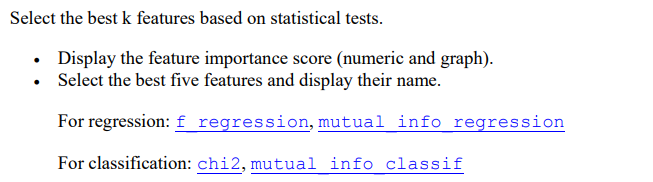
# CSE 4020 - MACHINE LEARNING

# Lab 29+30

# Feature Selection

# Submitted by: Alokam Nikhitha(19BCE2555)

**Question:**

****

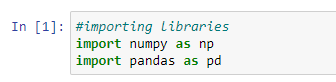
**Dataset Used:**

https://www.kaggle.com/uciml/autompg-dataset

**Procedure:**

* We first import the dataset into our workspace the use of pandas.
* Then we fill missing values using Mean Strategy
* We divide dependent and independent variables
* We split the data into training and Testing sets
* Then we use f\_regression to identify the most co-related five attributes in our dataset
* Next, we use mutual\_info\_regression to identify the most co-related five attributes in our dataset.
* And we find scores of each attribute
* And make plots

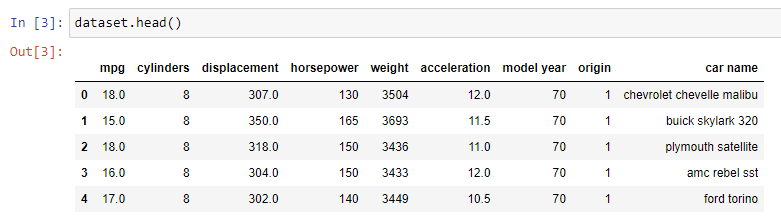
**Code Snippets and Explanation:**



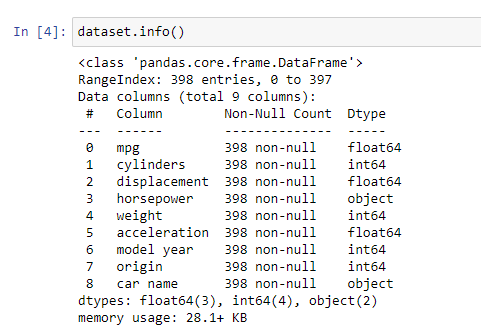
Here we are importing the libraries.



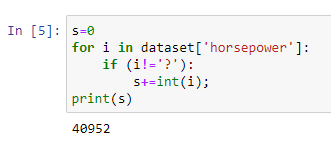
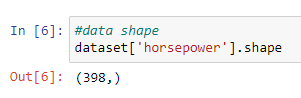
We're importing the dataset into our workspace

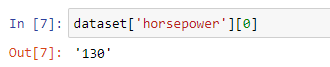


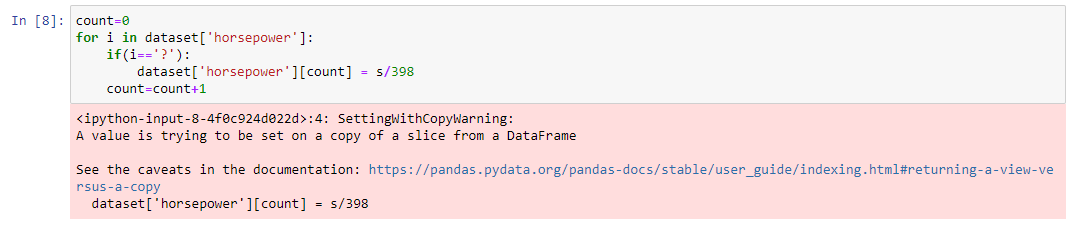
Printing the first few rows of the Dataset



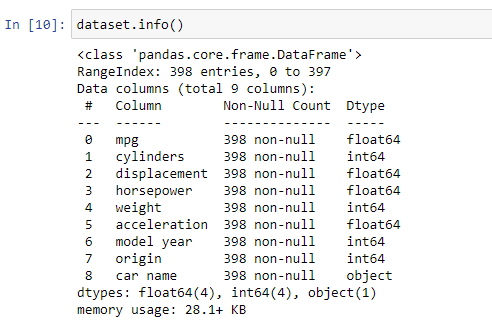
Info of the dataset

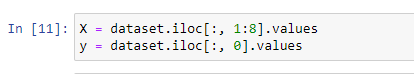




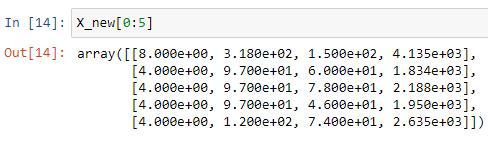
Here we're first off filling withinside the missing values withinside the horsepower attribute. We recognize that the missing values are marked as ? and we're using mean as our replacement value. Here we've calculate the sum of all of the values in horsepower barring the ? after which filled in the ones values divided through total occurrence in our ? marked value.

Since all the values are filled in, we can now convert it into float datatype and on seeing if the conversion has occurred we can see that the Data type of horse power attribute is not float64.

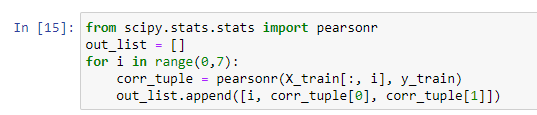
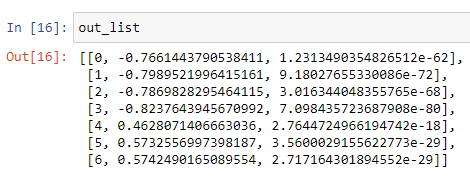
 

Here we have defined set of dependent and independent attributes and split into Test and Training data sets.

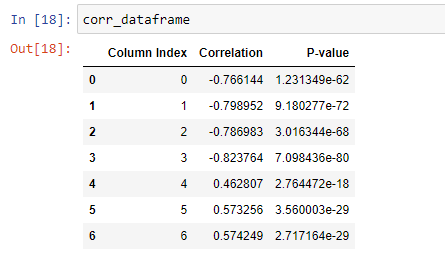
 

Here we have used the f\_regression method to identify the best 5 columns in our training set on basis of their correlation to y attribute of training set.

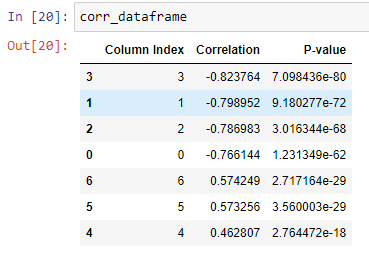
The attributes identified are ‘cylinders’, ‘displacement’, ‘horsepower’, ‘weight’ and ‘origin’.

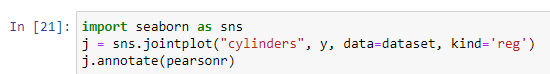
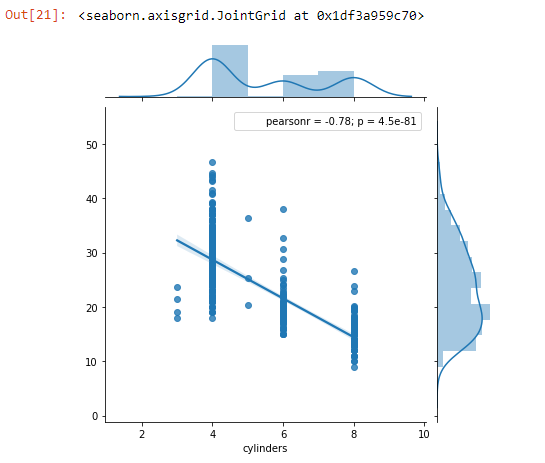
 

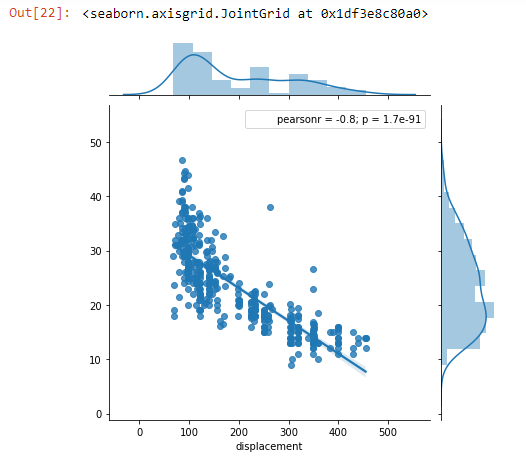
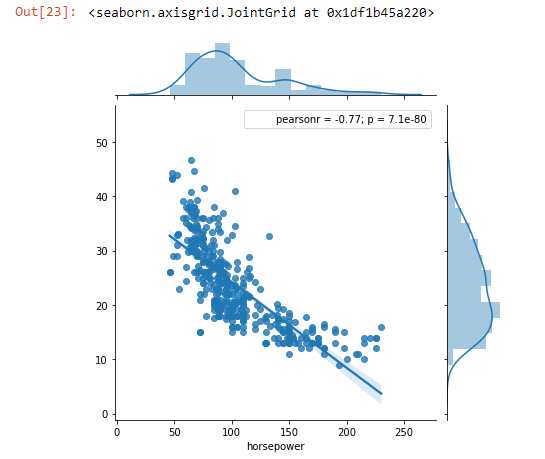
Similarly, we have used mutual\_info\_regression. The identified best attributes are ‘cylinders’, ‘displacement’, ‘horsepower’, ‘weight’ and ‘model year’.

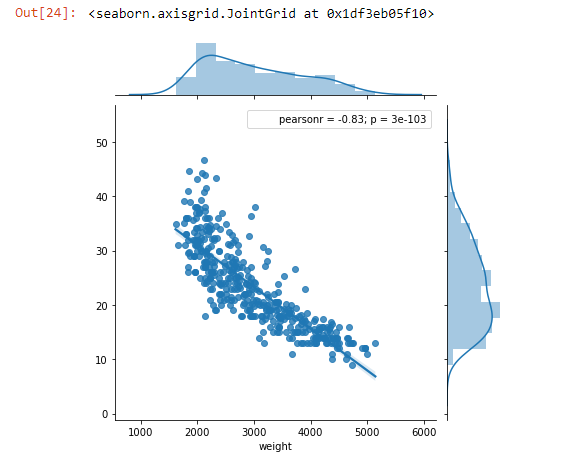
 

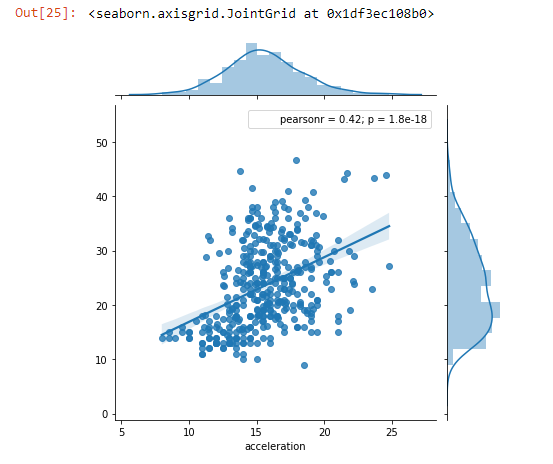
Here we're locating the Pearson’s coefficient and p fee of every attribute. The decrease the p fee the higher the correlation and better the importance of that attribute.

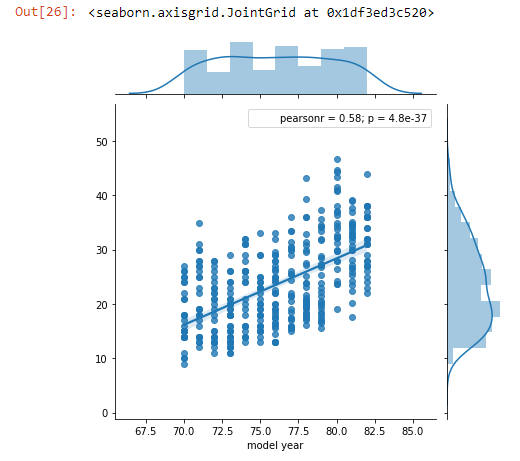
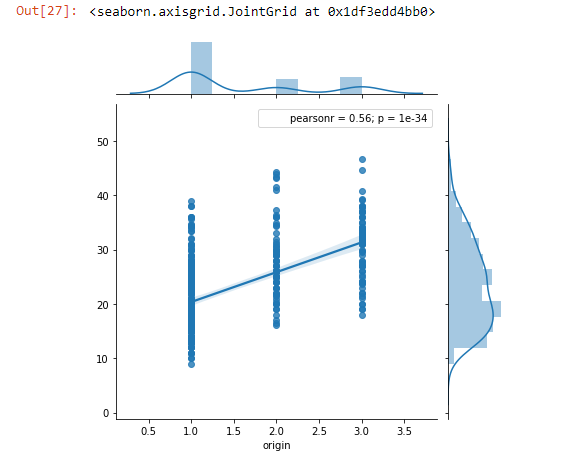
 

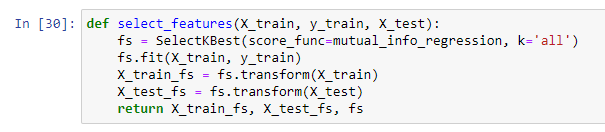
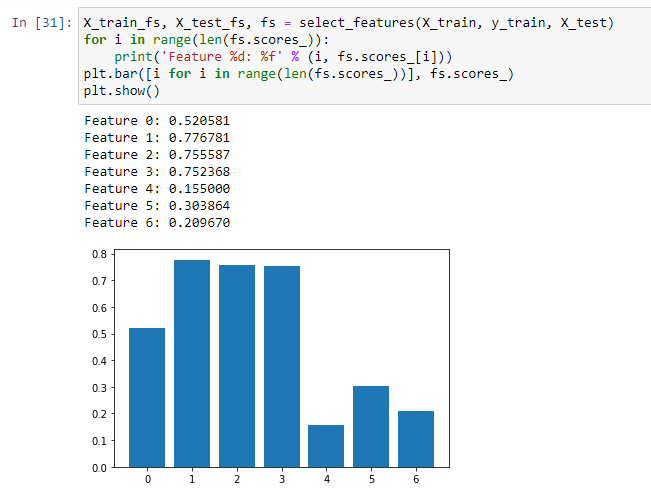
   

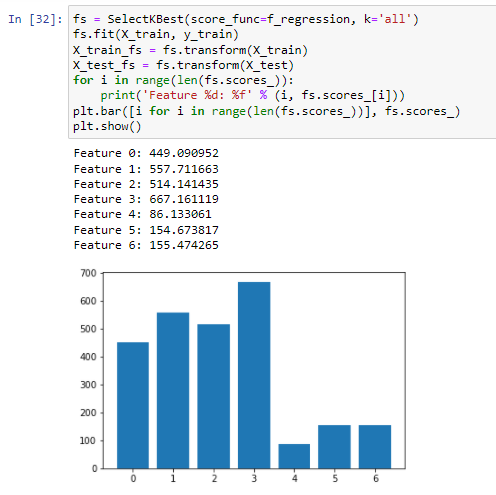
Here we have visualised all the attributes and their correlation with increase/decrease in y values of training set. We can see that attribute including ‘cylinders’, ‘displacement’, ‘horsepower’ and ‘weight’ are highly correlated with y values. There is not much correlation between the other attributes but the closes to fifth spot comes in with ‘origin’ and ‘model year’ attributes.

Here we are trying to plot the scores graph of mutual\_info\_regression.

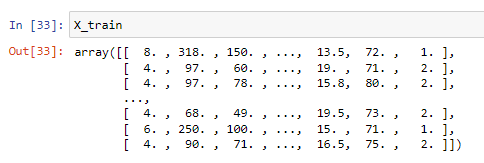
We can see that the attributes with highest scores in order are:

1. Displacement
2. Horsepower
3. Weight
4. Cylinders and
5. Model Year



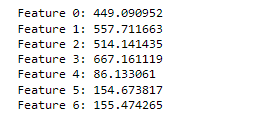
Similarly, here we have plotted the numerical and graphical values of scores. The highest scores in order here are:

1. Weight
2. Displacement
3. Horsepower
4. Cylinders and
5. Origin



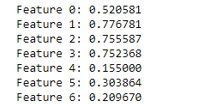
Results and Conclusion

The five features according to f\_regression are:



* Weight 🡪 667.16
* Displacement🡪 557.71
* Horsepower🡪 514.14
* Cylinders 🡪 449.09
* Origin🡪 155.47

The five features according to mutual\_info\_regression are:



* Displacement🡪 0.7767
* Horsepower🡪 0.7558
* Weight🡪 0.7523
* Cylinders 🡪 0.5205
* Model Year🡪 0.3038