DATA MINING
MSC IN BUSINESS ANALYTICIS AND DATA SCIENCE

LAPSING DONOR CLUSTERING ANALYSIS FOR MARKETING STRATEGIES

Ikram Bouziri M20200753
Luisa Crumley Barral M20201045
María Luisa Noguera Lecompte M20201005

NOVA IMS
DECEMBER 2020

I. INTRODUCTION

“PVA is a non-profit organization that provides programs and services for US veterans with spinal cord injuries or disease”

Paralyzed Veterans of America, relies on its sweepstakes and promotional mailing for donor engagement. A large portion of their database has fallen into the category of “Lapsing”, which means that for more than a year and up to 24 months **by the end of 2017**, the donor has not sent any donations to either one of PVA’s programs: (In House, P3, and Planned Giving) or through the purchase of merchandizing (sweepstakes).
Given the importance of the support they provide to the Veterans, PVA decided to consult data scientists to come up with a strategy for grouping and profiling such customers in order to approach them again with a different marketing strategy,
A great deal of the data provided to said data scientists was a combination of their own information on donor history and other characteristics stored in third party databases. Such information also associated outdated data (especially the one resulting from census) to most of their customers and further coded data that increased the complexity for analyzing their information.
Although the specialists were provided with a large dataset with over 450 features per customer, most of the information could not be used for the analysis because of its formatting, the lack of definition of the variable, and/or amount of missing information per variable.

II. METHODOLOGY

1. Understanding the metadata

The file provided by PVA in plain text explained very little about the features assigned to each donor. Some of their titles were self-explanatory or had a description long enough to understand what they contained. However, many of them did not and made it difficult to interpret.
Furthermore, the content on several variables had different format, or did not follow a standardized form of information. To help us with the understanding of the data and our analysis we used the describe function of the DTale library, which is similar to Pandas Profiling. Then the group was able to separate all variables into type of information and whether they had a good description but poor quality of content, a good description and good content (minimal or no missing values and simple formats), those with good description but irrelevant information, and those who had a poor description that automatically ruled them out. When reviewing the titles of each donor for example (TCODE), the codes contained in this feature had several codes that repeated the same information but in different languages or military and political denominations.
Other variables had straightforward descriptions but made it clear they were not relevant to the scope of the project, such as DATASRCE which shower which third party database they were acquiring the information from.
An example of a feature that neither its title nor its description was helpful, was PEPSTRFL since it contained blanks or X and its description only mentioned “Indicates PEP Star RFA Status”. The relationship between one concept and the other could not be determined.
Other type of inconsistencies in the data was the timeline for some pieces of information. Census information associated with the donors was outdated since it belonged to the 2010 census, and the last year of the database (the project’s “today) was the end of 2017.
After careful review of the entire metadata file, the features the group found to be relevant and contained the most amount of information were kept for the subsequent analysis.

2. Manual Feature selection

Reviewing the metadata helped the team select an initial set of 23 variables containing both numerical and categorical variables and that referred to their financial information, customer behavior, how they engaged with PVA and further demographics.
The only column of the original data that was included, not as a feature, but as the index is CONTROLN, since is a unique identifier of each donor.

Initial set: *ODATEDW, STATE, MAILCODE, DOB, RECINHSE, RECP3, RECPGVG, RECSWEEP, DOMAIN, GENDER, HIT, MAXADATE, NUMPROM, NUMPRM12, RAMNTALL, NGIFTALL, MINRAMNT, MAXRAMNT, LASTGIFT, LASTDATE, TIMELAG, AVGGIFT, RFA_2F*
Final set: *STATE, MAILCODE, DOB, RECINHSE, RECP3, RECPGVG, RECSWEEP, DOMAIN, GENDER, HIT, MAXADATE, NUMPROM, NUMPRM12, RAMNTALL, NGIFTALL, MINRAMNT, MAXRAMNT, LASTGIFT, LASTDATE, FISTDATE, TIMELAG, AVGGIFT, RFA_2F*

Nonetheless and given that the preprocessing is an iterative process, as the project progressed the set of variables selected was also adjusted to findings and realizing some other information initially left out was more appropriate or relevant. Thus, the final features selected for analysis and clustering were:
Other variables considered throughout the process were those that contained information on the donor's preferences provided by a third party. However, they were not used for the clustering analysis.

3. Data preprocessing

After selecting the variables manually, we proceeded with the cleaning and preprocessing of the data, depending on what each of them needed.

A. Dropping rows (reducing the noise)

- For the variable MAILCODE, the rows having "MAILCODE"='B' were dropped since this value means that the address is not correct. (1.47% of the data was dropped). Since PVA relies on currier, having incorrect addresses means that their promotions and/or any communications are not reaching the donors.
- In the variable STATE, the rows having values that did not correspond to any state in the United States of America in real life were dropped. (0.13% of the data was dropped)

B. Feature engineering (transforming variables and treating missing values)

- **DOB:**

The date of birth was converted to age using the year 2017 as "today" (year when the database was created). It was first split into year, month and day to then convert the variable to numeric and then calculate the years. Even though, after transforming this variable, 23295 missing values were found, it was kept because age is an important demographic feature for customer (donors) cluster analysis.
The missing values of this variables were filled using KNN imputer with 5 neighbors.

- **RECINHSE, RECP3 and RECPGVG:**

These three variables give information about In House, P3 and PGPV programs. First, empty spaces were eliminated to concatenate all of the three variables together deriving a new variable called Program_Donor. The values of the newly created variables are:

- " " = the person has not donated to any of the 3 programs
- X= the person donated to 1 program
- XX= the person donated to 2 programs
- XXX = the person donated to 3 programs

- **DOMAIN:**

Since this variable was composed of two bytes (the first byte describing Urbanicity level of the donor's neighborhood and the 2nd byte defining the socio-economic status of the neighborhood), the feature was split this into two new variables and DOMAIN was dropped:

  - Urbanicity level: 2202 missing values were found in this variable and they were all filled with the mode.
  - Socio-economic status: 2202 missing values were found in this variable and they were all filled with the mode. This variable was converted to numeric.

- **GENDER:**

For this variable, the values C and A were dropped since they have no meaning in the metadata. Blanks were filled with U merged with the category “J” which had the same meaning.
The final values in this variable are:

  - F = Female
  - M = Male
  - U = Unknown

- **HIT:**

The original HIT data was kept since it gives an information of how many mail orders the donor has responded to. Using HIT, another variable called Hit binary was derived to simply identify if the donor had responded to mailing or not. Hit binary became a yes/no feature. This new variable was used to identify the profiles after clustering.

- **FISTDATE:**

This variable was used to calculate the number of years since the first donation. The newly created variable was called "Tenure" or how long has the donor been engaging with the organization. FISTDATE was dropped after the transformation.

- **MAXADATE:**

After converting it to date type, MAXADATE was transformed to the number of months since the last time the donor was sent a promotion into a new variable called “Months_last_promotion”. MAXADATE was dropped after the transformation.

- **LASTDATE**

After converting it to date type, LASTDATE was transformed to the number of months since the last time a donor sent a gift to PVA, under a new variable called “Months_last_gift”.
LASTDATE was also dropped.

- **TIMELAG**

9866 missing values were found in this variable. Since it is the number of months between first and second gift, all the missing values were replaced with 0 because the donor only donated once (one time donors)

- **Preferences**

For the preferences variables, first, the blanks were replaced with nans. Then, all missing values were replaced with 0. Lastly, all the “Y” values were replaces with 1. In this case 0 represents not interested while 1 represents interested.

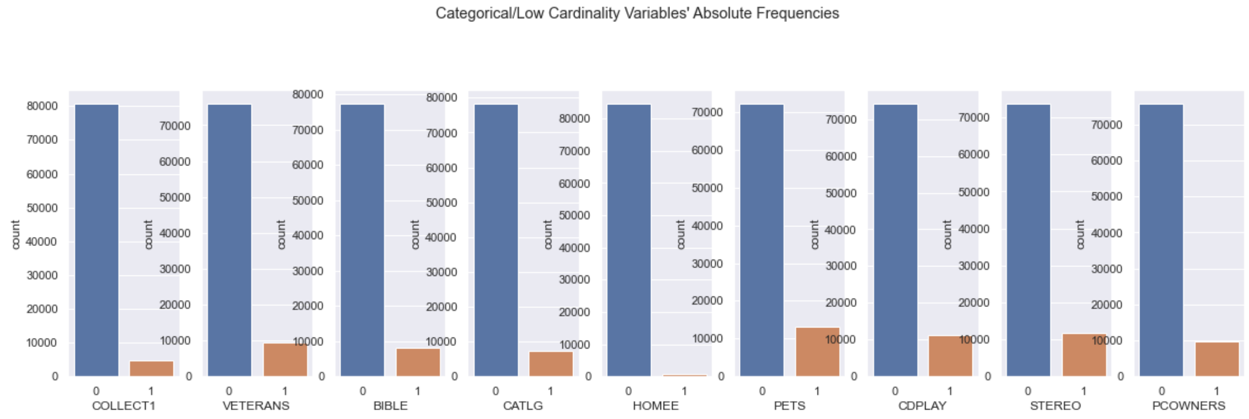


Table 1. Frequencies of Preferences.

C. Split data into Metric, Non-metric and Preferences

Since some of the variables had coded names that hardly explained their content, they were renamed with more obvious titles.
{'STATE':'State','RECSWEEP':'Sweepstakes_donor','GENDER':'Gender',’HIT':'Responses_topromotions' ,'NUMPROM':'Num_Promotions_total',’'NUMPRM12':'Num_Promotions_12months','RAMNTALL':'Total_donations','NGIFTALL':'Total_number_donations','MINRAMNT':'Smallest_donation_value', "MAXRAMNT": 'Largest_donation_value','LASTGIFT':'Last_donation_value',’ AVGGIFT': 'Average_donations'}

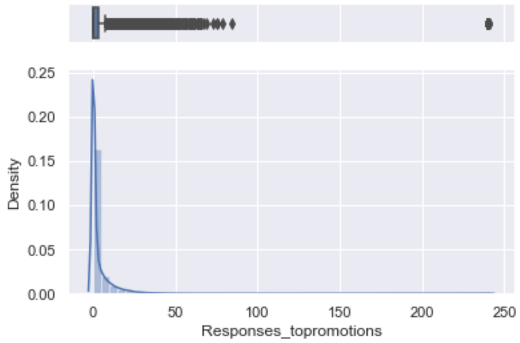
The variables were then split into metric, non-metric and preferences in order to continue with normalization and encoding, followed by cluster analysis.

D. Visualizations and Outlier removal

D.1. Numerical variables:

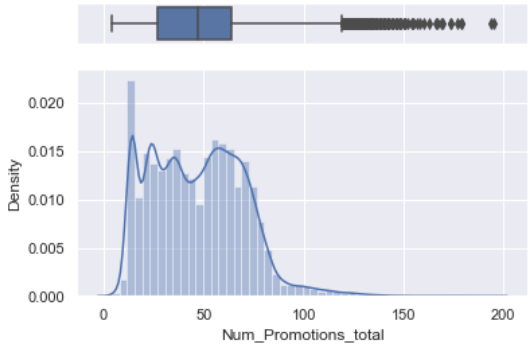
After plotting the charts for each variables, the group noticed that there were outliers that should be removed (further reducing the noise).
The first approach to remove the outliers from the variables was the IQR method, however, it proved to exclude a considerable percentage of information (approximately 18%), which was of course not ideal. Instead, the group decided to remove them manually after plotting each of them and still considering the IQR.

- **Responses_topromotions**



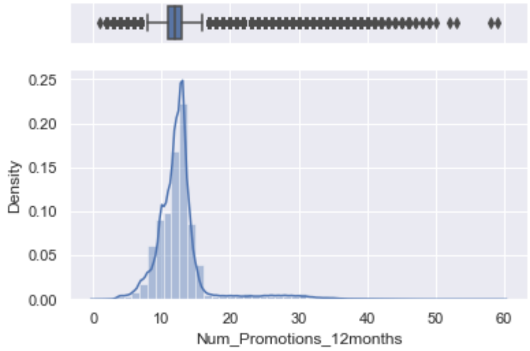
According to the chart and the boxplot above, all the rows that have values above 50 were dropped. (0.2% of the data was dropped)

- **Num_Promotions_total**



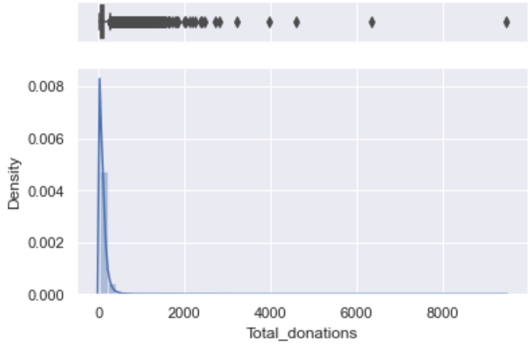
According to the chart and the boxplot above, all the rows that have values above 125 were dropped. (0.27% of the data was dropped)

- **Num_Promotions_12months**



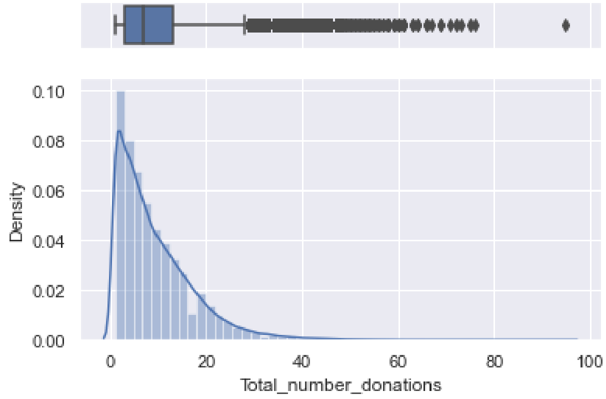
According to the chart and the boxplot above, all the rows that have values below 5 and above 50 were dropped. (0.48% of the data was dropped)

- **Total_donations**



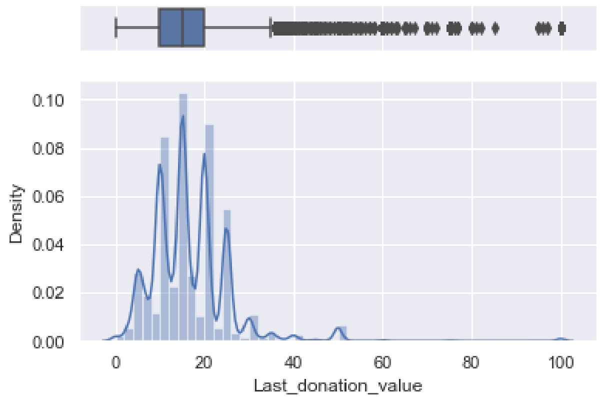
According to the chart and the boxplot above, all the rows that have values above 500 were dropped. (0.83% of the data was dropped)

- **Total_number_donations**



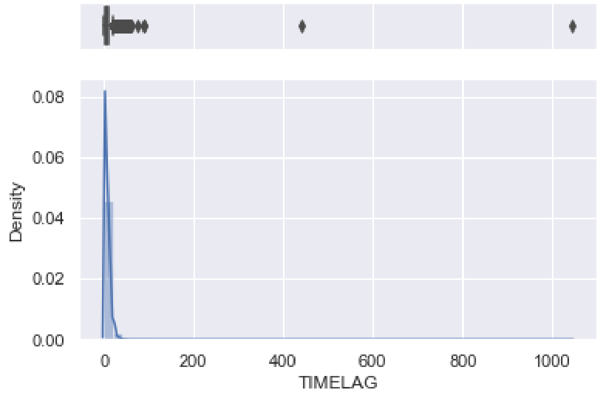
According to the chart and the boxplot above, all the rows that have values above 35 were dropped. (1.26% of the data was dropped)

- **Last_donation_value**



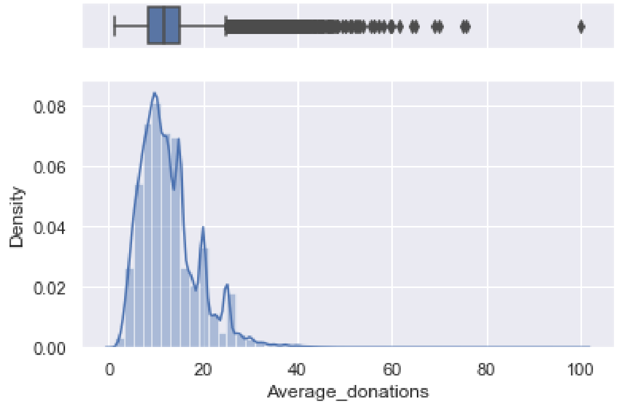
According to the chart and the boxplot above, all the rows that have values above 50 were dropped. (0.46% of the data was dropped)

- **TIMELAG**



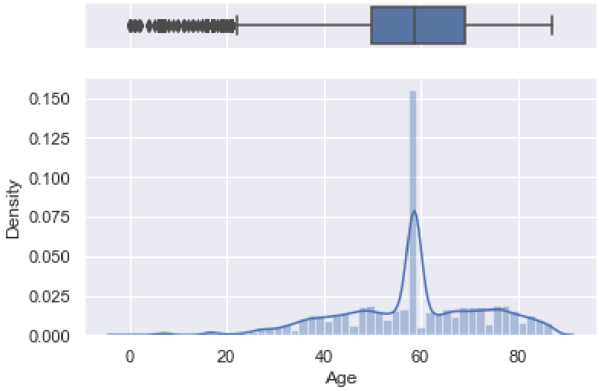
According to the chart and the boxplot above, all the rows that have values above 40 were dropped. (0.13% of the data was dropped)

- **Average_donations**



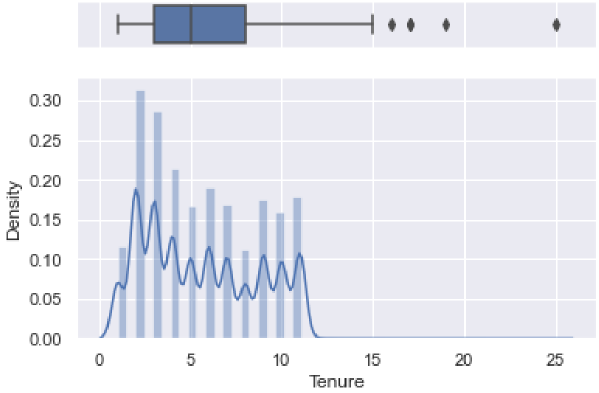
According to the chart and the boxplot above, all the rows that have values above 30 were dropped. (1.1% of the data was dropped)

- **Age**



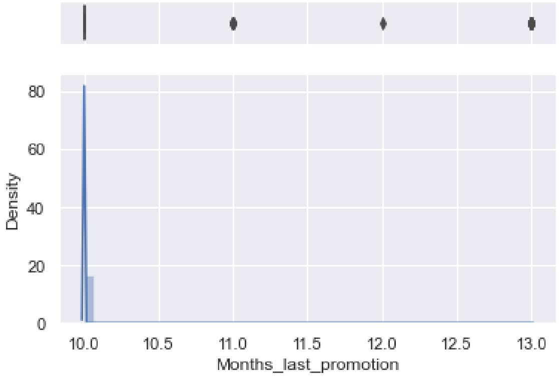
According to the chart and the boxplot above, all the rows that have values above 22 were dropped. (1.1% of the data was dropped)

- **Tenure**



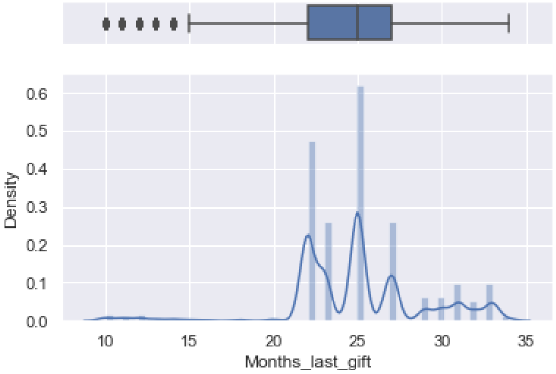
According to the chart and the boxplot above, all the rows that have values above 15 were dropped. (0.01% of the data was dropped)

- **Months_last_promotion**



According to the chart and the boxplot above, all the rows that have values above or equal 11 were dropped. (0.03% of the data was dropped)

- **Months_last_gift**



According to the chart and the boxplot above, all the rows that have values below 15 were dropped. (2.95% of the data was dropped)
The manual outlier removal for the numerical variables, resulted in losing 11% of the data.

D.2. Categorical variables:

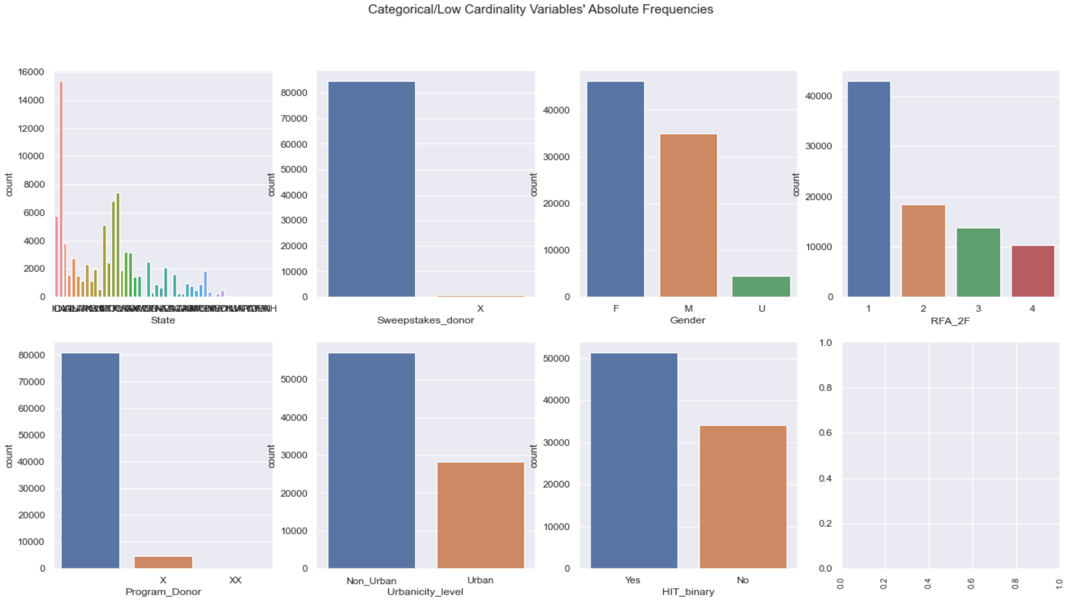
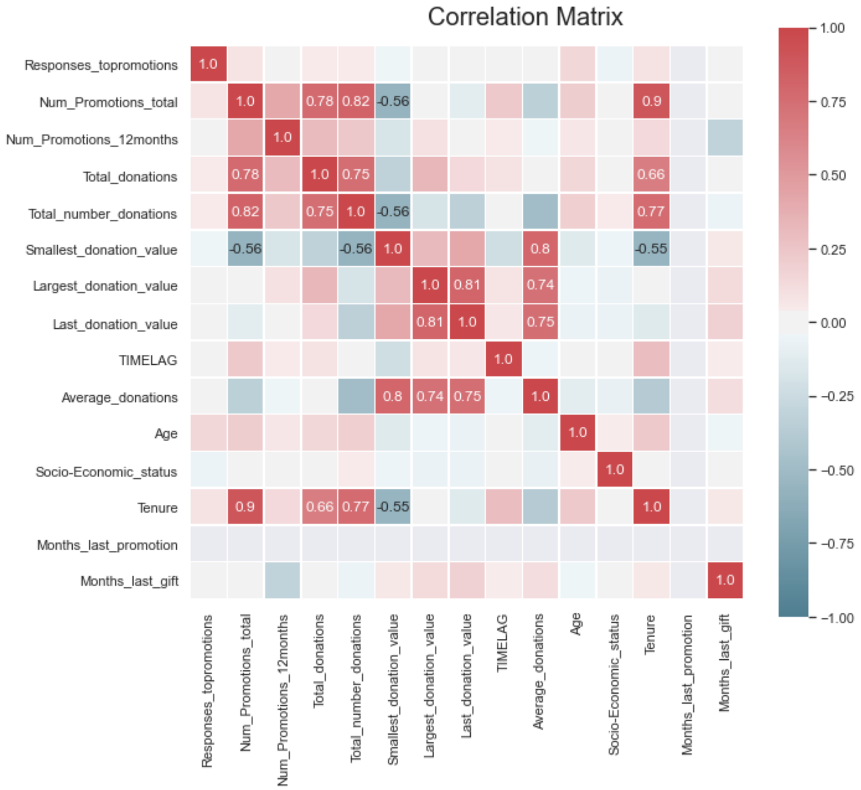


Table 2. Frequencies for Categorical variables

Thanks to the visualization of the categorical variables, we already know that important information to build engagement (marketing) strategies are that: Women are the majority, most of the lapsing donors only donated once, most of them respond to mail promotions, and also reside on suburban areas.

E. Correlation Matrix



Because the relationship between the correlated values were logical -i.e., the total number of donations and the total number of promotions sent to the donor, since the more you encourage a donor to participate (donate), the more likely it is for them to actually contribute to the programs- the only variable dropped after this analysis was the number of months since the last promotion (PVA sent to the donor) since this feature had only two values: 10 and 12 moths.

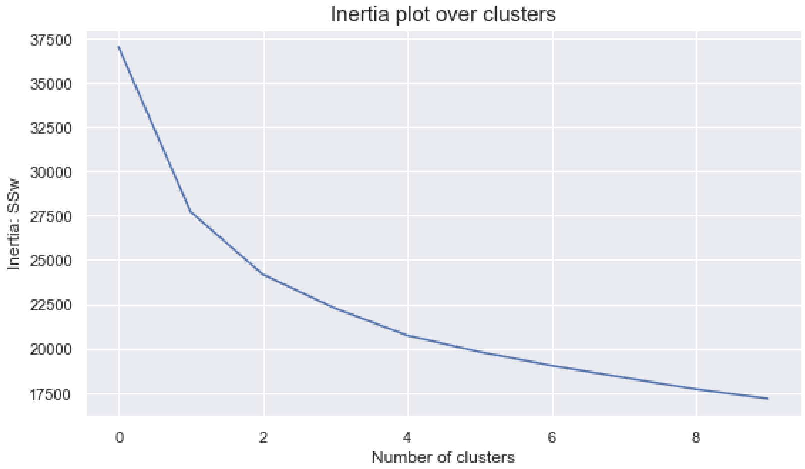
Just as the existing relationship between how many promotions were sent with how many donations each person made, was positive and logical, it also made sense for this group to be in the Lapsing category for 2017 since only one promotion was sent to them in almost a year.

III. RESULTS

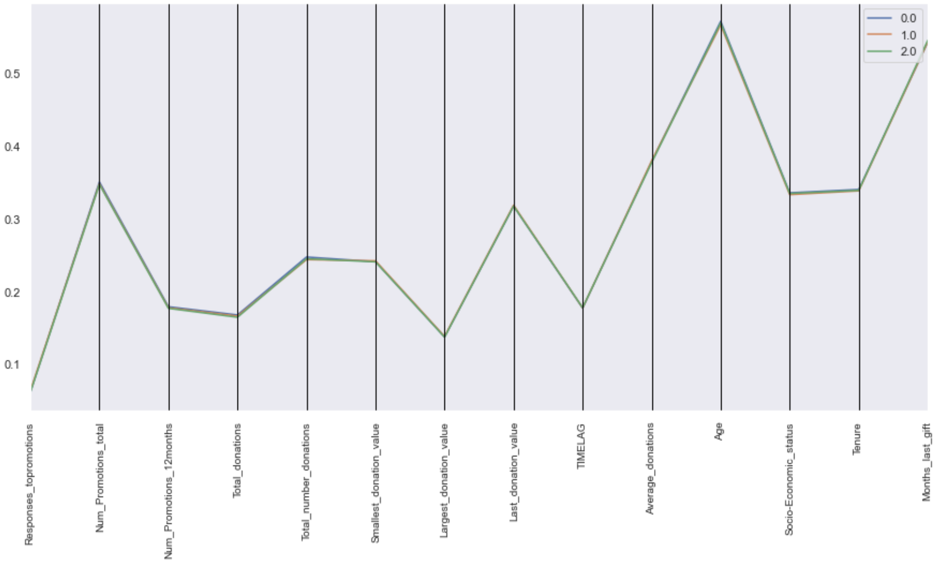
The algorithms applied to the preprocessed data, for clustering were KMeans and a combination of KMeans with Hierarchical clustering. The MinMaxScaler was used to normalize the metric variables of the dataset before running the algorithms.

1. Clustering analysis

a. KMeans

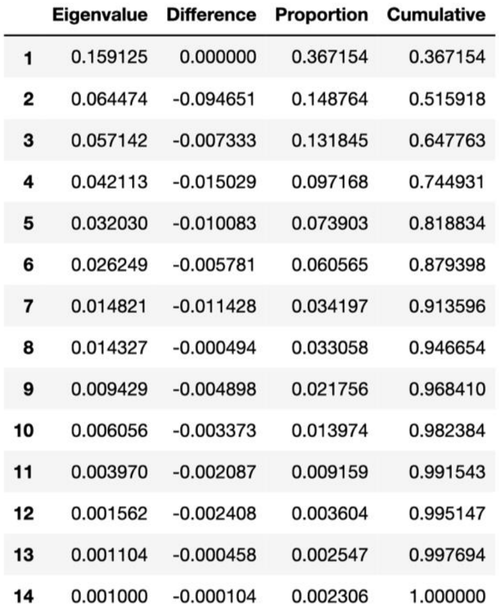


Because of its simplicity, we first implemented the KMeans algorithm and chose the number of clusters after running it a first time, and use the elbow method to identify an appropriate number of clusters. We also relied on the Silhouette score to determine the appropriate number, and both methods coincided. According to the line plot of inertia, and the Silhouette scoring (average Silhouette score for 2 0.223), the optimal number is two (2) clusters. Despite this and because the goal of the project are marketing campaigns, the group agreed that there should be more.

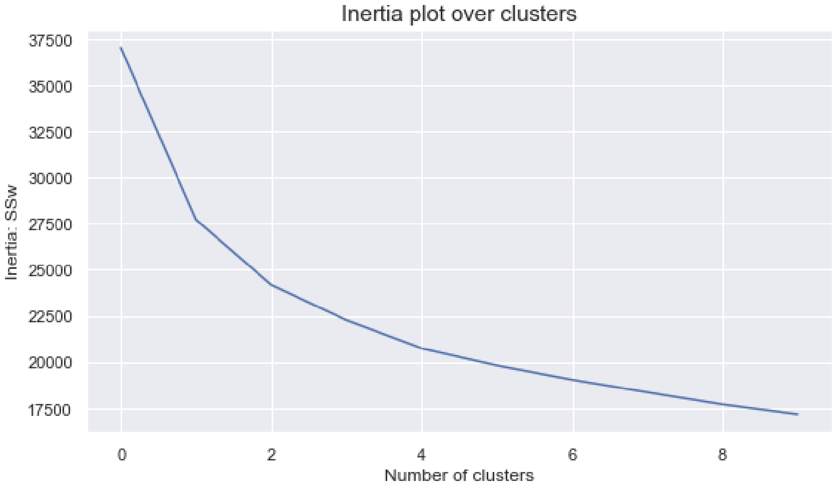


Nontheless, the final cluster solution for this method alone (as illustrated by the line plot above), showed that metric variables for all the clusters had the same behaviour, thus rendering us unable to identify any distinguishing factors between clusters with this algorithm. In this case, the application of Kmeans failed to segregate the customers properly. Because the quality of the outcome relies on the quality of the input, the team revalated the preprocessing and decided to test Kmeans using the resulting Principal Coponents (PCA). Although striving to keep as much information about each donor as possible, Kmeans becomes less effective as the dimensions increase. One of the strategies to reduce dimensionality to improve Kmeans' performace is PCA, however the combination showed exactly the same behaviour.

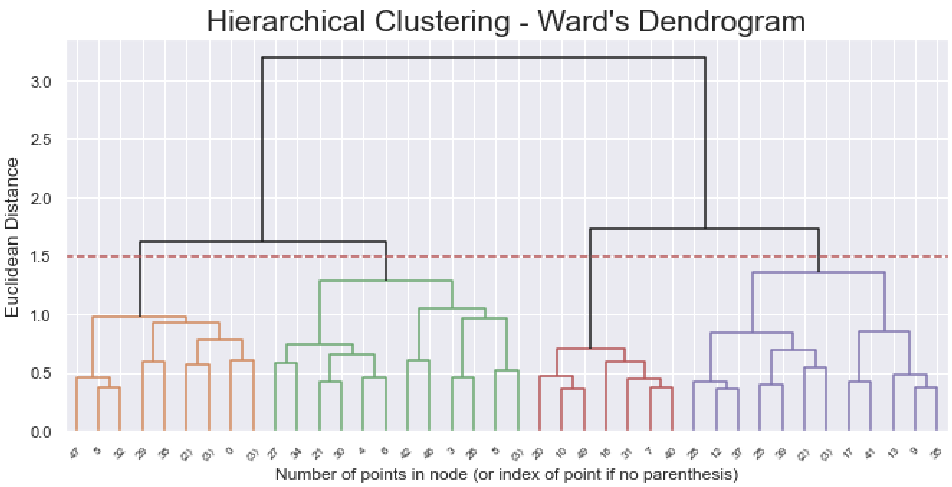
Since no significant difference was found, the original Kmeans analysis was kept as a reference for further model comparison. In hindsight, PCA could have been used for a larger dataset (more features from the 467 originally contained in the database) and found more appropriate components.



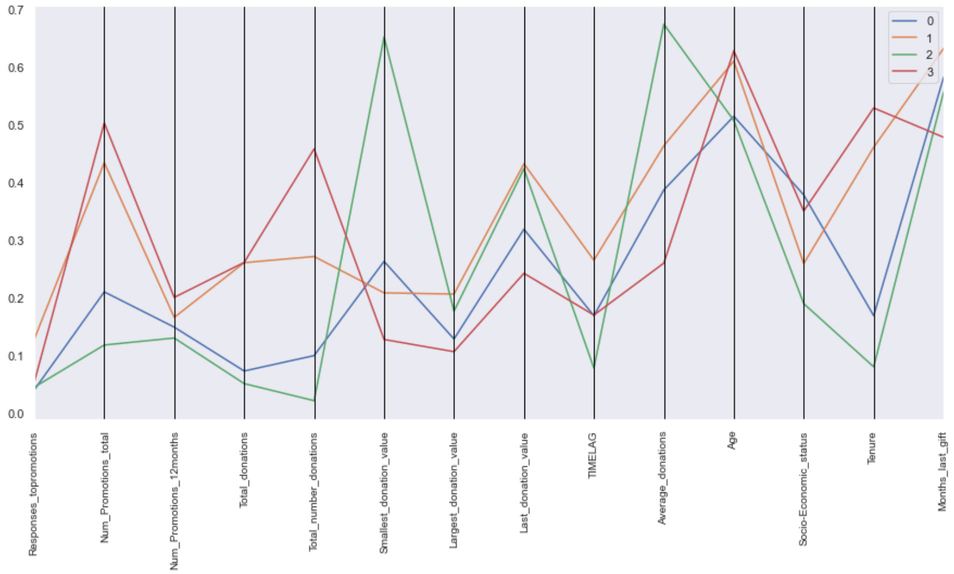
	Eigenvalue	Difference	Proportion	Cumulative
1	0.159125	0.000000	0.367154	0.367154
2	0.064474	-0.094651	0.148764	0.515918
3	0.057142	-0.007333	0.131845	0.647763
4	0.042113	-0.015029	0.097168	0.744931
5	0.032030	-0.010083	0.073903	0.818834
6	0.026249	-0.005781	0.060565	0.879398
7	0.014821	-0.011428	0.034197	0.913596
8	0.014327	-0.000494	0.033058	0.946654
9	0.009429	-0.004898	0.021756	0.968410
10	0.006056	-0.003373	0.013974	0.982384
11	0.003970	-0.002087	0.009159	0.991543
12	0.001562	-0.002408	0.003604	0.995147
13	0.001104	-0.000458	0.002547	0.997694
14	0.001000	-0.000104	0.002306	1.000000



a. Kmeans and Hierarchical clustering

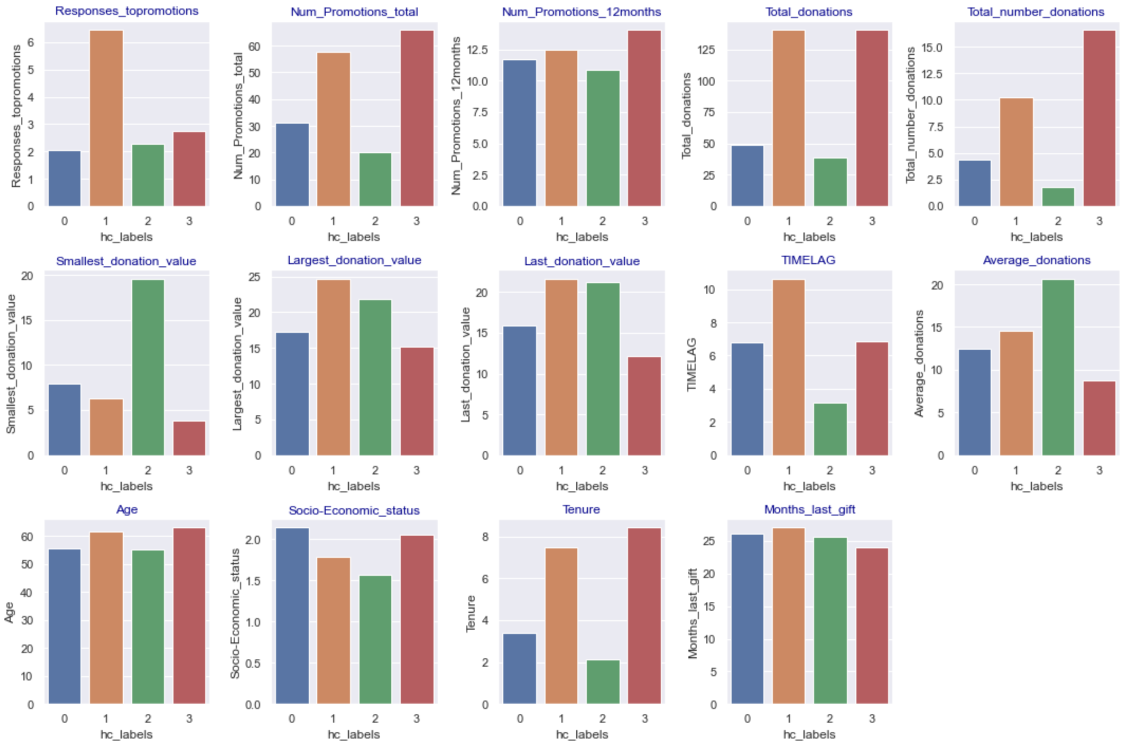


The second algorithm implemented was a combination of KMeans and Hierarchical Clustering (agglomerative). Initially, the KMeans algorithm was set to split the dataset in fifty (50) clusters and defining centroids for each of them. Subsequently, Hierarchical Clustering was applied on those fifty centroids to later identify clusters through the dendrogram. Because of the resulting structure the number of clusters was set 4.



The final characterization of each cluster is (given by the Line plot above), showed a better differentiation in cluster behavior for each metric variable, than the KMeans method on its own.
Although it is relatively simple to implement, Hierarchical clustering was not performed on its own since it would take a significant amount of computing power. Running it in combination with Kmeans reduced the need for computing and using both methods also reduced the risk of a poor clustering solution.
Clustering could have been improved by further data cleanup and engineering. However, given the scope of the project and the number of features, the majority of the variables were not considered.

b. Final cluster description



In order to implement different marketing campaigns to engage lapsed donors we will need to describe the most relevant details about each cluster we obtained.

For the first cluster (**cluster 0)** we have individuals that have been donors for not very long and have donated few times and with neither extremely high nor extremely low amounts of money.

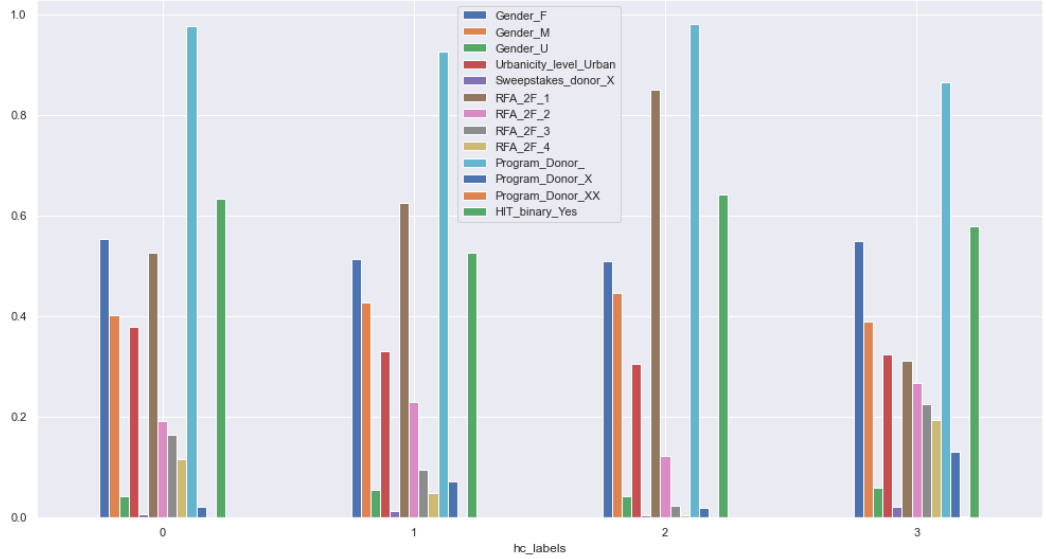
For **cluster 1** we have donors that respond the most to other mail order offers that are not PVA's and they receive a lot of promotions from PVA. These are also one of the biggest donors in terms of total amount of dollars given and dollar amount of largest donation to date, yet on average they do not donate much more than the other segments, since they also have donated various times and have given small donation values as well. They have one of the longest histories of donations to PVA with a long donating hiatus between their first and second donation.

For **cluster 2** we have donors that have given very few times but always with substantial amounts of money, they received the least number of promotions from PVA. Their first donation was only a couple of years ago, therefore they have been lapsed for most of the time they have been donors.

For **cluster 3** we have donors that have donated the most times, but with the lowest values, they have also received the most promotions from PVA and have been donors for the longest time.

Between 2015 and 2016 the proportions of donors that have made 1 or 2 or 3 or 4 or more donations are similar in cluster 3, are overwhelmingly only 1 in cluster 2.

The donors in all four segments are in general middle aged, mostly female, of average socio-economic status, mostly from non-urban neighborhoods, received 10-13 promotions from PVA in the last 12 months, mostly non program donors, mostly have responded to other mail order offers other than PVA's and a larger proportion of donors are from California.



2. Marketing Strategy

Now that we know the behavior of each donor cluster, we recommend the following personalized marketing strategy that will be more effective to retain this population and make sure they will keep donating in the future. Since cluster 1 and 3 donors have a long history with PVA, the organization should invest more resources on them. We suggest sending a 'Thank You' note in which we share what their earlier support has helped the organization achieve. Particularly for cluster 3 we should advertise a recurring giving program where they can donate a pre-determined amount of money, of their choosing, every two months. For cluster 1 we should share the real-life impact of the organization's recent activities, so they can be aware what their gifts would go toward. Cluster 0 and 2 donors gave very few times before lapsing. Notably cluster 2 has given high values in all donations, so we should also advertise a program for giving a pre-determined amount, but in this case only once a year, possibly only during the holiday season. For cluster

0 we suggest sending a letter reminding them of their past contributions and explain how small donors are invaluable to PVA's mission.

IV. REFERENCES

Arbel, I. *What is PCA and how can I use it*. Retrieved from: https://www.bigabid.com/blog/data-what-is-pca-and-how-can-i-use-it

Bektas, M. (2020, May). *Customer Segmentation with Clustering Algorithms in Python*. Retrieved from: https://medium.com/@mbektas/customer-segmentation-with-clustering-algorithms-in-python-be2e021035a

Ibrisevic, I. (2020, October). Step by Step Guide For Regaining Your Lapsed Donors. Retrieved from: https://donorbox.org/nonprofit-blog/regaining-your-lapsed-donors/

Aldecoa R., Marín I. (2010, July). *Jerarca: Efficient Analysis of Complex Networks Using Hierarchical Clustering*. PLoS ONE 5(7): e11585. https://doi.org/10.1371/journal.pone.0011585

Statquest, (2018, April) Josh Stamer, Principal Component Analysis, https://www.youtube.com/watch?v=FgakZw6K1QQ