Dimensionality Reduction and Clustering Techniques

By : Sarvesh MEENOWA

Supervised by: Dr Matthieu Cisel

Bachelor of Data Science by design

November 19, 2021

Contents

List of Figures

List of Tables

1 Identifying correlations in the variables

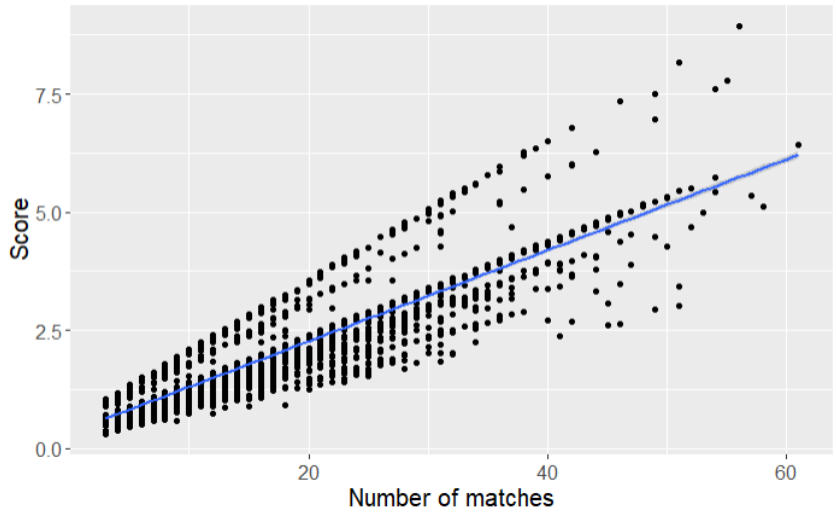

Figure 1: Scatter plot of Score against the number of matches.

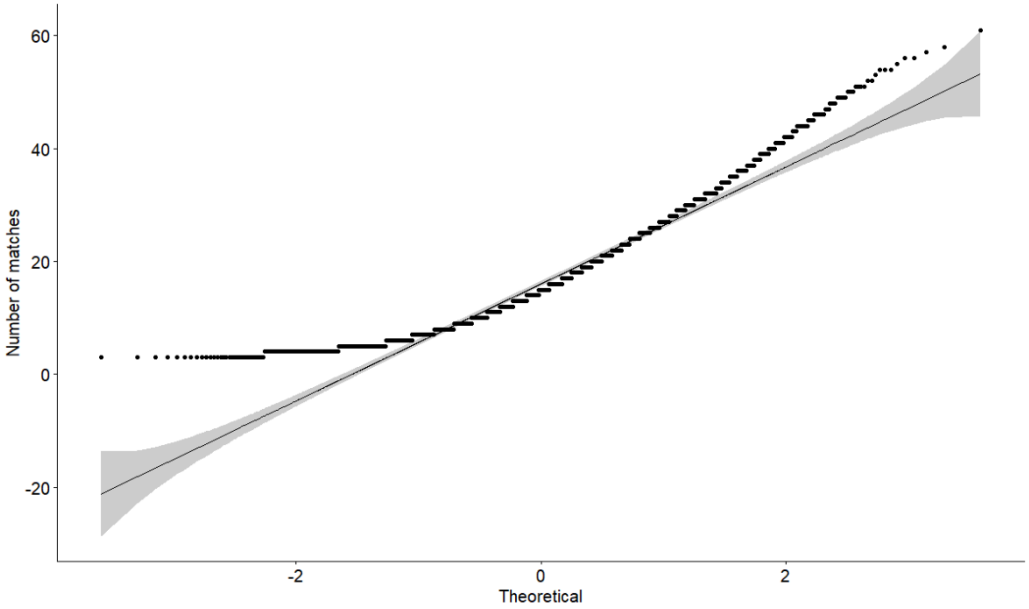

Figure 2: Quantile-Quantile(QQ) plot for the number of matches variable.

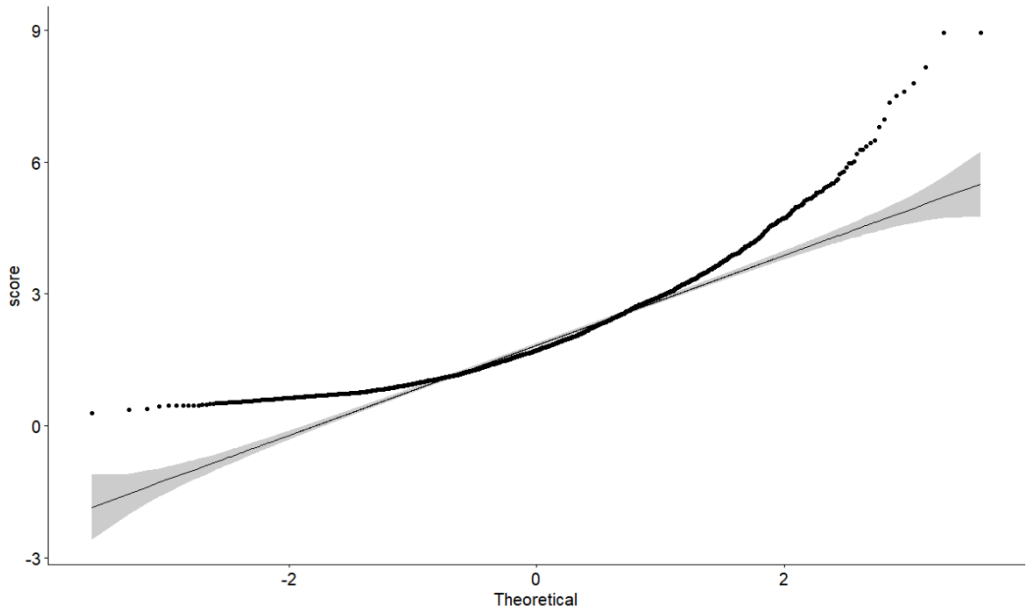

Figure 3: Quantile-Quantile(QQ) plot for scores variable.

From Figure 1, we can see that the relationship is monotonically increasing and non-normality can be assumed based on Figure 2 and Figure 3 where the points don't follow a straight line. We can perform the non-parametric correlation test using the Spearman method to test the relationship between score and the number of matches. From Figure 1, it indicates a very strong positive correlation(rho=0.92, p-value $< 2.2e^{-16}$). This means as the number of matches increases, the score also increases.

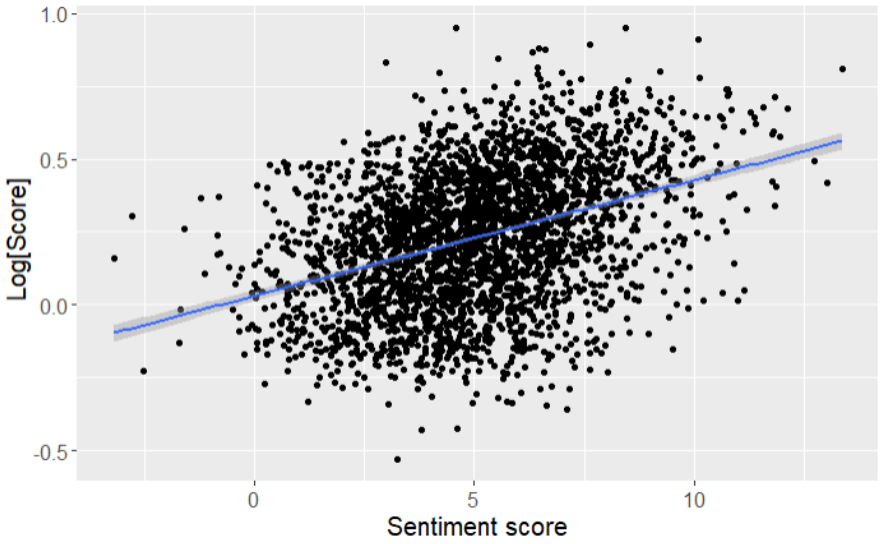

Figure 4: Scatter plot of log(Score) against the sentiment scores.

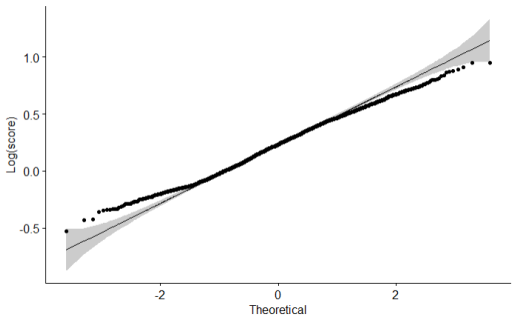

Figure 5: Quantile-Quantile(QQ) plot for the log of scores.

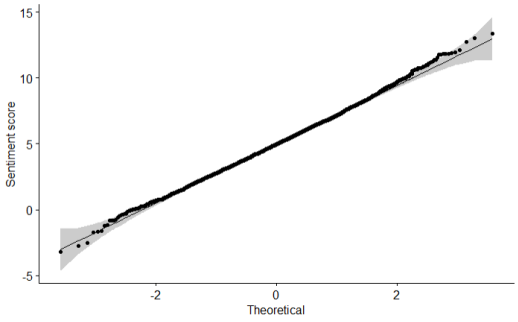

Figure 6: Quantile-Quantile(QQ) plot for sentiment score variable.

From Figure 5 and Figure 6, we can assume normality for the log of the variable scores and sentiment score.We therefore perform the parametric correlation test using Pearson method to test the relationship between the log of scores and the sentiment score. From Figure 2, we can see there is a positive correlation between the log of scores and the sentiment scores,albeit only a moderate relationship (r= 0.38 , p-value $< 2.2e^{-16}$) .

2 Dimensionality Reduction

2.1 Circle of correlations of variables

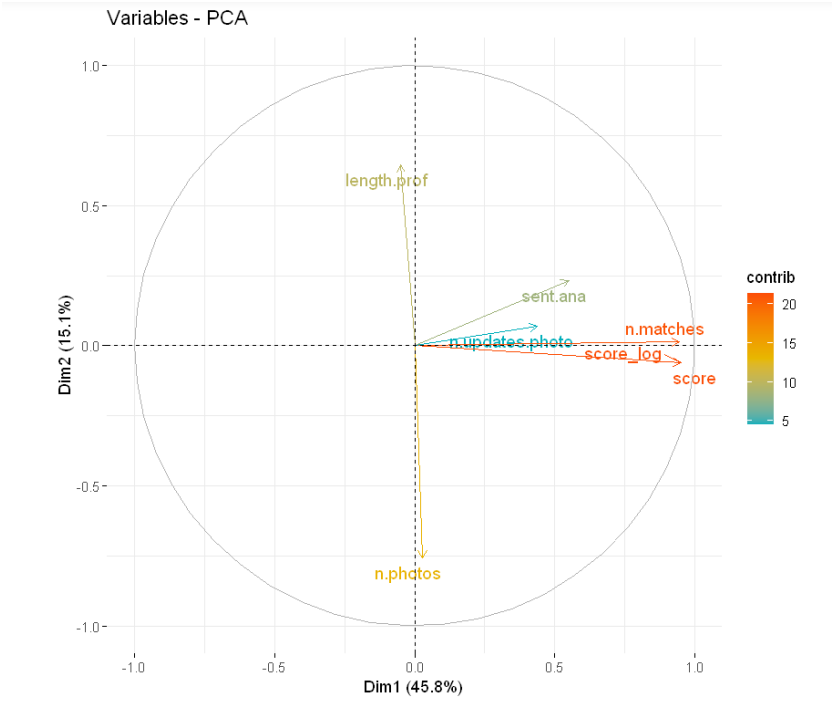

Figure 7: Circle of correlations of variables.

Figure 7 represents 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 the dataset. The variables in Figure 7 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 dimensions, 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 the high negative correlation.

More specifically, the longest arrows in the x-direction are the 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 the number of photos(n.photos) contribute the most to PC2. The number of matches and scores are very highly correlated since the angle between the two arrows is very acute, while the length profile is almost uncorrelated to other variables(almost perpendicular) except the number of photos, which is highly negatively correlated to the latter. Supplementary variables, in this case, the number of profile pictures updated(n.updates.photo) have no influence on the principal components of the analysis, they only help to interpret the dimensions of variability.

2.2 Selecting and naming principal components

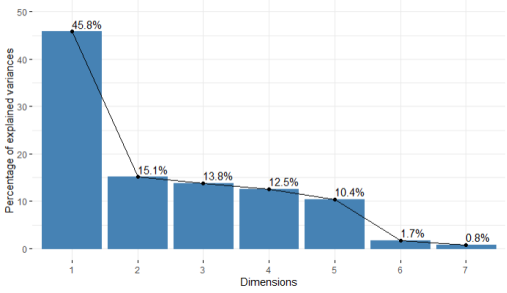

Figure 8: PCA scree plot with the percentage of explained variance.

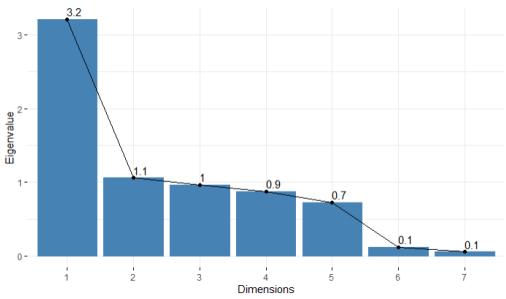

Figure 9: PCA scree plot with eigenvalues.

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 bend 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 8, the percentage of explained variance of the chosen PCs should be able to explain between 70-80% of the data.
- From Figure 9, using Kaiser rule, the PCs with eigenvalues greater or equal to one are selected.

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

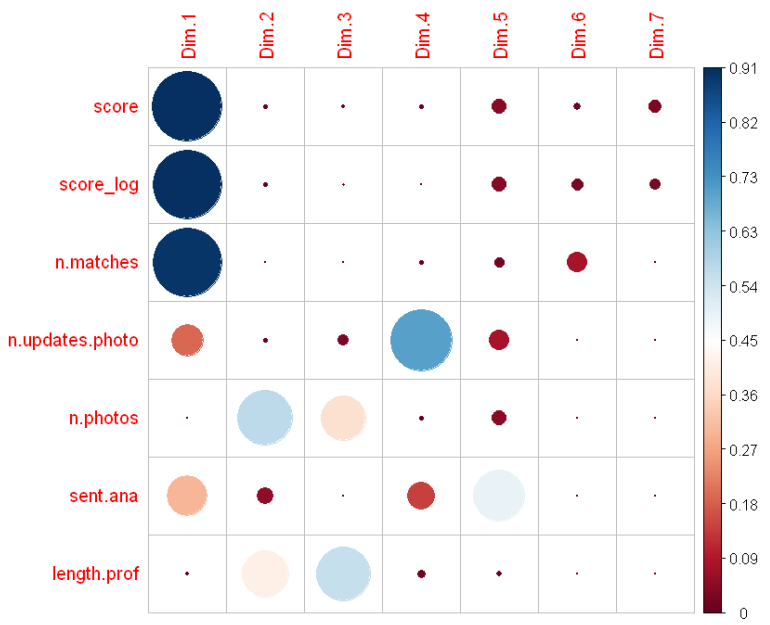

Figure 10: Cos2 of variables on all the dimensions.

The quality of representation of the variables on the 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 10, score, log of score and number of matches(n.matches) are well represented on PC1, the number of photos(n.photos) and length profile(length.prof) are the variables that are represented the best in PC2 and PC3.

We would therefore name PC1 : number of matches and score, PC2 and PC3 : number of photos and length of profile text.

2.3 PCA biplot and loadings table

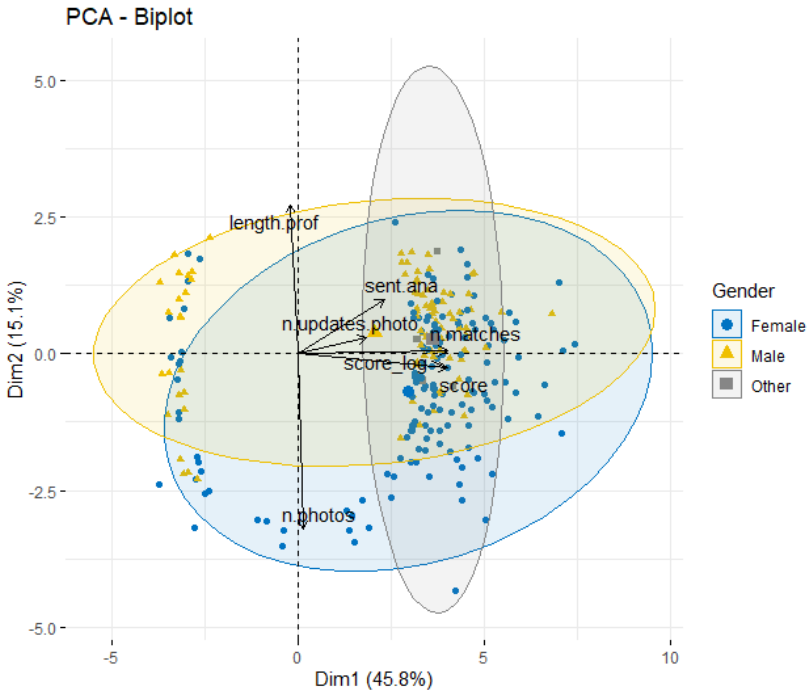

Figure 11: Biplot with a limited number of individuals with ellipses.

A PCA biplot merges individuals and the loadings plot. Using a sample of individuals grouped by their gender, from Figure 11, we can see that individuals consisting of a great proportion of females and other genders and an intermediate proportion of males share high values for the variables: the number of matches(n.matches), score and sentiment score(sent.ana) and low values for the variables: length profile(length.prof) and the number of photos(n.photos).

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 8 and Figure 9, three principal components were selected, so a closer look is given to those three.PC1 has

	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.

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).

2.4 MCA

Multiple correspondence analysis (MCA) is an extension of simple correspondence analysis(CA) to summarize and visualize a data table containing more than two categorical variables. It can also be thought of as a way of generalizing principal component analysis when the variables to be investigated are categorical rather than quantitative.

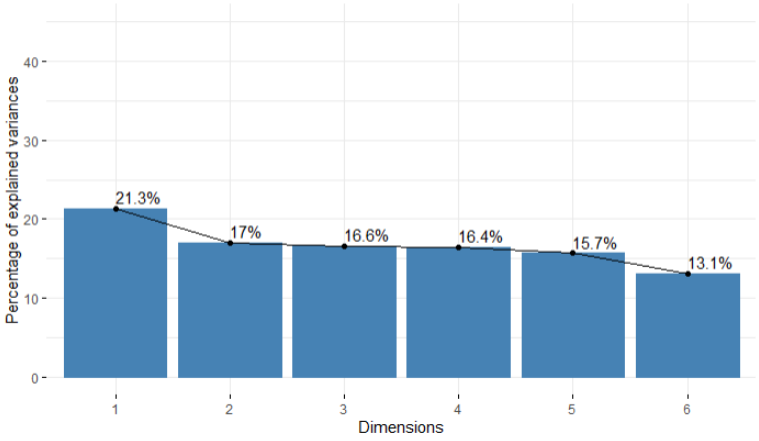

Figure 12: Scree plot with the percentage of explained variance of MCA.

To visualize the percentages of inertia explained by each MCA dimension, a scree plot is used as shown in Figure 12. We can see that from Figure 12 that the two dimensions 1 and 2 are sufficient to retain 38.3% of the total inertia (variation) contained in the data. Not all the points are equally well displayed in the two dimensions. In order to have an acceptable percentage of explained variance, at least 4 dimensions must be taken into account, which would then explain around 70% of the total inertia retained by the data.

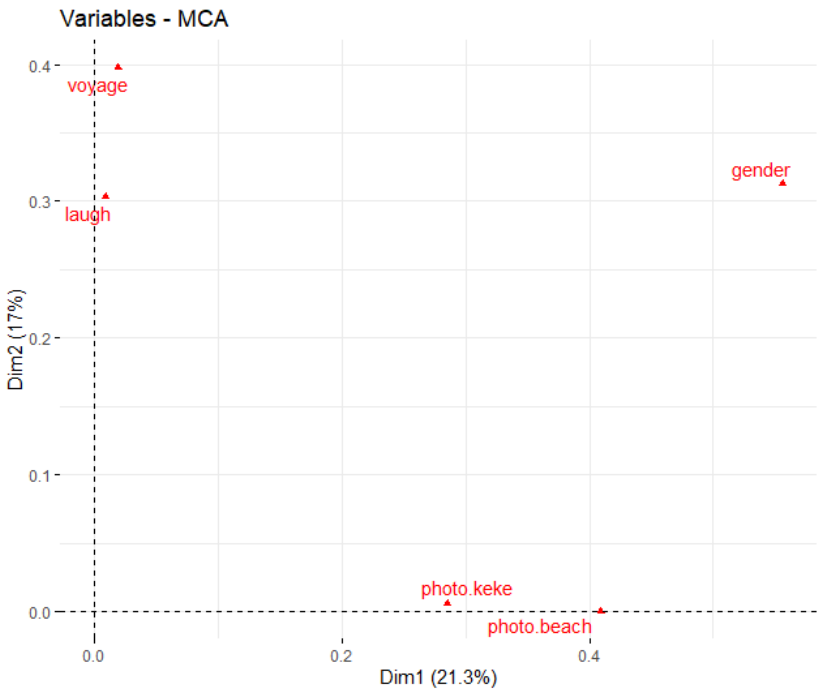

Figure 13: Correlation between variables and MCA principal dimensions.

Figure 13 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.

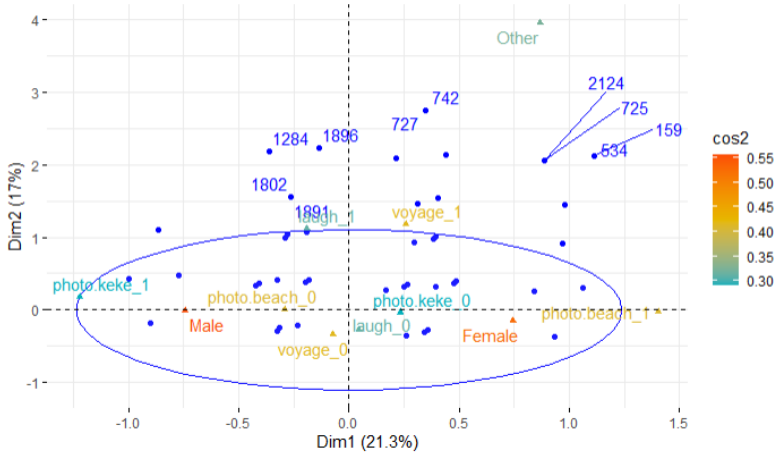

Figure 14: MCA biplot with variable categories and individuals.

Figure 14 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 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. 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

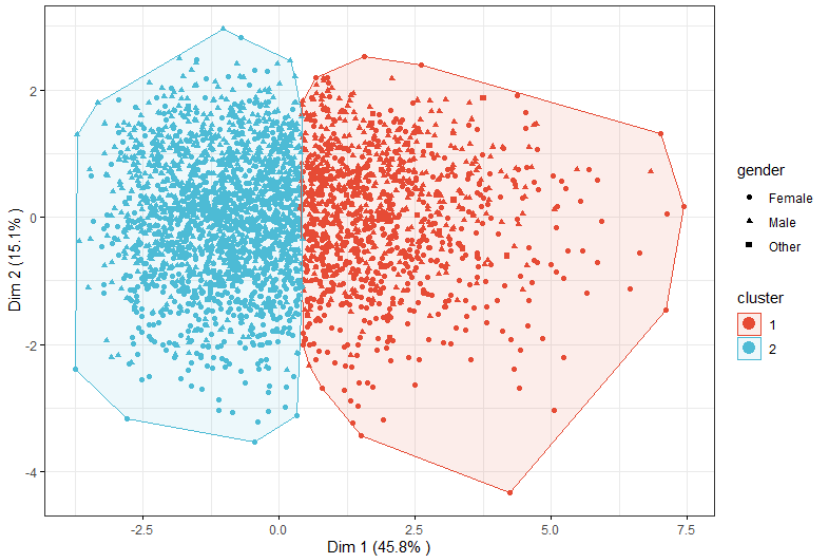

Figure 15: K-means clustering on principal components one and two.

	score	score_log	n.matches	n.updates.photo	n.photos	sent.ana	length.prof
1	1.01	1.01	1.03	0.47	0.00	0.57	-0.06
2	-0.58	-0.58	-0.59	-0.27	-0.00	-0.33	0.03

Table 2: Table with centers of each variable in each cluster.

We perform k-means clustering on the principal components one and two. From Figure 15, the number of optimal clusters deemed was 2, the reason is explained in Section 3.3. From Table 2, we can associate the centers for each variable in the clusters shown in Figure 15. Therefore, we can derive that cluster 1 seems to indicate that it has high values of scores, the number of matches with moderate values of number of profile pictures updated and sentiment scores and low values of number of photos and number of words in the profile text. While Cluster 2 indicates that it has individuals with moderate scores, number of matches, number of profile picture updates, sentiment scores and low number of photos and length profile.

3.2 How k-means work?

1. Specify the number of clusters (K) to be created.
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.
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 the "cluster assignment step". Note that, to use correlation distance, the data are input as z-scores.
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.
5. Iteratively minimize the total within the sum of squares. 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

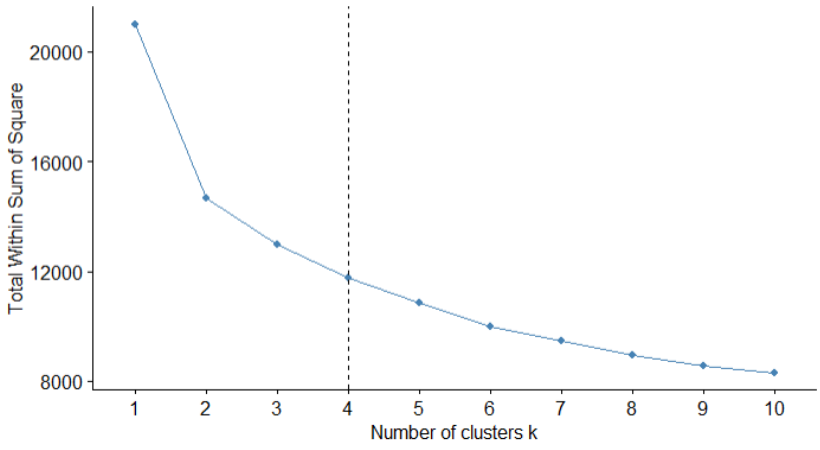

Figure 16: Elbow method to find the optimal number of clusters of k-means algorithm.

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 16, 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 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 squares (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 16, the optimal number of clusters is 4, however, the elbow method is sometimes ambiguous which is the case here.

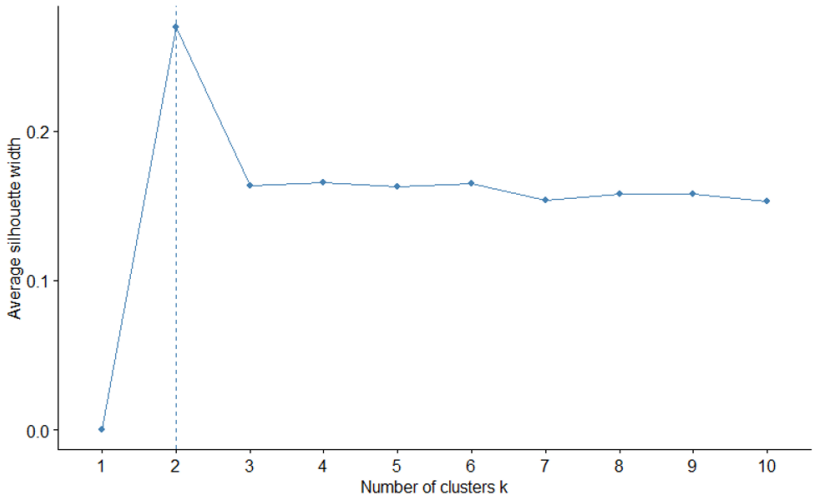

Figure 17: Silhouette method to find the optimal number of clusters for k-means algorithm.

An alternative method is the average silhouette method, it measures the quality of clustering by determining how well each object lies within its cluster. The average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the

one that maximizes 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 17, 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).

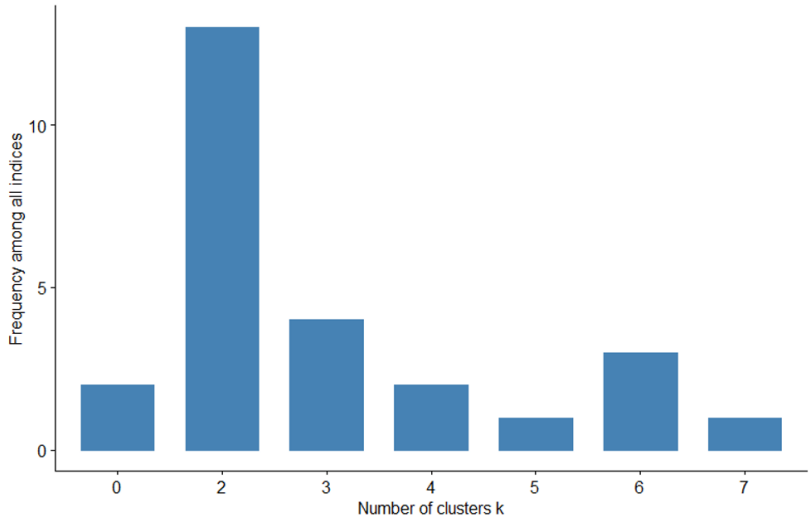

Figure 18: NbClust() function - 30 indices for choosingthe best number of clusters

Another approach is to use the 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 18, 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

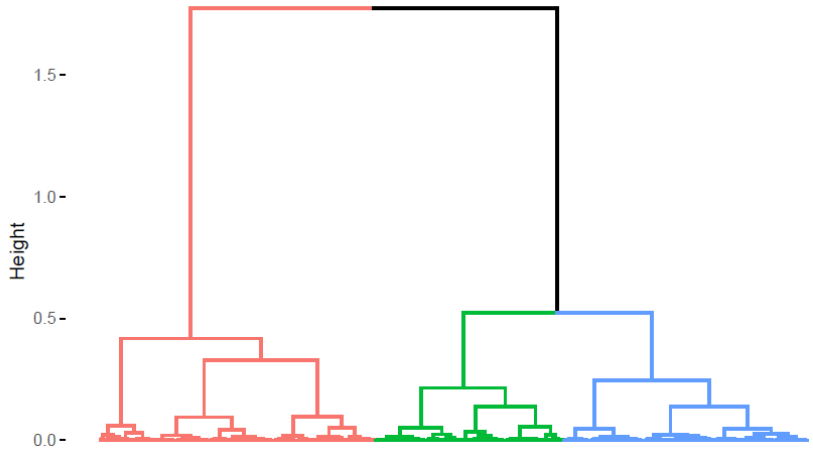

Figure 19: Dendrogram generated by hierarchical clustering performed on the first three principal components.

	v.test	Mean in category	p.value
1.length.prof	8.10	0.17	<0.001
1.n.updates.photo	-13.57	-0.29	<0.001
1.sent.ana	-13.58	-0.29	<0.001
1.n.photos	-23.73	-0.50	<0.001
1.n.matches	-32.35	-0.69	<0.001
1.score	-32.63	-0.69	<0.001
1.score_log	-35.84	-0.76	<0.001
2.n.photos	32.95	1.01	<0.001
2.score_log	-2.07	-0.06	0.04
2.n.updates.photo	-5.76	-0.18	<0.001
2.score	-6.86	-0.21	<0.001
2.length.prof	-8.23	-0.25	<0.001
2.n.matches	-8.46	-0.26	<0.001
2.sent.ana	-10.04	-0.31	<0.001
3.n.matches	42.43	1.14	<0.001
3.score	41.22	1.11	<0.001
3.score_log	40.11	1.08	<0.001
3.sent.ana	23.95	0.65	<0.001
3.n.updates.photo	19.88	0.54	<0.001
3.n.photos	-5.90	-0.16	<0.001

Table 3: Table of quantitative variables that describe the most each cluster.

Unlike k-means clustering, the number of clusters does not need to be specified beforehand. We have already computed the principal components and we keep only the first three principal components. Next, hierarchical clustering is applied on the result of the PCA. The objects are categorized into a hierarchy similar to a tree-like diagram which is called a dendrogram as shown in Figure 19. The dendogram suggests a 3 cluster solution. It indicates both the similarity and the order that the clusters were formed. From Table 3, we can see the variables associated to each of the three clusters. The means are the scaled means, hence the values can be negative and the base overall mean is taken to be zero.We can see that score, number of matches and number of photos are most significantly associated with cluster 1 and cluster 3 and for cluster 2, the variable number of photos is the one which is most strongly associated to it.

3.5 How Hierarchical clustering(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 saves 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 to 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 finds the pair of clusters that leads to a minimum increase in total within-cluster variance after merging.

3.6 Pros and cons of HC compared to k-means

Advantages:

- It produces a series of clusterings in various granularity. This reveals inherent hierarchical structures hidden in the data set and provides rich information and insights into the data set.
- It does not necessitate the number of clusters as an input argument. It allows the user to choose the best decomposition of X from the dendrogram.

Disadvantages:

- Hierarchical clustering requires the computation and storage of an n×n distance matrix. For very large datasets, this can be expensive and slow.
- The order of the data has an impact on the final results.
- Very sensitive to outliers

4 Appendix

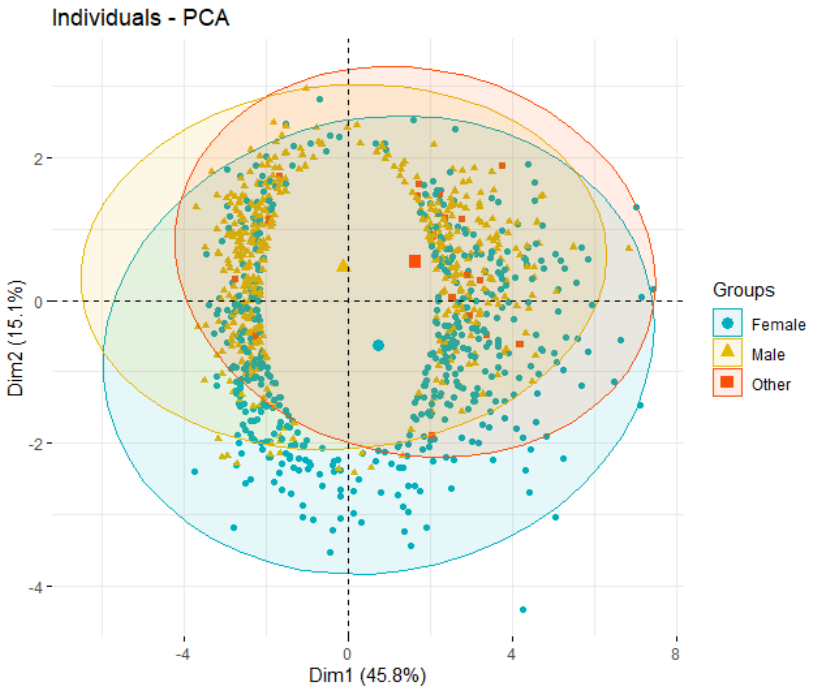

Figure 20: Individual map with a sample of individuals with ellipses

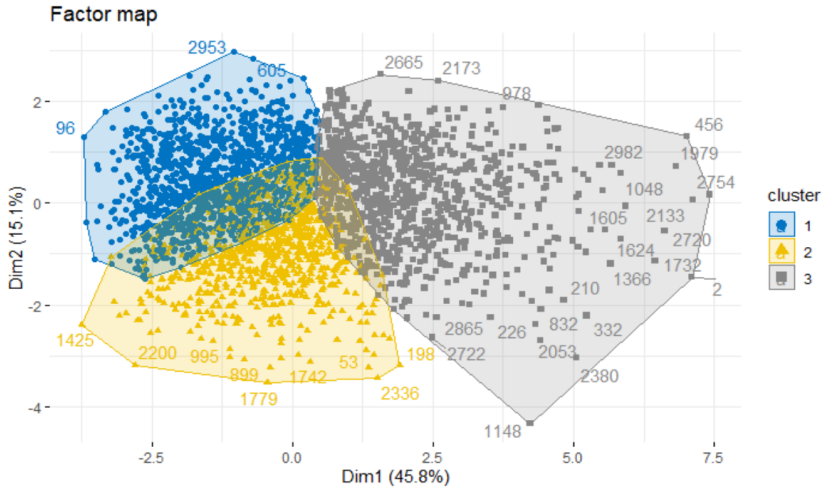

Figure 21: Three dimensional plot combining the hierarchical clustering and the factorial map

5 References

Charrad, Malika. "(PDF) NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set." ResearchGate, June 2012, https://www.researchgate.net/publication/267210575$_NbClust_An_RPackage_for_Determining_the_Rel$

"The Complete Guide to Clustering Analysis: K-Means and Hierarchical Clustering by Hand and in R." Stats and R, https://statsandr.com/blog/clustering-analysis-k-means-and-hierarchical-clustering-by-hand-and-in-r/.

Desagulier, Guillaume. "Multivariate Exploratory Approaches." HAL, 10 Feb. 2020, https://halshs.archives-ouvertes.fr/halshs-01926339v3.

"Interpret All Statistics and Graphs for Principal Components Analysis." Minitab, 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/.

"Interpretation and Visualization." Interpretation and Visualization • SOGA •, 1 May 2017, https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/Principal-Component-Analysis/principal-components-basics/Interpretation-and-visualization/index.html.

Kassambara, Alboukadel. “Chapter 12 : Determining the Optimal Number of Clusters.” Practical Guide to Cluster Analysis in R Unsupervised Machine Learning, STHDA, Frankreich, 2017, pp. 128–137.

Kassambara, Alboukadel. Practical Guide to Principal Component Methods in R. STHDA, 2017.

Kaufman, Leonard, and Peter Rousseeuw. “Finding Groups in Data: An Introduction to Cluster Analysis.” Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley Sons, 1990, https://www.researchgate.net/publication/31839302

Lee, I, and J Yang. “Hierarchical Clustering.” Hierarchical Clustering - an Overview | ScienceDirect Topics, 2009, https://www.sciencedirect.com/topics/mathematics/hiera clustering.

Rick Wicklin on The DO Loop. “What Are Biplots?” The DO Loop, 6 Nov. 2019, https://blogs.sas.com/content/iml/2019/11/06/what-are-biplots.html.

Santini, Marina. Advantages Disadvantages of k-Means and Hierarchical Clustering (Unsupervised Learning), 2016, http://santini.se/teaching/ml/2016/Lect$_1$0/10$c_U nsup$e

Soetewey, Antoine. The Complete Guide to Clustering Analysis: k-Means and Hierarchical Clustering by Hand and in R, 2020,https://statsandr.com/blog/clustering-analysis-k-means-and-hierarchical-clustering-by-hand-and-in-r/.