MINI PROJECT K-MEANS

Project Summary

Humberto Carvalho – March 2017

EXERCISE 0 – INSTALL PACKAGES

As required, we install the packages **cluster**, **rattle**, and **NbClust**.

These packages are need for the following purposes:

- **cluster** – providing the methods for cluster analysis;
- **rattle** – providing the data to be analysed, it means in our case, the data from 13 chemical analyses of 178 Italian wines;
- **NbClust** – determining the best number of clusters for the analysis.

EXERCISE 1 – SCALE DATA

After installing the packages and loading the wine dataset, from **rattle** package, we scale the data, in order to standardize it, and remove the 1st column. For this goal, we use the **scale()** function, creating a new dataset named **df_wine**.

EXERCISE 2 – FIND THE NUMBER OF CLUSTERS TO BE USED – METHOD 1

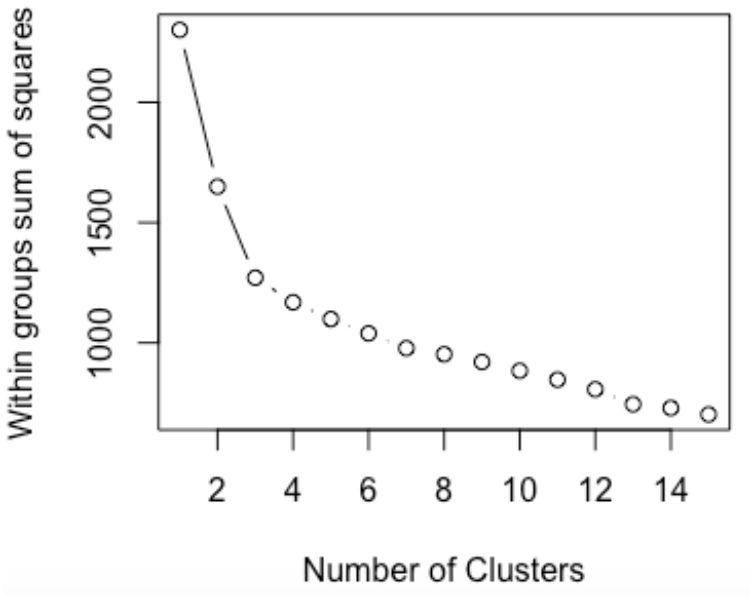
In opposite to other clustering techniques, the k-means clustering requires that the number of clusters to be extracted is specified in advance.

For determining the best number of clusters to be used, we apply a method based on a plot of the total within-groups sums of squares against the number of clusters in a k-means solution. For this purpose, we code the **wssplot** function.

For this function, we use, as data parameters, the dataset to be analysed (**df_wine**), the maximum number of clusters to be considered, and a random-number seed generated by the seed() function.

Mainly, the **wssplot** function applies the k-means clustering technique to the dataset **df_wine**.

Running the above-mentioned plot, we get the output as follows:



From this plot, we verify that there is a significant drop in the within-groups sum of squares (y-axis) when moving from one to three clusters (x-axis). After three clusters, the variation on the y-axis is not so significant as before. Hence, we could conclude that a three-cluster solution could be a good choice.

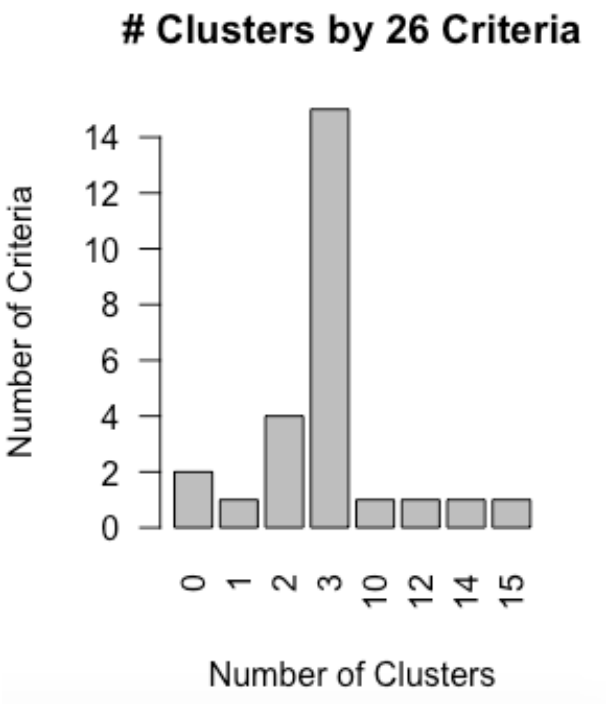
Therefore, this method suggests using three clusters.

EXERCISE 3 – FIND THE NUMBER OF CLUSTERS TO BE USED – METHOD 2

The 2nd method is based on the **NbClust** package, which runs many experiments and gives a distribution of the potential number of clusters.

The **NbClust** provides thirty indices for determining the number of clusters and suggests the best clustering scheme based on different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods. However, not all thirty criteria can be calculated for every dataset.

For our case, 14 of 24 criteria provided by this method suggest a 3-cluster solution, as shown on the below output.



Therefore, the 2nd method also suggests using three clusters, like the 1st one.

Exercise 4 – Run K-means Using The Suggested Number Of Clusters

As required, we run the k-means clustering based on the suggested number of clusters, three, and assign the output into a variable named **fit.km**.

Exercise 5 – Evaluate How Well This Clustering Does

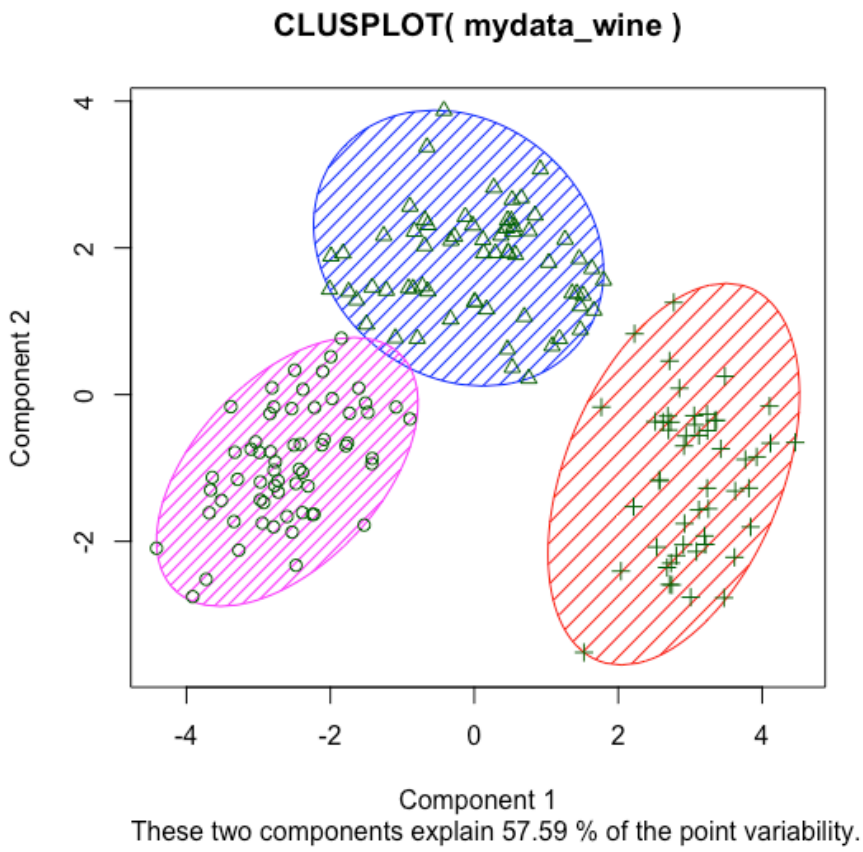
For this goal, using the **table()** function, we compare the clusters in **fit.km$clusters** with the actual wine types in **wine$type**, based on a cross-tabulation of these two variables.

```
> fit.km <- kmeans(df_wine, 3, nstart = 25)
> table(x=wine$Type, y=fit.km$cluster)
   y
x    1  2  3
  1 59  0  0
  2  3 65  3
  3  0  0 48
```

Based on the above output, we conclude that this is a good clustering taking into account the most samples (172 of 178, in total) fell on the diagonal of the crosstab matrix which means a value of almost 97% correct classifications.

EXERCISE 6 – VISUALIZE THE CLUSTERS USING FUNCTION CLUSPLOT()

```
> mydata_wine <- data.frame(df_wine, cluster = fit.km$cluster)
> clusplot(mydata_wine, fit.km$cluster, color = TRUE, shade = TRUE, labels = 0, lines = 0)
```



Reinforcing the previous conclusion, from the exercise 5, the visual aspect of the 3 clusters shows that we could consider this as a good clustering.