CLUSTER ANALYSIS

CLUSTER ANALYSIS

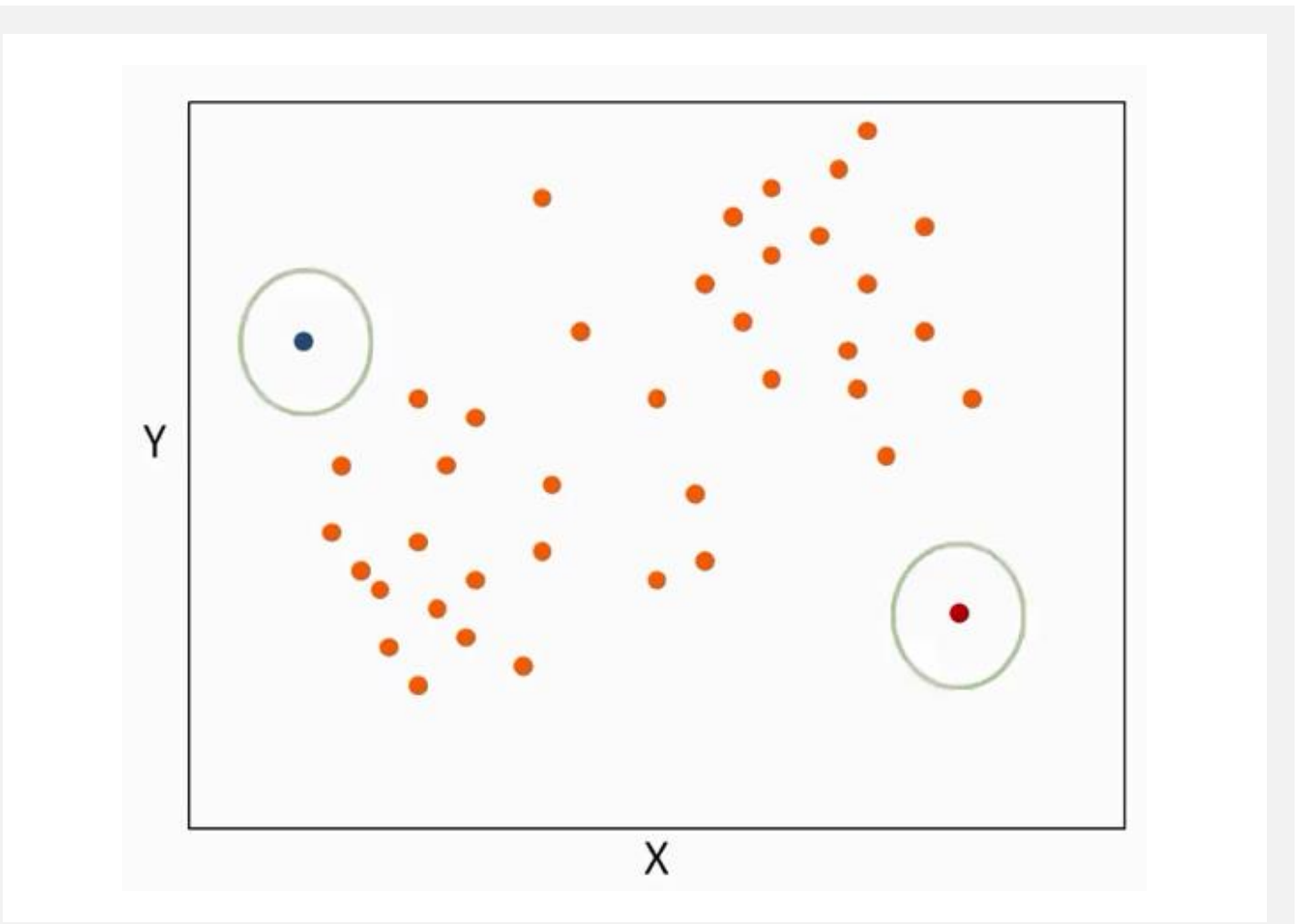
- The goal of cluster analysis is to group or cluster observations into subsets based on the similarity of responses on multiple variables. Observations that have similar response patterns are grouped together to form clusters.
- Each observation only belongs to one cluster.
- Unsupervised learning method, no specific response variable included in the analysis.
- Types of groceries people buy often used to group people together based on buying patterns. Then used for target marketing, or market segmentation.
- Clusters should have little variation within a cluster and more between clusters.

CLUSTER ANALYSIS

- K-Means clustering algorithm is one of the most commonly used.
- It is conducted by creating a space that has as many dimensions as the number of input variables.
- Input variables are designated with the notation P, so p-dimensional space is formed.
- The distance between observations in this space is used to determine how the data are partitioned. Cluster analysis measures the distance between points in the p-dimensional space, and groups those observations that are close to eachother.
- There are multiple ways to calculate that distance between observations. The most common K-means cluster analysis, is called Euclidean distance.

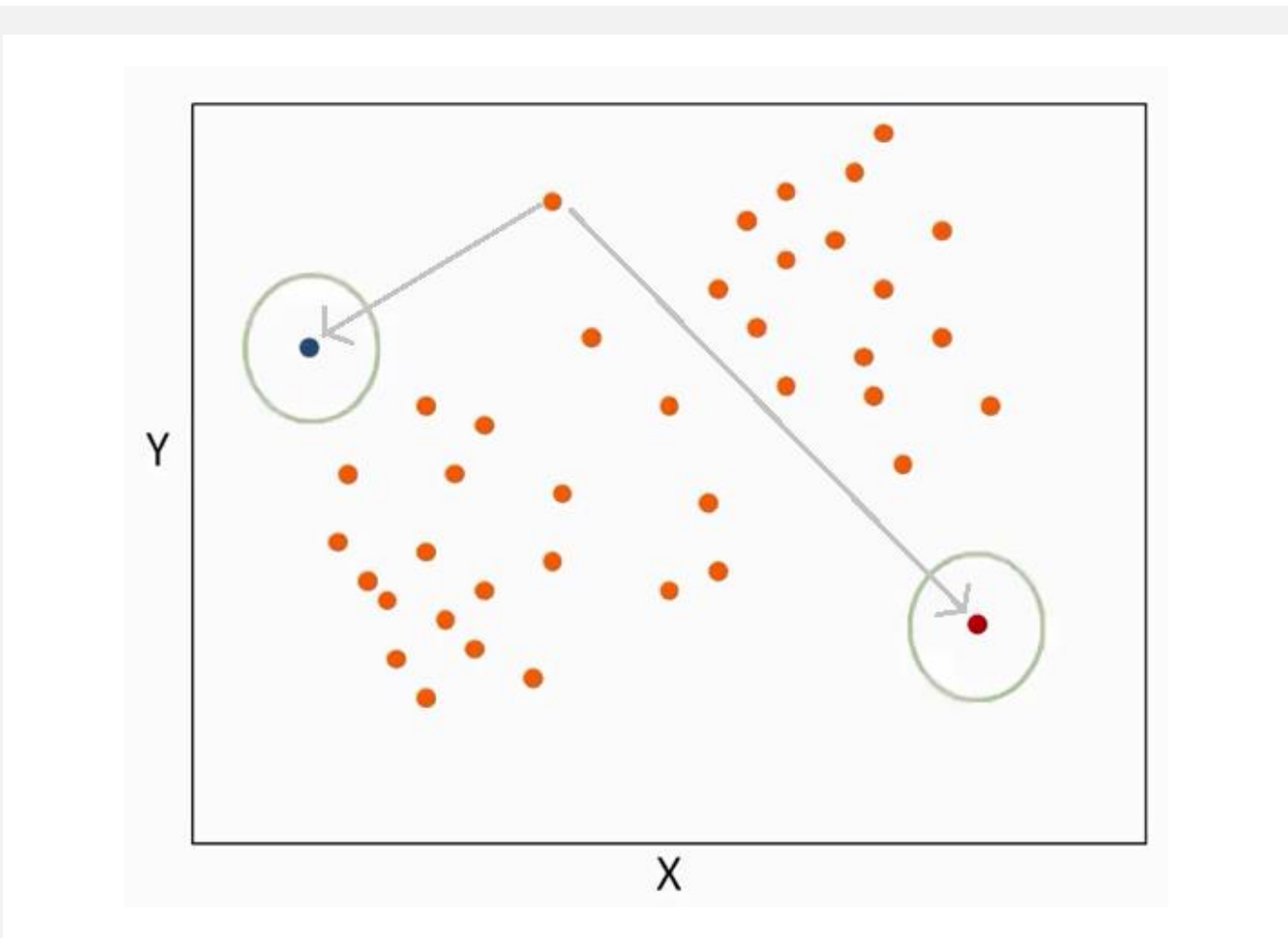
CLUSTER ANALYSIS

- P = 2 variables.
- Each dot is an observation.
- Step 1: randomly choose 2 points as the initial centroids and calculate distance between points and the centroids.



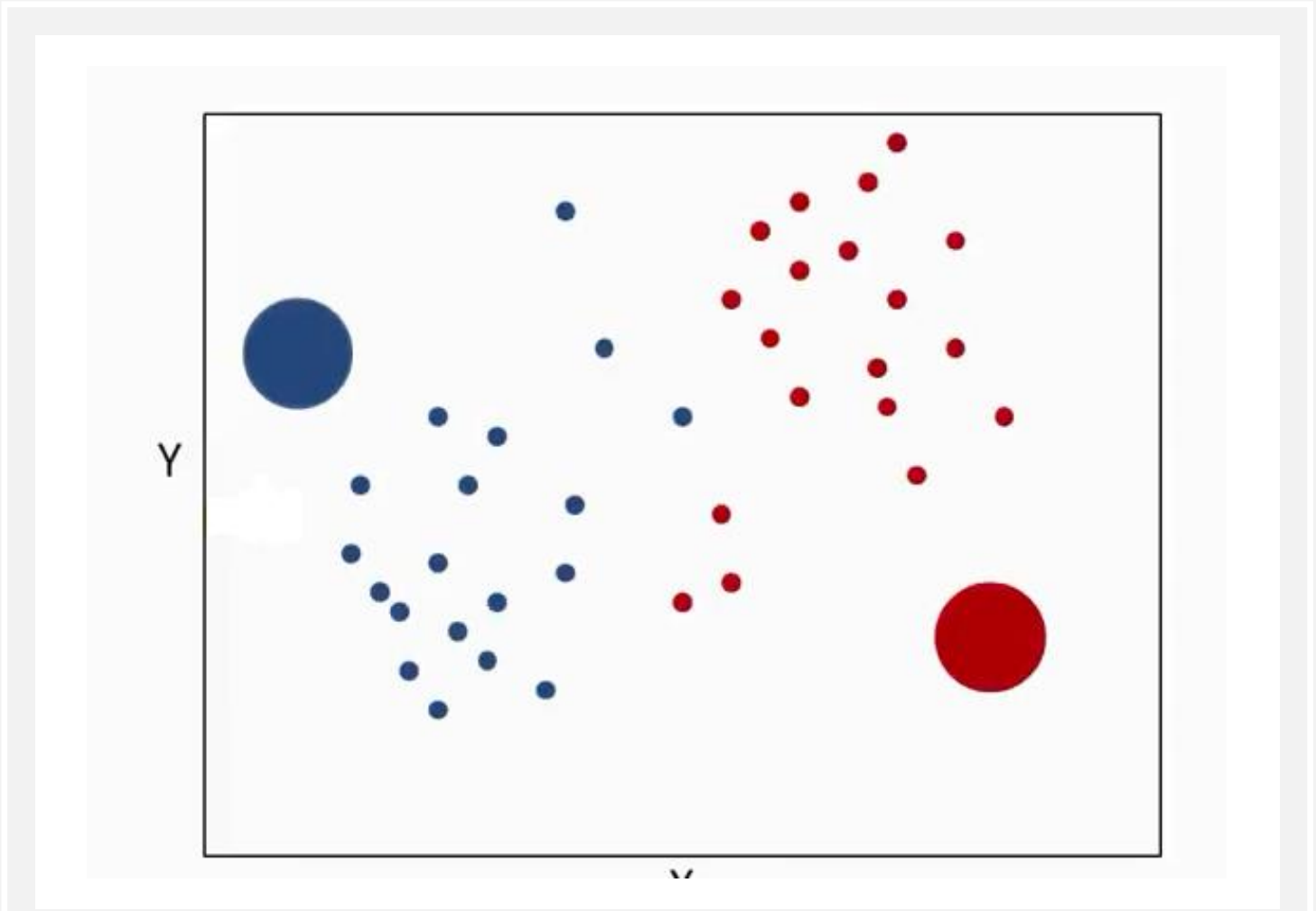
CLUSTER ANALYSIS

- Step 2: take one point and measure the distance between it and each centroid.



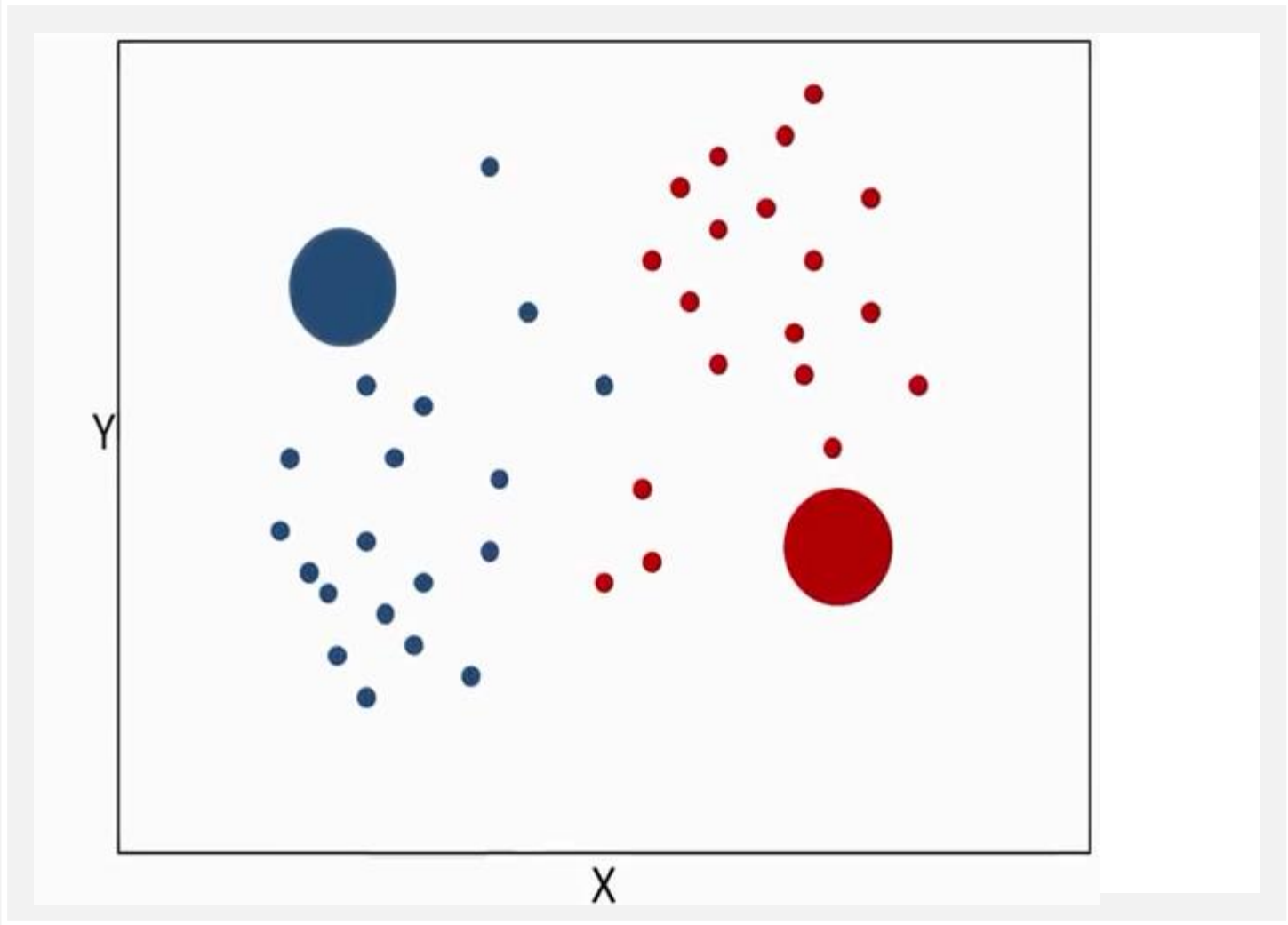
CLUSTER ANALYSIS

- Step 3: The point is then assigned to the nearest centroid, in this case the blue one.
- This is done for every point or observation.



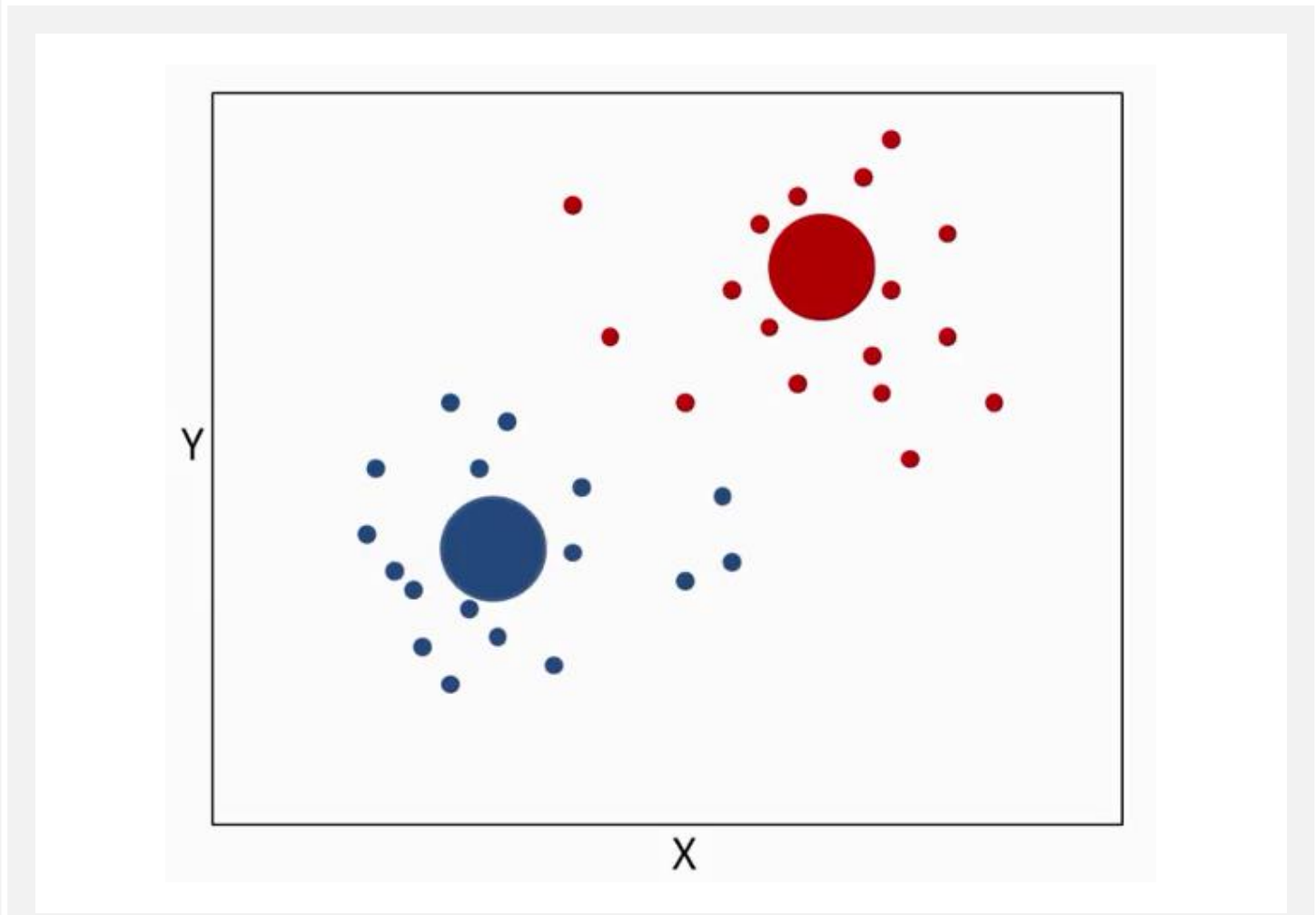
CLUSTER ANALYSIS

- After this first formation of the clusters, the original centroid for each cluster is recalculated based on the location of the points that were assigned to it.
- It is relocated to the place at which all the points and the centroid are smallest.



CLUSTER ANALYSIS

- Then the process starts all over again by calculating the distance between each observation and the new centroids.
- Then relocating the centroid again to the place where the sum of the new distances for the points assigned to the cluster is at the minimum.



CLUSTER ANALYSIS

- This process is repeated using multiple iterations until the location of the centroids doesn't change very much.
- During the process, observations that were originally assigned to one cluster may end up in a different cluster.

CLUSTER ANALYSIS

- The example we will look at characteristics of adolescence that could have an impact on school achievement.
- The ultimate goal may be to develop a few targeted interventions to improve academic achievement that are targeted to the needs of specific student population subgroups based on the characteristics of students in clusters.
- We will use 2 binary variables for alcohol and marijuana use, along with several quantitative variables such as alcohol problems, deviant behaviour, violence, depression, self esteem, parental presence, parental activities, family connectedness and school connectedness.
- Grade point average will be excluded so that we can use it to validate our clusters. We expect to see differences between clusters in grade point average.

CLUSTER ANALYSIS

- In the lab the add health data file is used.
- We standardise the data as we want each variable to have equal weighting. Similar to how we approached using weight and height in our classification model in Excel.
- We then split the dataset into a test set of 30% of the data and a training set of 70% of the data.
- We use a plot to show the average minimum distance of the observations from the cluster centroids for each cluster solution. In our example the average distance decreases as the number of clusters increase.

CLUSTER ANALYSIS

- Our goal is to minimise the distance between observations and their assigned clusters, we want to chose the fewest number of clusters that provides a low average distance.
- This can be a little subjective so further analysis is conducted to see whether they do not overlap, whether the patterns of means on the clustering variables are unique and meaningful, and whether there are significant differences between the clusters on the external validation variable GPA.