Data Mining

Group Project Report

Source: https://www.thibautderoux.com/

Group I

Davide Farinati (m20201080@novaims.unl.pt)

Diana Furtado (m20200590@novaims.unl.pt)

Hiromi Nakashima (m20201025@novaims.unl.pt)

1st Semester 2020/2021

TABLE OF CONTENTS

INTRODUCTION

We were challenged to analyze a dataset provided by the Paralyzed Veterans of America (PVA), consisting of a sample of the results of one of its recent fundraising appeals, in order to develop a customer segmentation to help PVA understand and forecast donors behaviour and perceive potential donation opportunities, producing a marketing report.

This customer segmentation is made by separating the donors into groups (clusters) based on the characteristics they share.

This report intends to go through all the process and reasoning behind the decisions made both on the data cleaning and preprocessing and on the clusters' definition and interpretation; ending with the marketing approach recommended for each cluster defined.

The GitHub repository where all the present analysis is saved can be accessed through the following link: https://github.com/Fari98/DataMiningProject

DATA ANALYSIS

The data preprocessing phase is perhaps the most crucial one the data mining process [1]. This process, while working with "real-world" projects, defines the quality of data mining results [2]. For good measure, we have gathered efforts to do a deep analysis and improve the given data.

In the first moment we analysed the data to provide a better understanding. At this step we have done an initial feature selection, in order to reduce the number of the features and make possible a better analysis. Firstly we looked at missing values columns, we've considered the columns with space (' ') in the cells were considered also missing values, then with this 111 columns were dropped.

Secondly, we identified the correlation between the variables. As references [3] and taught in class, discarding the variables that either highly or lowly correlated with each other. To make the process more efficient, two functions were created. The *auto_feat_selection_hcorrelation* checks, for each pair of variables, if the correlation between them is higher than 0.9 or lower than -0.9, in which case it selects one of the two variables randomly and drops it. The *auto_feat_selection_lcorrelation* function counts how many times a column has a correlation under 0.1 compared to the other columns.

To better cluster analysis, we decided to split the data in categories, more specifically: demographic, gift promotions and military. Then delete the data that had more than 90% of 0.1 correlation or less. The final result of the cleaning process into Table 1.

Category	Number of features
Demographic	268
Gift Promotion	14
Military	15

Table 1- Table with the division of number of features in each category.

As the table shows, the demographic category demanded additional feature selection processing, therefore we have applied LassoCV Regression, a L1-norm regularization [4]. Lasso works especially well when dealing with a large number of features as it shrinks the less important feature's coefficient to zero, removing them together [5]. In sum, what this analysis does is to first, count how many times a variable is considered not significant for predicting other variables (where the coefficient is zero) and, second, shows the most relevant variables for this prediction purpose (higher coefficient). Even with the great results from the other categories, we applied this method in all to confirm our choice.

After logic reduction of the features, the dataset still remains with 74 metric variables and 53 non metric variables. Therefore, to perform a better cleaning of the data, we looked more deeply into each column individually and the metadata to then make selection decisions based on meaningfulness and redundancy - to drop columns that will not be useful for the clustering analysis of the domain (PVA donors). Based on that, we have decided to change the CARDPM12 and NUMPRM12 for MAJOR and NUMPROM. The final result of feature reduction, the name of the columns were changed to better understand, and the data set was adjusted to 55 columns and 95412 rows.

Considering the first step of feature reduction, the next step was "Data Cleaning". Beginning by fixing the columns data types, in other words, convert the string dates to datetime type.

Define a clustering demand good understanding of the data, then we proceed by searching the data inconsistencies, duplicates values, *Non Lapsed Donors*, coherent donor. Through metadata, our data should be composed by Lapsed Donors, it means, only those which the donation were between 13 and 24 months ago. Also, to be a donor the person should be more than 0 days, then we subtract the *most recent donation* with the *donors_first_donation* and *data_first_donation* since both have the same meaning. We decided to keep only the *most recent donation* to maintain the data coherency. The final coherence analysis was to identify if the columns *data_last_promotion_mailed* and *data_most_recent_promotion_received* refer to the same promotion and the second one subsequent to the first one. We couldn't validate this coherence, we dropped the *date_most_recent_promotion_received*.

Before proceeding with the data cleaning we retrieve some metric features from the categorical data, like the time passed from the first and last donation and the age of every donor.

The data pre-process was based on [1] and [2]. The diagram below (Figure 1), summarize the steps of this process, which going to be described:

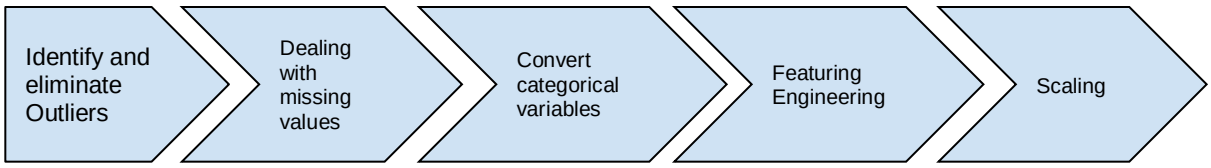


Figure 1- Flow of the pre-process step of data.

As shown in the diagram, firstly we worked with the outliers using the understanding of univariate and multivariate statistical analysis [6]. This logic was coherent to our project, since univariate analysis emphasizes description while multivariate, hypothesis testing and explanation, this last one can be very useful to a clustering analysis.

The outlier identification and elimination were done using three methods: Zscore, Test statistic of Median Absolute Deviation (MAD) and InterQuartile Range Method (IQR) [8][9]. The Zscore measures how many standard deviations away a data point is from the mean, which we used the threshold as 3. MAD, measures the median of the absolute deviations from the data's median, this method can't be applied in columns with 50% or more of the same values, with a threshold set to 3. And the least mentioned was IQR, which uses the extreme quantiles and the interquartile range, with the multiplier set to 3 (rather than the usual 1.5). The application of those methods was done through functions described in the notebook.

Using the three univariate methods described below, we opted to drop the 3,144 points (almost 4% of the dataset) considered as outliers by the three methods simultaneously.

After checking for univariate outliers, we checked for multivariate ones through three different analyses: the Local Outlier Factor (LOF), One-Class Support Vector Machine (SVM) and Isolation Forest [6][7].

The LOF detects outliers by comparing the density of the neighborhood of a point to the ones of its neighbours. [8]

The One-Class SVM is an unsupervised learning algorithm that is trained only on the 'normal' data, it learns the boundaries of these points and is therefore able to classify any points that lie outside the boundary as outliers. [9] In this algorithm, the expected percentage of outliers is defined in advance, in this case we have set it to the figure resulting from LOF (0.01157).

Finally, in the Isolation Forest, in order to isolate data points to detect anomalies. It recursively generates partitions on the sample by randomly selecting an attribute and then randomly selecting a split value for the attribute, between the minimum and maximum values allowed for that attribute. When all the trees are grown, outliers are identified as points easier to isolate, therefore with a smaller path length in the tree, being closer to the root. [10] Similarly to what happened for the One-Class SVM, in the Isolation Forest we have also set the percentage of points expected to be outliers to the one retrieved from LOF.

To better comprehend univariate and multivariate outliers, we reduced the data to 2 dimensions using PCA

and a contingency table.

Keeping the processing in the data cleaning diagram, we proceed treating the missing values. We used the K-nearest neighbor Imputer (KNN Imputer) to look at the distance as weight and considered the 100 neighbors and return the best value.

After we treated the categorical variables, by transforming them in binary variables. On this step, we created only two genders (female and male) and if it did not correspond to any it was treated as the joint account. Then converted the state code 4 territories values (NorthEast, West, MidWest, South, and if all 0s means other) and to finalize, we transformed the variables of *homeowner* and *major_donor_flag*. The same treatment was applied on the outliers dataset.

The feature engineering process was characterized only to combine some features that we had in the data set. Those columns are *donations_per_month*, *donations_per_promotions*, *avg_promotions_per_year*, *amount_donated_per_promotion* and *amount_donated_per_year*. The last step of data pre-processing was scaling the dataset using the Standard Scaler, because we intended to use algorithms based on euclidean distance.

At this stage, to avoid keeping redundant data, we looked again at the features' correlation matrix (Figure 2), considering all features with a positive or negative correlation stronger than 0.6. We have then decided to drop the 15 features listed below:

number_promotions_received, *tot_card_prom*, *max_donation_amnt*, *perc_househ_wfam*, *perc_housev_100h+*, *min_donation_amnt*, *percent_WW2_vet+*, *percent_vietnam_vet_census*, *perc_rural*, *perc_black*, *perc_male_mil*, *perc_males_vet*, *perc_local_gov*, *perc_state_gov*, *perc_fed_gov*.

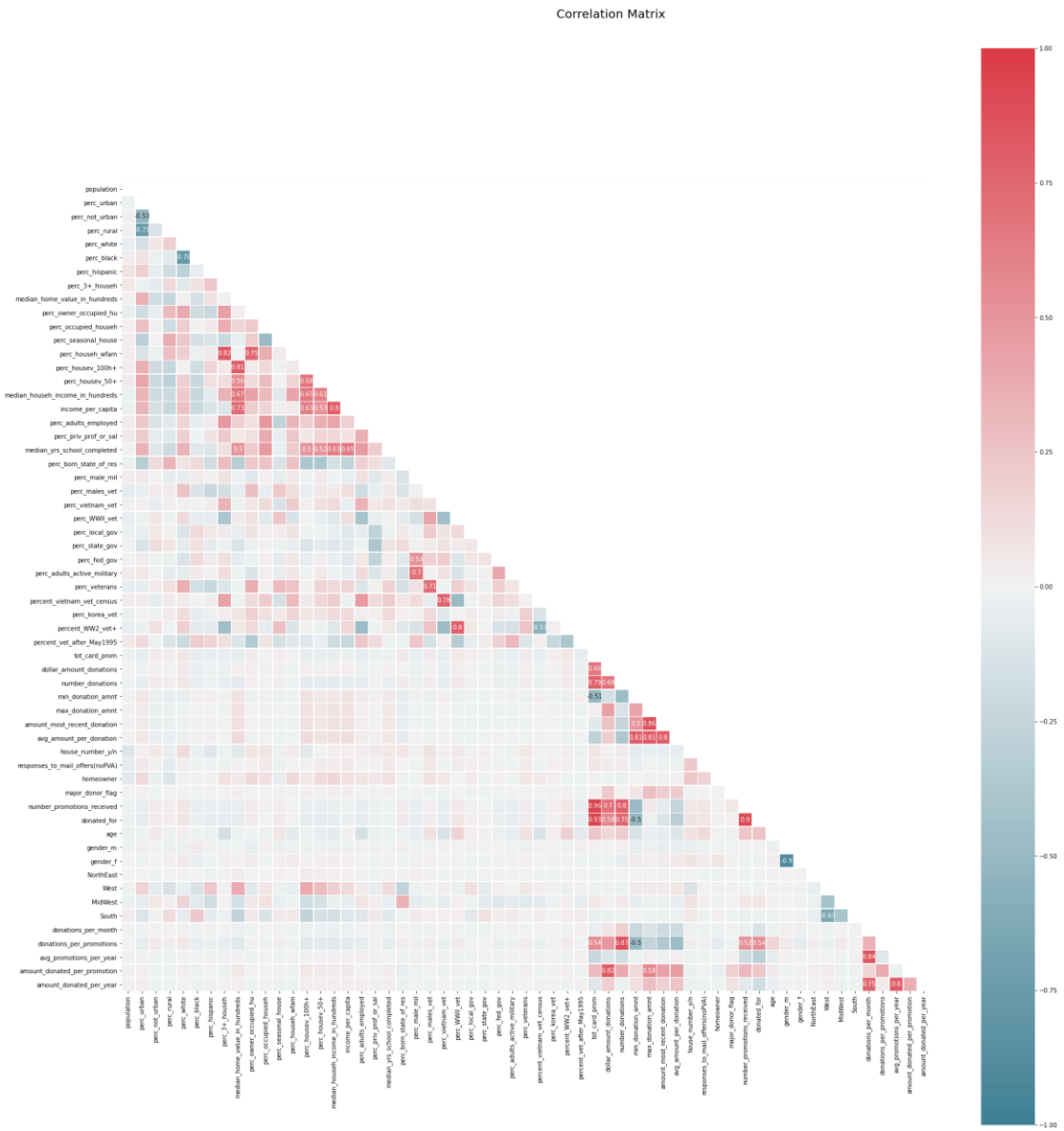


Figure 2 - Correlation matrix.

At last, we split the metric features into three different categories for the clustering:

- Gifts_promotions: including a total of 13 features;
- Military: including 6 features in total;
- Demographic: including the remaining 25 features.

Once again we have used the Standard Scaler to scale the dataset.

In *Appendix III - Clustering features brief description* there is a table containing the description of the final set of features for clustering.

K-means clustering

We started by testing the k-means clustering, being the most reliable method. In order to make that process more efficient we have defined a few functions which are explained below:

1. ***kmeans_elbows_plot*** - this function generates the optimum number of clusters and evaluates the quality of the clusters (that ideally should be well defined - distant enough from each other and with the points not spread within the cluster), through 3 different plots (Figure 3). The first one, Inertia plot, shows the dispersion of the points within the cluster, meaning a small inertia is the best outcome. The second plot analysed was the average silhouette plot that determines how well each object lies within its cluster, being the best outcome a higher number. At last, we have looked at the Davies-Bouldin score plot which index is based on a ratio between distances within the cluster and distances between clusters, the best outcome is the smallest index.

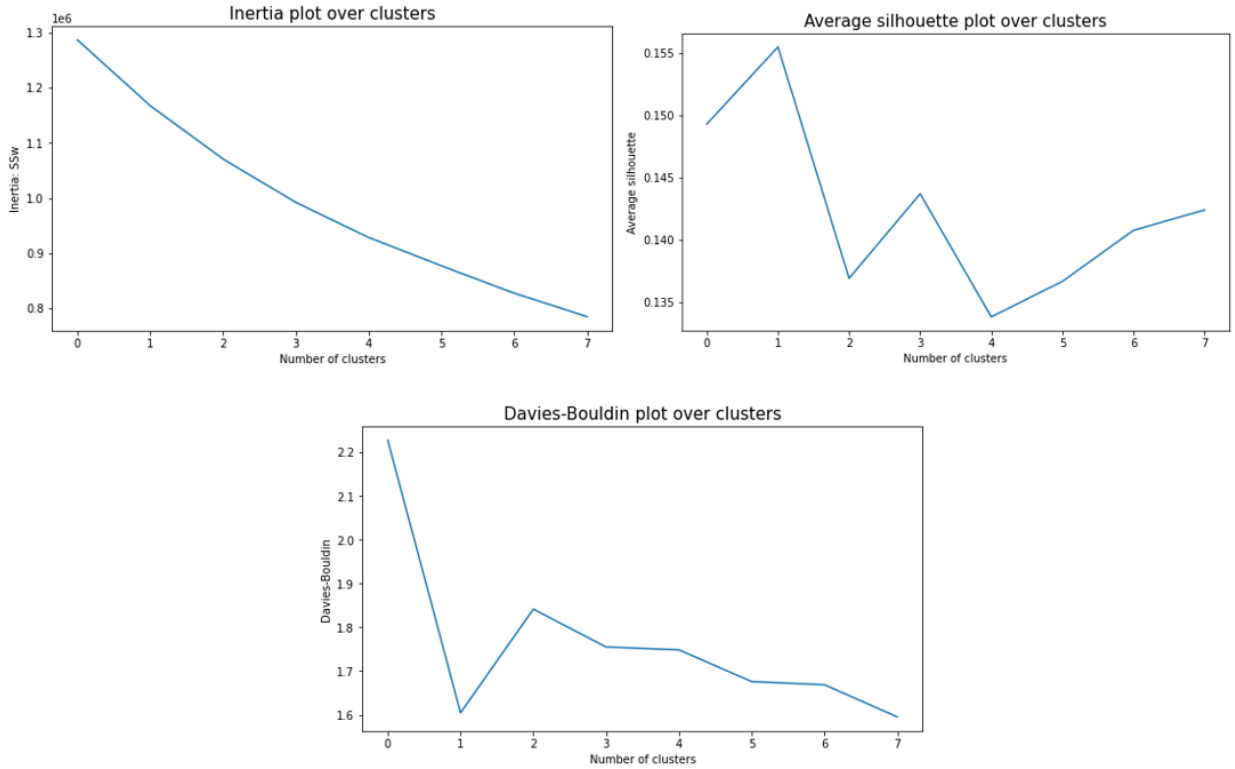


Figure 3 - Inertia, average silhouette and Davies-Bouldin plots example for k-means (Demographics).

2. ***complete_sh_score*** - This function produces silhouette plots to show the quality of the number of clusters defined by the previous function, the closest to -1 or 1 the figure is, the best quality the cluster has. An example of this plot is provided in Figure 4 below.

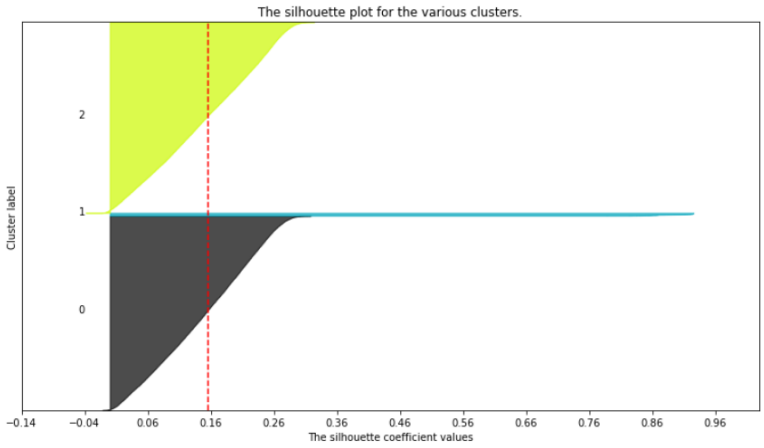


Figure 4 - Silhouette plot example for k-means (Demographics, 3 clusters).

3. ***clust_interpretation*** - This function characterizes the clusters by its features' means. The outcome of this function is a table with the mean of every feature for each cluster.
4. ***get_ss*** - This function returns the result of the sum of the r squared.
5. ***r2*** - This function measures how good is the approximation of the regression predictions and the real data points through the computation of the proportion of variance in the dependent variable that is predictable from the independent variable, the closer this value is to 1, the better.
6. ***cluster_profiles*** - For the chosen number of clusters, this function performs the cluster profile according to the clusters' features. The output, Figure 5, shows the mean for each feature in each cluster and the absolute frequency for each cluster. This helps understanding the consistency of the data. In the example provided in Figure 5, it shows that cluster 1 is very noisy and it can be considered as outliers and therefore it can be dropped.

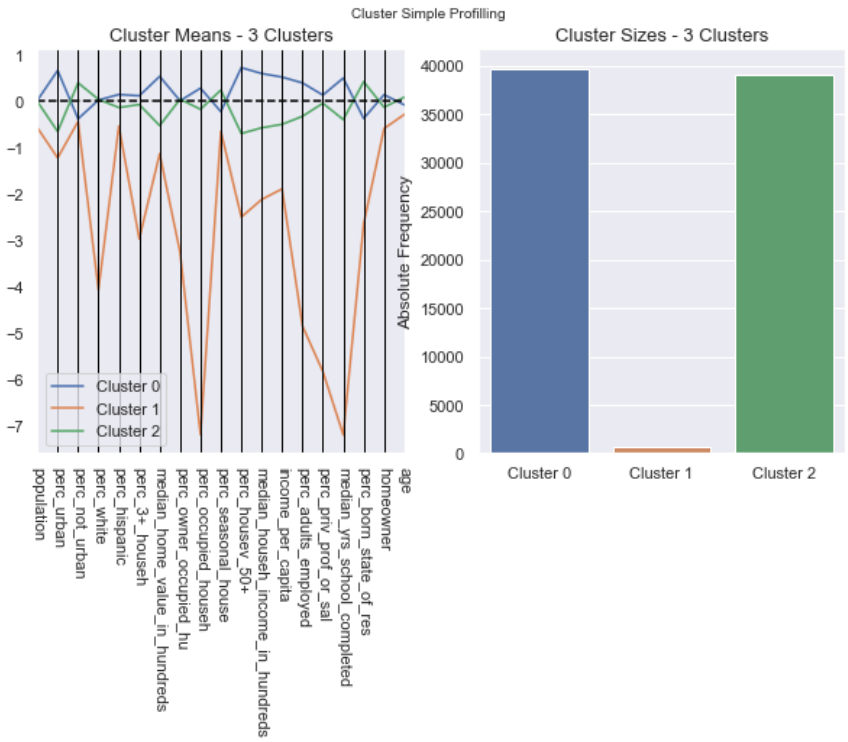


Figure 5 - Example of output of cluster_profiles function (Demographic, 3 clusters).

7. ***get_ss_variables*** and ***r2_variables*** - the first function computes the sum of squares to be used in the second function, to compute r squared for each variable.

The functions explained previously were applied to each of the three categories of variables defined for the dataset (Demographics, Military and Gifts_promotions). The results of the first category analysed (Figure 3, Figure 4 and Figure 5), demographic, we could identify the best number of clusters as 3, from the plots, from the silhouette plot.

Although, as previously mentioned, after running the *cluster_profiles* function, Figure 5, we could identify that the cluster numbered as 1 was not representative as it showed a lot of noise (inconsistent profile). This way we opted for keeping only the 2 remaining clusters (numbered as 0 and 2) for the demographic sub-dataset.

A similar analysis was made on the *Gifts & promotions* and *Military* sub-datasets. For the first category we got an optimum number of clusters, after plots' analysis, of 4 and decided to drop cluster 1 after looking at the clusters' profiles. For the second category, *Military*, the same happened: initially the optimum number of clusters was identified as 2 and, after looking at the clusters' profiles, we have decided to drop cluster 0 due to its noise.

Given the noise found in every category of the dataset, we found it important to revisit the outliers' cleaning, and repeat the clustering process with the newly clean dataset. The result was then 4 clusters for *Demographic* category, 3 clusters for *Gifs & Promotion* category and 3 clusters for *Military* category.

Hierarchical clustering

To improve the consistency of the analysis, we opted to try a different clustering methodology, hierarchical clustering, for the Military category. Similarly to what happened for K-Means, some new functions were created to optimize the clustering process:

1) ***get_r2_hc*** - Similarly to function ***r2***, this function computes the r2 for the hierarchical clustering.
2) ***hc_clust_method_eval*** - This functions selects which of the agglomerative hierarchical approach methods suits best (starting with 10 clusters): Ward's method, Single linkage, Complete linkage and Average linkage - varying between them on the way the points are agglomerated to create clusters. The number of clusters is chosen by looking at the output plot (Figure 6) and checking where there is the first big decrease on the R2 metric. In this case, for the Military category, we have considered 3 clusters.

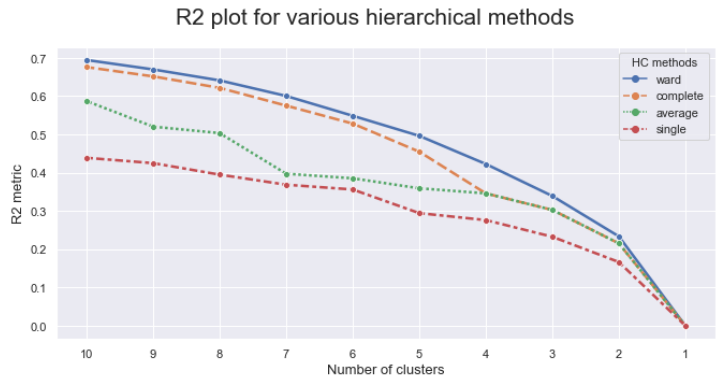


Figure 6 - Example of output from function hc_clust_method_eval for Military category.

3) ***full_tree*** and ***full_tree_visual*** - these functions produce the dendrogram of the hierarchical clustering analysis. An example using Ward's method to link the clusters is presented in Figure 7.

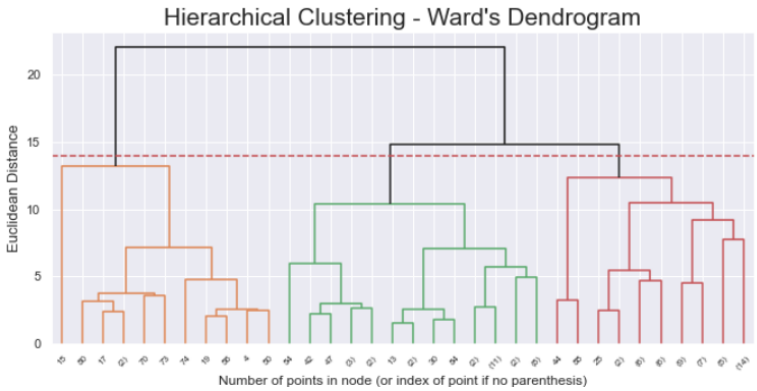


Figure 7 - Example of output from function full_tree_visual for Military category, with Ward's method used to link the clusters.

After applying these functions to the Military category, we've used the previously defined *clust_interpretation*, *cluster_profiles* and *r2* to evaluate the 3 clusters defined for this sub-dataset.

4) ***gmm_n_components*** - This function checks the best number of components and covariance type for the gaussian mixture algorithm, using the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC).

Since the K-means and hierarchical are the main types of classification, the first used with a predefined class and the second, unknown to help determine an optimal number [11]. We used the functions below to apply this cluster method on the military dataset.

Firstly, the *hc_clust_method_eval* returned the best hierarchical method to use in this dataset, which was *ward*. Doing the same steps as in the K-means process, the cluster_profiles return a similar cluster analysis which took us to apply a fuzzy clustering method, Gaussian Mixture Model. Figure 8 plots the best covariance versus the number of clusters which was 2.

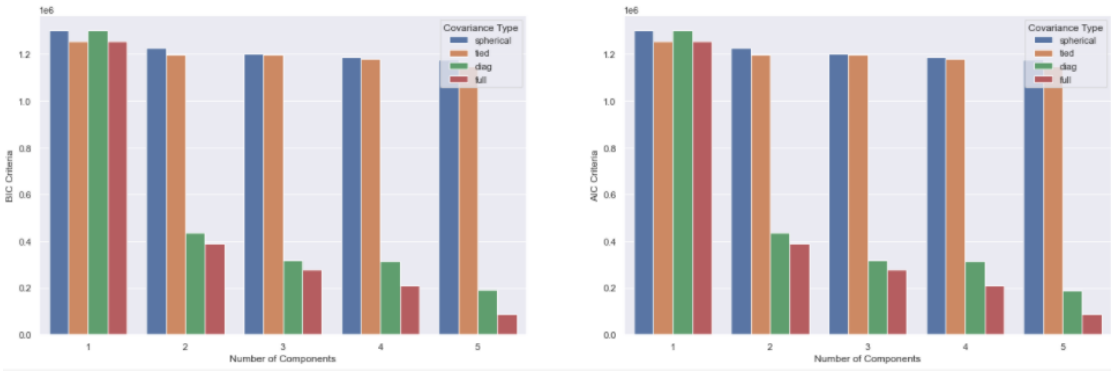


Figure 8 - Covariance versus the number of clusters.

After the classic clustering methods, we tried the Self Organizing Maps (SOM) with 100 units followed by the K-means on the fitted units. This returned a better solution of the clustering to the military group (Figure 9).

Figure 9 - Cluster profiling - Military.

Having all the groups clustered we create a contingency table to analyse the data points in each cluster category defined (demographic, gifts & promotions and military). We used this method to analyse and merge some clusters by a low number of elements to balance the data points in each cluster. Another merge method used was the hierarchical clustering similarity, by using the hierarchical tree distances, this returned 9 clusters. The characterization and interpretation of those 9 final clusters is detailed in the Results section.

RESULTS

In this section we present the cluster interpretation based on the individual clusters' means. On *Appendix I - Cluster means table* we present the means table and, to make the analysis more efficient we have produced some graphical visualisation of each cluster's means against the other clusters' means. On *Appendix II - Cluster means, graphical visualization*.

Figure 10 shows the centroid of the clusters after a reduction to a two-dimensional space, which was very helpful to understand similarities/ differences between clusters.

Analysing the bigger picture and the results achieved through this analysis, we were able to get some insights about the clusters which are demonstrated below.

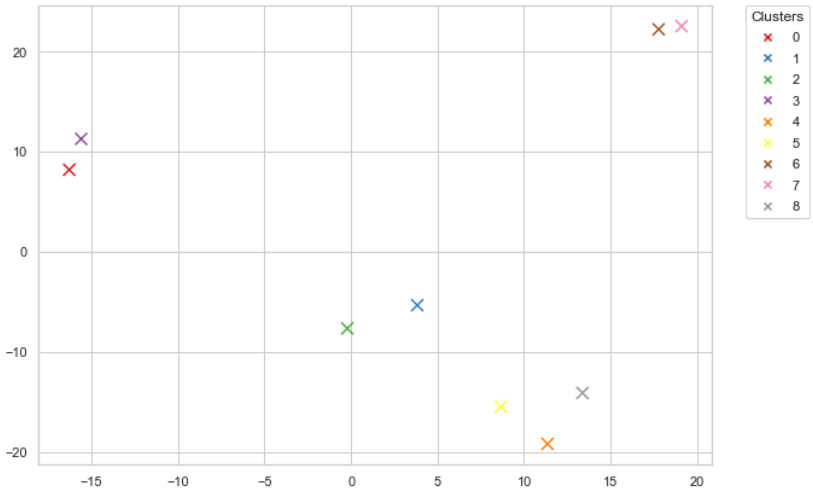


Figure 10 - Clusters' centroids location.

- ***Cluster 0***: This is the 2nd largest cluster and, demographically, it is the group with the 2ndhighestrate of population living in non urban areas. This is also mostly composed by white population and the 2nd least educated citizens. Military wise, this group has the 2nd lowest veterans percentage and it does not seem to be very responsive to mail offers. Other than that, in the Gifts & Promotions perspective, this group is average.

- ***Cluster 1***: Demographically, this cluster shows the lowest percentage of population living in non urban areas, the 2nd lowest percentage of white population (just over 84%) contrasting with the 2nd highest hispanic percentage. Regarding the Military indicators, this group has the highest percentage of adults active in the Military, mostly of Korea and Vietnam veterans. In terms of Gifts & Promotions this group does not show anything in particular that distinguishes it from the other clusters' average.

- ***Cluster 2***: Demographically and Gifts & Promotions side, this cluster is very similar to Cluster 1 described previously. However, Military it shows the highest percentage of veterans after May 1995.

- ***Cluster 3***: Both Demographic and Military indicators of this Cluster are similar to Cluster 0. Regarding Gifts & Promotions, this group shows the lowest dollar amount of donations and it is also the least responsive group to mail offers. Additionally, this is the 2nd youngest group.

- ***Cluster 4***: The main characteristic of this cluster is the high income per capita and highest home values. Probably correlated with the high median of years of education completed. This cluster is the 2nd lowest percentage of veterans after May 1995. They are the 2nd group that donates the highest amounts per year, are concentrated in the MidWest area and are the most recent donors.

- ***Cluster 5***: In the demographic perspective, this cluster is very similar to Cluster 4, the difference being that they are the 2nd eldest group amongst all. Also, it is possible to highlight the lowest percentage of veterans after May 1995. This group has the highest average of time between first and last donation (over 7 years) and has been donating the highest amounts and in more times, also this group has the best response to emails offers.

- ***Cluster 6***: This cluster has a high percentage of white population (contrasting with the low percentage of hispanic population), low percentage of adults employed and low percentage of population born in the same state of residence. Regarding the Military indicators, this group has a high percentage of veterans (mainly from WWII) and its citizens are concentrated in the South. This group seems to be very generous when it comes to the number and amount of promotions. However, regarding the responses to mail offers, this group does not seem to be particularly responsive.

- ***Cluster 7***: Regarding the Demographic (apart from the age) and Military indicators, this group is very similar to Cluster 6. In terms of Gifts & Promotions, this is the group donating less frequently and the smallest amounts.

- ***Cluster 8***: This is the youngest cluster and probably because of that, with the highest percentage of households with 3 or more elements, the 2nd highest median of household income and the highest percentage of adults employed. Military wise, this cluster is very similar to Cluster 1. The responses to mail offers is above other clusters' average and its population is mostly concentrated in the West area. The remaining Gifts & Promotions indicators are within the average.

OUTLIERS CLASSIFICATION

In this section we present the results of the outlier classification explained in the *DATA ANALYSIS* section. The first one (Figure 11), decision tree with depth equal 4. Figure 12 shows the optimal parameter which is depth equal 7 and we are able to predict over 82% of the customers correctly.

```
dt_plot = DecisionTreeClassifier(max_depth=4, random_state=42)
dt_plot.fit(X_train, y_train)
print("It is estimated that in average, we are able to predict {0:.2f}% of the donors correctly".format(dt_plot.score(X_test, y_t
```

```
It is estimated that in average, we are able to predict 71.88% of the donors correctly
```

Figure 11 - The result of parameters for the depth of 4.

```
X = donors.drop(columns=['merged_labels'])
y = donors.merged_labels
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify = y
)
dt = DecisionTreeClassifier(max_depth=7, random_state=42)
dt.fit(X_train, y_train)
print("It is estimated that in average, we are able to predict {0:.2f}% of the donors correctly".format(dt.score(X_test, y_test)*
```

```
It is estimated that in average, we are able to predict 82.20% of the donors correctly
```

Figure 12 - The result of optimal parameter for the data analysis of this project.

CONCLUSION AND MARKETING APPROACH

Looking at the customer segmentation and its analysis presented in the previous section, it is possible to see some relations which are interesting in the marketing point of view:

- The groups that seem to **react more to mail offers** are mainly **urban residents**, not necessarily with the highest income or education level. We can also conclude that the most responsive groups tend to reside in the **West area**.
- The groups (**Clusters 4, 5**) with high income also tend to be **more responsive to mail offers**. Cluster 4 is composed of more recent donors, having recent donations indicators more generous; while Cluster 5 has long time donors, making high volumes/ amounts of donations per promotion.
- The groups showing a **higher percentage of WWII veterans** (**Clusters 6 and 7**) live in **non-urban areas**, in the **South**, and they are **not very responsive to mail offers**. Although, Cluster 6 has a great frequency of donations per promotion and month, and the cluster 7 does not seem to have the same behaviour.
- **Clusters 0 to 3**, have the **highest percentages of veterans after May 1995**. In terms of Gift & Promotions they are average showing a frequency during all year, but without highlights of amounts. Alghouth, in Demographic terms, those clusters concentrate the lowest percentages of white people and high percentage os hispanics also they concentrate highest percentages of residents from the same state that was born.
- **Cluster 8** is the most singular cluster. The **youngest cluster** and with one of the **highest incomes** per capita, however when it comes to their **response to PVA campaigns this group is average**, they do not seem to be specially willing to donate, probably because this cluster represents **younger families on working age**.

Marketing Approach

Based on the analysis shown throughout this report, we would advise PVA to invest in promotions targeting the West to return the best donations. Although to keep a good return of donations during the all year, PVA also should reach the South region, focusing on veterans in general, to reach mostly the cluster 6 in this area.

In order to improve the interest over the PVA, should develop a different communication approach focusing on reaching the clusters with more diverse ethnicity (Clusters 0 to 3).

The strategy proposed would guarantee current donations during all year with highlights periods with great amounts and a long term result of new donors for PVA.

REFERENCES

[1] Aggarwal, C (2015). *Data Mining - The Textbook*, 2015th Edition ed. Switzerland, Springer International

[2] Han, J.; Pei, J. and Kamber, M., (2011). *Data Mining - Concepts and Techniques*, 3rd ed. Amsterdam, Elsevier Science LtdPublication

[3] Zhang, Z 2019, Feature Selection Why & How Explained, Towards Data Science, viewed 9 December 2020, <https://towardsdatascience.com/feature-selection-why-how-explained-part-1-c2f638d24cdb>

[4] Kumar, A 2020, Lasso Regression Explained with Python Example, Data Analytics, viewed 9 December 2020, <https://vitalflux.com/lasso-ridge-regression-explained-with-python-example/>

[5] Nagpal, A 2017, L1 and L2 Regularization Methods, Towards Data Science, viewed 29 December 2020, <https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c>

[6] Hall, S 2017, Similarities of Univariate & Multivariate Statistical Analysis, Sciencing, viewed 20 December 2020, <https://sciencing.com/similarities-of-univariate-multivariate-statistical-analysis-12549543.html>

[7] Soeteway, A 2020, The complete guide to clustering analysis, Towards Data Science, viewed 27 December 2020, <https://towardsdatascience.com/the-complete-guide-to-clustering-analysis-10fe13712787>

[8] Ge, Y 2020, sklearn.neighbors.LocalOutlierFactor, Scikit learn, viewed 27 December 2020, <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html>

[9] Dawson, C 2018, Outlier Detection with One-Class SVMs, Towards Data Science, viewed 27 December 2020, <https://towardsdatascience.com/outlier-detection-with-one-class-svms-5403a1a1878c>

[10] Lewinson, E 2018, Outlier Detection with Isolation Forest, Towards Data Science, viewed 27 December 2020, <https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e>

[11] Donges, N 2019, A Complete guide to the random forest algorithm, Builtin, viewed 29 December 2020 <https://builtin.com/data-science/random-forest-algorithm#feature>

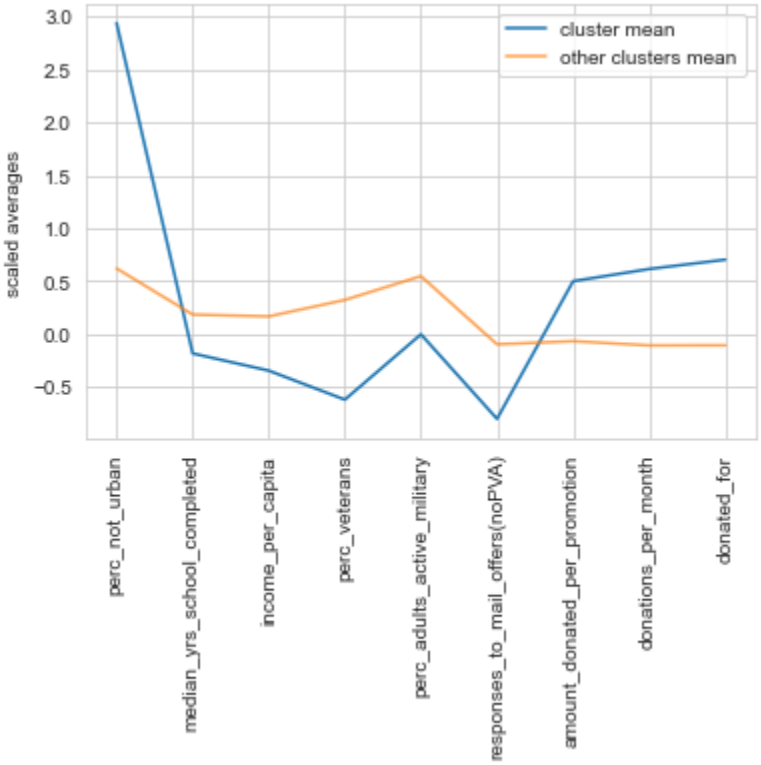
APPENDICES

Appendix I - Cluster means table

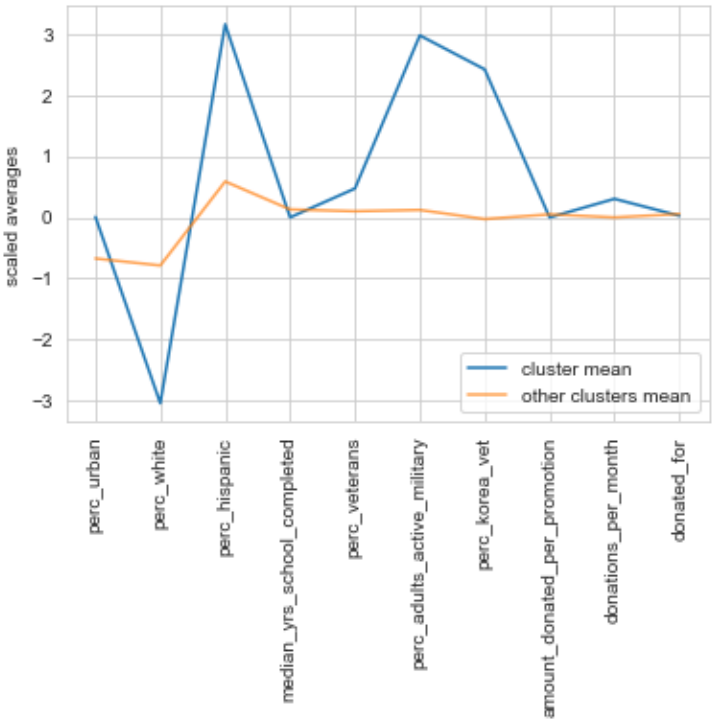
Feature	Cluster								
	0	1	2	3	4	5	6	7	8
population	3288.70	2447.77	2869.17	3571.66	2352.37	2618.89	3061.36	3053.42	2362.74
perc_urban	2.24	84.71	90.79	2.31	92.42	91.70	41.47	43.85	84.27
perc_not_urban	37.18	0.96	1.07	41.80	1.97	2.61	13.67	13.08	2.80
perc_white	92.01	84.31	83.07	89.74	91.21	91.53	95.79	95.91	90.95
perc_hispanic	3.29	9.21	9.33	3.94	4.88	4.60	3.33	3.44	4.90
perc_3+_househ	42.81	43.91	38.69	41.27	38.66	39.12	24.01	24.00	47.40
median_home_value_in_hundreds	530.27	926.91	957.45	516.70	2722.00	2585.08	916.25	995.67	2290.63
perc_owner_occupied_hu	75.45	70.86	66.85	74.29	79.31	80.37	78.69	78.90	83.62
perc_occupied_househ	90.17	92.70	92.81	89.40	94.27	94.90	71.72	70.52	95.06
perc_seasonal_house	16.27	9.71	9.05	16.52	12.12	11.65	59.50	60.13	10.47
perc_housev_50+	49.14	80.37	79.44	46.86	98.47	98.35	78.53	79.96	98.38
median_househ_income_in_hundreds	252.33	335.06	311.84	239.24	602.38	585.34	262.67	276.21	598.06
income_per_capita	11597.60	14675.94	14928.40	11407.23	32835.66	30655.49	15934.51	17058.63	27329.40
perc_adults_employed	61.29	66.43	62.63	59.20	65.85	65.30	43.45	43.89	69.96
perc_priv_prof_or_sal	66.86	71.55	71.88	66.60	67.41	66.99	68.73	68.86	67.86
median_yrs_school_completed	121.24	125.80	125.85	121.15	151.60	151.24	124.20	125.23	149.49
perc_born_state_of_res	72.41	56.21	57.94	72.60	47.68	47.95	31.14	30.93	47.77
perc_vietnam_vet	26.78	30.50	20.08	22.64	19.68	19.94	17.35	17.42	30.85
perc_WWII_vet	36.80	28.73	47.69	44.70	48.72	49.39	57.82	58.32	26.52
perc_adults_active_military	0.15	0.48	0.21	0.10	0.14	0.15	0.10	0.10	0.31
perc_veterans	15.84	17.37	15.96	15.61	16.34	16.71	22.02	22.13	17.22
perc_korea_vet	21.77	28.15	18.57	18.47	20.47	20.19	18.48	18.41	27.99
percent_vet_after_May1995	7.35	7.90	8.15	7.64	4.19	4.11	5.20	5.00	4.41
dollar_amount_donations	128.78	99.62	98.16	62.51	68.43	188.97	138.72	65.26	102.52
number_donations	14.37	10.01	9.59	5.14	4.72	17.47	14.64	5.09	8.60
amount_most_recent_donation	14.59	16.18	16.59	17.76	20.55	15.60	13.99	18.98	18.21
avg_amount_per_donation	10.65	12.41	12.72	13.80	16.19	11.35	10.54	14.63	14.38
house_number_y/n	0.51	0.53	0.54	0.49	0.48	0.51	0.52	0.52	0.49
responses_to_mail_offers(noPVA)	2.87	3.82	4.06	2.52	4.44	4.95	3.41	2.86	3.95
homeowner	0.42	0.63	0.61	0.44	0.66	0.63	0.53	0.55	0.64
major_donor_flag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
donated_for	70.69	51.14	50.35	32.33	31.36	87.74	67.22	38.34	46.80
age	61.04	59.19	60.45	57.89	58.76	63.95	66.31	63.35	57.28
gender_m	0.54	0.55	0.55	0.53	0.53	0.51	0.52	0.50	0.51
gender_f	0.40	0.40	0.40	0.42	0.43	0.41	0.42	0.45	0.44
NorthEast	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
West	0.24	0.46	0.41	0.23	0.58	0.57	0.33	0.35	0.59
MidWest	0.34	0.20	0.28	0.30	0.18	0.17	0.08	0.07	0.12
South	0.41	0.34	0.30	0.47	0.23	0.24	0.58	0.57	0.27
donations_per_month	0.21	0.20	0.20	0.19	0.17	0.21	0.26	0.11	0.19
donations_per_promotions	0.23	0.19	0.18	0.14	0.13	0.26	0.25	0.12	0.17
avg_promotions_per_year	0.96	1.13	1.12	1.29	1.25	0.81	1.23	0.85	1.13
amount_donated_per_promotion	2.13	1.98	1.96	1.69	1.83	2.76	2.39	1.62	2.07
amount_donated_per_year	2.02	2.19	2.15	2.25	2.36	2.28	2.73	1.36	2.30

Appendix II - Cluster means, graphical visualization

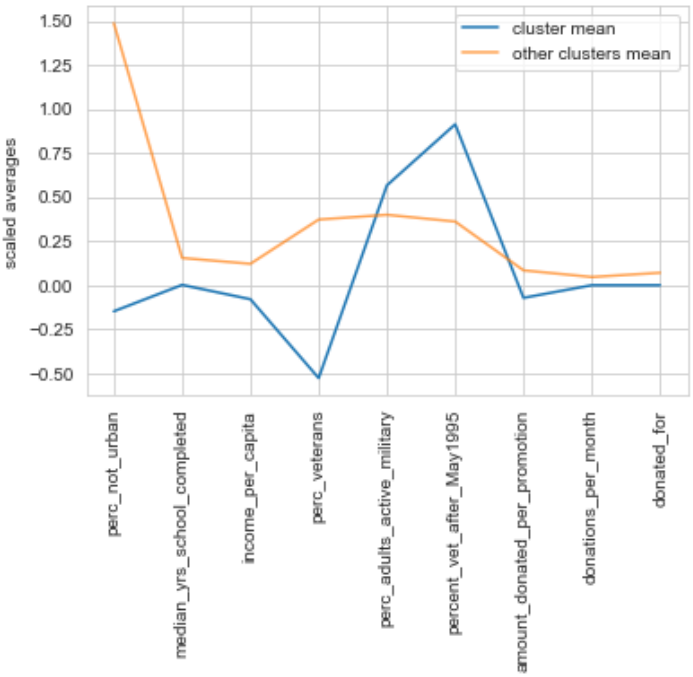
- ***Cluster 0***



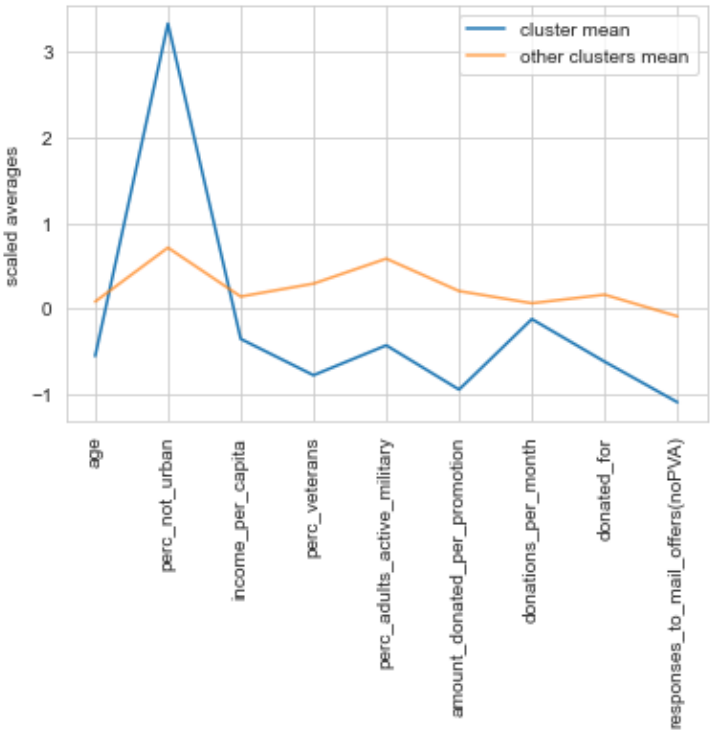
- ***Cluster 1***



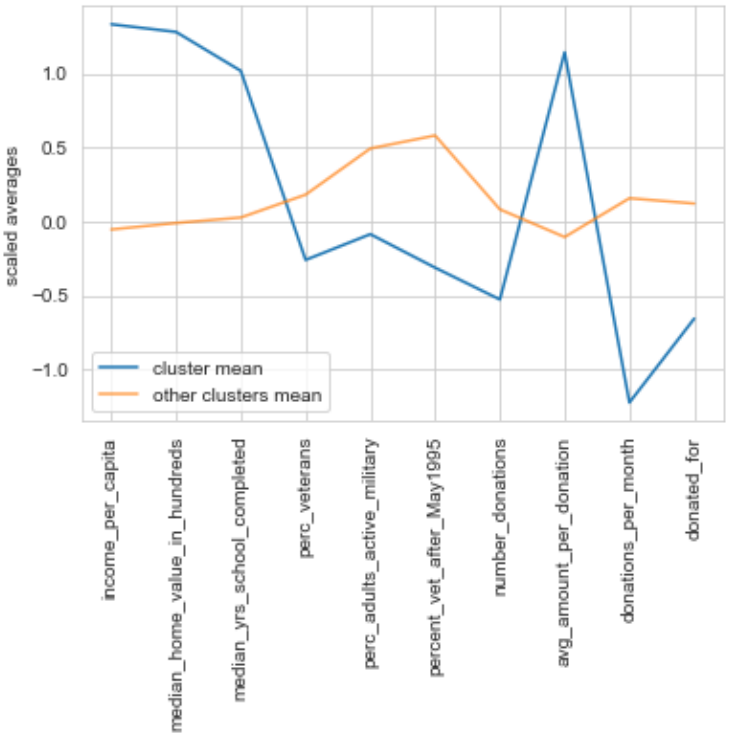
- ***Cluster 2***



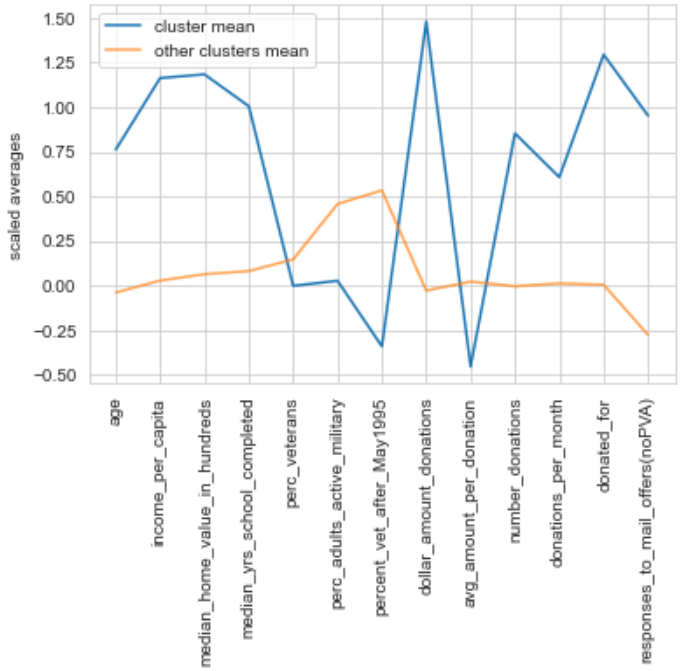
- ***Cluster 3***



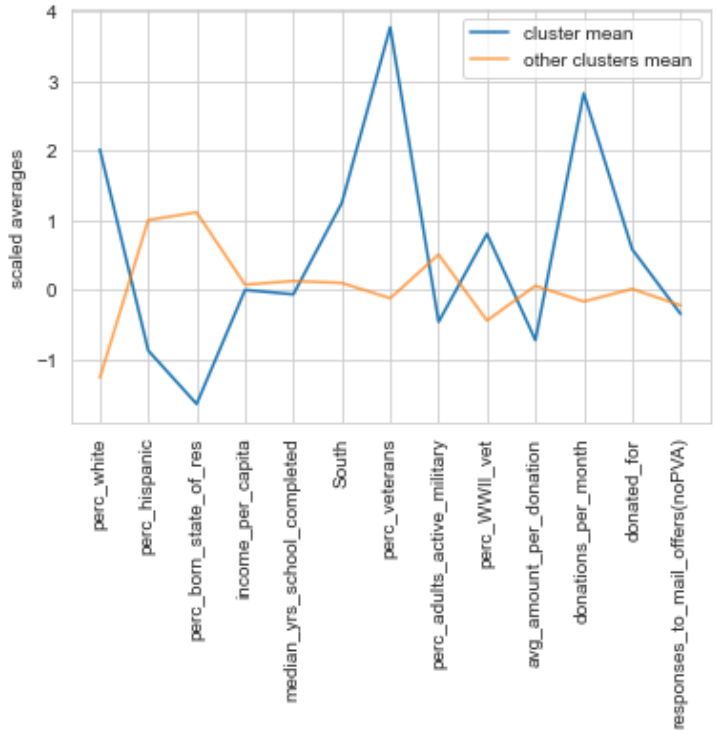
- ***Cluster 4***



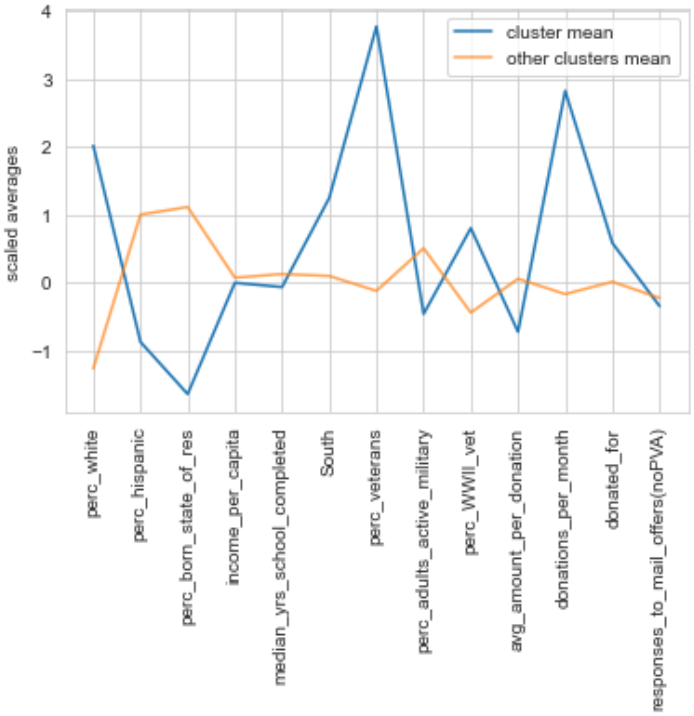
- ***Cluster 5***



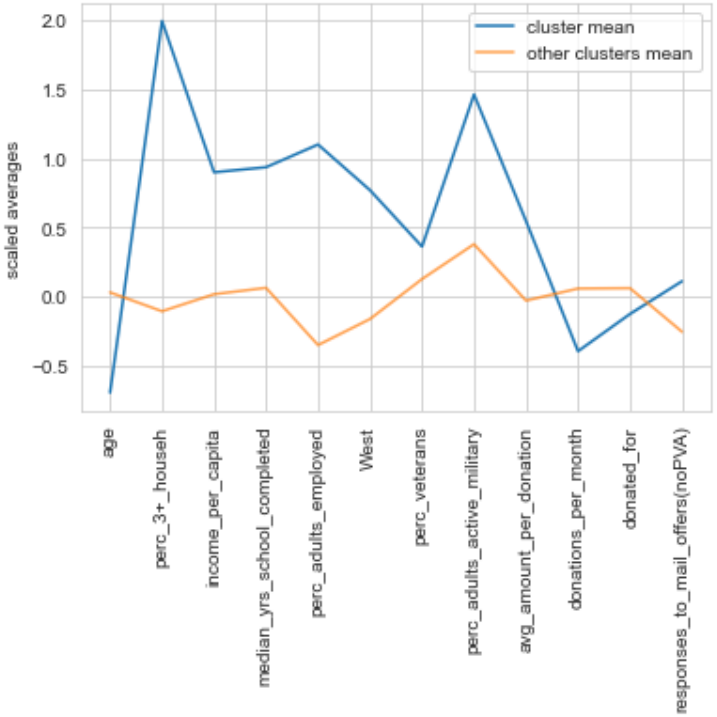
- ***Cluster 6***



- ***Cluster 7***



- ***Cluster 8***



Appendix III - Clustering features brief description

Feature	Description
population	Cluster size
perc_urban	Percentage of population living in urban areas
perc_not_urban	Percentage of population living in non urban areas
perc_white	Percentage of white population
perc_hispanic	Percentage of hispanic population
perc_3+_househ	Percentage of households with 3 or more elements
median_home_value_in_hundreds	Median home value (in hundreds)
perc_owner_occupied_hu	Percent owner occupied housing units
perc_occupied_househ	Percent occupied housing units
perc_seasonal_house	Percent seasonal/recreational vacant units
perc_housev_50+	Percent home value >= $50,000
median_househ_income_in_hundreds	Median household income in hundreds
income_per_capita	Per capita income
perc_adults_employed	Percentage of adults employed
perc_priv_prof_or_sal	Percentage of private profit wage or salaried workers
median_yrs_school_completed	Median years of school completed by adults 25+
perc_born_state_of_res	Percentage born in state of residence
homeowner	Percentage of homeowners
age	Age (months)
gender_m	Percentage of male population
gender_f	Percentage of female population
NorthEast	Percentage of residents in the NorthEast region
West	Percentage of residents in the West region
MidWest	Percentage of residents in the MidWest region
South	Percentage of residents in the South region
dollar_amount_donations	Dollar amount of donations
number_donations	Number of donations
amount_most_recent_donation	Amount of the most recent donation
avg_amount_per_donation	Average amount per donation
house_number_y/n	Contact number (1 if there is a record of the donor's telephone number; 0 otherwise)
major_donor_flag	Major donor flag
donated_for	Time between the first and last donation (in months)
donations_per_month	Number of donations per month
donations_per_promotions	Number of donations per promotions
avg_promotions_per_year	Average number of promotions per year
amount_donated_per_promotion	Amount donated per promotion
amount_donated_per_year	Amount donated per year
responses_to_mail_offers(noPVA)	Total number of known times the donor has responded to a mail order offer other than PVA's
perc_vietnam_vet	Percentage of Vietnam veterans
perc_WWII_vet	Percentage of WWII veterans
perc_adults_active_military	Percentage of adults active in the military
perc_veterans	Percentage of veterans
perc_korea_vet	Percentage of Korea veterans
percent_vet_after_May1995	Percententage of veterans after May 1995