DATA MINING

Paralyzed Veterans of America

NOVA INFORMATION AND MANAGEMENT SCHOOL

DIOGO GONÇALVES | 20200632
HENRIQUE COSTA | 20200652
NUNO PAIS | 20200576

TEACHERS:

FERNANDO BAÇÃO
DAVID SILVA
JOÃO FONSECA

TOTAL PAGES w/ CONTENT [17]

Table of Contents

1. List of Figures

2. Introduction

The Paralyzed Veterans of America was founded on 1946 by veterans from the World War II, wounded soldiers that sought to contribute to society by helping other veterans- "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."[1]

PVA's expertise focuses on high quality healthcare, research on spinal cord injuries, benefits and civil rights/ opportunities for its veteran members.

According to PVA's 2020 Financial Report[2], available currently on their website, the biggest share of revenues throughout the year consisted of public contributions, which totaled approximately $72.9M. This represents a decrease of 3.19% in public donations, from 2019 to 2020. Regardless, 2020 saw a resurge in contributed services, investment income and other income, making it overall a financially better year than the previous.

The biggest share of the organization's expenses is funneled into 'Public Education and Awareness', marketing campaigns (such as the one further developed in this report) making use of public service announcements and other mediums to promote donations, highlighting the struggles lived by those with disabilities. In 2020, the two main advertising mediums used were TV campaigns (77.3% of total 'Public Education' expenses) and printed campaigns (the remaining 22.7%). These marketing expenses equal $68.379M. The other main section of expenses comprises benefits for veteran members.

Understanding the organization's financial and operational standings enhances the activities of this work, putting the data into a more holistic view of the business goals. A core part of the data science process, in any supervised or unsupervised (such as the current case) problem, is identifying the business objectives inherent to the activity.

The business objectives are to segment the donors database, utilizing a sample of 95.412 individuals who have previously donated to the organization. This segmentation helps the company understand the kind of donors they have, what defines them and how they behave. This is of special importance within the Marketing context of the firm, which, as has been referred, represents a major share of its cost structure.

By understanding which customers donate the most and the least, which have donated recently, and which are 'lapsed', the firm can identify sections of its customer matrix that can be specifically targeted through marketing campaigns, made exclusively to capture them. This is also done through the help of the multitude of socio-demographic data available about the donors.

Another important focus is the recapturing of the 'lapsed' donors- individuals who have donated the last time more than a year ago, as the more time passes the less likely will these individuals donate again.

One methodology taken to explore the customer segments was to conceptually divide the given dataset, according to its variables. The dataset contained globally 475 variables, and we divided them according to whether they were referring to donation activities (or donation-related information) or referring to socio-demographic information. The dataset containing donation-

[1] PVA's Mission Statement, taken from https://pva.org/about-us/mission-statement/ (seen on 18/12/2020)

[2] https://pva.org/about-us/financial-information-and-governance/ (seen on 18/12/2020)

related information was named Donors (130 variables) and the dataset containing the individual's information was named Info (344 variables)- before preprocessing.

For purposes of brevity, we will not exhaustively describe all variables in the dataset, but summarizing:

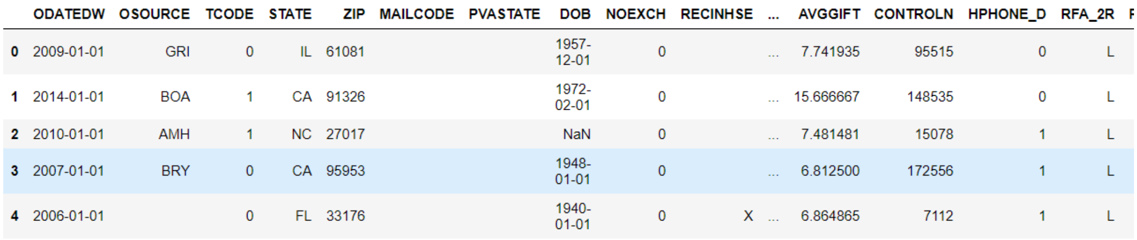
- Donors- Contains variables referring to the individual's donation amounts, dates when the donations took place (the first, the second and the last), dates when certain promotions were mailed and then when the donations associated to that promotion were made. The most important features are metric and concern donation amounts and how long ago those donations occurred.
- Info- Contains a large portion of information related with the individual's neighborhood, according to the 2010 US census. Includes variables denoting the donator's interests, hobbies and personal information (such as age, gender, income, living location, etc…). These Info variables have very relevant data to be used for marketing targeting, specially within the context of a clustering analysis.

The segmentation was conducted through a Cluster Analysis, a branch of unsupervised learning that seeks to group data points that are more related within themselves than they are with other points belonging to other groups of data (clusters). The specific algorithms used will be discussed later.

Ideally, the objective of the work is to find several well-defined and differentiated clusters that can explain the customer structure in a meaningful way. Then, according to those groups of data, analyze each one and prepare a marketing strategy specifically to target them, given their average characteristics.

3. Data Exploration

The first thing to do before assessing any model should be to understand and explore the dataset. A quick first analysis tells that the dataset is quite big, actually composed of 476 columns and 95412 rows, as previously stated on Introduction.



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

Figure 1 Data exploration - head()

The 476 columns contain important information to further analyse the details of donations. Besides, there are also features that will play a major role on clustering information, about the donors' personal information as, for example, characteristics about theirs neighbours, cultures and their connection to the US Army. At this, step, visualizing data would be very heavy and no major conclusions would be possible to make so, in that case, visualizations will only be made after dividing the dataset, on chapter 0.

4. Data Preparation

1. Data Adjustments and Feature Engineering

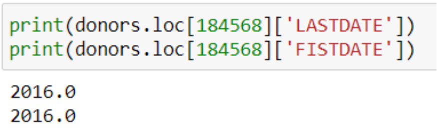
Being a real raw dataset, it’s quite usual that it comes with a lot of noise and scattered information among it. Therefore, there were made some adjustments and preparation on the data. First, the output of some features was rearranged, based on their metadata. For example, the feature MAJOR, which indicates whether a donor is a major donor or not, has a blank space for donors that are not major. As this would make the imputation process messy and confused, this preparation had to be made before any other procedures.

Another example that explains this necessity is the feature RECPGVG, which measures if the donor is a planned giving donor, filling it with an X if he is, and leaving the space blank if not. This way, when counting the missing values and further on the imputation process, these blank spaces that contain information would be treated as an error and, therefore, the value of the information obtained on the clustering process (chapter 5) would not reliable.

Besides this handy process, there were still 76 features with plus 40% of values being missing values, which represents more than 38 000 rows being errors. For these reasons, they were deleted, once not even the robust methods used for imputation, as KNN Imputer, would be able to accurately predict them. Later, the tolerance threshold for missing values percentage was adjusted to 70%, accounting for more information but allowing some biasedness in the imputation.

The dataset also contained information about dates, which is important to measure whether a donor is relatively young or if has been assumed a donor for a long time. In that way, a function called get_data was created, which returns the date in float type, becoming way more handy dealing with this type of information (now, instead of an actual date, dates were translated in how long it has passed since that specific date[3]). A variable called age was also created, easier to deal with and more intuitive than the date of birth itself.

One important detail is that NEXTDATE and TIMELAG have the same number of missing values. It was assumed that this fact meant that there were 9973 individuals that only donated 1 time and, for that reason, there was no record of a second gift. To prove this, it was checked whether the date of the last gift was the same as the first gift ever, which the fig. 2 suggests being true.



```
print(donors.loc[184568]['LASTDATE'])
print(donors.loc[184568]['FISTDATE'])
```

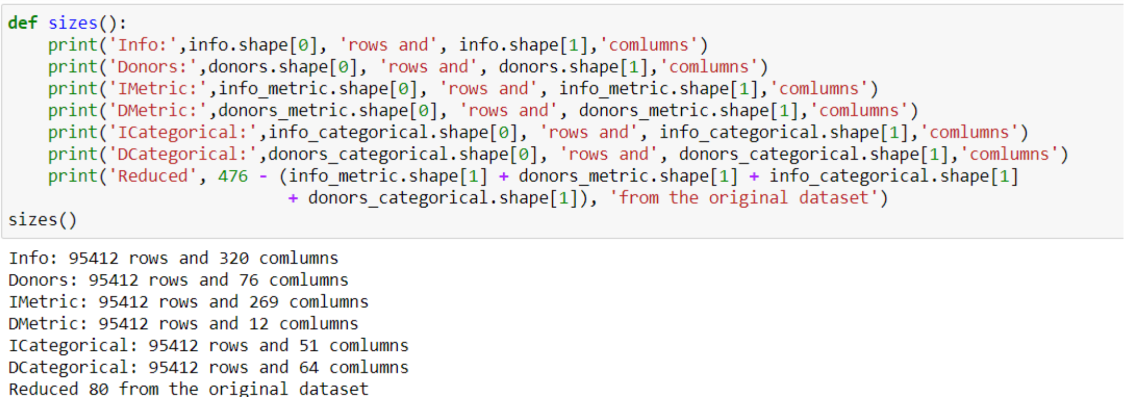
```
2016.0
2016.0
```

Figure 2 Confirmation before adjustments

[3] This process transforms date variables (categorical) into metric, like age. Example: 2016-05-24 now becomes 4-> the number of years passed since this date occurred.

2. Categorical and Metric Variables

As previously stated on Data Exploration, an important way to distinguish information about the donation process and the donors themselves is to split and divide the dataset in DONORS and INFO. The dataset was naturally also divided into metric and categorical features. At this moment, data is divided into 4 parts, as the figure 3 demonstrates.



```
def sizes():
    print('Info:',info.shape[0], 'rows and', info.shape[1],'comlumns')
    print('Donors:',donors.shape[0], 'rows and', donors.shape[1],'comlumns')
    print('IMetric:',info_metric.shape[0], 'rows and', info_metric.shape[1],'comlumns')
    print('DMetric:',donors_metric.shape[0], 'rows and', donors_metric.shape[1],'comlumns')
    print('ICategorical:',info_categorical.shape[0], 'rows and', info_categorical.shape[1],'comlumns')
    print('DCategorical:',donors_categorical.shape[0], 'rows and', donors_categorical.shape[1],'comlumns')
    print('Reduced', 476 - (info_metric.shape[1] + donors_metric.shape[1] + info_categorical.shape[1]
                            + donors_categorical.shape[1]), 'from the original dataset')
sizes()
```

```
Info: 95412 rows and 320 comlumns
Donors: 95412 rows and 76 comlumns
IMetric: 95412 rows and 269 comlumns
DMetric: 95412 rows and 12 comlumns
ICategorical: 95412 rows and 51 comlumns
DCategorical: 95412 rows and 64 comlumns
Reduced 80 from the original dataset
```

Figure 3 Size and shape of current dataset

3. Missing Values

The ideal and more scientifically efficient process to deal with outliers and missing values would be to do a simple imputation first, handle outliers and, after that, applying a more robust imputation. Nevertheless, this process was giving worst results on the clustering process, as further discussed on the Conclusion & Discussion.

That way, it was decided to use right away the KNN Imputer, a very common known robust method for imputation, and imputed the missing values on Donors Metric (only 2 variables had significant missing values) and Info Metric (12 variables with meaningful missing values out of 269). On the other hand, the imputation of categorical values was made through Random Sample Imputer, from scikit learn, which is also accepted as quite a good fit for non-metric variables. The Random Imputer works for both categorical and numerical variables replacing the missing values with a random sample obtained from the variables.

4. Outliers

On the Outlier process, a combination of Standard Deviation and Interquartile Range was used to check univariate outliers. To check multivariate outliers, Local Outlier Factor was used. These methods were combined and were chosen which to use based on the percentage of outliers

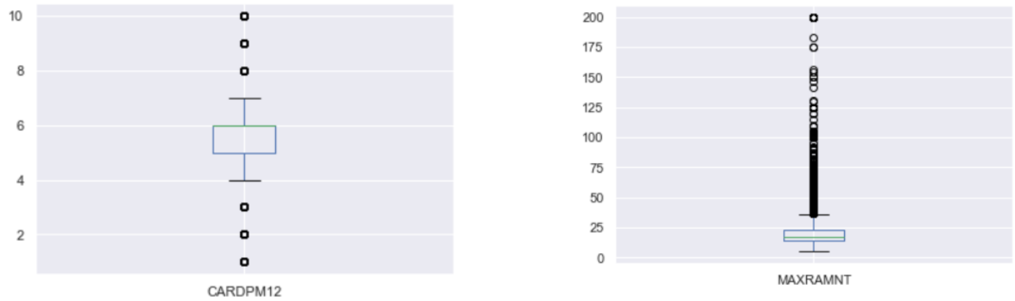


Figure 4- Visualizations used to remove outliers

removed. On Info metrics, these 3 methods did not produce very good results, so it was decided to complement it with manual removal through visualizations, using boxplots and histograms.

5. Redundancy and Irrelevance

Analysing the clusters based only on the initial raw dataset did not produce any good results, as expected. It is very important that the number of features is reduced in order to be able to handle features that are meaningful and neither too highly nor too lowly correlated with other features.

Once the data was already normalized during the previous process, using MinMaxScaler (which also contributed better for the results over the Standard Scaler), there was no preoccupation whether to use the Spearman or the Pearson correlation

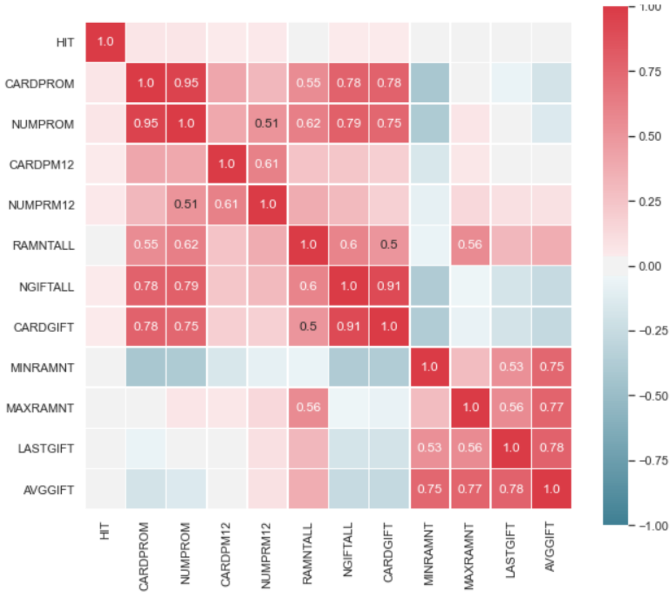


Figure 5 Correlation Matrix for Donors - metric

Although keeping in mind metadata and some critical thinking, the majority of features with correlations above 0.85 and below 0.05 were dropped. After that, this process was complemented with an analysis of the variance of each feature, looking for features that didn't vary its values over 5%, which means it won't differentiate the donors on the clustering process and, for that reason, were no longer important.

As neither Spearman and Pearson correlation measure directly correlation between categorical features, Cramér's V test was applied, which is based on a nominal variation of Pearson's Chi-Square Test and allows measuring the association between two categorical features. This alternative but robust method spotted 5 features on Donors categorical and 3 on Info categorical which would not add any efficiency to the clustering analysis. Though, it was decided not to remove the features that had very few percentages of correlation with any other, once there is no obligation that a categorical feature needs to be removed by not being sufficiently correlated with others.

Combining Pearson's correlation, Cramér's V test and variance analysis, the data set is now cleaner and more reduced, ready for the cluster analysis.

As is common in the data science process, certain preprocessing steps should be revised in light of the results achieved in later parts of the work. After conducting a clustering analysis for several sets of variables and by re-working specific preprocessing steps, the ‘Final Adjustments’ section was completed. This small section includes the main preprocessing steps, which allowed us, throughout the clustering process, to very quickly change crucial parts without having to erase, modify (and run all the code) the previous larger and more scientifically accurate data processing part. The main difference from the Final Adjustments is just that fewer variables were eliminated, giving the clustering process more information. Some features were conceptually very important for the business objectives of the process (the marketing analysis), although scientifically redundant. That being said, the group still thinks that the more complete preprocessing steps, despite not giving the highest results, are still the most relevant and scientifically accurate, having resulted in an interesting analysis. The Final Adjustments merely summarizes the essential core-steps of the process.

5. Clustering

One important step that preceded clustering was choosing the most important and relevant variables for the analysis itself. Producing clusters with every variable may utilize all information available, but a large percentage of them is irrelevant to the analysis and barely adds any variance or offers any cluster distinction.

This processing of cluster feature selection was made by trial and error, by running the clusters with different sets of variables and testing for the ones that provided relevant insights or differences. One important note here is that different clustering methods “select” different variables, but it was important, for the sake of consistency, to remain faithful to only one specific set of variables to be tested on all clustering algorithms.

Another important measure used to select relevant variables was the application of a function (drop_lowvar), which removed variables from the dataset that displayed very low variance throughout the dataset. The individual R^2 for all variables was also taken into consideration (keeping variables with high R^2 and removing ones with very low to null), after achieving the first cluster solution.

Conceptual analysis was also done- which variables are most relevant for the business goals in mind? The final set of variables was an intersection of both a redundancy/ relevancy analysis, as has been discussed.

The final variables are defined, in code, by the names **new_d_metric** and **new_i_metric**.

- **new_d_metric** | Comprises date variables (regarding how long ago certain donations took place or when certain promotions were mailed), the total donation amounts, average donation amounts and number of promotions received.
- **new_i_metric** | Comprises neighborhood information regarding population, geographical location, racial/ethnicity diversity, household income, home value, etc…

The strategy the group initially conceived was the following: Run K-Means to find, through an inertia plot and overall analysis, what the best initial number of clusters should be, and then using that value as the initial-seed for Hierarchical Clustering. However, going through this process the group realized this approach was more useful for the instances where K-Means reveals a large amount of clusters (which was not the case) as then the Hierarchical Clustering could merge those clusters into an optimal final (smaller) number. Given that this is not the case, it was decided that K-Means and HC should be run independently, seeing which one could output the best results (higher cluster differentiation).

Afterwards, Mean Shift Algorithm, Density Based Spatial Clustering, a Gaussian-Mixture Model and Self Organizing Maps (with K-Means) were used to also compare results. At the end, the final clusters for the Donors dataframe and Info dataframe should be visualized using t-SNE and then merged altogether.

1. KMEANS

Very briefly, K-Means uses given initial points as ‘means’ and forms several clusters in regard to the distance between surrounding points and those initial means. After defining the clusters, the centroids of those clusters become the new ‘means’ and the process repeats itself until a certain threshold (error tolerance) is met.

This algorithm proved to be the most computationally efficient and provided the best cluster distinction for the selected variables, both for the donors and for the info.

Note: As has been referred previously in the preprocessing portion, also mentioned in the notebook, the cluster profiling visualizations may, at first glance, seem to indicate that the clusters behave very similarly. This happens as a result of the usage of the MinMax scaler, which turns all metric data in the scale of [0,1]. Naturally, using StandardScaler would generate a more differentiated (visually) cluster profiling, as values would be situated in the range [-1,1]. Nonetheless, at the end, MinMax gives a much better final cluster, with much more diverse proportions (StandardScaler would incentivize most data to fall under just 1 cluster, for the Donors dataset) as can be seen in the following part ‘Cluster Interpretation’.

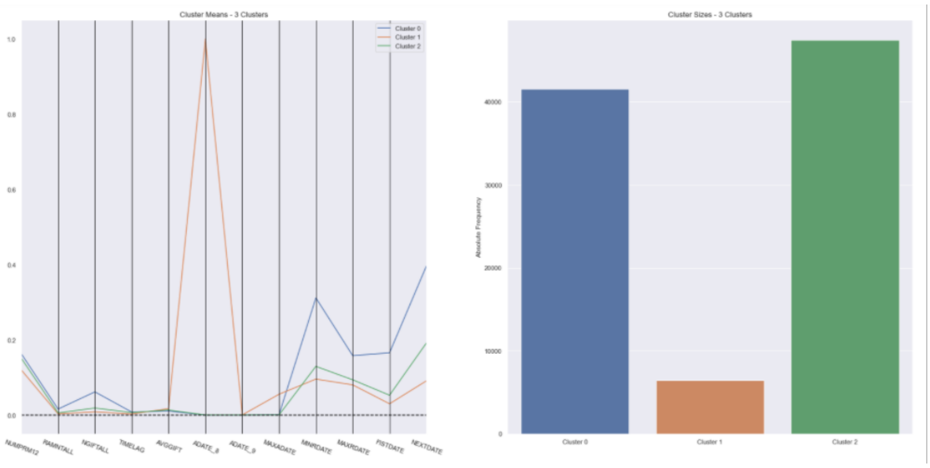


Figure 6 Donors Cluster Profiling using K-Means

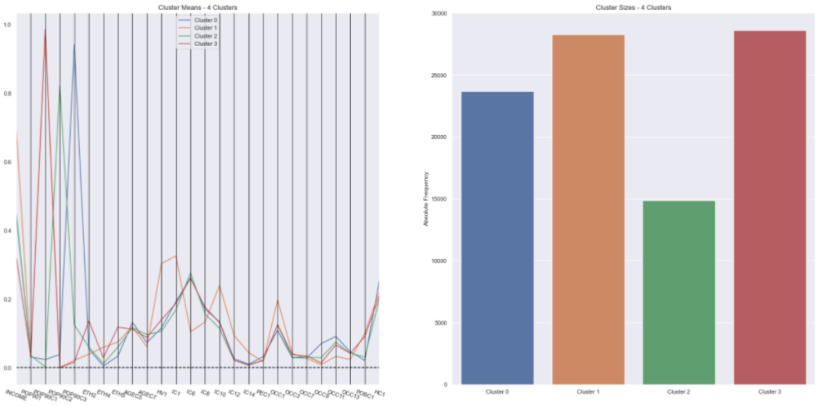


Figure 7 Info Cluster Profiling using K-Means

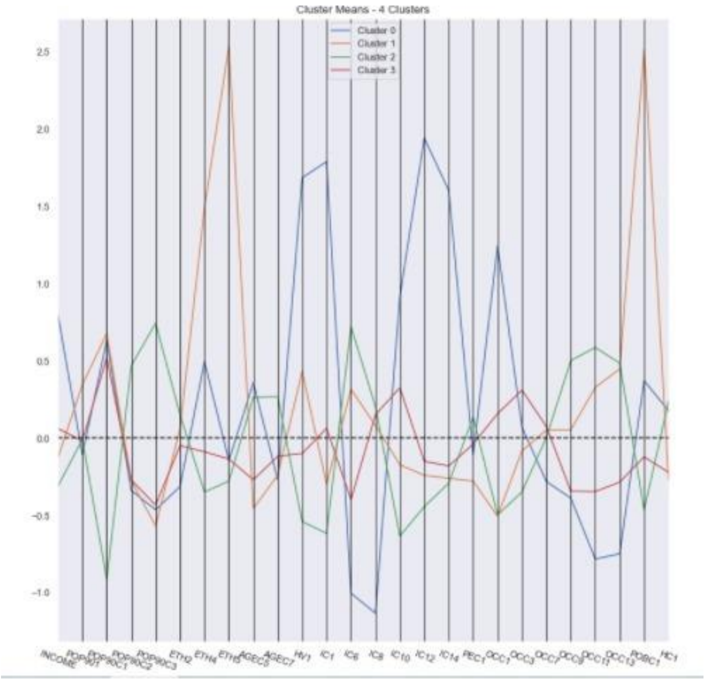


Figure 8- For purposes of comparison, Info Cluster using K-Means, under Standard Scaler.

2. Hierarchical Clustering

HC was implemented in a ‘bottom-up’ approach- Agglomerative Clustering, where each data point is its own cluster, and data points (mini clusters) are consequentially merged together. A threshold is defined by the biggest first “jump” in the dendrogram, signifying that beyond that line, the clusters to be merged are less alike (and are more distanced) than the clusters merged before it. Ideally, the clusters should have a high Sum of Squares Between and very low Sum of Squares Within. The distance between a random point and the centroid of the cluster that point belongs to should be minimal, while the distance between the centroid of a cluster and the centroid of the data should be high- more cluster distinction. This results in a higher R^2 coefficient, the data is more highly explained by the model in place.

Through the threshold line, the optimal number of clusters is defined, considering the Euclidean distance between each cluster.

This algorithm proved from the get-go to be extremely computationally inefficient for the given dataset. The clusters defined by the algorithm couldn’t be fully reliable, as running it on the whole dataset would require tremendous amounts of computer memory, largely surpassing the average available RAM of even modern computers. One way to get around this issue was defining clusters on just 20 thousand individuals and checking the results. There are two problems with this approach: 1- Only 20K donors are assigned with cluster labels; 2- The clusters are defined taking into consideration only 20% of the data, which may output biased results. To avoid these problems, the solution thought was to randomly sample 20K rows at a time, run the clusters and cluster profiling on the selected rows and repeat the process several times, to observe whether different data would arise different cluster results. If it yielded similar output, then the remaining steps would be to choose a set of clusters assign the remaining 75K donors to the preexistent clusters.

$$SST = \sum (y - \bar{y})^2$$

$$SSR = \sum (y' - \overline{y'})^2$$

$$SSE = \sum (y - y')^2$$

Figure 9 - Sum of Squares Total, Sum of Squares Within and Sum of Squares Between.

The results presented here are one of the sets of clusters for just 20K records.

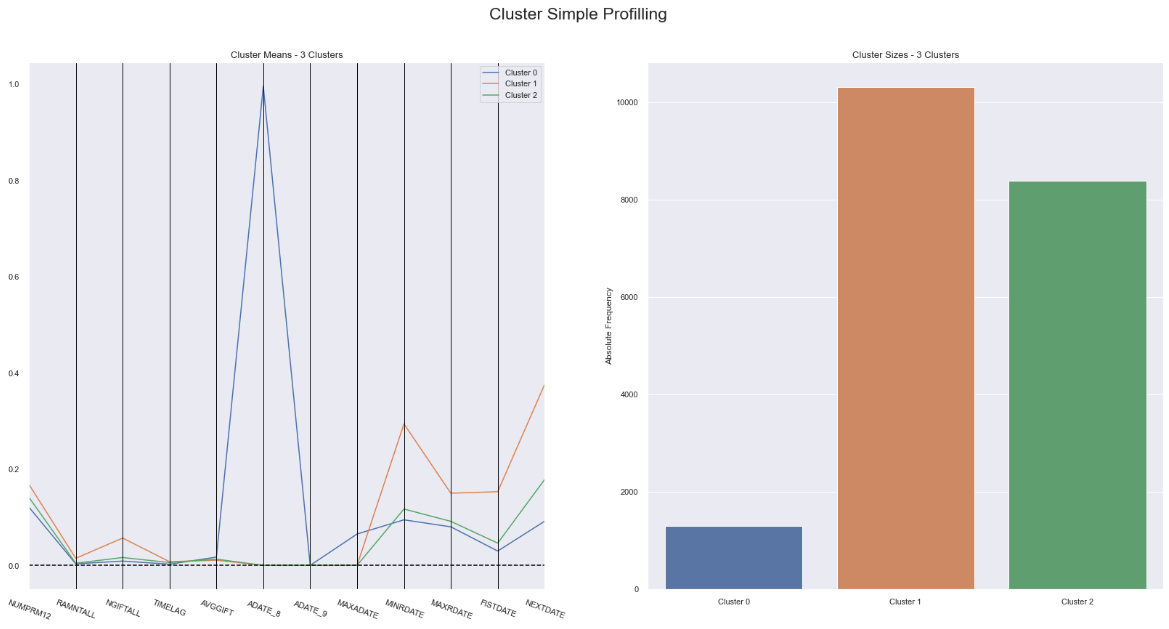


Figure 10 Donors Cluster Profiling using HC

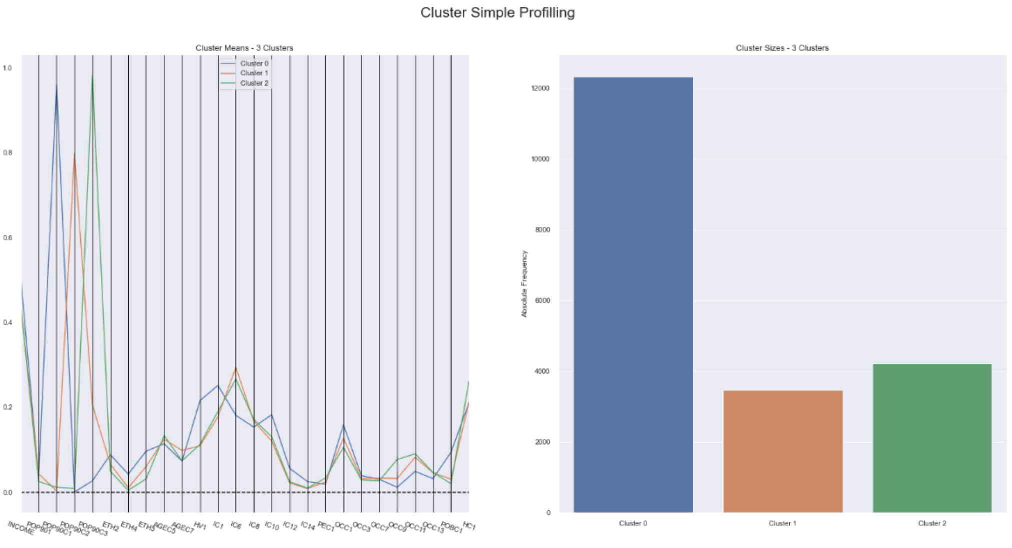


Figure 11 Info Cluster Profiling using HC

A preliminary cluster analysis, over the previous two methods, shows: For the Donors, variables NUMPRM12, RAMNTALL, NGIFTALL, TIMELAG,AVGGIFT have very similar values altogether, between the two methods. The only real distinction is in variables MINRDATE and MAXRDATE where the gap between clusters is wider, only slightly.

For Info, the realization that K-Means provides better results is even more evident, with better cluster distinction and having four clusters with similar absolute frequency of data points, unlike HC in which cluster 0 contains a large percentage of them. Naturally, the fact that these clusters used only 20K observations might induce some error in this analysis.

Overall, considering these computational inefficiencies and poorer results, the group decided to stick with K-Means over Hierarchical Clustering.

3. DBSCAN | GMM | SOM w/ K-MEANS

For purposes of brevity, the remaining clustering methods used will be summarized here. DBScan was useful for outlier detection, but the clusters defined were easily the worst. GMM was computationally efficient and provided relevant results, to be compared against the ones from K-Means. The Self Organizing Maps provided a very useful way to visualize the variables, through Component Planes, and, along with K-Means, provided results very similar to the original K-Means computed before.

DBScan measures density as the number of points surrounding a particular point, defined by a radius (eps) input. The other user given input is the number of points that define whether a point is considered to be a part of the cluster or not. This algorithm was also very inefficient, requiring tremendous amounts of space and CPU memory to be run on the whole dataset. Additionally, defining the eps and min_samples (number of points) according to the "elbow-method" and considering the number of dimensions in the set also gave very bad cluster results. A trial and error process was then required to set hyperparameters that gave reasonable results, but they were still very poor in comparison with other models. Setting the eps at the seemingly correct elbow value would return 3 clusters, one of which was noise points, cluster 0 being a very large cluster (made up of almost the entirety of points in the set) and cluster 1 with a very low absolute frequency. The same occurred when the eps was decreased, to guarantee that all points weren't fit to the same cluster. Running the algorithm with a smaller eps would increase the cluster number to five, but again, one cluster would contain most data points and the other ones a very low percentage. The number of min_samples was set to 25 (twice the number of dimensions, the heuristic rule), then 70 (to account for the largeness of the dataset) and even 120, and the results were still very similar. For all of these reasons, the DBScan results were disregarded by the group in this analysis.

The Gaussian Mixture Model is a soft clustering method that introduces the concept of the probability that a specific point belongs to a certain cluster, being that cluster associated to a gaussian distribution. This makes use of the maximum likelihood estimation- reassuring that the gaussian distribution is able to explain the data that is within the clusters. This algorithm was very efficiently applied and provided relevant results. The number of components was selected through the plotting of AIC and BIC, choosing the nº of components at the elbow. Donors was applied a GMM with 4 components and Info the same.

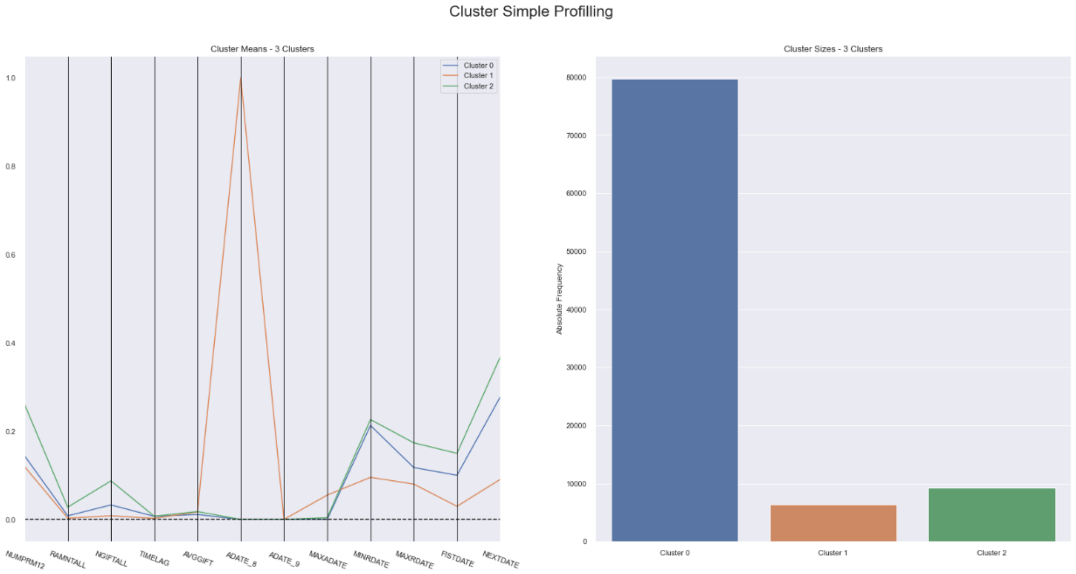


Figure 12- Donors Cluster Profiling using GMM

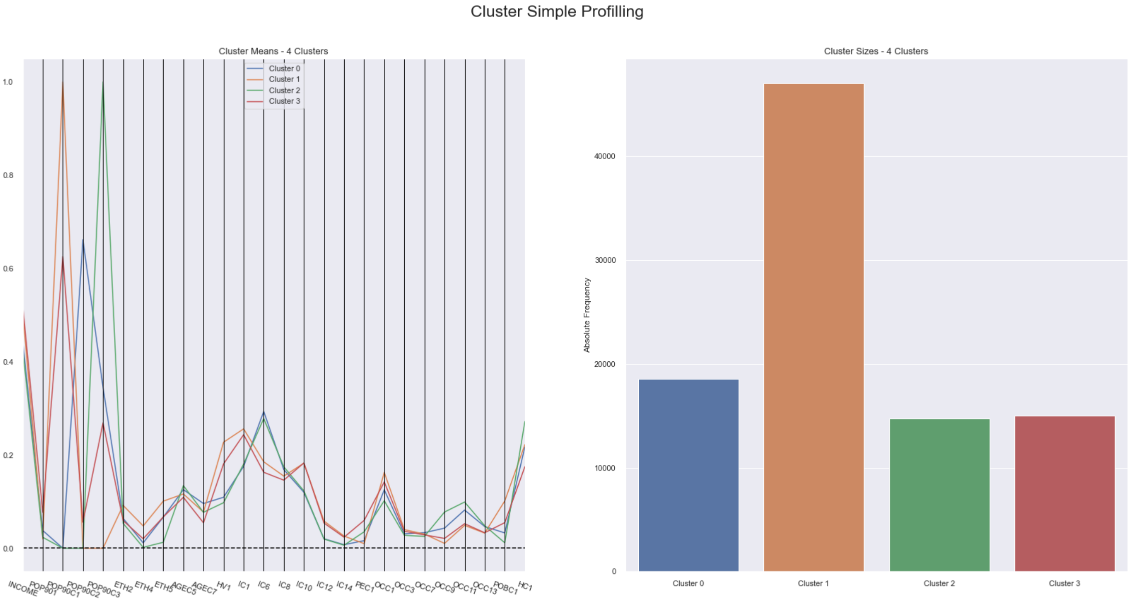


Figure 13- Info Cluster Profiling using GMM.

The Self Organizing Maps allowed for an interesting visualization of the component planes for each metric variable used in the clustering process and also provided a good output over the K-Means, even though it was computationally expensive, so the standard K-Means made more sensed to be utilized.

The final clusters for Donors and Info were merged through a process of hierarchical clustering, joining the total seven clusters into five. This can be seen through the dendrogram, where the biggest leap in Euclidean distance (threshold of 1.4) creates five distinct clusters.

4. Cluster Interpretation

T-SNE was used to visualize the final clusters, outputting the following:

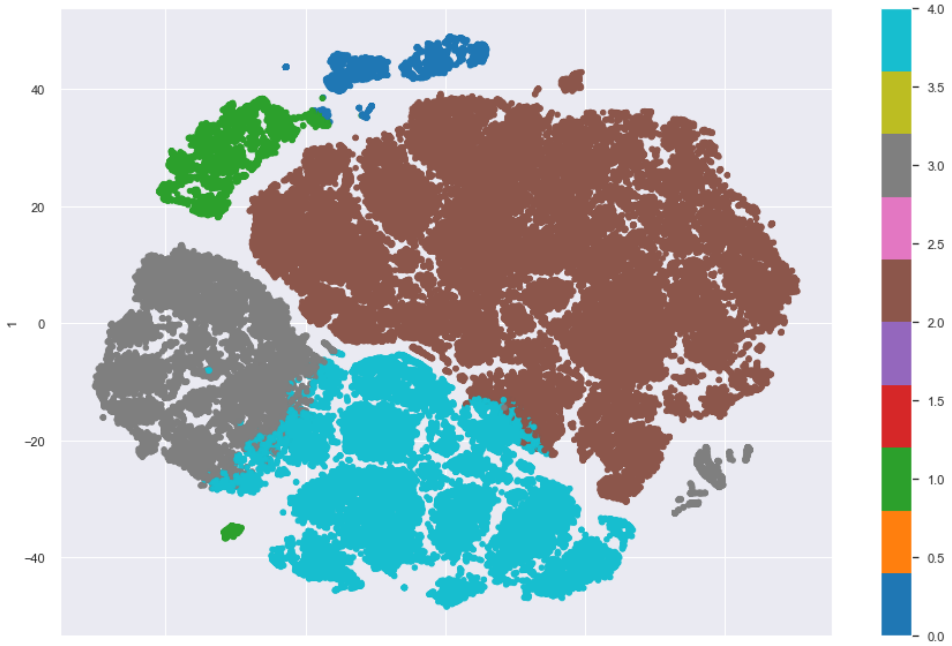


Figure 14 Visualization of the final clusters, using t-SNE

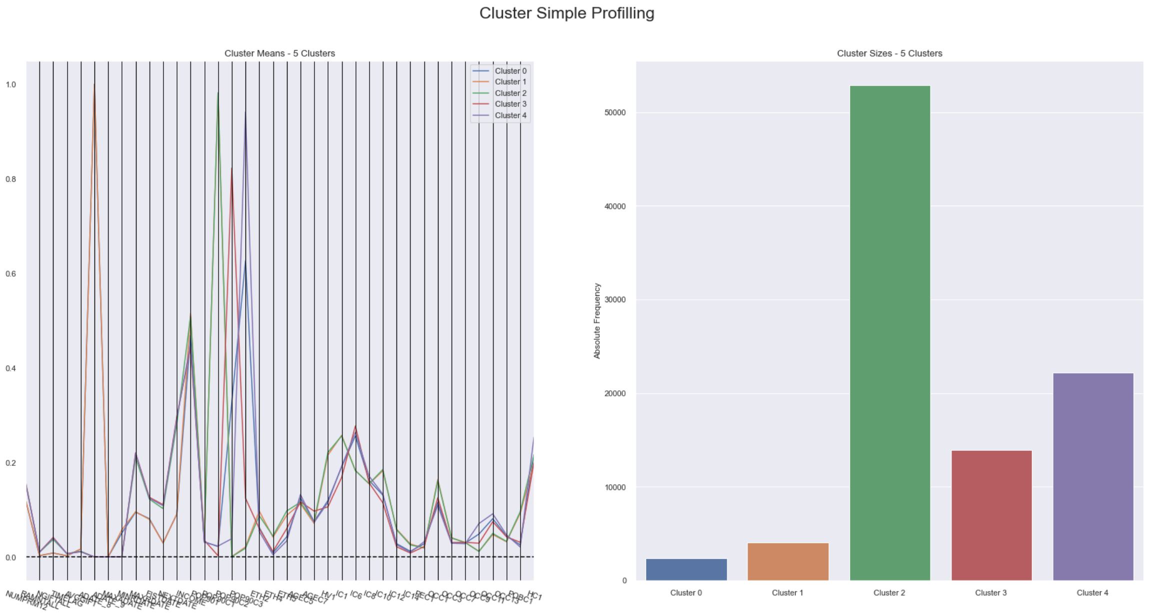


Figure 15- Final Clusters Profiling

Cluster 1 | ADATE_8 defines how long ago the promotion 96GK was mailed. There are currently 89K donors who were mailed the promotion 4 years ago, which, after normalizing is defined as 0, and 6411 donors who were mailed the promotion 5 years ago. This is a big distinction for cluster 1, but this characteristic itself seems partially conceptually irrelevant. The interesting part is that cluster 1 also seems to have lower average MINRDATE, MAXRDATE and FISTDATE. This implies that this cluster of donors has donated their lowest value gift, highest value gift and overall first donation more recently than the others, despite having being mailed the promotion 96GK longer ago. In TIMELAG, we can see that Cluster 1 has lower values, meaning that the period between the first and second donation was smaller than average.

Cluster 2 | Is by far the largest, with more than 50 thousand donors. It is distinguished by a high POP90C1, meaning urban residents, given that these donors live in neighborhoods with a high percentage of the citizens in highly urbanized environments- this allows for a very relevant analysis.

Cluster 3 | Is defined by a high POP90C2, which means that these individuals live in neighborhoods outside urban regions. These can be seen as areas with less population density, away from the main larger cities and suburbs. One important variable here is AGEC7, where this cluster shows superior values to all other clusters, meaning higher number of senior citizens.

Cluster 4 | Is defined by a high POP90C3, meaning donors that live in neighborhoods in rural areas. These outweigh the number of individuals living outside urban areas (cluster 3) but are still definitively smaller than the number of individuals in urban environments (cluster 2). It is also important to note that **Cluster 0**, the smallest, has a high percentage of individuals living in rural areas, so both cluster 4 and 0 can be geographically located in such regions.

This preliminary analysis concludes that geographical location is a big cluster determinant, not only obviously due to the geographical differences (someone living in a place cannot simultaneously live in another), but the mere fact that different clusters refer to different locations might suggest that these differences also lead to divergent behavioral patterns, which is what will be examined now.

Clusters 1 & 2 vs. Clusters 0 & 3 & 4 | One pattern that can be seen is that, in most variables regarding individual information, cluster 2 and 1 go usually together, while the remaining 0, 3 and 4 tend to behave similarly. The next significant divergence in behavior is in ETH4, where Cluster 1 and 2 have individuals living in neighborhoods with higher percentage of Pacific Islander/Asian individuals, possibly because these clusters are geographically located inside the main city landscape, with higher population density numbers, so it is statistically more probable for these clusters to have higher percentages of the widest racial and ethnic diversity. This is also proved by the variable POBC1, which gives the percentages of citizens born in foreign countries, in the neighborhood, also high for clusters 1|2. Another even larger difference in this distinction between Clusters 1|2 and Clusters 0|3|4 is in HV1 and IC1. These are highly important variables to consider in this marketing and data analysis, as both refer to the individual's income (indirectly). HV1 measures the median home value of the donor's neighborhood, which is a huge signifier of financial power and surrounding culture. Individuals in Clusters 2|1 live in higher-value neighborhoods, which is consistent with the geographical location where they reside, as we've already discussed- inner-city urbanized environments. IC1 measures the median household income of the donor's neighborhood. Here, the discrepancy between the two cluster groups is smaller, but still relevant. Clusters 1 and 2 are made up of higher income individuals, an inference that can be made due to the fact that they live in higher home value and higher income neighborhoods. Confirming the previous results, in IC6, it is evident to see that the second group of clusters has a higher percentage of households with less than $15K dollars a year of income. Along with the household income increase, so do the percentages of higher income families in the neighborhoods of clusters 1|2. These clusters also seem to have a higher percentage of 'Professional' citizens, a broad term to denote any that work in highly specialized professions. Very interestingly and expected, cluster 4, characterized by living in rural areas, have a higher percentage of farmers than all other clusters, even cluster 0.

Note: In certain variables, it is possible that we see no difference between clusters, if those variables required (during preprocessing) a lot of missing value imputation. Also, it is logical that the kind of imputation method employed might originate different results.

Cluster Interpretation of Categorical Variables

Gender: All clusters seem to contain more women than men, disregarding the individuals who did not disclose their gender.

Wealth 2: Cluster 2 contains 28101 donors who belong at the highest income rating class. This is larger than all other individuals in the lower wealth-rating classes combined, for this cluster. WEALTH2 measures income relative to State (location), which confirms that Cluster 2 certainly has the largest high-income group. Nonetheless, Cluster 2 seems to also have the highest income inequality, resulting in a quite financially heterogeneous set of donors. Clusters such as 0 and 1 have a very homogeneous economic structure, with the majority of donors also belonging to rating 9 (the highest one). Cluster 4 is the most wealth-diverse group of donors. The results from Wealth 1 also confirm these results.

Interests: Working from home, boats, buying children’s products, card buying and plate collection were all determined to be irrelevant for cluster distinction, as virtually no individuals engaged with either one of those activities.

Cluster 4 has the highest share of fans of collectables, at 6.11%; interest in veterans- 12%; engaged in bible reading- 11.43%; shopping by catalog- 9%; owner of household pets- 18.22%; Crafts- 10.41%; Fishing- 10.54%; Gardening- 16.84% (this activity seems very relevant for all clusters, donors are seemingly interested in this hobby).

Cluster 2 has the highest share of owners of CD players and, naturally, CDs - 13.58% and 13.94%. They are PC owners (11.93%), photography enthusiasts (5.2%, despite not representing a significant niche at any cluster) and walk for health purposes (11.39%) more than any other cluster.

Regarding the number of times the donors have responded to mail order offers of different kinds, such as related to Gardening or Books, on average no cluster showed any relevant difference. All clusters seem to indicate that, on average, every individual does not bother to reply to mail offers related to their hobbies, so the average number of times is 0.

Summary

There seems to be a clear geographical distinction between the clusters, which has implications regarding the purchasing power of each group, their respective profession and educational levels.

Cluster 1 | Recent donors, have made most donations more recently than all others, more recently acquired by PVA. Live in urban environments, in higher income neighborhoods and possibly middle/upper-middle class. Contains a financially very homogenous group of citizens, with high wealth ratings.

Cluster 2 | The biggest share of donors belong here, residing in urban areas with higher salaries and higher median home values- having a really large share of individuals belonging at the highest wealth ratings. These live in areas that consist of mainly 'Professionals', likely with superior education. Neighborhoods of these clusters tend to contain higher racial diversity and foreign born citizens. Also, despite the significant proportion with high income, financial diversity is still present. Their hobbies consist of more traditional urban interests (namely in the cultural field), such as music and photography. Additionally, they exercise for health benefits and are more in-touch with technology.

Cluster 3 | Donors of this cluster generally live outside urban areas. It is distinguished by a higher percentage of citizens above 75 years old. Tends to display very similar behavior to Cluster 4.

Cluster 4 | These individuals live most predominantly in rural areas, outside large cities. These groups tend to have lower income and lower value homes. A considerably higher than average share of people in this cluster are farmers. Simultaneously, their sizes of residence are also higher. A small MAXADATE implies they received their last promotion more recently than other clusters, and a high MINRDATE means the date linked to their smallest donation was made longer ago than other clusters. In comparison with other clusters, this group of citizens is significantly more interested in war veterans, Bible reading, fishing, gardening and owner of animals (extremely consistent data considering their geographical location).

Cluster 0 | These also live in rural areas, though in less percentage. Have a higher percentage of farmers, just like Cluster 4, but in smaller size. They are also the smallest cluster in the data. Unlike Cluster 4 though, these have a higher MAXADATE, having received their last promotion mailing longer ago.

5. Lapsed Donors

All donors in the dataset are defined to be Lapsed, as they have been characterized as Lapsed according to the last promotion they received. This is explicit in RFA_2R (x10_L in encoding).

That being said, there are certain data inconsistencies, given the relation between the 'Lapsed' status and the date of the last donation/ gift. A Lapsed donor is an individual that has made their last donation more than one year ago and less than two years ago, meaning that LASTDATE should be $1< x < 2$, given that, as has been seen, all 'date' variables were translated to numeric form, by calculating how long ago they occurred (in years). If the dataset was originally from Dec 2017, the data would still be inconsistent, given that every donor should have donated their last gift during from Jan-2016 to Dec 2016. If we consider the base year as 2020, the more inconsistent it gets. Nonetheless, we figured this is a relevant variable (LASTDATE) for cluster differentiation, allowing us to have a general relative idea of which clusters donated most recently, which may provide valuable insights. These results are presented in the ending portion of the code notebook.

6. Marketing Analysis

Marketing is a process by which companies create value for customers and build strong customer relationships in order to capture value from customers in return.

To create a marketing strategy relevant to their clients, companies need to define their own marketing mix. The marketing mix is the set of tools (4 Ps) the firm uses to implement its marketing strategy. This set includes Product, Price, Promotion, and Place. Once PVA is an organization where the biggest share of revenues throughout the year consists of public contributions, as seen on Introduction, the only tool that can be used is promotion and placement.

Considering the characteristics of the customers in the final clusters, the group recommends some marketing approach to better serve each cluster.

Cluster 1 & 2 | These 2 clusters contain a group of citizens with high wealth ratings. They represent younger donors and mostly with superior education. For this kind of citizens, the best marketing option is through **digital marketing-** advertising on social media and E-mail marketing, which involves sending highly targeted, highly personalized, relationship-building marketing messages via email. This strategy is also justified by these cluster's relations to technology, being owners of personal computers and engagements with urban culture. Here, artistic activities are more highly valued, so marketing campaigns that include cultural events might be more well-regarded. This can be leveraged by the organization by displaying ads on the mediums that these clusters most typically consume.

Since these individuals are the biggest share of donors, a reward might be given depending on the personal goals and values of their typical donor.

For some that seek recognition, a list can be included on the website or in a special publication as well as a personalized 'thank you' note. For others, that might be motivated to give because they see their contribution as giving them access to a community that they would like to be a part of and share with others, they may be given the opportunity to attend a yearly gala **dinner** with other donors and veterans. The last reward to another kind of donors might be nothing at all, since some donors do not like the idea of donating for a reward, making the act of donating feel less genuine and generous. Furthermore, many donors will feel their donation is less impactful because it also pays for a gift.

The placement of these advertisement campaigns, whenever physical (such as through posters or billboards) should be in cities, highly urbanized environments. A strong suggestion is also made for the advertisements to be located in specific services or stores related to the clusters' cultural interests (music stores, photography and other arts shops) or to their professions (given that the clusters are made up of 'Professionals', these can include private clinics, lawyers offices, financial services, etc…).

Cluster 0 & 3 & 4 | Once these individuals live most predominantly in rural areas, PVA can conclude that they will be less technologically evolved. Actually, cluster 3 has the higher percentage of citizens above 75 years old. A mature audience is more likely to be influenced by strong, straightforward messaging. Keeping text clear and using relevant and plain language will do the job. These clusters must be reached through an old fashion way. Since this kind of audience spends a lot of time watching **television**, it might be a good idea to advertise in local tv channels and setting some posters in places where the community gathers the most. Beyond that, since this group of citizens, especially cluster 4, is significantly more interested in war veterans, a **lunch** can be organized with the donors and veterans, to raise culture and dynamic friendships.

A great part of these individuals shares a common interest in reading and fishing. So, perhaps some advertising in fishing and book **magazines** might appeal to donate to PVA. It is also recommended to engage with local churches or Bible study groups, as means of achieving more awareness for the issues of veteran healthcare, given that these clusters tend to be more involved with religious practices. The placement for physical advertising should be in rural areas and on the outskirts of urban environments. Further analysis can be done on the kind of political views these clusters tend to have, but one strong assumption (based on statistical data from the 2020 Presidential Elections) is that population living in less urbanized environments, being also more strongly religious, tends to agree with predominantly more republican ideologies, which can be included in the messaging nature of the ad, evoking a more proponent sense of American nationalist spirit in the help of the troops.

7. Conclusion & Discussion

Although the ending clusters were crisply defined with highly consistent data, in the preprocessing the final missing values imputation did not follow a scientifically ideal process, due to the resulting lack of quality on the final results (that would arise as a result of following more complex approaches), and the applied methods were also not as good as the group expected them to be. This may have occurred due to the redundancy analysis, which probably reduced too many features, some that could play an important role during the clustering analysis.

One other important detail that can be misguiding data too is the outlier methods used, once manual removal could have removed too many rows, making the data more homogenous and therefore difficult for algorithms to cluster the groups.

The final methodology used, despite more basic, resulted in highly consistent and relevant clustering groups, tied to a simpler algorithm (K-Means). As was analysed previously, the clusters were very crisply defined visually and also conceptually- the clusters' geographical locations, hobbies, income and house values were all very consistent with other data, displaying no irregularities or contradictory information. Having these results provided the group with the potential of designing capable marketing strategies, specifically catered to each clusters' own conditions.

8. References

2.1. Gaussian mixture models, scikit-learn [online] Available at < https://scikit-learn.org/stable/modules/mixture.html> [Accessed 22 December 2020]

David Silva & Susana Paco. Data Mining's Final Project [online] Available at <https://github.com/DavidSilva98/DataMiningFinalProject > [Accessed 20 December 2020]

Feature-engine.readthedocs.io. 2020. *Randomsampleimputer — Feature-Engine 0.6.1 Documentation*. [online] Available at: <https://feature-engine.readthedocs.io/en/latest/imputers/RandomSampleImputer.html> [Accessed 26 December 2020].

Gaussian Mixture Modelling (GMM), towards data science [online] Available at < https://towardsdatascience.com/gaussian-mixture-modelling-gmm-833c88587c7f> [Accessed 23 December 2020]

KNN Imputer, scikit-learn [online] Available at: < https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html> [Accessed 26 December 2020]

Local Outlier Factor, scikit-learn [online] Available at: < https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html > [Accessed 26 December 2020]

Machine Learning Project Customer Segments, Ritchieng [online] Available at: <https://www.ritchieng.com/machine-learning-project-customer-segments/> [Acessed 30 December 2020]

ML| Types of Linkages in Clustering, GeeksForGeeks [online] Available at < https://www.geeksforgeeks.org/ml-types-of-linkages-in-clustering/> [Accessed 2 January 2021]

Paralyzed Veterans of America, 2021. Financial Information and Governance [online] Available at: < https://pva.org/about-us/financial-information-and-governance/> [Accessed 15 December 2020]

Paralyzed Veterans of America, 2021. Our Missions & History [online] Available at: < https://pva.org/about-us/mission-statement/> [Accessed 20 December 2020]

Self Organizing Maps, Abhinav Ralhan [online] Available at < https://medium.com/@abhinavr8/self-organizing-maps-ff5853a118d4> [Accessed 23 December 2020]

Self-organizing map, Wikipedia [online] Available at < https://en.wikipedia.org/wiki/Self-organizing_map> [Accessed on 22 December 2020]

sklearn.cluster.DBSCAN, scikit-learn [online] Available at < https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html> [Accessed 23 December 2020]

Univariate and Multivariate Outliers. Statistics Solutions [online] Available at: <https://www.statisticssolutions.com/univariate-and-multivariate-outliers/ > [Accessed 22 December 2020