Data Visualization lab

CSE613

K-Means Clustering

- Clustering is a data exploratory technique used for discovering groups or pattern in a dataset.
- **clustering** is defined as **grouping objects in sets**, such that objects within a cluster are as similar as possible, whereas objects from different clusters are as dissimilar as possible.
- k-means clustering aims to partition **n - observations** into **k clusters** in which each observation belongs to the cluster with the **nearest mean**, serving as a prototype of the cluster. (source: Wikipedia)
- It requires the analyst to specify the number of optimal clusters to be generated from the data.
- The algorithm of Hartigan and Wong (1979) is used by default in **R software**. → uses **Euclidean distance measure** between data points to be determined the with-in and between-cluster similarities.

K-Means Clustering - Algorithm

- Step- 1: Specify the number of clusters (K) to be created (by the analyst)
- Step-2: Select randomly k-objects from the dataset as the initial cluster centers or means
- Step-3: Assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid.
- Step-4: each of the k-clusters update the **cluster centroid** by calculating the new mean values of all the data points in the cluster. The centoid of a K_{th} cluster is a vector of length *p* containing the means of all variables for the observations in the K_{th} cluster; *p* is the number of variables.
- Step -5: Iteratively minimize the total within sum of square. That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached. By default **R** uses 10 as the default value for the maximum number of iterations.

R function for K-Means Clustering

- K-means Clustering can be performed on data, in which all variables are continuous. → K-means algorithm uses variable means.
- Type → ?kmeans command in Rstudio → useful information
- The standard function for **K-means clustering** is **kmeans(),** which is defined in stat package. Rstudio has stat package by default. No need to load it.

```
kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong",
"Lloyd", "Forgy", "MacQueen"), trace=FALSE)
```

- x → numeric matrix, numeric data frame or a numeric vector
- **centers**: Possible values are the number of clusters (**k**) or a set of initial (distinct) cluster centers. If a number, a random set of (distinct) rows in x is chosen as the initial centers.
- **iter.max**: The maximum number of iterations allowed. Default value is 10.
- nstart: The number of random starting partitions when centers are number
- algorithm: 4-types of algorithm.
- trace: logical or integer number →used in Hartigan-wong algorithm. → if positive → producing progress of algorithm execution.

R function for K-Means Clustering

- kmeans() returns a list including:
  - cluster: A vector of integers (from 1:k) indicating the cluster to which each point is allocated.
  - centers: A matrix of cluster centers (cluster means)
  - totss: The total sum of squares(TSS). TSS measures the total variance in the data.
  - withinss: Vector of within-cluster sum of squares, one component per cluster.
  - tot.withinss: Total within-cluster sum of squares, i.e. sum(withinss)
  - betweenss: The between-cluster sum of squares, i.e. totss−tot.withinss
  - size: The number of observations in each cluster
  - iter : The number of (outer) iterations.
  - ifault: integer: indicator of a possible algorithm problem – for experts