Universidade NOVA de Lisboa
NOVA Information Management School

PARALYZED VETERANS OF AMERICA DATA MINING PROJECT

Authors:
Eleonora Sbrissa (20200628)
Luís Reis (20200636)
Pedro Godeiro (20200398)

Lisbon
January 2021

1. INTRODUCTION

To guide us through this project we decided to use the The CRoss Industry Standard Process for Data Mining, else known as CRISP-DM. It's like a set of guardrails to help you plan, organize, and implement your data science project. It has since become the most common methodology for data mining, analytics, and data science projects. [1]

2. BUSINESS UNDERSTANDING

This dataset was provided by the Paralyzed Veterans of America (PVA). PVA is a non-profit organization that provides programs and services for US veterans with spinal cord injuries or disease. The organization raises money through promotions sent by mailing, as well as donations done by people (the money the organization receives is called a "gift").

The organization would like to better understand their donor segmentation in order to understand how the donors behave and identify the different characteristics of potential donors, in order to raise the highest amount of money possible. The dataset we analyzed is a sample of the results of one of PVA's recent fundraising appeals, containing data related to 95412 donors that made their last donation 13 to 24 months ago, so called lapsed donors.

The task itself consisted in aggregating the donors into clusters, extracting their behaviour and insights on how to better approach them, for a future marketing approach on each cluster.

3. DATA UNDERSTANDING

The initial dataset contains 95412 rows × 475 columns. We have information related to the gifts done by donors to the organization, but we also have info related to the donors' interests, as well as characteristics of their neighborhood (data collected from the 2010 Census). We have different types of variables: categorical, metric variables and dates. as we can see on Figure 3.1.

Figure 3.1 - Visualization of the Dataset

	ODATEDW	OSOURCE	TCODE	STATE	ZIP	MAILCODE	PVASTATE	DOB	NOEXCH	RECINHSE	...	AVGGIFT	CONTROLN	HPHONE_D	RFA_
0	01/01/2009	GRI	0	IL	61081			01/12/1957	0		...	7.741935	95515	0	
1	01/01/2014	BOA	1	CA	91326			01/02/1972	0		...	15.666667	148535	0	
2	01/01/2010	AMH	1	NC	27017			NaN	0		...	7.481481	15078	1	
3	01/01/2007	BRY	0	CA	95953			01/01/1948	0		...	6.812500	172556	1	
4	01/01/2006		0	FL	33176			01/01/1940	0	X	...	6.864865	7112	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
95407	01/01/2016	ASE	1	AK	99504			NaN	0		...	25.000000	184568	0	
95408	01/01/2016	DCD	1	TX	77379			01/01/1970	0		...	20.000000	122706	1	
95409	01/01/2015	MBC	1	MI	48910			01/01/1958	0		...	8.285714	189641	1	
95410	01/01/2006	PRV	0	CA	91320			01/05/1960	0	X	...	12.146341	4693	1	
95411	01/01/2008	MCC	2	NC	28409			01/01/1938	0	X	...	96.794872	185114	1	

95412 rows × 475 columns

We know that all donors in this dataset are lapsed donors, as we can see in the feature “RFA_2R” related to Donor's RFA status on the last promotions mailed. It contains only ‘L’ value, which stands for Lapsed Donors, as shown in figure 3.2:

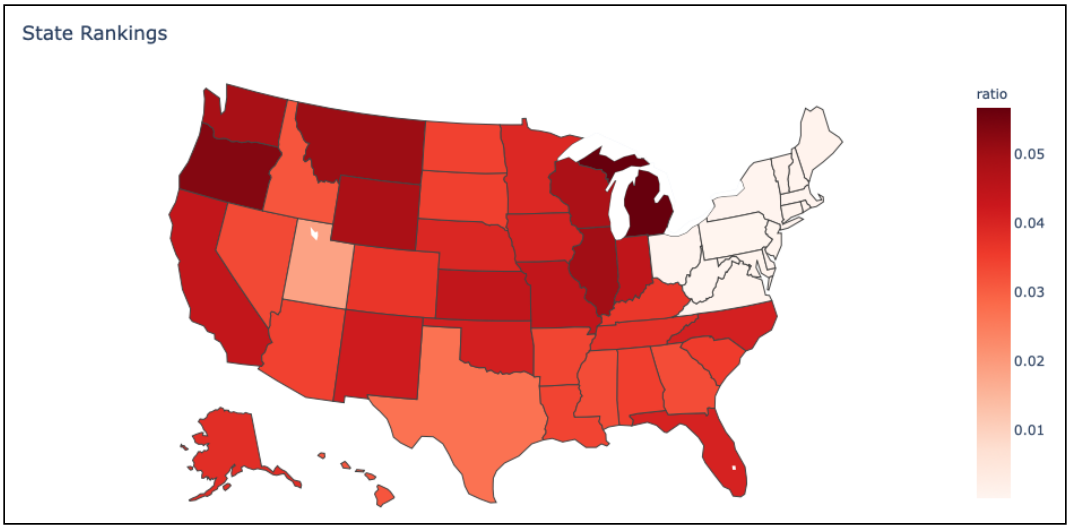
Figure 3.2 - RFA_2R Status

```
1 data1['RFA_2R'].value_counts()
```

```
L    95412
Name: RFA_2R, dtype: int64
```

Another interesting thing we can get out of the dataset is to understand where the lapsed donors are coming from. First of all it is important to notice that some states in the US naturally have a bigger population than other ones, so it wouldn't make much sense to do an analysis on the amount of lapsed donors in which state, instead we are analyzing how many percent of the population of each state is a lapsed donor for PVA. This information is available on Figure 3.3.

Figure 3.3 - USA Heatmap for Proportion of Donors [9-10]



As we can analyze on Figure 3.3, PVA has very few lapsed donors on the northeast Countries, which raised some hypothesis: for example we can assume that the fact that PVA´s headquarters is on this Region (Washington DC) may influence the donations or even that PVA´s marketing strategies on this region is working well. Meanwhile, in the north and northwest of the country we have a lot of lapsed donors, what it may mean is that PVA´s marketing strategies are not working well in these regions. This kind of information may be really important for PVA to understand what does and doesn't work regarding their marketing strategies.

4. DATA PRE PROCESSING

As the dataset had a considerable amount of variables, we decided to divide them into small datasets in order to analyse the correlation between the different features and drop the ones that were highly correlated. The 4 macro subsets created for this purpose are: Personal Info, Household, Neighborhood and State. We then had a closer look on the Giving History File and to all info related to Mails.

Inside each one of these subsets we then grouped the variables according to logic in order to do an initial analysis of data types, do some data transformation, exclude incomplete variables (with more than a half of missing values), and find some kind of correlation. A closer look to all this analysis is available in the notebook.

Through data transformation, feature engineering and dimensionality reduction we found the following variables that will be used for the clustering:

PERSONAL INFO:

- GENDER_FEMALE: Binary variable created from the categorical feature “Gender”. 0 is considered as male, 1 is female. We have 5037 null values (the joint account was interpreted as null values, we will then input the missing values before the clustering)
- AGE: Created from Date of Birth, using as reference year 2017. The average age of this dataset is 60 years. We have 23883 null values
- PEP_status: Transformed into a binary variable. In this dataset we have 45267 Politically Exposed People (47%).

HOUSEHOLD

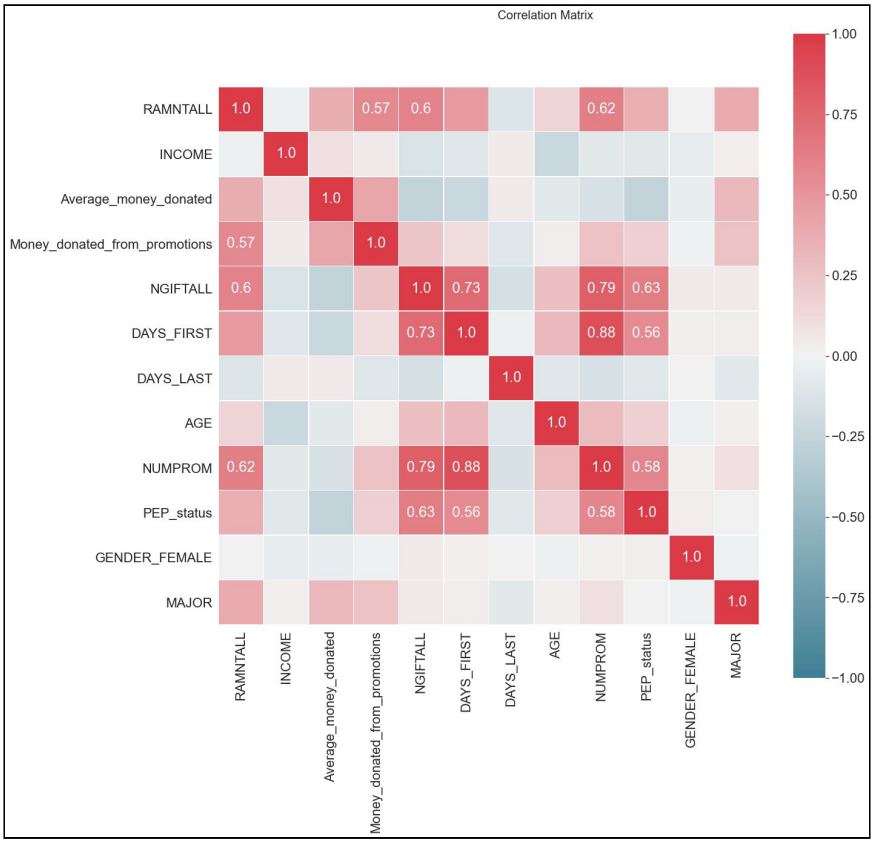
- INCOME: Household Income, from 1 to 7. We assume that 7 stands for richer people, 1 for poorer ones. We have 21285 missing values

GIVING HISTORY FILE

- MAJOR: Transformed into a binary variable. We have 292 major donors in this dataset
- Money_donated_from_promotions: Amount of money donated from promotions. Created from the horizontal sum of all the columns related to RAMNT, total amount donated for each promotion.
- NUMPROM: Number of promotions received.
- RAMNTALL: total dollar amount of lifetime gifts to date
- NGIFTALL: Number of lifetime gifts to date
- Average_money_donated: Average amount of money donated to the organization. Created from the ratio between RAMNTALL (total amount of gifts) and NGIFTALL (total number of gifts).
- DAYS_FIRST: number of days since the first gift. The reference date is 31/12/2017. Created from FISTDATE
- DAYS_LAST: number of days since the last gift. The reference date is 31/12/2017. Created from LASTDATE

This decision was made after having a look at all the metric variables that were available after the preprocessing (213 columns) and deciding which were the ones that would be more important for a customer segmentation and therefore for clustering. We can have a look at their correlation on Figure 4.1.

Figure 4.1: Correlogram for Variables used on Clustering



We can see that NUMPROM, which is the number of promotions received up to date, is highly correlated to DAYS_FIRST, which is the number of days since the last promotion. We will then exclude NUMPROM from the variables used for clustering.

4.1 Missing Values

After the selection of the variables, we need to check for the presence of missing values in our features. We can see that on Figure 4.2.

Figure 4.2 - Presence of missing values on the features used for clustering

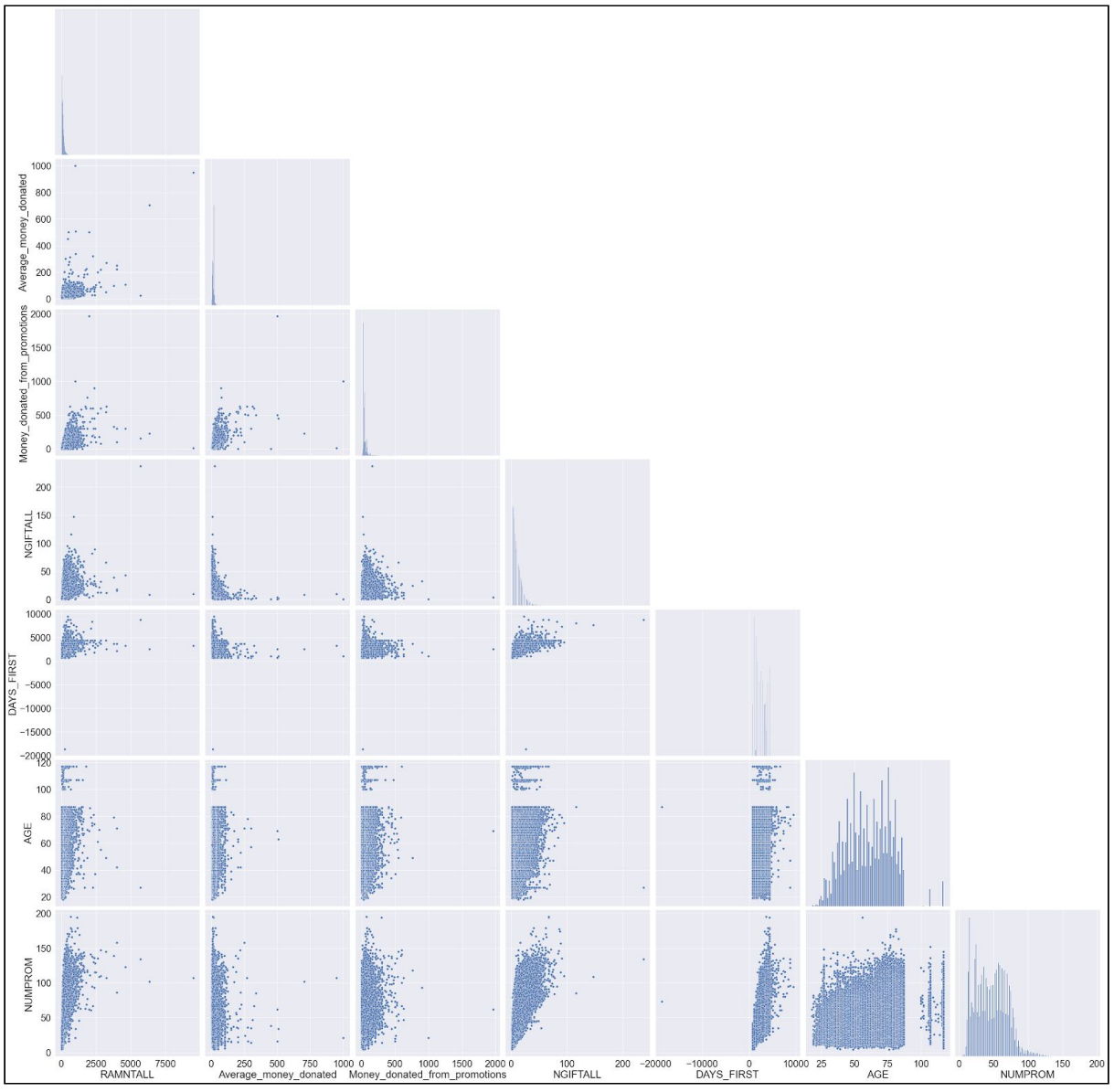
RAMNTALL	0
INCOME	21285
Average_money_donated	0
Money_donated_from_promotions	0
NGIFTALL	0
DAYS_FIRST	2
DAYS_LAST	0
AGE	23882
PEP_status	0
GENDER_FEMALE	5037
MAJOR	0

So realizing that we have missing values, we decided to input them using KNN Imputer, which takes into consideration the mean value from the nearest neighbor. We decided to keep the standard values for the imputer with 5 neighbors and consider that all points in each neighborhood are equally weighted [2].

4.2 Outliers

As you can see on Figure 4.3, we have many outliers in different variables.

Figure 4.3 - Outliers



As a way of detecting which observations are outliers, we decided to use the Local outlier factor [3] with standard parameters to detect them, and the algorithm detected 3% of the data as such.

After detecting the outliers, we took them out of the dataset before the clustering, and after that we inserted them back and allocated them to a specific cluster by a Decision Tree. To decide which classifier we would use, we splitted the data into training and validation and checked the performance of a Default Decision Tree Classifier, Random Forest and Gradient Boosting. The decision was made based on the Accuracy score and the time taken to run, available on Figure 4.4.

Figure 4.4 - Accuracy and Lapsed time of Classifiers

MODEL	ACCURACY	TIME TO RUN
Random Forest	95.2%	16 sec
Gradient Boost	95.3%	4mins 30sec
Decision Tree	90.7%	-

Even if the Gradient Boosting Classifier was the best one in terms of score, we opted to use the Random Forest Classifier since it took way less time to run.

4.3 Data Standardization

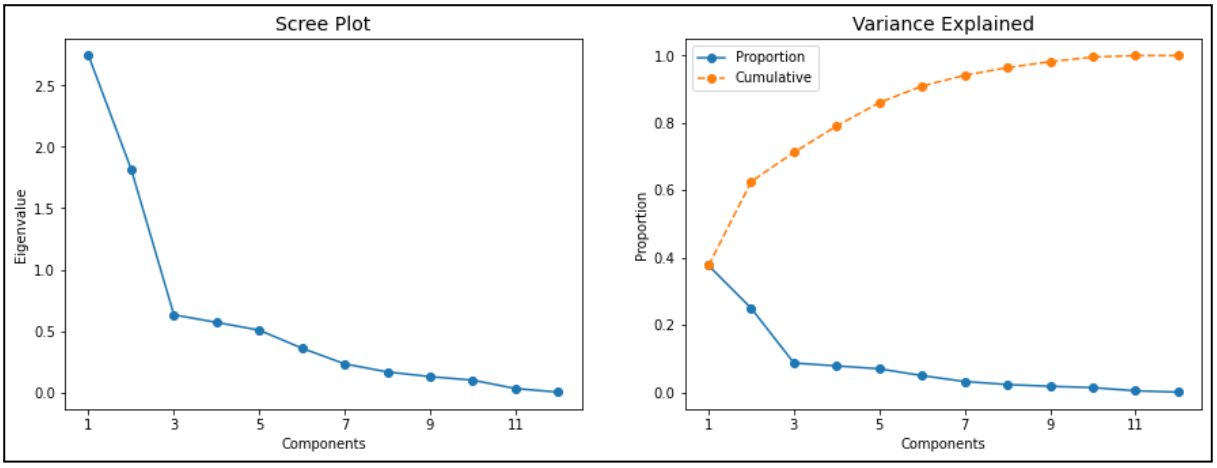
Before performing PCA we need to standardize our data [4-5] to provide a fair comparison between the explained variance in the dataset. We opted for the Robust Scaler[11] as it's the best one when we have a wide dataset, with different types of data and even if we removed 3% of the dataset with the local outlier factor we are sure there is still some data that may be outliers but were not detected by the algorithm.

4.4 Reducing Dimensionality

One good thing to do before clustering when you have many variables is to reduce the dimensionality of it. Considering that we have 11 variables to cluster, and we are running algorithms that may take a lot to run, we decided to use PCA to reduce the dimension and avoid redundancy of our clustering dataset.

PCA is a way of reducing dimensionality in which you get a new dataset with the same number of Principal Components as you have of features, but now they are sorted on the order from the most important component (PC0) to the least important (PCN as in N the number of features you have).[7] The key here to reduce dimensionality is to choose a specific number of Top PCA´s, and one good way to do so is by observing the elbow of the Scree Plot and the amount of Variance Explained by the Top PCAs, as we can observe in Figure 4.5.

Figure 4.5 - Scree and Variance Explained Plot of PCA



By analyzing Figure 4.5, we can see that the first elbow of the Scree Plot is on the Top 3 PCAs, but if we decided to choose only the Top 3, we wouldn't have even 80% of the variance explained. Taking this into consideration, we decided to use the second Elbow on 7, in which there is more than 90% of the variance explained and we used it for Clustering the Top 7 PCAs.

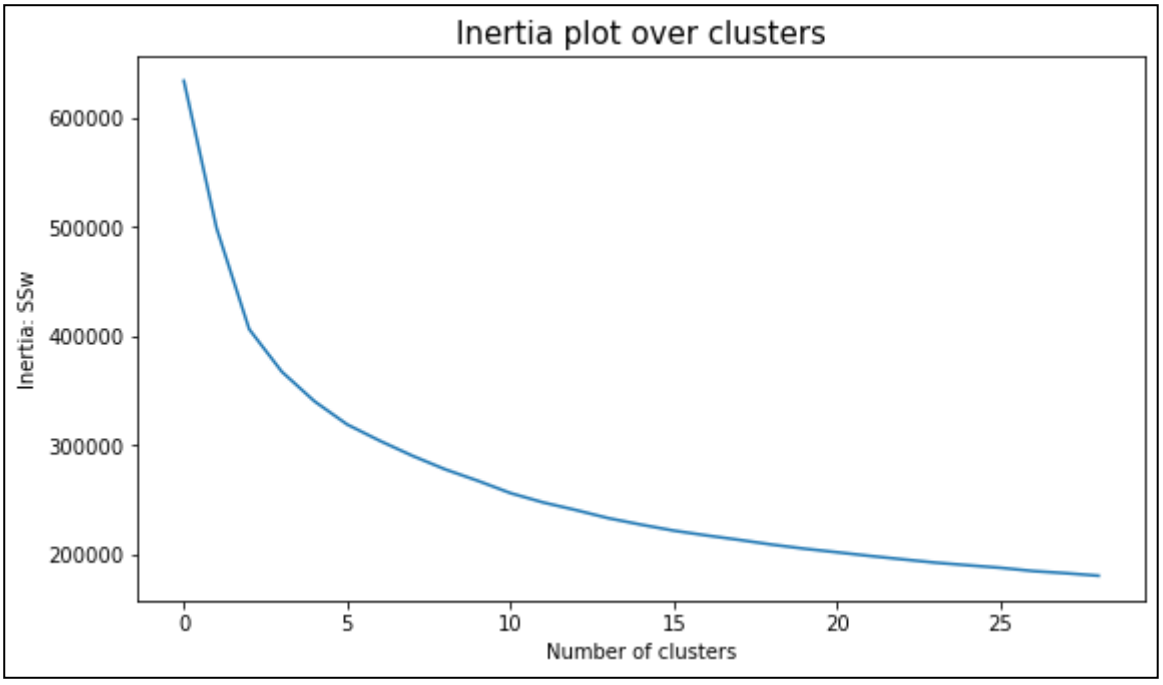
5. MODELING

5.1 Clustering Technique

After doing all of the pre-processing on the data, we needed to start working on how we are going to generate value to PVA via Marketing Segmentations. The way to do the marketing segmentations is by dividing the different kinds of lapsed donors into different clusters.

There are many ways to cluster data, and the method we decided to use for this Project was k-means on top of PCA. One of the key attributes of the k-means algorithm is the apriori number of clusters we are using on it. Having that in mind, one way to decide how many clusters are we using is by doing an inertia plot and looking for the “elbow” of the curve. The inertia plot can be seen on Figure 5.1.

Figure 5.1 - Inertia plot over k-means clustering



And with the Inertia plot we could see that there was no clear elbow on the data. At first we could suppose from the Inertia curve that we could have an elbow around 3 clusters, but when we tried to use 3 clusters, since the clusters contained such a big number of people, we couldn't identify differences between the clusters, so it would be impossible to create marketing segmentations. We experimented using different numbers of clusters and reached the conclusion that 15 clusters were potentially the best number of clusters. When we started analyzing the differences between the 15 clusters, we could see some interesting patterns, so we decided to use 15 clusters in this study.

5.2 Evaluation of the Clusters

To identify the characteristics of each cluster we decided to use the marketing personas. A persona in the context of marketing refers to the ideal customer or customers for your business. It’s a semi-fictional representation of your ideal customer based on market research and real data about your existing customers.[8]

By analysing all the features in the dataset (the ones that were used for clustering and the other ones) we could create 15 different profiles based on the characteristics we found in the neighborhood, in the donors interests, and in the donors’ amount of money donated to the organization.

5.2.1 Persona of the Cluster 0 (13177 people)

Meet Jane, she is 40 years old, mum of 4 children and a University professor while her husband is an Entrepreneur. They need to have 2 cars for the daily routine but one of them is big enough to gather all 6 of them when they want to go on family trips. The other one is an electric car. She is a member of the book club where she goes regularly.

5.2.2 Persona of the Cluster 1 (7572 people)

This is Edward, currently living in California with his cat Lucy and his dog Soul. He is a vet, living in a beautiful apartment in a wealthy neighborhood with people coming from different parts of the US and even from outside. He cares about the PVA organization and he donates from time to time.

5.2.3 Persona of the Cluster 2 (6164 people)

Get to know Jorge, he moved to the US from Colombia at the age of 23, currently working in a food restaurant. He is living in a suburban Area, in a small apartment with another friend. He likes to buy second hand clothes and furniture. In his free time he likes to construct pots. He loves to collect stamps and other objects. He is really interested in knowing other people's stories, and he is really interested in knowing the life of people that went to war.

5.2.4 Persona of the Cluster 3 (3165 people)

Here we have Jude, 32 years old living in a small town in Texas. He is a freelancer doing translations for external companies. In his free time he likes to read magazines in order to always keep up to date with all the trends. As he doesn't need to move out of town for work, when he needs to go out he uses public transports.

5.2.5 Persona of the Cluster 4 (4954 people)

Meet Manuel Santiago and Chloe, a couple from Argentina that started to work as gardeners, and now after a lot of years of work they are owners of a company that outsource gardeners to all the State, and even outside it. Sometimes they need to move outside the State to meet potential new customers. They now own a really nice house in a rich neighborhood, and even if they are super busy with their business, sometimes they like to donate to the veterans cause.

5.2.6 Persona of the Cluster 5 (583 people)

This is Janet, a 42 years old woman who is part of an elite social group. She's white, like all of her friends. A good part of her friends are married to veterans, but Janet is not. Janet's husband works in a regular corporate job, but he's paid really

well. By consequence, they have a lot of money, they just don't donate a lot of money. Everytime they get promotions, they are always donating, but it is rare situations in which they donate a good amount of money in just one donation, even though it has already happened.

5.2.7 Persona of the Cluster 6 (161 people)

Here we have Walter, a 64 years old retired executive from a big bank on a commercial center. He is married to Margareth, who is 62 years old. They are both retired and they decided to live in a luxurious condominium away from the city. Walter studied a lot, till he got an PhD on Finance when he was 27. Walter used to donate really frequently since a long time ago, but stopped donating a while ago too. He never had big money contributions, his key thing was about donating really frequently. One of the hobbies Walter is really interested in now is gardening.

5.2.8 Persona of the Cluster 7 (775 people)

Meet Adonis, a 34 year old man who lives on the suburbs in a neighborhood well known for being a black community. While growing up, Adonis never had the opportunity to go deep on studying, so he never went to college. Now he is working between different gigs, mostly on things related to agriculture, so he spends a lot of time on public transportation to reach his workplace. Regarding his donor behaviour, Adonis is donating frequently, but always on small amounts.

5.2.9 Persona of the Cluster 8 (10858 people)

This is Alex, a mid age woman who comes from a job with high income. Alex is the kind of donor who tends to donate a lot of money each time she is donating, but she usually doesn't donate frequently. One of the things Alex has as a hobby is crafting.

5.2.10 Persona of the Cluster 9 (11760 people)

This is Ross, a 55 years old guy , owner of a luxurious car, who has been working in a high end job for his whole life, which means he has a lot of money. He does donate frequently, but even though he has a lot of money, he doesn't tend to donate big amounts of money.

5.2.11 Persona of the Cluster 10 (2661 people)

Meet Jack, 49 years old, he works in Hollywood as an actor and lives in a mansion in the suburbs. His wife is also in the entertainment area. Their mansion has a lot of built in technology. They own 2 nice cars but they rarely use them, they mostly travel by uber or other fancy transportation. One of the key things about this

family is that they don’t seem to help very often organizations like PVA and when they do the amount is also low.

5.2.12 Persona of the Cluster 11 (10579 people)

Janet, 39 years old, single mom of 2 boys. They live in a modest house in the suburbs, she owns a low price car, which is their means of transportation. Her husband died in war and she now struggles to earn money, working in multiple jobs and always searching for better ones.

5.2.13 Persona of the Cluster 12 (1568 people)

Get to know Carlos, 60 years old, mexican immigrant, lives in Texas with his family, has a nice house and cars. Carlos has multiple hobbies and collects rare items, he and his family are christians so he often helps monetarily good causes like PVA. His 2 sons are finishing college so they have a lot of technological material to work with.

5.2.14 Persona of the Cluster 13 (11723 people)

Meet William, 52 years old, american with parents coming from Nigeria. He is a family man, married and 2 daughters. He is the main provider of the family, going from gig to gig so he can provide a good future for his children.

5.2.15 Persona of the Cluster 14 (9688 people)

Meet Kenny, 54 years old, lives in the city with his wife, they are both professors in a nearby college. They have 2 sons, who are currently finishing high school. They are christians, so they go to church every sunday. They own an apartment and 2 cars. Kenny and his wife really like helping organizations like PVA.

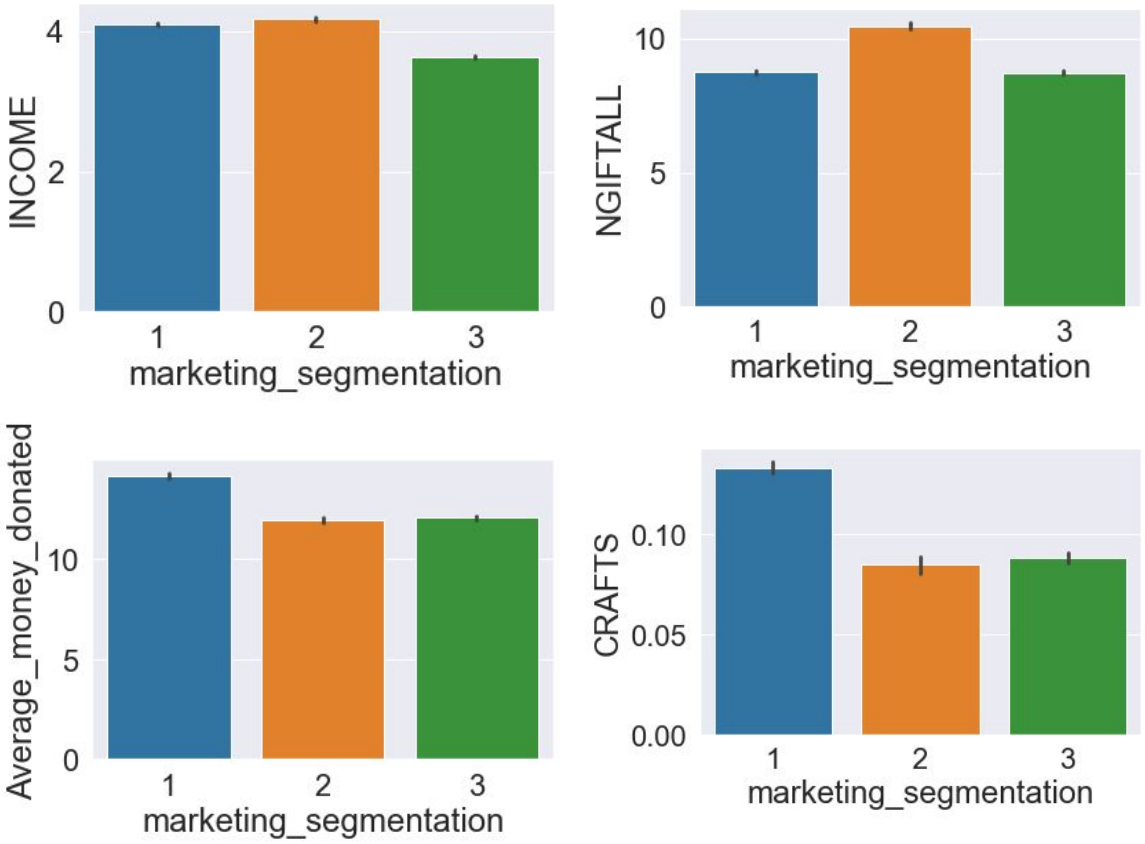
5.3 Deployment: Marketing Segmentations

As said on 4.1, the reason why we are clustering the data is because we want to have different marketing segmentations of the PVA´s lapsed donors. The problem that we are facing is that with less clusters we couldn't find patterns in our data and having too many clusters as 15 clusters, it would be difficult for PVA to create 15 different marketing campaigns.

Having that in mind, we decided that after clustering our data, we would group similar clusters in a way that we would have less Marketing Segmentations. By doing that, we managed to reduce by a lot the number of Marketing Segmentations that we had, going from 15 to 3 different marketing segmentations.

Another thing is that we managed to find interesting patterns on the different groups of marketing segmentations, as we can see on Figure 5.2

Figure 5.2 - Different variables for different Marketing Segmentations



5.3.1 Marketing Segmentation 1 (34660 people)

For this marketing segmentation we decided to have the wealthier people who didn't use to donate frequently, but when they do it, it would be with large amounts. On this marketing segmentation we find clusters 1, 4, 8, 12 and 14.

We can assume that the problem we face with this group is about making the act of donating as a habit, as we can see that when they donate the average amount of money is high. Another approach that may be good to get them back to active donor status is the interest in crafts.

Assuming that, we can first use the fact that they are interested in crafts and look for a partnership with crafts companies to bring the lapsed donors back. After that, we can create some kind of fidelization plan, in which if they donate more frequently they would have access to exclusive rewards, as an example PVA´s events such as special auctions.

5.3.2 Marketing Segmentation 2 (25681 people)

In marketing segmentation 2 we still have a group of wealthier people, but in this case we have a group that donates frequently but when they donate they don't donate large amounts. For this marketing segmentation we have Clusters 0, 5, 6 and 9.

For that case, the problem is that this group of clusters is composed of people who have a lot of money and can donate large amounts but they actually dont. Taking that into consideration, one good hypothesis related to the reason why they are not donating bigger amounts is the fact that they are not as connected to the cause as they could.

Considering that the problem is how connected this group is with the cause, one good idea is having paralyzed veterans telling their stories through interviews and using the provided contacts of the lapsed donors to show this kind of content would be something that may connect this kind of people more to the cause.

5.3.3 Marketing Segmentation 3 (35067 people)

In marketing segmentation 3 we have a group of people who have a smaller income, which makes them donate less in terms of frequency and average money donated. This marketing segmentation is for clusters 2, 3, 7, 10, 11 and 13.

These donors are definitely the more consistent ones, as they have the highest percentage of donations (since they are a bigger group than Marketing Segmentation 2), though their amount donated each time is low compared to other groups of donors. The thing that we should be more aware of is the fact that this group donates less not because they don't have the habit neither are not connected to the cause, the key factor here is money limitations.

So, to get these people to donate more we need to invest in promotions, and a valid approach could be sending at first a promotion that includes a card, and at each next donation they make they can participate in sweepstakes with the possibility to win multiple prizes.

6. CONCLUSIONS

This dataset contained a wide variety of variables that came from different sources. We found the metadata a bit poor of information in some parts and all this made the preprocessing long and hard to do. In the middle of a lot of noise, the selection of the metric features for the clustering was also a hard decision. After finding the clusters, we then decided to analyse not only the metric features, but also all those variables that were kept after the preprocessing, in order to have a wider knowledge of our clusters. All the analysis done is available in the notebook. We approached this project with a business eye, by trying to make sense out of the clustering from a marketing point of view, applying some knowledge we have from previous experiences. We think that the 3 marketing approaches may help the PVA to gather more money, and to approach future donors as well.

7. REFERENCES

1. https://www.datascience-pm.com/crisp-dm-2/
2. https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html#sklearn.impute.KNNImputer
3. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html
4. https://builtin.com/data-science/when-and-why-standardize-your-data
5. https://scikit-learn.org/stable/modules/preprocessing.html
6. https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#sphx-glr-auto-examples-preprocessing-plot-all-scaling-py
7. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
8. https://digitalagencynetwork.com/how-to-create-personas-for-marketing/
9. https://wellsr.com/python/creating-python-choropleth-maps-with-plotly/
10. https://plotly.com/python/county-choropleth/
11. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html