# Dimensionality Reduction and Clustering Techniques

By : Sarvesh MEENOWA

Supervised by: Dr Matthieu Cisel

Bachelor of Data Science by design

November 17, 2021

# Contents

# List of Figures



# List of Tables

# 1 Identifying correlation in the variables

Search for correlations among (cor.test). Use both parametric and non-parametric techniques (Pearson vs. Spearman for correlation between continuous variables). Design a couple of plots featuring correlated variables.

# 2 Dimensionality Reduction

- add subsections, abstract(possibly) , rename the captions

  - refer to pca with R book (multivariate analysis)

  - Practical Statistics for Data Scientists: 50 Essential Concepts

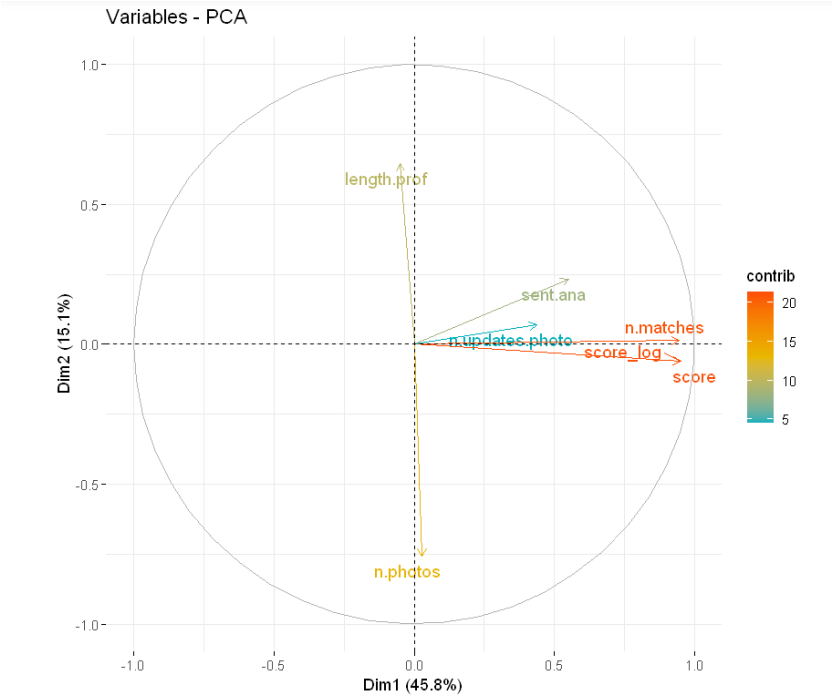


**Figure 1: Circle of correlations of variables**

Figure 2 represent the variable correlation plot which shows the relationship between all the variables. The first two dimensions,PC1 and PC2 retain 60.9% of the total inertia which is the total variance of dataset i.e. the trace

of the correlation matrix. The variables in Figure 2 are represented as vectors(arrows). The arrows start at the origin(0,0) and extend to coordinates given by the loading vector. The vectors can be interpreted in three different ways : The orientation (direction) of the vector, with respect to the principal component space, in particular, its angle with the principal component axes: the more parallel to a principal component axis is a vector, the more it contributes only to that PC, the length in the space; the longer the vector, the more variability of this variable is represented by the two displayed principal components; short vectors are thus better represented in other dimension, the angles between vectors of different variables show their correlation in this space: small angles represent high positive correlation, right angles represent lack of correlation, opposite angles represent high negative correlation.

More specifically, the longest arrows in the x-direction are number of matches(n.matches), score and the log of the score(score_log). This means that these three variables contribute the most to PC1. Along the y-direction, the length of profile(length.prof) and number of photos(n.photos) contribute the most to PC2. Number of matches and scores are very highly correlated since the angle between the two arrows is very acute, while length profile is almost uncorrelated to other variables(almost perpendicular) except number of photos, it is highly negatively correlated to the latter. Supplementary variables(number.updates.photo) have no influence on the principal components of the analysis. They are going to help to interpret the dimensions of variability.....TBC

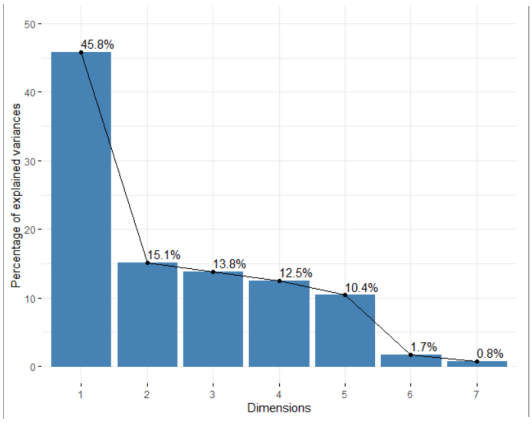


**Figure 2: Scree plot with percentage of explained variance**

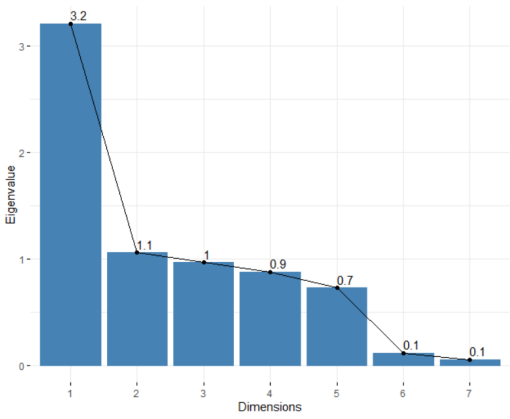


**Figure 3: Scree plot with eigenvalues**

- NB: Reshape figure

In order to know whether PCA works with our data, a scree plot, a diagnostic tool is used. The principal components are generated in the order of the amount of variation they capture: PC1 retains the most variation, PC2 - the second most, and so on. Each of them contributes some information about the data, and in a PCA there are as many principal components as there are features. By excluding the PCs, information is lost. In an ideal scree plot, the curve should be steep, then bends at an "elbow" which is the cutting off point and after that it flattens. While it might be hard to know visually at which principal component to put the cutting off point, there are 2 conventions that can be used :

- From Figure 2, the percentage of explained variance of the choosen PCs should be able to explained between 70-80% of the data.
- From Figure 3, using Kaiser rule, the PCs with eigenvalues greater or equal to one are selected.

In our case, the first three PCs are choosen explaining 74.7% of the data with eigenvalues 3.2, 1.1 and 1 respectively.

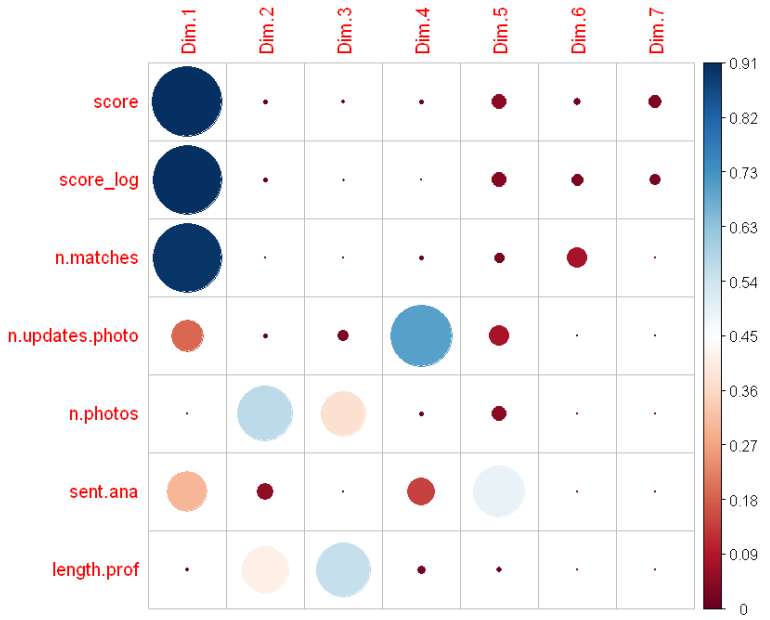


**Figure 4: Cos2 of variables on all the dimensions.**

The quality of representation of the variables on factor map is called cos2 (squared cosine).A high cos2 suggests a good representation of the variable on the principal component while a low cos2 indicates that the variable is not well represented by the PCs. From Figure 2, score, log of score and number of matches(n.matches) are well represented on PC1, number of photos(n.photos) and length profile(length.prof) are the variables that are represented the best in PC2 and PC3.

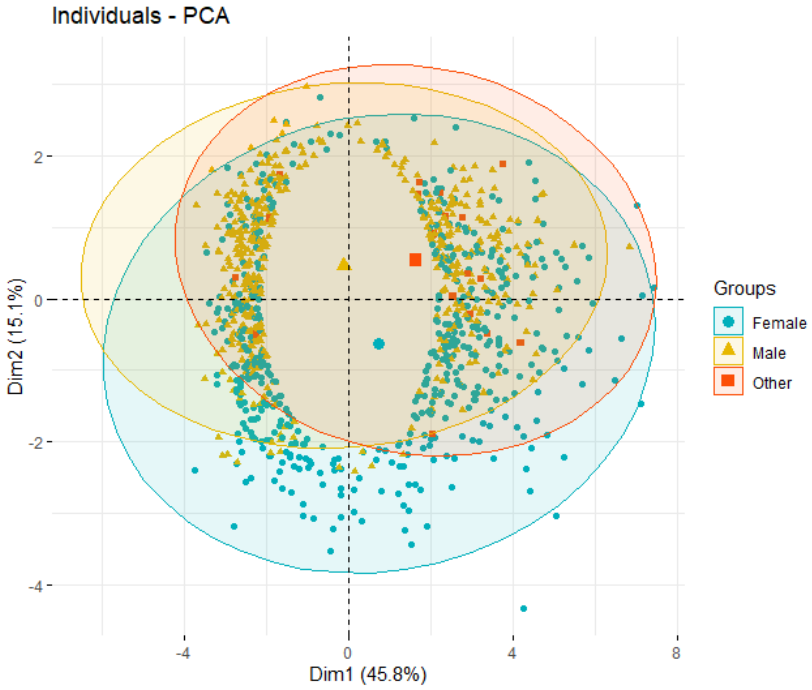


**Figure 5: Individual map with a sample of individuals with elipses**

From Figure 5,a sample of individuals with the higher contribution to the plane construction(first two PCs) were plotted. The individuals are coloured and grouped by their genders category.

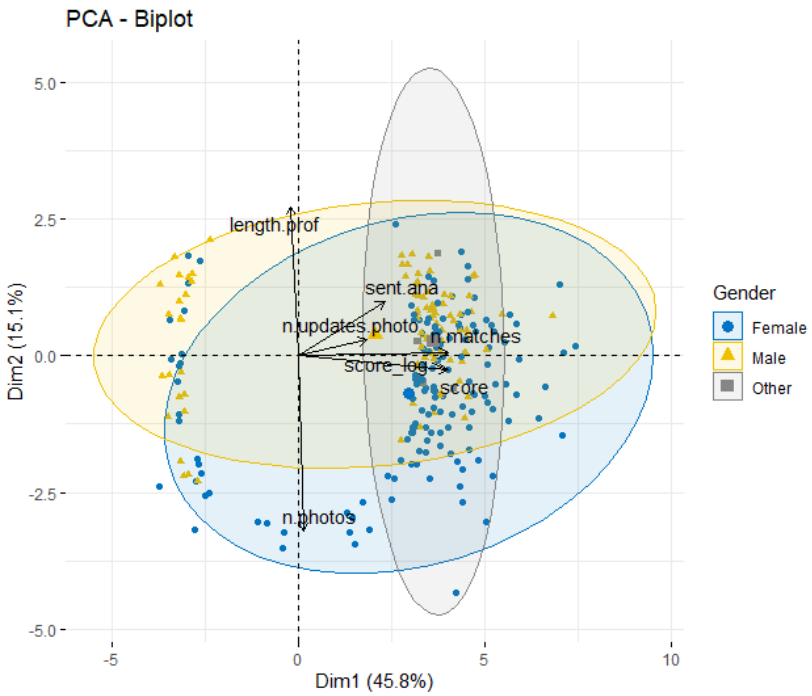


**Figure 6: Biplot with a limited number of individuals with ellipses**

A PCA biplot merges individuals and the loadings plot. Using a sample of individuals grouped by their gender, from Figure 6, we can see that individuals consisting of a great proportion of females and other genders and intermediate proportion of males share high values for the variables : number of matches(n.matches), score and sentiment score(sent.ana) and low values for the variables : length profile(length.prof) and number of photos(n.photos).

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|---|
| score | 0.53 | -0.06 | 0.05 | -0.07 | 0.24 | 0.30 | 0.75 |
| score_log | 0.53 | -0.06 | 0.04 | -0.03 | 0.23 | 0.48 | -0.65 |
| n.matches | 0.53 | 0.01 | 0.01 | -0.07 | 0.17 | -0.82 | -0.11 |
| n.updates.photo | 0.24 | 0.07 | -0.15 | 0.90 | -0.33 | 0.01 | 0.03 |
| n.photos | 0.02 | -0.74 | 0.62 | 0.07 | -0.24 | -0.05 | -0.01 |
| sent.ana | 0.31 | 0.23 | -0.01 | -0.41 | -0.83 | 0.07 | 0.00 |
| length.prof | -0.03 | 0.63 | 0.76 | 0.13 | 0.09 | 0.01 | 0.00 |

**Table 1: Table describing the loadings of each Principal Component**

Loadings are the correlations between the original variables and the unit scale components, i.e they are the linear combination weights (coefficients) whereby unit-scaled components or factors define or "load" a variable. From Table 1, the loadings for the first principal component can be calculated from the standardized data using the weights listed under PC1 : PC1 = 0.53*score + 0.53*score_log + 0.53 *n.matches + 0.23*n.updates.photo + 0.02 * n.photos + 0.31*sent.ana - 0.03*length.prof. This is done for all the others PCs.From the scree plots in Figure 2 and Figure 3, three principal components were selected, so a closer look is given to those three.PC1 has strong positive loadings for score, log of scores(score_log) and number of matches(n.matches) and , PC2 has high positive loadings for length profile(length.prof) and high negative loadings for number of photos(n.photos) and lastly, PC3 has high positive loadings for number of photos(n.photos) and length of profile(length.prof).

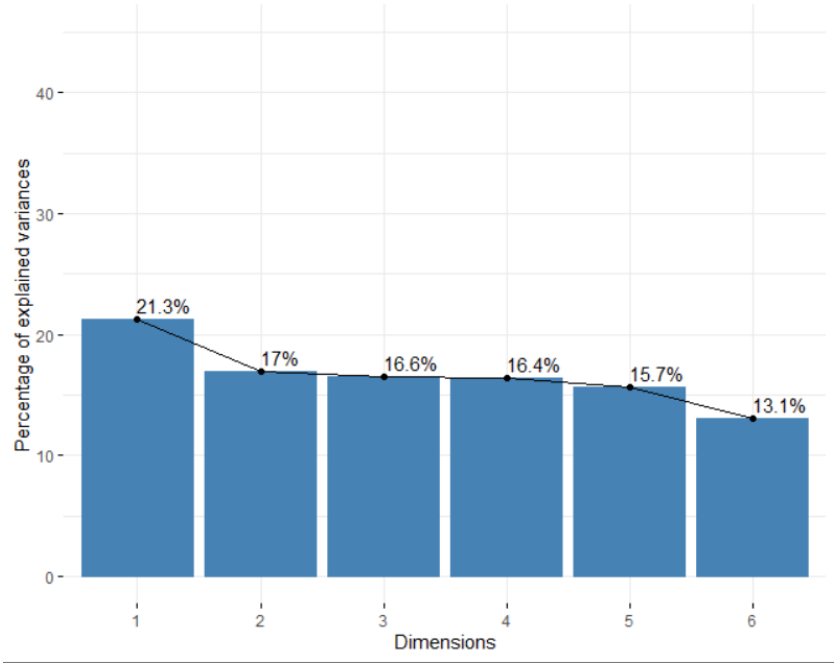


**Figure 7: Scree plot with percentage of explained variance of MCA.**

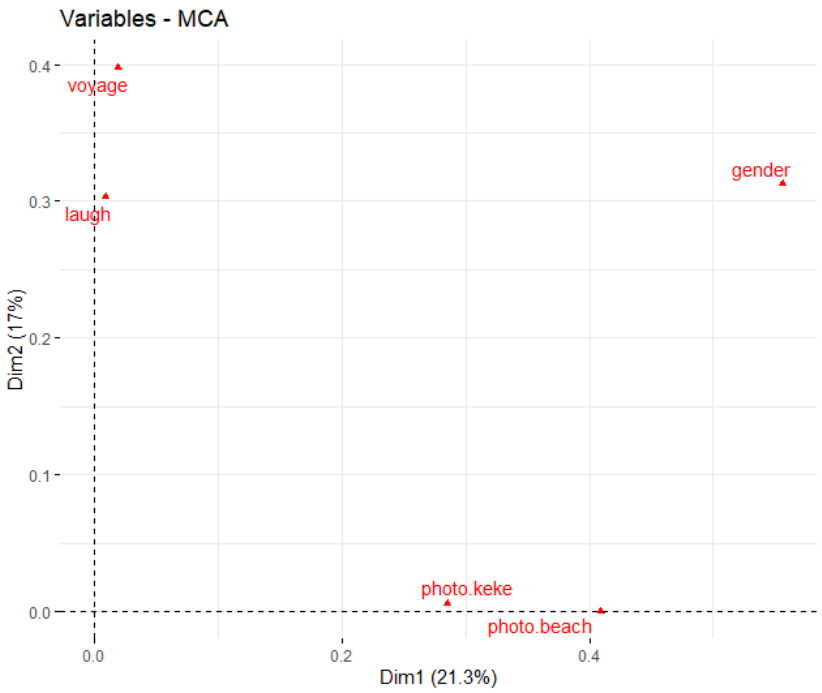


**Figure 8: MCA variable map**

Figure 8 helps in identifying variables that are most correlated with each dimension. The squared correlations between variables and the dimensions are used as coordinates. The variables photo.keke and photo.beach are the most correlated with dimension one while the variables voyage and laugh are most correlated with dimension two.

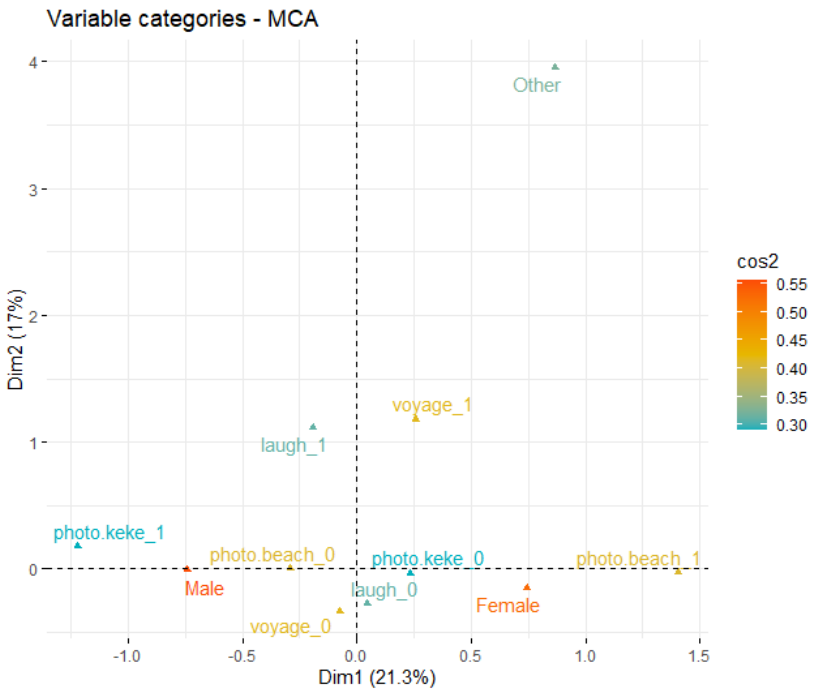


**Figure 9: MCA variable categories map**

Figure 9 illustrates the relationship between variable categories. Variable categories with a similar profile are grouped together. Negatively correlated variable categories are positioned on opposite sides of the plot origin(opposed quadrants). The distance between category points and the origin measures the quality of the variable category on the factor map. Category points that are away from the origin are well represented on the factor map. The two dimensions retain 38.3% of the total inertia(variation) and from Figure 9, the quality of representation is measured by the squared cosine.The better a category is represented by two dimensions, the closer the sum of cos2 is to one. In this case, the categories Female and Male are the most the well represented by the first two dimensions. The categories photo_beach_1 and female have an important contribution to the positive pole of the first dimension, while the categories photo_keke_1 and male have a major contribution to the negative pole of the first dimension. The categories voyage_1, Other

and laugh_1 represent an important contribution to the second dimension.

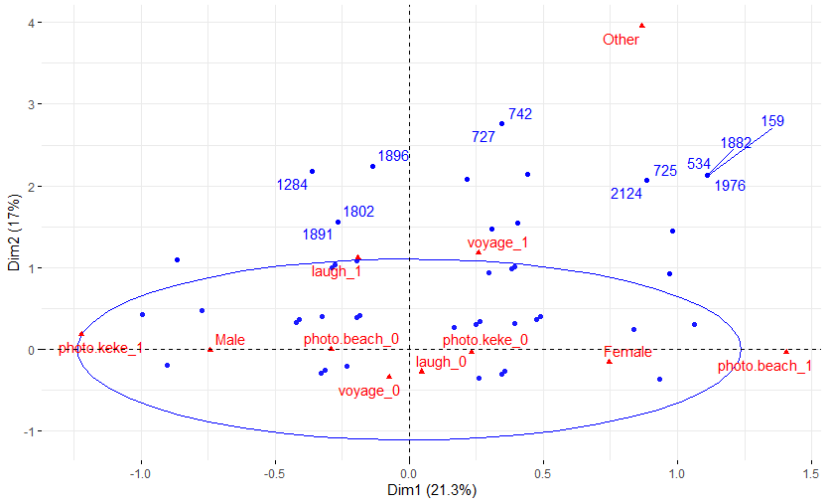


**Figure 10: MCA biplot**

In the MCA biplot from Figure 10, the variable categories are displayed in red and the individuals in blue. Females tend to post beach photos, and females who post photo keke also have the laugh keyword in their profile text. In the negative pole of the horizontal axis, similarly males have profiles with photo keke and males who post pictures on the beach also have the keyword voyage in their profile text.

# 3 k-means and Hierarchical Clustering

## 3.1 K-means clustering on principal components

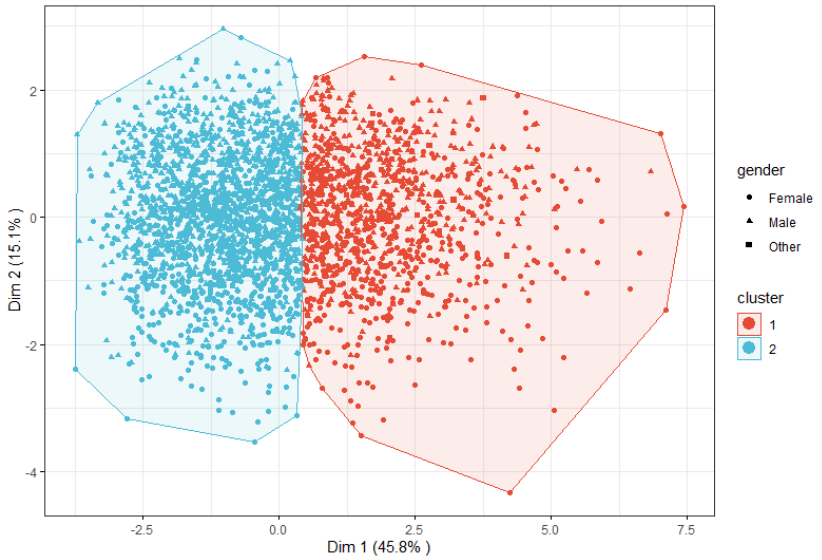


**Figure 11: K means on pca**

## 3.2 How k-means work?

- Step 1 : Specify the number of clusters (K) to be created.
- Step 2 : The algorithm starts by randomly selecting k objects from the data set to serve as the initial centers for the clusters. The selected objects are also known as cluster means or centroids.
- Step 3 : Assigns each observation to their closest centroid,where closest is defined based on the Euclidean distance between the object and the centroid(cluster mean). This step is called "cluster assignment step". Note that, to use correlation distance, the data are input as z-scores.
- Step 4 : After the assignment step, the algorithm computes the new mean value of each cluster.(The centroid of a Kth cluster is a vector of length p containing the means of all variables for the observations in the kth cluster; p is the number of variables.) The term cluster "centroid update" is used to design this step. Now that the centers have been recalculated, every observation is checked again to see if it

might be closer to a different cluster. All the objects are reassigned again using the updated cluster means.

- 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, the R software uses 10 as the default value for the maximum number of iterations.

## 3.3 Choice of the number of clusters

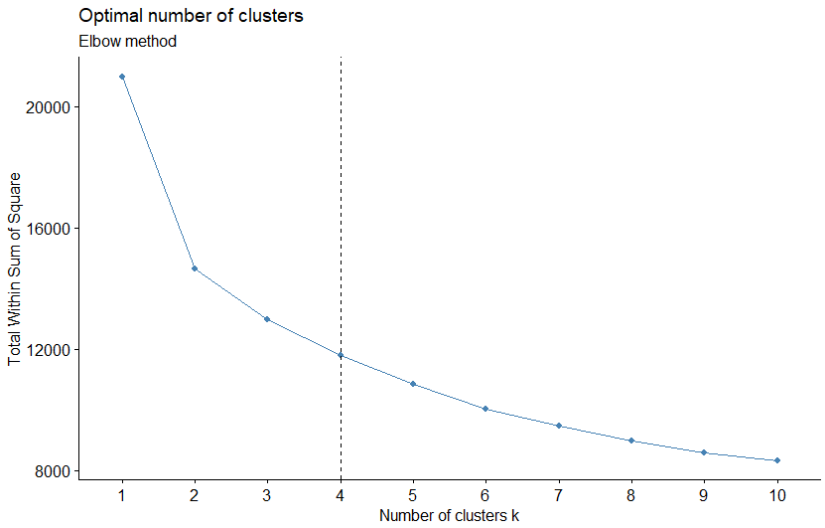


**Figure 12: Elbow kmeans**

K-means clustering, is to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total WSS measures the compactness of the clustering and we want it to be as small as possible. With the Elbow method as shown in Figure 12, it looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.

The optimal number of clusters can be defined as follow:

1. The k-means clustering algorithm is calculated for different values of k. For example, by varying k from 1 to 10 clusters.
2. For every k,the total within-cluster sum of square (WSS) is calculated .

3. The curve of WSS is plotted according to the number of clusters k.
4. The location of a bend in the plot is generally considered as an indicator of the appropriate number of clusters.

From Figure 12, the optimal number of clusters is 4, however the elbow method is sometimes ambiguous which is the case here.

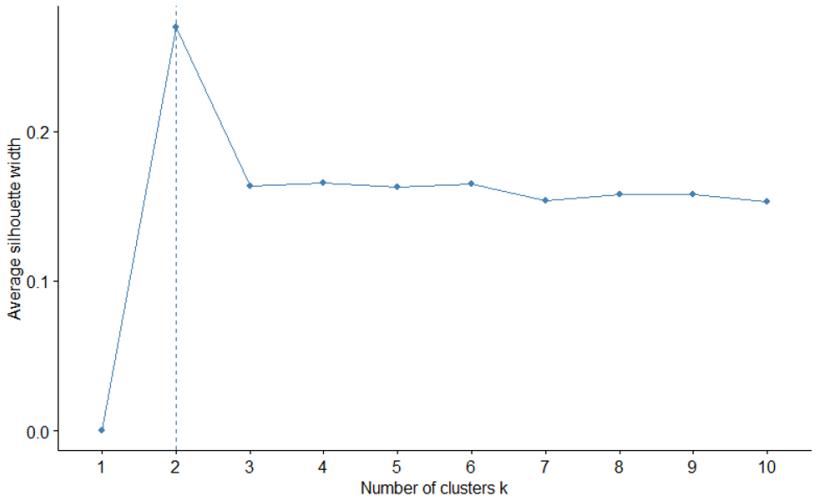


**Figure 13: Silhouette kmeans**

An alternative method is the average silhouette method, it measures the quality of clustering by determining how well each object lies within its cluster. Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k (Kaufman and Rousseeuw 1990).

The steps are similar to the elbow method and are as followed :

1. The k-means clustering algorithm is calculated for different values of k. For example, by varying k from 1 to 10 clusters.
2. For each k, the average silhouette of observations (avg.sil) is calculated.
3. The curve of avg.sil according to the number of clusters k is plotted
4. The location of the maximum is taken as the appropriate number of clusters.

From Figure 13, the optimal number of clusters is 2 where the maximum average silhouette width is around 0.3. The closer the average silhouette is to 1, the better the observation is grouped( (all observations are close to

cluster center). According to Kaufmann and Rousseuw (1990), a value below 0.25 means that the data are not structured. Between 0.25 and 0.5, the data might be structured, but it might also be deceptive. The values are indicative only and are not theoretically defined (it is not based on some statistical tests and associated p-values).

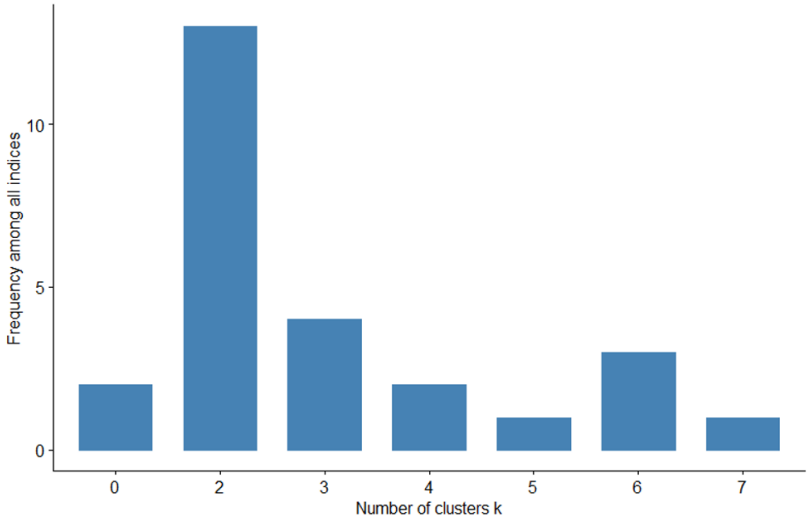


**Figure 14: NB clust**

Another approach is to use NbClust() function (Charrad et al. 2014). With one function call, it computes about 30 methods to determine the appropriate number of clusters and provides the best clustering scheme amongst the multiple results. From Figure 14, by the majority rule, the algorithm suggests that the optimal number of clusters is 2, having the highest frequency among the 30 indices.

## 3.4 HCPC on continuous variables

## 3.5 How HC works?

The difference with the partition by k-means is that for hierarchical clustering, the number of classes is not specified in advance. Hierarchical clustering will help to determine the optimal number of clusters.

The following shows how the ascending hierarchical clustering works step by step:

1. It starts by putting every point in its own cluster, so each cluster is a singleton
2. It then merges the 2 points that are closest to each other based on the distances from the distance matrix. The consequence is that there is one less cluster
3. It then recalculates the distances between the new and old clusters and save them in a new distance matrix which will be used in the next step
4. Finally, steps 1 and 2 are repeated until all clusters are merged into one single cluster including all points.

There are 5 main methods to measure the distance between clusters, referred as linkage methods:

1. Single linkage: computes the minimum distance between clusters before merging them.
2. Complete linkage: computes the maximum distance between clusters before merging them.
3. Average linkage: computes the average distance between clusters before merging them.
4. Centroid linkage: calculates centroids for both clusters, then computes the distance between the two before merging them.
5. Ward's (minimum variance) criterion: minimizes the total within-cluster variance and find the pair of clusters that leads to minimum increase in total within-cluster variance after merging.

- https://statsandr.com/blog/clustering-analysis-k-means-and-hierarchical-clustering-by-hand-and-in-r/

## 3.6 Pros and cons of HC compared to k-means

Advantages 1. Hierarchical clustering outputs a hierarchy, ie a structure that is more informaHve than the unstructured set of flat clusters returned by k-means. Therefore, it is easier to decide on the number of clusters by looking at the dendrogram.

Disadvantage: 1. Hierarchical clustering requires the computation and storage of an n×n distance matrix. For very large datasets, this can be expensive and slow.

With a large number of variables, K-means compute faster The result of K-means is unstructured, but that of hierarchal is more interpretable and

informative It is easier to determine the number of clusters by hierarchical clustering's dendrogram

# 4 Appendix

# 5 references

- interprete circle of correlations https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/Principal-Component-Analysis/principal-components-basics/Interpretation-and-visualization/index.html

- interprete biplots https://blogs.sas.com/content/iml/2019/11/06/what-are-biplots.html

- Interprete loadings : https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/principal-components/interpret-the-results/all-statistics-and-graphs/

- http://strata.uga.edu/8370/lecturenotes/principalComponents.html

-http://factominer.free.fr/factomethods/principal-components-analysis.html

- https://stats.stackexchange.com/questions/143905/loadings-vs-eigenvectors-in-pca-when-to-use-one-or-another

- https://halshs.archives-ouvertes.fr/halshs-01926339v3/document - https://www.researchgate

- https://www.abebooks.fr/9781542462709/Practical-Guide-Cluster-Analysis-Unsupervised-1542462703/plp