

**Task** – Predict the petrol consumption of a country given various factors such as tax, distance of highways, and population with driving license.

**Input** (Independent Variable)- Tax, distance of highways, average income, and population with driving license.

**Output** – Expected petrol consumption.

Steps –

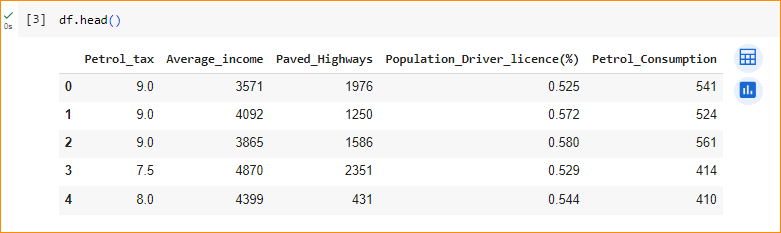
1. Import libraries



1. Load CSV file



1. Show Data

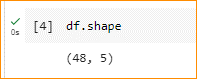


Input (independent variable) – green box

Output – red box

1. Data Shape

Around 48 entries in the data base.



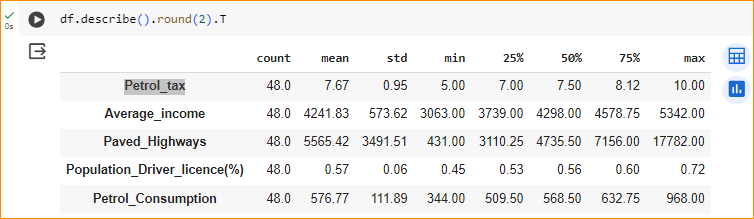
1. Round all data

df.describe(): Generates descriptive statistics of the DataFrame.

round(2): Rounds the numerical values to two decimal places.

T: Transposes the DataFrame, swapping rows and columns.

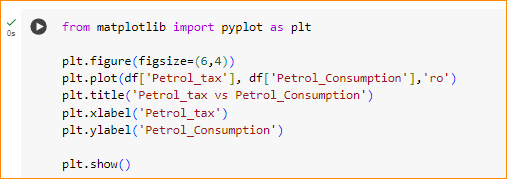
In here it show – total count, mean value, standard deviation

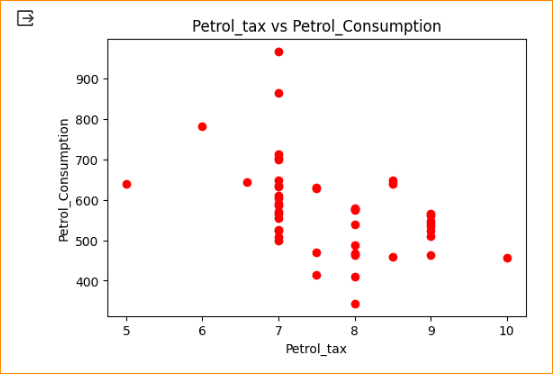


1. Petrol\_tex vs Petrol\_Consumption

Now we have multiple dimension and it is very difficult to show the plot.

In here plot tax vs consumption



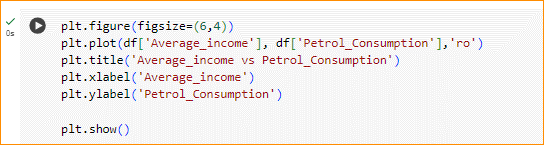


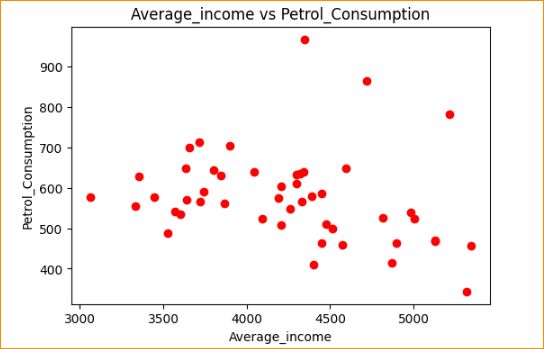
In here can see weak linear relationship

1. Average\_income vs Petrol\_Consumption

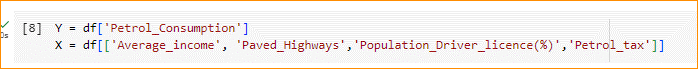
Next try income vs consumption

Here can see strong relationship





Here we can see strong relationship and there is outliers



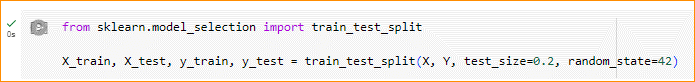
In here lets separate our input(X) and output(Y).

1. Training model

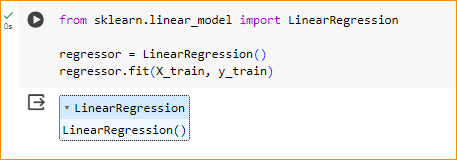
Sklearn there is a function – train\_test\_split

Give both input and output and give the test size. 20% for testing and 80 % training.

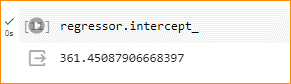
When you define integer variable 42 – this is for good shuffle.



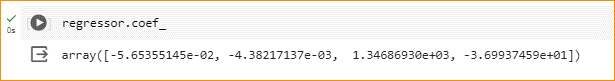
1. Use linear Regression



1. Get intercept



1. Get coefficient – using all attribute

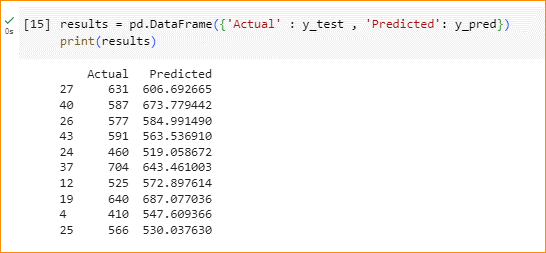


1. Get prediction – provide test data

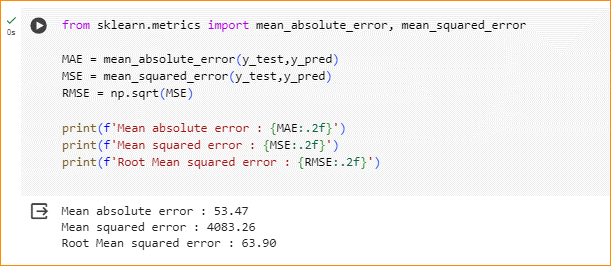


1. Compare to the actual value

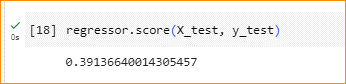
Here can see a good prediction not all time best but good.



1. Evaluate the model



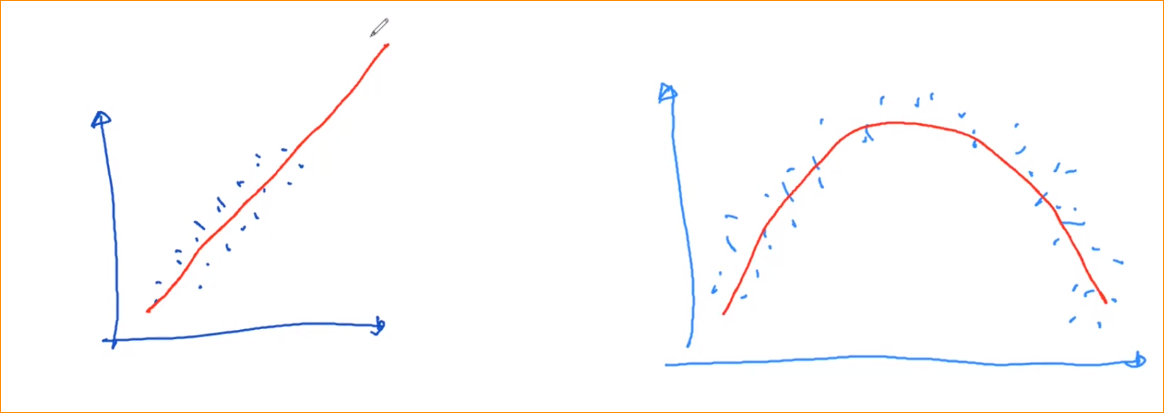
1. Get R square score value for test data



When you give test data it is performing very poor.

Reason – Give very small set of data

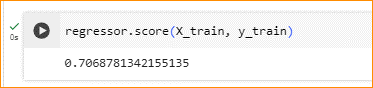
Model have to learn real relationship of the data. It can not recognize what actual graph of the data



Here small data set large data set

Here call generalization of the data set.

1. Get R square score value for train data



* This near to 1 mean good model and near to 0 mean bad value.
* Here this is good model.