CY Tech Sciences et Techniques

Dimensionality Reduction and Clustering Techniques

Author
Anh Thu DOAN

November 26, 2021

Contents

List of Figures

1 Presentation of the data set

The data used in this study is an artificial dataset containing data on user-profiles and conversations from an imaginary dating app. We are going to use dimensionality reduction and clustering techniques to explore this synthetic dataset(artificial as in created by the instructor for the needs of the class))

userid :	id of the user
date.crea :	date of the creation of the account
score :	score of the profile (how liked is the profile by other users)
n.matches :	total number of matches the user has had since account creation (with conversation)
n.photos :	number of photos on the profile
last.up.photo :	last time the user updated profile pictures
last.pr.update :	last time the user updated profile text
last.connex :	last time the user updated profile text
gender.	O is male, 1 is female. 2 is ”other” (notably if the user did not want to specify gender)
sent.ana :	sentiment score for the text of the profile
length.prof :	Number of words in the profile text
voyage :	Keyword voyage found in the profile text
laugh :	Keyword laugh found in the profile text
photo.keke :	one of the profile pics comports a photo without a T-shirt / with sunglasses / selfie in an elevator
photo.beach :	one of the profile pics comports a photo taken on the beach

2 Identifying correlations in the variables

2.1 Pearson correlation test

To evaluate the association of the relationship between the scores of one's profile and how many matches the user has, we use a correlation test to answer the question.
The correlation coefficient of 0,902 suggests a strong positive correlation between scores and matches. As the p-value for the test is much smaller than 0.05, the null hypothesis (r=0) is rejected. There is strong evidence to suggest that the correlation coefficient is different from 0.

Check that the variables are typically distributed, we are using histograms. From Fig. 1, we can see that both variables are normally distributed.

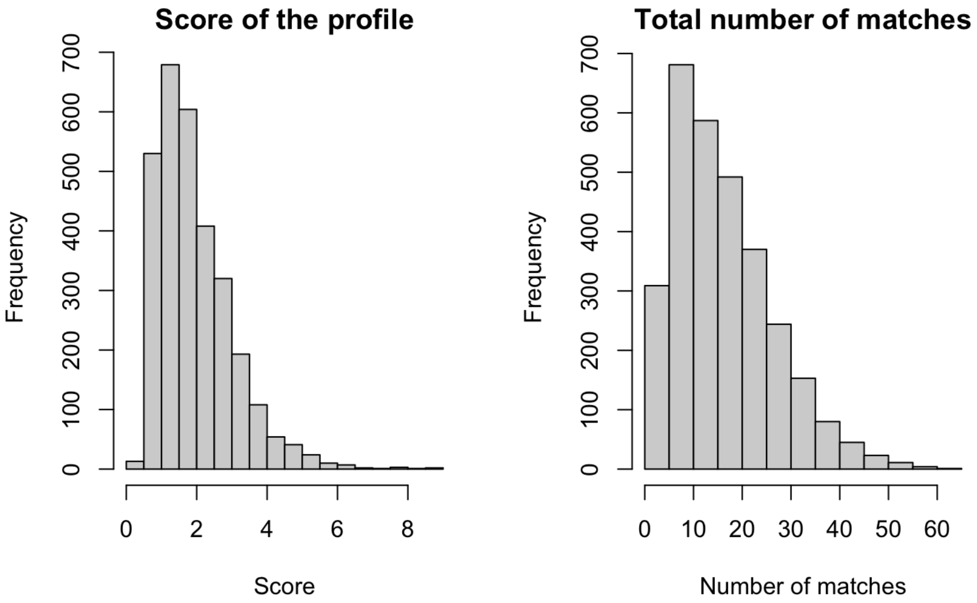

Figure 1: Histograms of Score and Matches variable

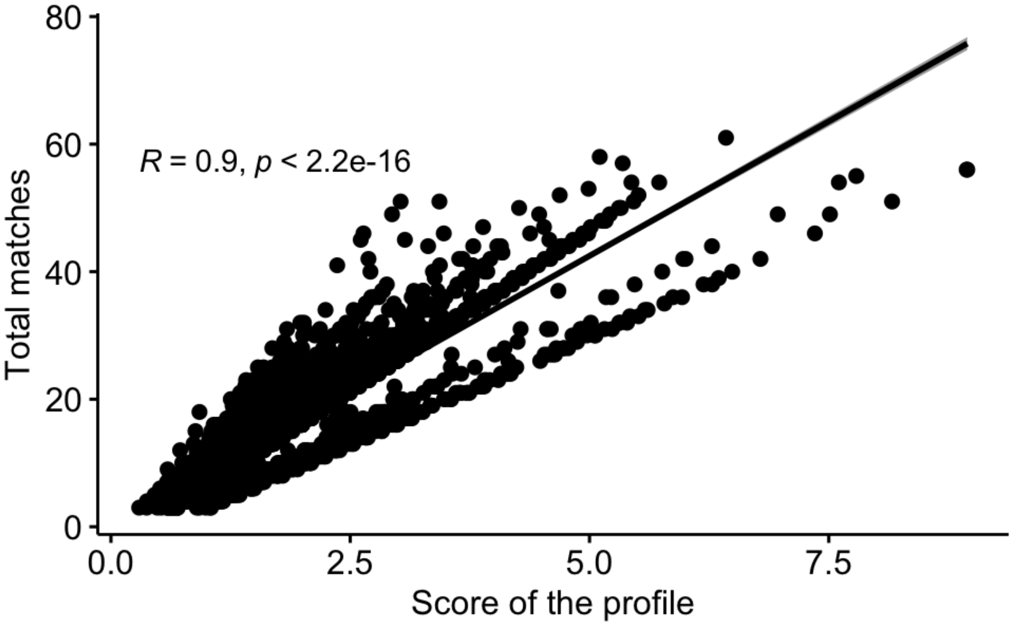

Figure 2: Scatter plot between Score and Matches variables

The presented Fig. 2 illustrates the correlational relationship between the two variables has been mentioned above. As the score increases, the total of matches the user gets will also increase. Moreover, the correlation seems weaker in the higher values.

2.2 Spearman rank correlation test

The correlation coefficient between gender of the users and do they have the photo keke in their profile are -0.145, and p-value equal to 1.165e-15, indicating a strong negative relationship.

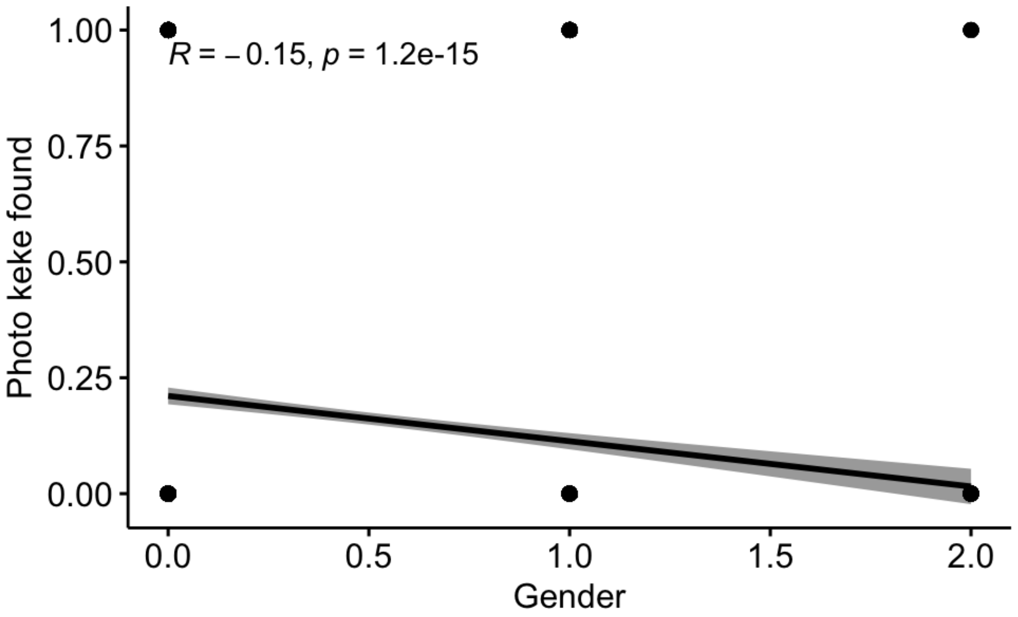

Figure 3: Scatter plot between Gender and Photo keke variables

From the Fig. 3, we can clearly see that the negative correlation between Gender and Photo keke. There is no relationship between them both.

2.3 Reporting correlation

The correlations between several variables can be displayed in Table 1. As we can see from Table 1, there are primarily positive numbers which mean that also positives relationship between them. The negative correlation always happened with the Number of photos and the other one. From that, we can say that the Number of photos does not correlate with the other variables.

	Score	N.matches	N.u.photos	N.photos
Score	1.00	0.90	0.29	0.05
N.matches	0.90	1.00	0.32	-0.01
N.u.photos	0.29	0.32	1.00	-0.02
N.photos	0.05	-0.01	-0.02	1.00

Table 1: Table of correlations between several variables

3 Dimensional Reduction

3.1 Scree plot (PCA)

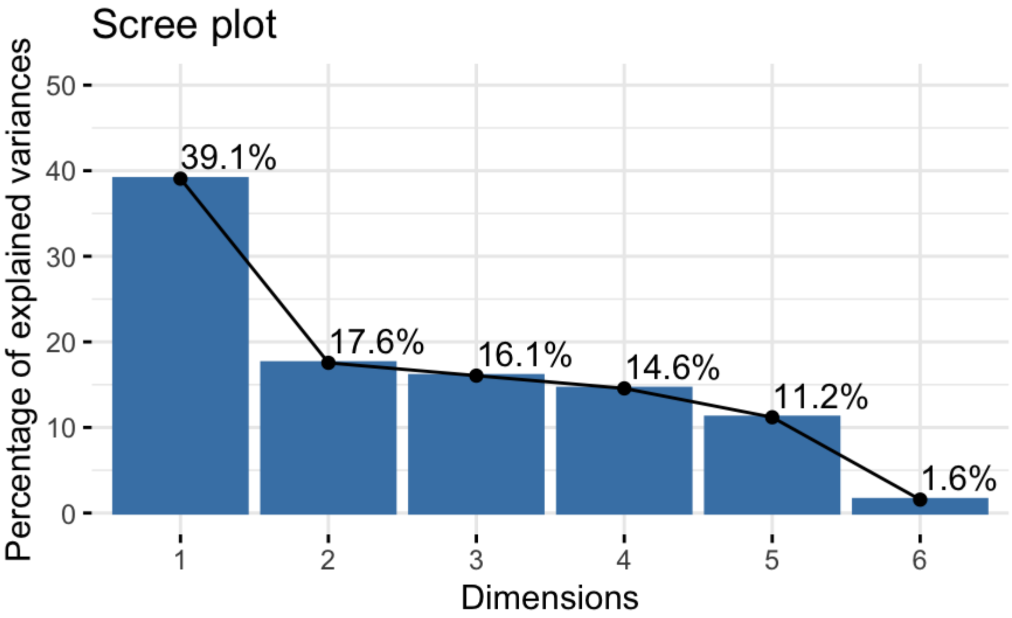

Figure 4: Scree plot

From Fig. 4, we might want to stop at the third principal component. The first three principal components explain 72.8% of the variation. This is a significant enough percentage to be considered acceptable. The fraction of variance that can be explained reduces dramatically after the first component. It implies that the second component's addition has an effect on the explained variance.

3.2 Variables factor map (PCA)

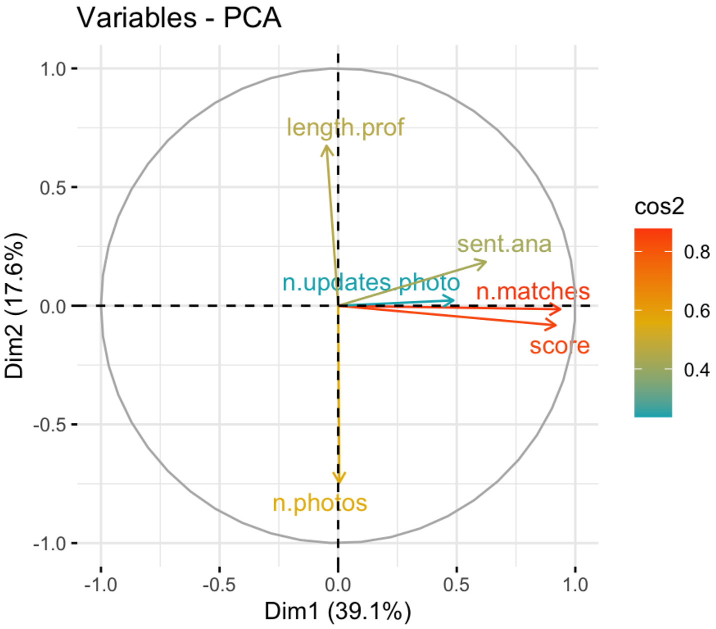

Figure 5: Variables factor map (PCA)

In Fig. 5, the colors for variables are automatically controlled by their qualities of representation. The Variables factor map presents the first principal component explaining 39.1% of the total variation, and the second principal component an additional 17.6%. So the first two principal components explain nearly 56.7% of the total variance. The first two-component correlate almost perfectly with the variable Number of matches and Length.prof of the user variables, respectively.

3.3 The Individuals factor map (PCA)

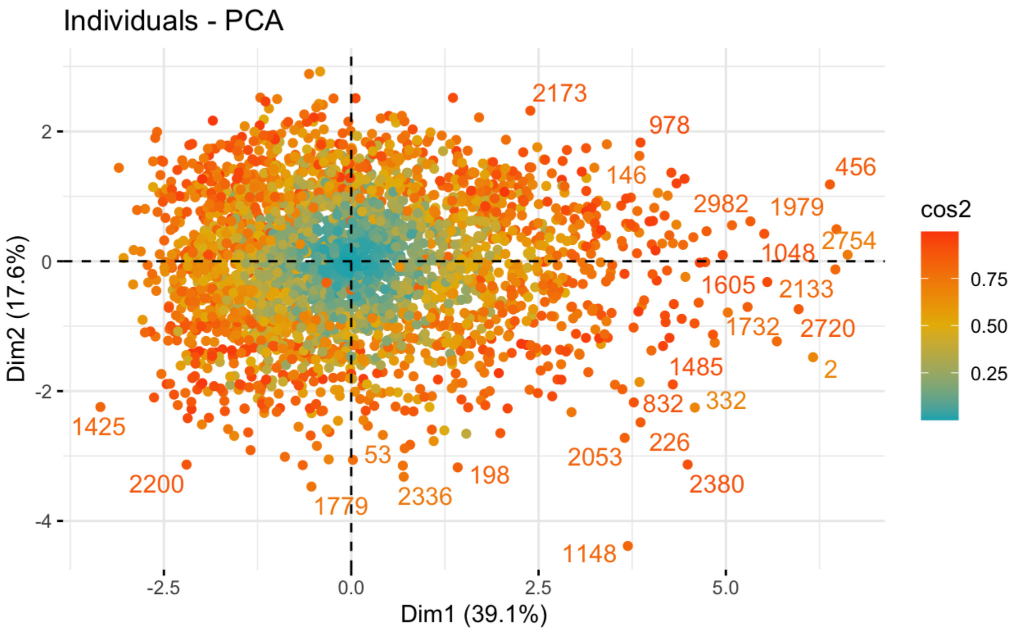

Figure 6: Individuals factor map (PCA)

In Fig. 6, the colors for individuals are automatically controlled by their qualities of representation. Similar individuals are grouped on the plot. In addition, we make another Individual factor map, but now we group it by gender to research the difference between gender. As it is shown in Fig. 7, there is an overlap between three different gender type, but we can still see that female has the most significant area in the plot. Furthermore, the individual from Female scores high on component 1, and component 2 seems to be connected with the other group.

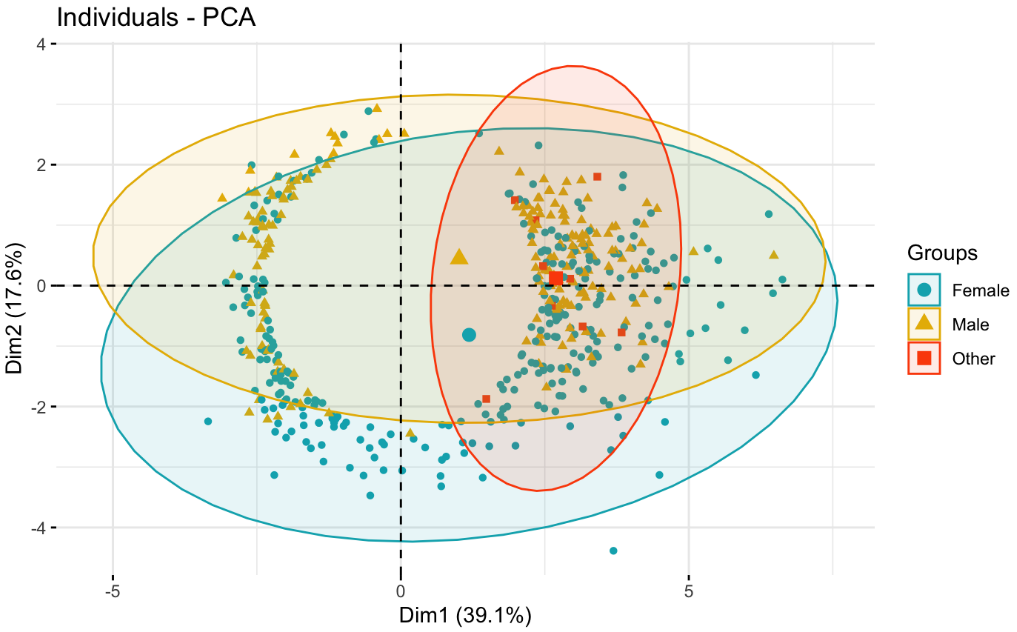

Figure 7: Individuals factor map group by gender (PCA)

3.4 Biplot (PCA)

Fig. 8 represents the Individuals and Variables factor in the same plot called Biplot. Based on the Fig. 4 Scree plot, it is found that the 1st component and 2nd component explain 56.7% of the variance of components so that the results obtained by the biplot analysis are pretty good in explaining the variance of the data.

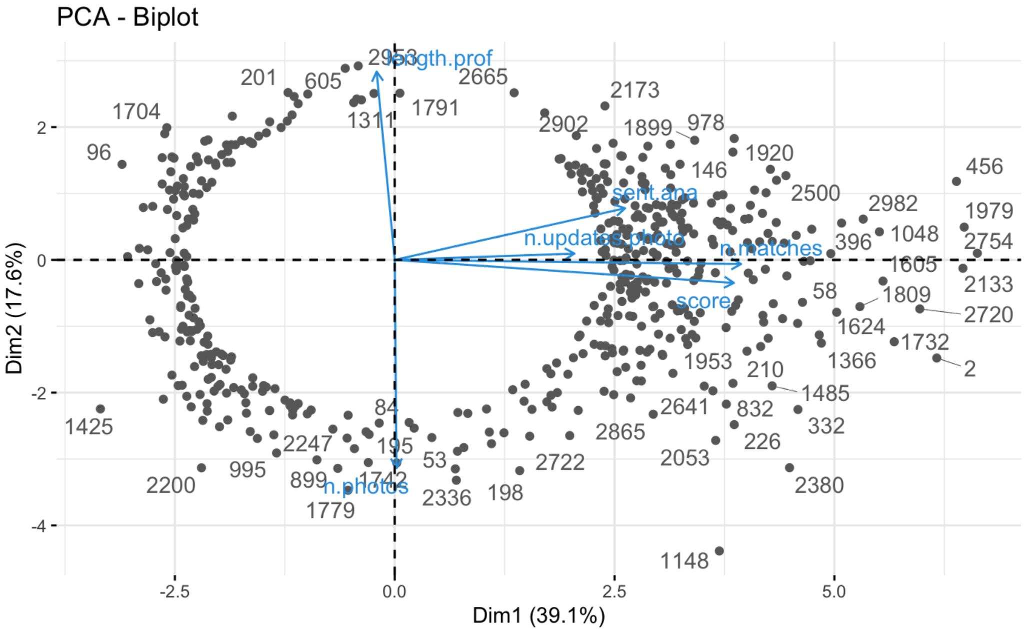

Figure 8: Biplot (PCA)

3.5 Table of loadings (PCA)

	PC1	PC2	PC3	PC4	PC5	PC6
score	0.60	-0.08	0.06	0.08	0.38	-0.69
n.matches	0.61	-0.01	0.03	0.07	0.32	0.72
n.updates.photo	0.32	0.02	-0.12	-0.88	-0.34	-0.02
n.photos	0.00	-0.73	0.66	-0.04	-0.18	0.04
sent.ana	0.41	0.18	0.02	0.45	-0.77	-0.04
length.prof	-0.03	0.66	0.74	-0.13	0.08	-0.01

Table 2: Table of loadings

Table 2 of loadings of all variables for each of the principal components that were studied. As we can see from Table 2, score loadings are considered the most important for the fifth principal component. Similarly, with n.matches variables has strong positive correlations, and the first component (as we mentioned before in 3.2. In short, those loadings that are considered the most important for each principal component

3.6 MCA

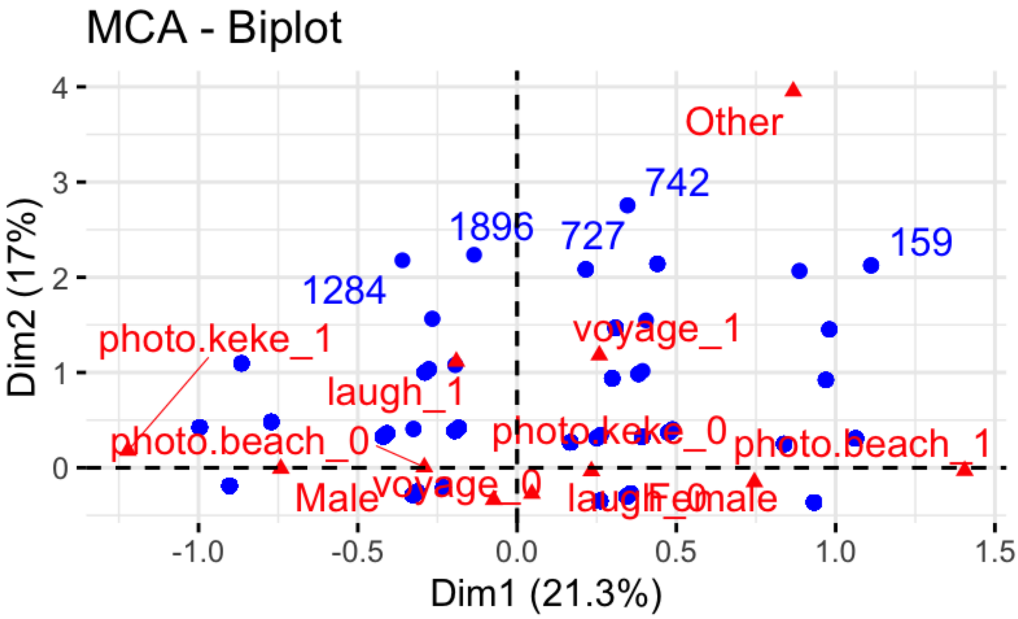

Figure 9: Biplot (MCA)

Fig. 9 shows the individuals and variables categories of the data by Biplot of Multiple Correspondence Analysis (MCA). The MCA results are interpreted as the results from a simple correspondence analysis. Individual variables are represented by blue points and Variable categories by red triangles.

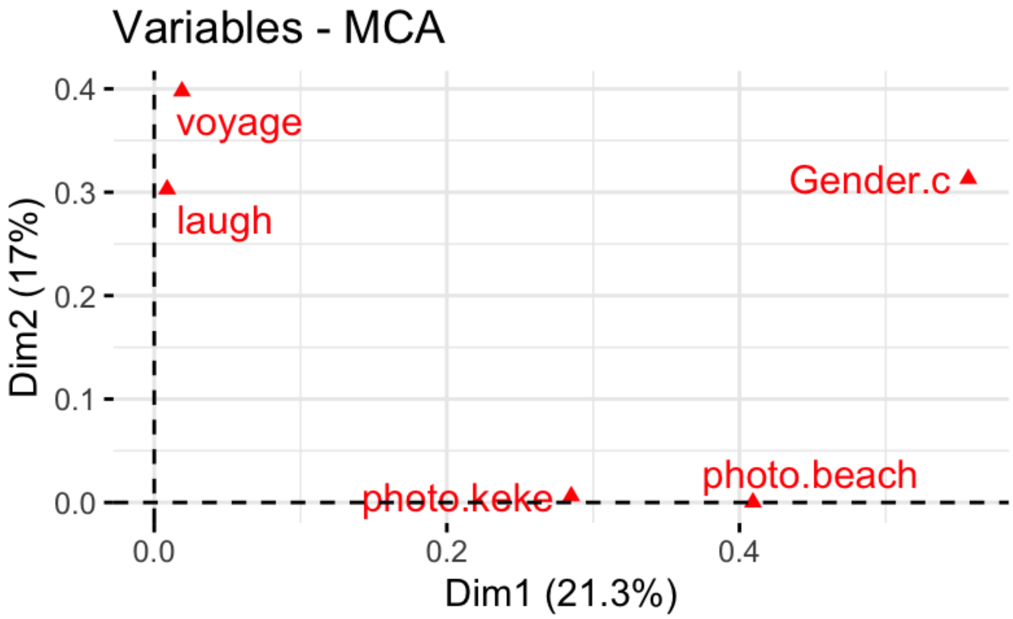

Figure 10: Variable categories (MCA)

It can be seen that, in Fig. 10 the variables photo.keke and photo.beach are the most correlated with dimension 1. Similarly, the variables are the most correlated with dimension 2.

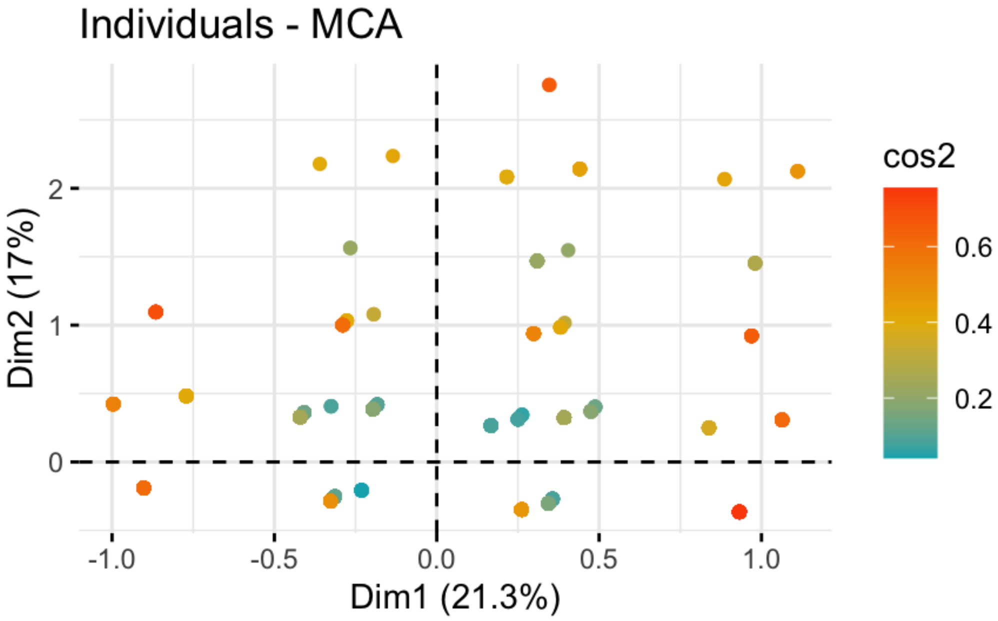

Figure 11: Individuals (MCA)

The Individuals-MCA (Fig. 11) gives an idea of what pole of the dimensions the individuals are actually contributing to. For instance, the colors for individuals in Fig. 11 are automatically controlled by their qualities : the individuals with low cos2 values will be colored in “blue” the individuals with mid cos2 values will be colored in “orange” the individuals with high cos2 values will be colored in “red”

4 K-means and Hierarchical Clustering

4.1 K-mean Clustering

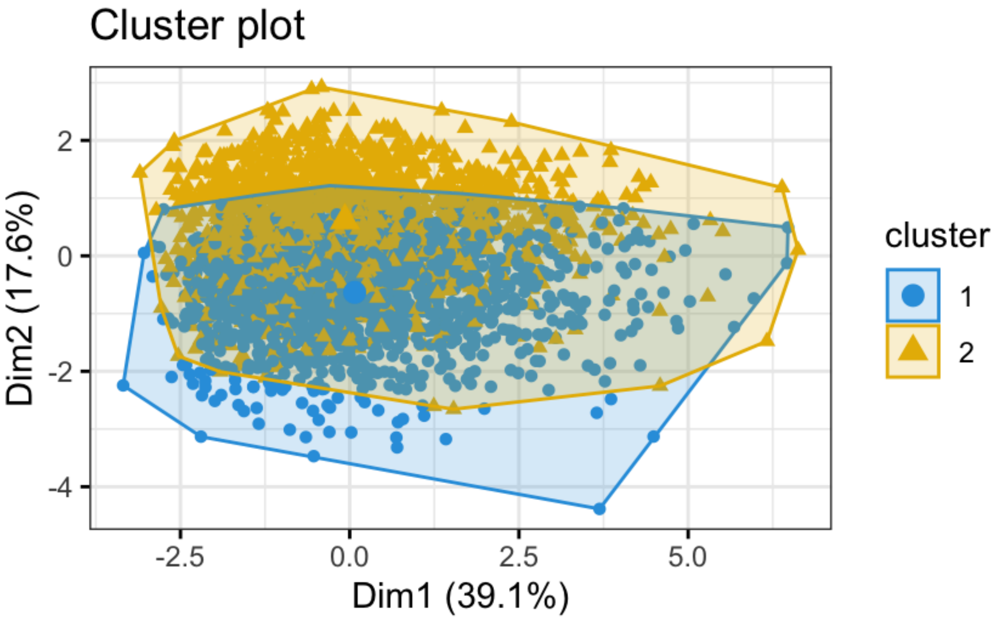

Figure 12: K-mean Clustering

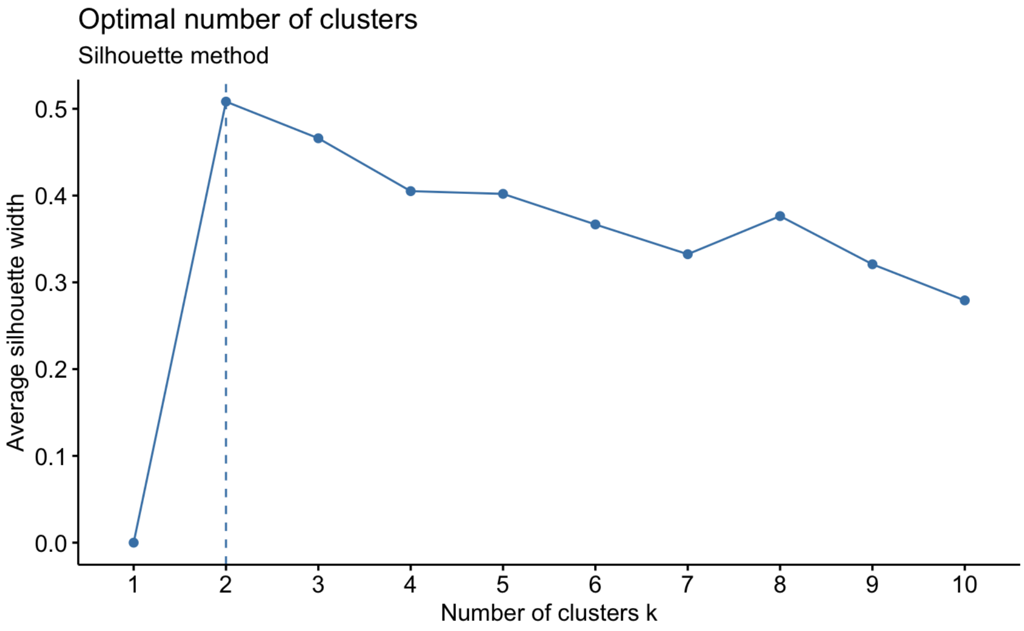

Figure 13: Silhouette method to find the optimal number of clusters for k-means algorithm.

Fig. 12 shows the representation of data of two different items. the first item has shown in blue color and the second item has shown in yellow color. Here we are choosing the value of K randomly as 2. Using Silhouette method to select the right number of clusters, this indicates that the quality of the model is no longer improving substantially as the model complexity (i.e. number of clusters) increases. If we take a look at Fig. 13, we set k equal to the number of dimensions at the location of the maximizes the average silhouette over a range of possible values for k which is 2.

4.1.1 How K-means works?

K-means clustering attempts to group similar types of items into clusters. It detects similarities between objects and groups them into collections. The K-means clustering algorithm operates in three stages: Choose the k values; Set up the centroids. Find the average for the group.
The total within-cluster sum of squares is used as the measurement by 'k-means' in R. The best model is the 'k-means' run with the lowest total within the cluster sum of squares. The sum of squares within a cluster is simple to compute. Calculate the squared distance from the observation to

the cluster center for each cluster in the model and each observation assigned to that cluster. This is simply the Euclidean distance squared from the plane geometry class. Add up all of the calculated squared distances to get the total within-cluster sum of squares.

4.2 Hierarchical Clustering

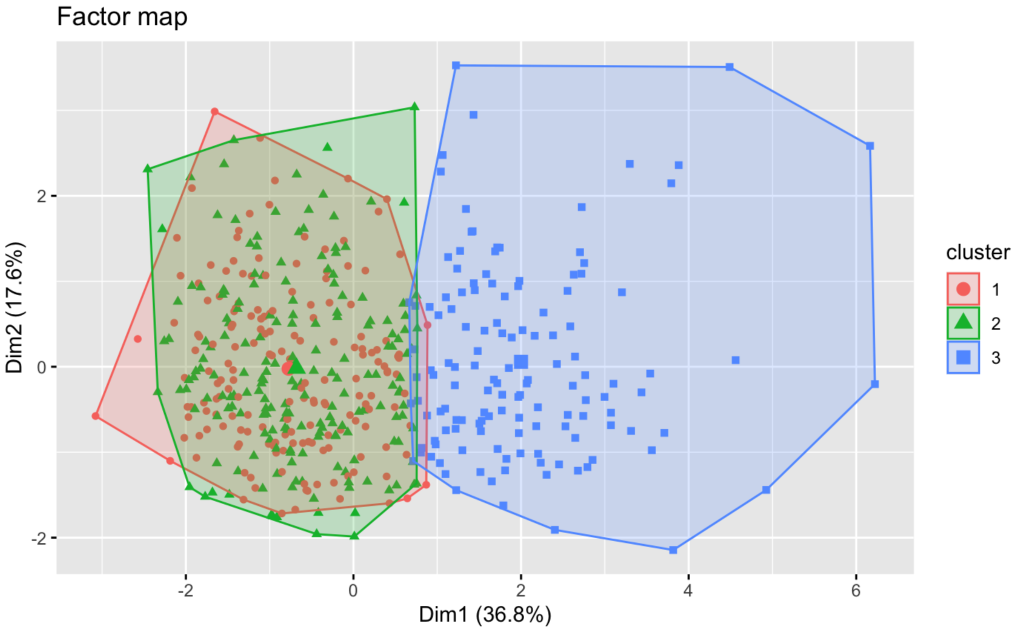

Figure 14: Factor map Hierarchical Clustering

Fig. 14 visualize individuals on the principal component map, and there are three different color individuals according to the cluster they belong to.

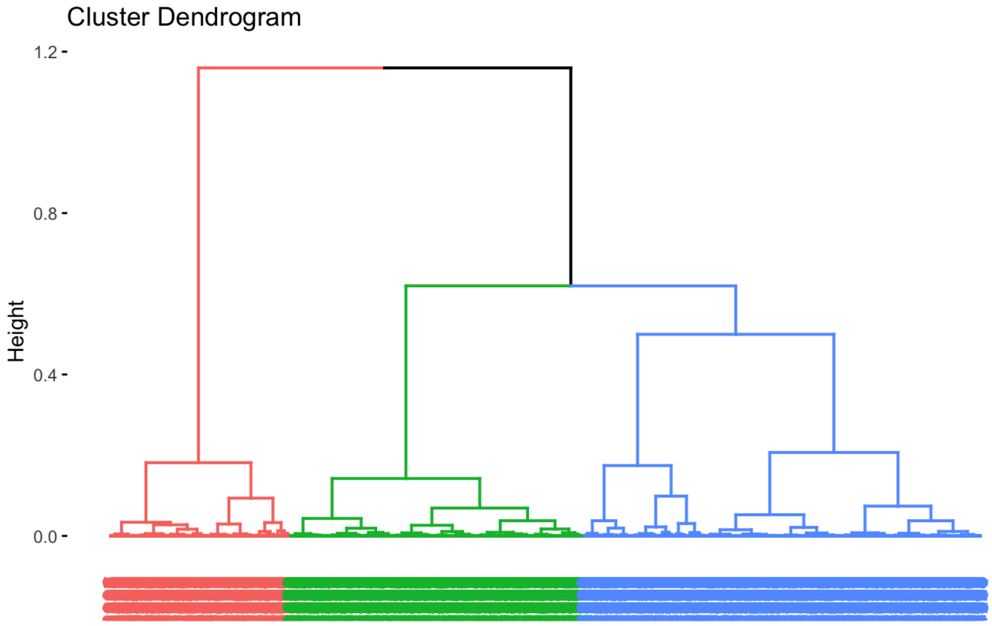

Figure 15: Hierarchical clustering

In Fig. 15, the horizontal axis represents the clusters. The vertical scale on the Dendrogram represents the distance or dissimilarity. Each joining (fusion) of two clusters is represented on the diagram by splitting a vertical line into two vertical lines. The vertical position of the split, shown by a short bar, gives the distance (dissimilarity) between the two clusters.

4.2.1 How Hierarchical clustering(HC) works?

Hierarchical clustering is a method of clustering data by grouping objects into hierarchical clusters based on the similarity of objects in the data. The hierarchical order of these clusters forms a dendrogram tree structure. The algorithm implementation process can be summarized in the following steps:

1. Consider each object, each observable object as a child cluster or single cluster. Now a cluster has only one observation

2. Calculate the distance between individual clusters and highlight the smallest values to make clustering easier.

3. Proceed to include clusters of pairs of clusters according to the minimum distance rule.

4. Step 3 continues until there is only 1 cluster containing all objects.
5. Display the results on the Dendrogram and determine the location of the branch break to determine the number of clusters we want to find.

4.2.2 K-means vs Hierarchical clustering

K-Means

Advantages: There is a guarantee of convergence. Specializing in clusters of various sizes and shapes.

Disadvantages: It is difficult to predict the K-Value. It did not work well with the global cluster.

Hierarchical clustering

Advantages: Ease of dealing with any form of similarity or distance. As a result, it applies to any attribute type.

Disadvantage: Hierarchical clustering necessitates the calculation and storage of an nxn-dimensional distance matrix. This can be costly and time-consuming for enormous datasets.