Paralyzed Veterans of America Donors - Data Mining Project

Andreia Taboleiros (m20200581@novaims.unl.pt)

Filipe Lourenço (r20170799@novaims.unl.pt)

Guilherme Neves (r20170749@novaims.unl.pt)

Abstract: The purpose of this project is to provide a meaningful customer segmentation of the 95 000 donors in the available dataset. To do so, this dataset was exhaustively cleaned and prepared in order to have good quality data for the clustering algorithms. Furthermore, we used some dimensionality reduction techniques to avoid the curse of dimensionality, while elaborating different perspectives of features. Then, we applied 3 different cluster algorithms and chose the best one to construct the clusters to be later profiled.

Key Words: Donors; Data Preparation; Variables; Missing Values; Outliers; Standardization; Dimensionality Reduction; Self-Organizing Maps; K-means; Mean-Shift; Clustering; Profiling; Marketing Strategies

I. Introduction

Paralyzed Veterans of America is a non-profit organization which provides services for US veterans with health problems. It is also one of the largest direct mail fundraisers in the US.

PVA's running is due to several donations throughout the years and an important factor to study is the donations time lag. It is known that the longer someone goes without donating, the less likely they will donate again. For that reason, it is important to study how donors behave, therefore existing the need to develop a Customer Segmentation. This segmentation has to be able to give us a better understanding on how donors behave and which citizens will be donating again.

We were given a sample of PVA's recent fundraising appeals, containing 95 412 donors. This is a significant amount of people which originates several data. To reduce this data after the preparation of our dataset and divide the variables into metric and non-metric variables we still needed to reduce the information that we had to do the clustering in a smarter way.

II. Methodology

Importation and initial exploration of the dataset

The data to be analyzed was provided by the Paralyzed Veterans of America (PVA) and everyone included in this database had at least one prior donation to PVA.

We started by importing the necessary libraries and importing the dataset. Then, we did some exploration on the dataset visualizing the number of features and observations, the number of duplicated rows, the names and data types of each variable as well as the variables with the highest number of missing values.

Data Cleaning and Preparation

1. Remove Variables Manually

In order to do the segmentation, we proceeded to data cleaning and preparation. Firstly, we started with the manual removal of some variables. We could see that RDATE_ and ADATE_ variables have a high number of missing values, namely each RDATE_ variable has more than 68 000 missing values, so we decided to remove them all since if we were to fill them, we would bias the data, due to the high number of missing values in relation to the total data. Furthermore, alongside the RDATE_ variables, RAMNT_ variables also present a high number of missing values, although we decided to keep these variables and treat them later.

Then, we also dropped some variables, such as AC1, AC2, WEALTH_1, WEALTH_2, GEOCODE, HHN4, HHN5, HHN6, AGE903 or AGE906, since there were variables on the dataset with similar information or with a more significant value to our analysis. Another group of variables dropped was from HC11 to HC16, since they were referring to the type of electricity the house has and we think this type of data is not relevant, according to the scope of our study. In addition, we also removed all the variables related to other types of mailing order offers as they had a significant number of missing values. Moreover, CHILD03, CHILD07, CHILD13, CHILD18 and NUMCHLD were eliminated, as they had too many missing values. In total, 73 variables were dropped, in addition to all RDATE_ and ADATE_ variables (check in the first section of the Appendices - Variables which were manually eliminated - for the whole set of deleted features).

2. Binarization

Some features in the dataset held a nominal value for either the presence or absence of that same feature. Thus, in order to improve their interpretability, we decided to replace both ‘Y’ and ‘X’ with ‘1’ and ‘N’ and ‘_’ with 0. SOLP3 and SOLIH had a category represented by a blank space, therefore we decided to fill them in with ‘20’, due to ‘1’ being already assigned to another category. On the other hand, HOMEOWNR categories were also transformed, being ‘H’ (home owner) converted to ‘1’ and ‘U’ and ‘ ‘ to ‘0’, while ‘ ‘ (Address is OK) and ‘B’ (Bad Address) in MAILCODE were converted to ‘1’ and ‘0’, respectively (check in the second section of the Appendices - Variables which were converted to binary variables - for the whole set of features which were converted to binary ones).

3. Missing Values

After doing all these cleaning in the dataset, we proceeded to fill in the missing values. Firstly, we decided to remove the rows with missing values in variable FISTDATE since there were only 2 rows. Secondly, the RAMNT_ variables' missing values were replaced by 0, since we assumed that every 'nan' in those features corresponds to the non-existence of a donation, thus being the dollar amount equal to zero. Then, GENDER blank spaces were assigned to an existing category ('U', which means 'Unknown') and the remaining blank spaces for the entire dataset were filled in with 'nan', in order to KNN imputer' algorithm to assume them as missing values. Finally, we detected RFA_ features also had missing values, thus being substituted with 'U0U', meaning 'U' absence of Recency and Amount and '0' absence of Frequency.

After the previous step, five features were left to impute their missing values with KNN imputer: NEXTDATE, GEOCODE, DOB, DOMAIN and INCOME. Accordingly, we created five new lists to aggregate the variables that were more correlated to each one of the features to be imputed:

- For NEXTDATE, we chose variables related to the time gap between donations, variables that give us the minimum date and maximum date of donation, as well as variables related to the summary of donations lifetime.
- For GEOCODE, we used variables related to the number of persons in the donor neighborhood and the percentage of population in rural and urban areas in that same neighborhood.
- For DOB, we used variables related to the percentages of population for each age group.
- For DOMAIN, we used features related to income and the type of house plumbing as that is a good indicator of which area people live (whether is urbanized or rural).
- For INCOME, we used the variables related to the percentage of income and wealth of the donors and neighbors.

To check if the variables included in each list were related to each other, we used the Pearson's correlation to evaluate if those variables had a strong relationship between them or not. However, we could not check the correlation between categorical variables.

Following the creation of lists mentioned above, we resorted to LabelEncoder to transform each category of the categorical variables with missing values to integers, in order to KNN imputer perform the missing values' imputation, since it does not accept categorical data. The 'nan' were also assumed by LabelEncoder to be a category, which is not desired. Despite this, the 'nan' are assigned to the category with the highest integer value, so these values could be easily converted again to 'nan'. Following this step, we could finally perform KNN imputer for each data frame containing all the variables correlated to the feature to be imputed, by transforming each missing value of that same feature with the value of the most similar individual, considering the values for the other variables in that data frame. Being the missing values filled in, we transformed the values of the features used in KNN imputer into the original ones and appended the filled in data frames to the original dataset. In order to not run again all the missing values' steps, we decided to save the original dataset in a csv file.

4. Transform and Creating Variables

As the dataset had a high number of features, one important step was to aggregate similar variables and create new ones. The first step was to create AGE from the DOB variable, calculating it relatively to the year of the most recent date in the dataset. Then, we created the variable FIRSTDATE, getting the oldest date between FISTDATE and ODATEDW variables, since they had the same information.

In addition, we decided to join some variables related with percentages of the population with a certain characteristic, aiming to reduce the number of variables. Accordingly, we joined the IC variables (percentage of household income), AGEC variables (percentage of adults belonging to a certain age range) and ANC variables (percentage of the population with a specific ancestry). After this merge of features, we were able to drop 18 variables. The variable DOMAIN was also transformed, being divided into DOMAIN_URBAN and DOMAIN_SOCIO, where DOMAIN_URBAN assumes value '1' when the donor lives in an urban, city or suburban neighborhood and assumes value '0' when the donor lives in a town or rural neighborhood.

Regarding the RFA_ columns, we started to replace some wrong values in variable RFA_23 with 'U0U', due to not existing the first byte for Recency. Then, we divided each RFA_ variable into three variables, where the first byte was saved into a new variable RFA_R (Recency), the second one into RFA_F (Frequency) and the third one into RFA_A (Amount). In RFA_A and RFA_R variables, we grouped some values in order to facilitate their use and access. Another variable which was also transformed was TCODE. It mentions the titles the different citizens have, however, there are several categories with few people associated to it, so we decided to just leave the titles with more frequency ('_', 'MR.', 'MRS.' and 'MS.') and aggregate the others into a single category ('1000').

When checking the STATE variable, we had to remove some rows, since there were more states on the dataset than the ones that actually exist. After that, we grouped the states in four different regions (North-East, Centre-West, South and West) for an easier use afterwards. Relatively to GENDER, it had a lot of categories, so we decided to transform it into a binary feature, called FEMALE, which holds '1' for female and '0' for the remaining ones.

Moreover, GEOCODE2 was transformed into a new variable, called Urbanized_County, where it assumes value '0' when the county size code is 'D' (meaning 'Rural County') and assumes '1' otherwise. Relatively to datetime variables, we decided to calculate the number of months between each date and the most recent date on the dataset, in order to perform some further calculations based on datetime features.

5. Outliers

In this section, we check the existence of any outliers, which are extreme observations whose values are far apart from the remaining ones of other observations. Accordingly, we started by splitting all the variables into nominal, ordinal, binary and metric features, since we can only apply Local Outlier Factor algorithm for metric variables.

The Local Outlier Factor algorithm is an unsupervised method for detecting outliers based on the local density deviation of an observation relatively to its neighbors. In other words, it compares the local density of that individual with the local densities of its neighbors, being this observation an outlier if its local density is lower than the ones of its neighbors or if the anomaly score for that same individual is higher than the ones corresponding to the individuals in its neighborhood. By applying this outlier detection technique, we dropped 1.4% of the data. Despite that, we wanted to check for each metric variable if there were still extreme observations and we found some of them, which resulted in 0.8% of eliminated data. This way, 2.2% of the total data was dropped out from the original dataset.

6. Coherence Checking

In this section, our group searched for possible inconsistencies among the data and features (check in the third section of the Appendices - Incoherencies found - for the inconsistencies).

7. One Hot Encoding

Furthermore, we decided to use One Hot Encoding in ordinal and nominal variables to help us on the profiling of the clusters. Relatively to the hyperparameters, only ‘drop’ was set to a different label from the default, as it was set to ‘first’, in order to avoid perfect multicollinearity between binary variables originated from categories of the same feature.

8. Data Standardization

The dataset has some features whose scales are very different from each other, which makes it harder to perform an objective analysis. To solve this problem, we had to perform some techniques for scaling the data, namely Normalization, MinMax Scaler, Standard Scaler and Robust Scaler. Finally, we decided to use Standard Scaler since outliers were already removed, because this scaler «does not guarantee balanced feature scales» (geeksforgeeks.org, 2020), and it transforms the distribution to be centered around 0 (mean equals 0 and variance 1), an assumption required by some algorithms.

Dimensionality Reduction

1. Correlations per perspective

This step is one of the most important ones in the entire Knowledge Discovery Database process, since the clustering solutions highly depend on the discrimination ability and relevancy of the set of variables upon which clustering is made.

Firstly, we decided which perspectives would be the most relevant to assign features to, aiming to further perform clustering on top of each one and apply Principal Components Analysis. Hence, our group chose to adopt a more conservative approach by choosing all features thought to be more related to each one of the groups and thus assigning to them.

The chosen perspectives were:

- **'vars_ClientValue'**: Features related to the client value for PVA, thus variables which relate to recency, frequency and amount of a given, for instance.
- **'vars_Demographic'**: Features which convey information about the percent of population in a neighborhood for each age range, number of people living in a certain area, people's ancestry provenience and people's nationality, among others.
- **'vars_Social'**: Features related to various social indicators, such as percent of population in each education degree, household conditions and composition, number of units per housing unit or marital status.
- **'vars_Economical':** Feature which have to do with income and employment, such as household income, percent of population in a neighborhood employed in several sectors, housing rent or housing value.

After this initial choice of features, several variables were recursively dropped out from each group, by analyzing the Pearson Correlation heat map for each perspective. Accordingly, it was only kept a set of few features for each group, the ones with higher relevancy and correlation among them, excluding redundant features, in order to not fall in the curse of dimensionality and to obtain more accurate clustering solutions (check in the fourth section of the Appendices - Pearson's Correlation Heat Maps for each Perspective - for the final correlation matrices with the variables chosen for each perspective)

Following the previous selection, our group proceeded to combine both metric features and binary features for each perspective, ensuring that we could try different clustering solutions for each perspective, so we could then choose which set of variables would enable to achieve the best one.

2. Principal Components Analysis

In this sub-section, we implemented PCA, a dimensionality reduction technique which could be very useful to reduce the number of features to work with, while guaranteeing much of their variance (information) would be explained by the Principal Components (PC's). This new set of linearly uncorrelated set of features were obtained from the whole set of features chosen in the previous sub-section, ensuring PC's had as much information as possible from each perspective.

Assuming the previous statement, the eigenvalues (variance explained by each one of the PC's) and the percentages of total and cumulative variance for each PC were computed, so we could determine a suitable number of Principal Components. Accordingly, the goal was to not keep too many of them, to not incur into the curse of dimensionality, and not too less, so it could be lost important information. Therefore, we based on three criteria to choose how many PC's to retain, namely:

- **Kaiser criterion**: keep all Principal Components whose eigenvalue (explained variance) is higher than 1. Consequently, we would choose 5.
- **Proportion of explained variance**: keep all PC's which explain at least 80% of the total variance. This criterion shows 9 is the suitable number of PC's.
- **Scree Plot**: by looking at the scree plot, we should choose the number of PC's to retain which corresponds to the elbow. Resorting to this technique, we would choose 5 PC's.

Considering the previous criteria, we decided to choose 5 Principal Components, as Kaiser criterion and Scree Plot show, although the proportion of explained variance criterion advised to choose 9 PC's, a number which might be too high, hardening the interpretation of the PC's. Despite this, with 5 Principal Components, 70% of the total variance of the original variables is still retained. After that, a new data frame was built, holding the values for each PC, followed by calculating the loadings (correlations between the PC's and the variables from which they were built), which allowed to assess the relationships between the Principal Components (check in the fifth section of the Appendices – Principal Components Analysis – for the table which shows the eigenvalues and the proportion of variance explained by each cluster, as well as the Scree Plot and the loadings' table).

A-priori Grouping - Cohort Analysis

After applying dimensionality reduction techniques, we decided to perform a Cohort Analysis, both for absolute retention numbers and retention rate (check in the sixth section of the Appendices – A-Priori Grouping - Cohort Analysis). The main goal of this technique is to better understand the behavior of PVA donors through time, by perceiving the donations trend from each cohort group. By crossing those two types of analysis, it can be seen there were months which registered a higher number of donations, such as December 2014 and December 2015, due to Christmas Time, a time of the year when usually people donate more; July and September 2014; April and June 2015. On the other hand, in the months of February, March, July and August of 2015 were when people donated less.

Clustering

1. K-means

One of the approaches for performing clustering which our group attempted was to use k-means algorithm. Accordingly, the lists of variables created previously for each perspective were used to do a segmentation of the individuals based on each one of them, in order to assess which perspective would deliver the best results (the one which could lead to better interpretations and higher R-squared).

For every perspective, the inertia and the average silhouette coefficient for each number of clusters were computed, so it could be assessed which number of clusters could suit the best to the given problem. Therefore, an inertia plot was drawn in order to find the elbow, corresponding to the number of clusters which minimizes the within-cluster sum of squared distances, ensuring the individuals in each cluster are similar. An average silhouette coefficient plot was also built for each number of chosen clusters, being the goal to maximize this coefficient, meaning less individuals are wrongly assigned to each cluster and that each individual in a cluster is far away from the ones belonging to neighboring clusters. An average silhouette coefficient graphic was also plotted for finding the elbow corresponding to the number of clusters which maximizes this coefficient.

After combining the information given by each plot, the appropriate number of clusters was chosen and k-means was performed. Since the results of this algorithm highly depend on the initial seeds, we set the hyperparameter ‘init’ to be equal to ‘k-means++’, so the initial seeds would be far apart from each other.

Three different approaches for assessing the quality of the clustering solution for each perspective were made. Firstly, each observation was assigned to each cluster and the mean values of each cluster for each variable chosen to perform clustering were computed, hence, evaluating if the mean values of each cluster could distinguish their individuals. Secondly, the number of observations pertaining to each cluster was checked and thirdly, the R-squared of the clustering solution was computed.

Analyzing the results delivered by k-means algorithm for each perspective, it can be said that they were very poor, since the mean values for each cluster were very similar to each other and the R-squared never surpassed 0.03, meaning the individuals in each cluster were not similar to each other and they were not very different from the individuals belonging to other clusters.

2. Self-Organizing Maps

Self-Organizing Maps (SOM) are 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 (Abhinav Ralhan, 2018). Besides being a dimensionality reduction technique, SOM is applied for multidimensional data visualization, cluster detection, market segmentation or outlier detection.

In our case, we decided to use a SOM with 2500 units (50x50 grid), a random initialization and a gaussian neighborhood function where we would apply K-means and Hierarchical Clustering on top of those units. The training of the SOM was done in two phases since it becomes more effective. In unfolding phase, the main goal is to spread the units in input space and to get the data patterns and, in fine tuning phase, the goal is to reduce the quantization error (the difference between the data points and the neurons that represent them).

As explained previously, in Dimensionality Reduction part we divided all the features in four different perspectives and chose the most relevant ones, ending with two set of features for each perspective, making a total of eight sets. So, we decided to apply SOM in each one of the eight sets of features and, also, in the principal components resulting from the Principal Components Analysis. For each set of features, we applied K-means on top of the resulting 2500 units where we chose the number of clusters based in the inertia plot and, in the same way, we applied Hierarchical Clustering on top of the 2500 units where we chose the number of clusters based on the dendrogram.

3. Density-Based Clustering – Mean-Shift

Besides k-means algorithm and SOM, our group resorted to another clustering algorithm to get an appropriate clustering solution, called Mean-Shift Algorithm. This algorithm tries to find areas with higher density of points by moving the centroids of each sliding window towards those zones. It has one important hyperparameter called ‘bandwidth’, which defines the size of the sliding windows and thus, the calculation of the mean of the points inside of them for computing the centroids. The appropriate number for this hyperparameter can be estimated through a function called ‘estimate_bandwidth’, which given a predetermined quantile, determines a suitable bandwidth. Finding an appropriate quantile is also fundamental, since «the estimated bandwidth increases with an increase in quantile resulting in a smaller number of clusters. Similarly, decrease in quantile decreases the bandwidth and hence higher no. of clusters.» (https://stackoverflow.com/, 2015). Therefore, after trying several quantiles, it was checked which one could deliver the best clustering results for each perspective, taking into account the number of observations in each cluster, as well as the R-squared.

In the end, the results for each perspective were evaluated. The achieved results were not great, since almost all the observations were gathered in a single cluster (around 95%) for all perspectives, with minor differences, and the R-squared for each clustering solution in each perspective was also very low, being 0.2476 the maximum achieved R-squared, for the Income and Employment Perspective, segmenting the observations in 10 clusters.

III. Results and Discussion

Self-Organizing Maps (SOM) – Choosing the best clustering solution

Since the results obtained with K-means and Mean-Shift algorithms were not so good, we decided only to evaluate and to compare the results between the different perspectives using SOM.

In order to assess and choose the best clustering approach, we used the R-squared coefficient to understand the proportion of variance that each clustering approach can explain. In the graph below, the R-squared values are plotted for the set of features that achieved the highest values for the different perspectives and, also, for the principal components. In general, clustering with K-means on top of the SOM units delivered higher R-squared values than with Hierarchical Clustering. We can also see that using the principal components, we got poorer results when comparing with the set of features from the different perspectives. Relatively to the sets of features from the different perspectives, we obtained the highest values with the two sets of features from the economical perspective when the K-means was used, achieving both set of features a R-squared above 0.7. The difference between these two R-squared is really small, however, the Economical2 has a slightly better R-squared, so we chose this one where we got 7 clusters using K-means.

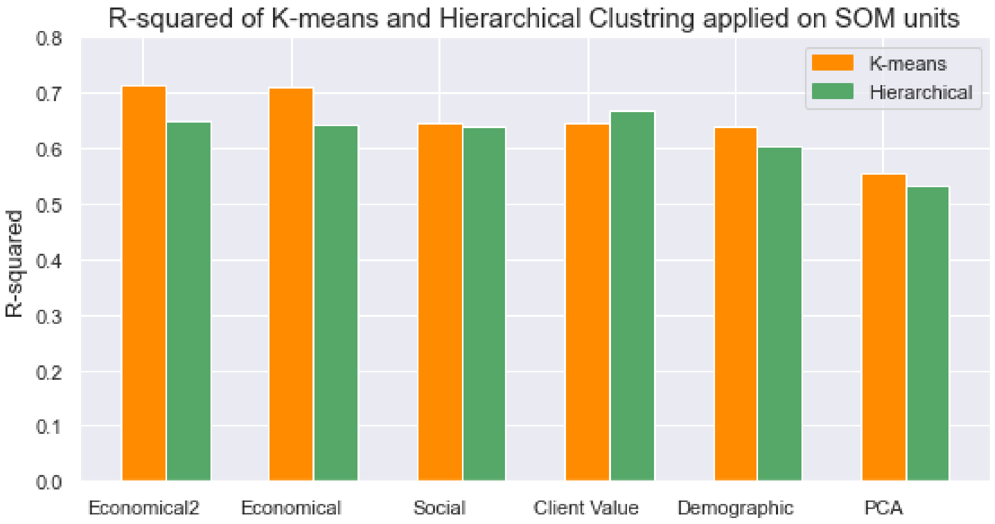


Figure 1 - R-squared of the best SOM

Cluster profiling

Our clustering solution has 7 different clusters, where clusters 4 and 5 contain more than 20 000 observations and clusters 0 and 1 are the smallest, containing each one around 6 000 observations, as shown in the graph below.

Furthermore, to choose the variables used for profiling those clusters, we took into account the standard deviation for each feature; the discrimination ability of the features by looking at the parallel coordinates plot; the meaning of each variable, ensuring that features with similar meaning would not be chosen for interpreting the clusters; the mean values of the non-standardized features for each cluster:

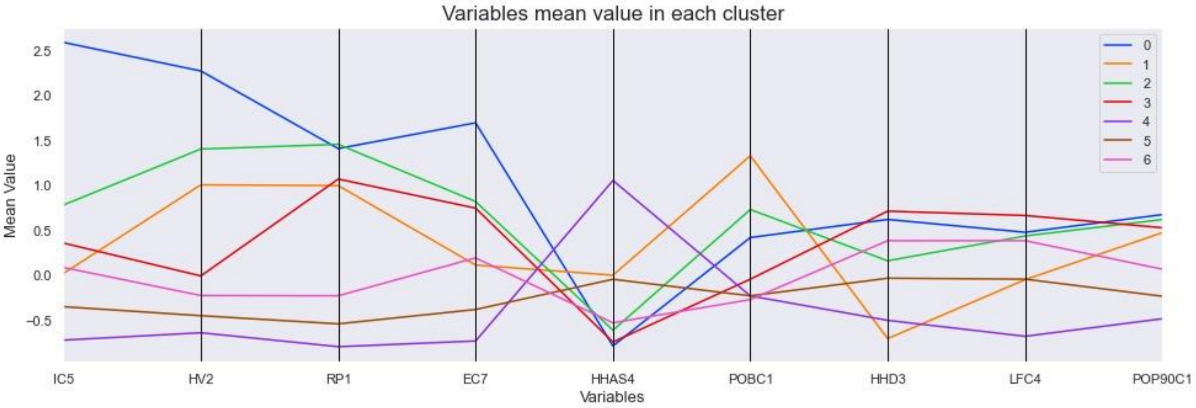


Figure 2 - Standardized variables mean values per cluster

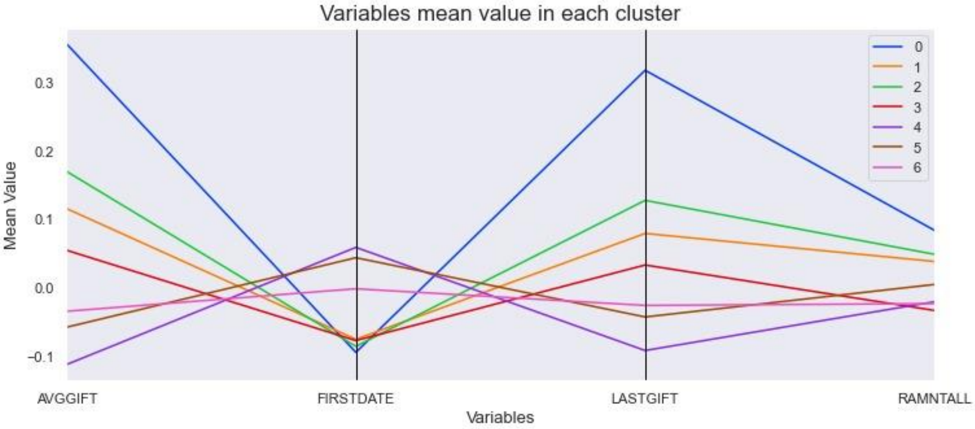
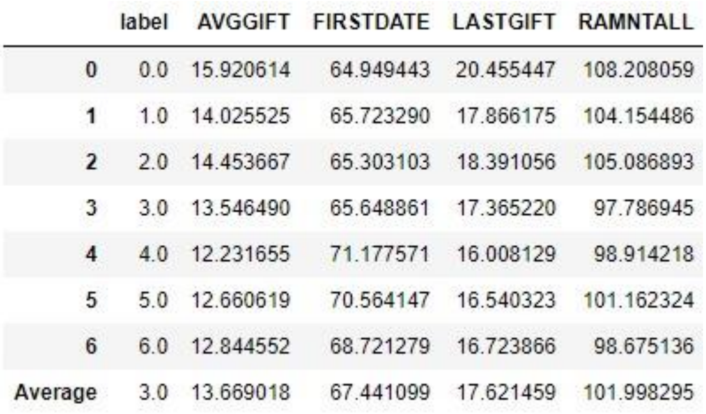
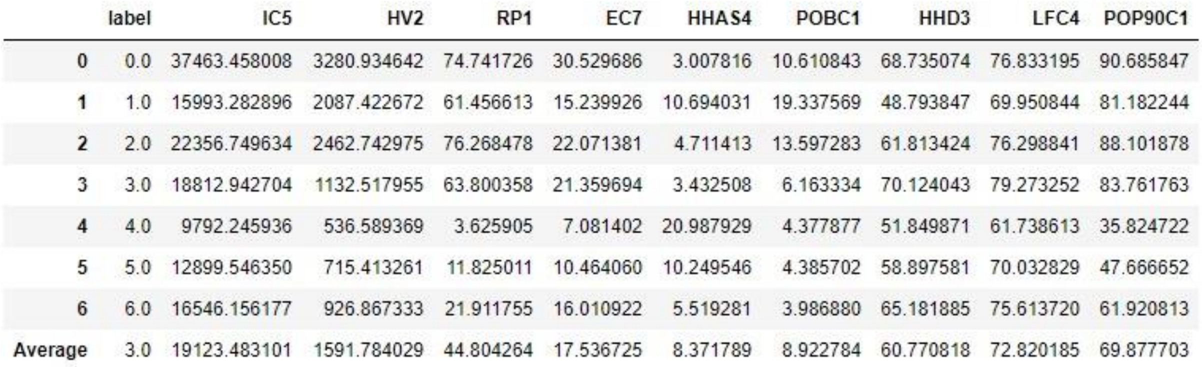


Figure 3 - Standardized variables mean value per cluster



	label	AVGGIFT	FIRSTDATE	LASTGIFT	RAMNTALL
0	0.0	15.920614	64.949443	20.455447	108.208059
1	1.0	14.025525	65.723290	17.866175	104.154486
2	2.0	14.453667	65.303103	18.391056	105.086893
3	3.0	13.546490	65.648861	17.365220	97.786945
4	4.0	12.231655	71.177571	16.008129	98.914218
5	5.0	12.660619	70.564147	16.540323	101.162324
6	6.0	12.844552	68.721279	16.723866	98.675136
Average	3.0	13.669018	67.441099	17.621459	101.998295

Figure 4 – Non-standardized variables mean values per cluster



	label	IC5	HV2	RP1	EC7	HHAS4	POBC1	HHD3	LFC4	POP90C1
0	0.0	37463.458008	3280.934642	74.741726	30.529686	3.007816	10.610843	68.735074	76.833195	90.685847
1	1.0	15993.282896	2087.422672	61.456613	15.239926	10.694031	19.337569	48.793847	69.950844	81.182244
2	2.0	22356.749634	2462.742975	76.268478	22.071381	4.711413	13.597283	61.813424	76.298841	88.101878
3	3.0	18812.942704	1132.517955	63.800358	21.359694	3.432508	6.163334	70.124043	79.273252	83.761763
4	4.0	9792.245936	536.589369	3.625905	7.081402	20.987929	4.377877	51.849871	61.738613	35.824722
5	5.0	12899.546350	715.413261	11.825011	10.464060	10.249546	4.385702	58.897581	70.032829	47.666652
6	6.0	16546.156177	926.867333	21.911755	16.010922	5.519281	3.986880	65.181885	75.613720	61.920813
Average	3.0	19123.483101	1591.784029	44.804264	17.536725	8.371789	8.922784	60.770818	72.820185	69.877703

Figure 5 – Non-standardized variables mean values per cluster

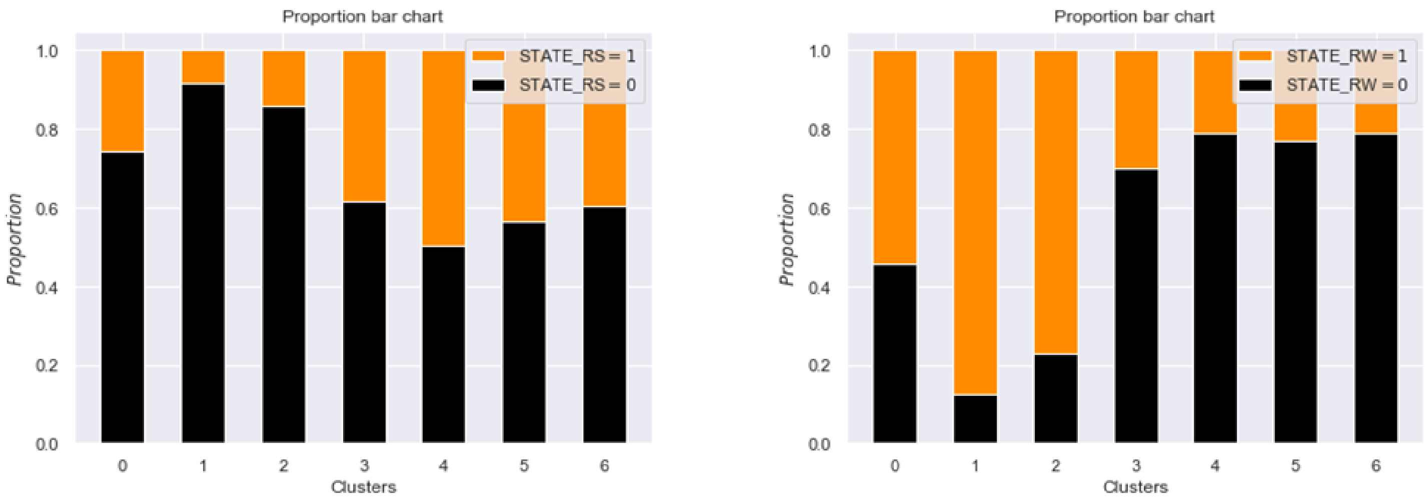


Figure 6 - Proportion of '1's and '0's for variables 'STATE_RS' and 'STATE_RW' per cluster

Considering the previous assumptions and plots shown above, the clusters' profiling was done as the following:

Cluster 0: This cluster is characterized as being composed by individuals who live in neighborhoods where, in average, the per capita Income is $37,463; the average home value is around $328,000; 75% of the renters pay more than $500 per month for their rent; 31% of the adults aged 25+ have a bachelor degree; just 3% of the people live below the poverty level; 90% of the population live in an urbanized area. Relatively to the Client Value features, these people donated, in average, $16 per gift; donated for the first time around 2.5 months later than the average; donated for the last time an amount of $20.5, in average; each person donated, in average, a total of $108 in his/her life to PVA. Finally, by looking at the plots showing the discrimination between classes of binary features, it can be said that the majority of the people in this cluster live in a Western state. Therefore, this cluster can be labeled as **'Wealthy and high education level donors who live in urbanized areas, having donated a large amount of money and made the first donation more recently'**.

Cluster 1: This group of donors can be described as being composed of donors who live in a neighborhood where, in average, the per capita Income is $15,993; the average home value is around $208,700; 61% of the renters pay more than $500 per month for their rent; 19% of the people were born in foreign countries; just 48% of the families are headed by a married couple; 81% of the population live in an urbanized area. Considering the Client Value variables, these individuals donated for the first time around 2 months later than the average; each person donated, in average, a total of $104 in his/her life to PVA. Now looking at the plots showing the discrimination between classes of binary features, it can be seen that the majority of the people in this group live in a Western state. Hence, this cluster can be labeled as **'Medium income donors who live in urbanized areas, where one-fifth of the population is foreign, and have donated a large amount of money'**.

Cluster 2: It can be defined as a cluster whose donors live in a neighborhood where, in average, the per capita Income is $22,356; the average home value is around $246,200; 76% of the renters pay more than $500 per month for their rent; 22% of the adults aged 25+ have a bachelor degree; just 5% of the people live below the poverty level; 14% of the people were born in foreign countries; 88% of the population live in an urbanized area. Considering the Client Value features, these individuals donated, in average, $14.5 per gift; donated for the first time around 2 months later than the average; donated for the last time an amount of $18.4, in average; each person donated, in average, a total of $105 in his/her life to PVA. Finally, by looking at the graphs showing the discrimination between classes of binary variables, it can be concluded that the majority of the people in this cluster live in a Western state. Consequently, this cluster can be labeled as **'Medium-high income donors who live in urbanized areas and have donated a large amount of money'**.

Cluster 3: This cluster is composed by donors who live in a neighborhood where, in average, 64% of the renters pay more than $500 per month for their rent; 21% of the adults aged 25+ have a bachelor degree; just 3% of the people live below the poverty level; 70% of the families are headed by a married couple; 79% of adult males are employed; 83% of the population live in an urbanized area. Taking into account the Client Value variables, these persons donated for the first time around 2 months later than the average; each person donated, in average, a total of $98 in his/her life to PVA. By analyzing the plots showing the discrimination between classes of binary features, it can be said that the majority of the people in this cluster live in a Western or a Southern state. In consequence, this cluster can be labeled as **'Medium income donors who live in urbanized areas, with less unemployment, having donated a small amount of money'**.

Cluster 4: The individuals pertaining to this cluster live in a neighborhood where, in average, the per capita Income is $9,792; the average home value is around $53,600; just 4% of the renters pay more than $500 per month for their rent; just 7% of the adults aged 25+ have a bachelor degree; 21% of the people live below the poverty level; just 4% of the people were born in foreign countries; just 61% of adult males are employed; just 35% of the population live in an urbanized area. Considering the Client Value features, these set of individuals donated, in average, $12.2 per gift; donated for the first time around 3.5 months earlier than the average; donated for the last time an amount of $16, in average; each person donated, in average, a total of $99 in his/her life to PVA. Considering the plots showing the discrimination between classes of binary variables, it can be said that most people in this cluster live in a Southern state. Hence, this cluster can be labeled as **'Low income and education level donors, born in the USA, who live in rural areas, with higher unemployment, having donated for longer, although donations have been small'**.

Cluster 5: This group is characterized as being constituted by donors who live in a neighborhood where, in average, the per capita Income is $12,899; the average home value is around $71,500; just 12% of the renters pay more than $500 per month for their rent; just 10% of the adults aged 25+ have a bachelor degree; just 4% of the people were born in foreign countries; 47% of the population live in an urbanized area. By looking at the Client Value features, these group of individuals donated, in average, $12.6 per gift; donated for the first time around 3 months earlier than the average; donated for the last time an amount of $16.5, in average. By analyzing the graphs showing the discrimination between classes of binary variables, it can be said that most people in this set of individuals live in a Southern state. Accordingly, this cluster can be labeled as **'Low-medium income donors, born in the USA, having donated for longer'**.

Cluster 6: This set of individuals can described as donors who live in a neighborhood where, in average, the per capita Income is $16,546; the average home value is around $92,600; 21% of the renters pay more than $500 per month for their rent; 6% of the people live below the poverty level; just 4% of the people were born in foreign countries; 62% of the population live in an urbanized area. Taking into consideration the Client Value features, these people donated, in average, $12.8 per gift; donated for the first time around 1 months earlier than the average; donated for the last time an amount of $16.7, in average; each person donated, in average, a total of $99 in his/her life to PVA. Considering the plots showing the discrimination between classes of binary features, it can be concluded that most people in this cluster live in a Southern state. Thus, this cluster can be labeled **as 'Medium income donors, born in the USA, who have donated a not so large amount of money comparing to their salary'**.

Note: For those variables whose clusters presented similar values to the mean values, they were not stated above for profiling the clusters.

IV. Conclusion / Marketing Strategies

Regarding **cluster 0**, it contains the wealthier donors and the ones who have donated a larger amount of money so it is important to retain and to encourage them to donate, for example, through a **membership card**, which could give access to some promotions in supermarkets, gyms or restaurants, etc., if they reach a certain amount or frequency of donations per semester. These membership cards could also be applied for donors belonging to **clusters 1 and 2**, since these people also donated a significant amount of money. Taking into account the Cohort Analysis done previously, we noticed that in December, due to Christmas Time, there are a higher number of donations. So, for these donors who have donated more (**clusters 0, 1 and 2**), it could be sent **emails** in other times of the year besides Christmas, to remind them the importance of donating, so PVA can survive.

Furthermore, **clusters 3 and 6** are composed of donors who earn a medium income, but do not donate such a high amount of money compared to their income level, thus, it is very important to encourage these individuals to donate more. Besides the already mentioned marketing approaches, a few ones could be created, aiming to specifically target these two groups. For instance, PVA could send **personalized e-mails and postcards** for their postal code, advising for donating more and aiming to raise awareness among these people for the urge of donating to PVA, so they could continue to develop their activity and reaching out more disabled military veterans.

Relatively to campaigns to **all PVA donors**, it could be launched **fundraising campaigns** in cities in Western and Southern states of USA, since the majority of the donors belonging to these clusters live in those states, in locations such as shopping centers, big supermarkets or principal avenues of each city. Another approach would be to place **billboards** on the main highways of these cities, thus reaching a higher number of people, as well as PVA organizing **charity events**, such as **solidarity concerts** and **friendly sports' matches**, whose price of the tickets would revert for PVA. On the other hand, PVA could also launch **announcements** on some TV channels and social media, so a high number of people could be reached out.

Note: We considered that every donor was a Lapsed Donor, since all values for the variable 'RFA_2R' are 'L'.

V. References

Hale J., 2019, “Scale, Standardize, or Normalize with Scikit-Learn”, 2020, https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02

GeeksforGeeks, 2020, “StandardScaler, MinMaxScaler and RobustScaler techniques – ML”, 2020, https://www.geeksforgeeks.org/standardscaler-minmaxscaler-and-robustscaler-techniques-ml/

Ralhan A., 2018, “Self Organizing Maps”, 2020, https://medium.com/@abhinavr8/self-organizing-maps-ff5853a118d4

StackExchange, 2019, https://softwareengineering.stackexchange.com/questions/388313/how-to-define-the-bandwidth-in-mean-shift-clustering

Stackoverflow, 2015, https://stackoverflow.com/questions/28335070/how-to-choose-appropriate-quantile-value-while-estimating-bandwidth-in-meanshift

Scikit-learn, 2020, “Selecting the number of clusters with silhouette analysis on KMeans clustering”, 2020, https://scikitlearn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

VI. Apendices

1. Variables which were manually eliminated:

- **Census variables:**

AC1, AC2, HC11, HC12, HC13, HC14, HC15, HC16, HHN4, HHN5, HHN6, CHILC1, CHILC2, CHILC3, CHILC4, CHILC5, AGE903, AGE906, HC3, HC4, HC5, HC6, HC7, HC8, VOC2, VOC3, HVP1, HVP3, HVP4, RP2, RP3, IC15, IC16, IC17, IC18, IC19, IC20, IC21, IC22, IC23.

- **Variables related with the number of known times the donor has responded to other types of mail order offers:**

PUBCULIN, PUBHLTH, PUBDOITY, PUBNEWFN, PUBPHOTO, PUBOPP, PUBGARDN, MBCRAFT, MBGARDEN, MBBOOKS, MBCOLECT, MAGFAML, MAGFEM, MAGMALE.

- **RDATE and ADATE variables:**

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, 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.

- **Other variables:**

WEALTH1, WEALTH2, MSA, DMA, ADI, ZIP, OSOURCE, GEOCODE, MDMAUD, RFA_2, DATASRCE, PVASTATE, CHILD03, CHILD07, CHILD12, CHILD18, NUMCHLD, LIFESRC.

2. Variables which were converted to binary variables:

- **Variables whose '1' was encoded as 'X' and '0' as '_':**
  RECINHSE, RECP3, RECPGVG, RECSWEEP, MAJOR, PEPSTRFL.

- **Variables whose '1' was encoded as 'Y' and '0' as '_':**
  COLLECT1, VETERANS, BIBLE, CATLG, HOMEE, PETS, CDPLAY, STEREO, PCOWNERS, PHOTO, CRAFT, FISHER, GARDENIN, BOATS, WALKER, KIDSTUFF, CARDS, PLATES.

- **Variables with another types of encoding:**
  NOEXCH, SOLP3, SOLIH, HOMEOWNR.

3. Incoherencies found:

1. The sum of 'Percent Population in Urbanized Area', 'Percent Population Outside Urbanized Area' and 'Percent Population Inside Rural Area' cannot be lower than 99% (because of rounding errors) or higher than 101% (because of rounding errors).
2. The sum of 'Percent of Male' with 'Percent of Female' cannot be lower than 99% (because of rounding errors) or higher than 101% (because of rounding errors).
3. The sum of 'Percent of Children Under Age 7', 'Percent of Children Age 7-14' and 'Percent of Children Age 14-17' cannot be lower than 99% (because of rounding errors) or higher than 101% (because of rounding errors).
4. The sum of the all AGEC variables cannot be lower than 97% (because of rounding errors) or higher than 103% (because of rounding errors).
5. The sum of 'Percent of Married', 'Separated or Divorced', 'Widowed' and 'Never Married' must not be lower than 99% (because of rounding errors) or higher than 101% (because of rounding errors).
6. 'Percentage of Households with a person older than 65' cannot be lower than 'Percentage of Households with a person older than 65 living alone'.
7. The sum of 'Percent of 1 Person Households' with 'Percent of 2 Person Households' and 'Percent of 3 or more Persons Households' must not be lower than 99% (because of rounding errors) or higher than 101% (because of rounding errors).
8. The sum of 'Percent Owner Occupied Housing Units' with 'Percent Renter Occupied Housing Units' cannot be lower than 99% (because of rounding errors) or higher than 101% (because of rounding errors).
9. The sum of 'Percentage of Occupied Housing Units' with 'Percentage of Vacant Housing Units' cannot be higher than 101% or lower than 99% (because of rounding errors).
10. 'Percent Households w/ Related Children' cannot be higher than the 'Percent Households w/ Families'.

11. ‘Percent Married Couples w/ Related Children’ cannot be higher than ‘Percent Married Couple Families’.
12. The sum of ‘Percent Persons in Family Household’ with ‘Percent Persons in Non-Family Household’ cannot be higher than 101% or lower than 99% (because of rounding errors).
13. The sum of ‘Percent Male Householder w/ Child’ with the ‘Percent Female Householder w/ Child’ cannot be higher than ‘Percent Single Parent Households + 1 (because of rounding errors) or lower than ‘Percent Single Parent Households’ - 1 (because of rounding errors).
14. The sum of ETHC1, ETHC2 and ETHC3 must not be lower than ETH1 - 1 (because of rounding errors) or higher than ETH1 + 1 (because of rounding errors).
15. The sum of ETHC4, ETHC5 and ETHC6 must not be lower than ETH2 - 1 (because of rounding errors) or higher than ETH2 + 1 (because of rounding errors).
16. ‘Percent Adult Veterans Age 16+’ cannot be lower than the sum of ‘Percent Male Veterans Age 16+’ with ‘Percent Female Veterans Age 16+’.
17. ‘Number of lifetime gifts to card promotions to date’ cannot be higher than the ‘Number of lifetime gifts to date’.
18. ‘Dollar amount of smallest gift to date’ cannot be higher than ‘Dollar amount of largest gift to date’.
19. Date of first gift’ cannot be more recent than ‘Date of second gift’.
20. The sum of all RAMNT cannot be higher than RAMNTALL.
21. RAMNT count (except RAMNT = 0) has to be lower or equal to NGIFTALL.
22. RAMNT count for each person cannot be higher than CARDGIFT for the following promotions: FS, GK, TK, SK, NK, XK, UF, UU.
23. MINRAMNT must not be higher than the lowest RAMNT (excluding 0).
24. MAXRAMNT must not be lower than the highest RAMNT.
25. FEMALE cannot be ‘1’ and TCODE being ‘1’ (donor title code equal to MR.).

4. Pearson's Correlation Heat Maps for each Perspective

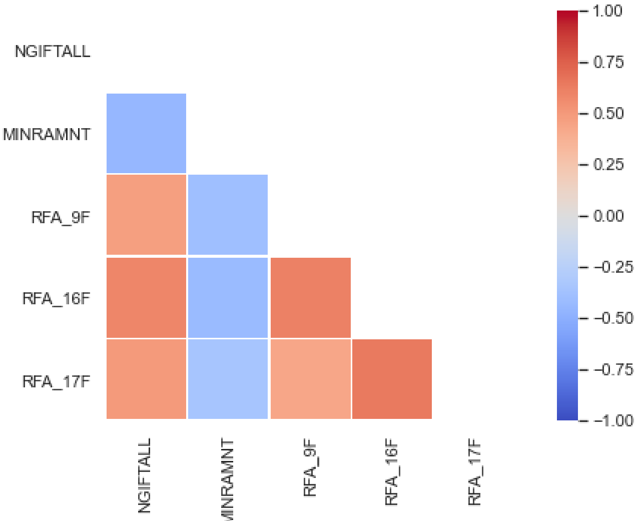


Figure 7 - Pearson's Correlation - ClientValue Perspective

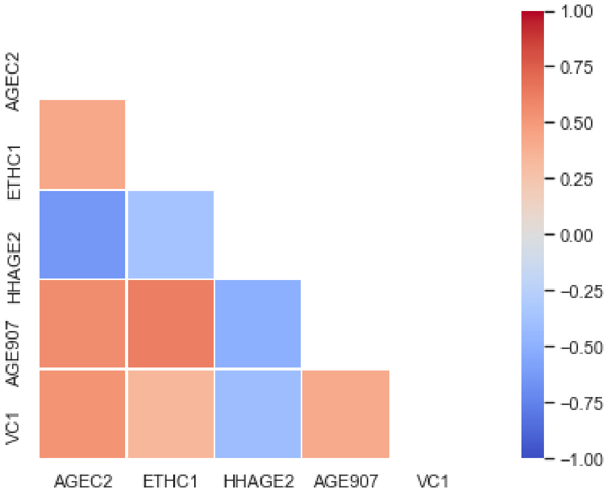


Figure 8 - Pearson's Correlation - Demographic Perspective

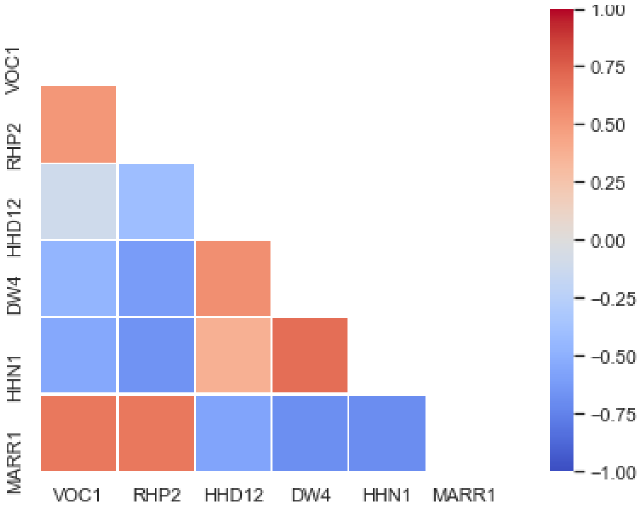


Figure 9 - Pearson's Correlation - Social Perspective

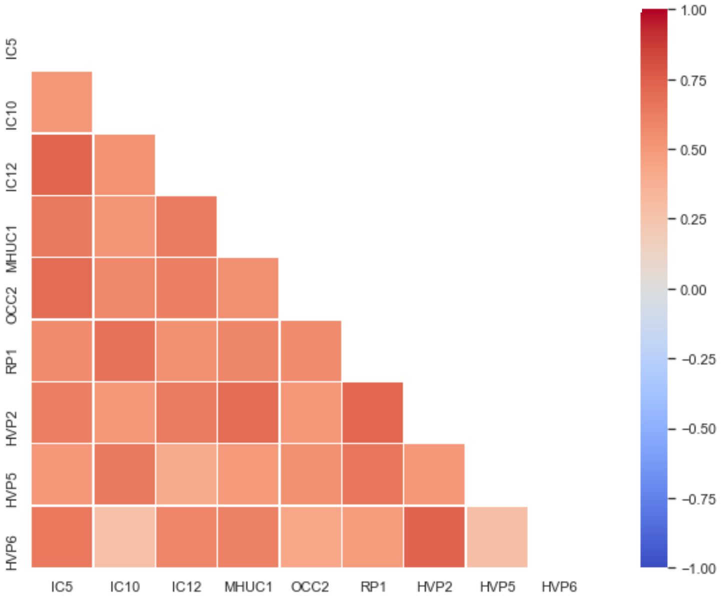
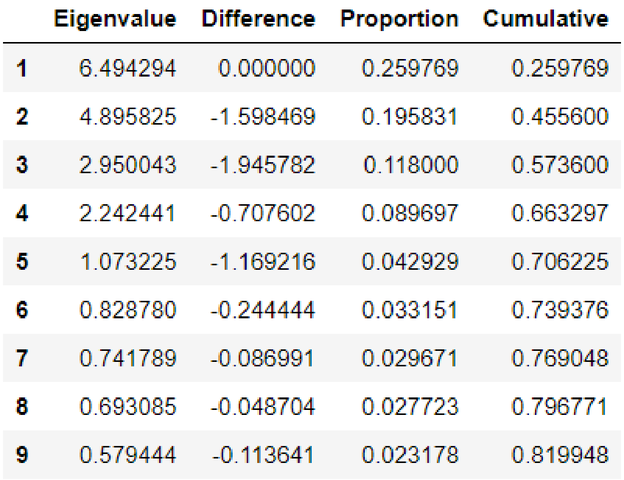


Figure 10 - Pearson's Correlation - Economical Perspective

5. Principal Components Analysis



	Eigenvalue	Difference	Proportion	Cumulative
1	6.494294	0.000000	0.259769	0.259769
2	4.895825	-1.598469	0.195831	0.455600
3	2.950043	-1.945782	0.118000	0.573600
4	2.242441	-0.707602	0.089697	0.663297
5	1.073225	-1.169216	0.042929	0.706225
6	0.828780	-0.244444	0.033151	0.739376
7	0.741789	-0.086991	0.029671	0.769048
8	0.693085	-0.048704	0.027723	0.796771
9	0.579444	-0.113641	0.023178	0.819948

Figure 11 - PCA Table (Eigenvalues and Variance Explained)

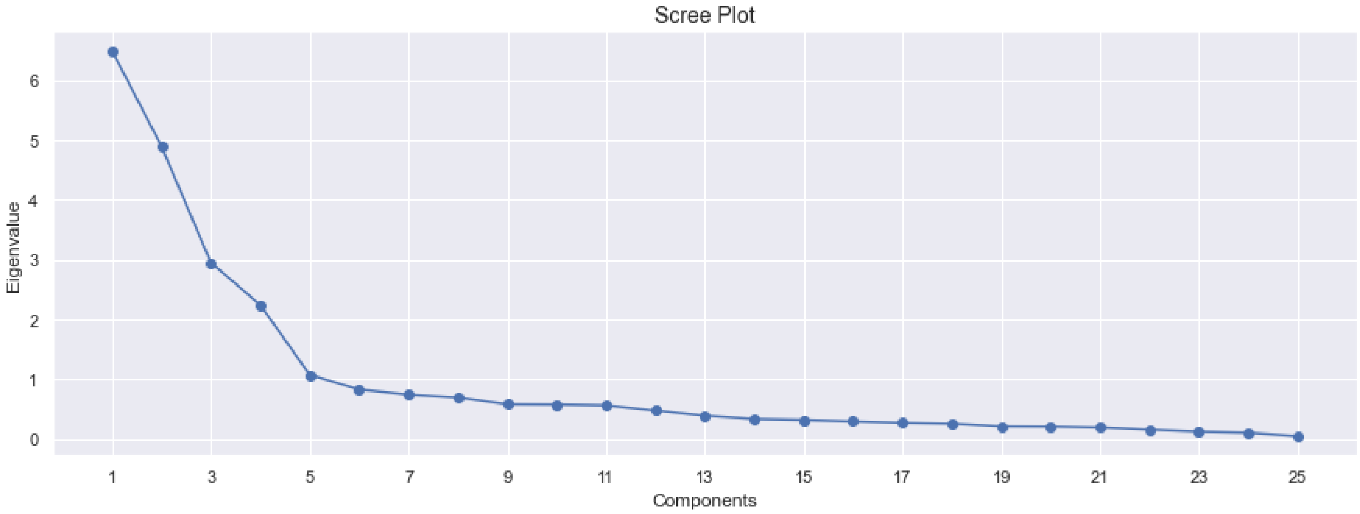


Figure 12 - PCA Scree Plot

Principal Components	**Highly Positively Correlated (>0.7)**	**Positively Correlated**	**Negatively Correlated**	**Highly Negatively Correlated (<-0.7)**
PC0		HHN1, HHAGE2	VOC1, RHP2, MARR1, OCC2, HVP2, HVP5, HVP6	IC5, IC10, IC12, MHUC1, RP1
PC1	DW4, HHN1	HHD12, HHAGE2, IC5, HVP2, HVP6	RHP2, MARR1	ETHC1, AGE907
PC2		MINRAMNT		NGIFTALL, RFA_9F, RFA_16F, RFA_17F
PC3	AGEC2	HHD12, DW4, VC1	MARR1, HHAGE2	
PC4		HVP6	VOC1, HVP5	

Figure 13 - Table with Variables Correlated with Principal Components

6. A-Priori Grouping – Cohort Analysis

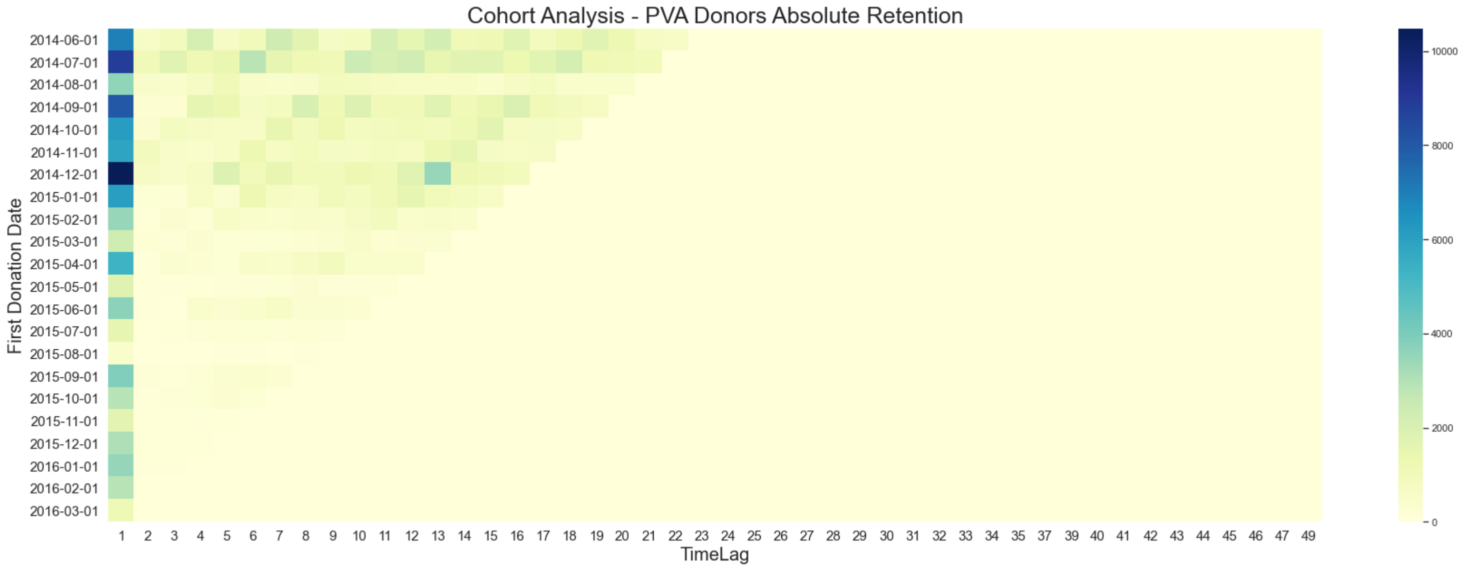
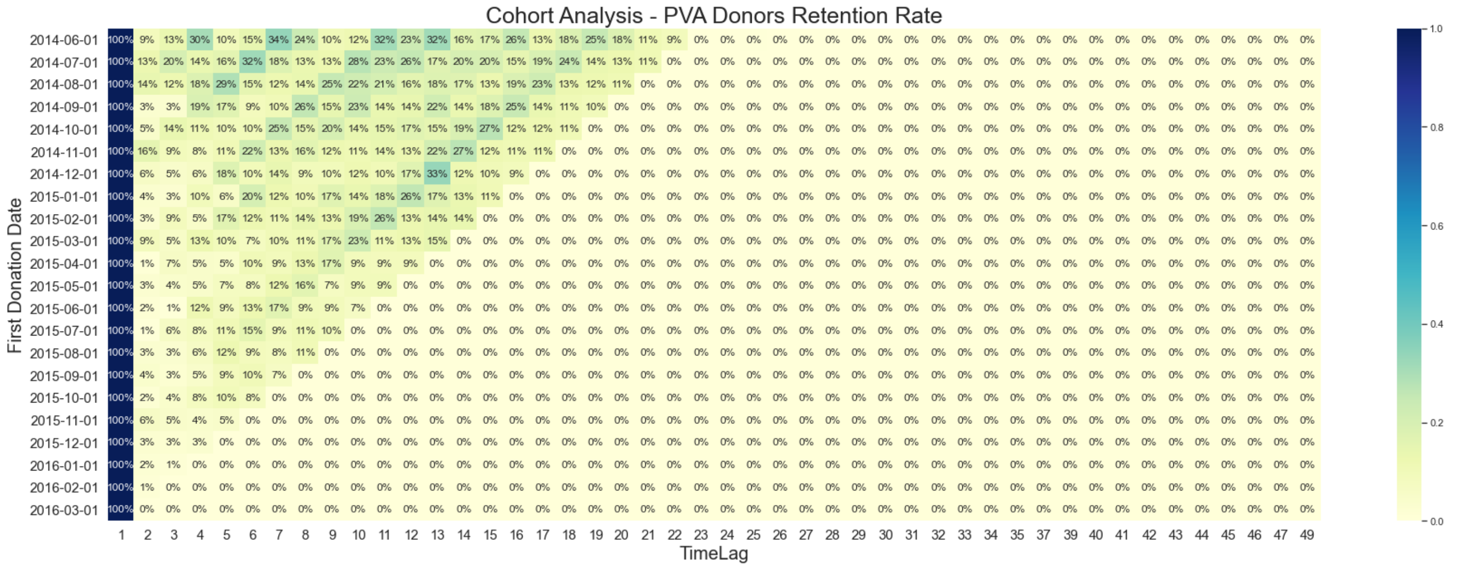


Figure 14 - Cohort Analysis - Absolute Retention



First Donation Date	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
2014-06-01	100%	9%	13%	30%	10%	15%	34%	24%	10%	12%	32%	23%	32%	16%	17%	26%	13%	18%	25%	18%	11%	9%	0%
2014-07-01	100%	13%	20%	14%	16%	32%	18%	13%	13%	28%	23%	26%	17%	20%	20%	15%	19%	24%	14%	13%	11%	0%	0%
2014-08-01	100%	14%	12%	18%	29%	15%	12%	14%	25%	22%	21%	16%	18%	17%	13%	19%	23%	13%	12%	11%	0%	0%	0%
2014-09-01	100%	3%	3%	19%	17%	9%	10%	26%	15%	23%	14%	14%	22%	14%	18%	25%	14%	11%	10%	0%	0%	0%	0%
2014-10-01	100%	5%	14%	11%	10%	10%	25%	15%	20%	14%	15%	17%	15%	19%	27%	12%	12%	11%	0%	0%	0%	0%	0%
2014-11-01	100%	16%	9%	8%	11%	22%	13%	16%	12%	11%	14%	13%	22%	27%	12%	11%	11%	0%	0%	0%	0%	0%	0%
2014-12-01	100%	6%	5%	6%	18%	10%	14%	9%	10%	12%	10%	17%	33%	12%	10%	9%	0%	0%	0%	0%	0%	0%	0%
2015-01-01	100%	4%	3%	10%	6%	20%	12%	10%	17%	14%	18%	26%	17%	13%	11%	0%	0%	0%	0%	0%	0%	0%	0%
2015-02-01	100%	3%	9%	5%	17%	12%	11%	14%	13%	19%	26%	13%	14%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-03-01	100%	9%	5%	13%	10%	7%	10%	11%	17%	23%	11%	13%	15%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-04-01	100%	1%	7%	5%	5%	10%	9%	13%	17%	9%	9%	9%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-05-01	100%	3%	4%	5%	7%	8%	12%	16%	7%	9%	9%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-06-01	100%	2%	1%	12%	9%	13%	17%	9%	9%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-07-01	100%	1%	6%	8%	11%	15%	9%	11%	10%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-08-01	100%	3%	3%	6%	12%	9%	8%	11%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-09-01	100%	4%	3%	5%	9%	10%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-10-01	100%	2%	4%	8%	10%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-11-01	100%	6%	5%	4%	5%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-12-01	100%	3%	3%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2016-01-01	100%	2%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2016-02-01	100%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2016-03-01	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

First Donation Date	24	25	26	27	28	29	30	31	32	33	34	35	37	39	40	41	42	43	44	45	46	47	49
2014-06-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2014-07-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2014-08-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2014-09-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2014-10-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2014-11-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2014-12-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-01-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-02-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-03-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-04-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-05-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-06-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-07-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-08-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-09-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-10-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-11-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2015-12-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2016-01-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2016-02-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2016-03-01	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Figure 16 - Cohort Analysis - Retention Rate