CY Tech Sciences et Techniques

Dimensionality Reduction
&
Clustering

Author
Buu NGUYEN

November 18, 2021

Contents

1 Data Pre-processing 2

2 Identifying correlations in the variables 2

3 Dimensionality Reduction 6

4 Clustering 10
4.1 K-Means 10
4.2 Hierarchical Clustering 11

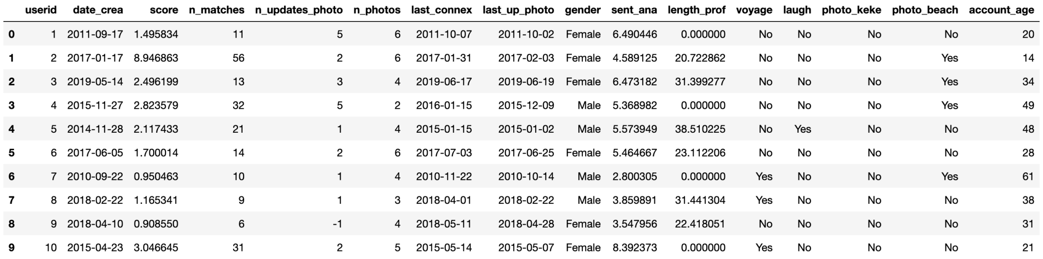
List of Figures

1 Dataset's head 2
2 Pearson correlation matrix 3
3 Cramers V correlation matrix 4
4 Account age by Gender 5
5 Number of photos by Gender 5
6 PCA Scree plot 7
7 Variables' correlations circle 8
8 Loadings plot of each Principal Component 9
9 K-Means Scree plot 10
10 K-Means Scatter plot 11
11 HC Dendograms 12
12 HC Scatter plot 13
13 Individual map with 25% sample of individuals 14
14 PCA Biplot with 10% sample of individuals 15

1 Data Pre-processing

There are fifteen variables in the original dataset.

- We use **userid** variable as the index.
- The **last.pr.update** column only contains NaN value, therefore we drop it.
- We create a new **account.age** variable which indicates how long this account exist from the creation date till the last connection date from **date.crea** and **last.connex**.



	userid	date_crea	score	n_matches	n_updates_photo	n_photos	last_connex	last_up_photo	gender	sent_ana	length_prof	voyage	laugh	photo_keke	photo_beach	account_age
0	1	2011-09-17	1.495834	11	5	6	2011-10-07	2011-10-02	Female	6.490446	0.000000	No	No	No	No	20
1	2	2017-01-17	8.946863	56	2	6	2017-01-31	2017-02-03	Female	4.589125	20.722862	No	No	No	Yes	14
2	3	2019-05-14	2.496199	13	3	4	2019-06-17	2019-06-19	Female	6.473182	31.399277	No	No	No	Yes	34
3	4	2015-11-27	2.823579	32	5	2	2016-01-15	2015-12-09	Male	5.368982	0.000000	No	No	No	Yes	49
4	5	2014-11-28	2.117433	21	1	4	2015-01-15	2015-01-02	Male	5.573949	38.510225	No	Yes	No	No	48
5	6	2017-06-05	1.700014	14	2	6	2017-07-03	2017-06-25	Female	5.464667	23.112206	No	No	No	No	28
6	7	2010-09-22	0.950463	10	1	4	2010-11-22	2010-10-14	Male	2.800305	0.000000	Yes	No	No	Yes	61
7	8	2018-02-22	1.165341	9	1	3	2018-04-01	2018-02-22	Male	3.859891	31.441304	Yes	No	No	No	38
8	9	2018-04-10	0.908550	6	-1	4	2018-05-11	2018-04-28	Female	3.547956	22.418051	No	No	No	No	31
9	10	2015-04-23	3.046645	31	2	5	2015-05-14	2015-05-07	Female	8.392373	0.000000	Yes	No	No	No	21

Figure 1: Dataset's head

Figure 1 shows ten first rows of our cleaned dataset.

2 Identifying correlations in the variables

We calculate the Pearson correlation coefficient between qualitative variables and achieve a correlation matrix as see in Figure 2.

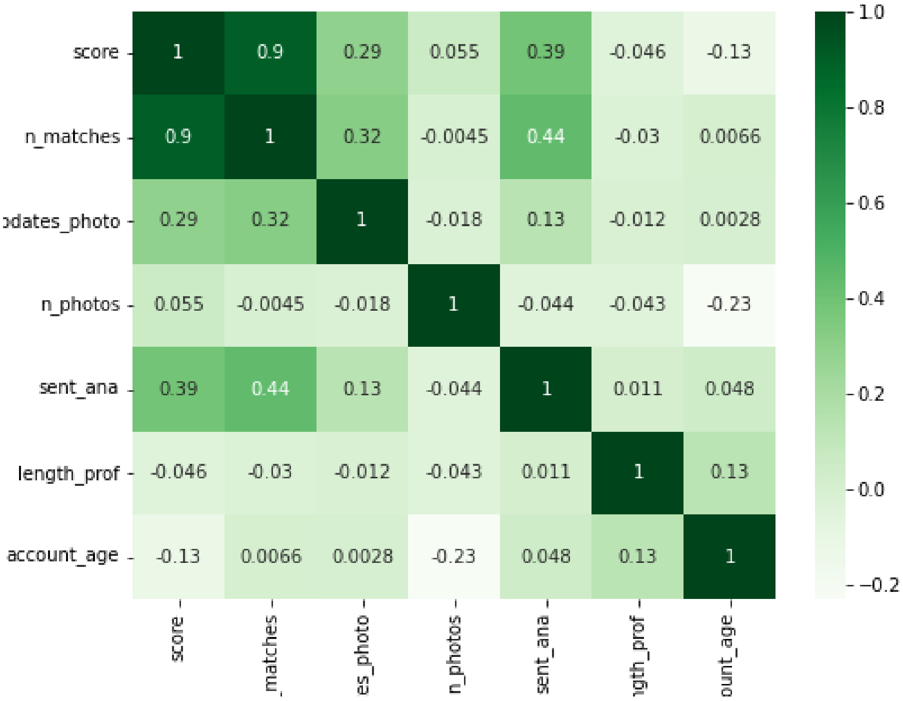


Figure 2: Pearson correlation matrix

From the correlation matrix, we can see:

- A very strong positive relationship between **score** and **n_matches** ($r = 0.9$).
- Both **updates_photo** and **sent_ana** correlates moderately and positively with **score** and **n_matches** ($r \in [0.40, 0.59]$). However, the relationship between them is very weak ($r = 0.13$).
- A very weak positive relationship between **account_age** and **length_prof** ($r = 0.13$).

For the quantitative variables, we calculate the Cramér's V point-biserial correlation. Figure 3 is the result correlation matrix.

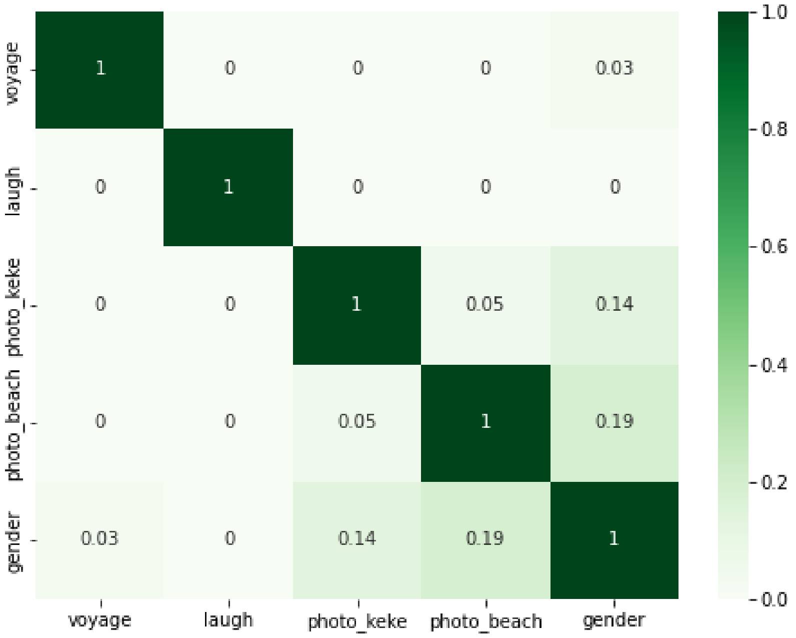


Figure 3: Cramers V correlation matrix

Figure 3 shows a moderate correlation between **gender** and **photo_keke** ($\phi = 0.14$), and a strong correlation between **gender** and **photo_beach** ($\phi = 0.19$).

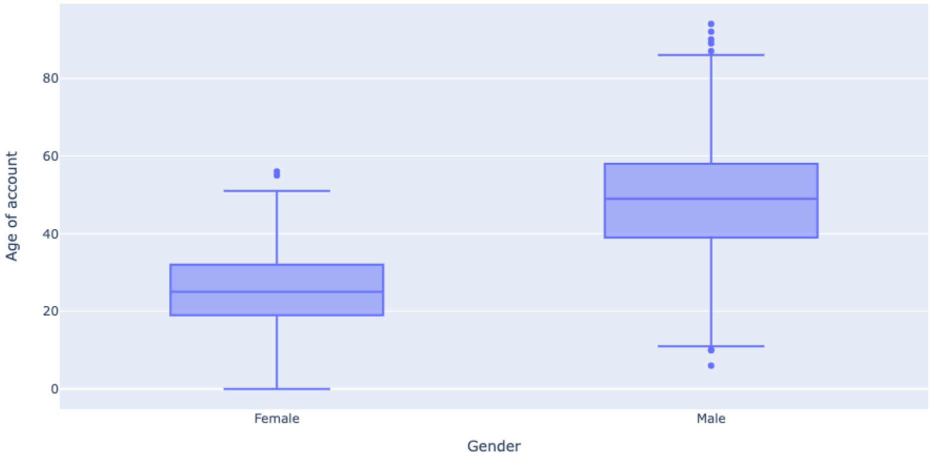


Figure 4: Account age by Gender

From Figure 4, we can see that Male users tend to use their account longer than the Female counterpart ($F = 2782.04$, $p << 0.001$).

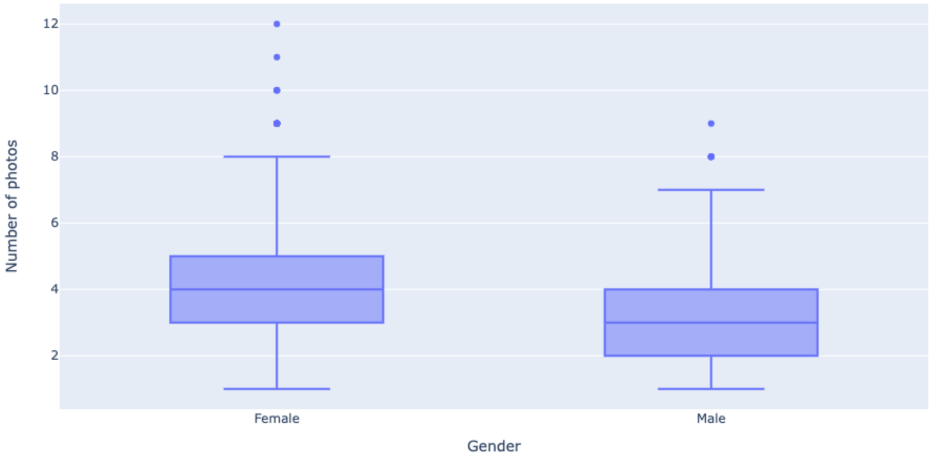


Figure 5: Number of photos by Gender

Figure 5 implies that Female users upload more photos than the Male counterpart in general ($F = 331.62$, $p << 0.001$).

3 Dimensionality Reduction

We perform Principal Component Analysis (PCA) on six continuous features: **n_matches**, **n_updates_photo** **n_photos**, **sent_ana**, **length_prof** and **account_age**. In addition, We use **scores** as the quantity supplementary variable.

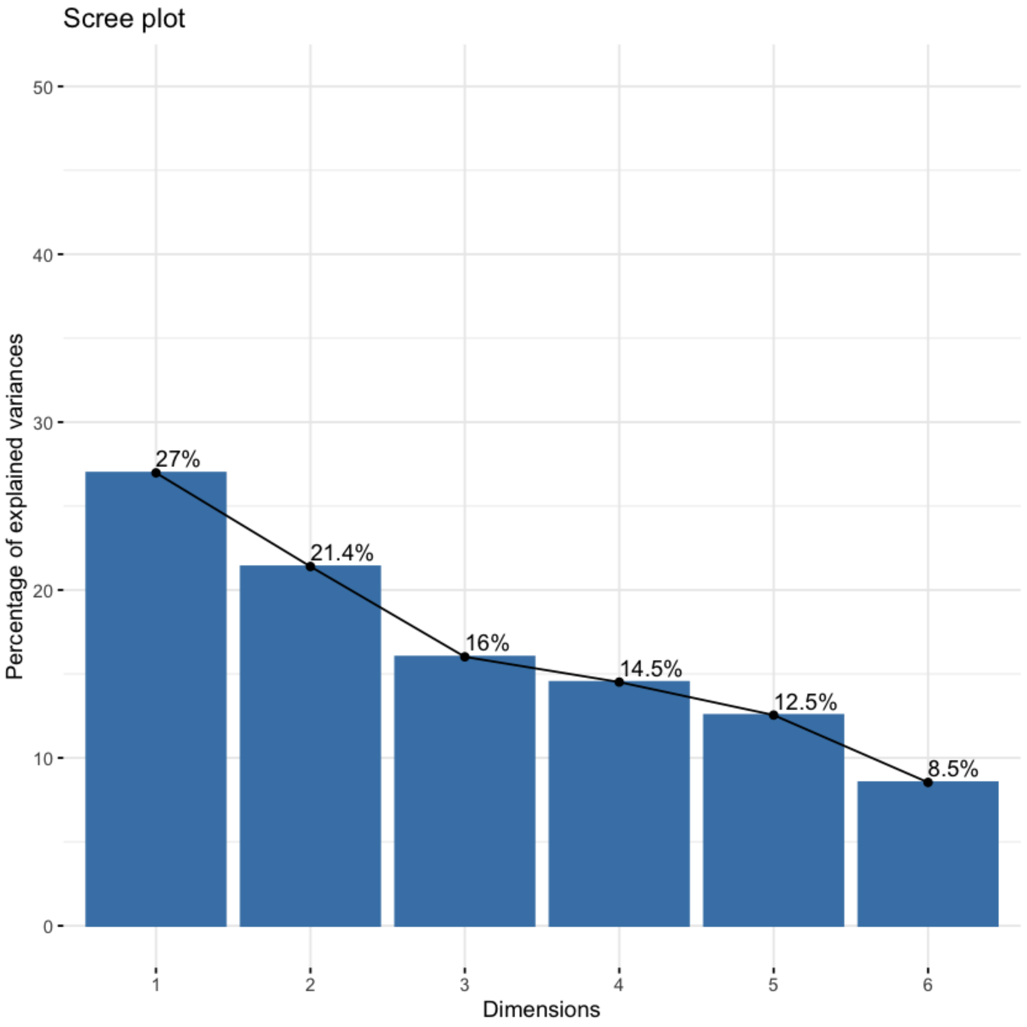


Figure 6: PCA Scree plot

From Figure 6, an estimation of the right number of dimension to interpret suggests to restrict the analysis to the description of the first two axis. We keep principal components that contain information more than one feature, In our case, it is component whose explained variance is higher than 16.67%: PC1 with 27% and PC2 with 21.4%. As a consequence, the description will stand to these components.

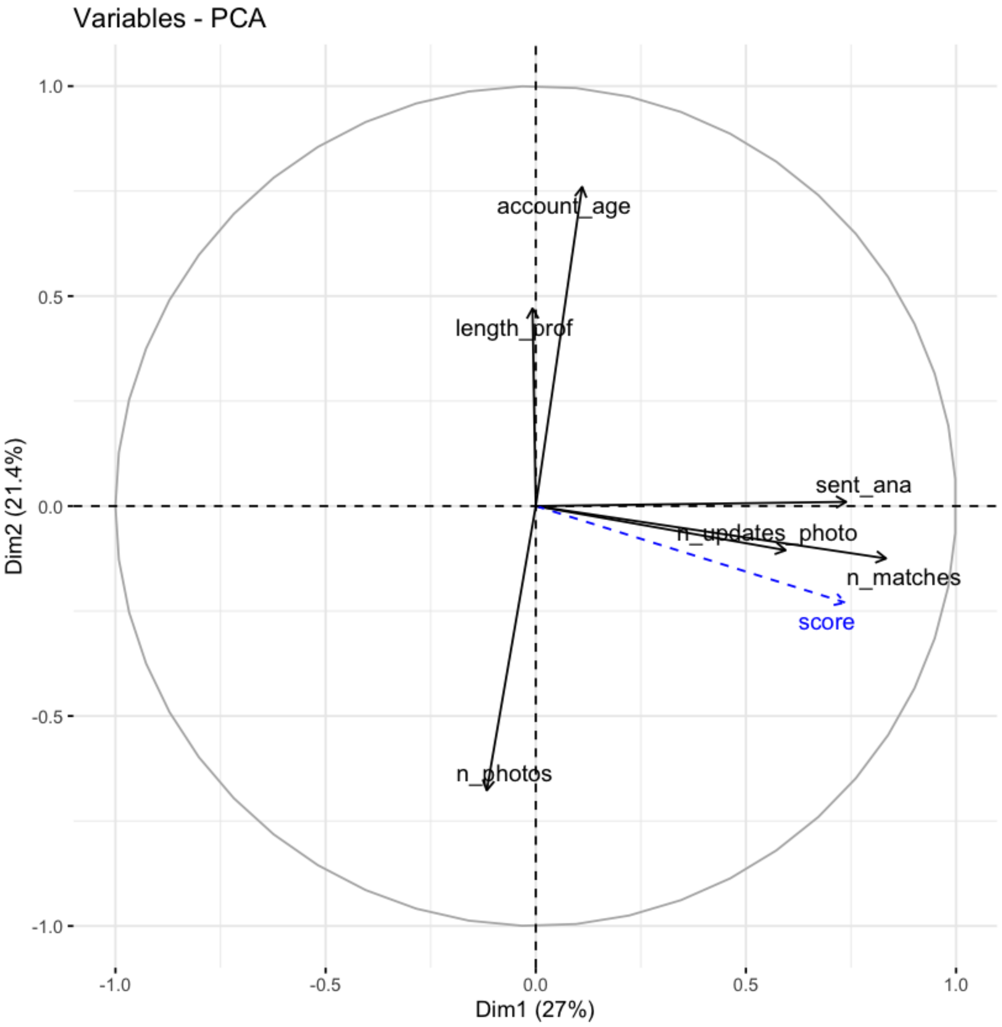


Figure 7: Variables' correlations circle

Figure 7 shows that the first principal component explains about 27% of the total variation, and the second principal component explains an additional 21.4%. So the first two principal components explain nearly 48.4% of the total variance.

The first component has a moderate positive relationship with **n_matches**, **score**, **n_updates_photos**, and **sent_ana**. Meanwhile, the second component correlates moderately and positively with **length_prof** and **account_age**. In addition, it also has a moderate negative relationship with **n_photos**.

We can see more clearly the coordinates between features and principle components in Figure 8.

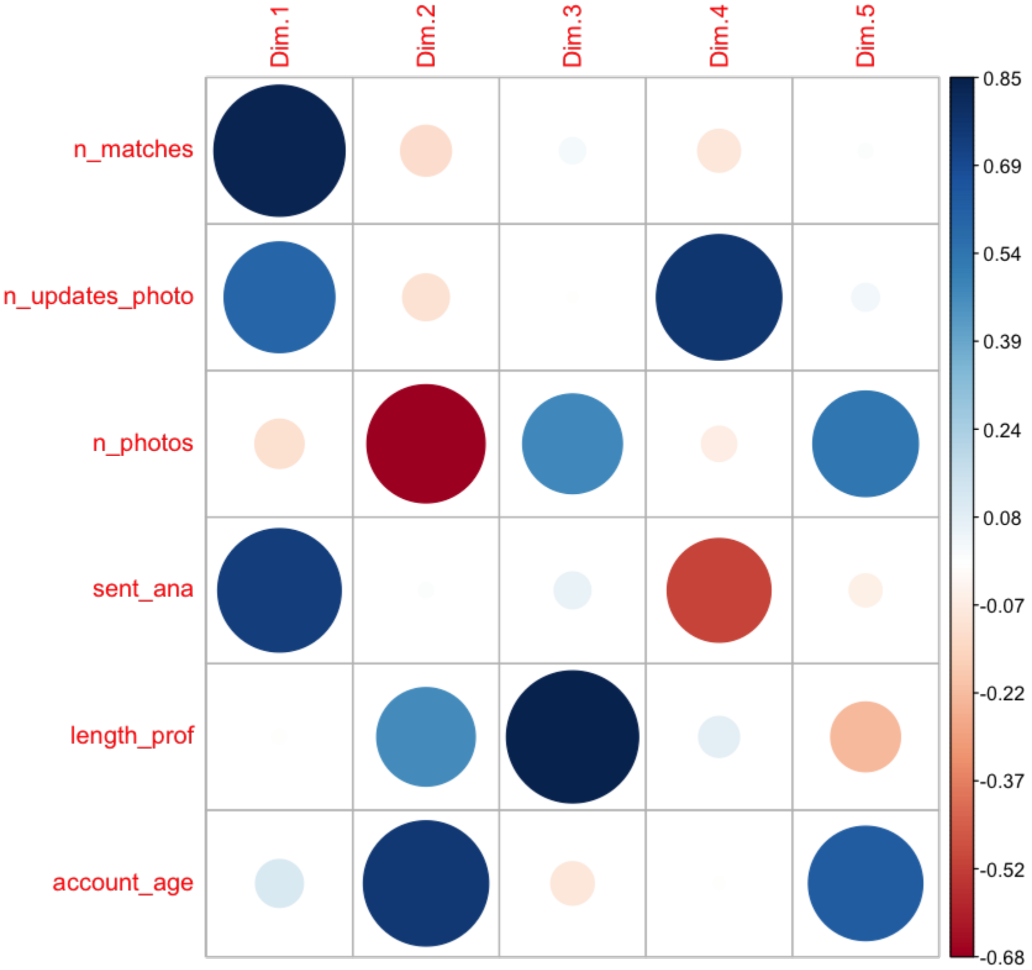


Figure 8: Loadings plot of each Principal Component

4 Clustering

4.1 K-Means

K-Means is a clustering algorithm that takes data points as input and groups them into k clusters. At the initiating state, we choose k centroid randomly and assign the data points to their closest centroid cluster. Then we move centroids to the center of their cluster and recalculate the points - centroids distance. We repeat this step until the centroids are at a steady-state.

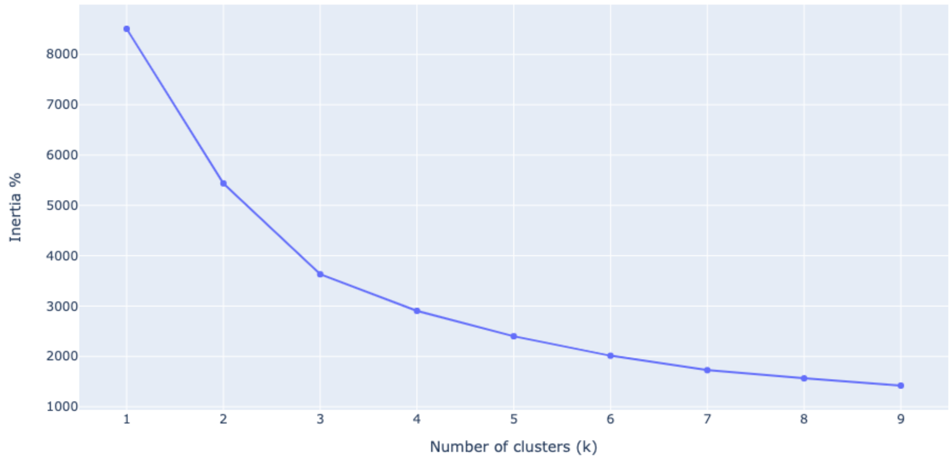


Figure 9: K-Means Scree plot

To determine the number of clusters for K-Means clustering, we generate a scree plot as see in Figure 9, then use the elbow method and got k=3 as the optimal one.

Applying K-Means clustering algorithm with k=3 on the two principal components, we got the clustering result in Figure 10 below.

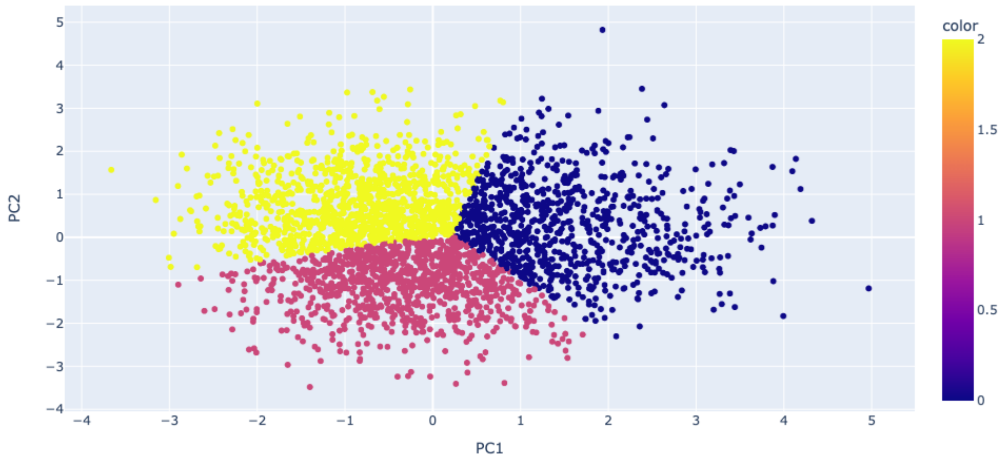


Figure 10: K-Means Scatter plot

We have 3 group of data:

- Group 1 (color=0): PC1 $>$ 0.4227
- Group 2 (color=1): PC1 $\leq$ 0.4227 and PC2 $\leq$ -0.1956
- Group 3 (color=2): PC1 $\leq$ 0.4227 and PC2 $>$ -0.1956

4.2 Hierarchical Clustering

In the Hierarchical Clustering algorithm, initially, each data point is considered as an individual cluster. Then, the similar clusters merge with other clusters at each iteration until one cluster or k clusters are formed. This algorithm does not contain randomness like K-Means, so the results are reproducible. However, it requires more computing power than K-Means and can be expensive and slow for massive datasets.

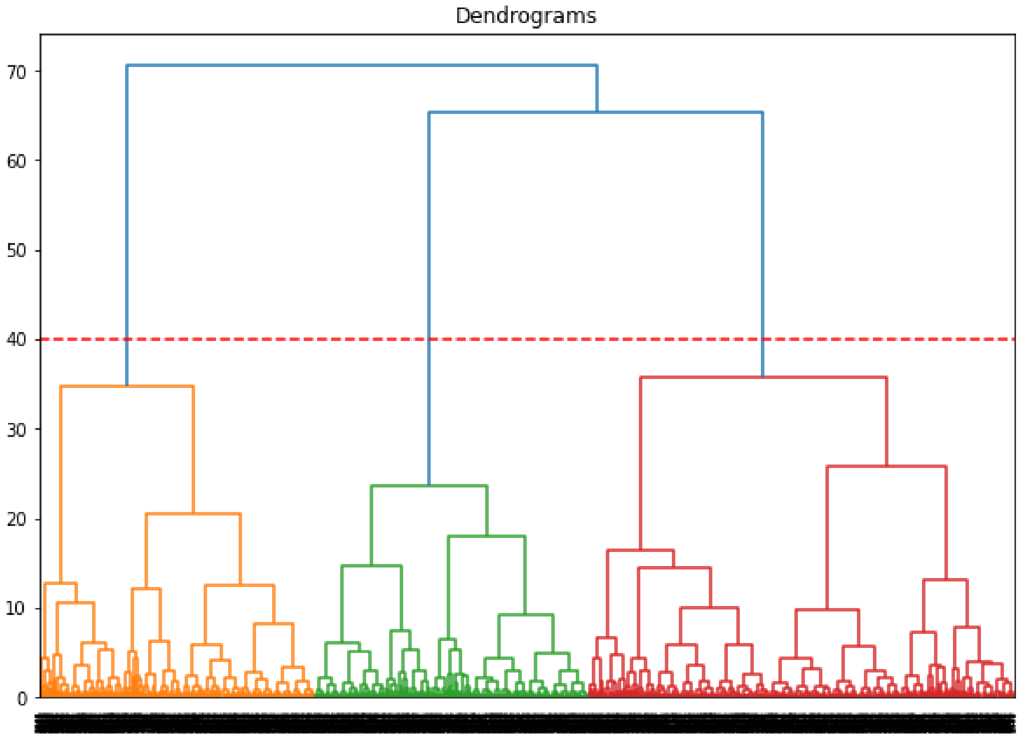


Figure 11: HC Dendograms

Figure 11 is the achieved Dendogram after we apply Hierarchical Clustering algorithm on the two principal components. We cut the tree at height equal 40 and got the clustering result in Figure 12 below.

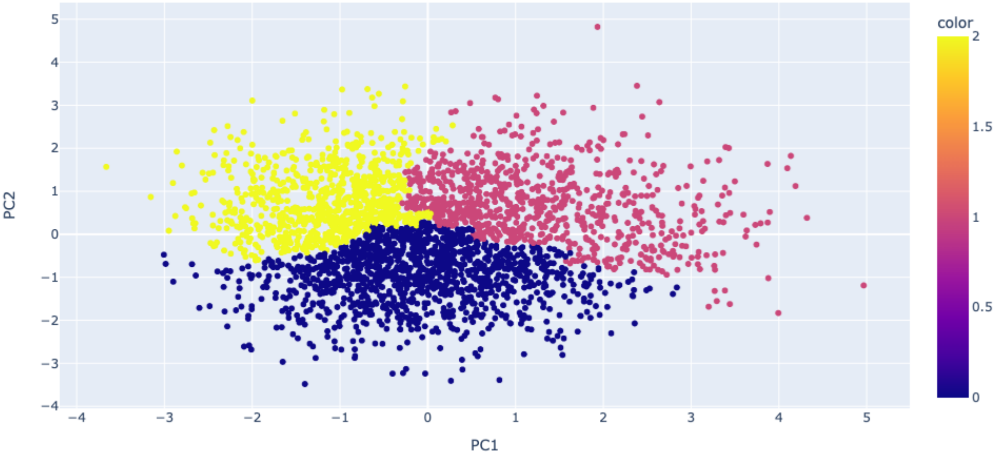


Figure 12: HC Scatter plot

We have 3 group of data:

- Group 1 (color=0): PC2 ≤ -0.2845
- Group 2 (color=1): PC2 > -0.2845 and PC1 > -0.2188
- Group 3 (color=2): PC2 > -0.2845 and PC1 ≤ -0.2188

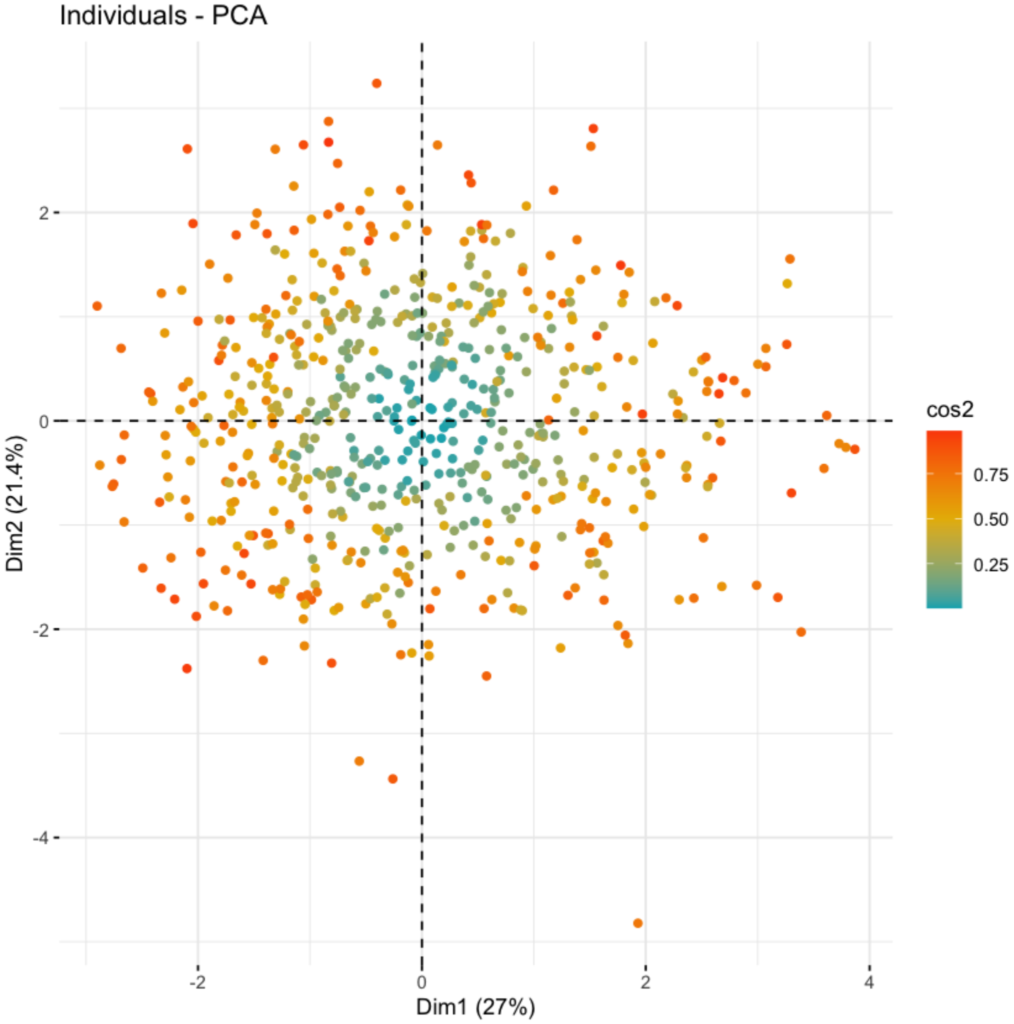


Figure 13: Individual map with 25% sample of individuals

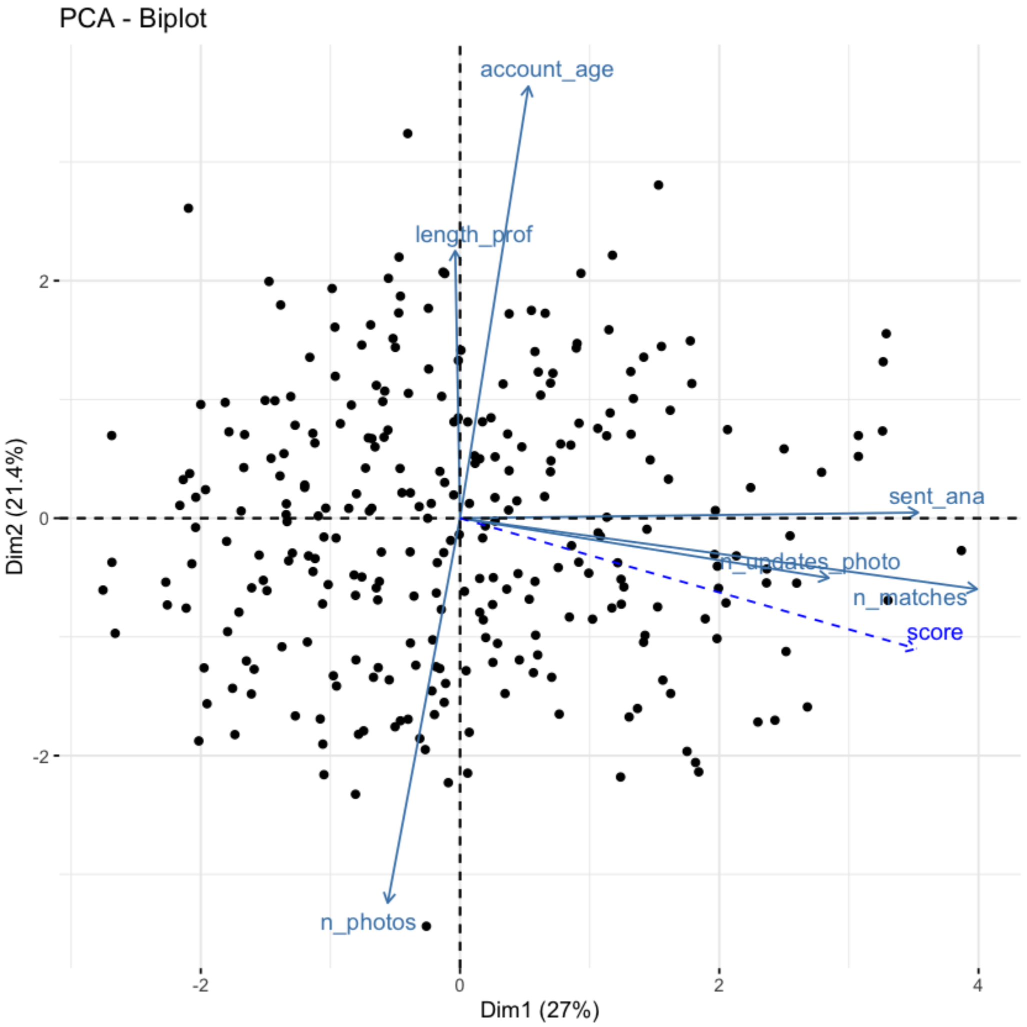


Figure 14: PCA Biplot with 10% sample of individuals