Motivation Challenge Test

September 2, 2023

Description

In this report we give our insights in a Clustering Project where, using the suitable techniques, we create clusters from a data set containing 49 observations (patients) and 12 features (motivations).
Each observation corresponds to a patient that gives a score ranged between the values -2 to 2 to each of the following motivations:

Power and influence,	Sense of community	Curiosity
Appreciation	Collector	Sense of purpose
Food	Movement	Emotional calm
Order	Eros and Beauty,	Competition

Questions

Question 1: It is possible to group patients according to their similarity in their answers? If so, how many groups would exist.

Question 2: It is possible to group motivations?

Methodology

Wrangling and cleaning data: The data set has been thoroughly examined, and does not contain missing values or duplicates.

Correlation in the context of clustering analysis Upon calculating the Pearson correlation for each pair of features, it became apparent that a strong or moderately strong correlation exists among the following features:

> *Power and influence*, *Competition*, *Eros and Beauty*, *Order*, *Curiosity* and *Food*

Dimensionality Reduction Due to the existent correlation among the features above mentioned, and to achieve an easy visualization of the clusters, we applied two techniques for reducing the number of features:

Principal Components Analysis (PCA): where we used the two first components, that explain 46% of the variance of the original features approximately, and

Multidimensional Scaling (MDS) where we use the correlation distance between each pair of features.

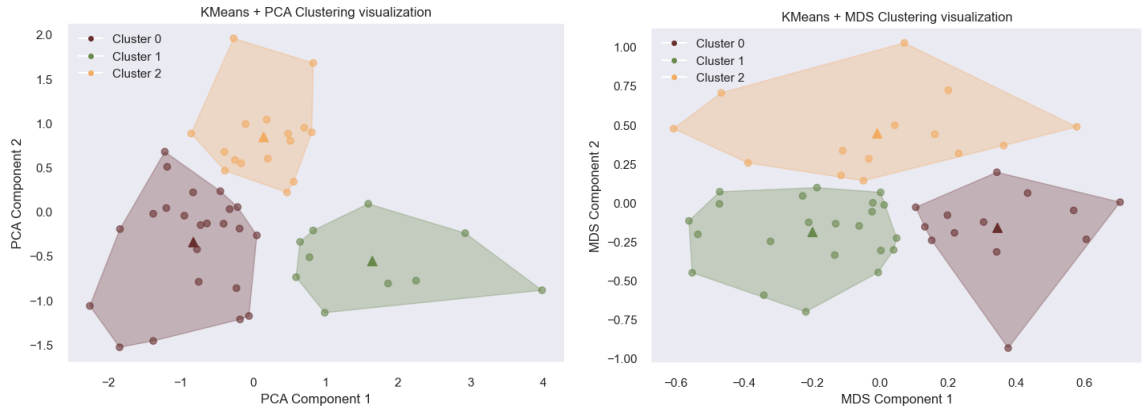
K-Means Clustering

Subsequently, we used the components from PCA and MDS to apply the K-Means clustering technique. We will refer to both variants explored in our project as K-Means + PCA and K-Means + MDS.

Despite its widespread use, K-Means has one drawback: it can not determine the optimal number of clusters for a given data set. We used well-known techniques to choose the number of clusters, such as the *Elbow Method* and the *Silhouette coefficient analysis*, which aim to minimze the within-cluster sum of squares of each cluster, and other statistical-based techniques like the *Gap Statistic calculation* (which contrasts evidence against the null hypothesis). As a result of all the calculations and analysis described above, we concluded to apply K-Means to create **3 clusters**. It is also important to note that the number of clusters is a choice that should also comply with the specific context of the project and the structure of the data set.

Question 1

Figure 1 shows the visualization of the 3 clusters found by K-Means+PCA (1(a)) and K-Means+MDS (1(b)).

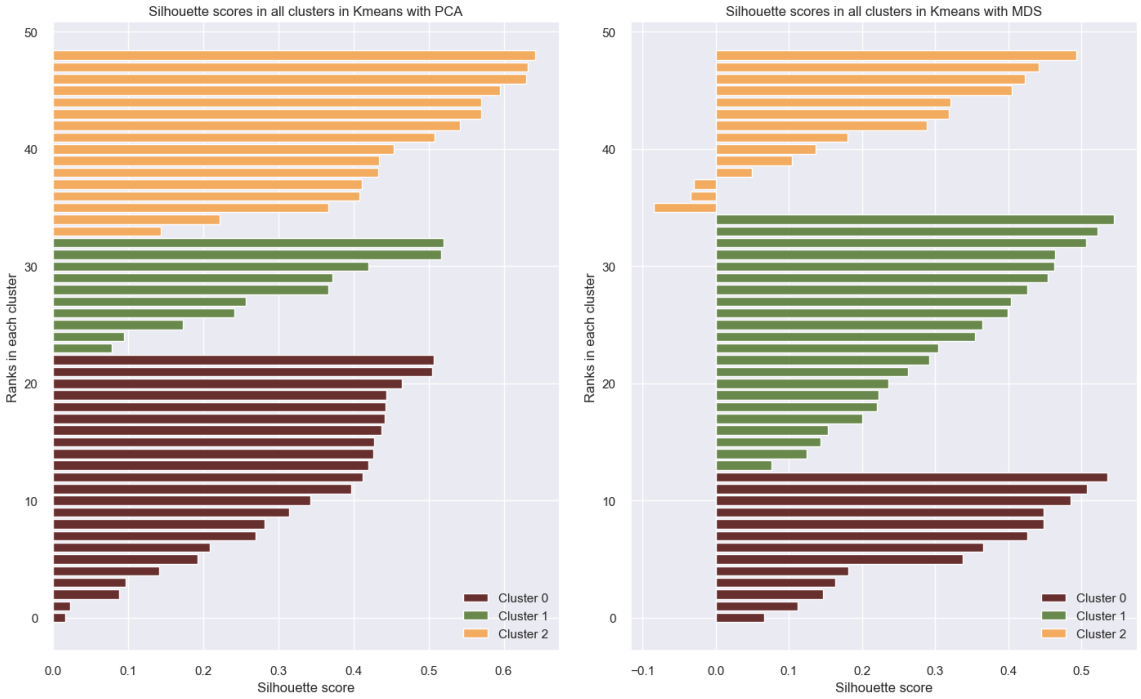


(a) 3 Clusters with K-Means with PCA.

(b) 3 Clusters with K-Means with MDS.

Figure 1: Clusters Visualization using K-Means

We calculated the silhouette coefficient to choose which of the two K-Means variants performs a better cluster assignment. Figure 2 shows the silhouette coefficient of all data points for each cluster when applying K-Means + PCA (2(a)) and K-Means + MDS (2(b)). Most of the silhouette coefficients values are relatively acceptable for both



(a) K-Means + PCA

(b) K-Means + MDS

Figure 2: Silhouette scores visualization using K-Means

variants. However, a notable finding is that three data points from Cluster 2 have negative silhouette coefficient values when using K-Means + MDS. Additionally, in both variants, every cluster contains some data points with silhouette coefficient values close to zero, indicating that they are located between clusters. As a result, it is conceivable that some clusters fail to keep a high enough level of isolation from one another.
The above stated is confirmed when we calculate the average silhouette coefficient to combine the cluster quality in a single number:

$$\begin{aligned} \text{silhouette score in K-Means + PCA} &= 0.366 \\ \text{silhouette score in K-Means + MDS} &= 0.293 \end{aligned}$$

Therefore, the variant **K-Means + PCA** provides the best (although not so good) cluster quality according to the quality metric silhouette score.

While it remains feasible to group patients according to how they respond to questions, the small sample size inevitably leads to a level of accuracy in the results that falls short of the desired standard, so the findings must be treated with caution. With the intention to improve the quality our results, we investigated other strategies including:

- **Different Clustering algorithms**: such as Agglomerative Hierarchical Clustering and DBSCAN
- **Tune K-Means parameters**: adjusting parameters like centroid initialization method or number of iterations

- **Increase the number of clusters**: increasing the number of clusters based on other techniques explored like the calculation of the Davies Bouldin and Calinski Harabasz scores.

Question 2

It is possible to classify the motivations, after all. In order to examine any potential linear relationships between the motivations, our initial strategy involved computing the correlation coefficient for each pair of motivations.

From figure 3, we observed a strong or moderate-strong linear relation between the

Figure 3: Correlation Heatmap

motivations:

Power and influence,	Competition,	Eros and Beauty,
Order,	Curiosity,	Food

Motivations grouping technique

To group the motivations we used what is called *Feature clustering* that is an unsupervised Machine Learning technique used to group similar features together based on their similarity or correlation. The idea is to identify groups of features that exhibit similar patterns or have high correlations.

We used **Agglomerative Hierarchical Clustering** as the technique to group the motivations. Figure 4 shows the dendrogram resultant from applying the linkage method **ward** and distance metric **correlation**:

We used a dendrogram to represent the hierarchical structure of patients motivations

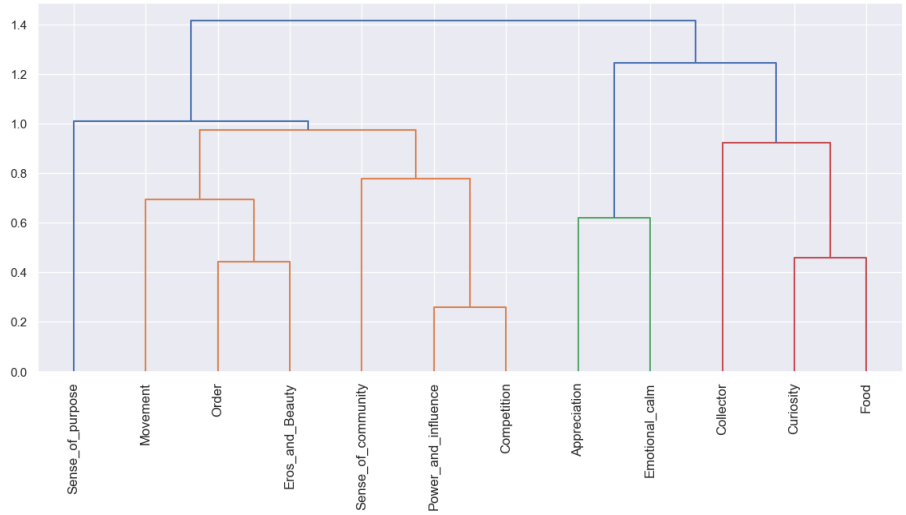


Figure 4: Dendrogram with motivations clusters

(Figure 4). All motivations were successfully grouped into four distinct clusters. We can see a relationship between the motivation "Sense of Purpose" and the group of motivations including "Movement, Order, Eros and Beauty, Sense of Community, Power and Influence, and Competition," which may show that people who value a sense of purpose also have stronger tendencies in these other areas. The next cluster includes "Appreciation" and "Emotional Calm" together, suggesting a possible connection between these two motivations. The final cluster of the motivations "Collector," "Curiosity," and "Food" suggests that these motivations may have similar underlying traits or tendencies. In general, this dendrogram offers a useful framework to represent the motivational interactions of patients. Therefore, based on these interactions, we created the following groups of motivations:

$$\textbf{Group 1} = \left\{ \begin{array}{lll} \text{Power and influence,} & \text{Competition,} & \text{Sense of community,} \\ \text{Eros and Beauty,} & \text{Order,} & \text{Movement} \end{array} \right.$$

$$\textbf{Group 2} = \left\{ \text{Collector,} \quad \text{Curiosity,} \quad \text{Food} \right.$$

$$\textbf{Group 3} = \left\{ \text{Appreciation,} \quad \text{Emotional calm,} \right.$$

$$\textbf{Group 4} = \left\{ \text{Sense of purpose,} \right.$$

Clusters Evaluation

We examined all motivations across the clusters found by K-Means + PCA. These clusters offer insightful information about how members of each group prioritize and perceive each motivation. In figure 5 we summarized the distribution of all motivation for each cluster. For instance, "Power and Influence" motivation (Figure 5) for patients in cluster 0 shows a compact box with its interquartile range (25% to 75%) almost falling between 0 and 1. This suggests a relatively consistent level of "Power and Influence" motivation among its members. Contrarily, Cluster 1 exhibits a wider box, with an interquartile range that

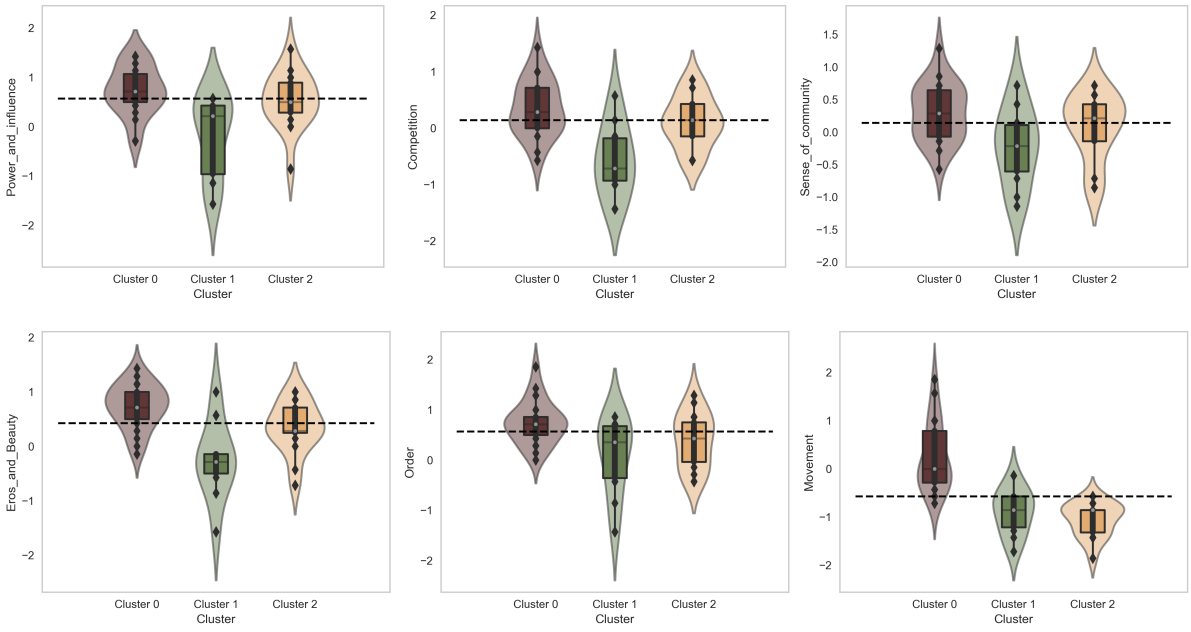
nearly ranges from -0.5 to -1, indicating a larger range in how people within this cluster visualize and valuate this motivation. Even though Cluster 2 is also tightly packed, its interquartile range is between 0.5 and 1.5, indicating a more narrowly focused set of responses and a higher overall level of "Power and Influence" motivation. Additionally, when we look at the overall medians, we see that for Cluster 0, the median is roughly in the middle of the line, suggesting a balanced perspective on this motivation. The median of Cluster 1 is below the line, indicating a tendency toward lower levels of "Power and Influence" motivation. On the other hand, Cluster 2 has a median that is above the line, indicating a greater overall tendency towards this motivation. These results provide crucial information for modifying interventions or strategies based on a person's cluster, enabling more effective participation and communication in a variety of settings.

Conclusion

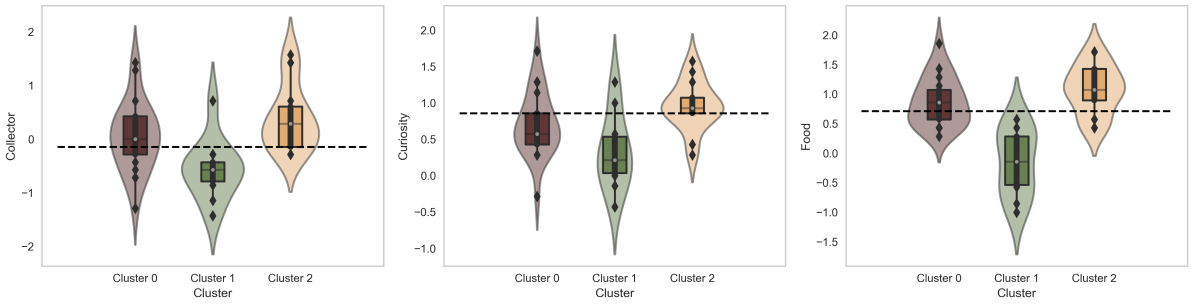
We used clustering techniques to identify trends and connections between 49 patients based on their answers to 12 motivational questions. We tried to answer two fundamental questions: whether patients could be grouped based on how similarly they answered questions, and whether motivations themselves could be clustered. We then looked at the relationships between motivations and discovered some relationships between some of them. The results showed that K-Means + PCA produced the best clustering outcomes. We came to the conclusion that 3 clusters were the best option for our data set, despite the drawbacks of the small sample size. With cluster 0 containing 23 patients, cluster 1e containing 10, and cluster 2 containing 16 patients (Table 1 in the appendix).

Additionally, the dendrogram analysis of motivations reveals the hierarchical relationships between motivational responses. The prioritization and perception of these motivations were further clarified by examining the motivations at each cluster, providing helpful advice for developing interventions and strategies that are tailored to particular patient groups.

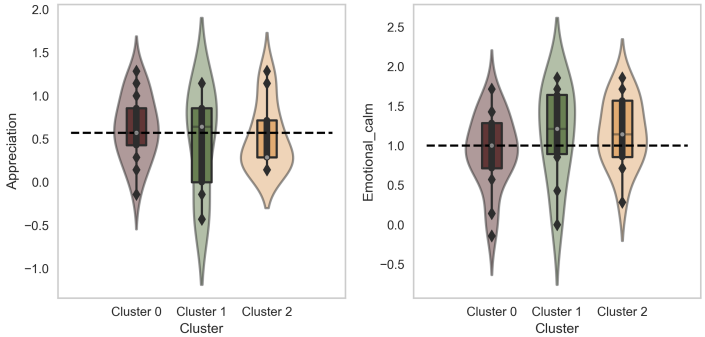
In conclusion, according to the patients' responses to various motivations, the clustering project successfully divided the patients into three different clusters. Although the small sample size may reduce the accuracy of the findings, the careful selection of clustering methodologies and quality metrics has enabled us to make defensible choices regarding the clustering strategy. The motivational hierarchy revealed possible connections between various motivational factors (Figure 4). Additionally, by looking at the motivations within each cluster, we can gain a deeper understanding of how patients within these groups prioritize and view their motivations, which provides useful information for tailored interventions and tactics. As figure 5 shows, some of the motivational responses are quite similar across clusters, but others are more different from each other. With more comprehensive data, we can identify, in more detail, motivational response changes across clusters, illustrate a more clear relationship between patient responses, and perhaps identify a clear pattern of how these motivations can influence patient groups. Basically, with more comprehensive data and a larger sample size, we could conduct a more in-depth analysis, uncovering additional insights into patients' motivational behavior under various circumstances.



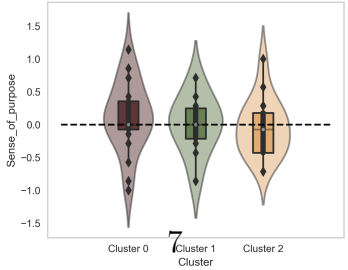
(a) Group 1



(b) Group 2



(c) Group 3



(d) Group 4

Figure 5: Distribution of motivations for each cluster

Appendix

Table 1: Cluster label for each patient ID

Patient_ID	Labels
P1	1
P2	2
P3	0
P4	0
P5	2
P6	1
P7	2
P8	0
P9	0
P10	2
P11	0
P12	1
P13	2
P14	0
P15	1
P16	2
P17	2
P18	2
P19	2
P20	0
P21	2
P22	2
P23	1
P24	0
P25	0
P26	1
P27	0
P28	0
P29	0
P30	0
P31	2
P32	1
P33	1
P34	1
P35	2
P36	0
P37	2
P38	0
P39	2
P40	0
P41	0
P42	0
P43	0
P44	1
P45	2
P46	0
P47	0
P48	0
P49	0