SEGMENTING CLIENTS: OUR ANALYSIS IN PYTHON

Matteo Bollettino
Madhurii Gatto
Lucia Gioria
Vittorio Nardi

The problem: Segmentation Analysis

The aim of this project is to analyse a dataset of bank clients in order to perform some clustering, to group them and extract the so called "Personas".

Personas are prototypes (not stereotypes!) used by banks to group people by their characteristics. This is done in order to make some customed financial advice, i.e. proposing to each individual a financial product that better suits his needs, but also for a lot of other useful applications.

Preprocessing

The first step is to preprocess the dataset. We apply One-Hot Encoding for the categorical variables and Min-Max-Scaling for all the numerical variables.

An overview of the preprocessed dataset:

	Age	FamilySize	Income	Wealth	Debt	FinEdu	ESG	Digital	BankFriend	LifeStyle	...	Job5	Area1	Area2	Area3	CitySize1	CitySi
0	0.065789	0.6	0.679599	0.705895	0.268264	0.770735	0.465122	0.718914	0.581720	0.612604	...	0.0	0.0	1.0	0.0	0.0	
1	0.368421	0.0	0.873299	0.919090	0.747693	0.892883	0.521675	0.986877	0.778748	0.868977	...	0.0	0.0	1.0	0.0	0.0	
2	0.250000	0.2	0.942846	0.902289	0.451701	0.504873	0.640388	0.772055	0.677446	0.761279	...	0.0	1.0	0.0	0.0	0.0	
3	0.631579	0.4	0.548115	0.425051	0.614591	0.512343	0.518146	0.607305	0.648808	0.337033	...	0.0	1.0	0.0	0.0	0.0	
4	0.184211	0.0	0.820609	0.734639	0.851100	0.889625	0.783674	0.730646	0.746853	0.915946	...	0.0	1.0	0.0	0.0	0.0	

Factor Analysis

We want to apply a strategy of dimensionality reduction but also to preserve interpretability: so, our natural choice is to apply Factor Analysis.

This method tries to find a reduced number of variables (factors), that are not directly observed, to explain the covariance structure of the original variables. In this way we group the existing variables by correlations between each other, with high correlation within a group and low correlation between groups. Each obtained group represents a measure of the factors responsible for the correlation, even if they were not directly observed.

Kaiser-Meyer-Olkin value

We begin by calculating the Kaiser-Meyer-Olkin (KMO) value, which represents how much data are suitable for a Factor Analysis and in particular represents the portion of common variance to be attributed to the potential factors.

The KMO value is 0.758: it is perfectly sound, so we proceed with this technique.

Factors

Similarly to PCA, a way to select the number of factors is to consider the ones with associated eigenvalues greater than 1. So, we initially choose to use 10 factors. After having performed the analysis, we notice that only 5 of them are actually relevant, since the other ones are correlated with some of the variables already correlated with the first five ones.

The loadings represent the correlation of each factor with each variable. By inspecting them we can give a real interpretation to the unobserved factors we are looking for.

Interpretation of the factors

	0	1	2	3	4
Age	-0.427474	0.040422	-0.020662	-0.024052	0.026739
FamilySize	-0.037592	-0.270304	-0.040750	0.002816	-0.126172
Income	0.639147	0.093202	0.076556	0.031103	0.006313
Wealth	0.620022	0.068681	0.094022	0.030352	-0.016236
Debt	0.634826	-0.231041	-0.020785	-0.015884	-0.118355
FinEdu	0.697423	0.088816	0.074368	0.019102	-0.011683
ESG	0.193161	0.276338	0.054809	-0.001031	0.147690
Digital	0.711058	0.110494	0.096361	0.027609	-0.002771
BankFriend	0.282857	0.293757	0.121519	0.041556	0.064014
LifeStyle	0.665501	0.133934	0.089048	0.028441	0.019506
Luxury	0.715873	0.192647	0.091299	0.052756	0.038171
Saving	0.503725	-0.186726	-0.022149	-0.014160	-0.061205
Gender1	-0.071087	0.051299	0.001706	0.018013	0.058618
Job1	-0.139827	0.172005	0.023449	0.074675	0.107124
Job2	0.547493	-0.579303	-0.083588	-0.050515	-0.420664
Job3	0.095335	0.132714	-0.001696	0.004840	0.176078
Job4	-0.008057	0.092711	-0.007612	0.004311	0.064389
Job5	-0.554632	0.325328	0.066807	-0.030208	0.173018
Area1	0.186228	0.367855	-0.963322	-0.073253	-0.128232
Area2	-0.149136	-0.115126	0.535155	0.039343	0.128471
Area3	-0.055485	-0.226462	0.302289	0.023042	-0.032007
CitySize1	-0.288092	-0.235177	-0.029061	-0.610571	-0.095066
CitySize2	-0.172045	-0.123405	-0.123862	1.000436	-0.062164
CitySize3	0.477610	0.346393	0.131204	-0.180665	0.143027
Investments1	-0.184130	0.033121	0.078288	0.021886	-0.301518
Investments2	-0.092110	0.370757	0.135019	-0.011229	-0.210796
Investments3	0.290998	-0.540727	-0.296902	-0.015030	0.793636

Interpretation of the factors

- Factor 1 is positively correlated with Financial Aspects and with whether the subject still works (positive with worker and negative with retired).
- Factor 2 is related to the type of Job and the type of Investments (negative with worker, negative with capital accumulation investments, positive with lump sum investments).
- Factor 3 is related to Area of Living (positive with Central/South area, negative with North area).
- Factor 4 is related to the Dimension of the City (positive with medium sized, negative with small sized).
- Factor 5 is related to Investments (positive with capital accumulation investments, negative with no investments and lump sum investments).

K-Means

As a first step, we exploit the classical K-Means as our clustering method, applied on the 5 factors found.

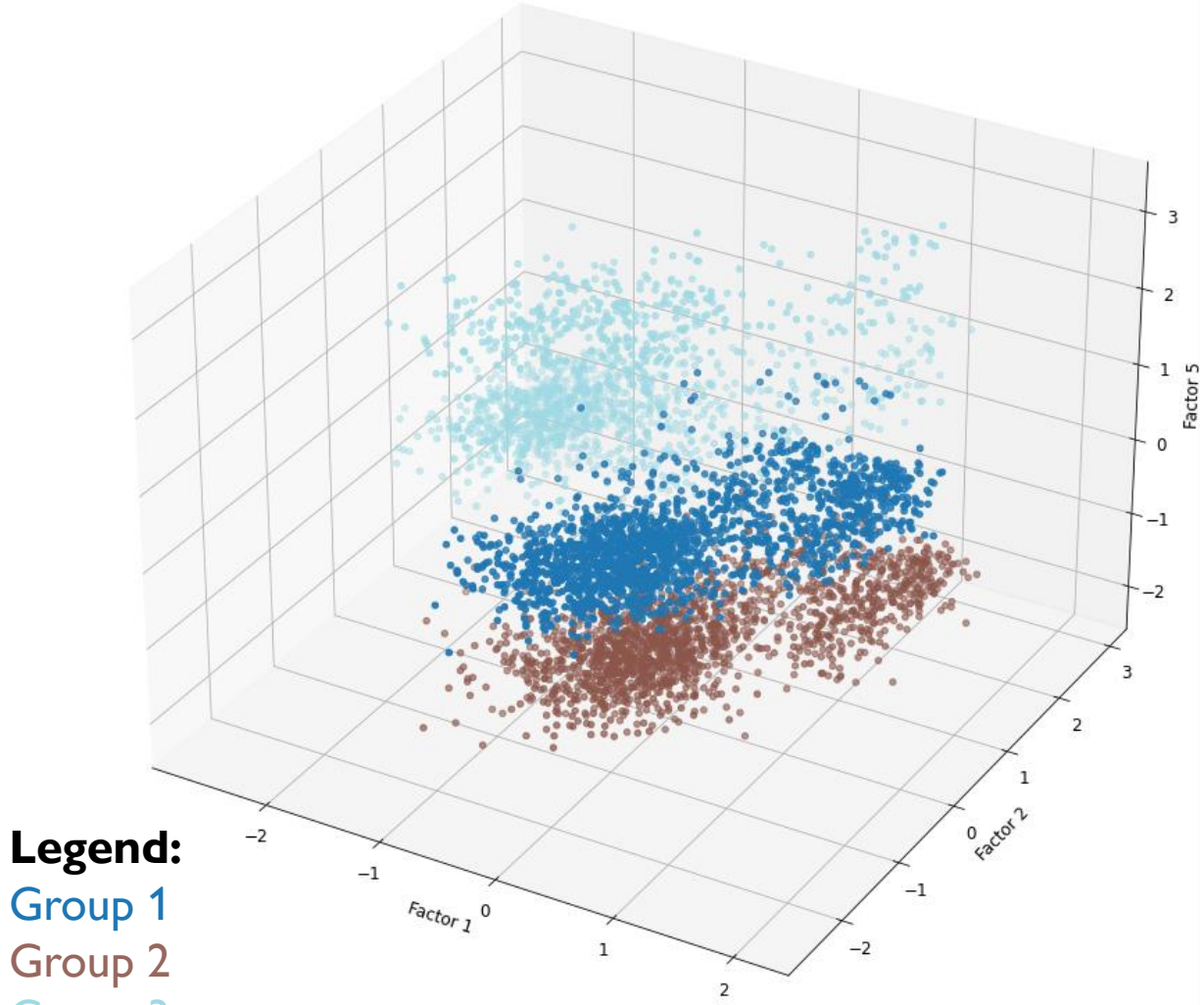
To find the optimal number of clusters we both:

- Search for the local maximum of the average silhouette
- Use a Knee-Elbow Analysis on the Within-Clusters Sum of Squares and the Between-Clusters Sum of Squares.

The average silhouette for K=3 is a local maximum, equal to 0.1717, so we choose to work with 3 clusters.
This choice is also confirmed by the slight elbow in the Knee-Elbow Analysis.

K-Means with 3 clusters

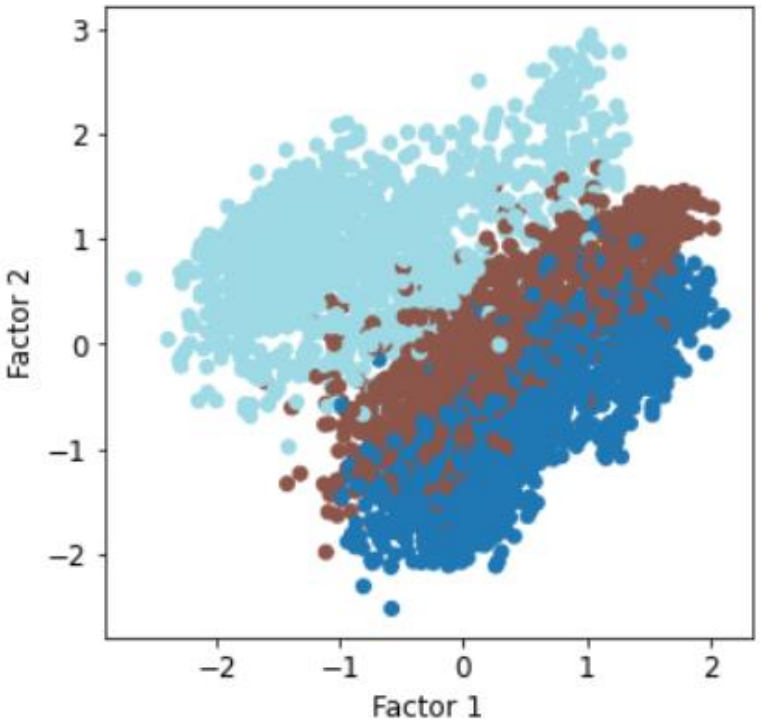
By checking the average values of the factors in the 3 clusters, we notice that the most important for discrimination are 1, 2 and 5. So we plot the clusters in the 3D space defined by these factors.

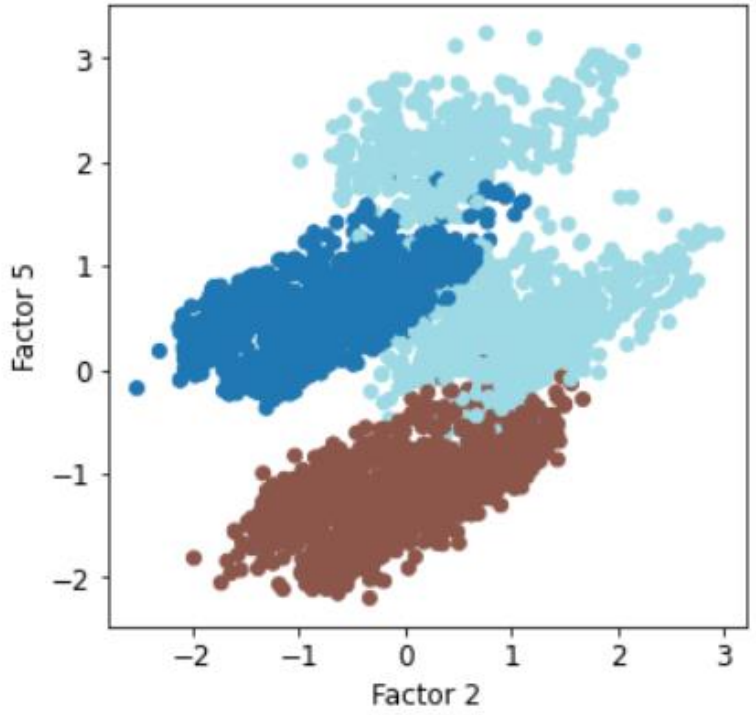


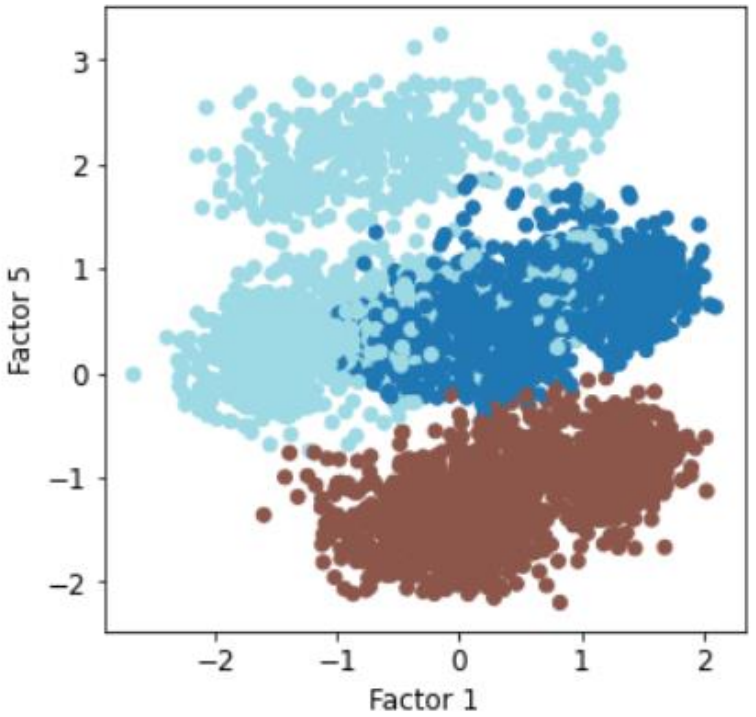
K-Means with 3 clusters

We also plot the clusters in the three 2D projections.

Legend:
Group 1
Group 2
Group 3







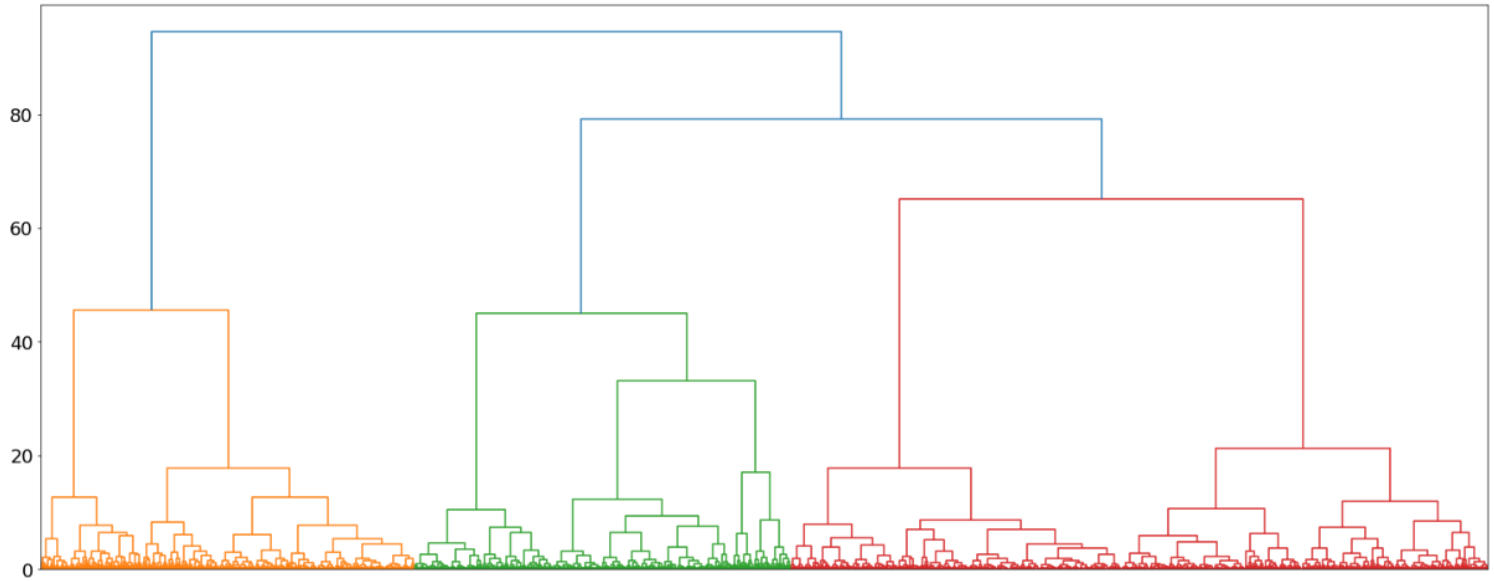
Hierarchical Clustering

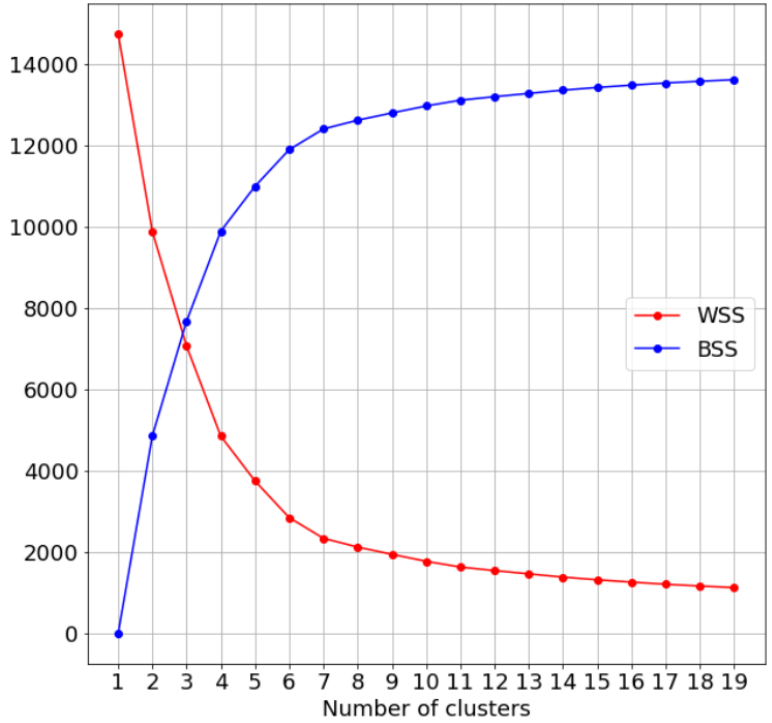
With K-Means we obtain a good segmentation. But can we do better?

We notice from the 2D projections that some clusters may be further separated. Also from a financial point of view, we may want a more specific division of the clients. Hence, we try a different method, Hierarchical Clustering, which gives us more insight on the number of clusters we should use.

Furthermore, in the following analysis, we only use the three most significant factors as coordinates.

Hierarchical Clustering: Dendrogram and Knee-Elbow Analysis

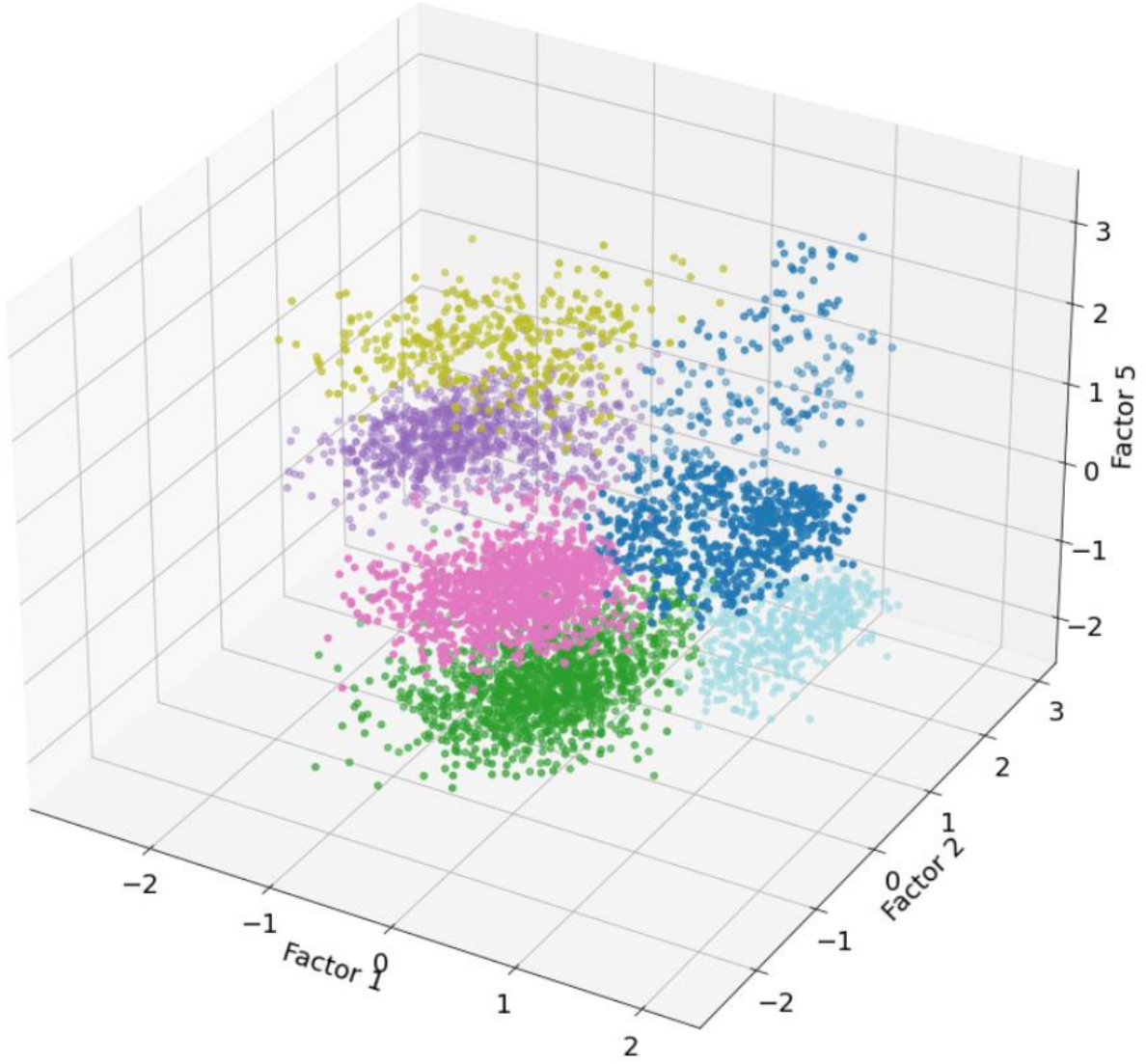




The two plots suggest a division in 6 clusters.

Hierarchical Clustering: 6 clusters

We plot the clusters in the space defined by the three factors.

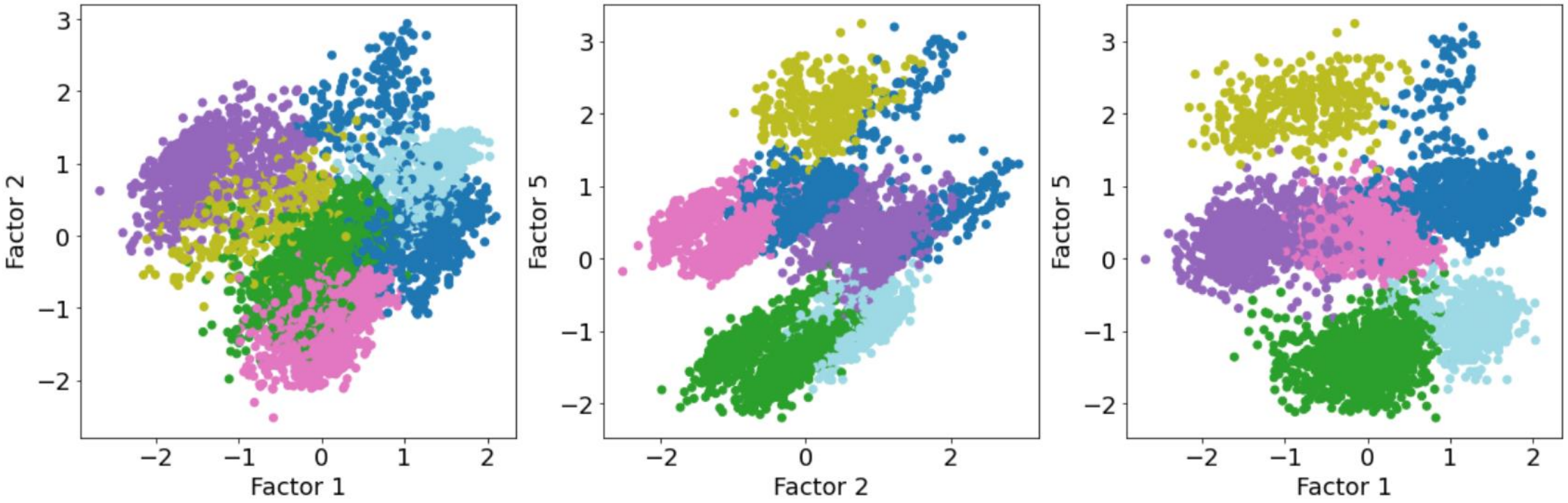


Legend:
Group 1
Group 2
Group 3
Group 4
Group 5
Group 6

Hierarchical Clustering: 6 clusters

We also plot the clusters in the three 2D projections.

Legend:
Group 1 Group 2 Group 3
Group 4 Group 5 Group 6



Interpretation of the obtained clusters

Profile 1: The Young City Caiman

- High score in factor 1 -> They have a high financial interest, they are mostly executives and entrepreneurs.
- High score in factor 5 -> They invest in capital accumulation.

Furthermore, they live in a big city in the North and are younger than average.

Profile 2: The Southern Village Worker

- Low score in factor 5 -> They do not invest or invest lump sums.
- Quite low score in factor 2 -> They are workers.

Furthermore, they have slightly below average financial aspects and they live in small or medium sized cities in the South area.

Interpretation of the obtained clusters

Profile 3: The Old Slacker

- Low score in factor 1 -> They are not interested in financial aspects.
- High score in factor 2 -> They are either unemployed or retired.

Furthermore, they are older than average, they do not invest or invest lump sums, they live in small or medium sized cities in the Central area.

Profile 4: "Average Joe", the Accumulator

- Low score in factor 2 -> They are workers and they prefer capital accumulation investments (confirmed by the positive value of factor 5 and by the above average value of Investments3).

Furthermore, they are spread on all the territory and have slightly below average financial aspects.

Interpretation of the obtained clusters

Profile 5: The Unconventional Elderly

- High score in factor 5 -> They invest in capital accumulation.
- Low score in factor 1 -> They are not interested in financial aspects.

Furthermore, they are either unemployed or retired, they are older than average, they live in medium and small sized cities in the Central area.

Profile 6: Daddy's Boy

- High score in factor 1 -> They have a high financial interest and they have a job, mostly workers but also other positions.
- Low score in factor 5 -> They invest lump sums.

Furthermore, they live in a big city in the North and are younger than average.

Conclusions

To sum up, with our client segmentation we obtained six personas:

- The conventional metropolis youngsters
- The workers
- The conventional unemployed/retired elderly
- The accumulators
- The unconventional unemployed/retired elderly
- The unconventional metropolis youngsters