DATA MINING PROJECT

REPORT

2020/2021

Master Program in Data Science and Advanced Analytics

NOVA IMS
Information Management School

GROUP MEMBERS
Inês Melo m20200624
Ricardo Nunes m20200611
Catarina Pinheiro m20200654

Table of Contents

Introduction

This project was developed as a part of the course of Data Mining from the master's in Data Science and Advanced Analytics at Nova IMS. We were proposed to analyse a sample of the results of one of PVA's recent fundraising appeals, containing 95412 donors.

PVA - Paralyzed Veterans of America - is non-profit organization that provides programs and services for US veterans with spinal cord injuries or disease.

Our job is to study a particular group of donors, lapsed donors. This group is represented by people who made their last donation 13 to 24 months ago, and so, it is our purpose to help recapture these former donors.

By the end of this project, we will know the different segments of potential donors in order to be possible to express a marketing approach for each of those segments so that lapsed donors are more likely to give again. This means clustering the data and understanding the donor's behaviour, defining, describing and explaining each chosen cluster.

We will discuss various approaches from more than one view and interpret the advantages and disadvantages of each decision to make the best possible final approach.

Data preparation

Feature Selection

After importing the given data from PVA we did a brief analysis of its structure. Realising the data had 476 features it was important to select only the ones appropriated for our assignment, so that it facilitates all a posteriori assessment. In the following section we are going to explain what features we have selected and why we have chosen them.

Our strategy for feature selection was based on two main criteria: demographic features, which tell us more about the donor's **life characteristics**; and gift features, which give us an idea about their behavioural pattern since they joined the PVA's database. The features selected were:

STATE (Categorical): Refers to the living state of the donors and gives us a geographical overview of them.

DOB (Date): Refers to the donor's date of birth which gives us a notion of the age.

DOMAIN (Categorical): Refers to the urbanicity donor's neighbourhood, whether it is urban, city, suburban, town or rural, and its socio-economic status.

GENDER (Categorical): Points out if the donation comes from a male, female, unknown or a joint account.

HIT: Refers to the total number of known times the donor has responded to a mail order offer other than PVA's.

AGE901 (Numerical): Denotes the donor's neighbourhood median age.

EC1 (Numerical): Denotes the donor's neighbours, who are older than 25 years, median years of school.

IC1 (Numerical): Points out the donor's neighbourhood median household income in hundreds.

MHUC1 (Numerical): Describes the donor's neighbourhood median homeowner cost with mortgage per month.

POP901 (Numerical): Refers to how many people live in the donor's neighbourhood.

PCOWNERS (Categorical): Indicates whether a donor has a computer or not.

NUMPRM12 (Numerical): Indicates the number of promotions that the donor received in the last 12 months.

LASTDATE (Numerical): Refers to the most recent gift date.

LASTGIFT (Numerical): Points out how much was the donor's last gift.

ODATEDW (Date): Indicates the date of the donor's first gift to the organization.

AVGGIFT (Numerical): Refers to the average dollar amount of gifts until the present date.

CARDGIFT (Numerical): Denotes the number of card promotions the donors engaged to.

RFA_2F (Numerical): Indicates the donor's frequency for the most recent promotion.

RFA_2A (Numerical): Indicates the donor's donation amount for the most recent promotion.

IC5 (Numerical): Points out the donor's neighbourhood per capita income

Other variables were considered, such as **NUMCHLD,** that refers to the number of children of the donor, **WEALTH2**, which refers to the median wealth rating of the family income of the neighbourhood's donor, **INCOME**, referring to the income rating of the donor. However, **NUMCHLD** had a very low cardinality (+90% of zeros), meaning the feature would not give us any additional information for clustering, **WEALTH2** presented a lot of missing values (45%), which gave us inaccurate inference about the data and **INCOME** was not going to be a right fit to cluster since it is a low cardinality feature, and for reference of the economic status we already have the feature **DOMAIN.**

Feature Engineering

What features can we simplify on our favour? In order to improve the clustering, we need to create and transform features based on the original ones, to help us interpret the given information, and improve the efficiency of the clusters.

- **AGE:** We transformed the feature DOB into **AGE,** having the data of the last promotion as the current year, indicating the age of the donors and not their birthday date.
- **SES** and **DOMAIN**: Since the **DOMAIN** feature had originally two bytes, we decided to break it into two variables: **SES,** socio-economic status, and **DOMAIN,** the donor's neighbourhood's urbanicity level.
- **PCOWNERS:** We transformed this categorical variable into a binary one. If the donor is a pc owner then this variable goes for 1, otherwise this variable goes for 0.
- **NOCARDGIFT:** This feature corresponds to all the donations that were made outside the promotions. We came across this variable by subtracting the values of the feature **CARDGIFT** to **NGIFTALL.**
- **LASTDATE:** This feature was in a form of date, but we transformed into months in order to realize how many months was the donor's last donation.
- **ODATEDW:** to compare better the one before we also transformed the date of the first donation into months.
- **EC1:** the original feature had the values (years of education) multiplied by 10, so we divided them by 10.

We gather up all the donations along the years of the promotions accordingly to their types, this gives us a notion of the promotions that had more adherence:

Total_NK joins all the donations of the mailings that are blank cards with labels.

Total_TK joins all the donations of the mailings that have thank you printed on the outside with labels.

Total_SK joins all the donations of the mailings that are blank cards with labels.

Total_LL joins all the donations of the mailings that have labels only.

Total_G1 joins all the donations of the mailings that have labels and a notepad.

Total_GK joins all the donations of the mailings that are general greeting cards (an assortment of birthday, sympathy, blank, & get well) with labels.

Total_CC joins all the donations of the mailings that are calendars with stickers but do not have labels.

Total_WL joins all the donations of the mailings that have labels only.

Total_X1 joins all the donations of the mailings that have labels and a notepad.

Total_XK joins all the donations of the mailings that are Christmas cards with labels.

Total_FS joins all the donations of the mailings that are blank cards that fold into thirds with labels.

Logical inconsistencies in Data

We know that lapsed donors are the group of people that do not make donations 13 to 24 months ago. Therefore, the first thing we did was to check if the feature **LASTDATE** had values in the interval mentioned above. We have only kept the values who checked the above condition, since the lapsed donors are the focus of the project. Then we removed all the duplicated values.

Regarding logical conditions of the variables, we can assume that for every donor, the variable **AGE** must be equal or higher than the date of the first donation, **ODATEDW**. Also, the date of the last donation has to be more recent than the date of the first donation. To assure the data's consistency, we eliminated observations whose data did not check the conditions mentioned above. Besides that, we verified that in the feature **ODATEDW**, there were no dates after the date of the last promotion.

GENDER: we excluded all the values that did not match the description, such as 'A' and 'C'.

STATE: having 57 different values, when there are only 50 different states, we checked the reaming 7 and came across either to island areas, US military post code or federal district:

AP -> U.S. Armed Forces - Pacific

AA -> U.S. Armed Forces - America

AE -> U.S. Armed Forces - Europe

VI -> U.S. Virgin Islands (Insular area)

GU -> Guam (Insular area)

DC -> District of Columbia (Federal district)

AS -> American Samoa (Insular area)

At the end, before moving on into dealing with missing values we categorized the data to metric and non-metric features. Being the metric features: HIT, AGE901, EC1, IC1, MHUC1, POP901, NUMPRM12, LASTDATE, LASTGIFT, ODATEDW, AVGGIFT, NOCARDGIFT, CARDGIFT, IC5, Total_NK, Total_TK, Total_SK, Total_LL, Total_G1, Total_GK, Total_CC, Total_WL, Total_X1, Total_XK, Total_FS, AGE and the non-metric: STATE, DOMAIN, GENDER, PCOWNERS, RFA_2F, RFA_2A, SES.

Dealing with missing values

Categorical features:

DOMAIN and **SES** are the two categorical features with missing values. Since those are categorical features, meaning we are not going to use in clustering, we decided to replace the missing values with the expression 'Unknown'. Later we can analyse these features in each cluster by determining the mode, for example.

Numeric features:

The only numeric feature with missing values was **AGE** with 23883 missing values, representing 25% of the data. We proceeded to try three different methods to fill the missing values:

- **KNN algorithm** with 7 n_neighbors. We chose this number because it is a prime number so there are no ties. We do not want a small number, because it would be very sensible to outliers and would create crisp frontiers. On the other hand, a large number would be computationally expensive and unable to detect small variations.

Still, through a visualization, we compared different values for n_neighbors, such as n=1, n=7 and n=50 to check the overfitting, which confirmed that our previous choice was the right one.

- Secondly, we used the linear regression, but we resign that since the score was too low.
- Lastly, we used the median since it is resistant to outliers, but if we used this method all the values would be equal and therefore, it would probably lead us to bias.

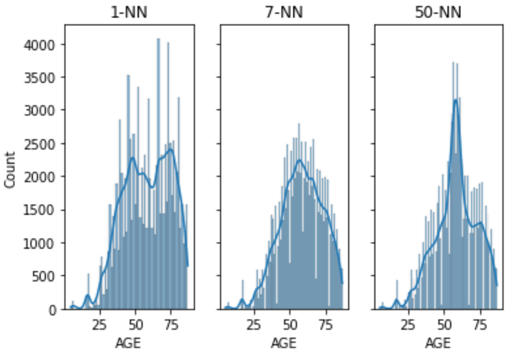


Figure 1- 1NN, 7NN, 50NN

Outliers

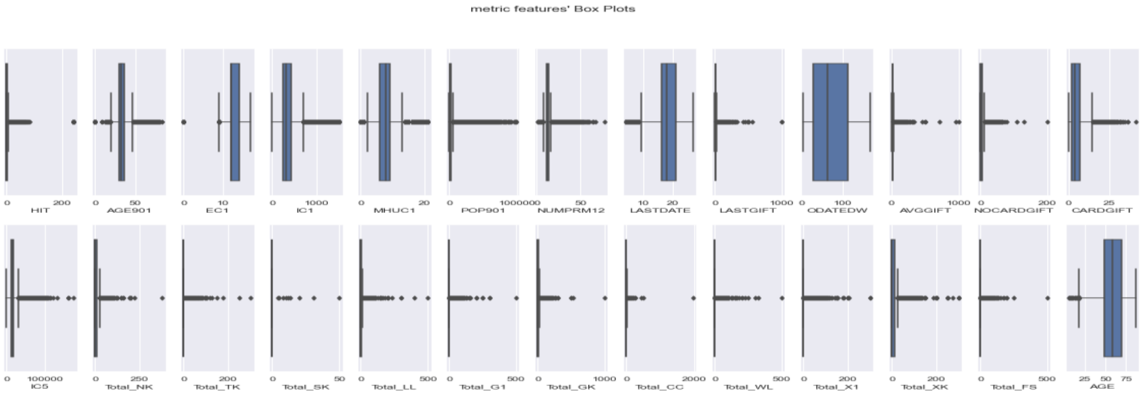
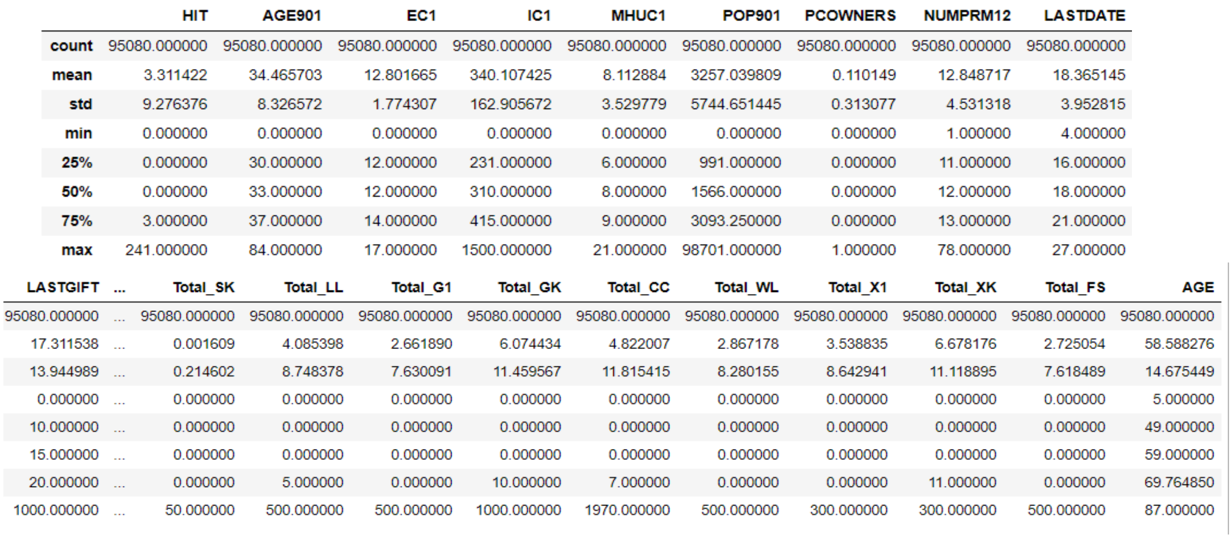


Figure 2- Boxplot for metric features

To find the outliers in our database, we began to analyse the boxplots of each one of the metric features. Looking at the boxplots, doing a first superficial analysis, we can say that a lot of features do not follow a normal distribution, and it looks that there are a lot of outliers. To help us understand that, we looked at the describe function:



	HIT	AGE901	EC1	IC1	MHUC1	POP901	PCOWNERS	NUMPRM12	LASTDATE
count	95080.000000	95080.000000	95080.000000	95080.000000	95080.000000	95080.000000	95080.000000	95080.000000	95080.000000
mean	3.311422	34.465703	12.801665	340.107425	8.112884	3257.039809	0.110149	12.848717	18.365145
std	9.276376	8.326572	1.774307	162.905672	3.529779	5744.651445	0.313077	4.531318	3.952815
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	4.000000
25%	0.000000	30.000000	12.000000	231.000000	6.000000	991.000000	0.000000	11.000000	16.000000
50%	0.000000	33.000000	12.000000	310.000000	8.000000	1566.000000	0.000000	12.000000	18.000000
75%	3.000000	37.000000	14.000000	415.000000	9.000000	3093.250000	0.000000	13.000000	21.000000
max	241.000000	84.000000	17.000000	1500.000000	21.000000	98701.000000	1.000000	78.000000	27.000000

LASTGIFT	...	Total_SK	Total_LL	Total_G1	Total_GK	Total_CC	Total_WL	Total_X1	Total_XK	Total_FS	AGE
95080.000000	...	95080.000000	95080.000000	95080.000000	95080.000000	95080.000000	95080.000000	95080.000000	95080.000000	95080.000000	95080.000000
17.311538	...	0.001609	4.085398	2.661890	6.074434	4.822007	2.867178	3.538835	6.678176	2.725054	58.588276
13.944989	...	0.214602	8.748378	7.630091	11.459567	11.815415	8.280155	8.642941	11.118895	7.618489	14.675449
0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	5.000000
10.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	49.000000
15.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	59.000000
20.000000	...	0.000000	5.000000	0.000000	10.000000	7.000000	0.000000	0.000000	11.000000	0.000000	69.764850
1000.000000	...	50.000000	500.000000	500.000000	1000.000000	1970.000000	500.000000	300.000000	300.000000	500.000000	87.000000

Figure 3- Summary for metric features

Doing a first analysis based on the box plot observations, and our prior knowledge of what makes sense to be considered an outlier on metric features, we can see that the promotions features are the ones with the most outliers, and that is probably because most of the donors do not respond to the promotions (reason there are so many zeros) and the ones who do respond give a large amount. Looking at this interpretation we can say that most of that values are not considered outliers. To deal with the outliers we tried 4 different methods:

- Manual way
- Interquartile Range method
- LocalOutlierFactor – algorithm from sklearn library
- Combination of the three methods mention above

Manual way:

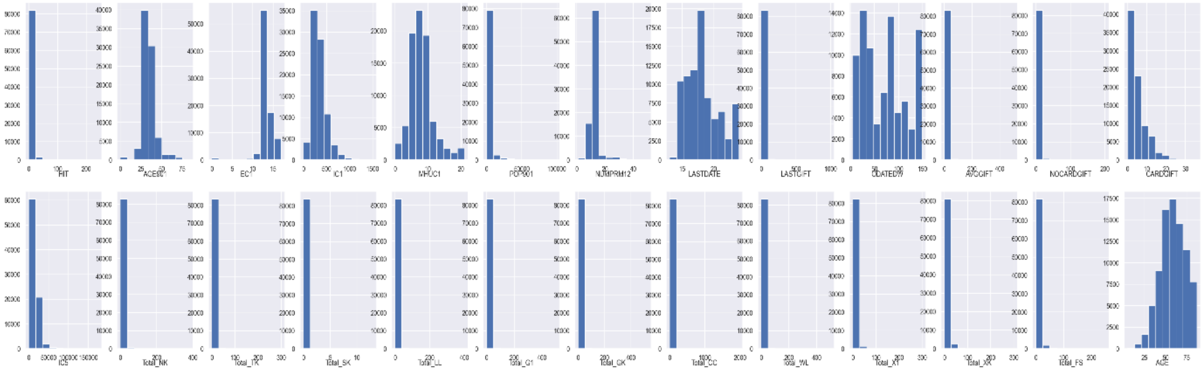


Figure 4- Histogram of each metric feature

Looking at the distribution of each feature, we kept these values for each one: HIT <= 60, NUMPRM12 <= 45, AVGGIFT <= 100, NOCARDGIFT <= 40, LASTGIFT <= 200, CARDGIFT <= 30, IC5 <= 70000, Total_NK <= 70, Total_LL <= 50, Total_CC <= 300, Total_GK <= 150, Total_XK <= 80, Total_X1 <= 80, EC1 >= 6, IC1 <= 1000, POP901 <= 50000. Percentage of data kept after removing outliers: 0.973

Interquartile Range Method: being the maximum: Q3 + 8*IQR and the minimum: Q1 -8*IQR. The percentage of data kept after removing outliers: 0.3318

LocalOutlierFactor: with n_neighbors=5. The percentage of data kept after removing outliers: 0.9725.

Combination of the three methods: we combined the three previous methods, making the condition that the filters should be at least in two of them. Percentage of data kept after removing outliers: 0.9655

At the end, the chosen method was the last one (Combination of the three methods).

Data Normalization

Using the StandardScaler () method, we scaled our data. This will help us to avoid inflations and false results lead by the distances between data points.

One-hot encoding

In order to encode the categorical features, we used One-HotEncoder. Although we are not going to use the categorical features to cluster, One-HotEnconding will help us, later, to interpret better the created clusters.

Multivariate Analysis

Before proceeding to clustering we decided to check the correlations between the metric features to see which ones are the most appropriated for clustering.

We want to eliminate the features that do not correlate to any of the other ones. That leaves us to continue further clustering without these features: Total_NK, Total_TK, Total_SK, Total_LL, Total_G1, Total_GK, Total_CC, Total_WL, Total_X1, Total_XK, Total_FS, AGE, HIT, POP901.

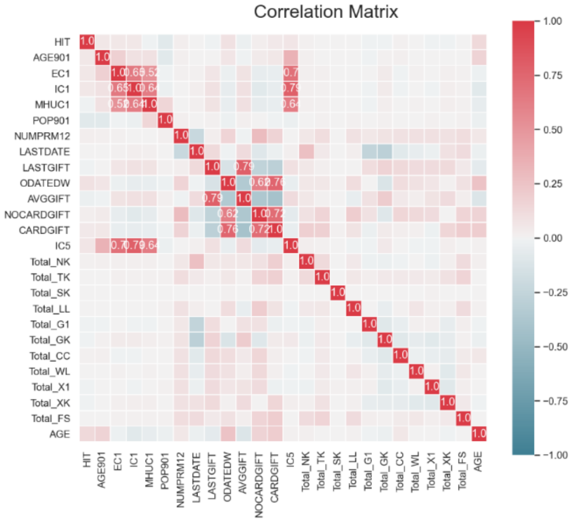
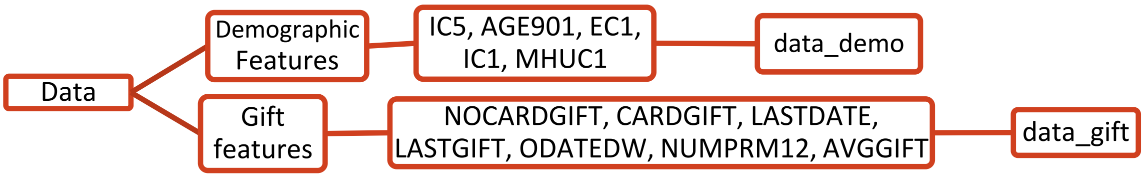


Figure 5- Correlation Matrix for metric features

Clustering

Since we took two approaches on features selection (demographic features and features related to gifts on PVA) we decided to use the same approach here, and split the features in the following perspectives and proceed the clustering for each one of them:


Data
Demographic Features
IC5, AGE901, EC1, IC1, MHUC1
data_demo
Gift features
NOCARDGIFT, CARDGIFT, LASTDATE, LASTGIFT, ODATEDW, NUMPRM12, AVGGIFT
data_gift

Techniques used:

- PCA & K-means
- K-means on top of Self-organizing maps
- Hierarchical on top of Self-organizing maps
- K-means
- DBSCAN
- Mean-Shift
- GMM (Gaussian Mixture Model)

The metric used to compare the quality of the techniques was the R^2 score.

Clustering for the demographic features

PCA & K-means

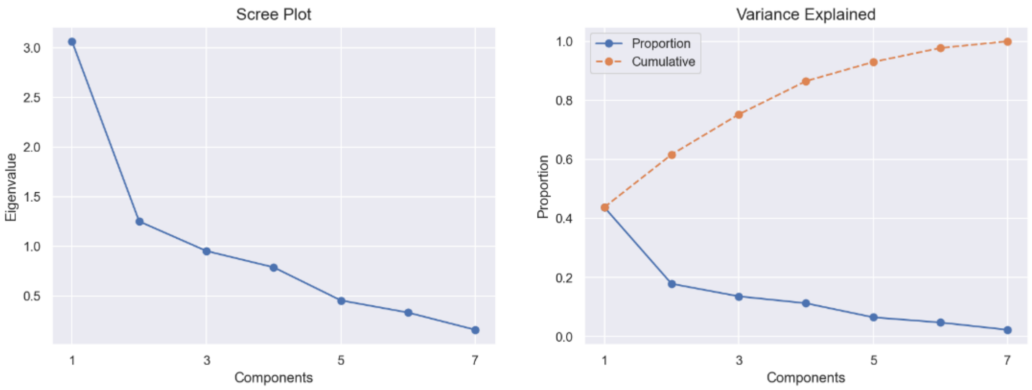


Figure 6- Scree Plot and Variance Explained plot for data_demo

After analysing the graphs above, we decided to keep 4 principal components due to the fact that more than 80% of the variance was explained with those PC. Applying the K-means on top of PCA gave us the R^2 scores:

2 clusters	3 clusters	4 clusters	5 clusters	6 clusters	7 clusters	8 clusters	9 clusters
0.391535	0.499902	0.582002	0.625669	0.664182	0.691631	0.714592	0.735293

Self-organizing maps

Being our dataset big, it occupies a huge amount of memory when we do the hierarchical clustering, so we firstly applied the self-organizing maps to our data. This will reduce dimensionality, being possible to run cluster techniques faster and will allow us to also visualize the univariate distances of the data (2D view).

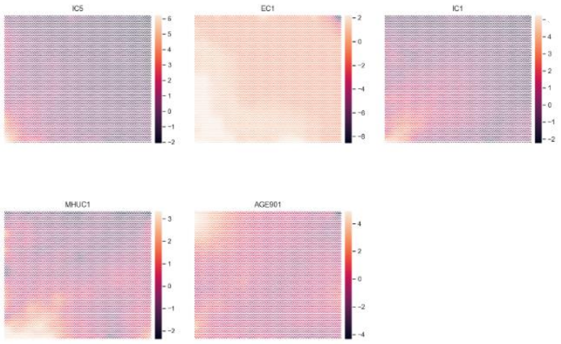


Figure 7- 2D View of Self Organizing Maps for data_demo

- **K-means and Hierarchical on top of Self-organizing maps**

To see the most appropriated approach, and the best linkage to use in the hierarchical clustering we computed the R^2 scores firstly and plotted them.

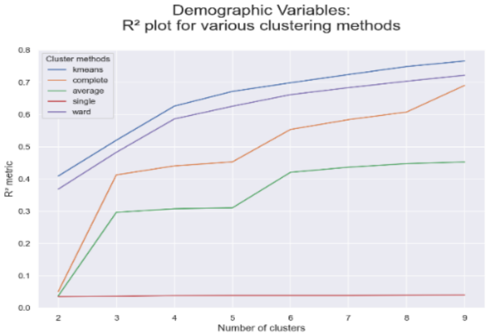


Figure 8- R2 Scores for different cluster methods- SOM

Seeing the plot, we can say that the best technique to use here is K-means, since it has the best scores in comparison to the others. The best linkage to apply to hierarchical clustering is the Ward linkage. The chosen number of clusters is unclear by only looking at this plot. To decide better we did an inertia plot for k-means.

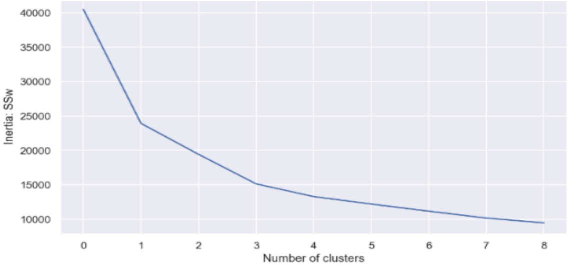


Figure 9- Inertia Plot over clusters – SOM of data_demo

Looking at the inertia plot, even though the elbow point is in the number 3, we were indecisive between 3 and 4 clusters. The best way to decide for us was to plot the average silhouette plot for both clusters and see the quality of the clusters of each decision.

Looking at the plots we can see that regardless the number of clusters we chose, there are always going to be negative points, meaning there are some values that are not perfectly fit for that cluster, and both are a bit unbalanced between clusters, meaning the size of the clusters are different. But, comparing one another, when we see the plot with 4 clusters, we can see that the clusters are more unbalanced than the plot with 3 clusters, and there are more negative points than the other one.

For that reasons, we decided that the final number of clusters chosen for K-means on top of self-organizing maps is 3. After applying the label to the data, that gave us a score of 0.4724.

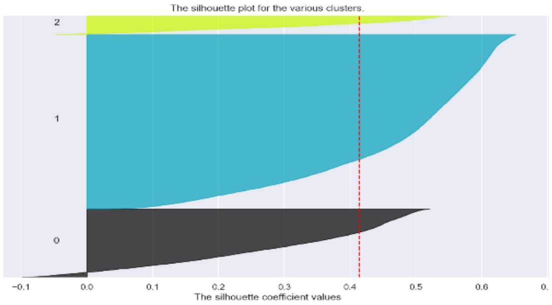


Figure 10- Average Silhouette Plot for 3 clusters - SOM

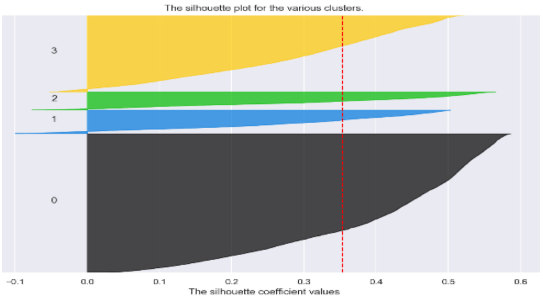


Figure 9- Average Silhouette Plot for 4 clusters- SOM

Hence, we also saw the Hierarchical clustering algorithm, to compare the results and to find similarities and patterns in the clusters:

After having generated the Dendrogram with Ward Linkage, we decided to draw the decision line at Euclidian distance = 90, meaning the point where donors are classified into 4 different clusters. Since the distance between 3 and 4 clusters is small, we checked the quality of them through an average silhouette plot.

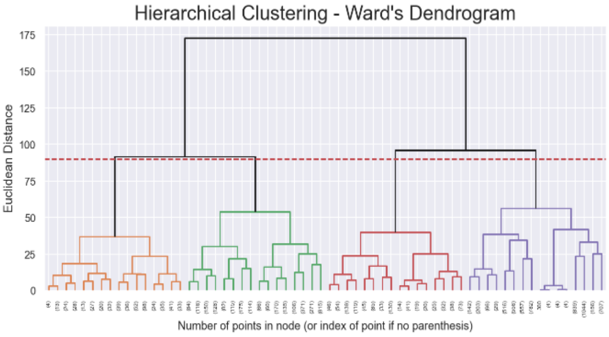


Figure 8- Ward's Hierarchical Clustering - SOM

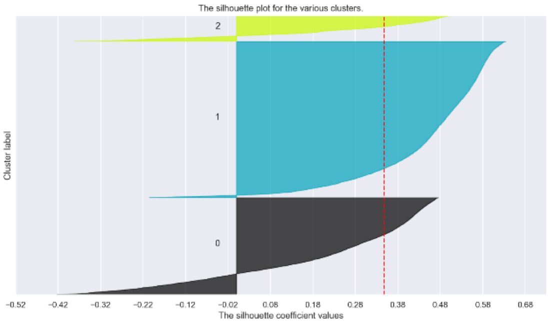


Figure 13- Average Silhouette plot for Wards Hierarchical for 3 clusters

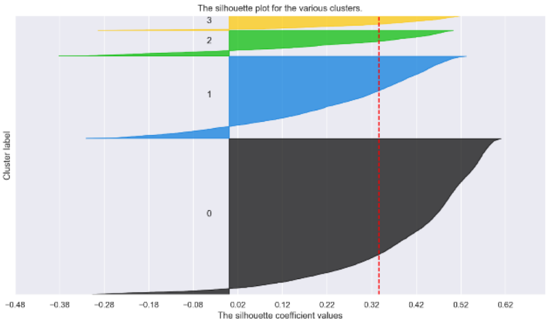


Figure 1411-Average Silhouette plot for Wards Hierarchical for 4 clusters

We can see that both have a lot of negative points, that is, there are a lot of points that are not perfectly fit for the cluster and all the clusters are unbalanced, with different sizes, being this method not very good for our data set.

K-means

Applying k-means without self-organizing maps on, we obtained these scores, and inertia plot:

2 clusters	3 clusters	4 clusters	5 clusters	6 clusters	7 clusters	8 clusters	9 clusters
0.378444	0.478616	0.575083	0.626340	0.665023	0.692975	0.714774	0.731008

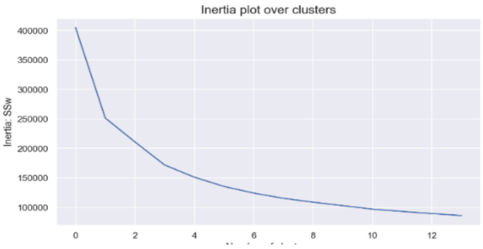


Figure 12- Inertia plot over clusters

Having these results, we can assume the best number of clusters for this method is 4, due to the fact that the R^2 scores started to stabilize from the number 4 and the elbow point is near 4 clusters. Still, to check the quality of those clusters we compared the average silhouette plot for 3 and 4 clusters.

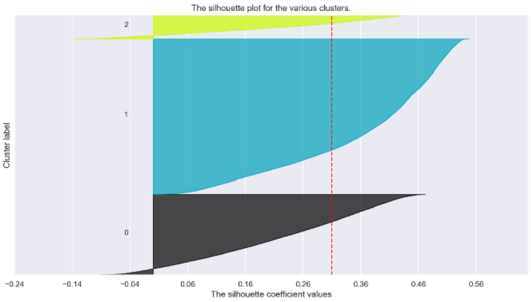


Figure 13- Average silhouette plot for 3 clusters

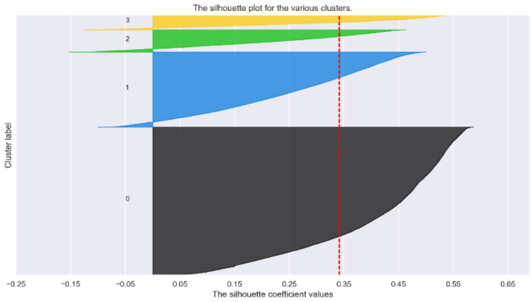


Figure 14-Average silhouette plot for 4 clusters

After looking at the plots we can see that the 3 clusters have more quality, since there are fewer negative points, meaning less values that do not perfectly fit the cluster. So, for this method, equally to the k-means on top of SOM, the optimum number for clusters is 3.

DBSCAN: Firstly, before applying the algorithm on our demographic features we checked the Nearest Neighbors to choose the optimal value for epsilon hyperparameter, giving us the estimated value of 0.75. Then applying the DBSCAN method, it gave us 3 estimated clusters, meaning 2 clusters and a group of outliers. Removing that group of assumed outliers, the R^2 score calculated for this method was 0.18. We can already see that is a very low number for clusters for this data.

Mean Shift: After exploring the values for the hyperparameters for this method, which means applying estimate_bandwidth to the data_demo and calculate the value for bandwidth, we applied it to our data set with the rest of the parameters bin_seeding=True, n_jobs=4. That gave us the estimated clusters of 29, with the value for the R^2 score of 0.4655. We can already see that is too many clusters for this data.

GMM (Gaussian Mixture Model): First we selected the number of components based on AIC and BIC. For k= 4, 5, 6, 7 we obtained the R^2 scores: 0.3322, 0.4245, 0.4590, 0.4886.

Decision: After applying all the clustering methods and comparing the efficiencies of the techniques used as well as the quality of the clusters, we have decided to choose the method **K-means** with 3 clusters for demographic features.

Clustering for the gift features

PCA & K-means

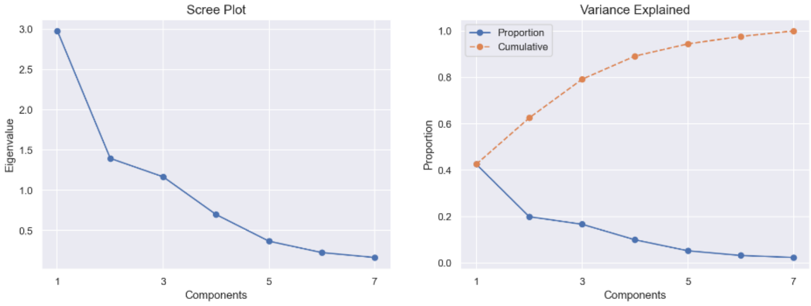


Figure 15- Scree Plot and Variance Explained for PCA

After analysing the graphs, we decided to keep 3 principal components since more than 80% of the variance was explained with those PC. Applying the K-means on top of PCA gave us the R^2 scores:

2 clusters	3 clusters	4 clusters	5 clusters	6 clusters	7 clusters	8 clusters	9 clusters
0.349053	0.473124	0.546759	0.602949	0.647931	0.689412	0.719818	0.745060

Self-organizing maps (Grid 100x100)

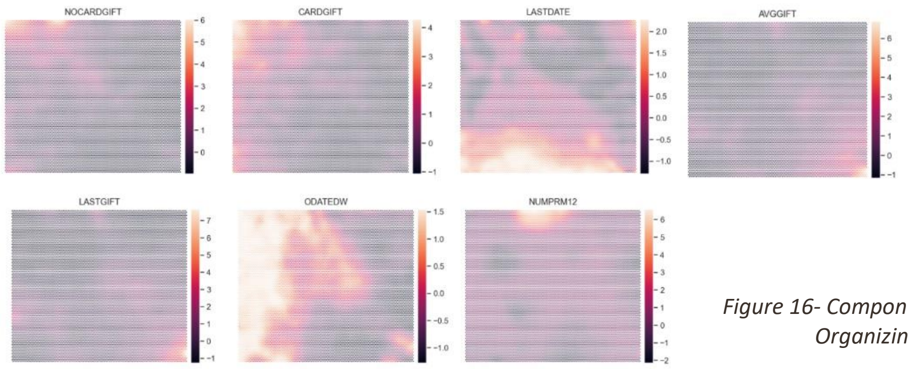


Figure 16- Component planes of Self Organizing Maps

- **K-means and Hierarchical on top of Self-organizing maps**

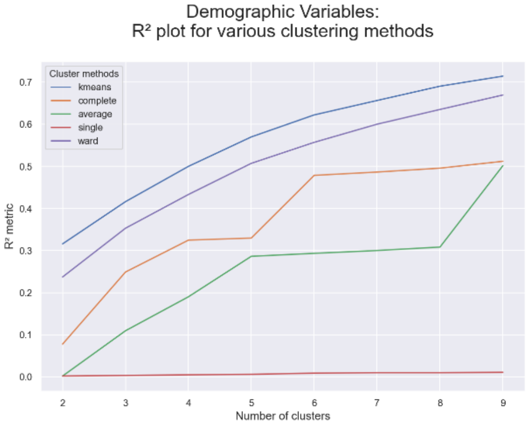


Figure 17- R2 Scores for different cluster methods- SOM

Following the same logic as data_demo, we computed the R^2 scores firstly and plotted them. Seeing the plot, equally to the data_demo, we can say that the best technique to use here is K-means, since it has the best scores in comparison to the others. The best linkage to apply to hierarchical clustering is the Ward linkage. The chosen number of clusters is unclear by only looking at this plot. To decide better we did an inertia plot for k-means.

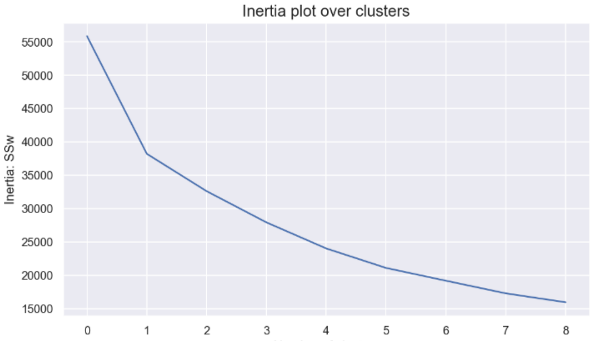


Figure 18- Inertia Plot

Looking at the inertia plot, we cannot easily see the elbow point. But combining this plot and the R^2 scores we guessed the optimum number for cluster was between 4 and 5

Once again, we also saw the dendrogram for the ward's linkage:

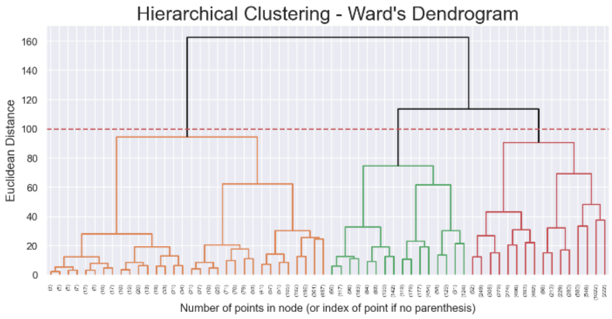


Figure 19- Ward's Dendrogram on top of SOM for data_gift

We decided to draw the decision line at 100 at the beginning, meaning the point where donors are classified into 3 different clusters. But since the distance between 3, 4 and 5 clusters are very low, we compared the average silhouette plots (see Appendix) and considering all the criteria we have been following we have decided to choose 5 for the number of clusters for this method.

K-means

Now that we know k-means is a good method we applied k-means without self-organizing maps on top and obtained these scores:

2 clusters	3 clusters	4 clusters	5 clusters	6 clusters	7 clusters	8 clusters	9 clusters
0.277628	0.376985	0.447573	0.512556	0.560108	0.597520	0.626467	0.650164

And this inertia plot:

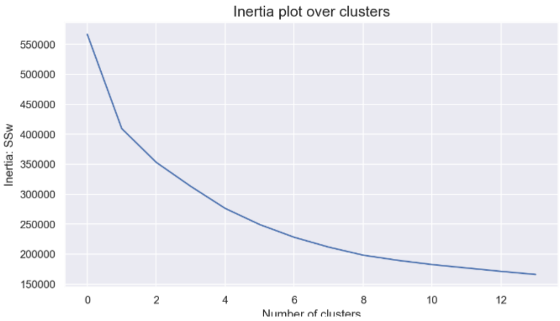


Figure 20- Inertia plot for data_gift

Even though the elbow point is not well defined, looking at the R^2 scores we were between 4 and 5 clusters. To help us decide we compared the average silhouette plot for 3 and 4 clusters.

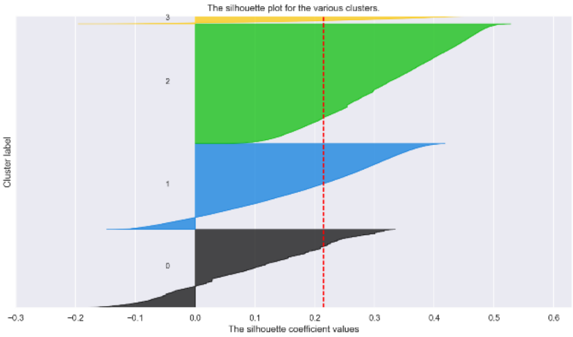


Figure 24-Average silhouette plot for 4 clusters

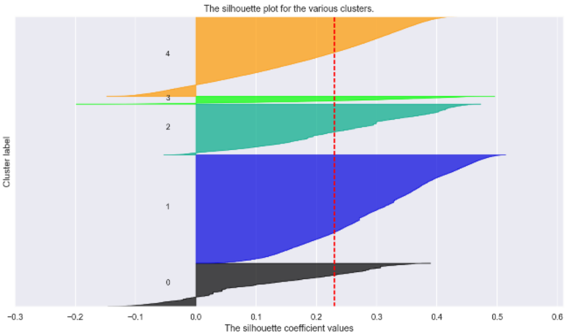


Figure 25- Average silhouette plot for 5 clusters

Looking at the plots, none of them are great. However, we can see that the 4 clusters have more quality, since the clusters are a bit more balanced. So, for this method, the optimum number of clusters is 4.

DBSCAN: Firstly, before applying the algorithm on our demographic features we checked the Nearest Neighbors to choose the optimal values for epsilon hyperparameter, giving us the estimated value of 1.1. Then applying the DBSCAN method, it gave us 3 estimated clusters, meaning two clusters with the R^2 score of 0.1644 and 2374 values identified as outliers.

Mean Shift: after exploring the values for the hyperparameters for this method, which means applying estimate_bandwidth to the data_gift and calculate the value for bandwidth, we applied it to our data set with the rest of the parameters bin_seeding=True, n_jobs=4. That gave us the estimated clusters of 134, with the values for the R^2 score of 0.3824. We can already see that is too many clusters for this data.

GMM (Gaussian Mixture Model): First we selected the number of components based on AIC and BIC. For k= 4, 5, we obtained the R^2 scores: 0.2736, 0.3524.

Decision: for this group of features, as well as in the demographic features, we have considered all the previous analysis of possible clustering methods and decided to choose the method **K-means** with 4 clusters for gift features. We have decided not to use PCA for dimensionality reduction in neither the perspectives because the scores are not significantly higher than the k means without PCA, and it would be a lot harder to interpret the clusters.

Merging the perspectives

To merge the clusters perspectives, we used the hierarchical clustering to combine the clusters centroids that are very close to each other, doing so the merge into one.

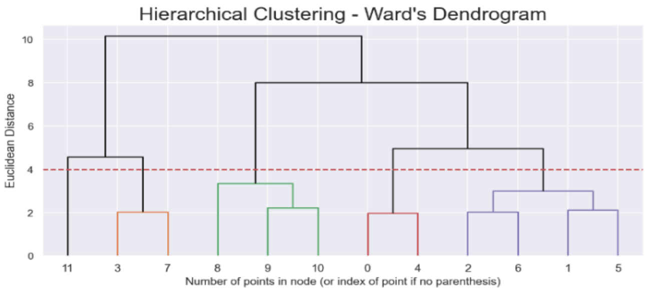


Figure 21- Ward Dendrogram for merging the perspectives

Interpreting the plot, we have decided to draw the decision line at the value 4, meaning the donors are classified into 5 different clusters. We have concluded that Euclidian distances bigger than 4 would lead us to a group of clusters with centroids that are not close to each other. After that we preceded to label the values from our data.

Reclassify Outliers

We used the Decision Tree method (see Appendix) to assign the outliers to a cluster. This method allowed us to know that in average, we were able to predict 89.27% of the customers correctly.

Cluster analysis

From now on we are considering the numerical order beginning at 1 (cluster zero in the plots is cluster 1), ending at 5 (cluster 4 in the plots is cluster 5).

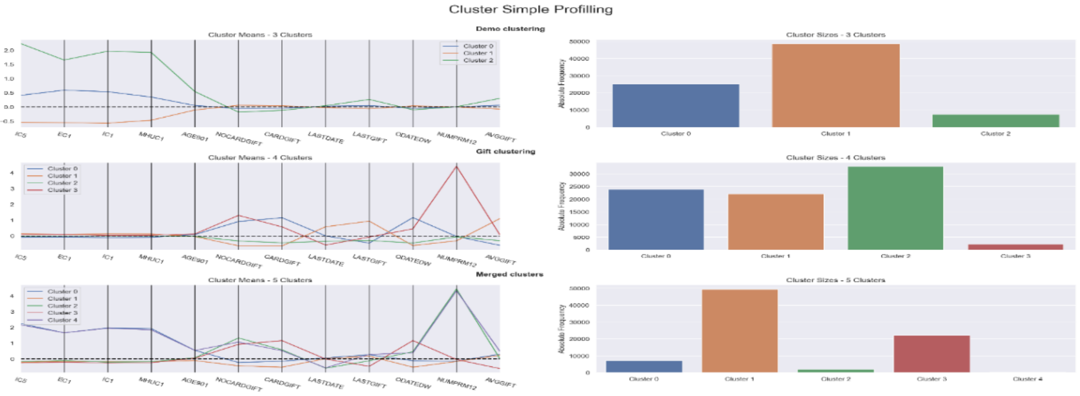


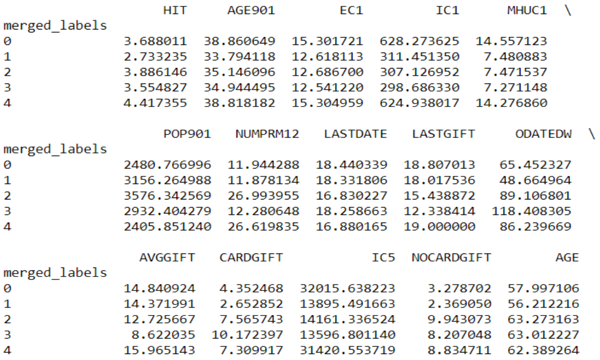
Figure 22 - Cluster Profiling

We plotted the cluster profiling to see the main differences between clusters, therefore we firstly calculated just the average between them to analyse that. We can see that in the merged clusters, in the demo features, there are two groups, one with cluster 1 and 5 and the other with clusters 2, 3 and 4. However in the gift features we can see that these groups mix themselves, forming groups of clusters that change throughout the features.

To see the features that most define the clusters we estimated the proportion of variance explained for each feature and sorted the values in descending order, giving us the features with the highest variance explained:

NUM PRM12	CARDGIFT	ODATEDW	IC5	NOCARD GIFT	IC1	MHUC1	EC1	AVG GIFT
0.553806	0.54073	0.54012	0.4977	0.4002	0.3850	0.3670	0.2723	0.1437

The number of promotions received in the last 12 months, the promotions donations and the neighbourhood income per capita, are the features which better define the different clusters. On the other hand, we were surprised to see that the age of the donors and the age of the neighbourhood's donors are features that do not contribute to the definition of clusters, being the proportion of variance explained 0.043747, 0.033471, respectively. When we verify the mean, it supports that statement, since the clusters have all a very similar age range.



```
                  HIT     AGE901        EC1         IC1      MHUC1  \
merged_labels
0            3.688011  38.860649  15.301721  628.273625  14.557123
1            2.733235  33.794118  12.618113  311.451350   7.480883
2            3.886146  35.146096  12.686700  307.126952   7.471537
3            3.554827  34.944495  12.541220  298.686330   7.271148
4            4.417355  38.818182  15.304959  624.938017  14.276860

                  POP901   NUMPRM12   LASTDATE   LASTGIFT     ODATEDW  \
merged_labels
0            2480.766996  11.944288  18.440339  18.807013   65.452327
1            3156.264988  11.878134  18.331806  18.017536   48.664964
2            3576.342569  26.993955  16.830227  15.438872   89.106801
3            2932.404279  12.280648  18.258663  12.338414  118.408305
4            2405.851240  26.619835  16.880165  19.000000   86.239669

                 AVGGIFT   CARDGIFT           IC5  NOCARDGIFT        AGE
merged_labels
0              14.840924   4.352468  32015.638223    3.278702  57.997106
1              14.371991   2.652852  13895.491663    2.369050  56.212216
2              12.725667   7.565743  14161.336524    9.943073  63.273163
3               8.622035  10.172397  13596.801140    8.207048  63.012227
4              15.965143   7.309917  31420.553719    8.834711  62.389264
```

Figure 23- Mean of the clusters of some Metric features

Cluster descriptions for marketing approaches

Based on the analyse of each cluster through plots for each feature, checking the distribution, the means, and others descriptive statistics (see Appendix), we sorted our descriptions of the clusters in this table:

	CLUSTER 1	**CLUSTER 2**	**CLUSTER 3**	**CLUSTER 4**	**CLUSTER 5**
NEIGHBORHOOD PER CAPITA INCOME	>$20000	<$20000	<$20000	<$20000	>$20000
YEARS OF EDUCATION	>=14	<14	<14	<14	>=14
NEIGHBORHOOD MEDIAN INCOME	>$45,000	<$45,000	<$45,000	<$45,000	>$45,000
NEIGHBORHOOD MEDIAN HOMEOWNER COST WITH MORTGAGE PER MONTH	>$1,000	<$1,000	<$1,000	<$1,000	>$1,000
RANDOM DONATION ENGAGEMENT	Low	Low	High	High	Low
ENGAGEMENT WITH CARD PROMOTIONS	Low	Low	High	High	High
LAST GIFT VALUE	> $10	> $10	< $10	< $10	> $10
TYPE OF DONOR ACCORDING TO THE FIRST DONATION	Undefined	Recent	Old	Old	Old
PROMOTIONS RECEIVED IN LAST 12 MONTHS	Few	Few	Many	Few	Many
SOCIAL ECONOMIC STATUS	High	Average	Average	Average low	High
URBANICITY (MOSTLY)	Suburban	Small town, rural	Undefined	Small town, rural	Suburban
DONOR'S AGE GROUP	Senior	Young	Young	Young	Middle age
NUMBER OF INDIVIDUALS	7090	49583	1985	22106	242
PROMOTIONS THAT THE DONOR ENGAGED	GK and XK	GK and XK	XK	NK and XK	GK and XK

Cluster 1:

Characterizes donors who live in an upper middle educated neighbourhood, mostly suburban. It contains middle age to senior people. When it comes to promotions, these donors do not usually engage card promotions nor random donations, but when they do it's a high value donation. The most answered promotions are GK, general greeting cards, and XK, Christmas cards. These types of donors can either be recent or older, which means that it is not really defined how long they have been donors.

Marketing approach: Since suburban areas are inhabited mostly by families, the marketing strategy should focus on the veterans and their families. As we referred above, these donors do not usually respond to mail promotions so a good approach would be the veterans participate in charitable events in the festive seasons.

Cluster 2:

This type of donor lives in a middle-class neighbourhood, where just a few finished high school. Lives in small towns or rural areas. Mostly composed by middle age people. As in cluster 1, this cluster aggregates donors who do not donate regularly. However, when they do, they give a considerable donation but not as high as the donors in the previous cluster. Also, these types of donors started to donate a few years ago. The most answered promotions are GK, general greeting cards, and XK, Christmas cards.

Marketing approach: Since we are talking about small towns and rural areas, we considered a possible approach going door to door with a focus on reintroducing the lapsed donors into our association.

Cluster 3:

Based on this cluster we can say that the donor's neighbourhood is constituted by middle-class people and just a few finished high school. The urbanicity is not well defined. They donate somewhat frequently with average values for both card promotions and random donations. These people started donating for quite some years, more years than the previous cluster.

The most answered promotion is XK, Christmas cards.

Marketing approach: As these donors don't respond to the many promotions that they received in the last 12 months; the association should stop sending so many cards. Instead, the association could reach out to the donors on their birthdays or special occasions.

Cluster 4:

The donors who belong to this cluster are the ones who live in a middle-class neighbourhood, where just a few finished high school. Lives in small towns or rural areas. This cluster contains the people who made their first donation the longest time ago. They give small value gifts but more frequently.

The most answered promotions are NK, blank cards, and XK, Christmas cards.

Marketing approach: Considering that this cluster contains lapsed donors who haven't been donating for the longest time the association should contact them and informed how their donation helped a veteran.

Cluster 5:

Profiles the donors who live in an upper middle educated neighbourhood, mostly suburban. The random donation engagement is way low when compared to the card promotions. The donations are frequent with relatively high value. These people started donating for quite some years. The most answered promotions are GK, general greeting cards, and XK, Christmas cards.

Marketing approach: As it happens in the cluster 3, these donors do not respond to the many promotions that they receive so the marketing strategy should be the same (decrease the number of promotions they send). In order to reactivate the lapsed donors, the association should invite them to an event or other cultivation opportunity. Although it is a cluster with low number of individuals, we can say that is an important cluster to reenforce to donate again due to the fact that this group of people donate high values frequently.

Conclusion

After analysing, cleaning and applying cluster techniques to the provided data set we came across a final costumer segmentation in such a way that is now possible to PVA to better understand all the donor's behaviour and identify the potential donors within their data base.

The cluster which the association should be more focused on is the number 4 because comparing this to the other ones it has a high number of individuals who donate frequently, even though it is the group with the lowest income per capita.

Based on our analysis of the donor's behaviour we can establish that one general smart approach would be to focus on the awareness of the PVA mission on the holiday's seasons, since it is the type of promotion that most of the donors, no matter the cluster, engage.

Now that the different clusters are made, the PVA can make more personalised approaches to recall all the lapsed donors. That can make the person itself feel that they are making a difference based on their unique actions, and the PVA is thankful for that. Gratitude will make the donor feel the need to help again.

References

- Rençberoğlu, Emre (Apr 1, 2019) ***Fundamental Techniques of Feature Engineering for Machine Learning***- https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114
- Campoy, Pascual (June 14-19, 2009). ***Dimensionality Reduction by Self Organizing Maps that preserve distances in Output Space***
- @ankurtripathi (16 May 2019). ***Mean-Shift Clustering*** - https://www.geeksforgeeks.org/ml-mean-shift-clustering/
- Mishra, Sanatan (May 19, 2017). ***Unsupervised Learning and Data Clustering*** - https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a
- ***Data Mining Practical and theoretical classes***
- ***Scikit learn library Documentation***

Appendix

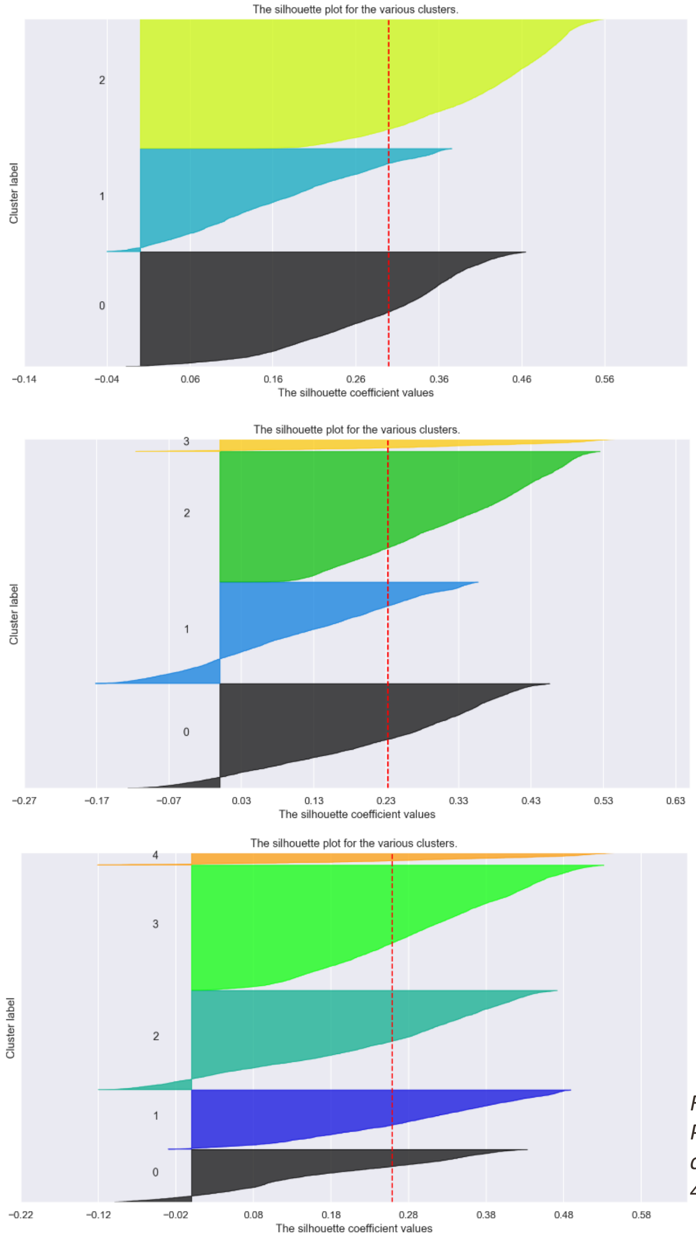


Figure 24- Average Silhouette Plots for Hierarchical clustering for data_gift (for 3, 4 and 5 clusters)

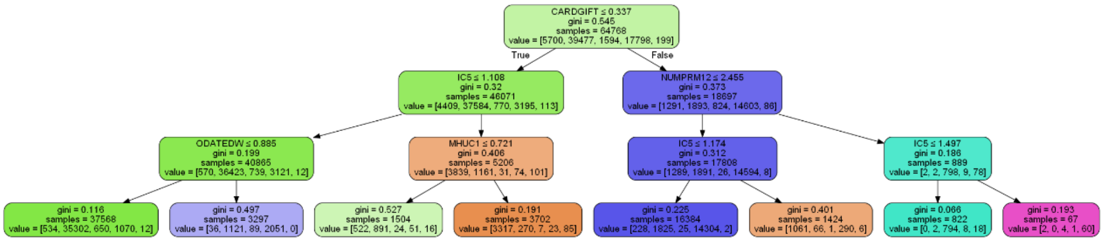


Figure 25- Decision Tree to Reclassify the Outliers

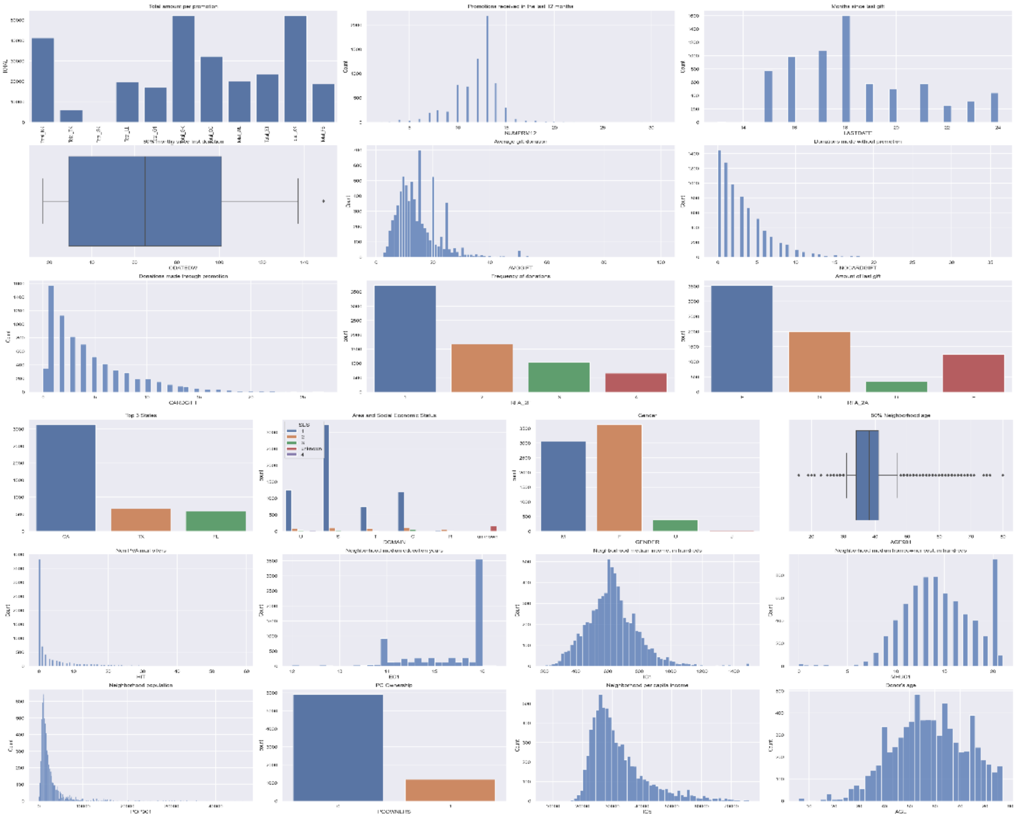


Figure 26- Descriptive plots cluster 1

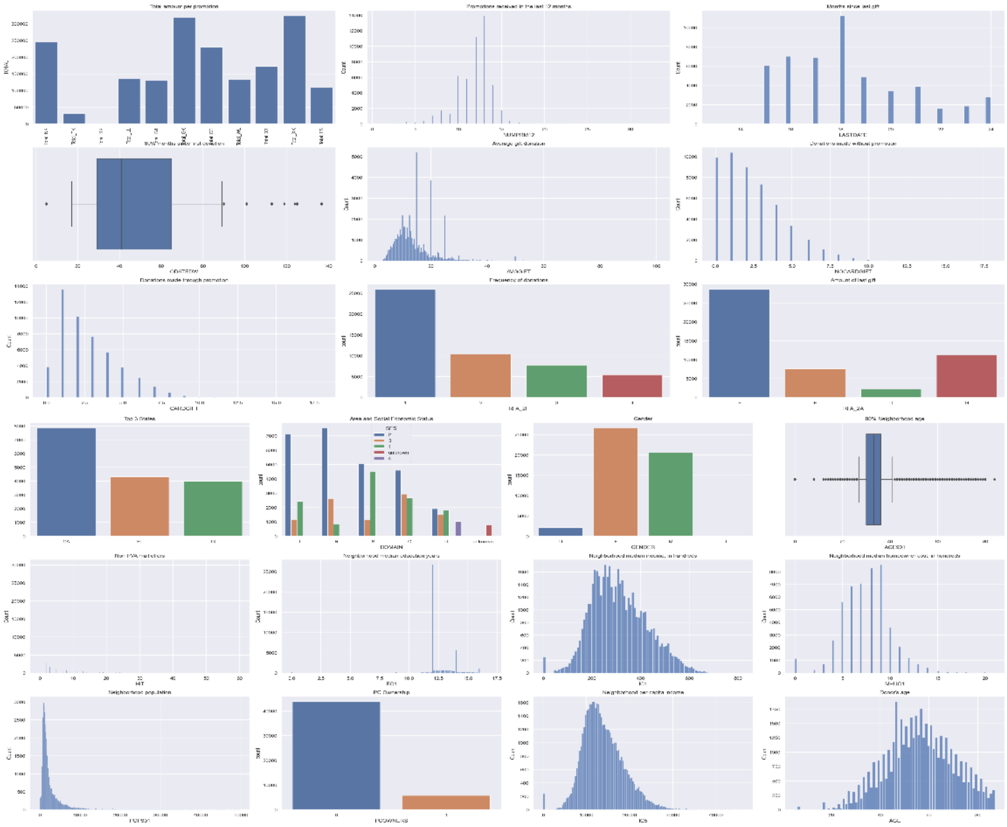


Figure 27- Descriptive plots Cluster 2

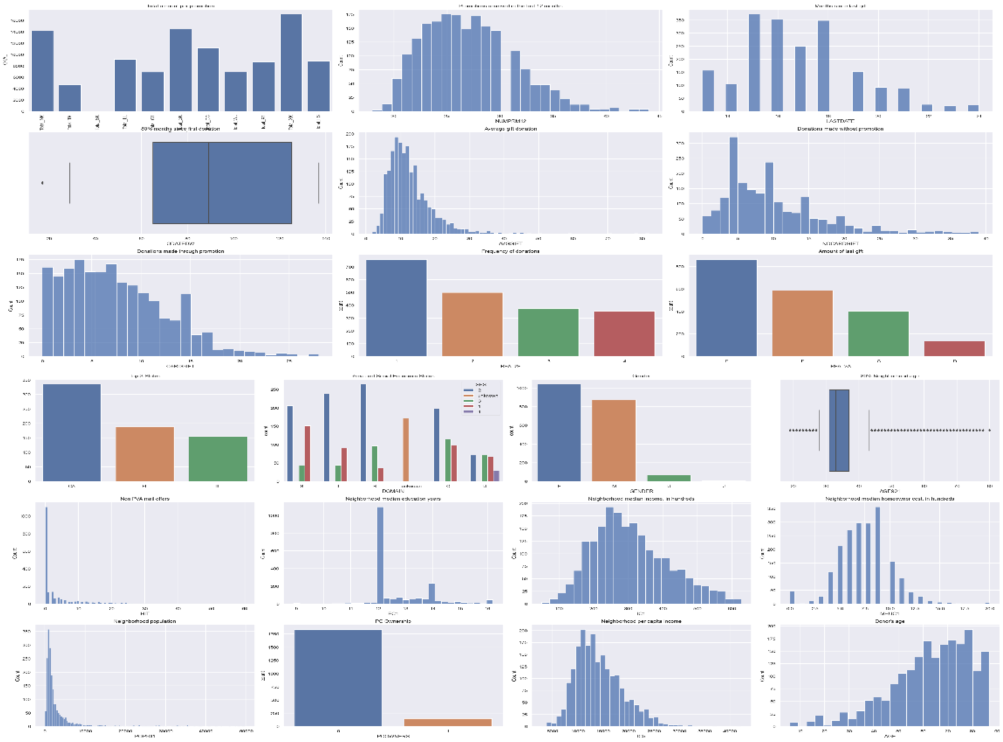


Figure 28- Descriptive Plots Cluster 3

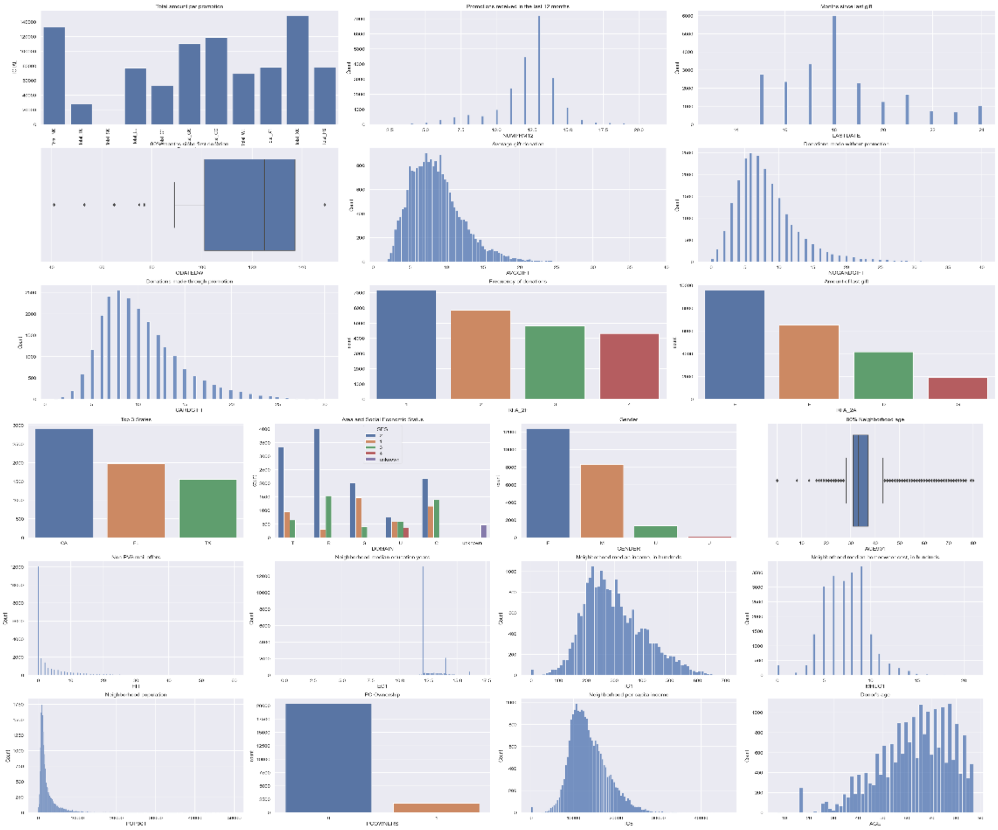


Figure 29- Descriptive Plots Cluster 4

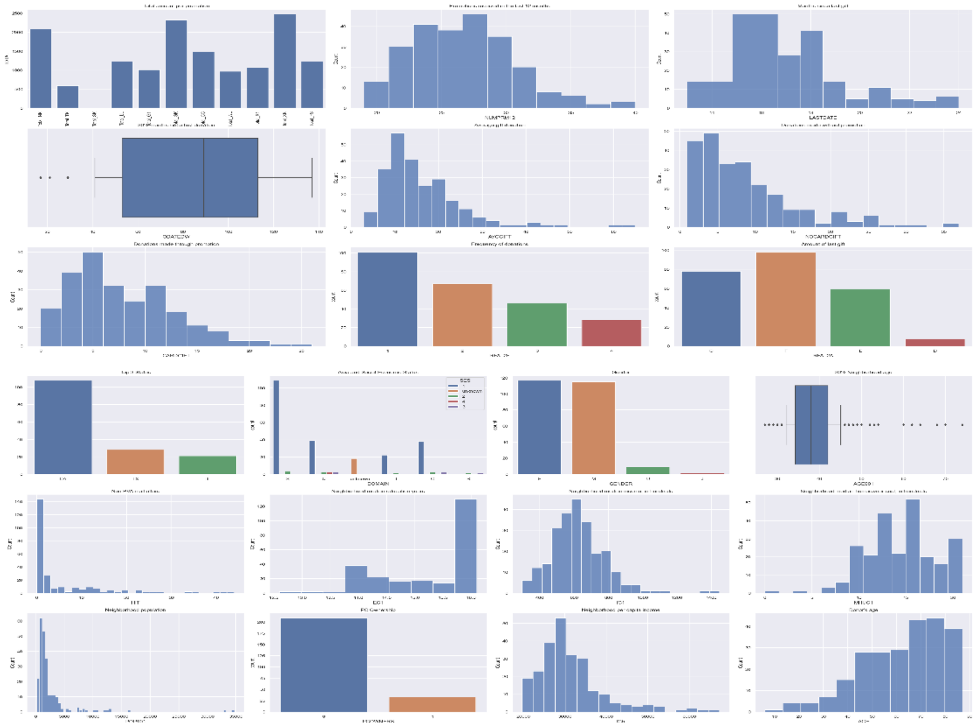


Figure 30- Descriptive Plots Cluster 5