NOVA IMS

# DATA MINING PROJECT

## PARALIZED VETERANS OF AMERICA (PVA)

BY:
MARIA LEONOR MORGADO R20170871
RAQUEL CASTRO R2015528
VALENTYNA RUSINOVA M20200591

# Index of Contents

## Index of figures

## Index of Graphs

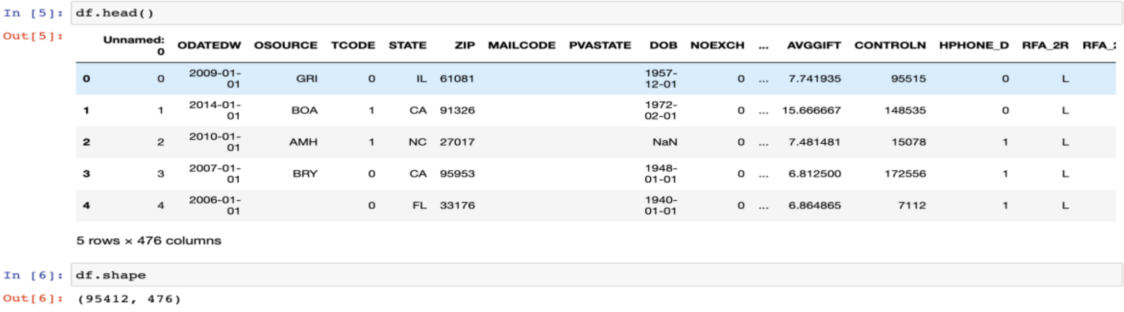
# Introduction

*"PVA was originally founded by a band of service members who came home from World War II with spinal cord injuries. They returned to a grateful nation, but also to a world with few solutions to the major challenges they faced.*
*These wounded heroes made a decision not just to live, but to live with dignity as contributors to society. They created Paralyzed Veterans of America, an organization dedicated to serving veterans—and to medical research, advocacy and civil rights for all people with disabilities".*
(PVA, Mission Statement)

Paralyzed Veterans of America (PVA) is a non-profit Organization that provides programs and services for US veterans with spinal cord injuries or disease since 1946. With an in-house database of 13 million donors, PVA is also one of the largest direct mail fundraisers in the United States of America.
As Data Mining/Analytics Consultants we were given the task of analyzing a sample of the results of one of PVA's recent fundraising appeals, containing 95 412 donors. This mailing was sent to a total of 3.5 million PVA donors who were on the PVA database. Everyone included in this database made at least one prior donation to PVA. One group that is of particular interest to PVA is *Lapsed Donors*, these are individuals who made their last donation to PVA 13 to 24 months ago.
Accordingly, to recapturing these former donors, that is a critical aspect of PVA's fundraising efforts, the PVA Organization asked us to develop a Customer Segmentation in such a way that it will be possible for them to better understand how their donors behave and identify the different segments of donors/potential donors within their database. In this study we defined, described and explained the segmentation and marketing approach elaborated for the PVA donors.

# The Data

The first step taken to elaborate the requested Segmentation Analysis was to know and understand the data made available for the study. The dataset (figure 1) is composed of a series of variables that correspond to demographic, economic, socio-economic data, data about donations and applied promotions, data about donors and data about the community and neighborhood where they belong.



```
In [5]: df.head()
Out[5]:
```

| | Unnamed: 0 | ODATEDW | OSOURCE | TCODE | STATE | ZIP | MAILCODE | PVASTATE | DOB | NOEXCH | ... | AVGGIFT | CONTROLN | HPHONE_D | RFA_2R | RFA_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 2009-01-01 | GRI | 0 | IL | 61081 | | | 1957-12-01 | 0 | ... | 7.741935 | 95515 | 0 | L | |
| **1** | 1 | 2014-01-01 | BOA | 1 | CA | 91326 | | | 1972-02-01 | 0 | ... | 15.666667 | 148535 | 0 | L | |
| **2** | 2 | 2010-01-01 | AMH | 1 | NC | 27017 | | | NaN | 0 | ... | 7.481481 | 15078 | 1 | L | |
| **3** | 3 | 2007-01-01 | BRY | 0 | CA | 95953 | | | 1948-01-01 | 0 | ... | 6.812500 | 172556 | 1 | L | |
| **4** | 4 | 2006-01-01 | | 0 | FL | 33176 | | | 1940-01-01 | 0 | ... | 6.864865 | 7112 | 1 | L | |

5 rows × 476 columns

```
In [6]: df.shape
Out[6]: (95412, 476)
```

*Figure 1 - Dataset*

Thus, the initial dataset had about 95,412 observations, with no duplicate data and 476 initial descriptive variables. The initial input variables were subjected to an exhaustive and individual initial analysis in order to understand their role and meaning given our purpose, and were subject to transformations and analyzes of variability, relevancy and redundancy with the aim of reduce the dataset and looking for the best performance and results of this segmentation analysis.

For a better understanding of the given variables and for better handling of them, we divided the dataset into Metric Features and Non-Metric Features, being the Metric Features divided in the following categories:

- ➔ **mili_gov** - data related with military and government workers;
- ➔ **neighbor** - data about donors neighborhood;
- ➔ **prom_hist** - history of promotions for the last four years;
- ➔ **amount_hist** - data about amount donated through the years;
- ➔ **doners_hist** - historical data about donors.

## Data Preprocessing

*"Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results".* (Han et al., 2012, 03:83)

*"Data has quality if they satisfy the requirements of the intended use. There are many factors comprising data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability".* (Han et al., 2012, 03:84)

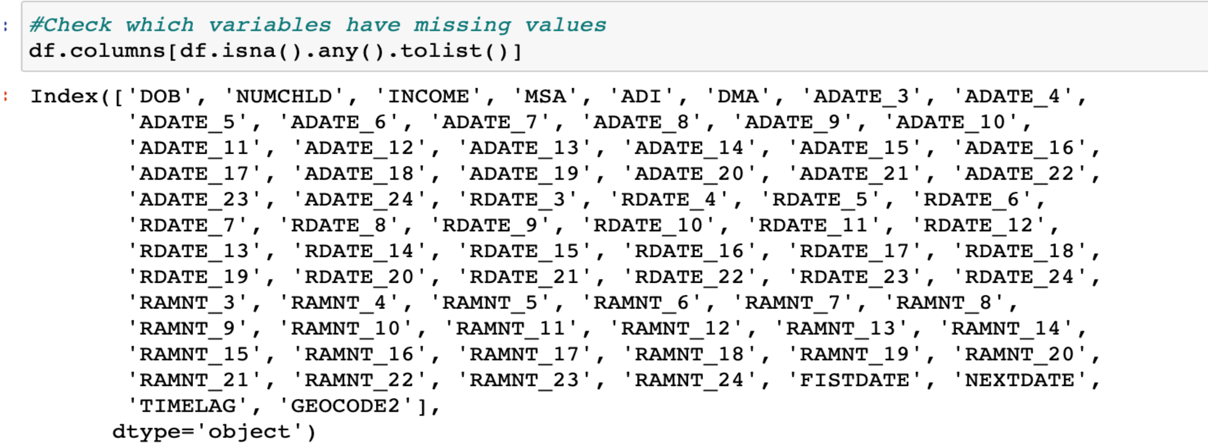
In order to overcome these inconveniences, were applied a series of steps that integrate the Data Preprocessing moment such as Detection and Treatment of Missing Values, Detection and Treatment of noise, such as Outliers, elaboration of Feature Engineering in order to find useful resources to represent the data and reduce the dimension of the dataset, we proceed with the Data Normalization and to complement the Feature Engineering. Lastly, to extract and select the most relevant and less redundant variables and in a way to reduce the number of variables for the elaboration of Clustering Analysis were carried out Correlation Analysis between the different numerical features and were applied the Principal Component Analysis Technique. To support some of our decisions we did not only use visualization techniques in code but also Pandas Profiling open source module.

Firstly, the features that were not going to be useful were removed from the dataset, since they had problems such as a lot of missing values and no variability. TCODE, CONTROLN, Unnamed: 0, ZIP, NOEXCH, RFA_2R, WEALTH2 are some examples of those features.

### Missing values

It is important to understand the concept of Missing Values in order to successfully manage data and their impact if the Missing Values are not handled properly.

Missing data are defined as values that are not available and that would be meaningful if they are observed. Missing data can be anything from, for example, missing sequence, incomplete feature, files missing, information incomplete, data entry error. Most datasets in the real world contain missing data. In the figure 2, we can see the variables that have missing values in our dataset.



```
#Check which variables have missing values
df.columns[df.isna().any().tolist()]
```

```
Index(['DOB', 'NUMCHLD', 'INCOME', 'MSA', 'ADI', 'DMA', 'ADATE_3', 'ADATE_4',
       'ADATE_5', 'ADATE_6', 'ADATE_7', 'ADATE_8', 'ADATE_9', 'ADATE_10',
       'ADATE_11', 'ADATE_12', 'ADATE_13', 'ADATE_14', 'ADATE_15', 'ADATE_16',
       'ADATE_17', 'ADATE_18', 'ADATE_19', 'ADATE_20', 'ADATE_21', 'ADATE_22',
       'ADATE_23', 'ADATE_24', 'RDATE_3', 'RDATE_4', 'RDATE_5', 'RDATE_6',
       'RDATE_7', 'RDATE_8', 'RDATE_9', 'RDATE_10', 'RDATE_11', 'RDATE_12',
       'RDATE_13', 'RDATE_14', 'RDATE_15', 'RDATE_16', 'RDATE_17', 'RDATE_18',
       'RDATE_19', 'RDATE_20', 'RDATE_21', 'RDATE_22', 'RDATE_23', 'RDATE_24',
       'RAMNT_3', 'RAMNT_4', 'RAMNT_5', 'RAMNT_6', 'RAMNT_7', 'RAMNT_8',
       'RAMNT_9', 'RAMNT_10', 'RAMNT_11', 'RAMNT_12', 'RAMNT_13', 'RAMNT_14',
       'RAMNT_15', 'RAMNT_16', 'RAMNT_17', 'RAMNT_18', 'RAMNT_19', 'RAMNT_20',
       'RAMNT_21', 'RAMNT_22', 'RAMNT_23', 'RAMNT_24', 'FISTDATE', 'NEXTDATE',
       'TIMELAG', 'GEOCODE2'],
      dtype='object')
```

*Figure 2 - Variables with missing values*

We detected the variables with Missing Values in the dataset under the study and verified that a total of 76 variables presented missing data, some were in fact missing but also there were zeros and blanks that needed to be replaced with missing values. To address this problem to the non-numeric variables, used a measure of Central Tendency and replaced the missing with the Mode. While, for the numerical variables we applied a more robust method, the K-Nearest Neighbor Algorithm which is based on imputation by finding the samples in the training set "closest" to it and averages these nearby points to fill in the value.

## Outliers Detection and Treatment

*""What is noise?" Noise is a random error or variance in a measured variable. (...) Outliers may represent noise".* (Han et al., 2012, 03:89)

It is important to clean the data sample to ensure that the observations best represent the problem, and sometimes a dataset can contain extreme values that are outside the range of what is expected and unlike the other data, these are called Outliers.

A simple approach to identifying outliers is to locate those observations that are far from the other examples in the input space. For the detection of Outliers for our metric features, we applied the method of Local Outlier Factor or LOF. To each point is assigned a scoring of how isolated the example is based on the size of its local neighborhood, the points with highest score are more likely to be Outliers. This method removed near 2,1% of our initial data. To remove categorical variables Outliers, we used histograms and boxplots to visualize them and we appealed once more to the analysis of Pandas Profiling, removing only about 1%. After this process we removed in total near 3% of our records in order to produce the study with the best quality possible.

## Feature Engineering

Feature Engineering is a crucial step in the Data Preprocessing process, since it uses the domain knowledge to extract features from raw data through Data Mining techniques. The main objective of these transformations is to improve the performance of the Clustering Analysis that will be performed further.

Accordingly, based on the DOMAIN variable we created two new variables: DOMAIN_URB which represents "Urbanicity Level" by 0 and DOMAIN_ECON which represents "Socio-Economic Status" by 1, finally excluding the DOMAIN variable.

In addition to these, we created two new variables: the variable LAPSED DONORS, since these are a group that is of particular interest to PVA, which distinguishes donors who made last donation to PVA 13 to 24 months ago and the variable FIRST_PVA, which takes the value of 1 if first gift dinnated was to PVA.

To the variable HOMEOWNR we replace the NULL values with Not_H.

We also converted the variables NEXTDATE, FISTDATE, LASTDATE, MAXRDATE, MINRDATE, ODATEDW and ADATE_2 to the format month-year.

Finally, the following variables were transformed into binary variables: VETERANS, PETS, CDPLAY, STEREO, PCOWNERS, GARDENIN and WALKER, variables that we replaced the NULL values with 0 and the "y" value with 1 and the TIMELAG, HIT (1 - Responded; 0 Didn't Respond), CHILD (1 - Have Child; 0 - Not Have Child), MAILCODE (1 - have correct mail; 0 - have wrong mail), RECINHSE (1 - PVA's In House program; 0 - Not PVA's In House program), RECP3 (1 - PVA's P3 program; 0 - Not PVA's P3 program), PEPSTRFL (1 - PEP Star RFA Status; 0 - Not PEP Star RFA Status).

## Data Normalization

*"Distance computations are sensitive to the value ranges of the features in the dataset. This is something we need to control for when we are creating a model, as otherwise we are allowing an unwanted bias to affect the learning process. When we normalize the features in a dataset, we control for the variation across the variances of features and ensure that each feature can contribute equally to the distance metric. Normalizing the data is an important thing to do for almost all Machine Learning algorithms".* (Kelleher et al., 2015, Chapter 5.4.3.)

In order to reduce data redundancy and improve data integrity, were applied the Min-Max Scaler Normalization to normalize the input features. Applying this method, all features will be transformed into the range that varies between 0 and 1 meaning that the *minimum* and *maximum* value of a feature is going to be 0 and 1, respectively.

## Correlation Analysis

Correlation Analysis is a statistical method used to evaluate the strength of relationship between two quantitative variables. At this stage of the Data Preprocessing Step, we applied a Correlation Matrix to evaluate the correlation between the selected metric features where high correlation means that two or more variables have a strong relationship with each other, while a weak correlation means that the variables are hardly related. Based on these assumptions, we excluded variables that had a correlation coefficient greater than 85%.

## Dimensionality Reduction

Dimensionality Reduction is the process of reducing the number of variables under consideration. Having a large number of dimensions in the input space may mean that the volume of that space is very large, and consequently mean that the observations that we have in that space often represent a small and non-representative sample.

This can dramatically impact the performance of the Clustering Analysis and fit on data with many input features, generally referred to as the "Curse of Dimensionality". Therefore, it is often desirable to reduce the number of input features. So, for this purpose, we proceed with the application of the Principal Components Analysis method, PCA for short, to reduce the number of variables, applying this technique to each group of numeric variables defined a priori: mili_gov, neighbor, prom_hist, amount_hist, doners_hist. Subsequently and after removing irrelevant variables in each of these categories, PCA was applied to the set of these variables simultaneously, which did not take any feature.

At this point and after all the transformations, removals, redundancy analyzes through the Correlations Matrix and relevance analysis using the PCA technique, we reduced the dataset of metric variables given initially from 476 to 156 variables.

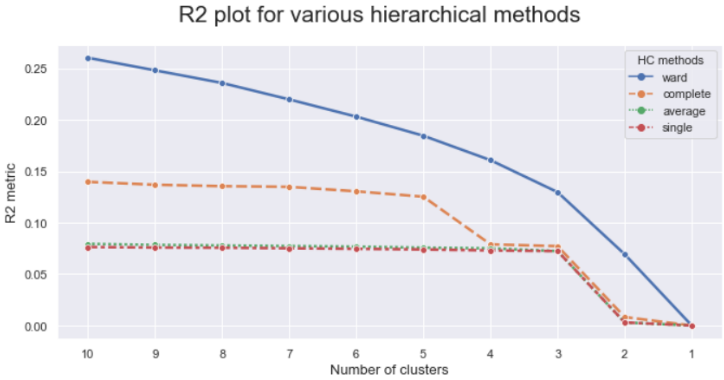
# Clustering

## Agglomerative Hierarchical Clustering

Hierarchical clustering is one of the most popular techniques and easy to understand.

*"In this technique, initially each data point is considered as an individual cluster. At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed."* (Patlolla, 2018)

With the Hierarchical clustering we try to find the best number of clusters. First, we select a sample which includes 10000 rows to calculate the $R^2$ for each HC method, and we can see in the graph 1 that the best method is 'ward' which minimizes the variance of the clusters being merged.



*Graph 1 - Various hierarchical methods*

With the sample we do the dendrogram and we can see in the figure 3 that for all variables (156) the best number of clusters is 5.

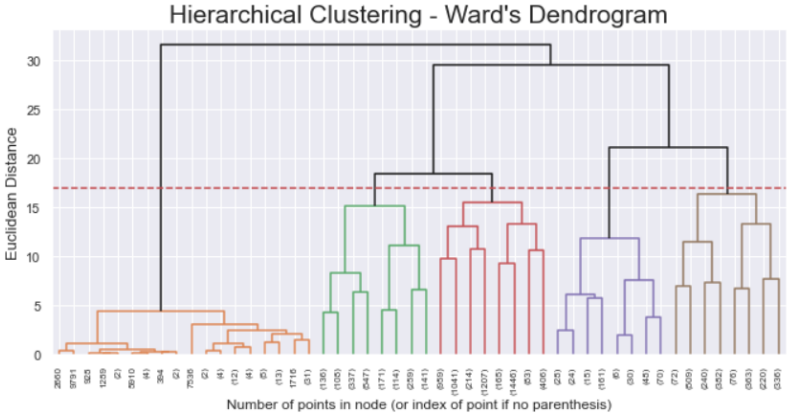


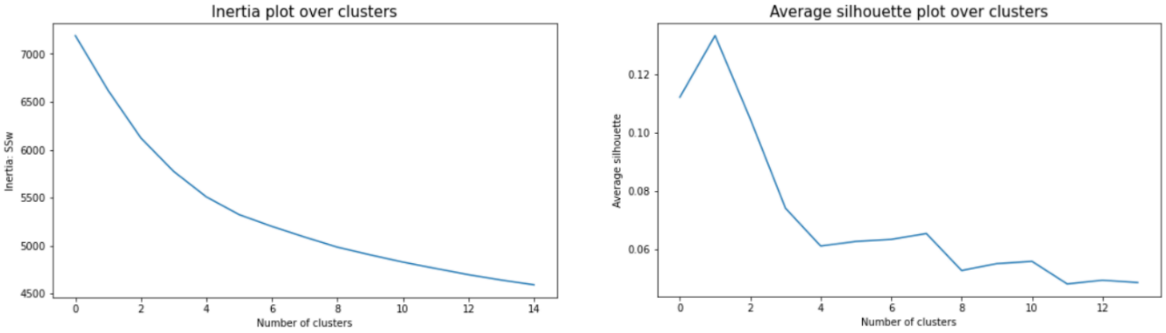
*Figure 3 - Hierarchical Dendrogram*

When we tried to use all variables the R$^2$ was very low (3,6%), therefore we only selected 15 variables that had the highest R$^2$. However, the R$^2$ of the sum of the variables remains low (16%), for this reason we only use this algorithm to define the number of clusters that we will use.

## Partitioning Clustering (K-Means)

*"K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible."* (Dabbura, 2018)

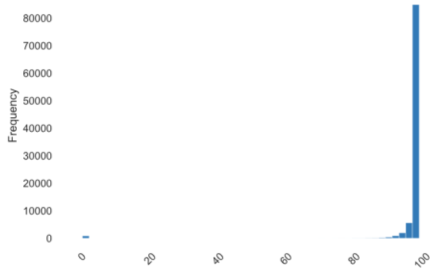
For defining the number of clusters, we looked to the inertia and the silhouette (graph 2), but with all variables it was difficult to understand witch number of clusters will be the better, thereupon we decide to use the dendrogram (figure 3) from the hierarchical clustering.



*Graph 2 - Inertia and Silhouette*

With all variables the R$^2$ using this algorithm is still low (23%). In this way, after testing various combinations, we select the best variables that had the highest R$^2$ in each variable.
We found that the R$^2$ of the variable 'HC20' was very high (94%) which means that K-means clustering was only focusing on this variable to make clusters. The 'HC20' doesn't have the

better distribution to be selected to do the clusters, as we can see in the graph 3. We removed this variable and re-selected the best variables based on R$^2$. After several combinations, we selected some variables to make the clusters. With the selected variables we made several attempts to see which ones provided the best distinction between the clusters and we obtained the R$^2$ with 64%.



Graph 3 - Distribution of 'HC20'

## Self-Organizing Maps (SOM)

*"A self-organizing map (SOM) is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction."* (Ralhan, 2018)

In this algorithm we use the variables that we select in the K-means algorithm and the number of clusters that was defined with the dendrogram.

First, we performed the algorithm with the mapsize = (10,10), in which we only get 100 rows with defined labels. Based on these rows, the cluster labels of the other observations were calculated. The number of clusters was maintained, however the result of R$^2$ was very low (14%). Then we try to calculate the algorithm with mapsize = (50.50), in which we obtained 2500 rows with defined labels. The number of clusters (from 5 to 3) and the value of R$^2$ (9%) has decreased. Using SOM, the R$^2$ was worse compared to R$^2$ of the K-means algorithm. This algorithm was not selected for the clusters, once the R$^2$ is low this algorithm doesn't base on any variable to perform the clusters.

## Density-Based Clustering

*"Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density."* (Chauhan)

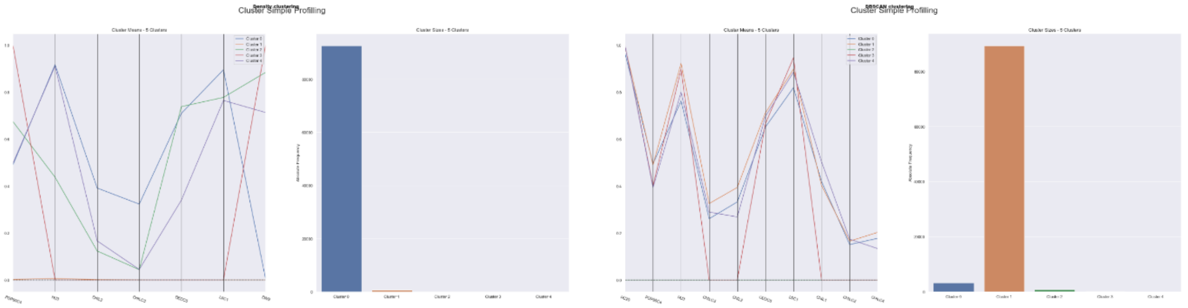
Like for others algorithms, we use all variables (156) in our first try and then we select the best variables with the best R$^2$. Unlikely, this algorithm estimates a large number for clusters, in this way, the number of clusters that we like to perform is needed given by us. After many combinations we select the best variables which performed the average R$^2$ with 30%.

## DBSCAN

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a *“base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.”* (Chauhan)

In the DBSCAN algorithm, after finding the best variables we obtain the $R^2$ with 36%.

Since these two algorithms are based on density, we decide to analyze which one will make better clusters. We can see in the graph 4 below that neither of the two algorithms divides the clusters well, they make an enormous cluster that contains almost all observations and 4 clusters that don’t even have 1000 observations.

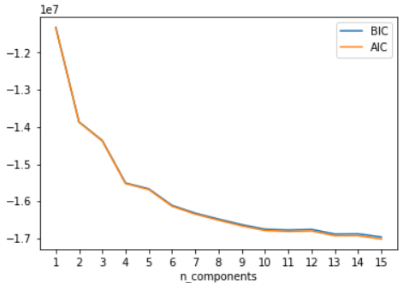


*Graph 4 - Density Clustering and DBSCAN Clustering*

## Gaussian Mixture Model (GMM)

*“A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.”* (Scikit-learn developers)

In GMM, the number of components (clusters) was based on AIC and BIC (graph 5). With all variables (156) the best number of components is 5, the same number that we obtained in the dendrogram.



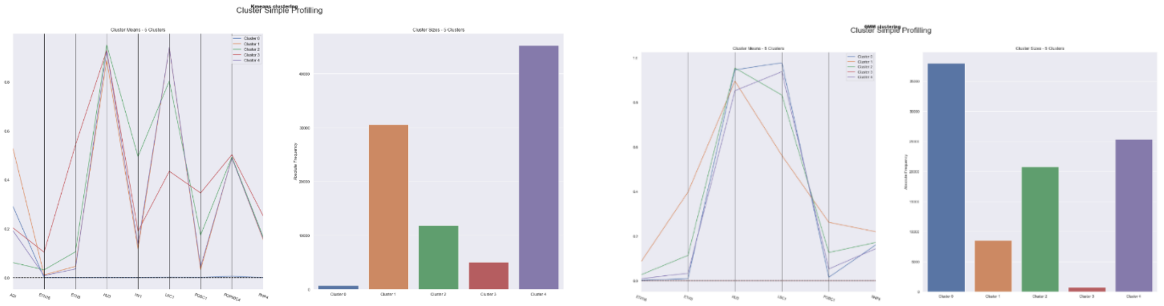
*Graph 5 - Number of components based on AIC and BIC*

Like in the K-means algorithm, GMM bases the formation of clusters only focusing on one variable, 'HC20' ($R^2$ with 94%), in this way we remove this variable and select the best variables based on $R^2$ of each variable.

We made another selection of variables by removing 'HC20' from the beginning and found that some of the best selected variables were different from the first attempt, when 'HC20' was removed after the first selection of variables. We put all the variables together and made a final selection of the best variables. And we obtain the $R^2$ with 61%.

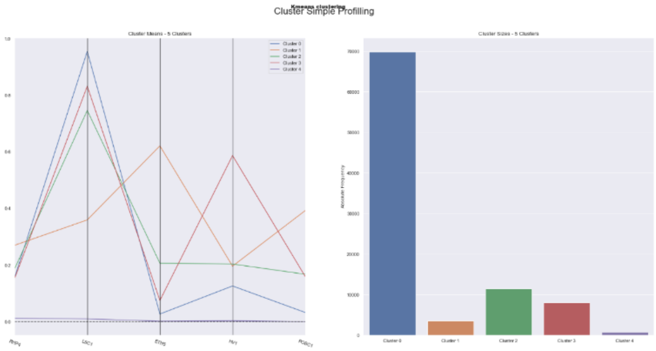
## Clusters Obtained

Considering the results obtained for each clustering algorithm, we choose the K-means and GMM, two who obtained the best $R^2$. In graph 6 we see how these algorithms perform clusters and they are similar.



*Graph 6 - K-means Clustering and GMM Clustering*

After analyzing each cluster, we found that only a few variables varied, and we decided to use only those variables to form the clusters. The algorithm that best distributed the clusters was K-means using the following variables, 'RHP4', 'LSC1', 'ETH5', 'HV1' and 'POBC1'.
In the graph 7, we can see how the clusters are distributed and in the figure 4 we can see how the observations are distributed using t-SNE.



*Graph 7 - K-means Clustering*

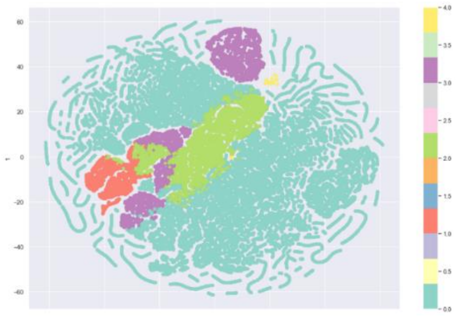
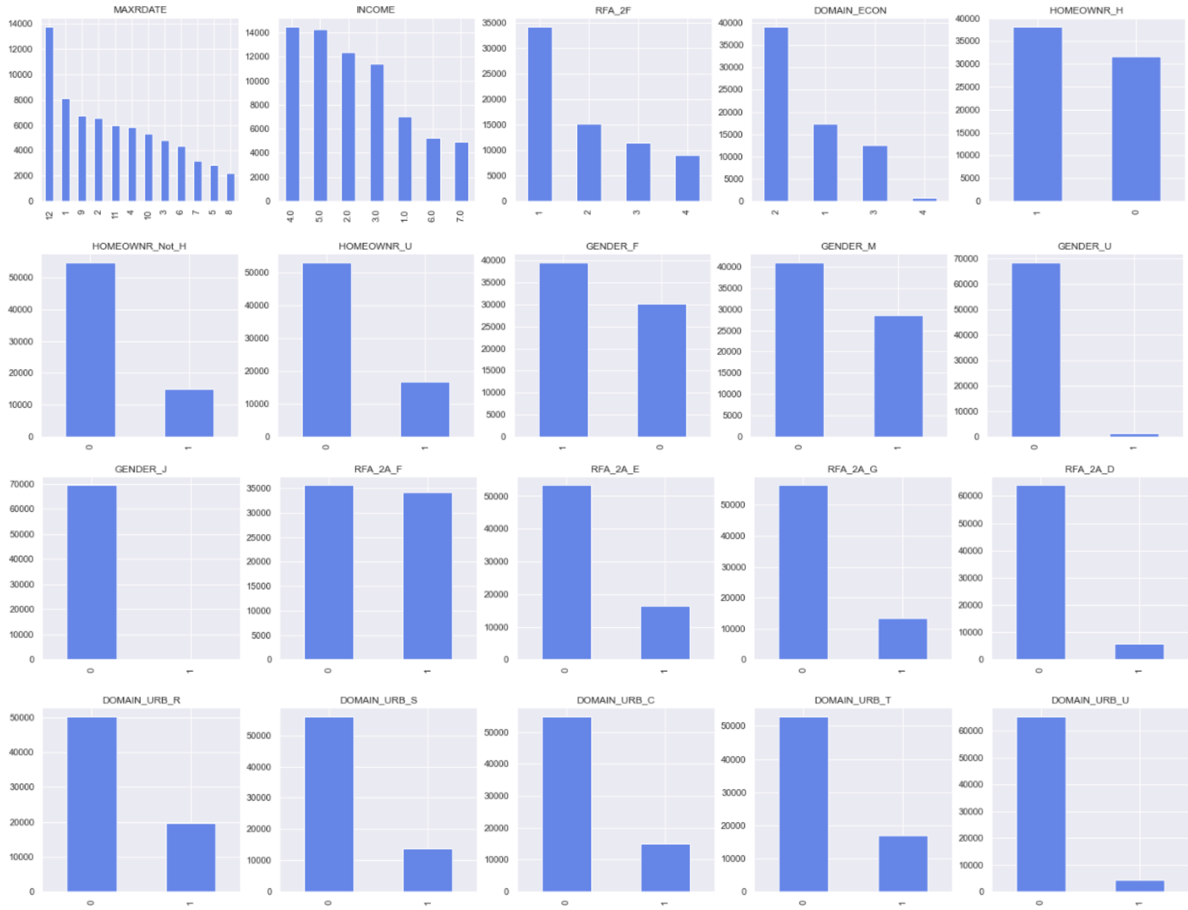


*Figure 4 - t-SNE visualization*

## Assigning/Interpreting Clusters

To help us understand the group of people in each cluster we used some of the categorical variables given to us. After selecting the ones that were more relevant and related with our study objective, which was understanding our population, we applied Dummy Encoding, to all except for the binary and ratio ones, in order to be able to visualize them and their values through the clusters. The chosen ones were: INCOME, RFA_2F, AGE, DOMAIN_ECON, HOMEOWNR, GENDER, RFA_2A, DOMAIN_URB.

Our final solution is composed by 5 clusters in which are included metric features mainly about the donors neighbourhood such as, hispanic presence, home value, number of persons per room, English only speaking and foreign born percentage and age of the donor as well. Cluster 0 is our biggest cluster and includes all types of people from our population, as we can see by the age feature. This group is English only speaking, there is no hispanic presence and so, all born in America. Furthermore, the medium home value is low. Therefore, we named this cluster **General** because it is not specific in a group. The profiling for the cluster 0 was as follows (figure 5):



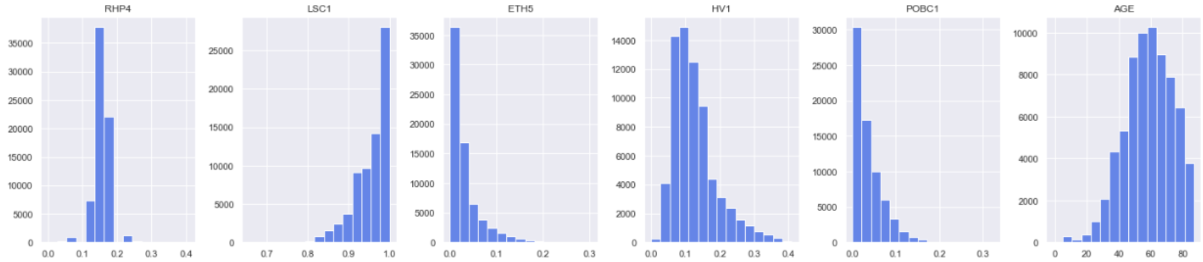


*Figure 5 - Profiling for cluster 0*

The name **low** was given to cluster 1. This cluster includes 3500 people, the home value is considered medium and this group distinguishes from the others because it has a considerable high hispanic presence which leads to people who speak other languages than English and some foreign born as well. The number of persons per room achieves the highest in this cluster and by all the mentioned above we can see why it is called low, because people have low socioeconomic status. The profiling for the cluster 1 was as follows (figure 6):

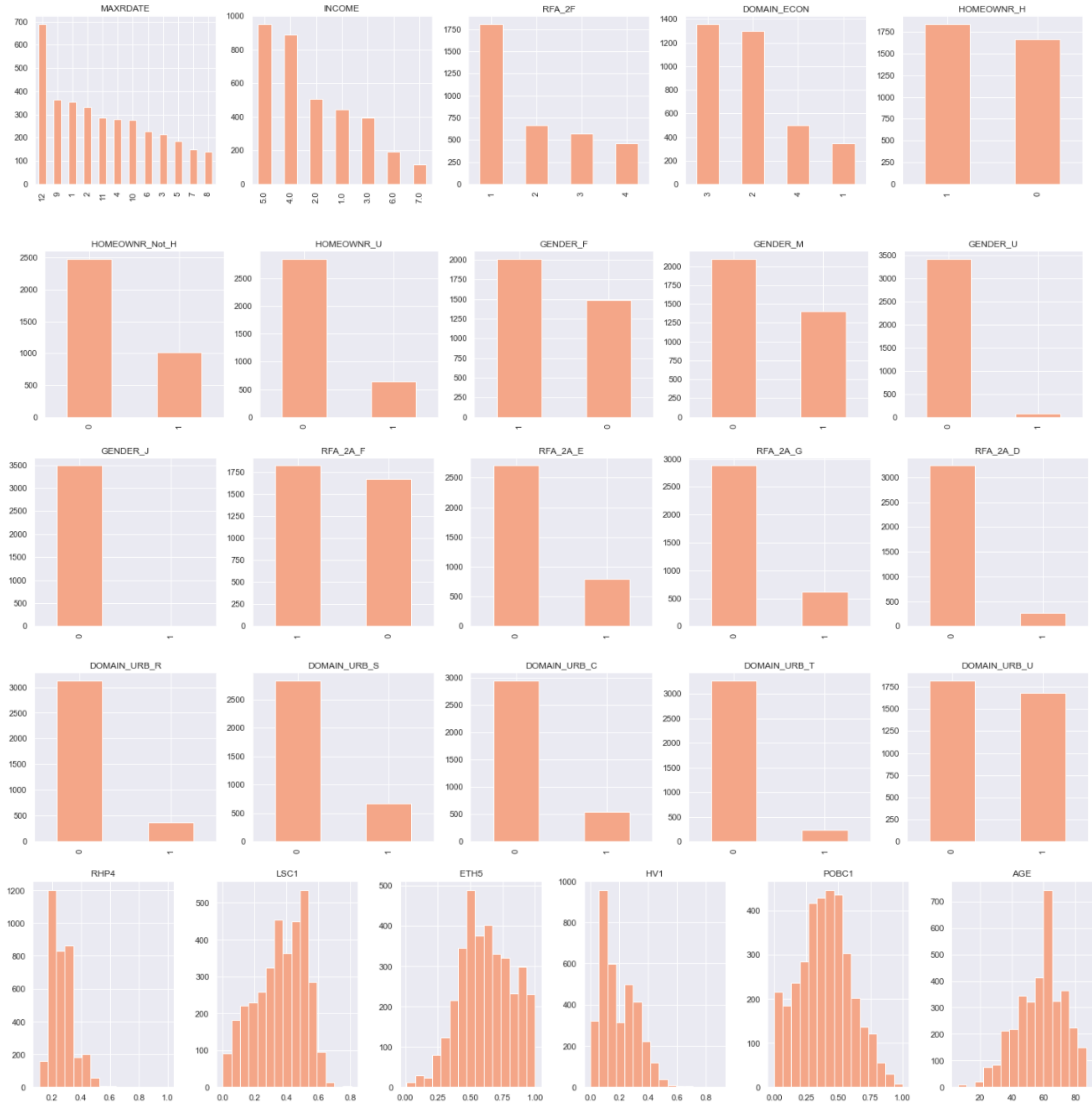


*Figure 6 - Profiling for cluster 1*

Speaking about cluster 2, the income and economic status are both considered medium, there are few hispanics which are the only ones that speak a language other than English and are foreign born. Home value is medium-low and there are a few people per room, 2 to 3. Thereby, this cluster is now called **medium**. The profiling for the cluster 2 was as follows (figure 7):

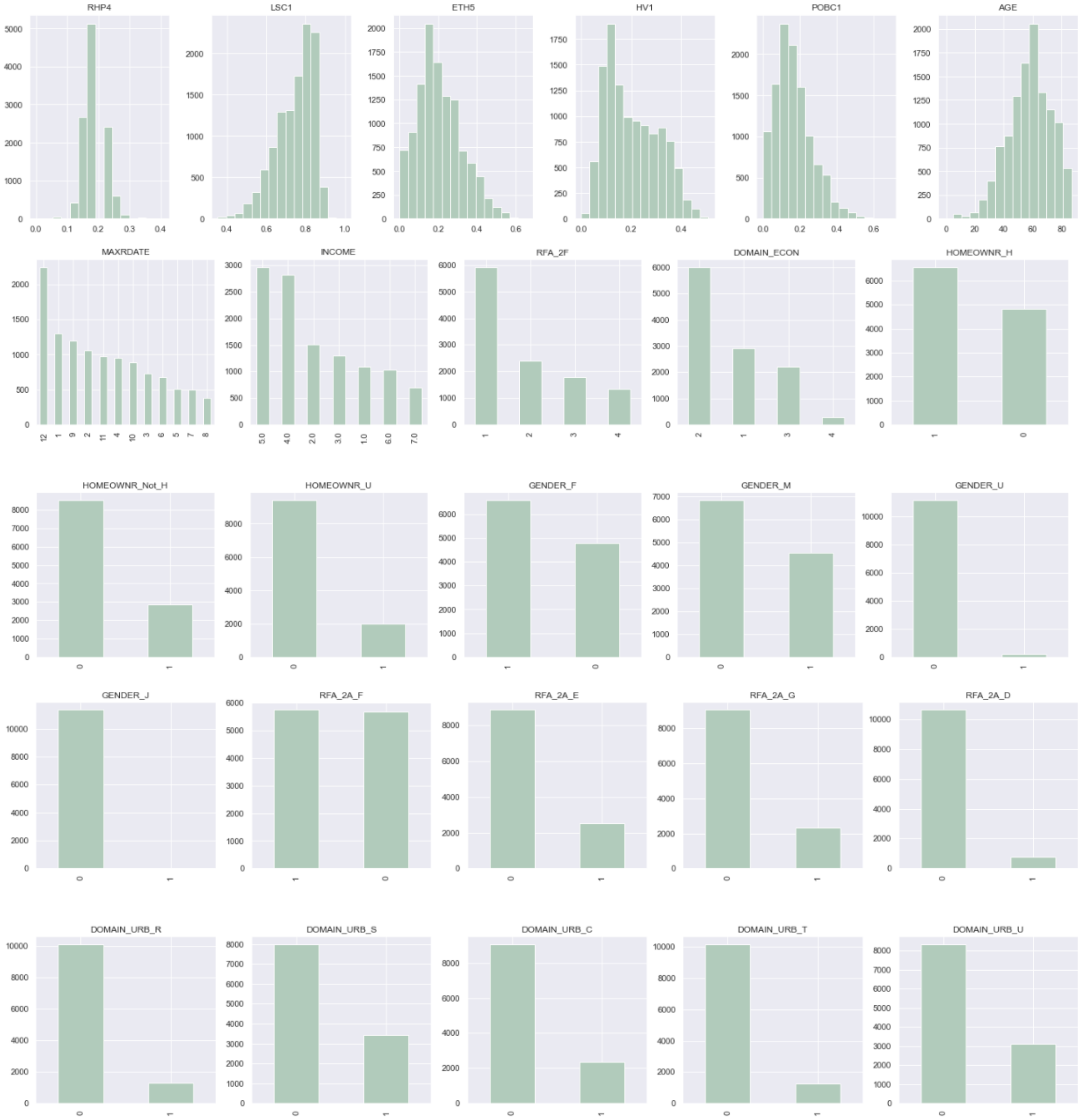


*Figure 7 - Profiling for cluster 2*

**High** is our cluster 3 and that is mainly because it has the highest economic status and income values.  That translates also in home value that varies a little bit from medium to high according to the donors urbanity, suburban or urban, respectively and there are few persons per room, 1 to 2.

In this there is a very small hispanic presence, almost everyone was born in America and only a few can speak other languages than English. The profiling for the cluster 3 was as follows (figure 8):

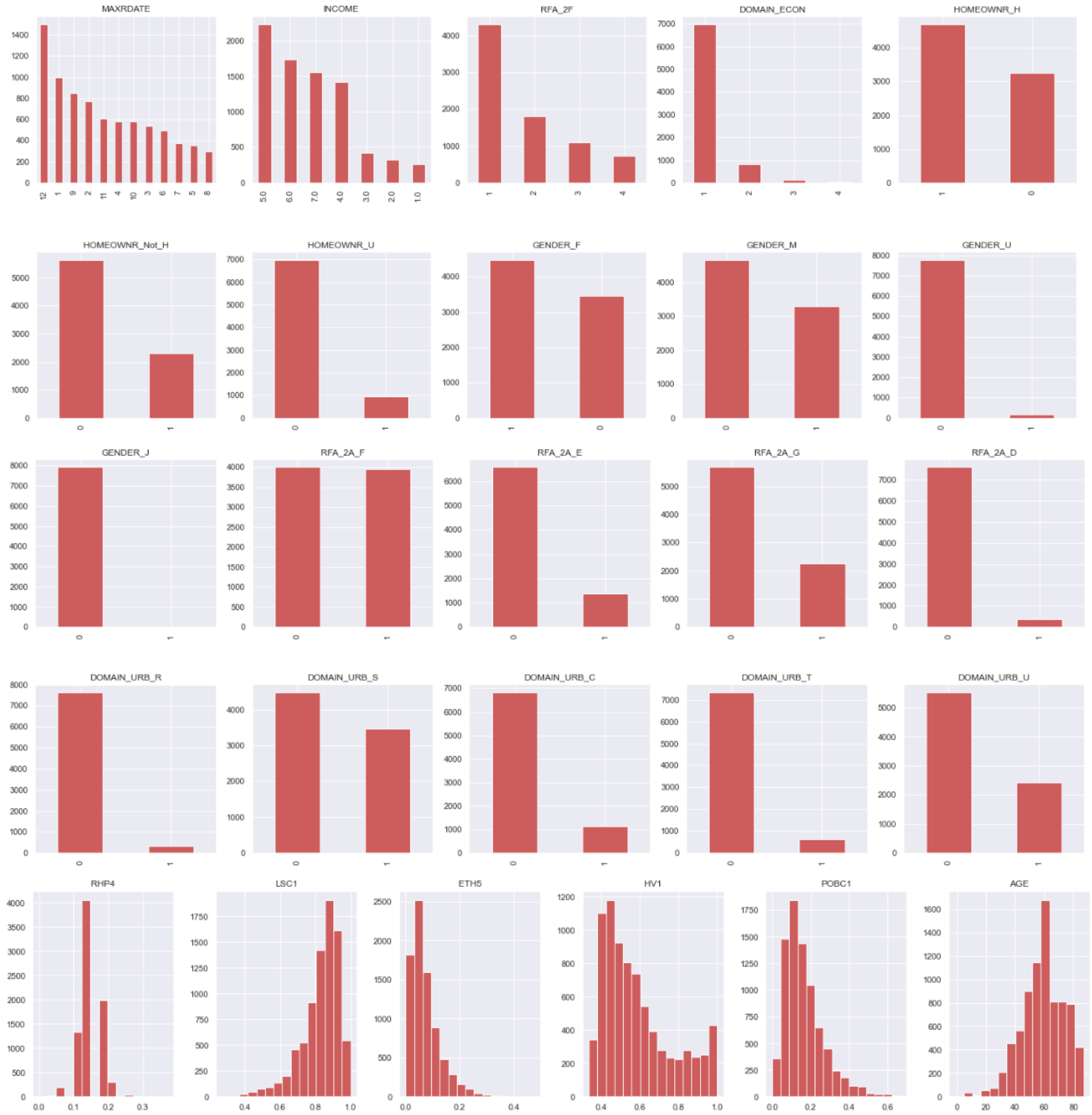
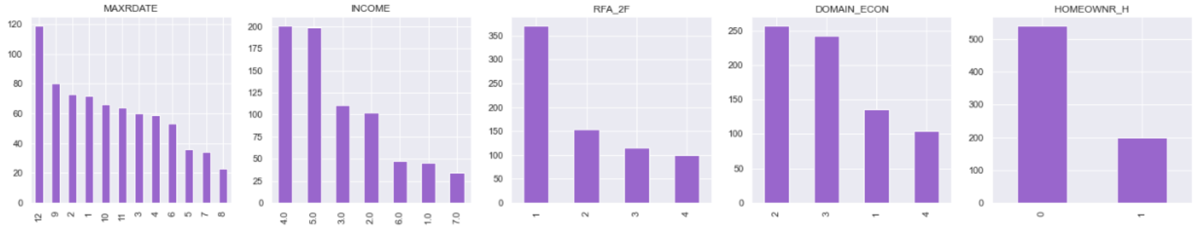


*Figure 8 - Profiling for cluster 3*

Finally, the last cluster, cluster 4 is the only from the above where the donor is not the owner of the house he lives in and the urbanity varies a lot. This is why there is not information about their neighbourhood and so we named the cluster **no info**. These donors can be moving from place to place and so this data analysed cannot be considered. The profiling for the cluster 4 was as follows (figure 9):



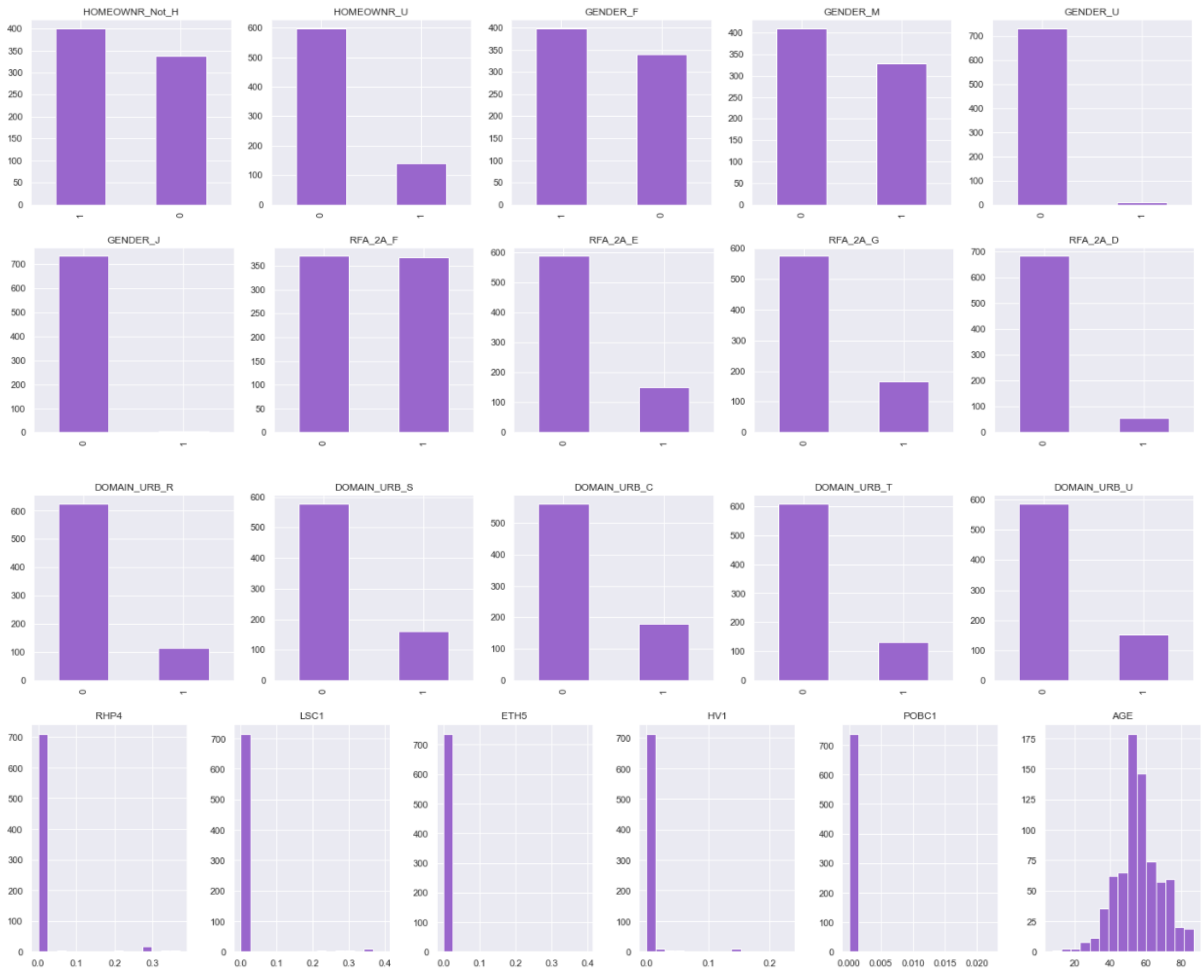


*Figure 9 - Profiling for cluster 4*

## Marketing Approaches

With this clustering analysis we realised some things were the same for the big majority of the population of donors.

December is the month when PVA receives the biggest number of donations, their amount is usually between 15 and 24.99 dollars and the majority of donors are around sixty years old. Also, the majority of donors have a Lapsed status, meaning that their last donation was 13 to 24 months ago.

Given this, the major fundraisers for all groups should be held in December and the amount requested should be that one.

For the General class we thought of doing a seminar with some life testimony of 2 or 3 of the amazing people PVA supports, this would show to all this group what PVA is all about and also is very good to reunite a large number of people. For this group and also the high one we could also sell goody baskets, since it is around Christmas time is a great present to give anyone.

Some crowdfunding though the internet is a very great idea to get to the people belonging to the low group. This way is possible to give small amounts that make a big difference. It is also an idea to sell PVA t-shirts to this class, it is always handy and by using it people promote this organisation as well.

For our second biggest group, a Christmas party should be done to raise funds. This party would have a lot of entertaining activities and will also show the PVA's work. Also, to target this group we will do some carol singing through these neighborhoods.

Since we do not have a lot of information about the No info group, we will call their attention by email, we can send cards monthly and a special card or video in December.

# Conclusion

A very big dataset with a lot of features and problems, that would damage the quality of any study, is what was given to us. Solving all these problems was the hardest part of our work, because the program used would take a lot of time to give the outputs and selecting only a few of our 476 features to apply the clustering method chosen was also difficult and required the test of a many different algorithms. Having a solution for the segmentation of our population, we labeled each group accordingly. Now we know the different people that contribute for the PVA's cause so it is possible to target them easily with specific marketing actions, leading, hopefully, to more donations for this organization.

# References

Brownlee, J. (2020, May 6). *Introduction to Dimensionality Reduction for Machine Learning*. Machine Learning Mastery. Retrieved December 20, 2020, from https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/

Brownlee, J. (2020, July 8). *4 Automatic Outlier Detection Algorithms in Python*. Machine Learning Mastery. Retrieved December 20, 2020, from https://machinelearningmastery.com/model-based-outlier-detection-and-removal-in-python/

Brownlee, J. (2020, August 17). *kNN Imputation for Missing Values in Machine Learning. Machine Learning Mastery.* Retrieved December 20, 2020, from https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/

Brownlee, J. (2020, August 18). *How to Remove Outliers for Machine Learning*. Machine Learning Mastery. Retrieved December 20, 2020, from https://machinelearningmastery.com/how-to-use-statistics-to-identify-outliers-in-data/

Chauhan, N. (n.d). *DBSCAN Clustering Algorithm in Machine Learning*. KD nuggets. Retrieved December 20, 2020, from https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html

Dabbura, I. (2018, September 7). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. Towards data science. Retrieved December 20, 2020, from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Han, J., Kamber, M., & Pei, J. (2012). Data Preprocessing. *Data Mining – Concepts and Techniques* (pp. 84-89) (3rd ed.). Morgan Kaufmann Publishers.

Kelleher, J., Namee, B., & D'Arcy, A. (2015). Data Normalization. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies* (2nd ed.). MIT Press.

Loukas, S. (2020, May 28). *Everything you need to know about Min-Max normalization: A Python tutorial.* Towards data science. Retrieved December 20, 2020, from https://towardsdatascience.com/everything-you-need-to-know-about-min-max-normalization-in-python-b79592732b79

Patlolla, C. (2018, December 10). *Understanding the concept of Hierarchical clustering Technique*. Towards data science. Retrieved December 20, 2020, from https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec
Paralyzed Veterans of America. (n.d). *OUR MISSIONS & HISTORY*. PVA. Retrieved December 20, 2020, from https://pva.org/about-us/mission-statement/

Ralhan, A. (2018, February 18). *Self Organizing Maps*. Abhinav Ralhan. Retrieved December 20, 2020, from https://medium.com/@abhinavr8/self-organizing-maps-ff5853a118d4