Unsupervised Learning and Agents

Report

Ka Po Chau (ka-po.chau@epitech.eu)
Dorian You (dorian.you@epitech.eu)

Overview

This project is done for the Epitech module Unsupervised Learning and Agents.

There are 5 tasks:

1. data distribution and the law of large numbers
2. meteorological data: dimensionality reduction and visualization
3. company clustering customers
4. exploitation / exploration compromise
5. application of unsupervised learning

Work distribution

Part 1: Ka Po Chau
Part 2: Ka Po Chau
Part 3: Ka Po Chau
Part 4: Dorian You
Part 5: Ka Po Chau

Part 1: data distribution and the law of large numbers

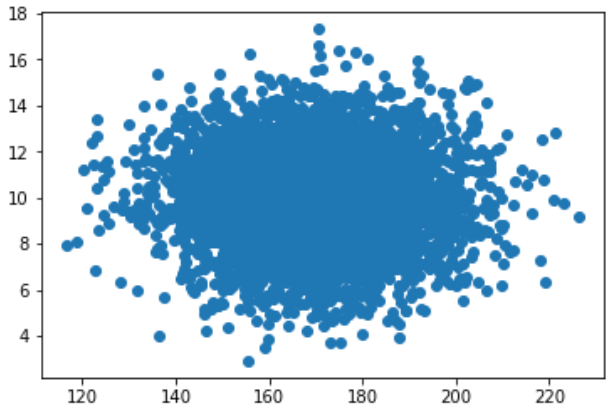
I propose X and Y as the height and finger length of a person in a population.

Both X and Y are continuous random variables.

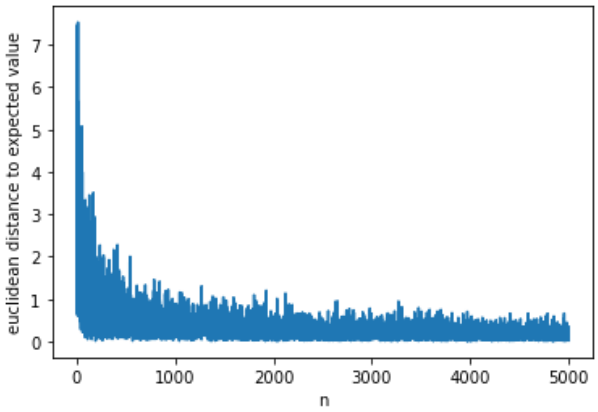
X, height, will have the mean of 170cm and STD of 15cm. Y, finger length, will have the mean of 10cm and STD of 2cm.

Z = (X, Y)

Now, 5000 points are sampled from the law of Z and plotted in a 2 dimensional figure:



After this, the euclidean distance to the expected value is plotted:



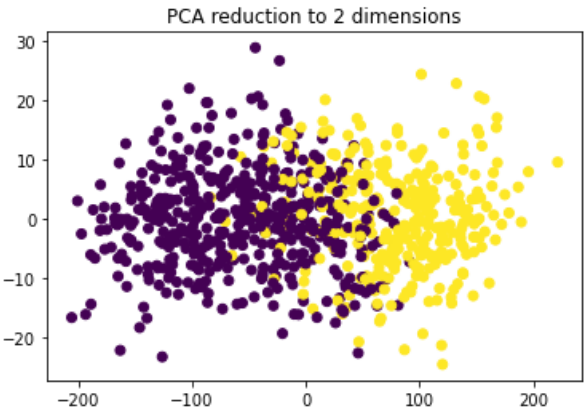
As we can see from the above graph, the bigger n is, the smaller euclidean distance is. This means that the empirical mean is closer to the expected value as n increases.

Part 2: meteorological data: dimensionality reduction and visualization

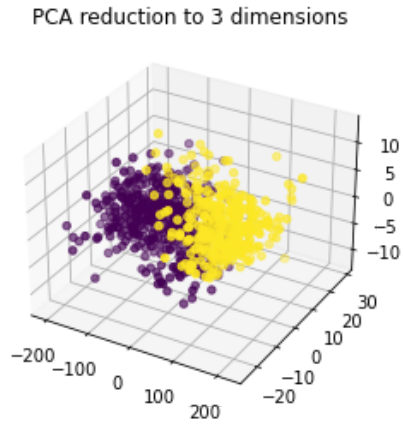
In this part, we will try to reduce the dimensionality of the given data into both dimensions 2 and 3, and then plot them onto scatter plots.

I used scikit-learn to plot the results with PCA method.

First, we will try a 2 dimension reduction with PCA, then plot the results.



After this, we try doing 3 dimensions with PCA:



As we can see from the 2 scatter plots, dimension 2 seems to allow to predict the label based on the projected components only, better than 3.

The answer is therefore 2.

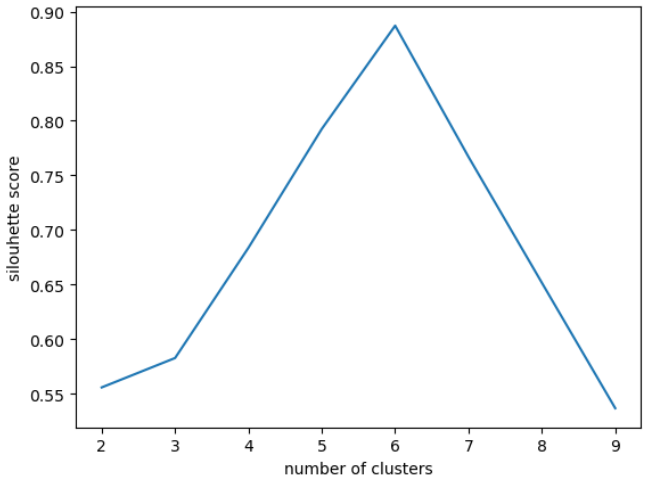
Part 3: company clustering customers

In this part, we will do 2 clustering methods: Kmeans and Spectral, and 2 cluster analysis methods: Silhouette and Elbow.

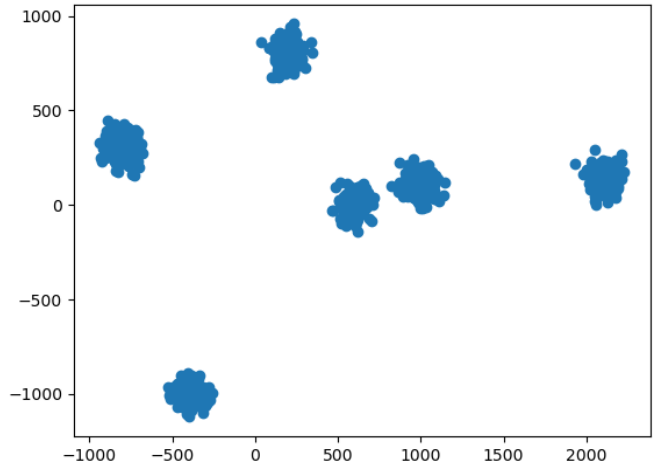
Silhouette Method with Kmeans clustering

First, we will use the Silhouette Method from sklearn to determine the best number of clusters.

I will run the sklearn silhouette_score from the range 1-10.

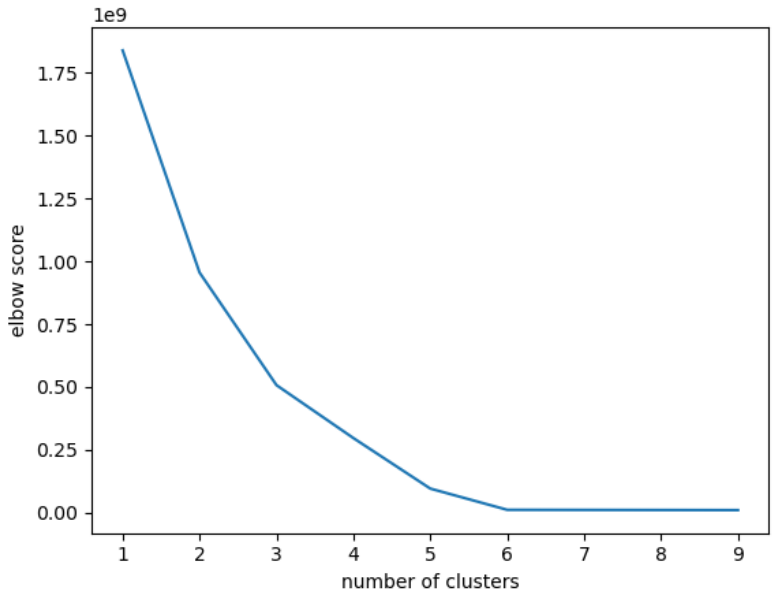


As we can see from the above plot, 6 clusters is clearly the biggest silhouette score, therefore we will now use 6 as the input number of clusters and do the Kmeans clustering method.

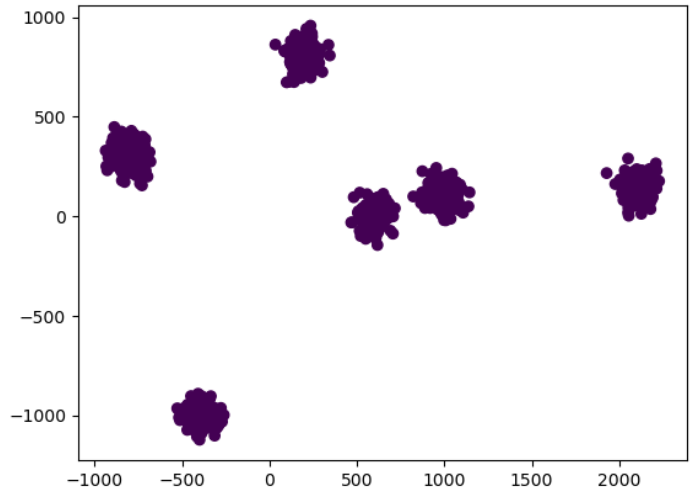


Elbow Method with Spectral Clustering

Next, we will determine the number of clusters with elbow method.



As we can see from the above Elbow plot, the optimal number of clusters is also 6. Now, we will use 6 as input for Spectral Clustering.



From the above 2 plots, both Kmeans+Silhouette and Spectral+Elbow gave very similar results.

But, Kmeans should be better for larger number of clusters and Spectral should be better for smaller number of clusters.

Part 4: Exploitation/Exploration Compromise

In this part, we will explain how the system of movement of our agent works.

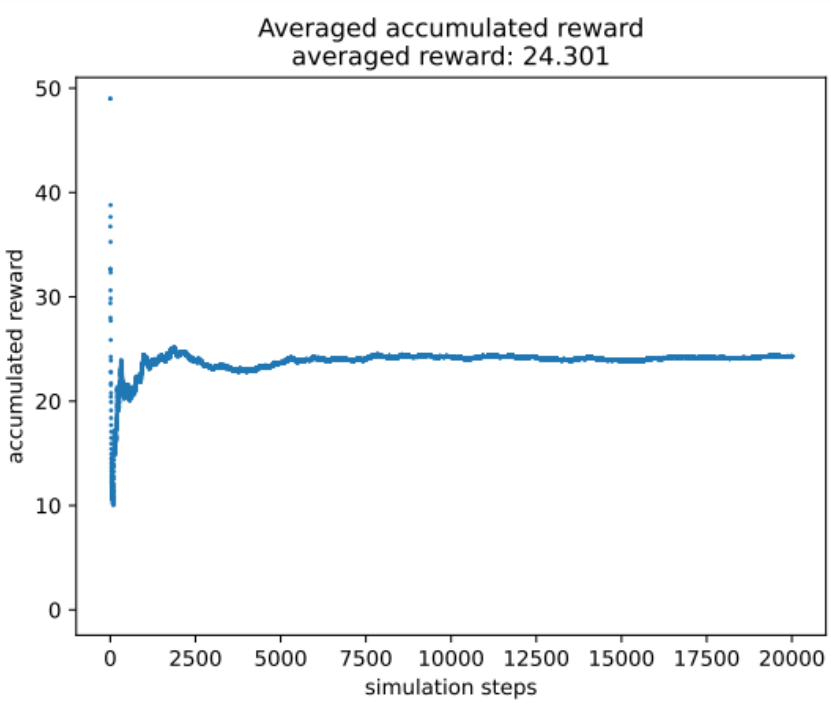
First of all we get all the information we need from our agent (his position and the emplacement of all rewards known.

With the position, we can determine the possible range of movement of our agent (He can't move right if he is at the border right and he can't move left if he is at the border left).

After getting the possible action for our agent, we start by selecting one randomly. This step is crucial to try to determine where the emplacement with rewards are.

If one of the cases next to our agent contains a known reward, we don't move randomly but we move in the direction of the case with the biggest reward.

With this we obtain a result like this,



We don't obtain the same result each time because there is randomness in the simulation and in the movement system.

Part 5: application of unsupervised learning

In this part, we will use the dataset "Cars".
(https://perso.telecom-paristech.fr/eagan/class/igr204/datasets)

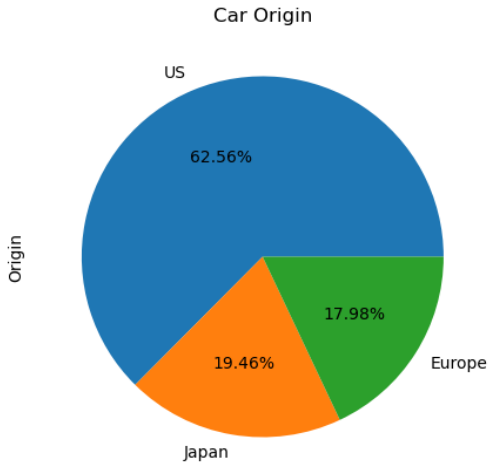
First, here is how the dataset actually looks:

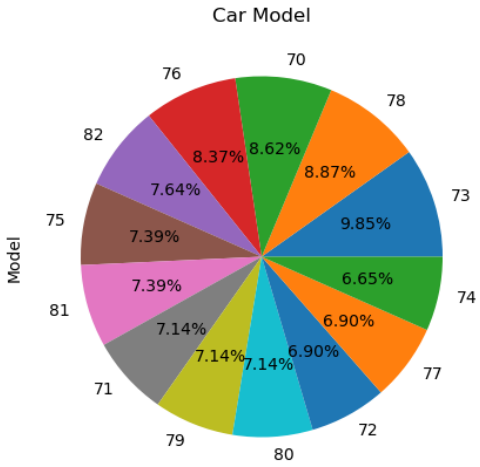
	Car	MPG	Cylinders	Displacement	Horsepower	\
0	Chevrolet Chevelle Malibu	18.0	8	307.0	130.0	
1	Buick Skylark 320	15.0	8	350.0	165.0	
2	Plymouth Satellite	18.0	8	318.0	150.0	
3	AMC Rebel SST	16.0	8	304.0	150.0	
4	Ford Torino	17.0	8	302.0	140.0	

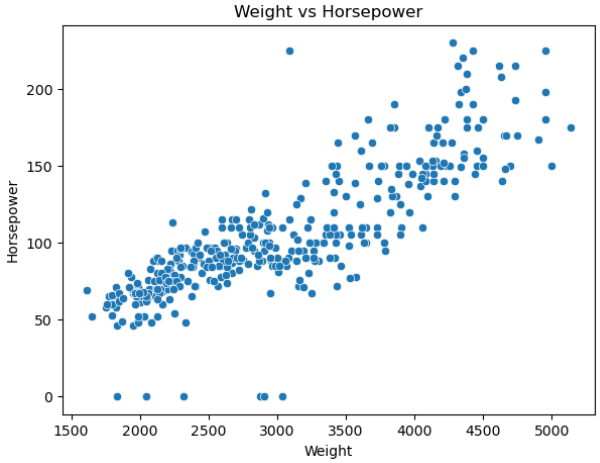
	Weight	Acceleration	Model	Origin
0	3504.0	12.0	70	US
1	3693.0	11.5	70	US
2	3436.0	11.0	70	US
3	3433.0	12.0	70	US
4	3449.0	10.5	70	US

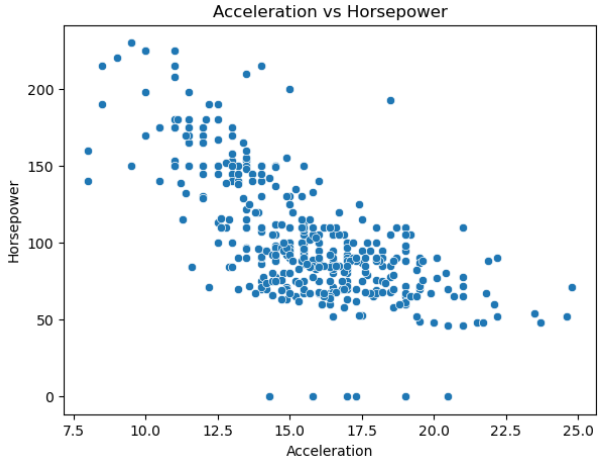
General Analysis of the dataset:

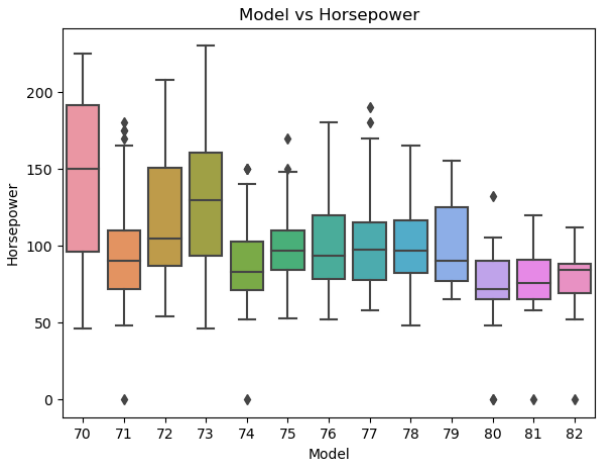
```
===============
Horsepower
===============

Maximum: 230.0
Minimum: 0.0
Mean: 103.5295566502463
Std: 40.52065912106341

===============
Weight
===============

Maximum: 5140.0
Minimum: 1613.0
Mean: 2979.4137931034484
Std: 847.0043282393509
```

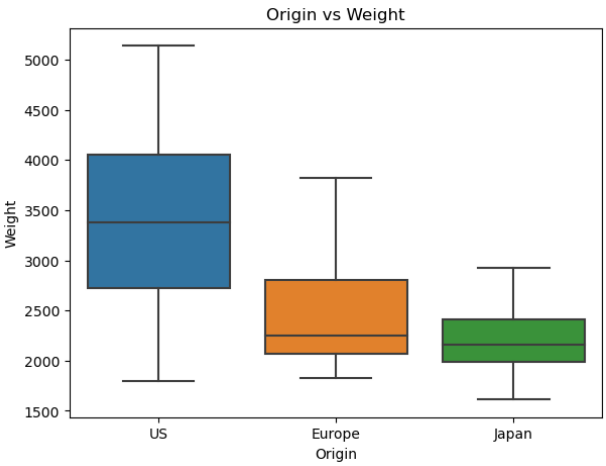

Car Origin
US
62.56%
Origin
17.98%
Europe
19.46%
Japan


Car Model
70
76
78
82
8.37%
8.62%
8.87%
7.64%
73
75
9.85%
7.39%
Model
6.65%
74
7.39%
81
6.90%
7.14%
7.14%
6.90%
77
71
72
79
80

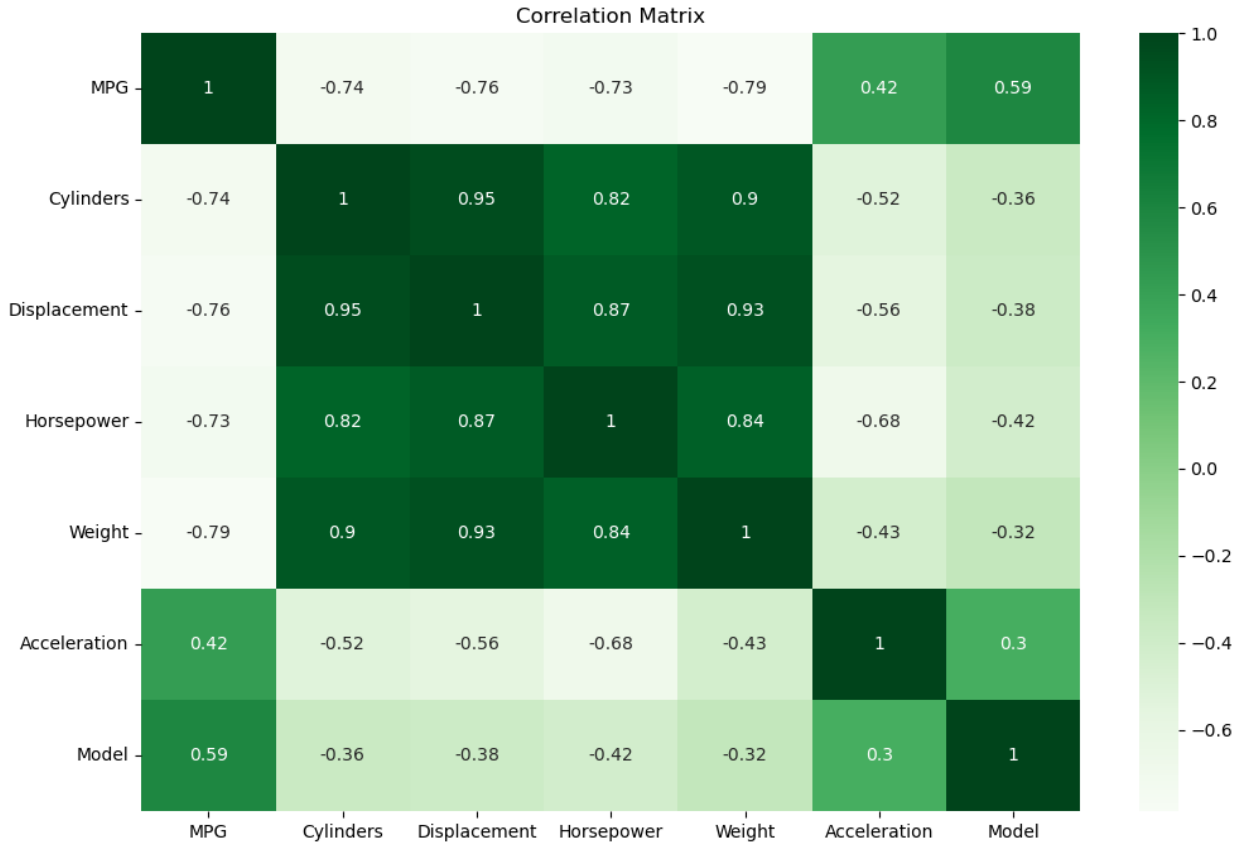

Weight vs Horsepower
Horsepower
Weight


Acceleration vs Horsepower
Horsepower
Acceleration


Model vs Horsepower
Horsepower
Model


Origin vs Weight
Weight
US
Europe
Japan
Origin

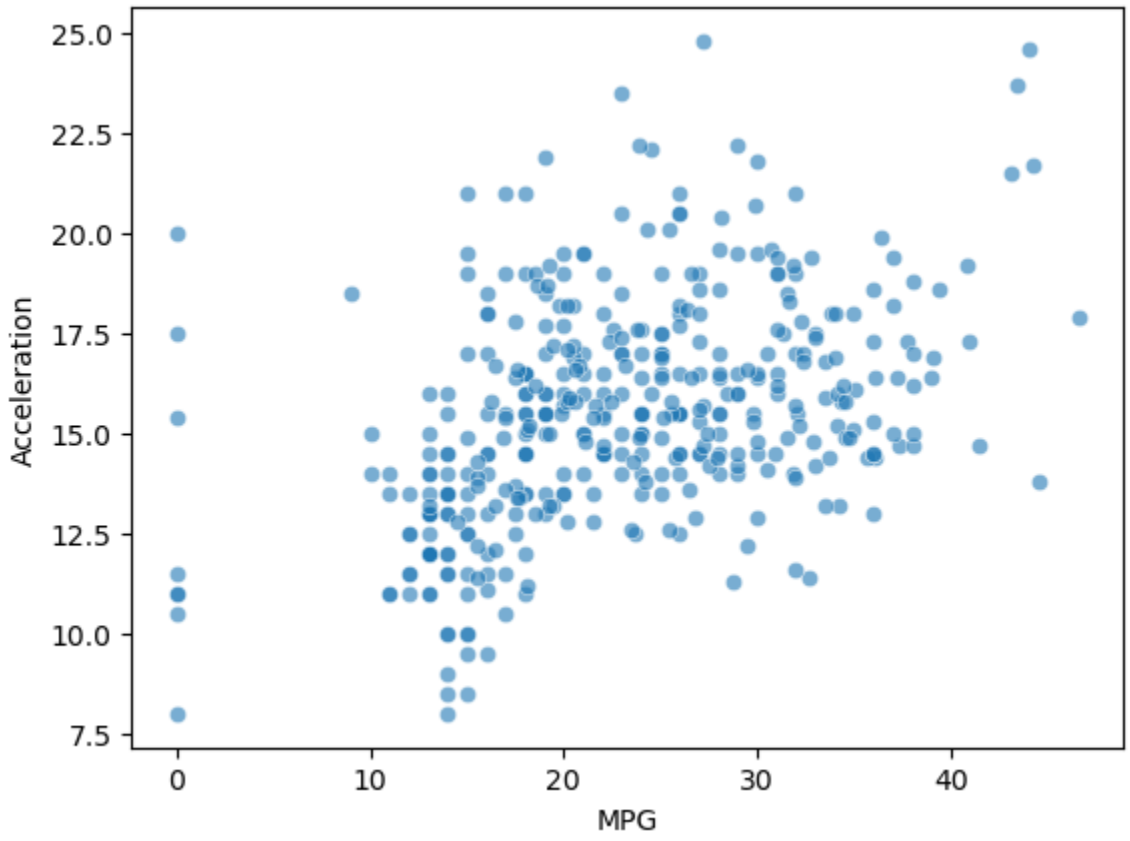
Correlation Matrix:

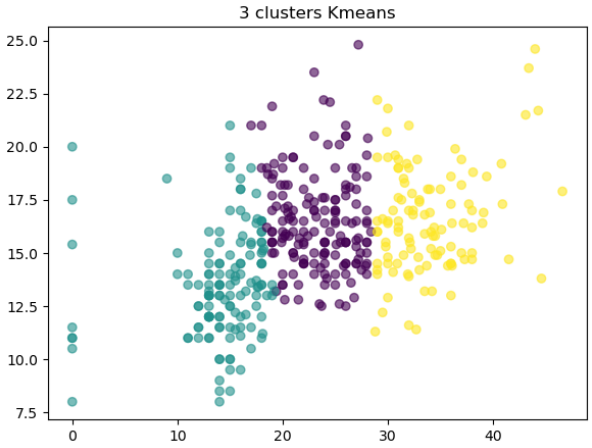


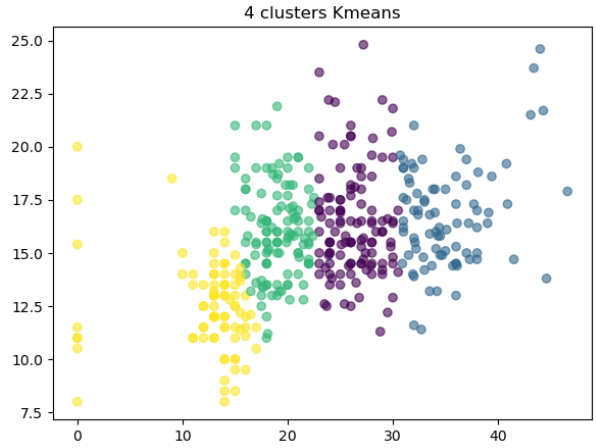
Kmeans clustering of MPG and Acceleration

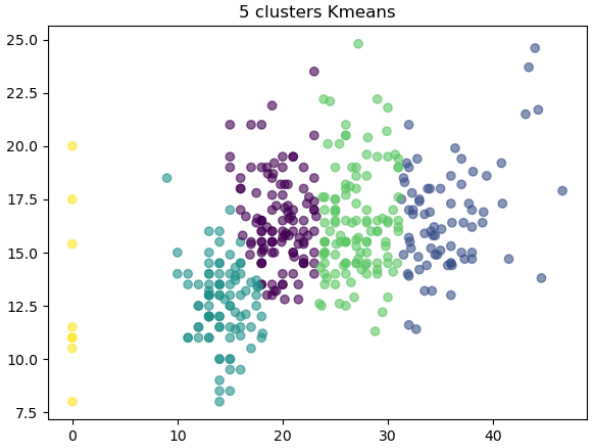
Now I will try to do a Kmeans clustering of the car's MPG and Acceleration.

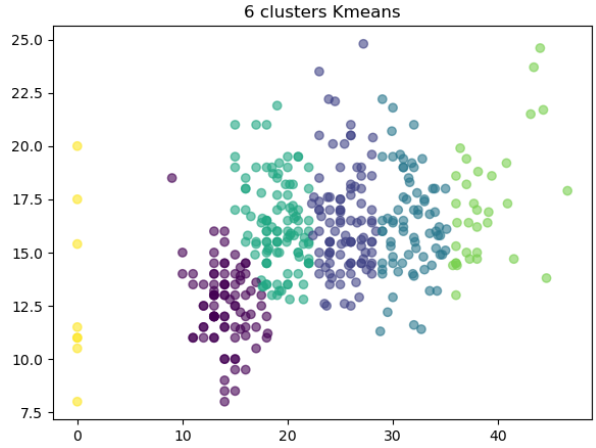
Here is the scatter plot of these variables.



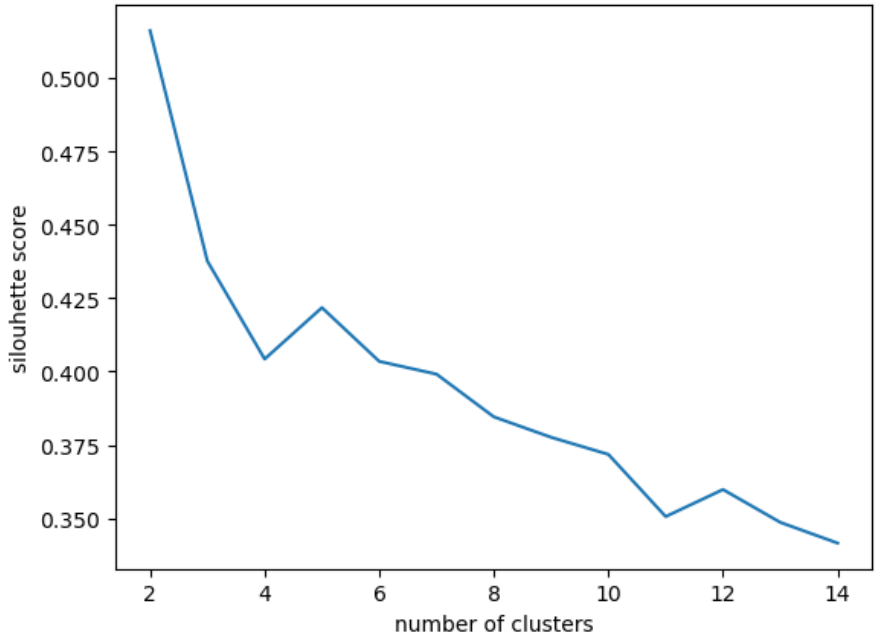

3 clusters Kmeans
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
0
10
20
30
40


4 clusters Kmeans
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
0
10
20
30
40


5 clusters Kmeans
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
0
10
20
30
40


6 clusters Kmeans
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
0
10
20
30
40

Silhouette Score of clustering



Elbow Method Score of clustering