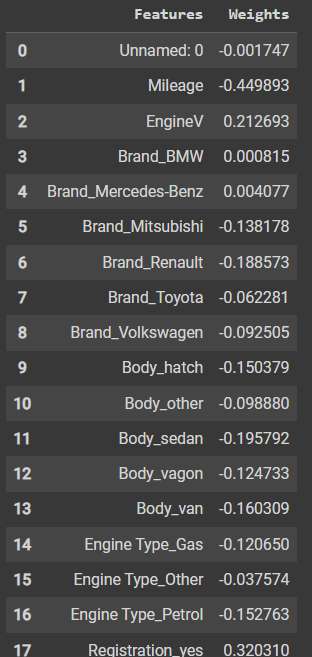


It is 74%.

### 2.6.4. Checking the weights and bias.

Weights play an important role in changing the orientation or slope of the line that separates two or more classes of data points.Weights tell the importance of the variable. biases are the learnable parameters of our model. Let’s check them.

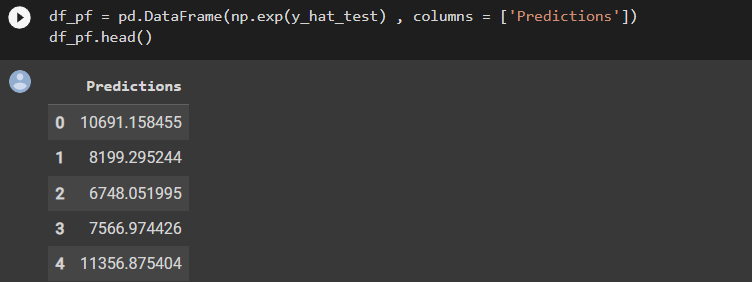


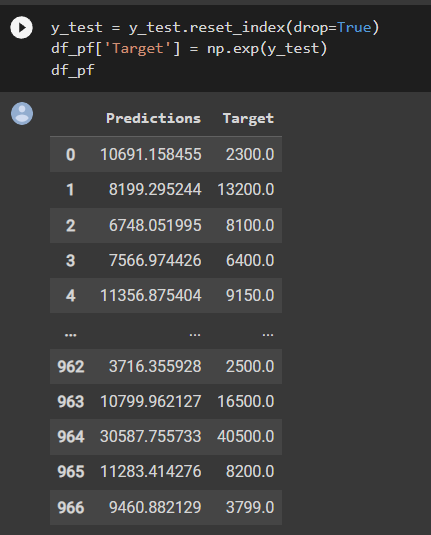


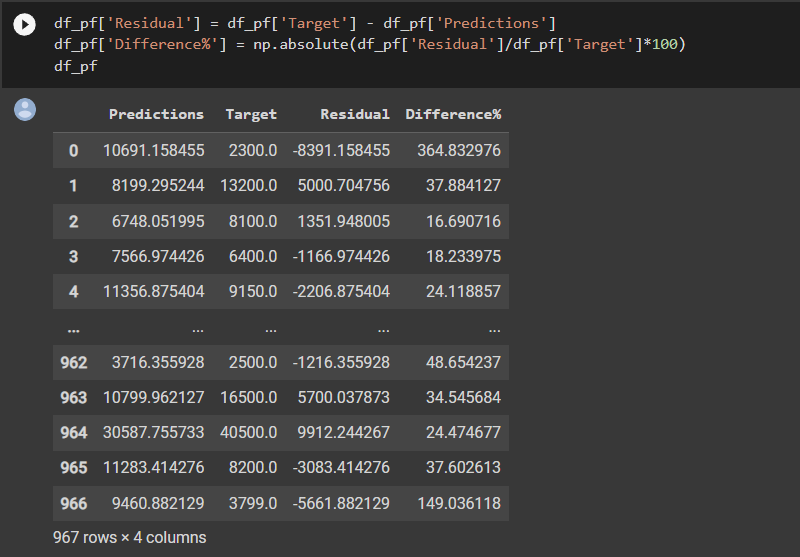
We can see from the summary table that some of the dummies are positive while others are negative. This refers to the relationship between these variables and "Price" or "log price," which is our target variable. When the dummy variable has a positive value, it is directly proportional to the target variable; when it has a negative value, it is inversely proportional to the target variable.

Now, observing the brands, the relation is a bit different. The values of *brands* are positive and negative based on the value of the "Audi" *brand*, as it was the benchmark variable while we assigned the dummies. This, for the positive value of a *brand,* means this brand car is more expensive than the "Audi '' *brand* car, and if it's negative means it's cheaper than the"Audi ".

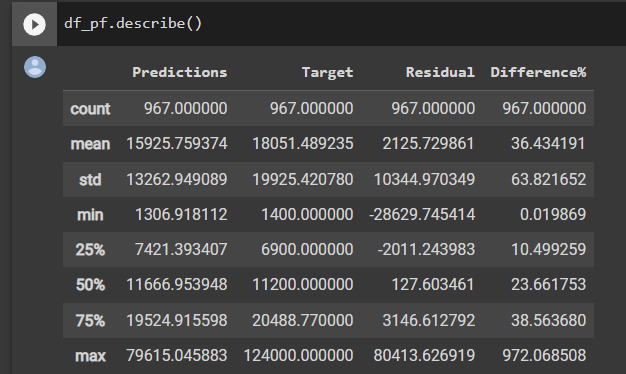
Let’s check how well our model performed compared to actual vs predicted.







From the below descriptive statistics of the df\_pf data frame, we can say that for most of the observations our predictions were quite good, but as there were a large number of outliers, our model didn't perform well there. This could indicate better treatment should have been done on outliers.



# Conclusion

From our analysis we can conclude that the linear regression model fits the datasets and predicts value quite well. We got a accuray score of 74%. It is not an outstanding model. We also noticed when evaluating the model, the residuals were affected due to the presence of outliers even after treating them. Our problem statement of “Predicting Second Hand Cars” is solvable using machine learning models. One research proposal could be fitting different machine learning models like decision trees and LASSO regression.

The code is available in my github profile.