Quantitative analysis methods for public policies

Clustering

Co-funded by the Erasmus+ Programme of the European Union

The idea

- “Help me understand our customers better so that we can market our products to them in a better manner!”
- You are not looking for specific insights for a phenomena, but what you are looking for are structures with in data with out them being tied down to a specific outcome.
- The method of identifying similar groups of data in a data set is called clustering. Entities in each group are comparatively more similar to entities of that group than those of the other groups.

Co-funded by the Erasmus+ Programme of the European Union

Example

Suppose, you are the head of a political party and wish to understand preferences of your voters to scale up your poll numbers. Is it possible for you to look at details of each voter and devise a unique strategy for each one of them? Definitely not. But, what you can do is to cluster all of your voters into say 10 groups based on their opinions / habits and use a separate strategy in each of these 10 groups. And this is what we call clustering.

Co-funded by the Erasmus+ Programme of the European Union

Types of Clustering

- **Hard Clustering: In hard clustering, each data point either belongs to a cluster completely or not. For example, in the above example each voter is put into one group out of the 10 groups.**
- **Soft Clustering: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, from the above scenario each voter is assigned a probability to be in either of 10 clusters.**

Co-funded by the Erasmus+ Programme of the European Union

Many types of clustering algorithms

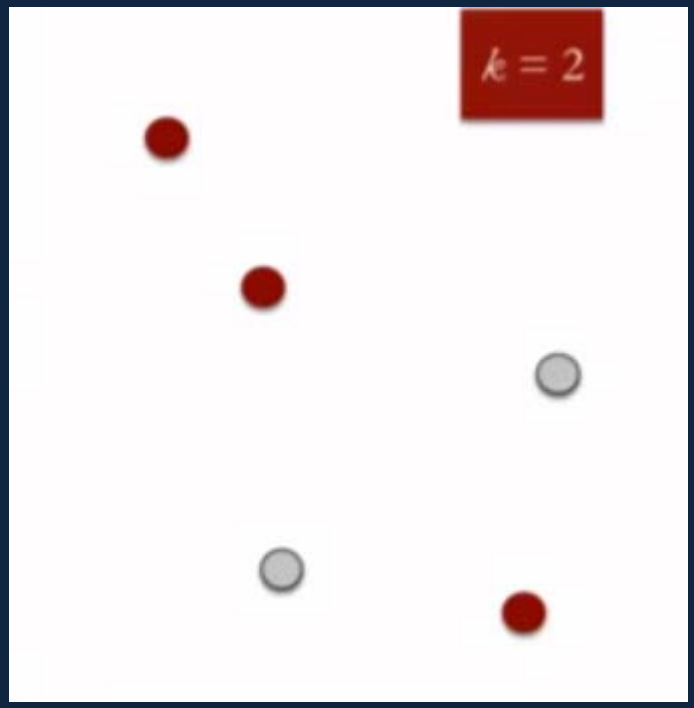
- K Means Clustering: The number of clusters is specified
- Hierarchical Clustering: Builds hierarchy of clusters.

- In practice K means is used more often

Co-funded by the Erasmus+ Programme of the European Union

K Means

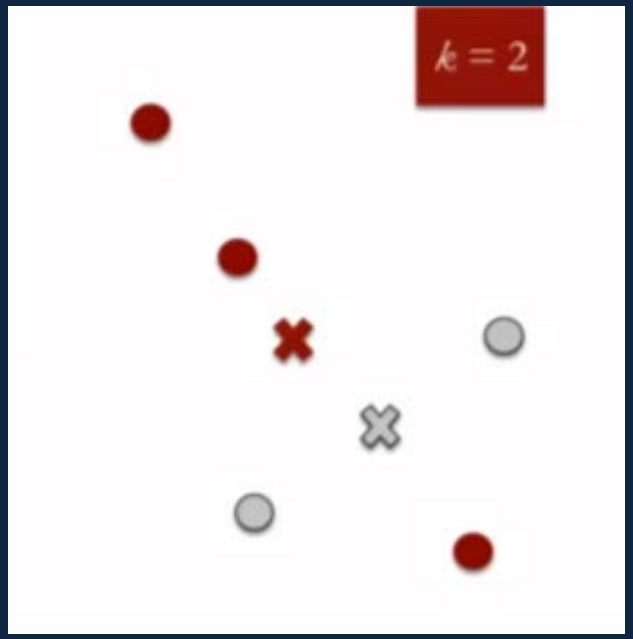
- Specify the desired number of clusters K
- Randomly assign each data point to a cluster



Co-funded by the Erasmus+ Programme of the European Union

K Means (2)

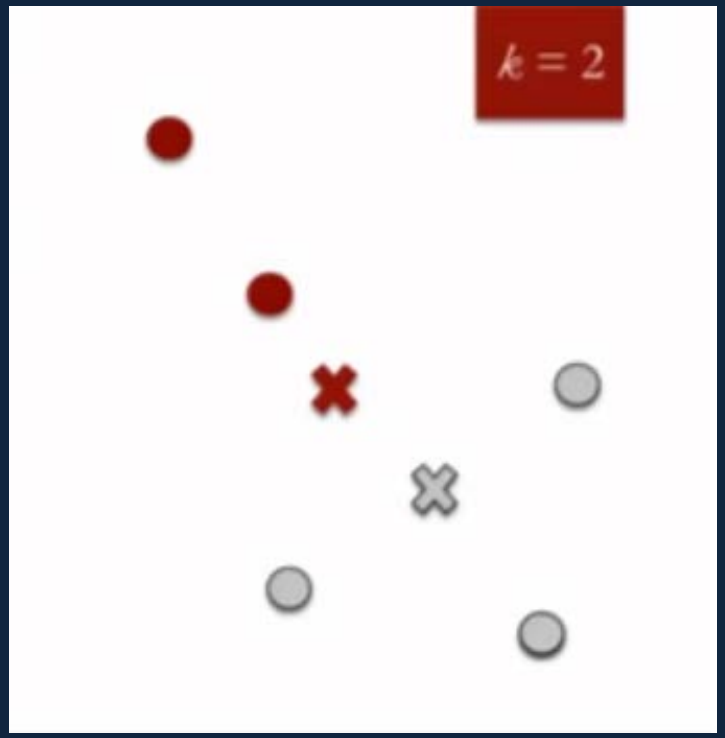
- Compute cluster centroids : The centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.



Co-funded by the Erasmus+ Programme of the European Union

K Means (3)

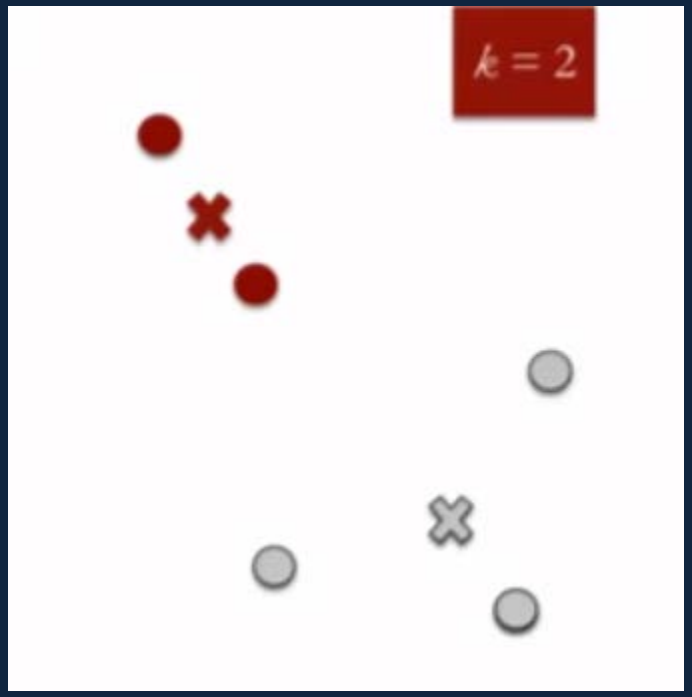
- **Re-assign each point to the closest cluster centroid**



Co-funded by the Erasmus+ Programme of the European Union

K Means (4)

- Re-compute cluster centroids
- Repeat last two steps until no improvements are possible



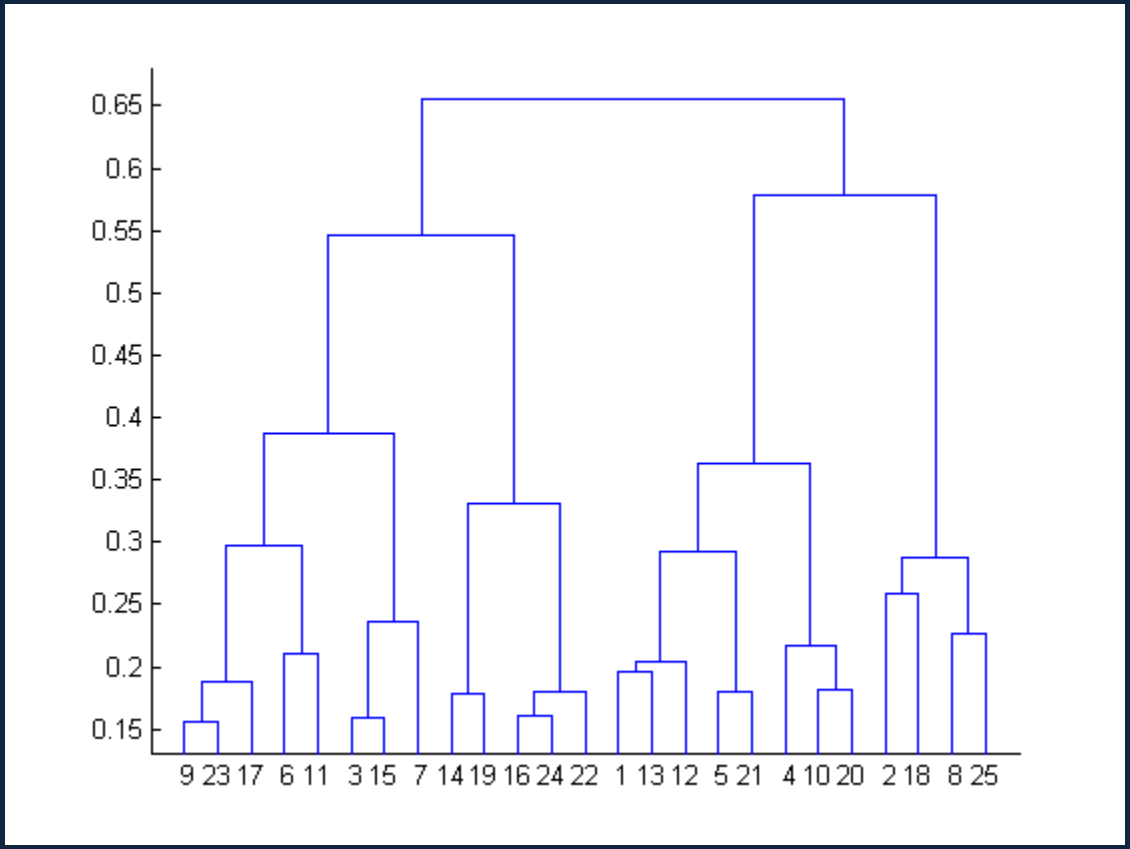
Co-funded by the Erasmus+ Programme of the European Union

Hierarchical Clustering

This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left.

Co-funded by the Erasmus+ Programme of the European Union

Hierarchical clustering – dendrogram



Co-funded by the Erasmus+ Programme of the European Union

Dendrogram

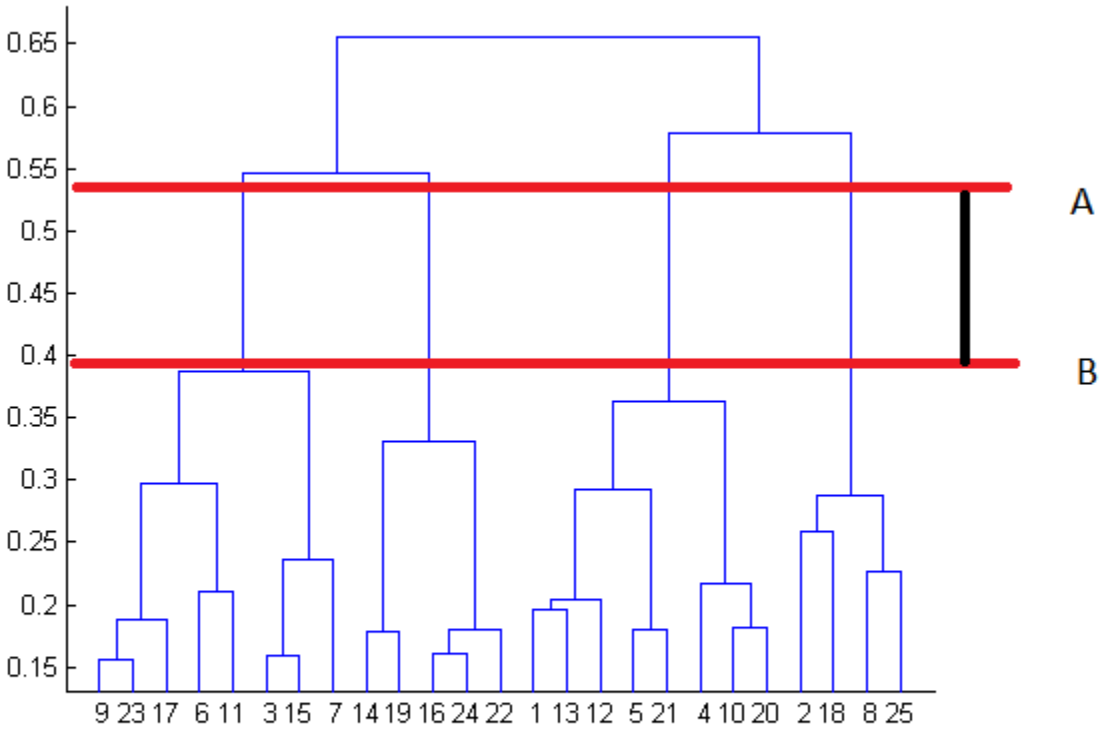
At the bottom, we start with 25 data points, each assigned to separate clusters. Two closest clusters are then merged till we have just one cluster at the top. The height in the dendrogram at which two clusters are merged represents the distance between two clusters in the data space.

The decision of the no. of clusters that can best depict different groups can be chosen by observing the dendrogram. The best choice of the no. of clusters is the no. of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.

In the above example, the best choice of no. of clusters will be 4 as the red horizontal line in the dendrogram below covers maximum vertical distance AB.

Co-funded by the Erasmus+ Programme of the European Union

Optimal number of clusters



Co-funded by the Erasmus+ Programme of the European Union

Differences between K Means and Hierarchical clustering

- Hierarchical clustering can't handle big data well but K Means clustering can.
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

Co-funded by the Erasmus+ Programme of the European Union