MSc. Data Science and Advanced Analytics 2020/2021

PVA DONORS

A Segmentation Project

DATA MINING 2020/2021

Professors: Fernando Bação | David Silva | João Fonseca

Authors:

Ana Paulino r20170743 | Soraia Cunha r20170806

Index

Table of Figures

Abstract

The following report exposes a segmentation model that aims to make clusters and understand the behaviour of the donors of the Paralyzed Veterans Association. To perform this study were analyzed 475 variables for 95412 individuals. Several processes were applied, including outlier's detection, data normalization, feature selection and the election of the best segmentation model from K-means, K-Prototypes and T-SNE. As a result, there were obtained four clusters, where for each it was designed a marketing campaign.

Keywords: Segmentation, Clusters, PVA, K-means, K-Prototypes, T-SNE, Feature Selection, PCA

1. Introduction

Paralyzed Veterans Association (PVA) is a non-profit organization that helps US veterans that during their missions have suffered from spinal cord injuries and disease. In this way, it was collected information about 95412 donors, so that could be performed a segmentation analysis to understand their behaviour and identify possible segments of the individuals.

PVA was founded by soldiers that fought in World War II, who wanted to contribute to their community by helping the other comrades that also suffered from spinal cord injuries. Their mission is to empower their gallant men and women to reconquer what they fought for, their independence and freedom.

In this research, we pretend to analyse the donors' dataset to explore and retrieve some insights about how we can make a partition regarding our donors' behaviour and other characteristics provided. Therefore, what we intend to know is "How we can divide the donors by clusters?" and "What will be the variables used to do this segmentation?". Also, we will elaborate on a marketing campaign for each segment found.

GitHub: https://github.com/SoraiaCunha/DM_PVA_PROJECT.git

2. Explore

Firstly, before analyzing the dataset, we must precise the business problem and perform an exploration of the data to understand how the dataset is composed, by how many variables, their data types, and other explorations. Moreover, we also used pandas_profilling to help us explore the variables and their distribution. Thus, this research aims to perform segmentation on the donors based on the collected data.

Next, we start the project itself by loading the Donors dataset and having a first look into the dataset, where we can state that we have 95412 rows (individuals) and 473 columns that correspond to the variables used to collect information. Then, using the method ".describe(include='all') ", we can get an overview of the dataset and see the count, frequency, mean, standard deviation and the quartiles. In here, we can realize two information's, the first is that if in a variable the difference between the value of the 75% quartile and the maximum value is big, we possibly have outliers in that variables, for example, "TCODE" has a value equal to 2 in the third quartile, but the max value is 72002. Also, what variables are categorical or not, based if they present information on the mean, standard deviation and in the quartiles. In the ".info()" method, we check that we have three data types in the dataset: float64, int64, object.

Afterwards, the next step is to check the missing and duplicated values. Considering, the duplicated values, we could observe that were not any rows repeated, and by examining the missing values, we detect that we have a lot of rows with missing data. In that way, the path we choose to deal with these missing values was first to eliminate columns with 50% or more cells with a null value, since these columns will not bring any relevant data and by filling the data it can lead to bias; secondly, we check the variables, where the cell is empty has a meaning, so we replaced by a letter, to know, that the cell is not a missing value. These modifications were made in the following variables: "GENDER" (we replace ' ', 'A', 'C' by 'U'), "HOMEOWNR" (' ' by 'U'), "MAILCODE"(' ' by 'O'), "CHILD03", "CHILD07", "CHILD12", "CHILD18" (np.nan by 'N'), "SOLP3" (np.nan by 'Default'), "SOLIH" (np.nan by 'Default'), "MAJOR" (np.nan by 'N') and "PEPSTRFL" (np.nan, 'N'). Thirdly, the other columns with empty cells without meaning, we replaced by np.nan and we used them as well as the columns already with NaN´s (Not a Number) to fill the missing values. For the non_metric_features (variables where the data type is an object) it was substituted by the mode and for the metric_features (variables where the data type is not object) it was replaced by the median since is not influenced by the outliers.

3. Modify

In this step, we inspect the correlation of the variables divided into categories, so that we could drop the most correlated variables (|Spearman correlation| > 0.90). We use Spearman because we cannot assume that the variables are linear. Then, transforming the variables' datatypes, followed by the creation of new variables. After that, we examined the incoherent records and then we checked again the correlation of the variables including the new ones. Thereafter, we scan the outliers using 3 criteria and then perceive the relevance of the variables in order to proceed with the analysis to the clustering phase.

3.1. Correlation of variables divided into categories

First, we explore data in order to drop some of the most correlated variables divided into 23 categories and then identify outliers. This was done since we had more than 400 variables in the original dataset, which was computationally expensive and time-consuming. In this section, we will focus on the categories and the results obtained regarding the corresponding dropping of the variables (the heatmaps with the correlations are shown in the Jupyter Notebook since there are many heatmaps to show in this report).

Subset of Variables	**Dropped Variables - |Spearman correlation| > 0.90**
Donor info	No correlation above the threshold was found, meaning that in these categories no variable was dropped.
Others	
Neighborhood Marital Status	
Neighbourhood Jobs	
Neighborhood Employment Sector	
Neighborhood Years Education	
Neighborhood Ancestry	
Promotion History File Summary	NUMPROM and CARDPROM have a correlation of 0.97; however, we opted not to drop any of them, since they are other variables depending on them.
Donor interests	Only composed by categorical variables, we did not drop any variable belonging to these subsets
Promotion History File	
Promotion RFA	
Well-being & House changes	MC1
Neighborhood Other	HC5 and HC7
Neighbourhood Population	POP902, POP903, ETHC5, POP90C5
Neighbourhood Population Age	AGE901, AGE902, AGE903, AGE905, AGE906, HHAGE1, HHAGE3 AC1 and AC2 were eliminated since they can be represented by AGEC5 and AGEC6
Households Number of People	HHN4, HHN5, HHN6, HHP1, HHD2, HHD6, HHD7
Neighborhood Housing Units	DW2, DW5, HU2, HU4

Neighborhood House Value	HV1, HV3, HVP1, HVP2, HVP3, HVP4, RP2, RP3, HC18
Neighborhood House Units – Rooms	RHP1, HUPA1, HUPA2, HUPA3
Neighborhood Household Income	IC1, IC2, IC4, IC15, IC18, IC19, IC20, IC21, IC22, IC23
Neighbourhood use of transports/time	TPE4
Neighbourhood Employment	LFC1, LFC2, LFC3
Neighborhood Military Service	AFC4

Table 1 - Groups of correlation and the dropped variables

3.1. Transforming Datatypes

In this step, we convert the variables with dates to 'datetime64' datatype, as it was the case of 'ODATEDW', 'DOB' and 'LASTDATE'.

3.2. Creation of New Variables

Intending to improve our analysis, we decided to create 4 variables from the original ones. In the table below it is shown the description and formula for each new variable.

New Variable	Formula
Age – the age of the donor at the time of the 2017 promotion was mailed	(pd.DatetimeIndex(donors['ADATE_2']).year - pd.DatetimeIndex(donors['DOB']).year)
Recency – days since last recent gift	(donors['ADATE_2'] - donors['LASTDATE']).dt.days.astype('int64')
Days_LastPromotion - days since most recent promotion received	(donors['ADATE_2'] - donors['MAXADATE']).dt.days.astype('int64')
Days_FirstGift - days since the first gift	(donors['ADATE_2'] - donors['ODATEDW']).dt.days.astype('int64')

Table 2 - New variables identification and formula

3.3. Incoherencies

We have checked the records regarding 2 possible incoherencies.

1) The first one, to check if there were donors younger than 17 years old

2) The second to grant that we only have records where the Recency is above zero.

In total, there were eliminated 1007 incoherent records, all correspondent to records where the donor is younger than 17 years old.

Regarding the Lapsed donors, we considered that not all donors from the database where Lapsed, maintaining all the individuals, independently if Recency is not in the interval between the 13 and 24 months.

3.4. Spearman correlation above 0.90 – Global drop of variables

In this step we decided to check again the Spearman correlation on all dataset, imposing a threshold of removal the pairs with |correlation| > 0.90. We end by eliminating 15 variables in this phase.

3.5. Outliers Detection

3.5.1. Inter-Quartile Range (IQR)

To detect outliers in our data, we decided to test this manual check, IQR; however, we have faced some limitations at the beginning due to a large amount of data. More specifically, using the usual range for IQR, the 25th and 75th percentile we ended by detecting as outliers a higher percentage of our data points, which would not be reasonable since it is recommended to eliminate between 3% up to 5% of outliers.

In this way, we decided to change the percentiles so that we keep a tolerable percentage of records in our analysis, being the selected ones the 3rd and 97th percentiles. Nevertheless, applying the previous values, it was not enough since it would still eliminate around 26.78% of records. Thereby, this outlier detection method was not the one selected in our analysis.

3.5.2. Local Outlier Factor (LOF)

The Local Outlier Factor (LOF) algorithm is an unsupervised anomaly detection method created with the goal of computing the local density deviation of a given data point about its neighbours. The algorithm is designed to recognize as outliers the samples that have a significantly lower density than their neighbours.

After the results we got by applying the IQR method, we decided to try the Local Outlier Factor, because this one measures the local deviation of a sample regarding its neighbours (k-nearest neighbours) in terms of density. The samples that show a significant lower density than their neighbours are considered outliers.

Hereupon, we have considered the 50 nearest neighbours and left the rest of the parameters as default and we detect around 1.90% of outliers, which is acceptable.

3.5.3. Z-score

In addition, we thought that we should also test another method of outlier detection since we have a large and disperse dataset. In this way, we applied Z-score that considers as outliers the observations that are the stated standard deviations above the mean value of what is being observed. Having in mind the shape and sparseness of our dataset, we declared as the threshold 15 standard deviations.

As a result of the Z-score, approximately 2.50% of records were considered as outliers.

3.5.4. LOF and Z-score

According to the results obtained in LOF and Z-score we elected both methods to be part of our outlier detection method, so we merged them and around 4.40% of records were dropped after being considered as outliers. Considering the dataset, we

have, we believe that this percentage of outliers will not be problematic, and it is following the rule of thumb of 3%-5% for outliers' removal.

3.5.5. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Besides the previous methods of outlier detection, we decided to try DBSCAN in our data. After trying different values as hyper-parameters, we chose an epsilon of 9.7 and 826 as min_samples, but it was not still a good way of detecting outliers in our dataset. The reason beyond that might be due to the large shape of our dataset, also our clusters might not have a similar density, which is the core of DBSCAN's good performance.

3.6. Data Standardization

As the metric features were in different scales and in order to be used in the clustering phase, we needed to standardize them. Therefore, we used Standard Scaler to standardize the features scaling them to zero mean and one standard deviation.

3.7. Feature Selection – Relevance of Variables

Focusing on the shape of our dataset as it is so large, before proceeding to the clustering phase we opted to check the relevance of our numeric variables regarding dispersion ration and mean absolute difference so that we can continue our analysis with the most relevant features. Unfortunately, for the Non-Metric features, we did not check their relevance since the methods that we have seen for these types of variables involved a target variable which is not applicable in our case since we are dealing with an unsupervised learning problem.

3.7.1. Dispersion Ratio and Mean Absolute Difference

Computing the Dispersion Ratio, that is the ration between the arithmetic mean and the geometric mean, we can get the most relevant features according to their dispersion, a higher dispersion is linked to a more relevant variable. So, we decided to apply this method in our data, and we got many variables with high dispersion.

Then, we also applied the Mean Absolute Difference (MAD) to our dataset, more specifically we computed the mean absolute difference from the mean for our metric features. In this way, the variables that have shown a high MAD are the ones that have a more discriminative power, meaning that they are the most relevant to our analysis.

Merging the two previous methods – Dispersion Ratio and MAD – we decided to eliminate from the study the variables that had the smallest values in MAD until the biggest jump and the ones that had a Dispersion Ratio equal to zero.

4. Data Redundancy (PCA)

After performing the feature selection, we ended up with 311 variables. So, to reduce the input space, we applied Principal Components Analysis, a dimensional

reduction technique. Thus, we selected 47 principal components following the Kaiser approach, where to find the number of principal components, we must retrieve all that have an eigenvalue higher than 1. Then, we apply a rule to the loadings where we would highlight a green loading with a value superior to 0.4 and a red the ones with a value inferior to -0.4, to help to define the factors. In the end, we discarded this possibility since the understanding of the factors is very subjective and we could not find a proper designation for each principal component. Besides that, the goal of the research is to do a segmentation so we would be adding an unnecessary level of difficulty to the clusters' interpretability.

5. Clustering

5.1. K-Means (All Metric Features)

K-means clustering algorithm focus on grouping samples, where each observation belongs to the cluster with the closest mean. The objective of K-means is to minimize the inertia, the within-cluster sum of squares in the sense of minimizing the intra-cluster distances and maximizing the inter-cluster distance. Besides, one particularity of this clustering algorithm is that it only works for numeric variables; however, we decided to apply this algorithm in our analysis because it is one of the most popular clustering algorithms that work well in large datasets, being faster, computationally talking, than hierarchical clustering. Furthermore, K-means is a simple algorithm to implement, is only required to calculate the centroids and compute distances.

In this sense, in our first analysis, we decided to see how K-means would work for the all set of metric variables, testing it accordingly. To achieve this, in our first trial we decided to plot the inertia for the range of one to fifteen clusters with init='K-means++' since it is a smart initialization process that speeds up convergence. After plotting the inertia plot over clusters, we decided to select 3 clusters.

In this way and based exclusively on the means of each cluster regarding the metric features (Figure 1 below), we can see that Cluster 0 clearly stands out with its low mean for the variables MARR1, VOC2, LFC7, TPE1, HUR2, AGEC4, CHILC4, ETHC1 and CHIL2. When it concerns to MARR2, MARR4, DW3, DW4, DW7, DW8, Cluster 0 behaves oppositely, standing out with its high mean. In addition, Cluster 1 stands out from the others for EC3, OCC11, IC16, IC6, IC7 with its low mean and for HV4, HUR2, IC3, IC11, HHAS3, OCC2, EC1, EC7 and MHUC1 with its high mean. Finally, Cluster 3 shows the most consistent behaviour around the mean; however, for the variables POP90C1, HV4, EC7, SEC5, POBC1 shows the lowest mean and for POP90C3, OCC10, EC4 and POBC2 shows the highest mean.

Based on this, we can denote that **Cluster 0 (Single Young Adults)** is composed of young adult donors that are single or divorced, who share a home and do not have their vehicle. Focusing on **Cluster 1 (Adults in a good financial situation)** we can say that is constituted by adults in managerial positions with a high household income and higher

education. Lastly, **Cluster 2 (Village people)** englobes people who were born in rural areas and have stayed there working on crafts.

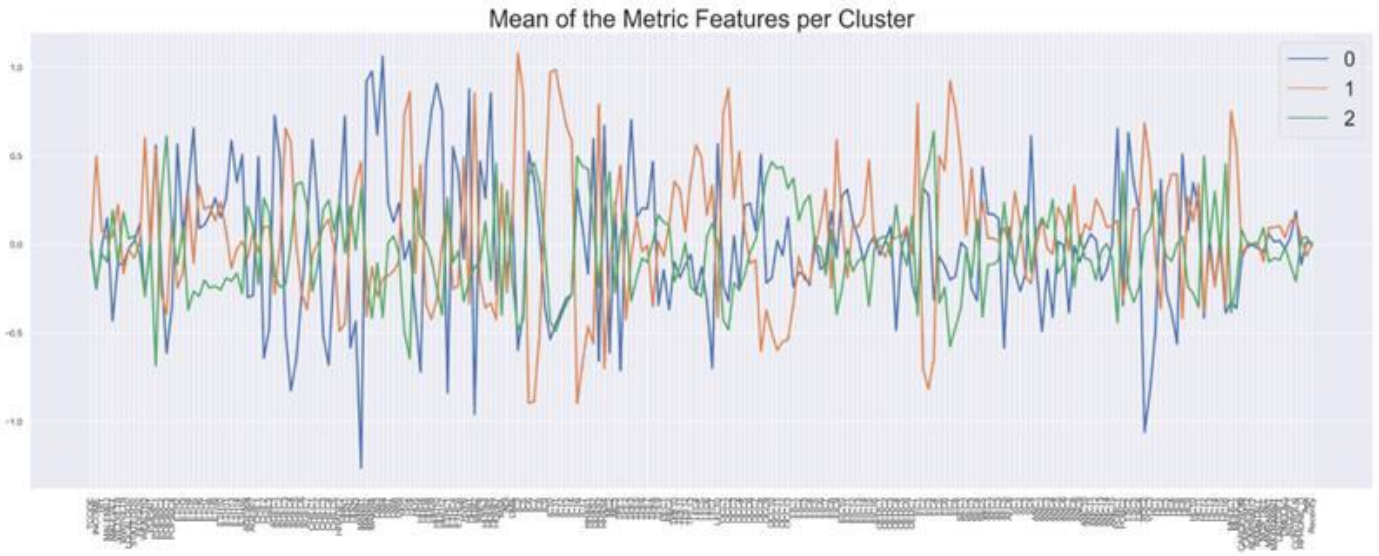

Figure 1 - Mean of Metric Features per Cluster (All features)

5.2. K-Means (Different Perspectives)

We decided to implement additionally K-means for different perspectives. In this way, as our data is almost divided into donors' data and donors' neighbourhood data, we have created those 2 perspectives – 1) donors features and 2) neighbourhood features. For each perspective, we selected the most relevant features and then we applied k-means algorithm, plotting the R-square for the different clusters in both perspectives. Thus, we considered that the best number of clusters would be 3 for both perspectives.

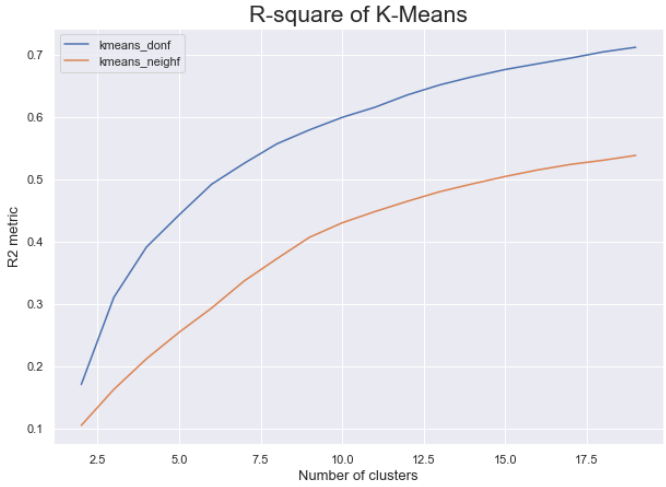

Figure 2 - R-squared for the number of clusters for the 2 Perspectives

Moreover, concerning the Donors' perspective, we can state that from a brief analysis (mean and distribution of the variables), based on the variables we have chosen, the 3 clusters show a similar behaviour regarding the variables Age, AVGGIFT and Income; however, when it comes to Recency, it is clear that Cluster 0 and 1 show values around the mean, but Cluster 2 is situated below the mean, which means that englobes more recent donors. Regarding NUMPRM_12 Cluster 2 is again the outstanding cluster situated above the mean, which suggests that such donors have received more promotions in the last 12 months. Concerning HPHONE_D (published home phone number) the 3 clusters show a different behaviour, being Cluster 0 above the mean, Cluster 1 bellow the mean and Cluster 2 around the mean.

When it concerns to Neighborhood perspective, regarding some variables the 3 clusters show similar behaviour, but in others they are different. More specifically, in EC1 and IC3, median years of school completed in the neighbourhood and average household income, respectively, Cluster 0 shows values above the mean, when Cluster 2 is the opposite and Cluster 1 shows values around the mean. So, Cluster 0 is composed of more literate people. Then, regarding EIC14 and EIC3, the percentage of people employed in Educational Services and Construction, respectively, Cluster 0 is bellowing the mean and Cluster 2 above the mean, and again Cluster 1 is around the mean value. Also, regarding the percentage of the population in the urbanized area there, some discrepancies since both Cluster 0 and 1 show values above the meanwhile Cluster 2 show values below the mean. Therefore, we can briefly say that Cluster 0 is composed by more literate people in a more stable financial condition, Cluster 1 is constituted by people in the average class and Cluster 2 englobes people in a low financial condition, living in rural areas.

5.2.1. Joining the 2 Perspectives

After analysing each perspective individually, we decided to join both in order to have a more meaningful definition of what represents each cluster. In this sense, we applied Hierarchical Clustering to reduce the number of clusters. Concretely, initially, we had 9 clusters (3x3), and after Hierarchical Clustering, we ended with 6 clusters with the help of the dendrogram. The dendrogram was useful because it has shown that 3 clusters (2, 8 and 5 in the image below), that had not many observations, were quite different from the remaining ones and, therefore, they could not be joint with the others.

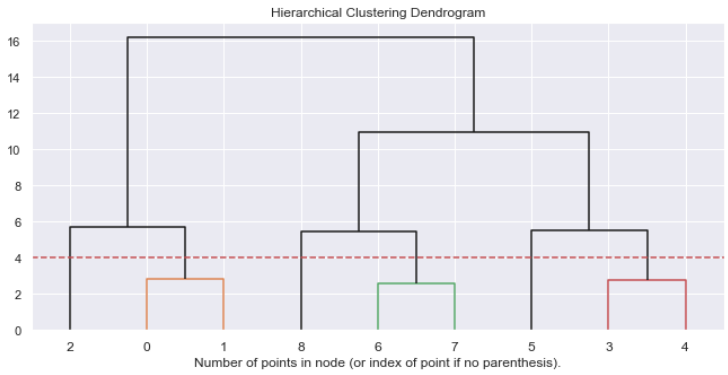

Figure 3 - Dendrogram k-means Joint perspectives

Concerning the composition of each cluster, Cluster 3, 4 and 5 are composed by older donors, the same clusters show as well that their donors have received more promotions in the last 12 months than in the other clusters and that their donors gave a gift more recently. Additionally, the donors of Cluster 3 spend more money per gift whilst the donors of Cluster 1 are the ones who spend the least. Concerning household income, Clusters 0 and 3 are the ones with higher values and Cluster 5 with the lowest. Moreover, Clusters 1 and 5 are composed of donors that live outside urbanized areas and that show neighbours with lower average household income in contrast with Clusters 0 and 5 that have the highest values. Ultimately, Clusters 0 and 3 are composed of more literate people while Clusters 1 and 5 are constituted by adults with less median completed school years. Under this, the 5 clusters can be designated as: **Cluster 0 -** More Literate and wealthy donors, **Cluster 1 -** Needy Rural donors, **Cluster 2 -** Middle-Class donors, **Cluster 3 –** Recent Literate and wealthy donors, **Cluster 4 –** Younger recent donors and **Cluster 5 –** Recent rural donors.

5.3. Silhouette Analysis for K-Means

We performed the silhouette analysis, to both k-means approaches and we get scores close to zero suggesting that the sample is on or very close to the decision boundary between two nearby clusters. Also, we have negative values implying that those samples might have been assigned to the wrong cluster. The graphics are on the appendix.

5.4. K-Prototypes

K-Prototypes is a clustering algorithm that offers an advantage comparing to K-Means of working with numerical and categorical data types since this one only performs using metric data. The algorithm calculates the distance between numerical features using Euclidean distance as K-means, but despite that, it also measures the distance between categorical features using the number of matching categories. It is an algorithm that joins the K-Means and the K-Modes. The types of K-Prototypes implemented to do the initialization are “HUANG”, “CAO” and “Random”. We decided to use the first one since we have a lot of variables and doing initialization with density might not be the best option regarding our dataset, also CAO is more computationally expensive. Also, because, in the cost function, “Random” got the same result as the “HUANG” initialization.

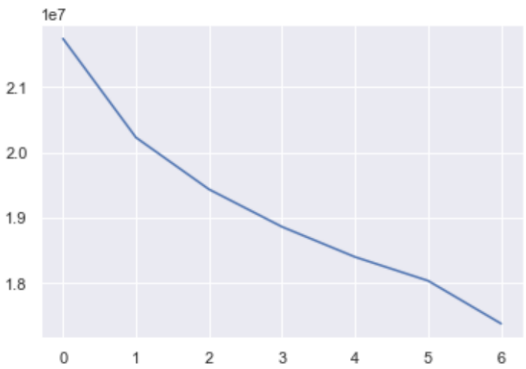

Figure 4 - Cost function initializatio HUANG

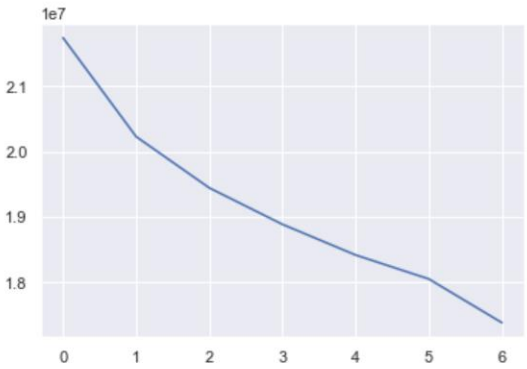

Figure 5 - Cost function initializatio RANDOM

The reasons that lead us to select this algorithm was for the fact that produced good results to the pretended segmentation using mixed data types.

Next, we ran the cost function in order to find the number of clusters, where we chose 4 clusters, given the result and since we could get some understanding about the clusters.

Subsequently, we fit_predict the dataset to obtain the label for each individual. Eventually, we had 18261 individuals in cluster 0, 9322 in cluster 1, 27285 in cluster 2 and 35635 in cluster 3.

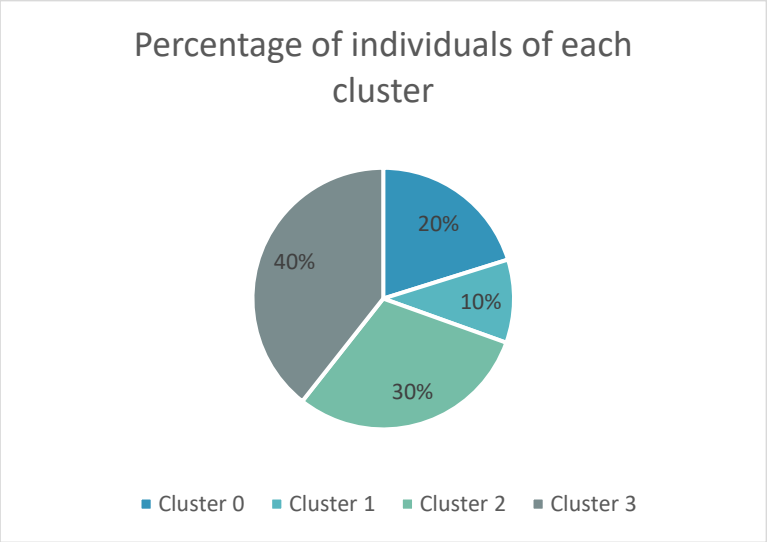

Figure 6 - Graphic with the % of individuals of each cluster

To help us to distinguish, we select a subset of variables by their relevance to the business problem, that is the following: Income, Recency, AVGGIFT, POP90C1, Age, HC13, PEC1, NUMPRM12, TPE3, IC3, AFC1, EC1, VC3, EIC5, EIC4 and DOMAIN (the graphics of the analysis are in the appendix).

5.5. T-SNE

T-SNE is a tool to visualize high dimensional data, that´s why we decided to use with that thought in mind. T-SNE differs from PCA by maintaining only small pairwise distances or local similarities while PCA is concerned with preserving large pairwise distances to keep the most variance. We decided to try it in our data because we can use to see our data points but also, we can see some clusters in the output with the right set of parameters. For this dataset, we set the number of components equal to 2 and perplexity equivalent to 5.

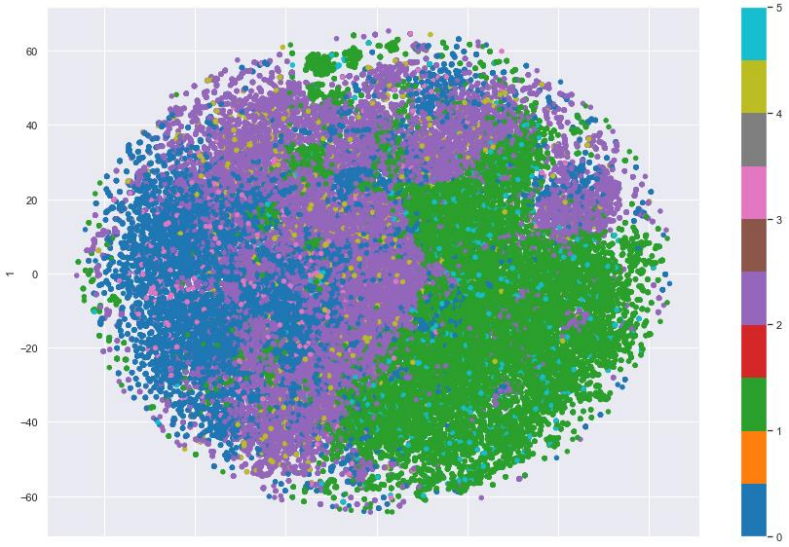

Figure 7 - T-SNE data visualization

5.6. Cluster Analysis for K-Prototypes:

After the four methods, we decided to proceed with our analysis with the output from K-Prototypes. Bellow, we have the clusters distinguished by the selected variables.

Variables	Cluster 0	Cluster 1	Cluster 2	Cluster 3
INCOME		The cluster has a larger range of incomes	Individuals with income above average	
Recency		Has more donors that donate recently		Has more donors that donate recently
AVGGIFT			Individuals that spent more per gift	Individuals that spent less per gift
POP90C1	A high percentage of individuals living in urbanized areas		A high percentage of individuals living in urbanized areas	A low percentage of individuals living in urbanized areas
Age		Older individuals		
HC13		More diversity and also a higher percentage of housing units heated electricity	Presents the lower percentage of housing units heated electricity	
PEC1	A lower percentage of individuals that work outside of residence area			
NUMPRM12		Has more individuals that received promotions on the last 12 months		
TPE3	Has more individuals that use			Has fewer individuals that use

	public transportation			public transportation
IC3	Low average income household		Higher average income household	Low average income household
AFC1	Has a higher percentage of individuals in active military service			
EC1	Has individuals with fewer years of education		Has individuals with more years of education	
VC3	A lower percentage of veterans from WWII	A higher percentage of veterans from WWII	A lower percentage of veterans from WWII	
EIC5		A lower percentage of individuals employed in transportation		A higher percentage of individuals employed in transportation
EIC4		A lower percentage of individuals employed in Manufacturing		A higher percentage of individuals employed in Manufacturing
DOMAIN	C3 – city, lowest social-economic status of the neighbourhood S2- suburban, average social-economic status of the neighbourhood	C2 - city, average social-economic status of the neighbourhood	S1 - suburban, high social-economic status of the neighbourhood	R2 – Rural, the average social-economic status of the neighbourhood

Table 3 - K-Prototypes cluster analysis

5.7. Clusters overview:

5.7.1. Cluster 0 – Life Beginners

Is composed by individuals that live in urban areas, with lower percentage regarding working outside of the residence area, use more public transportation, have a lower average income household, with more people in active military service, fewer years of education, most live-in city or in suburbs, belong to the lowest social-economic status of the neighbourhood or the average social-economic status of a neighbourhood, lapsing donors that send 1 gift in last 13-24 months and spent 15-24.99 and less percentage of veterans from WWII. This cluster has 18261 individuals.

Due to these characteristics, we can call this cluster “Life Beginners”, that is, people that are in the being of their independent adult life.

5.7.2. Cluster 1 – Older adults with a stable financial situation

Is constituted by individuals with higher income, older age, that has more people with housing units heated by electricity, more individuals that received promotions on

last 12 months, more donors that donated recently, live-in city, with the average social-economic status of a neighbourhood, lapsing donors that send 1 gift in last 13-24 months and spent 15-24.99 and more people who are WWII veterans. This cluster has 9322 individuals.

Due to these characteristics, we can call this cluster “Older adults with a stable financial situation”, that is, people of advanced age but with an average social-economic status.

5.7.3. Cluster 2 - Suburbs

Is formed by individuals that live in urbanized areas, that spent more money per gift, higher average income household, with more years of education, fewer veterans from WWII, live-in suburbs, with the high social-economic status of a neighbourhood, lapsing donors that send 1 gift in last 13-24 months and spent 15-24.99 and employed in the education sector. This cluster has 27285 individuals.

Due to these characteristics, we can call this cluster “Suburbs”, that is, people that live in the suburbs with a good social-economic status.

5.7.4. Cluster 3 – Middle-low class

Is represented by individuals that live in rural areas, that spent less per gift, use less public transportation, low average income household, live-in Rural, with the average social-economic status of a neighbourhood, more donors that have donated recently, lapsing donors that send 1 gift in last 13-24 months and spent 15-24.99 and in transportation and manufacturing. This cluster has 35635 individuals.

Due to these characteristics, we can call this cluster “Middle-low class”, that is, people that live in rural areas with a low average income of the household and average social-economic status.

5.8. Classification of the outliers

Later, as we decided to use the clusters from K-Prototypes as the solution for our business problem, we decided to see what would be the classification of the outliers that we removed previously in the process. In that way, to predict the labels, we used the Decision Trees classifier, where we considered 70% as the Train dataset and the remain as the Test dataset, also we insert the parameter stratify to have the same quantity of each label in each part. The model estimates that on average, it can predict 81.08% of the donors correctly. As a result, in the outliers dataset, we have 40 individuals that belong to the cluster 1, 3691 donors that were classified as cluster 2, 171 donors to cluster 3 and none to cluster 0.

6. Marketing Campaign

We noticed that the individuals from all clusters are Lapsed donors, that is, is the last time they send a gift between 12 to 24 months ago, considering the last promotion mail in 2017. So, we must do a campaign in order to encourage the increase in the number of times they donate and how much they donate.

Thus, considering **cluster 0**, they should incentive the donations through the social media, since they are younger people and because of that they spent more time in these networks and also since they usually use public transportation, PVA should a create a campaign where volunteers will go to the terminal to collect donations.

Regarding **cluster 1**, instead of sending by mail, PVA could phone the donors in order to keep an active line and strengthen the connection between both parts, allowing them to be more present and devoted to such cause in the organization.

Concerning **cluster 2**, the best approach to communicate with these individuals and keep them alert about the news would be to send the information through SMS since there is a more interactive of passing information. Also, since they spent more on average in the gift than the other clusters and good way to motivate them is to invite them to events, such as auctions.

Finally, to the **cluster 3**, since this is the one that we should focus more, a strategy to be closer to the communities and have a face-to-face communication is to send volunteers to collect donations on the festivals and events that occur in the rural areas.

7. Conclusion

Based on our donor's segmentation, we ended up with four clusters, being cluster 1 and 2 the ones with higher importance, there is, the ones that PVA should invest the most since they have bigger financial power, and they are willing to spend more on their gifts. Although these are the priority segments, PVA should not discard cluster 3 since they have a greater number of individuals. Additionally, apart from their low financial condition, they are willing to donate a percentage of their living, being the ones that have donated more recently, so it will be good for PVA to keep them engaged.

In that way, we come up with these conclusions because, in our research, we first focused on understanding the variables and their relationship. Then, we try to select the variables that were more meaningful to our purpose. Subsequently, we focus on trying different clustering algorithms to achieve the one that most contribute to the understanding of the donors' behaviour. In the end, we used the K-Prototypes to build the clusters.

To conclude, to answer our initial questions, we can say that we come up with four clusters by using the following variables to help to describe them: INCOME, Recency, AVGGIFT, POP90C1, Age, HC13, PEC1, NUMPRM12, TPE3, IC3, AFC1, EC1, VC3, EIC5, EIC4 and DOMAIN. Therefore, we had into account the household income of the donor as well as his/her age and the number of promotions he/she has received in the last 12 months. Regarding donors's neighborhood we take advantage of the percentage of population in the neighborhood that live in urbanized area, the percentage of house units heated by eletrivity, the percent working outside the state of residence, the percentage of neighbors using public transportation, the average household income of the neighborhood; also, the percentage of neighbors in active military service, the completed median years of school completed by donors' adult neighbors, the percentage employed in transportation and in manufacturing and the urbanicity level and socio-economic status of donor's neighborhood.

8. Appendix

8.1. K-means (Different Perspectives) graphics:

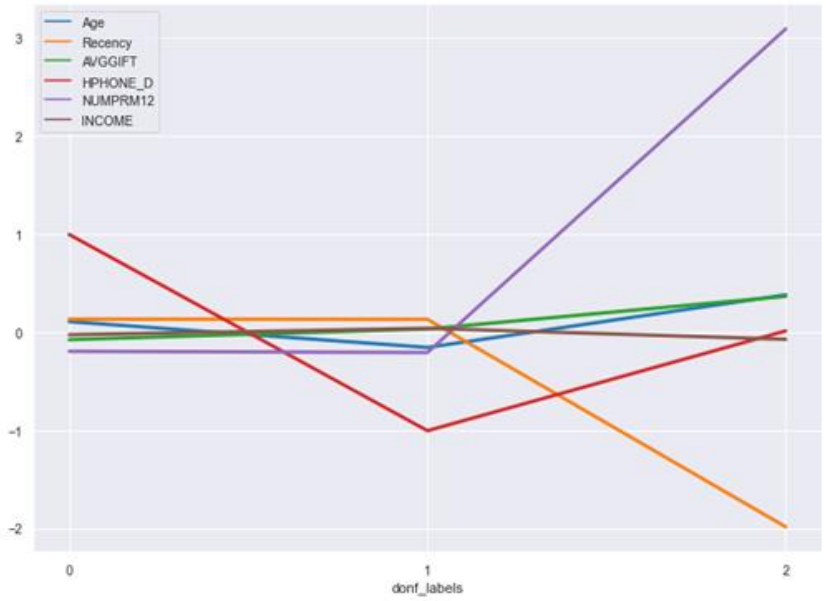

Figure 8 - Mean of Donors Variables per Cluster

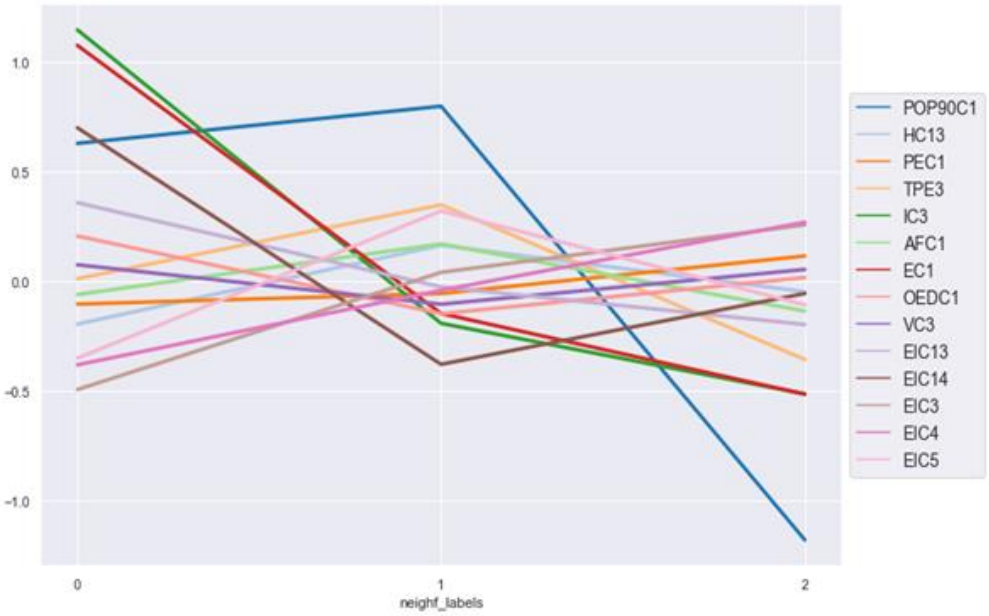

Figure 9 - Mean of Neighborhood Variables per Cluster

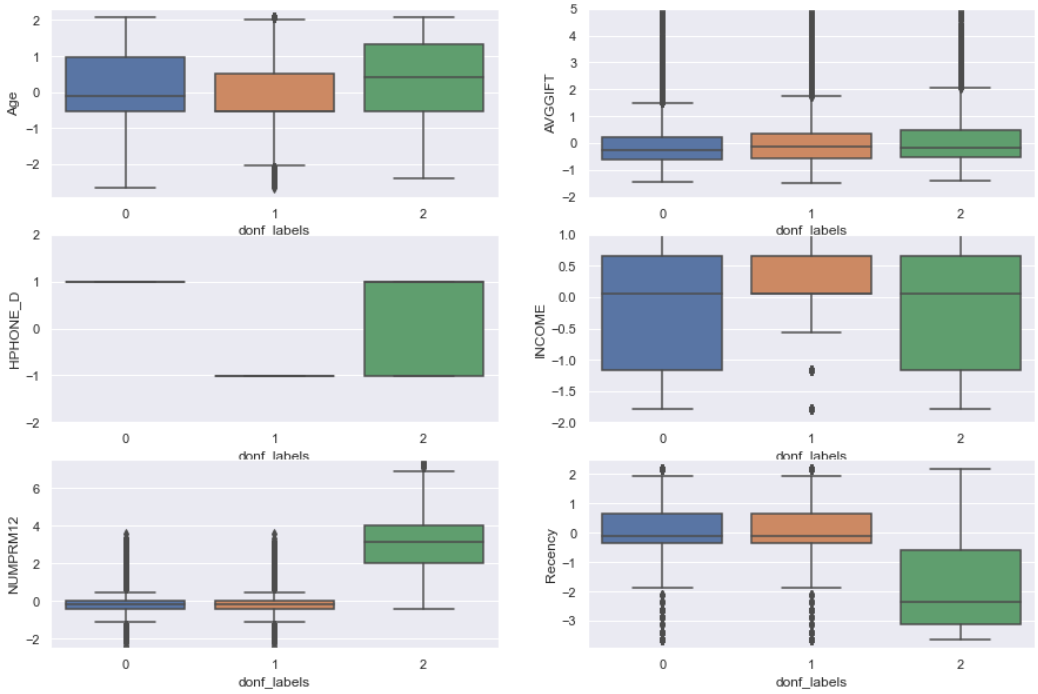

Figure 10 - Distribution of Donors Variables

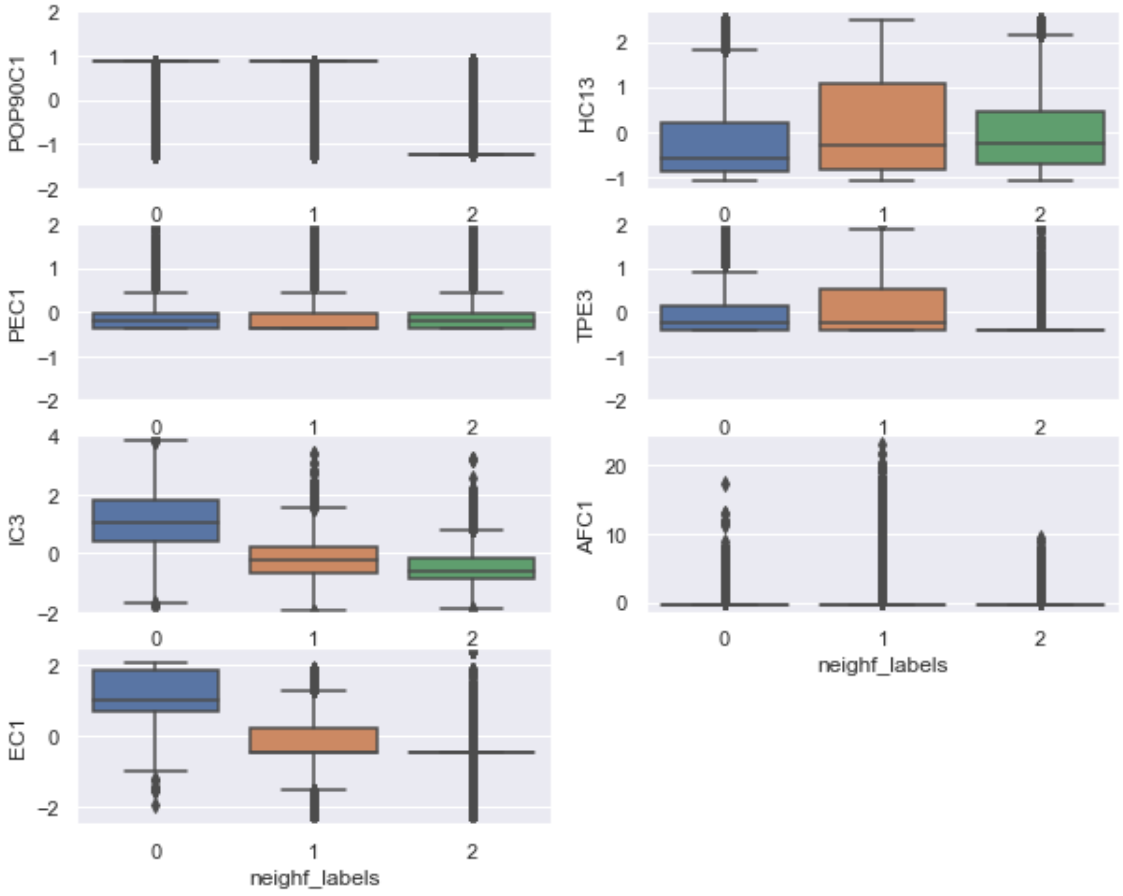

Figure 11 - Distribution of Neighborhood Variables

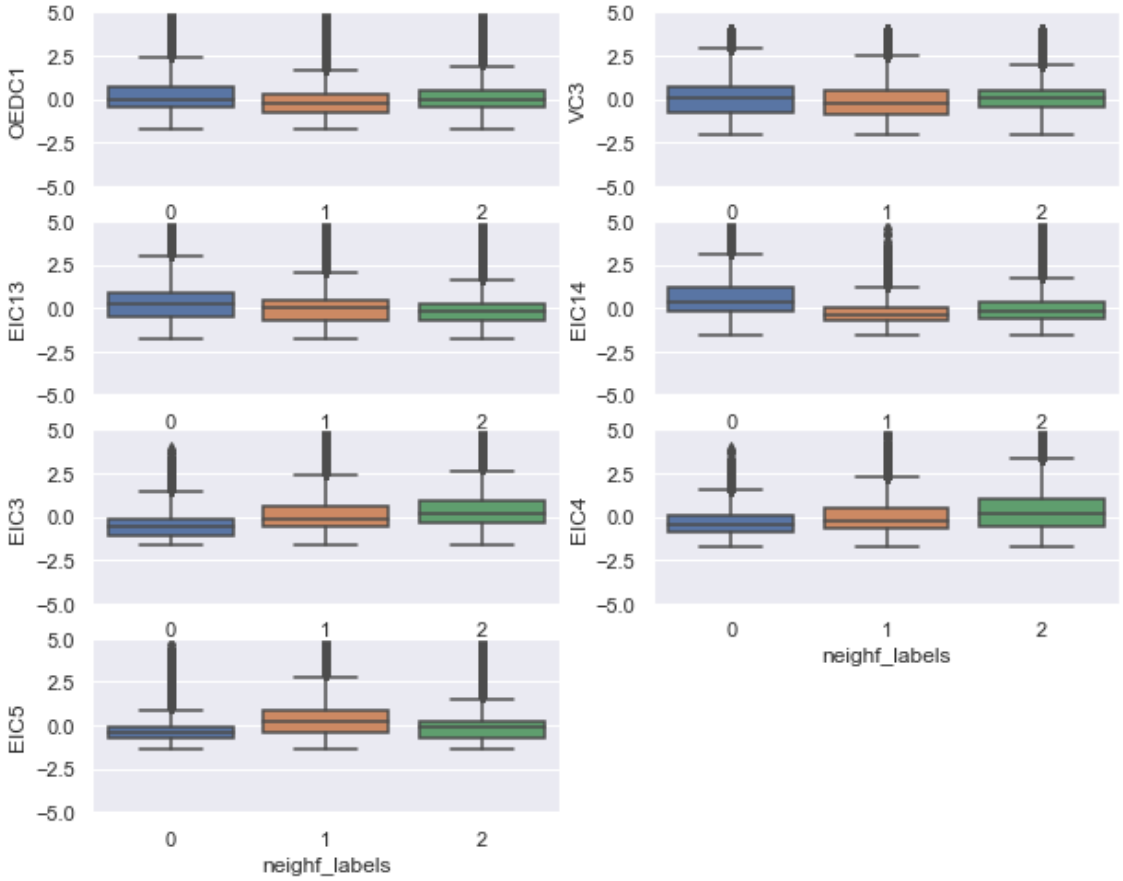

Figure 12 - Distribution of Neighborhood Variables Continuation

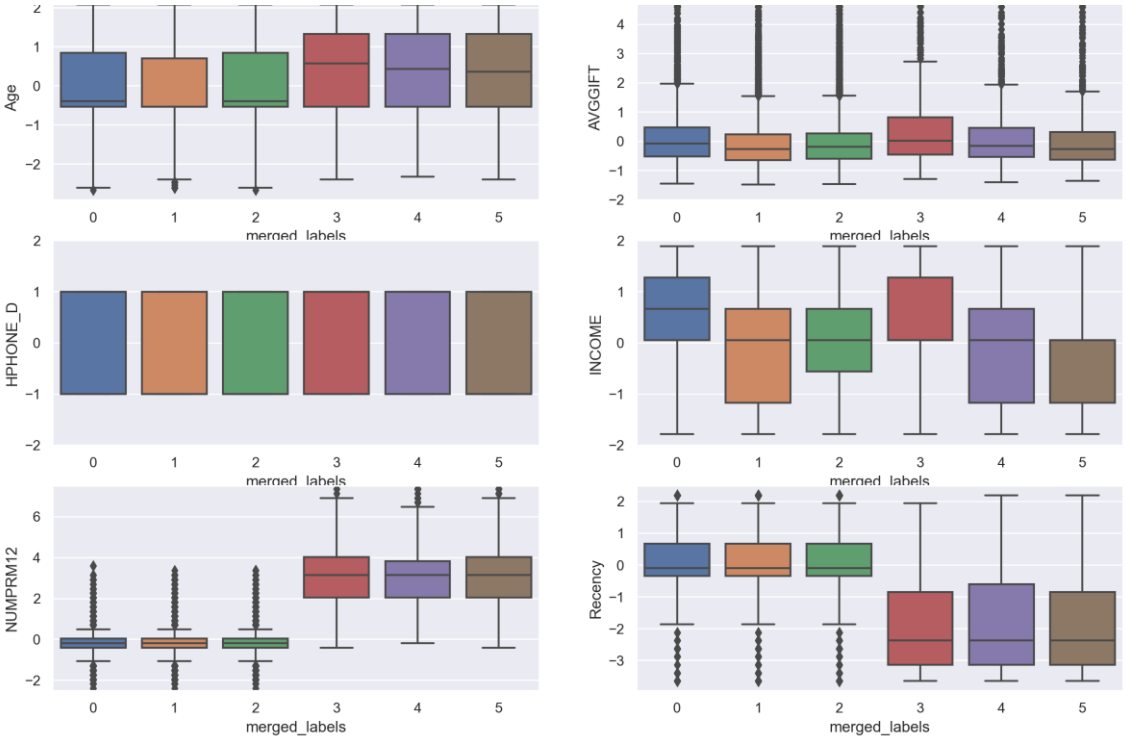

Figure 13 - Distribution of Joint perspectives K-means

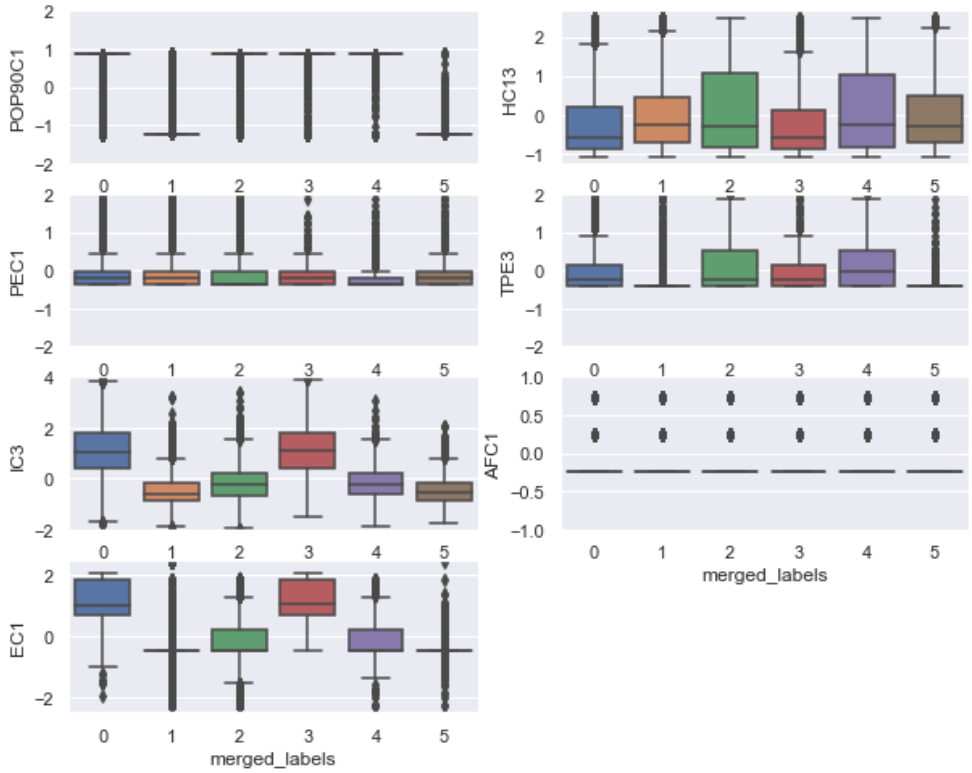

Figure 14 - Distribution of Variables Joint Perspectives Continuation

8.2. K-Prototypes graphics:

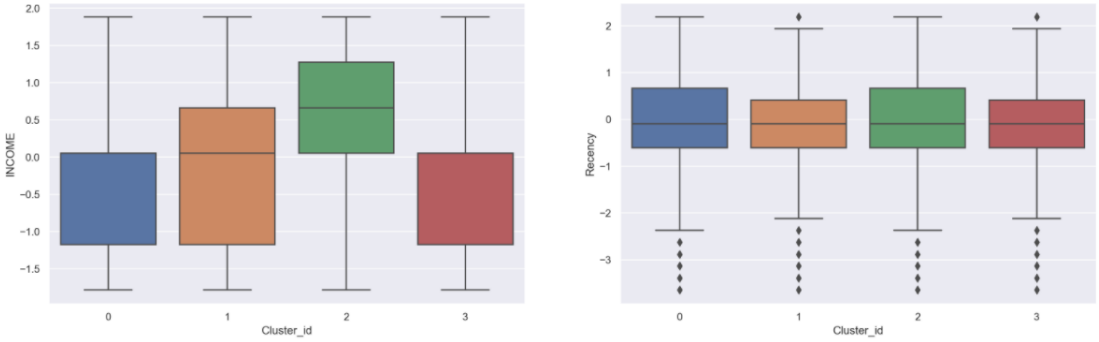

Figure 15 - Distribution of Income and Recency K-Prototypes

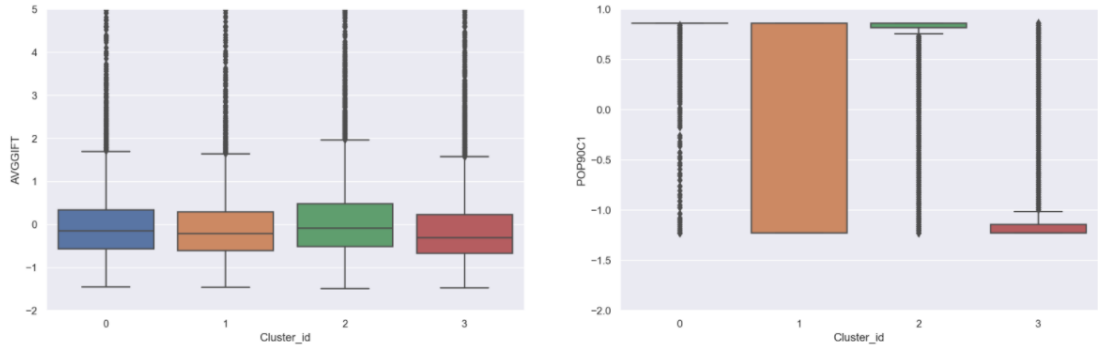

Figure 16 - Distribution of AVGGIFT and POP90C1 K-Prototypes

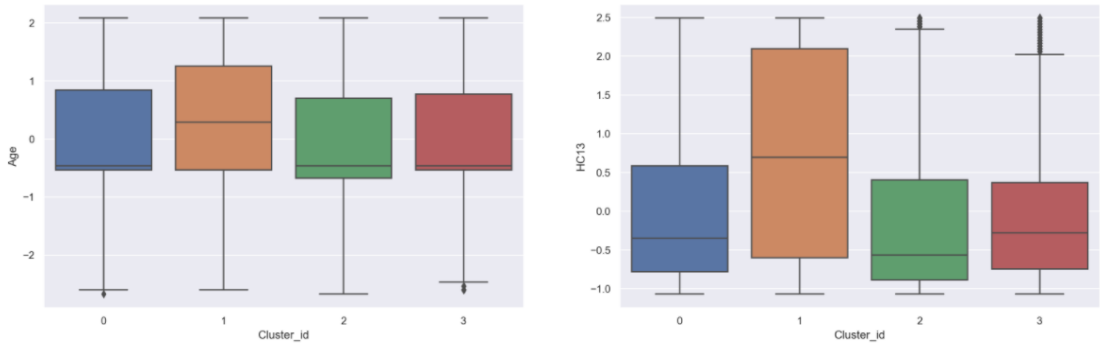

Figure 17 -Distribution of Age and HC13 K-Prototypes

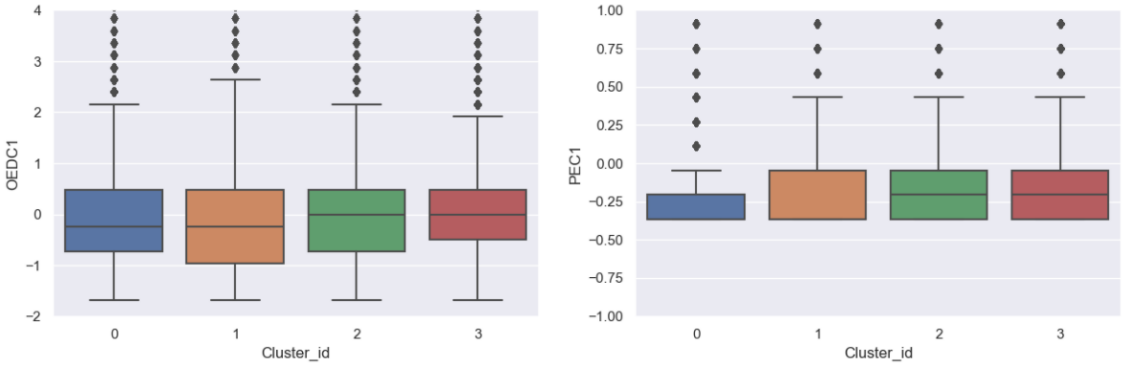

Figure 18 - Distribution of OEDC1 and PEPC1 K-Prototypes

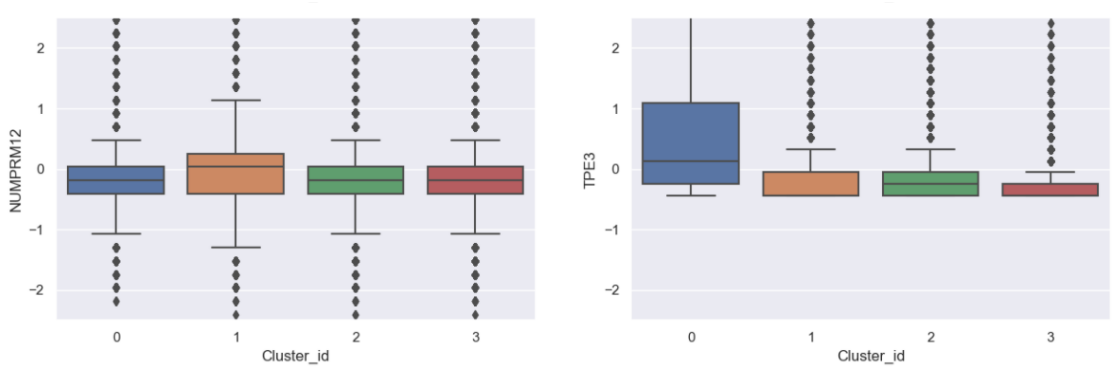

Figure 19 - Distribution of NUMPRM_12 and TPE3 K-Prototypes

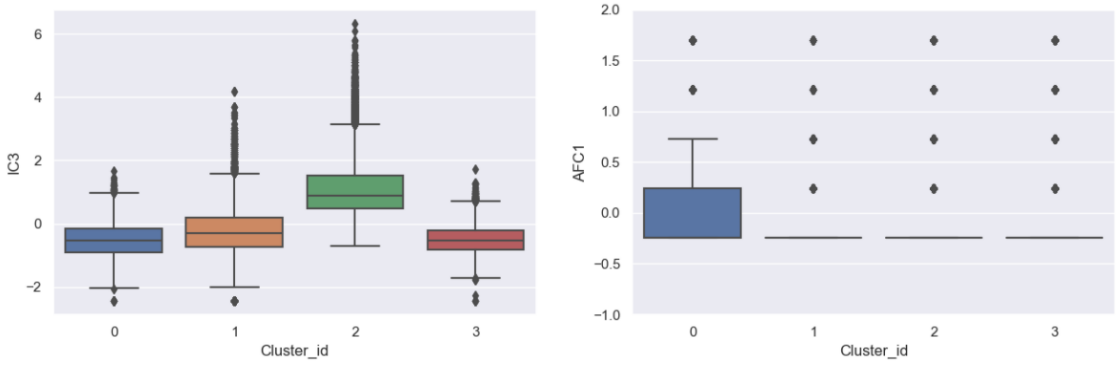

Figure 20 - Distribution of IC3 and AFC1 K-Prototypes

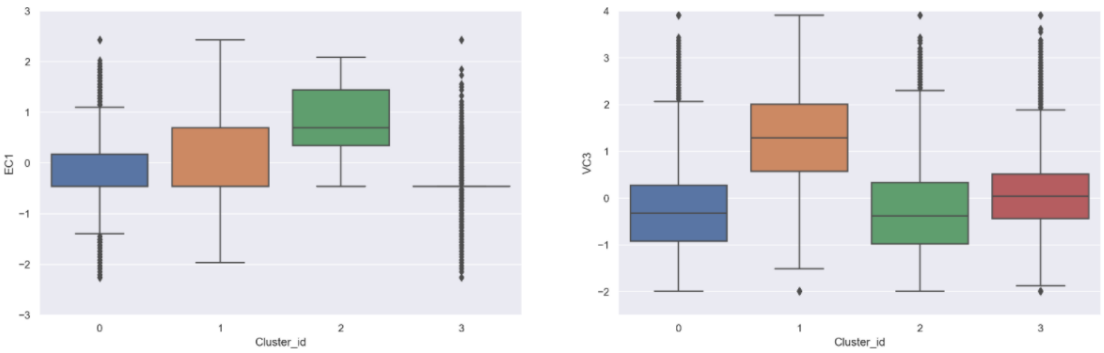

Figure 21 - Distribution of EC1 and VC3 K-Prototypes

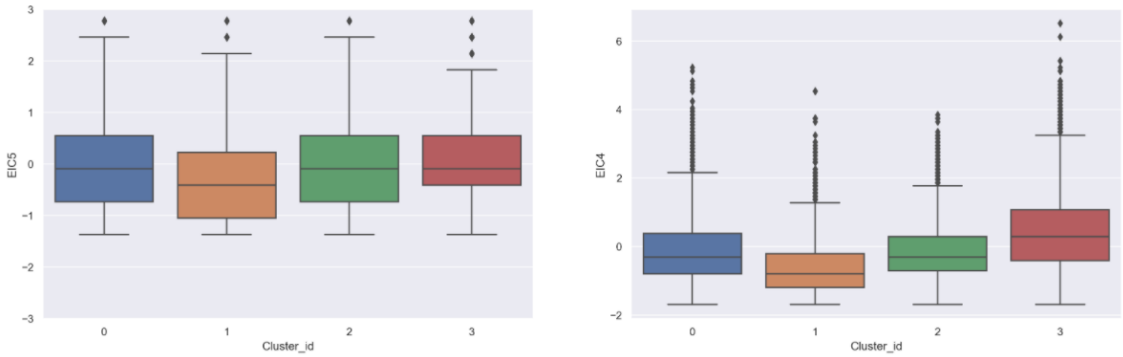

Figure 22 - Distribution of EIC5 and EIC4 K-Prototypes

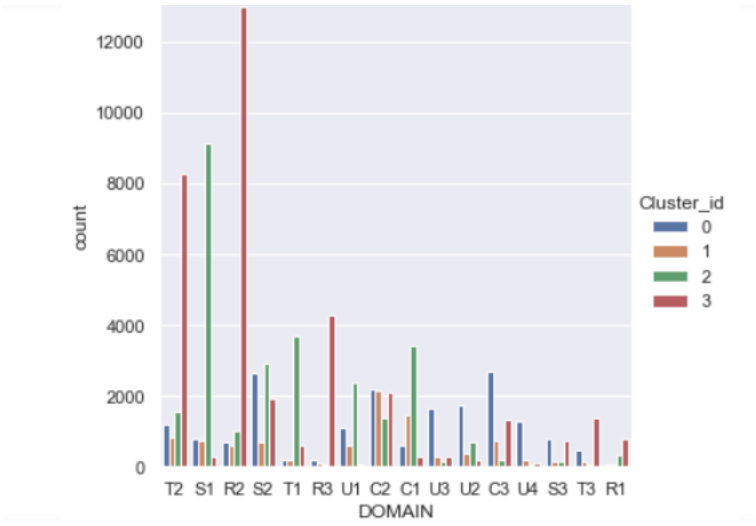

Figure 23 - K-Prototypes non_metric feature DOMAIN plot

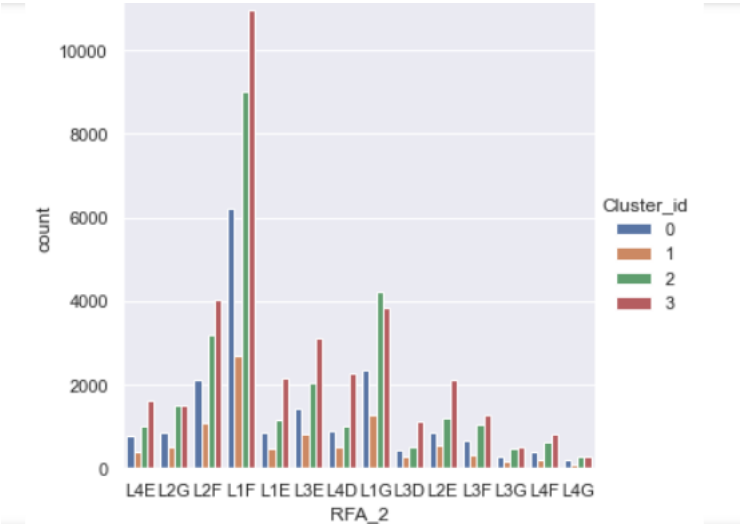

Figure 24 - K-Prototypes non_metric feature RFA_2 plot

8.3. Silhouette Graphics – K-means:

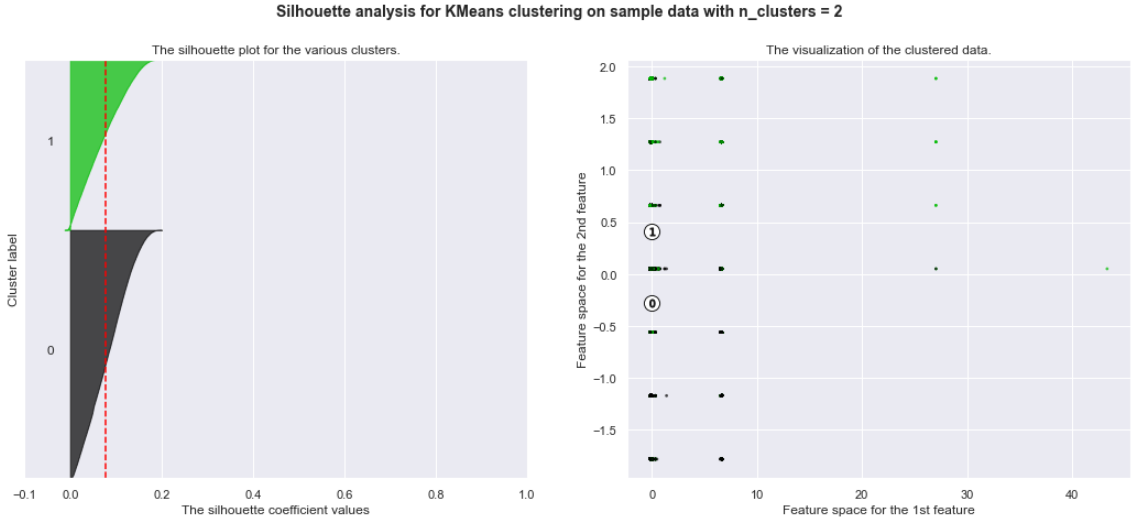

Figure 25 - Silhouette analysis for 2 clusters

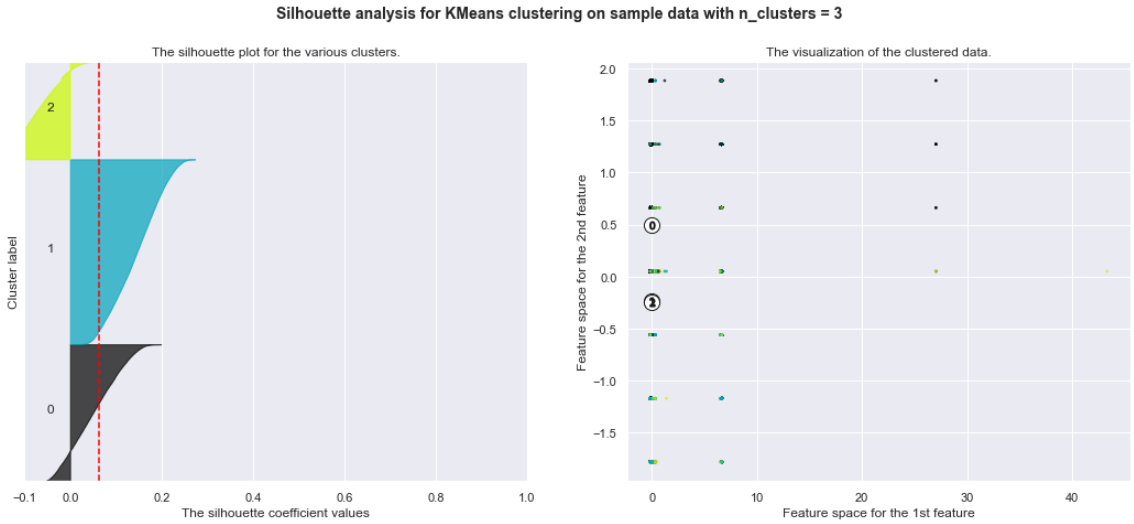

Figure 26 - Silhouette analysis for 3 clusters

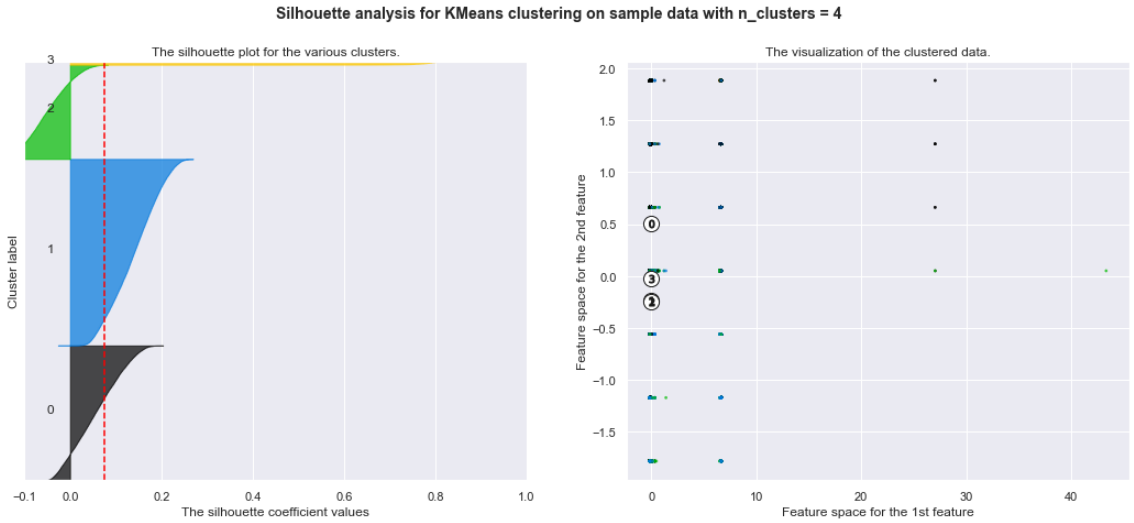

Figure 27 -Silhouette analysis for 4 clusters

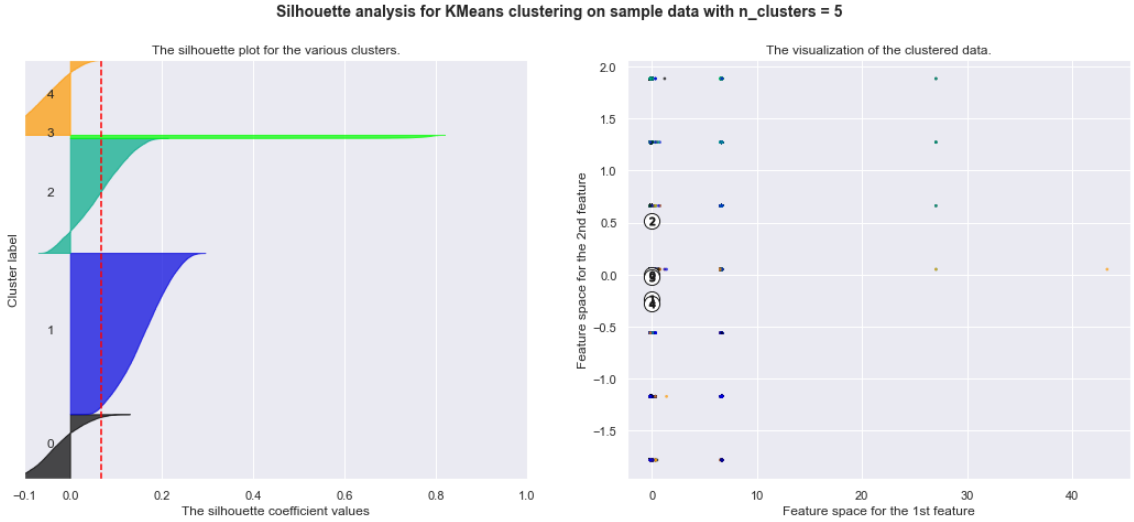

Figure 28 -Silhouette analysis for 5 clusters

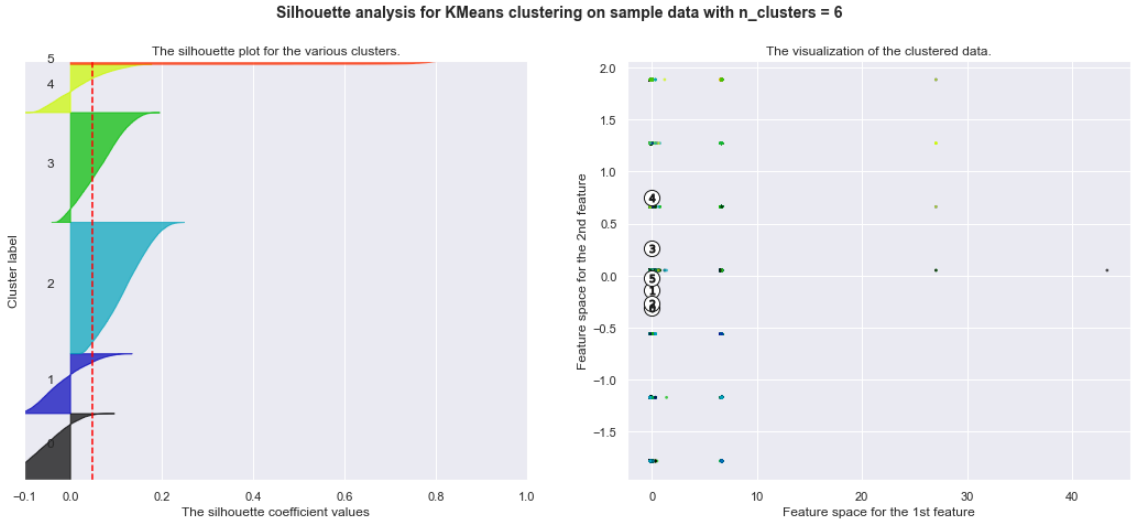

Figure 29 - Silhoutte analysis for 6 clusters

8.4.Silhouette Graphics – K-means with perspectives:

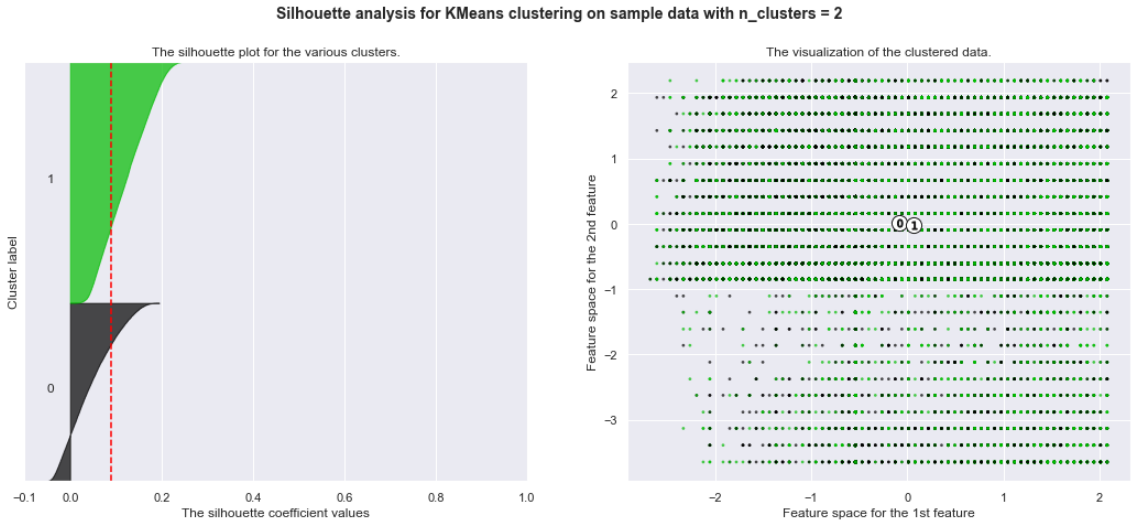

Figure 30 - Silhouette analysis for 2 clusters

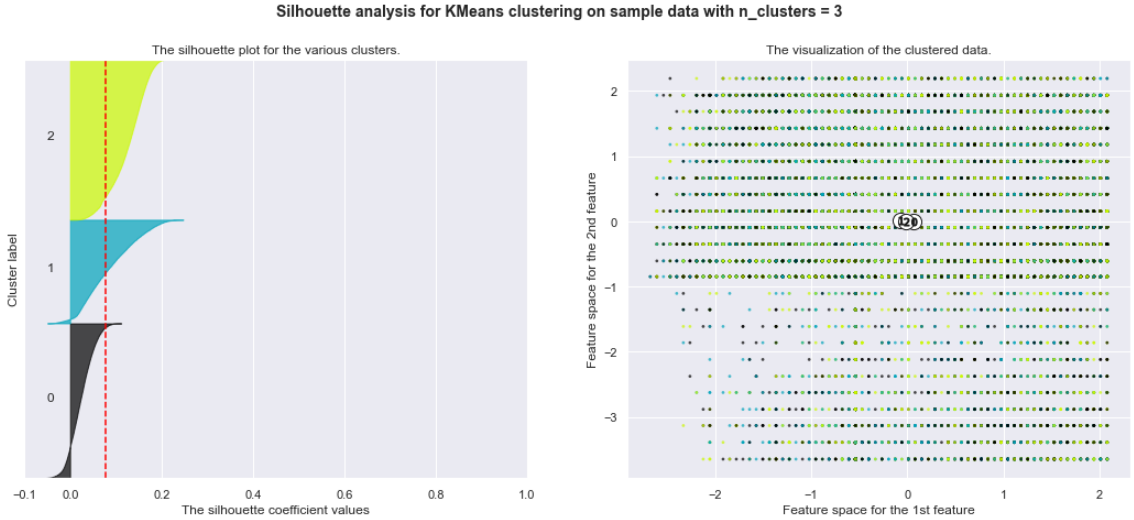

Figure 31 - Silhouette analysis for 3 clusters

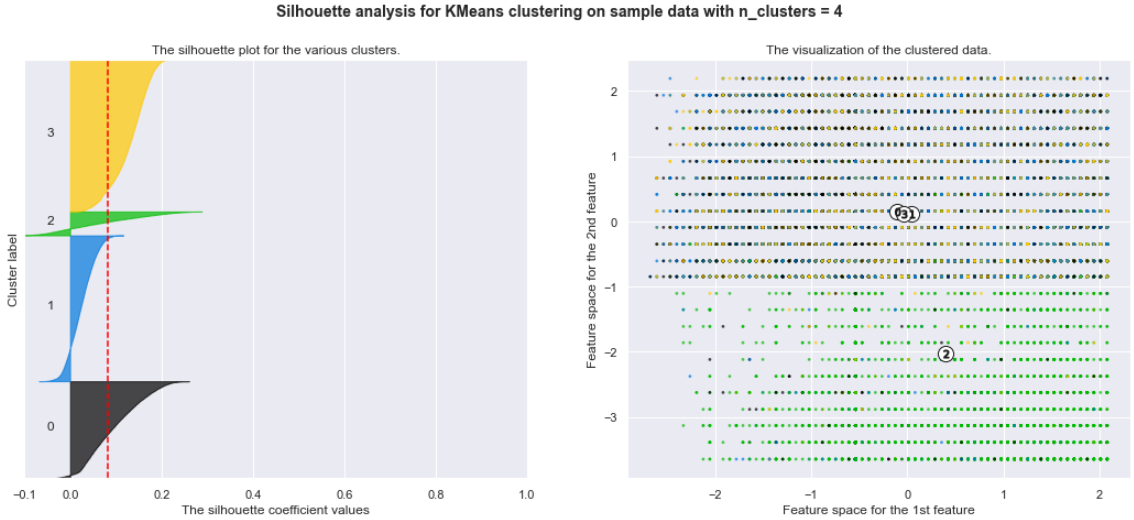

Figure 32 - Silhouette analysis for 4 clusters

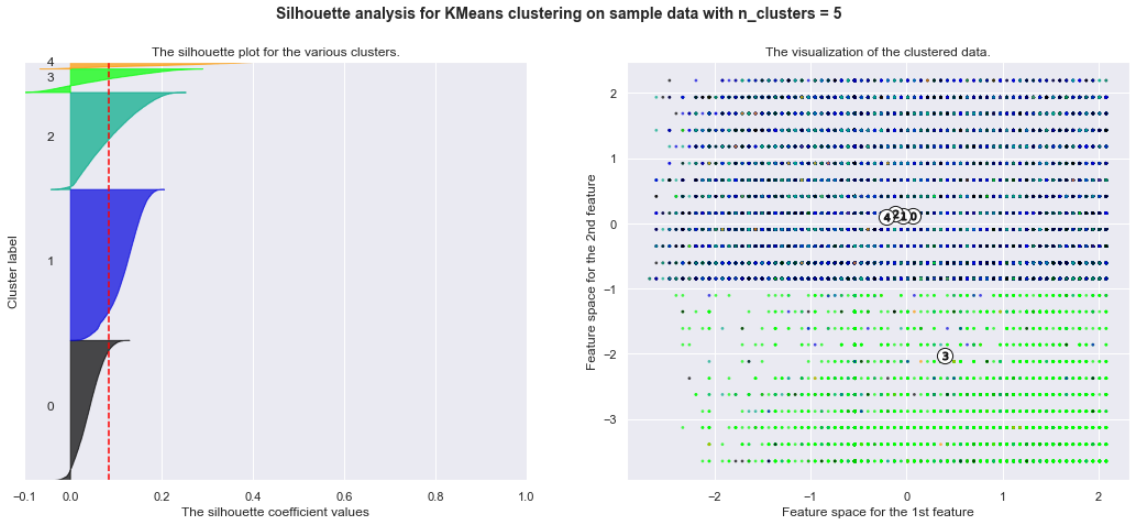

Figure 33 - Silhouette analysis for 5 clusters

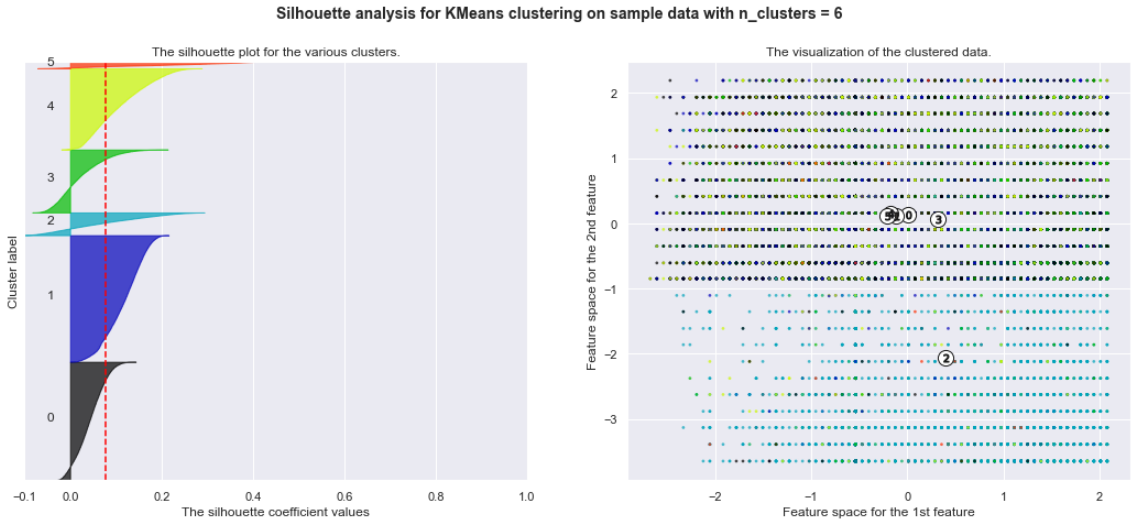

Figure 34 - Silhouette analysis for 6 clusters