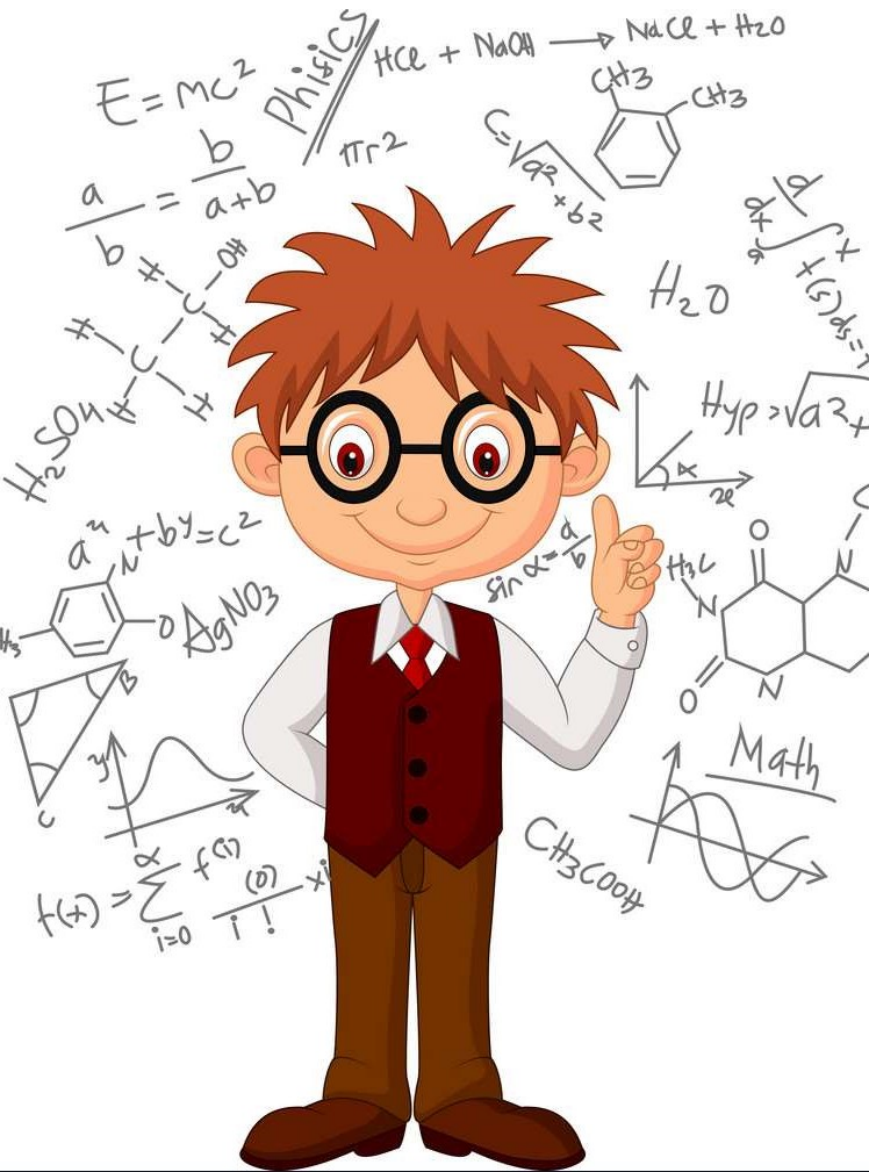
Predictive Data Analytics

Lecturer: Marina Iantorno
E-mail: miantorno@cct.ie
July 2022



In today's class we will cover:

- ❑ Hierarchical clustering
- ❑ K-means clustering
- ❑ Practice on Python

Recap about unsupervised learning

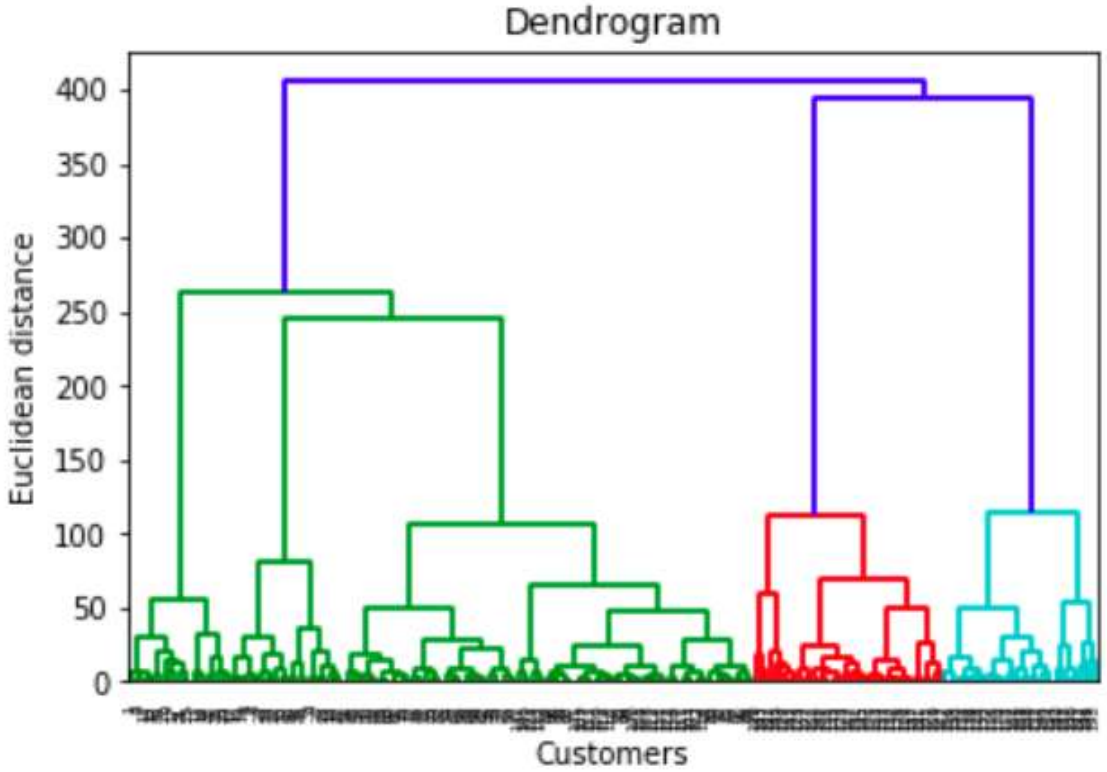
Unlike the supervised learning, with unsupervised learning we do not know our target, we do not have the target labels of classification, we do not know in advance how our data will be classified.

When we work with unsupervised learning, it is the responsibility of the researcher to conclude, identify and tag the groups after clustering. For instance, after clustering, you may discover that your customers are divided into groups based on their willing to buy a certain product.

Clustering

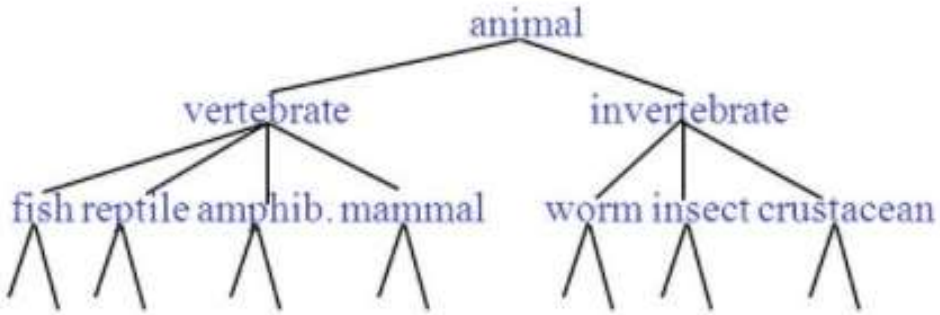
Clustering was originally developed by anthropologists to explain the origin of the human being. Later, other disciplines started using it, and now, in Big Data times, it is very common to use clustering techniques when we work with raw data.

There are two types of clustering:



Hierarchical Clustering

Hierarchical Clustering Algorithm

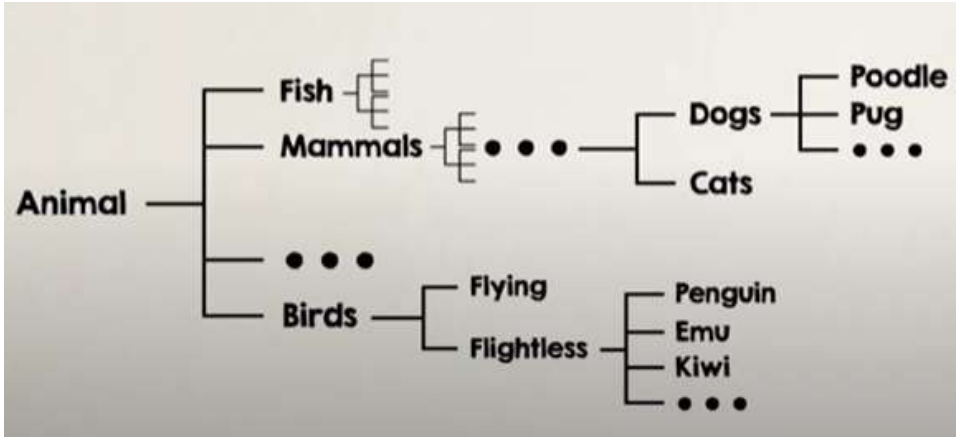


We will find two types of hierarchical clustering: Agglomerative (bottom-up) or divisive (top-down). With the divisive clustering we start with the premise that all the observations are in the same cluster, like this example with the animals.

Hierarchical Clustering Algorithm

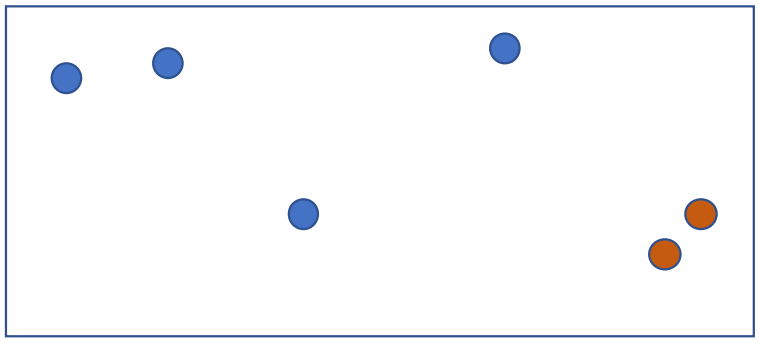
When we work with the agglomerative, we start from a certain point and we want to reach until the last species of each category.

Both methods should reach to similar result, but from the Mathematical perspective the agglomerative is easier to solve, and therefore, this is the most commonly used. What we will proceed is the agglomerative hierarchical clustering.



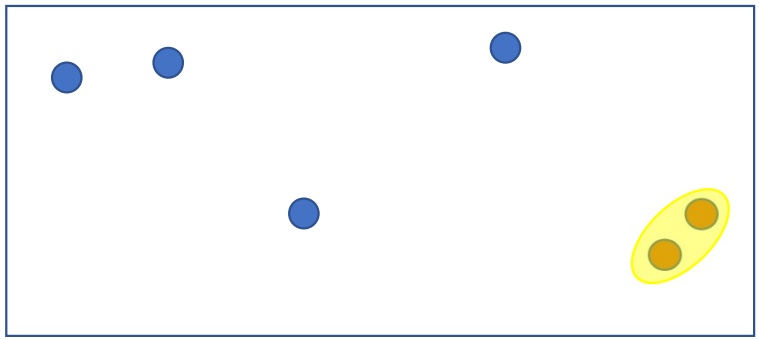
Hierarchical Clustering Algorithm

The hierarchical clustering method is unsupervised, and it is used to categorise or cluster the data into several categories.



Suppose that here each datapoint is considered as a cluster by itself. Let's say that we need to find the minimum distance between any two points.

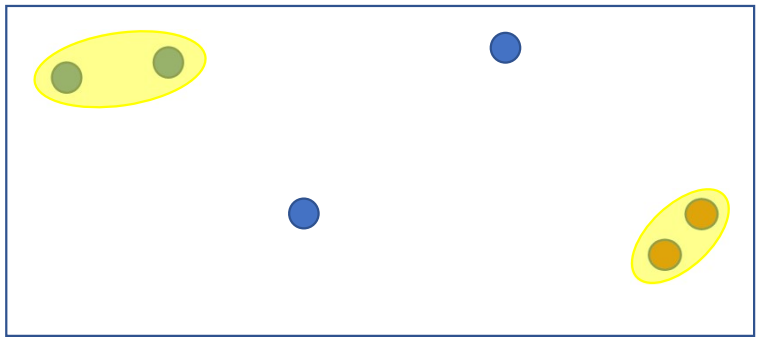
Hierarchical Clustering Algorithm



Let's say that the minimum distance is between these two datapoints. As you may already know, there are several methods to measure the distance between the dots, and one of them is the Euclidean method, the same one that we used in the K-NN and we will use the same one to build this model.

Based on the distance, this algorithm consider these datapoints as one cluster.

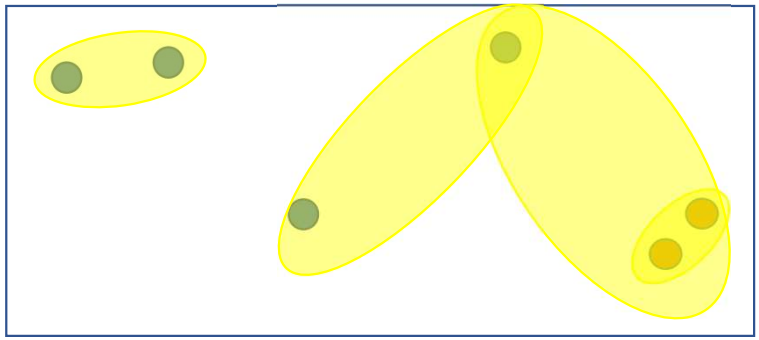
Hierarchical Clustering Algorithm



If we look at the minimum distance between any other two datapoints, and we found out that the minimum distance compare with all the distances is the one between these two datapoints selected.

Once again, the algorithm consider these datapoints as another cluster.

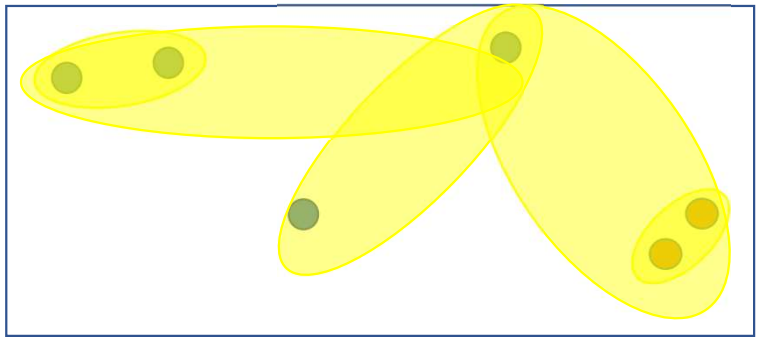
Hierarchical Clustering Algorithm



If we keep looking, the algorithm identifies that the datapoint is very close to the first cluster, so it considers everything as one cluster.

And later we can find out that the only dot left is close to the bug cluster, so the algorithm takes it and creates a bigger cluster.

Hierarchical Clustering Algorithm



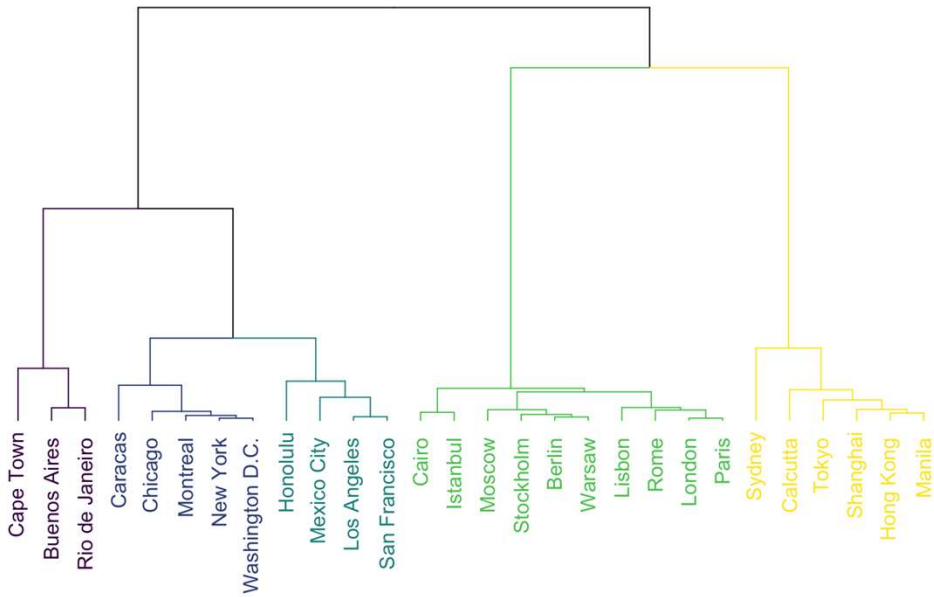
Finally, we could have everything grouped as a one cluster, but what would be the point of clustering the data if we would have everything in only one block?

The idea es to identify where is the best place to stop clustering the data, and to do that, there is a technique that the algorithm can use to determine how many cluster we will have.

Hierarchical Clustering Algorithm

The technique to identify how many clusters we will have is called “Dendrogram”.

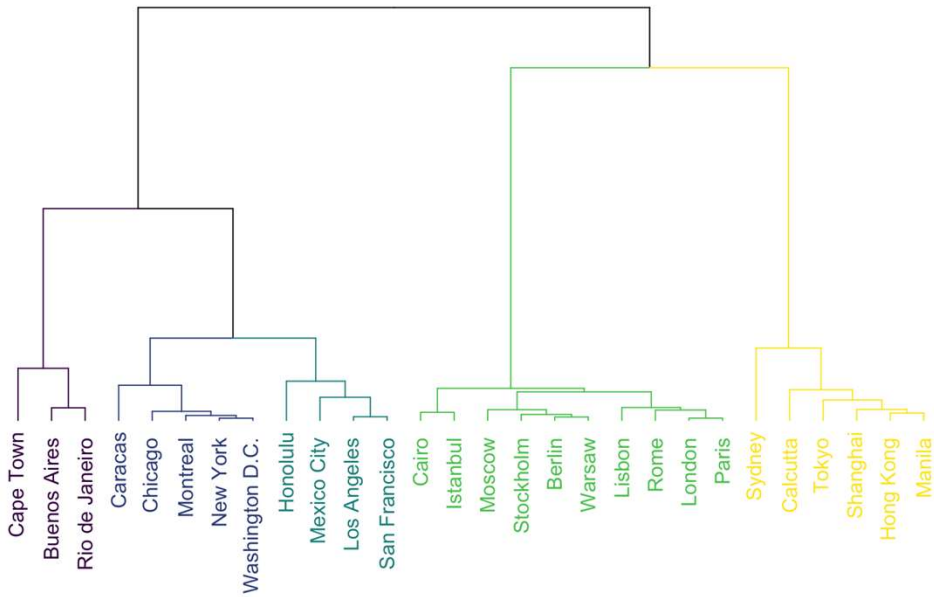
The dendrogram is a diagram that it is used to determine the optimal number of clusters.



Hierarchical Clustering Algorithm

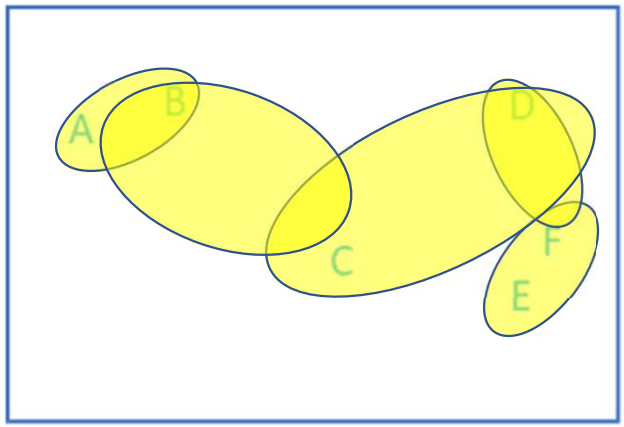
The technique to identify how many clusters we will have is called “Dendrogram”.

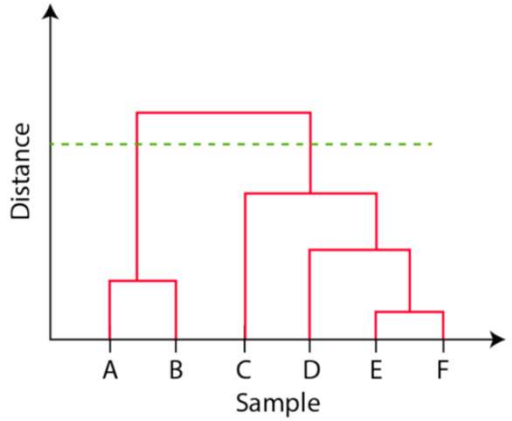
The dendrogram is a diagram that it is used to determine the optimal number of clusters.



Hierarchical Clustering Algorithm

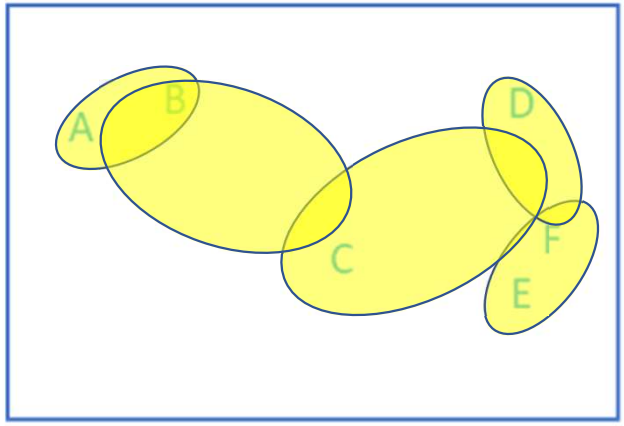
Imagine that we have the datapoints below and we want to create the Dendrogram

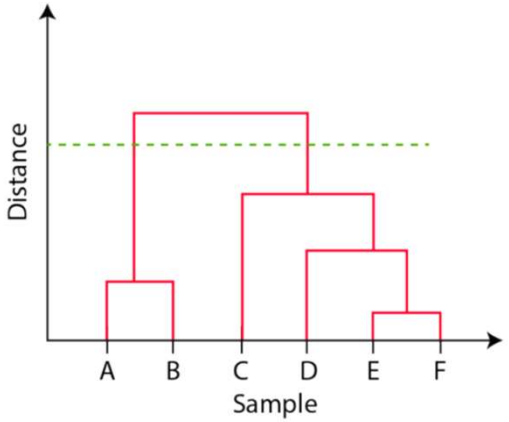




Hierarchical Clustering Algorithm

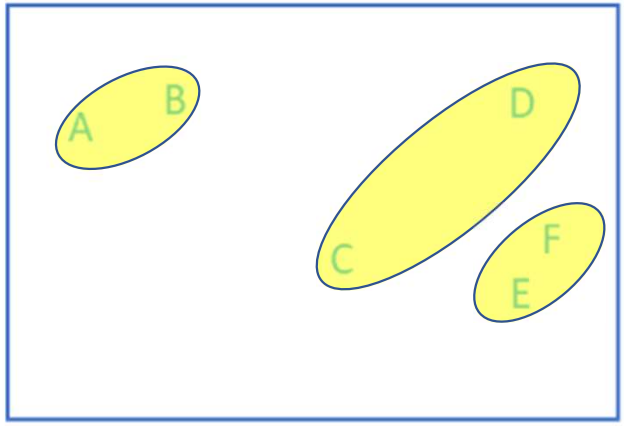
These two lines are the only ones that were not crossed by any horizontal line.

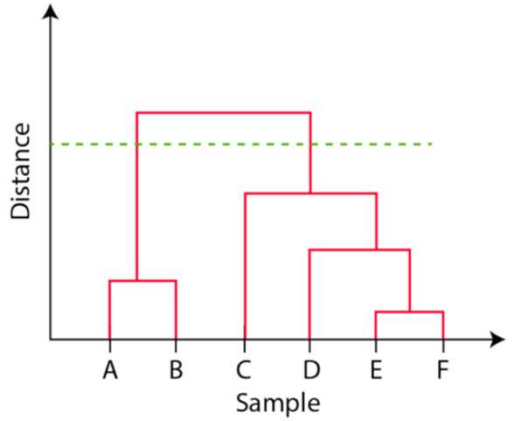




Hierarchical Clustering Algorithm

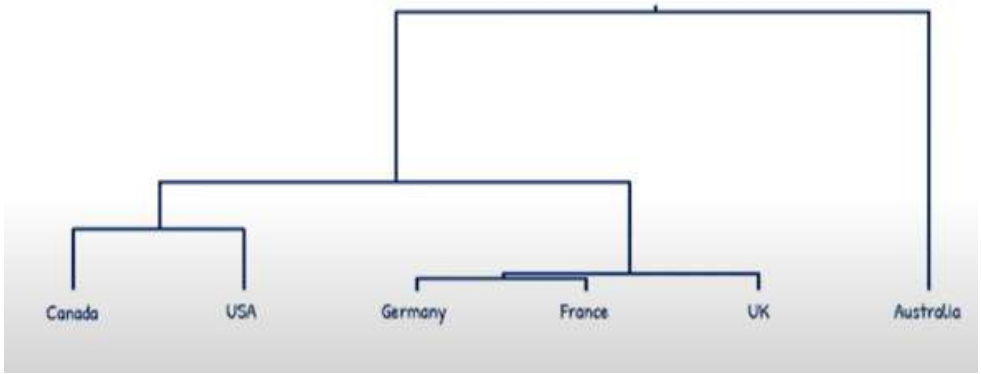
Imagine that we crossed a “virtual horizontal line” between the two vertical lines with the biggest scope, and the line is only crossing two vertical lines, therefore, our ideal number of cluster in this case would be 2.





Hierarchical Clustering Algorithm

There is not a real rule to draw the line of the dendrogram, but after working on that you will develop certain intuition. Let's analyse the following dendrogram together and let's see how would we divide the data.



Hierarchical Clustering Algorithm

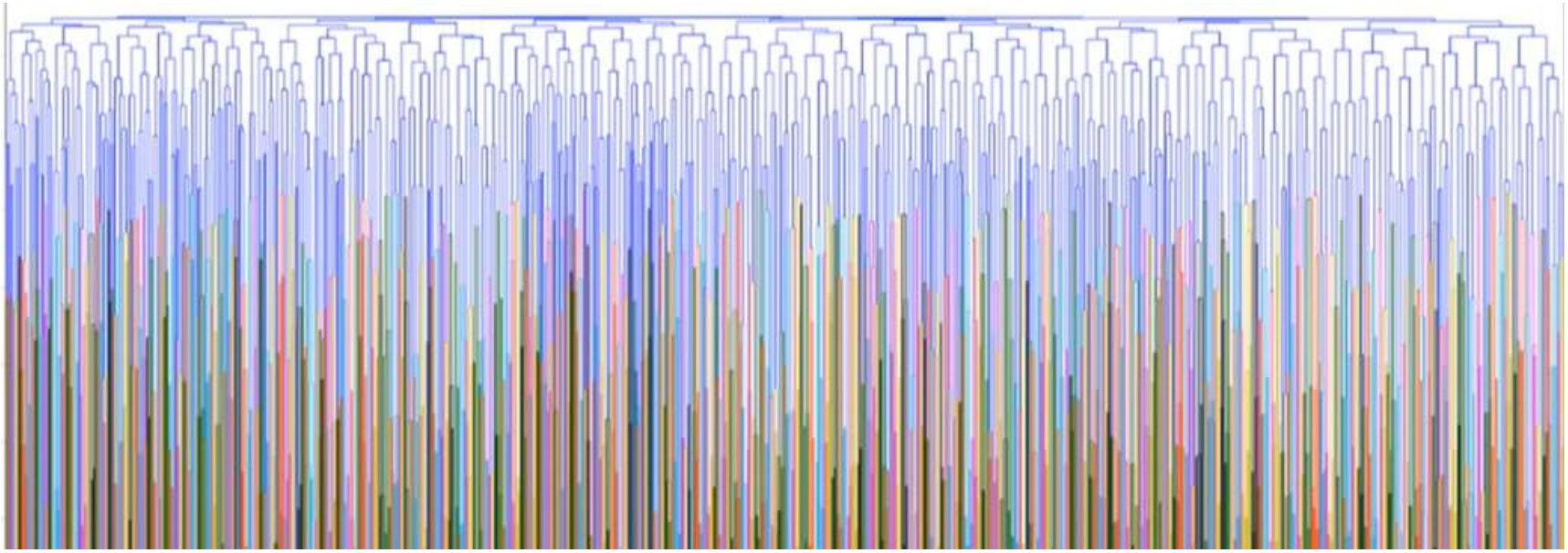
As a positive aspects we can mention:

- Hierarchical clustering shows all the possible linkages between clusters.
- It gives us the possibility to understand the behaviour of the data in an easy way.
- There is no need to pre- set the number of clusters, we find them on the way.

But the biggest con…

- It may not work very well when our dataset is so large.
- It may require a lot of money because at larger the dataset gets, as slower will Python work, and therefore we will need a better machine to support this model.

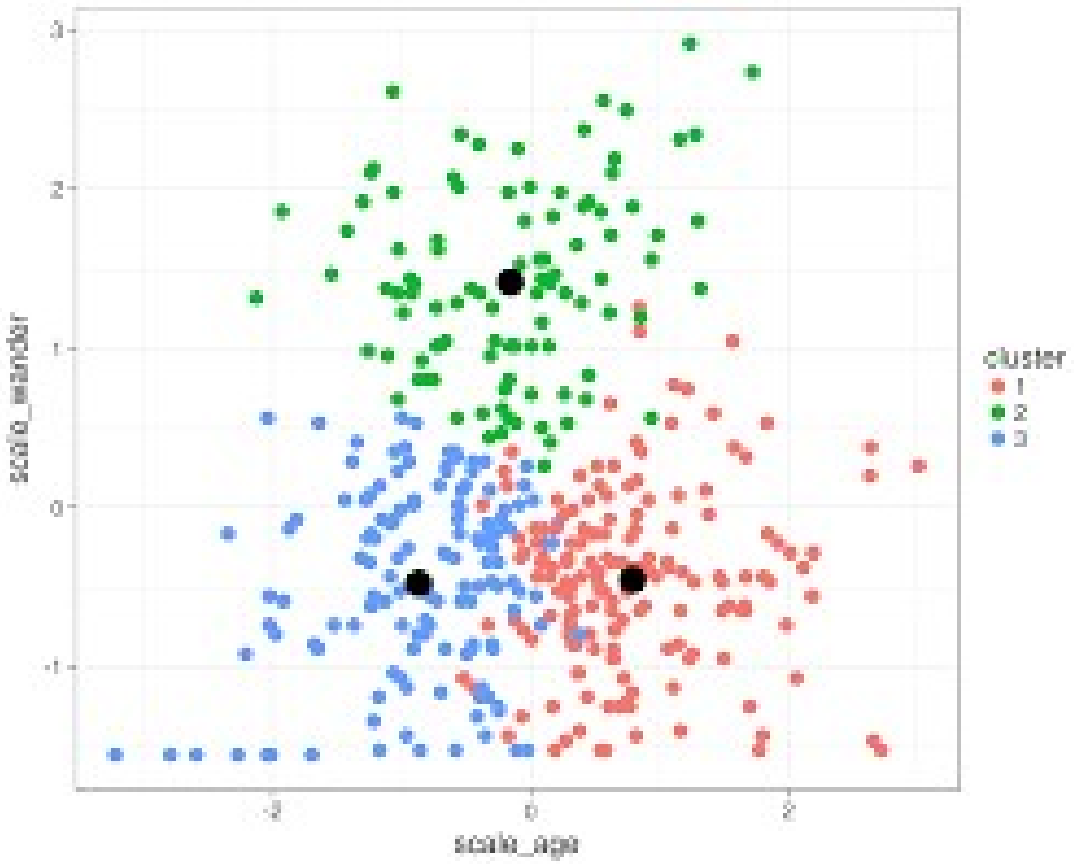
Hierarchical Clustering Algorithm



Hierarchical Clustering Algorithm

Let’s work now with an example, to make it clearer.

You will find on Moodle a dataset called “Movies.csv”. There, you will see a dataset with three columns that contain the Production Budget of certain movies, the worldwide gross income per movie and the genre. Let’s try it on Python.

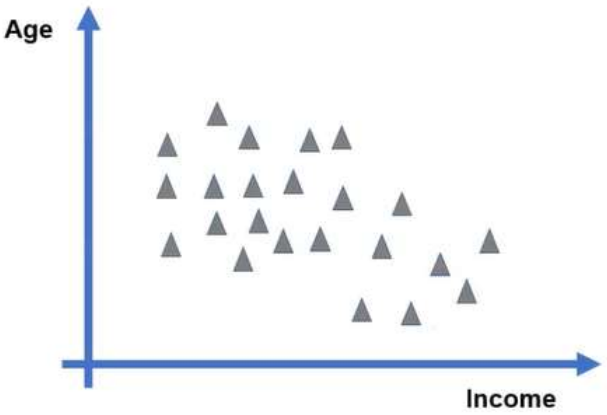


K-Means Clustering

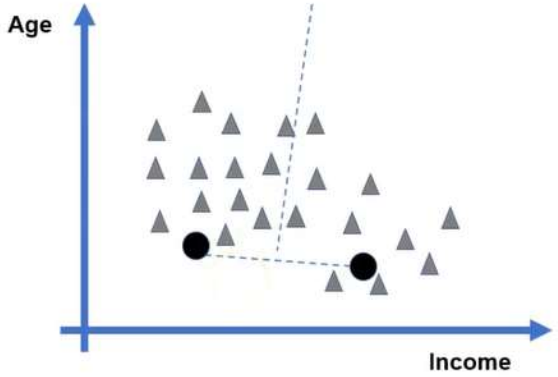
K-Means

It is common to confuse K-NN with K-Means, but they are different models. K-NN is a supervised model and it is used to classify the data, whereas K-Means is an unsupervised model that we use to relate groups and create clusters. Let's try to understand how this model works.

K-Means

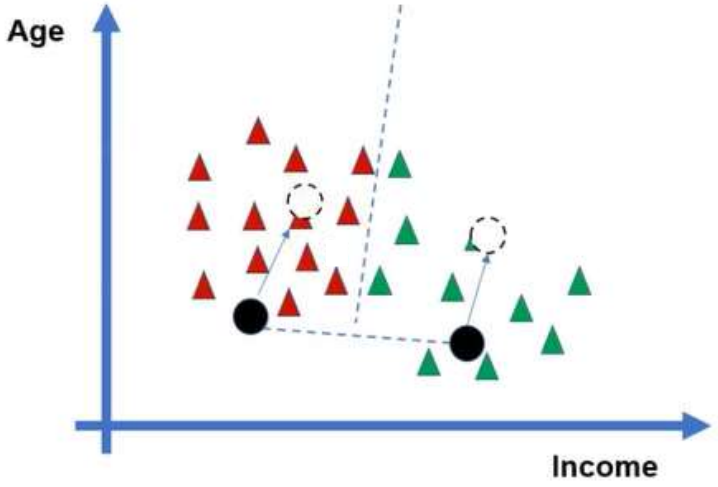


Suppose that we have datapoints related to these variables (Age and Income).

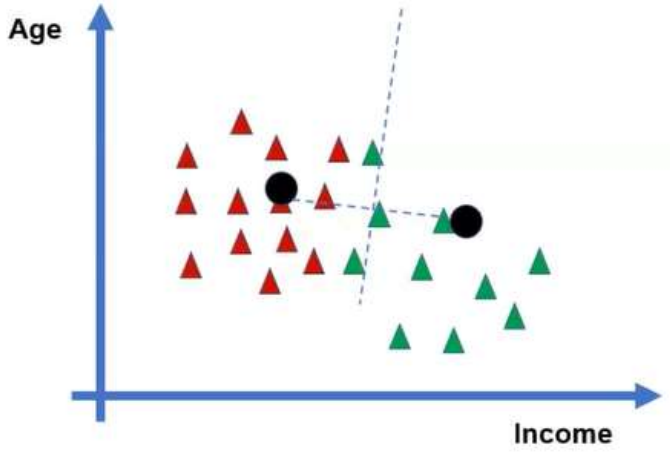


Now suppose that we have two dots and we draw a line with another line in the middle with the same distance from one point to another one.

K-Means

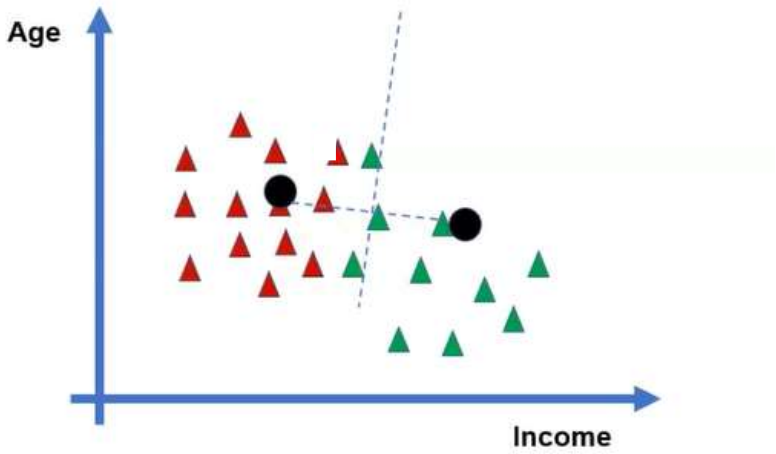


Based on these dots and the lines, we can divide out data in 2 clusters. Now we need to find the centre of these new clusters to see if some data is moving from the line.



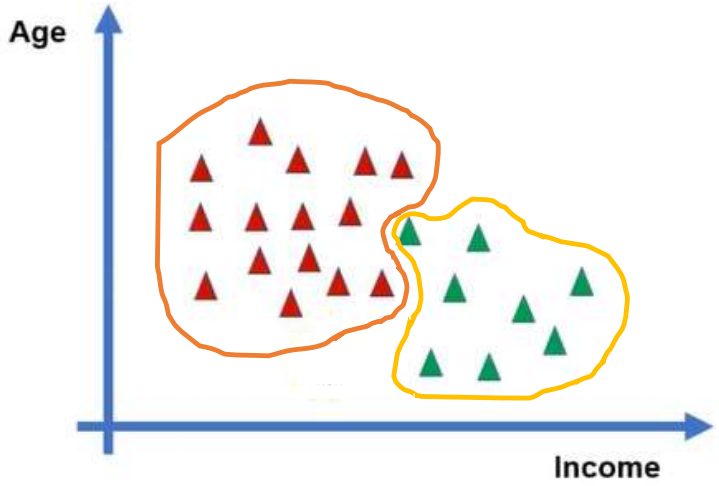
Now we repeat the practice of drawing the lines between these dots and try to identify if there are some movements.

K-Means



And we do the process again, and again, until there is no more changes in the data (no data is moving to one side or another one). At this stage we stop the process and we can say that we have “k” clusters (in the example, 2 clusters).

K-Means

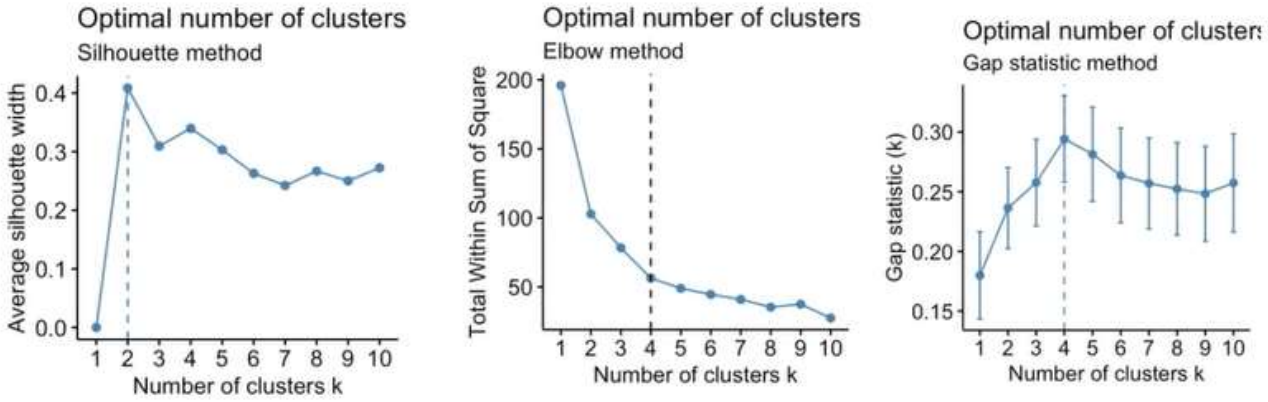


Now at this point we can imagine a couple of line bordering the clusters, and we reached to the point that we wanted: divide the data ensuring that there is no datapoint going to the other side of the line.

Now we may wonder, in the practice, how to choose the optimal number of clusters?

K-Means

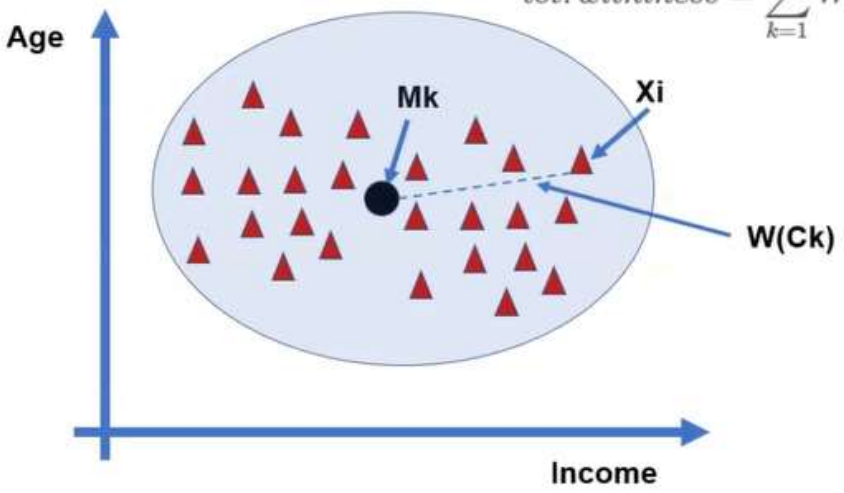
We need a method to select the optimal number of clusters. In this case, there are a few methods to choose.



Each of these methods has its own maths and procedure to reach to the “k”. We will use the Elbow Method, because it is efficient, it is similar to topics that we already know and it is straight forward.

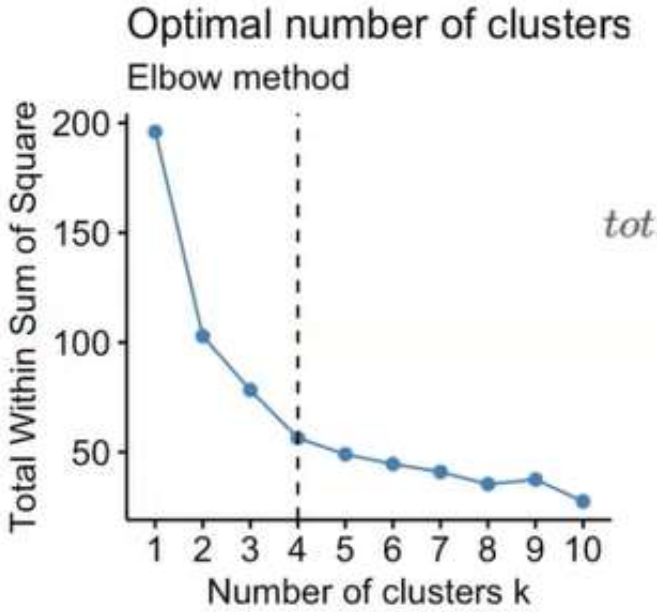
K-Means

$$tot.withiness = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$



We will have as many summaries as number of clusters.

K-Means



$$tot.withiness = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

In this case, the best number of clusters would be 4.

K-Means

Let’s work now with the “Movies.csv” dataset.

Practice Time!

Practice Time

You will find on Moodle a dataset called “Mall_Customers.csv”. Proceed to cluster with the Hierarchical algorithm and K-Means.

You will find the scripts with the results this weekend.

Best of luck!

THAT'S ALL FOR TODAY

THANK YOU