IMS Health & Quintiles are now IQVIA™

Day - 5
Cluster Analysis

January 12, 2018

What will I learn?

+ Concept behind clustering algorithms
+ How k – means algorithm works ?
+ How to build a robust clustering model and validate the accuracy using R ?

An introduction to cluster analysis

- Cluster analysis is the process of breaking down larger heterogeneous population / larger dataset into smaller homogenous groups.

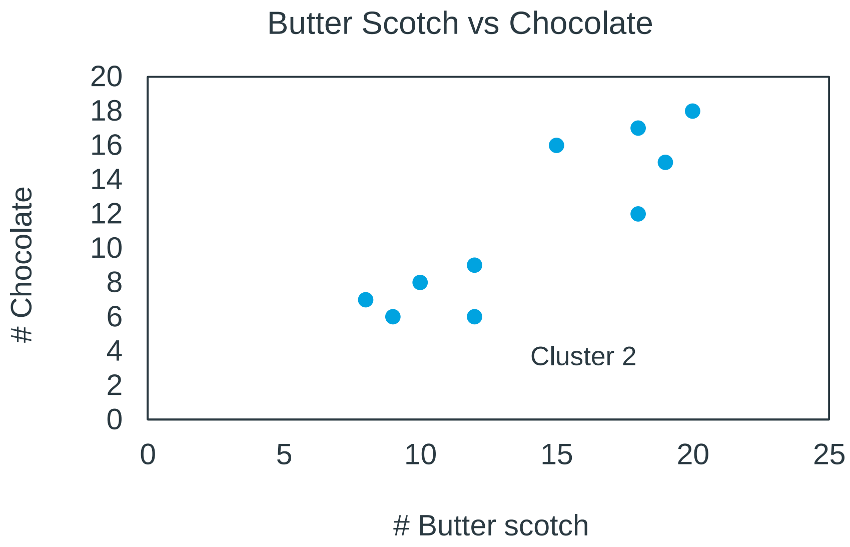
- Clustering allows us to break the larger population into smaller groups where each observation within a smaller group is more similar to each other than it is to an observation in other group. So, the idea is to group together the similar kind of observation into smaller groups from the population

Example :

Imagine that you own a chain of 10 cake shops, and sell 2 flavors of cake (Chocolate and Butterscotch). Now in the below table, we can see sales of both the flavors of cake across 10 stores

Shops	# Butter Scotch	# Chocolate
Shop 1	12	6
Shop 2	15	16
Shop 3	8	7
Shop 4	9	6
Shop 5	18	17
Shop 6	10	8
Shop 7	12	9
Shop 8	20	18
Shop 9	19	15
Shop 10	18	12

Very intuitive way of getting sense about this data is to plot this data in scatter plot

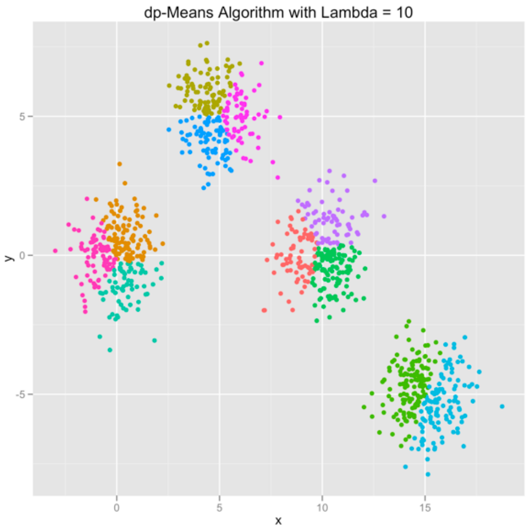


K-means clustering algorithms

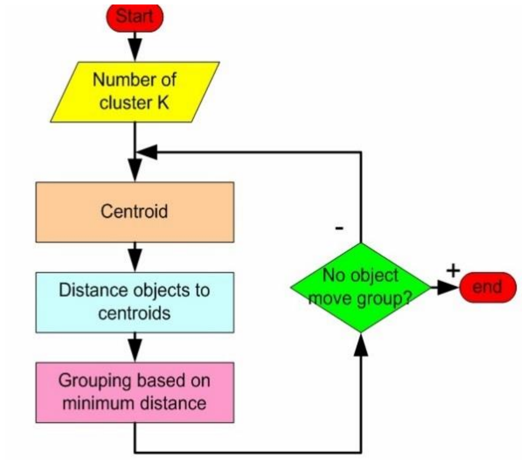
- K - means clustering algorithm is the most widely used clustering algorithms of all the clustering algorithms. The K in K means refers to the fact that algorithm is going to look for K different clusters, which means when applied on a dataset the algorithm is going to break the dataset into K different homogenous group. Incase if it is unable to find K different clusters it is going to break the dataset into K-1 different clusters.

Please Note : The value of K has to be defined to the algorithm before it starts. So, we have to decide how many clusters we want before starting the clustering algorithms

Let us take previous example to understand how this algorithm works. Once if we understand the logic, we can apply same logic across “n” dimensions. Let us assume that we want two different clusters

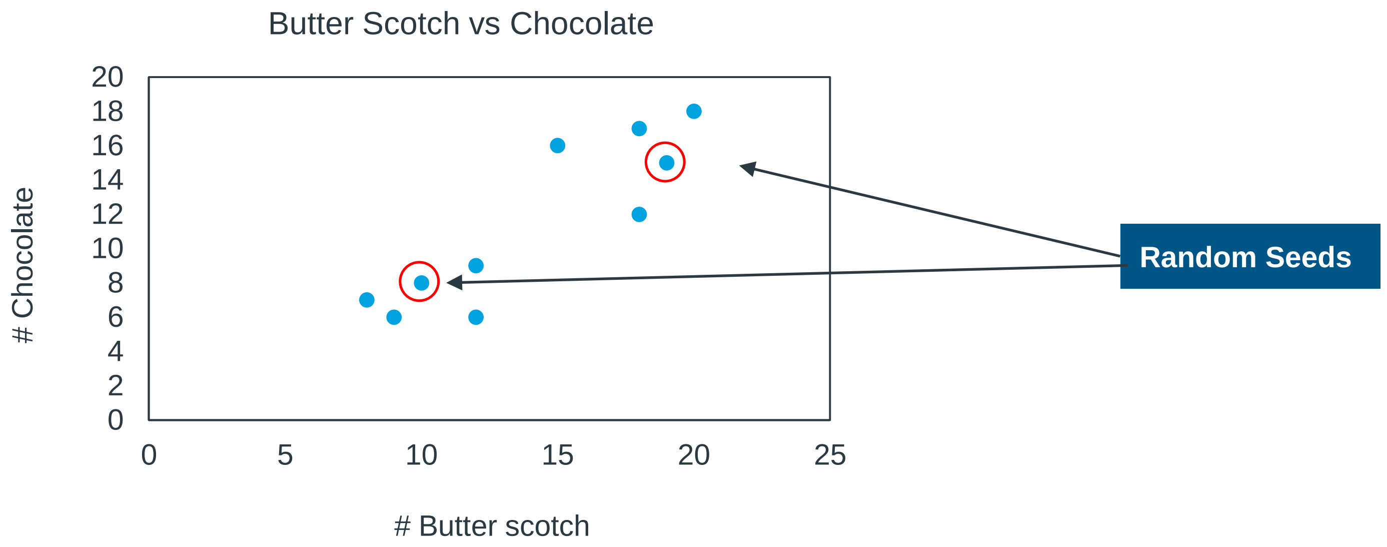


Steps Involved


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

Steps involved (1/2)

Step 1 : Algorithm has to identify two seeds. How the algorithm does this is, it identifies two random observations from the dataset and it assign them as seeds



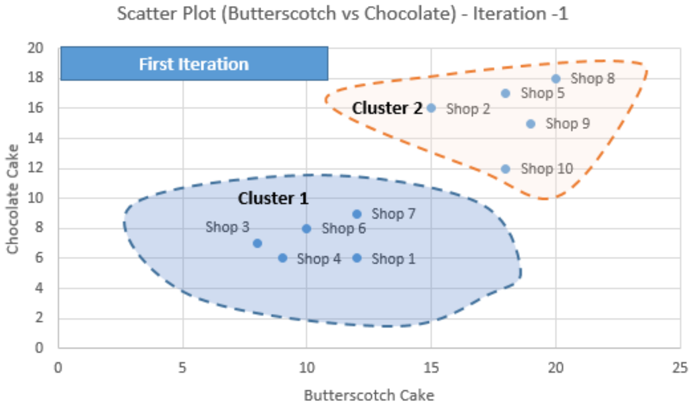
Step 2 : Assign all the other observations to one of these three seeds. How to assign observations to the seeds?

Based on Euclidian distance method:

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

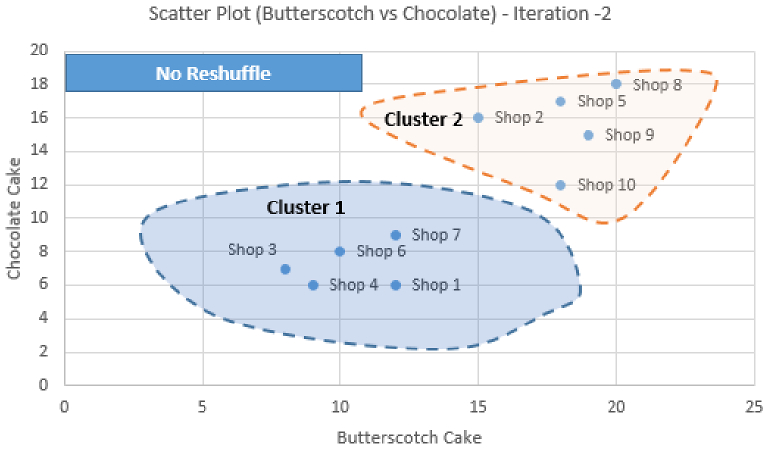
Steps involved (2/2)

Step 2 : Output



Step 3 : Calculate the centroid for each cluster and assign these centroids as seed. And restart the whole process of assigning each observations to the seeds (i.e. centroids). This process continues until the boundary of the cluster seizes to change. Once the boundary seizes to change it means that algorithm found the stable solution i.e. we found the optimal clusters

Final Cluster



Data preparation for clustering

1).Scaling:

Concepts of adjusting the values of the variables to take into account the fact that different variables are measured on very different scale

Methods of scaling :

- **Divide each variable by the range after subtracting the lowest value**
- **Divide each variable by the mean of the variable it takes on**
- **Normalization of variables / z scoring of variables**

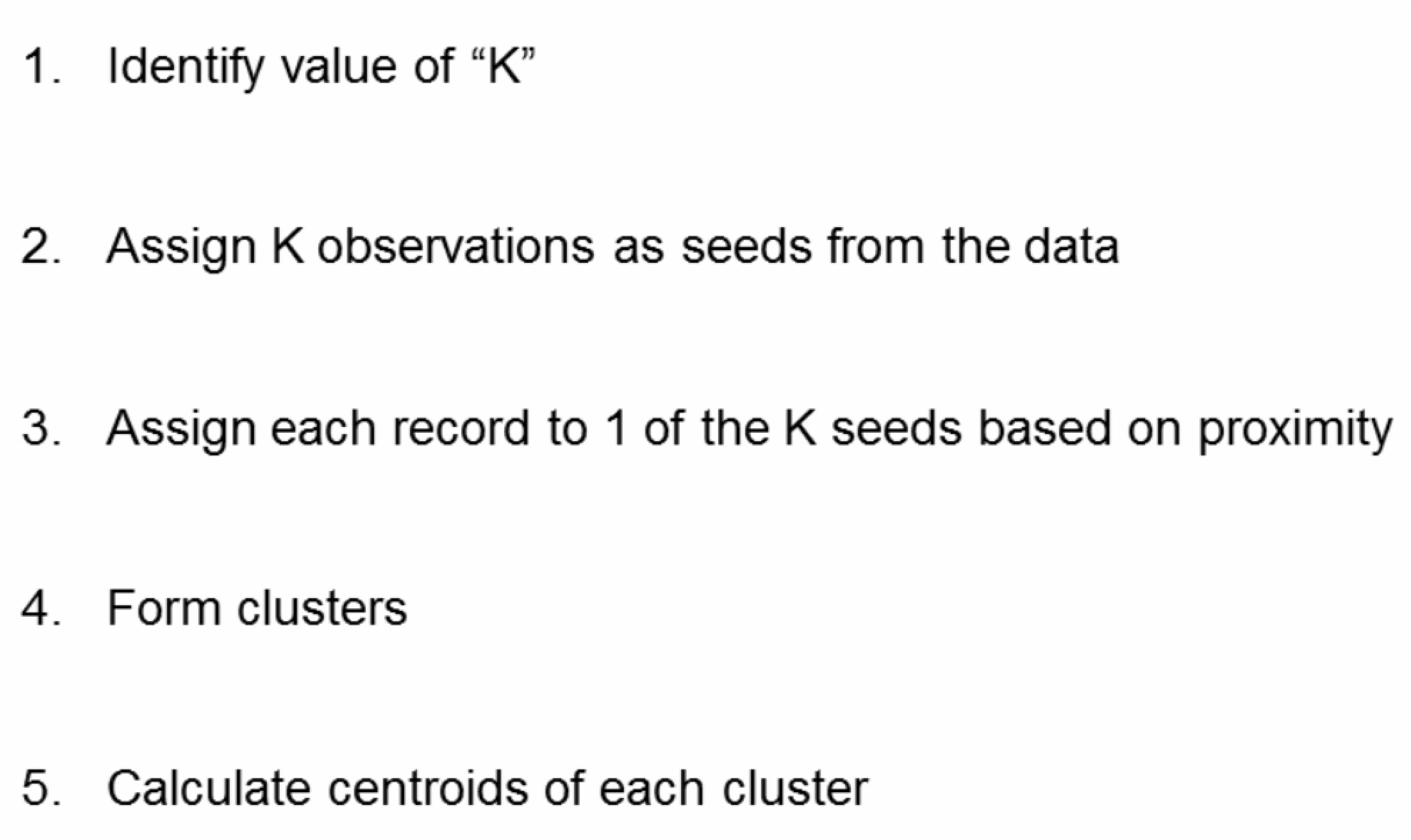
2).Weighting:

Concepts of adjusting the values of the variables as per their relative importance

Profiling for likely responders to a credit card offer :

Is # of cards owned as important as # no of times a person is married?

# of current credit cards	# of times married
1	1
3	3



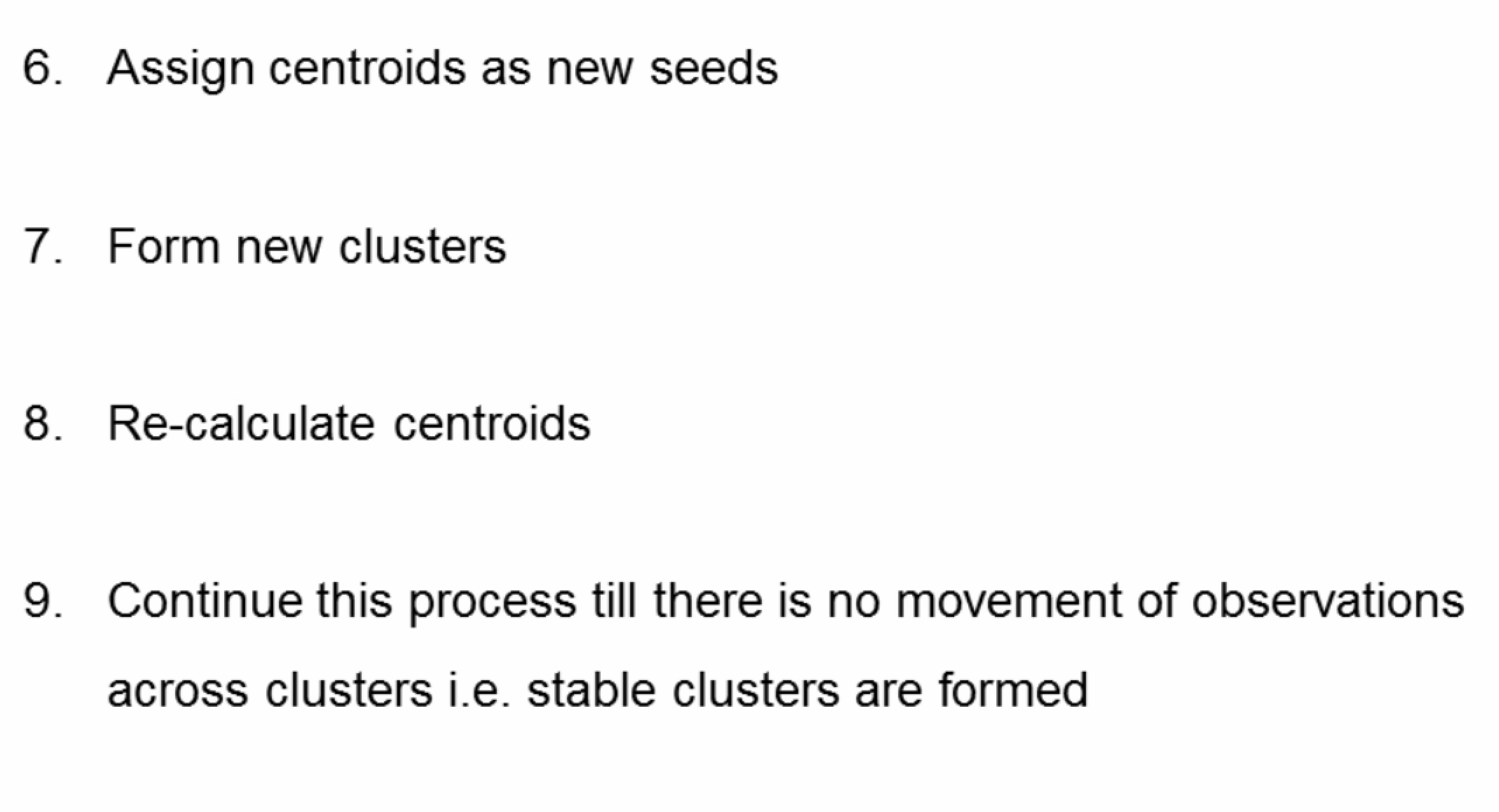
1. Identify value of “K”

2. Assign K observations as seeds from the data

3. Assign each record to 1 of the K seeds based on proximity

4. Form clusters

5. Calculate centroids of each cluster



6. Assign centroids as new seeds

7. Form new clusters

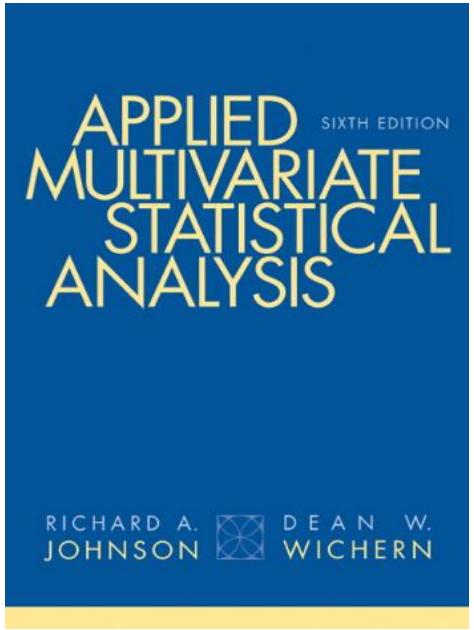
8. Re-calculate centroids

9. Continue this process till there is no movement of observations across clusters i.e. stable clusters are formed

Reference :

Applied Multivariate Statistical Analysis

- Johnson



IMS Health & Quintiles are now IQVIA™

Practical Session – Implementation in R