Clustering techniques applied in marketing strategy to reestablish donors from the PVA non-profit organization.

Débora Santos - 20200748

Rebeca Pinheiro - 20201096

Salim Bouaichi - 20200547

1 - INTRODUCTION

The application of clustering techniques has been used in several areas. One of the most used areas is Marketing. Understanding and segregating the customers to conduct more targeted marketing campaigns can be facilitated by applying these techniques.

PVA is a non-profit organization that provides programs and services for US veterans with spinal cord injuries or disease. With an in-house database of 13million donors, PVA is also one of the largest direct mail fundraisers in the United States of America. They have a particular group of lapsed donors who made their last donation to PVA 13 to 24 months ago. These donors made at least one prior donation to PVA and they represent an important group to PVA, since the longer someone goes without donating, the less likely they will be giving again.Therefore, recapturing these donors is a critical aspect of PVA's fundraising efforts.

The main objective of this work is to understand the behaviors of these donors, that can help identifying different segments of them, and present a marketing approach for each cluster. In order to achieve this goal, some clustering techniques were applied.

Another important objective of this study is to make a good pre-processing of the data. *It is well known that data quality is often a problem and has negative effects on the quality of the results* ***[1].***

The dataset for this work was provided by PVA and it included information related to 95412 donors with 476 attributes which represent donors gifts preferences, donors characteristics and donor's neighborhood data.

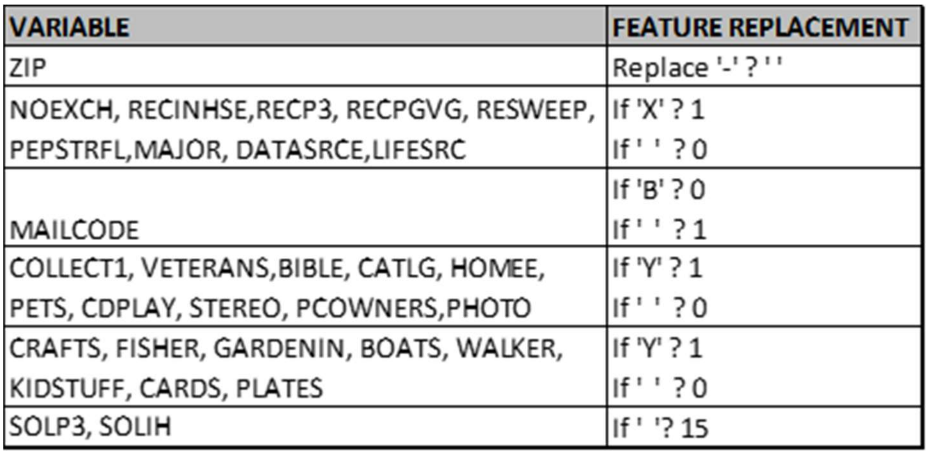
The main steps of this study were data preparation that consisted of data exploration, data cleaning, data transformation, then the implementation of cluster techniques and describing the clusters.

2 - METHODOLOGY

The first step consisted in doing an Exploratory Data Analysis. *'EDA' is all about understanding your data by employing, summarizing and visualizing techniques* ***[2]****.*

A primary overview in the data was done using the pandas profiling. Accessing it, we could understand some problems in the data: missing or empty values, high cardinality of variables, skewed numerical data and redundant data.

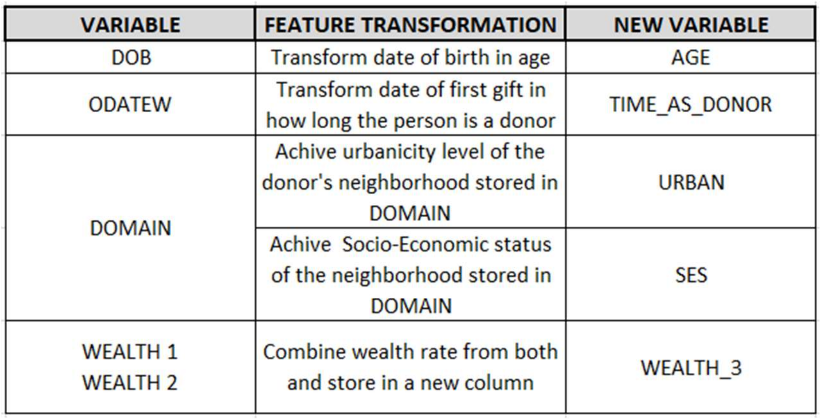
As we said, there were a lot of empty cells and some inconsistencies in the data. The table below shows some transformations done in order to have a better data:



VARIABLE	FEATURE REPLACEMENT
ZIP	Replace '-' ? ' '
NOEXCH, RECINHSE,RECP3, RECPGVG, RESWEEP, PEPSTRFL,MAJOR, DATASRCE,LIFESRC	If 'X' ? 1 If ' ' ? 0
MAILCODE	If 'B' ? 0 If ' ' ? 1
COLLECT1, VETERANS,BIBLE, CATLG, HOMEE, PETS, CDPLAY, STEREO, PCOWNERS,PHOTO	If 'Y' ? 1 If ' ' ? 0
CRAFTS, FISHER, GARDENIN, BOATS, WALKER, KIDSTUFF, CARDS, PLATES	If 'Y' ? 1 If ' ' ? 0
SOLP3, SOLIH	If ' '? 15

For the variable GENDER, we decided to replace the strange variables in gender for female gender, since the mode for that variable was female.

Some variables were transformed to be more useful. The table below show these variables:

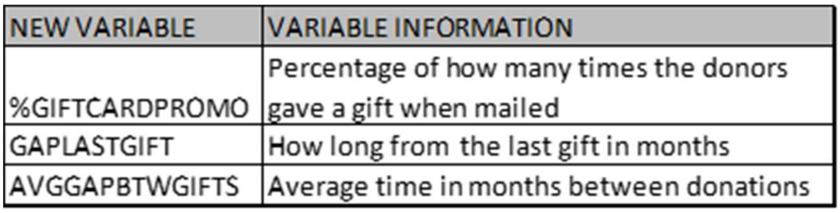


VARIABLE	FEATURE TRANSFORMATION	NEW VARIABLE
DOB	Transform date of birth in age	AGE
ODATEW	Transform date of first gift in how long the person is a donor	TIME_AS_DONOR
DOMAIN	Achive urbanicity level of the donor's neighborhood stored in DOMAIN	URBAN
	Achive Socio-Economic status of the neighborhood stored in DOMAIN	SES
WEALTH 1 WEALTH 2	Combine wealth rate from both and store in a new column	WEALTH_3

It is important to add some considerations regarding the new variables. The variable AGE and TIME_AS_DONOR used the variable ADATE_2 to calculate their values, since this variable stores when the last promotion mail was sent.

The variable WEALTH_3 was created combining the variables WEALTH_1 and WEALTH_2. The two existents variables inform the wealth rating and the combination happened as follow: If the rating was the same for both, WEALTH_3 will remain the same, if only one of them had the rating information, so WEALTH_3 will have store that one, if both of them had no information or both had different values, so WEALTH_3 will continue as missing value.

Three other variables were created to help in the cluster analysis:

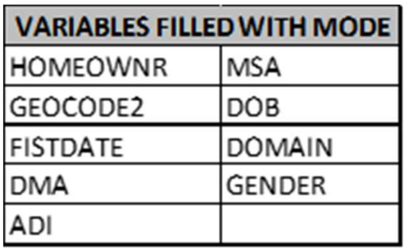


NEW VARIABLE	VARIABLE INFORMATION
%GIFTCARDPROMO	Percentage of how many times the donors gave a gift when mailed
GAPLASTGIFT	How long from the last gift in months
AVGGAPBTWGIFTS	Average time in months between donations

2.1 - MISSING VALUES

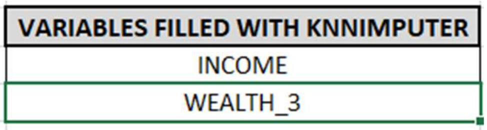
As said before, the quantity of variables with missing values were high. The list with all variables and their respective percentage of missing value is described in the table mentioned in the appendix.

The categorical variables were filled with the mode:



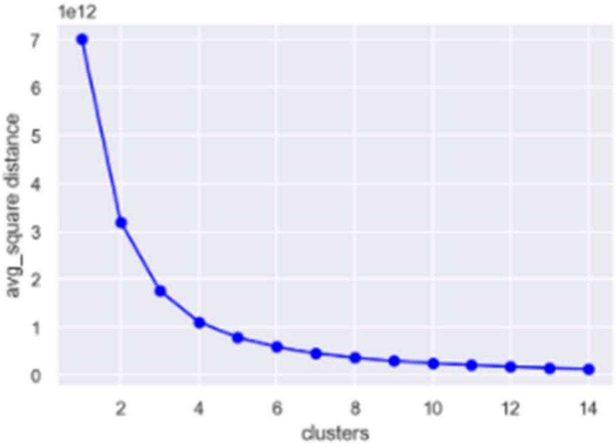
VARIABLES FILLED WITH MODE	
HOMEOWNR	MSA
GEOCODE2	DOB
FISTDATE	DOMAIN
DMA	GENDER
ADI	

Regarding the ordinal variables, we decided to fill them with KNNimputer, because they have a considerate number of missing values:



VARIABLES FILLED WITH KNNIMPUTER
INCOME
WEALTH_3

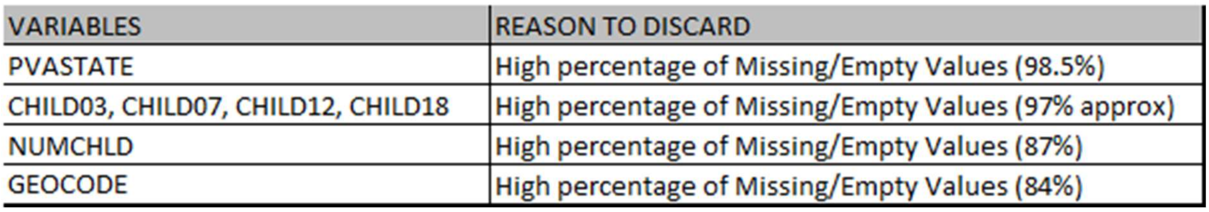
It is important to add that for those variables filled by KNNImputer, we used variables correlated with both to run the algorithm. And the number of neighbors (k=4) was defined by the elbow curve below.



2.2 - DROPPED VARIABLES

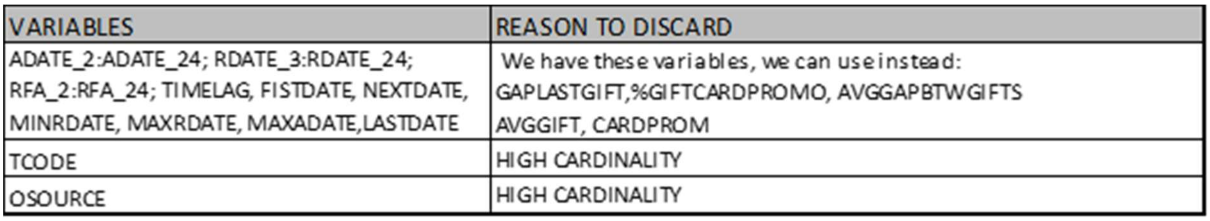
Another step of features transformation was to eliminate some attributes of the dataset. We used the following ways to do so:

1) High number of missing values:



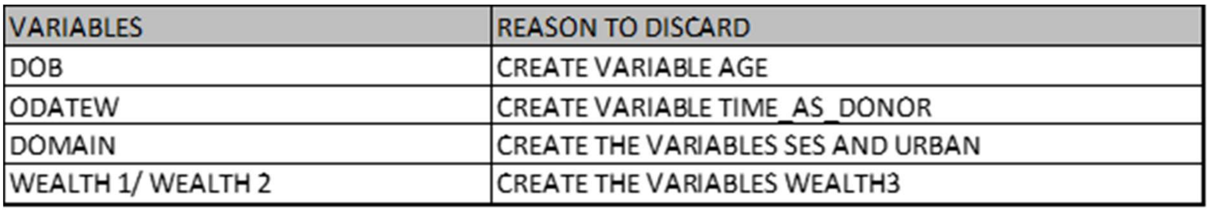
VARIABLES	REASON TO DISCARD
PVASTATE	High percentage of Missing/Empty Values (98.5%)
CHILD03, CHILD07, CHILD12, CHILD18	High percentage of Missing/Empty Values (97% approx)
NUMCHLD	High percentage of Missing/Empty Values (87%)
GEOCODE	High percentage of Missing/Empty Values (84%)

2) Irrelevance:



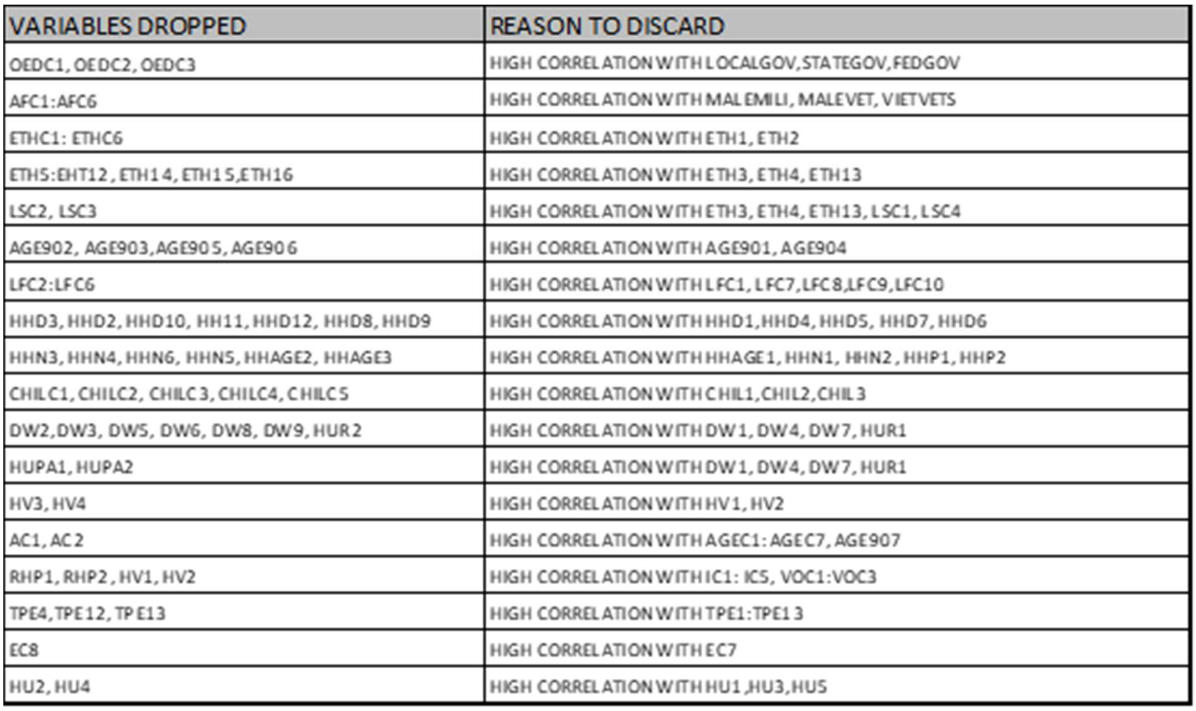
VARIABLES	REASON TO DISCARD
ADATE_2:ADATE_24; RDATE_3:RDATE_24; RFA_2:RFA_24; TIMELAG, FISTDATE, NEXTDATE, MINRDATE, MAXRDATE, MAXADATE,LASTDATE	We have these variables, we can use instead: GAPLASTGIFT,%GIFTCARDPROMO, AVGGAPBTWGIFTS AVGGIFT, CARDPROM
TCODE	HIGH CARDINALITY
OSOURCE	HIGH CARDINALITY

3) Create new variables



VARIABLES	REASON TO DISCARD
DOB	CREATE VARIABLE AGE
ODATEW	CREATE VARIABLE TIME_AS_DONOR
DOMAIN	CREATE THE VARIABLES SES AND URBAN
WEALTH 1/ WEALTH 2	CREATE THE VARIABLES WEALTH3

4) High correlation between variables



VARIABLES DROPPED	REASON TO DISCARD
OEDC1, OEDC2, OEDC3	HIGH CORRELATION WITH LOCALGOV,STATEGOV,FEDGOV
AFC1:AFC6	HIGH CORRELATION WITH MALEMILI, MALEVET, VIETVETS
ETHC1: ETHC6	HIGH CORRELATION WITH ETH1, ETH2
ETH5:EHT12, ETH14, ETH15,ETH16	HIGH CORRELATION WITH ETH3, ETH4, ETH13
LSC2, LSC3	HIGH CORRELATION WITH ETH3, ETH4, ETH13, LSC1, LSC4
AGE902, AGE903,AGE905, AGE906	HIGH CORRELATION WITH AGE901, AGE904
LFC2:LFC6	HIGH CORRELATION WITH LFC1, LFC7,LFC8,LFC9,LFC10
HHD3, HHD2, HHD10, HH11, HHD12, HHD8, HHD9	HIGH CORRELATION WITH HHD1,HHD4, HHD5, HHD7, HHD6
HHN3, HHN4, HHN6, HHN5, HHAGE2, HHAGE3	HIGH CORRELATION WITH HHAGE1, HHN1, HHN2, HHP1, HHP2
CHILC1, CHILC2, CHILC3, CHILC4, CHILC5	HIGH CORRELATION WITH CHIL1,CHIL2,CHIL3
DW2,DW3, DW5, DW6, DW8, DW9, HUR2	HIGH CORRELATION WITH DW1, DW4, DW7, HUR1
HUPA1, HUPA2	HIGH CORRELATION WITH DW1, DW4, DW7, HUR1
HV3, HV4	HIGH CORRELATION WITH HV1, HV2
AC1, AC2	HIGH CORRELATION WITH AGEC1: AGEC7, AGE907
RHP1, RHP2, HV1, HV2	HIGH CORRELATION WITH IC1: IC5, VOC1:VOC3
TPE4,TPE12, TPE13	HIGH CORRELATION WITH TPE1:TPE13
EC8	HIGH CORRELATION WITH EC7
HU2, HU4	HIGH CORRELATION WITH HU1 ,HU3,HU5

For the next steps of feature transformation, we divided the dataset in 4 parts:

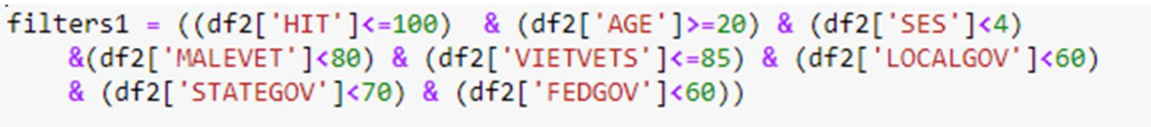
- Donors characteristics.
- Donors Preferences
- Promotions data
- Donors neighborhood (census)

2.3 - OUTLIERS REMOVAL

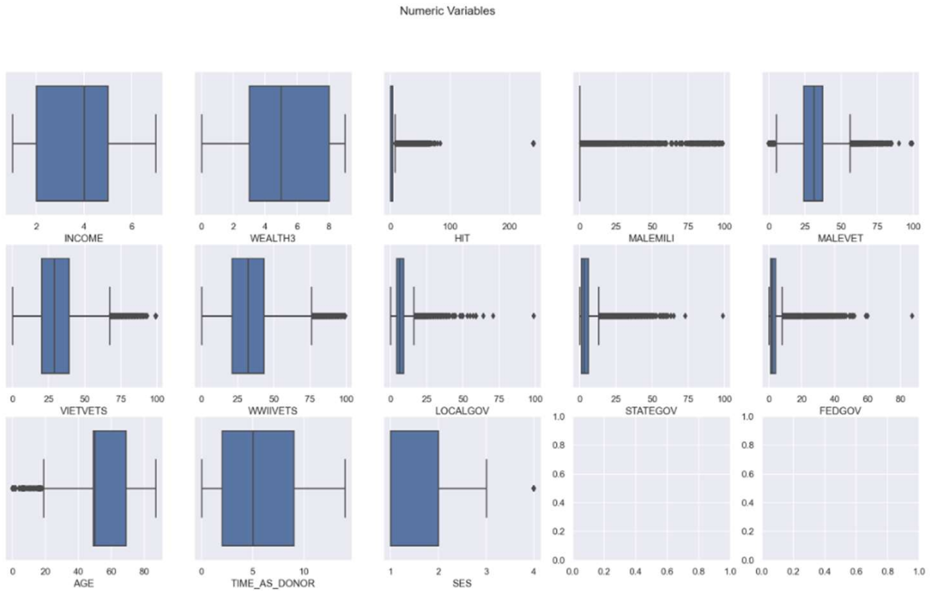
We detected outliers in donors' characteristics and promotions data subset.

For the donor's characteristics metrics features the approach that we use to identify the outliers was boxplot and the result can be find below:

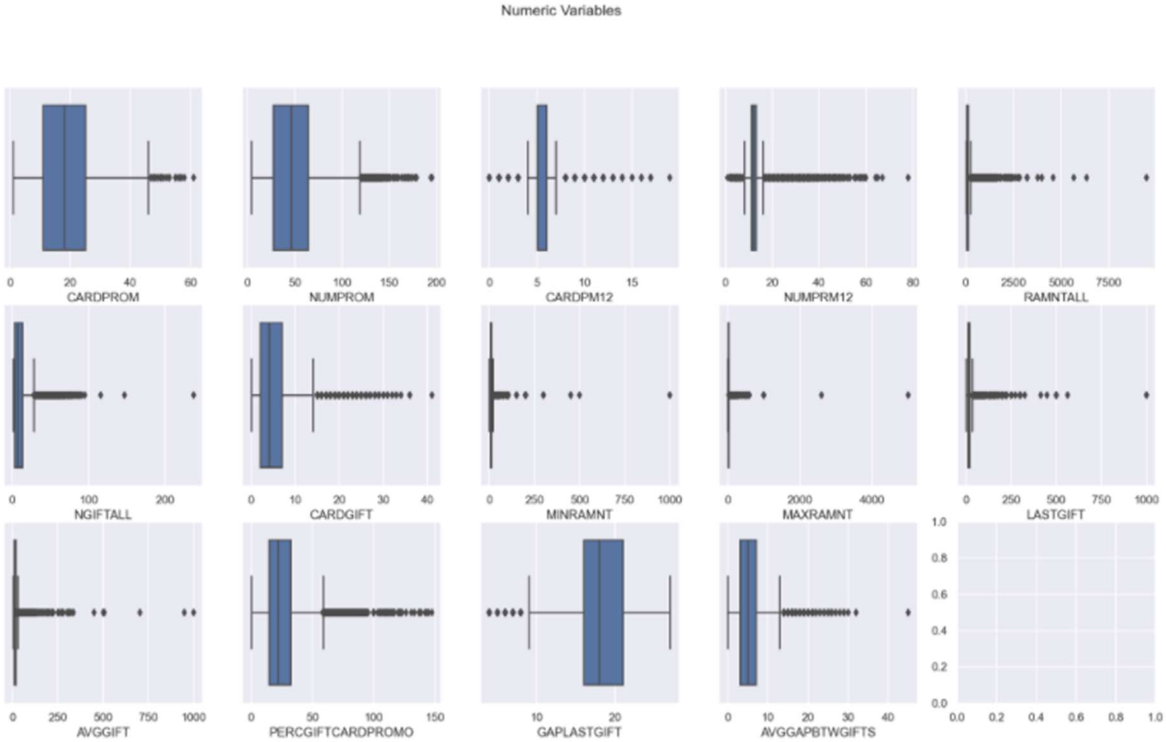
For some features we could apply the IQR method, but we decided to apply manual filters, as follow:



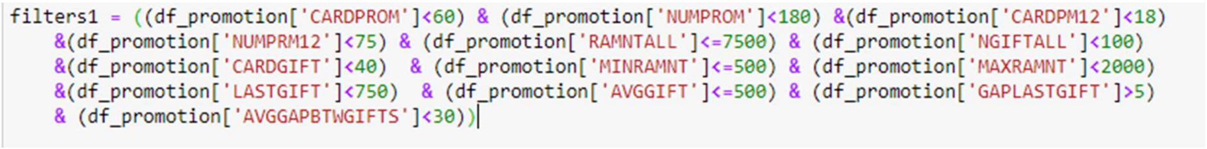
```
filters1 = ((df2['HIT']<=100)  & (df2['AGE']>=20) & (df2['SES']<4)
    &(df2['MALEVET']<80) & (df2['VIETVETS']<=85) & (df2['LOCALGOV']<60)
    & (df2['STATEGOV']<70) & (df2['FEDGOV']<60))
```



The promotion data, the outlier's result found was:

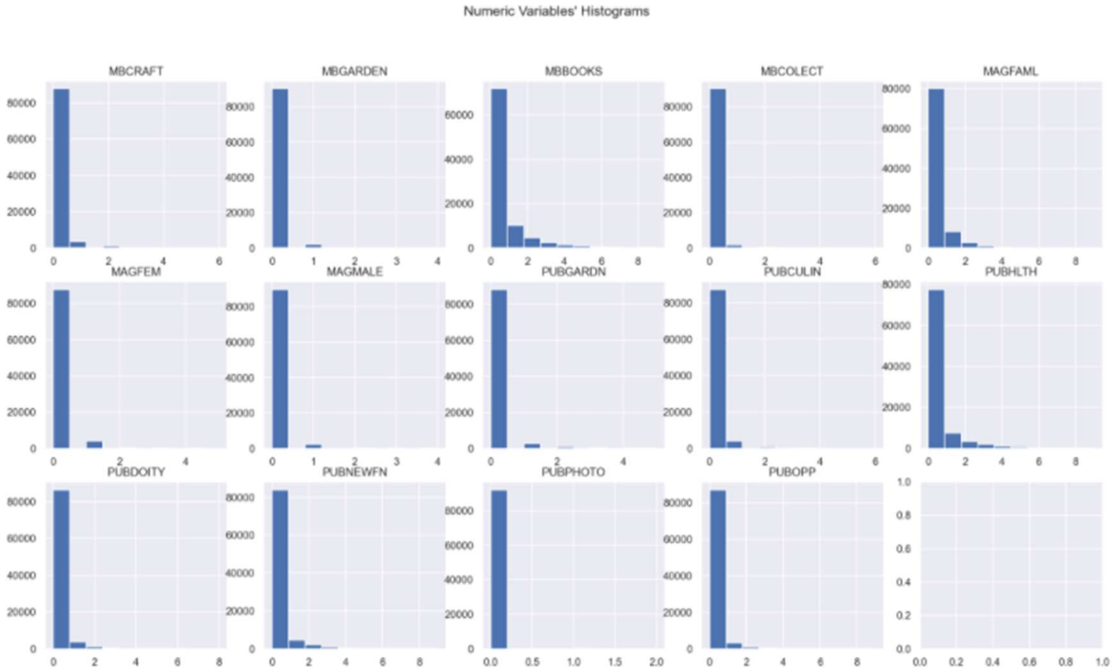


And the filters applied for promotion data were:


```
filters1 = ((df_promotion['CARDPROM']<60) & (df_promotion['NUMPROM']<180) &(df_promotion['CARDPM12']<18)
    &(df_promotion['NUMPRM12']<75) & (df_promotion['RAMNTALL']<=7500) & (df_promotion['NGIFTALL']<100)
    &(df_promotion['CARDGIFT']<40)  & (df_promotion['MINRAMNT']<=500) & (df_promotion['MAXRAMNT']<2000)
    &(df_promotion['LASTGIFT']<750)  & (df_promotion['AVGGIFT']<=500) & (df_promotion['GAPLASTGIFT']>5)
    & (df_promotion['AVGGAPBTWGIFTS']<30))
```

After applying the filters for both subsets, the data remaining was 95,2%.

On the preferences part, even if there were some skewed distributions, we understand that there were no outliers.



On the census part, as we had a lot of columns, we applied DBSCAN to detect outliers. The result will be shown in the next steps.

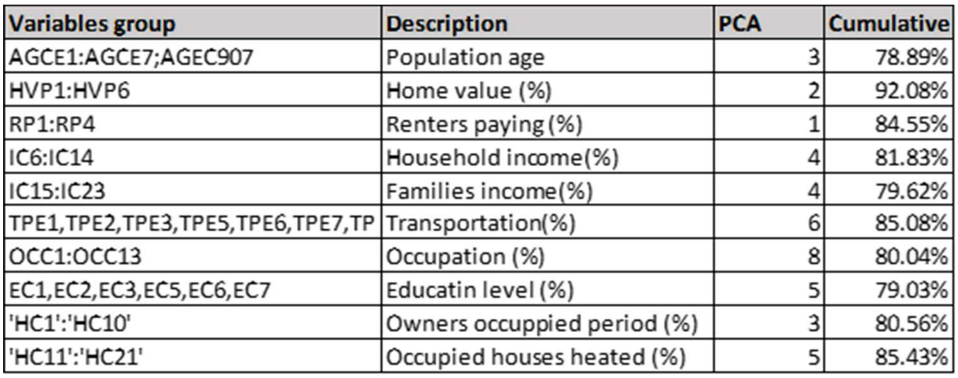
2.4 -DATA REDUCTION

On the census part, we decided to apply PCA in those variables that don't have high correlation with others in order to reduce the number of redundant variables.

A previous step of PCA is normalizing the data and for that we used the Standard Scaler.

We applied PCA in 122 variables in small subsets, according to the relationship between them. The final result was 31 original features and 41 Principal components. The criteria to select the principal components was the cumulative variance higher than 80%.

For the employment sector and ancestry subsets we believe that we could not use the principal components because as each feature has its own meaning and they are completely different from each other, even though the PCA shows some similarity, according to the features signification in practice it couldn't represent the reality, so we kept all the original features. The table below describes the features we grouped and their description:



Variables group	Description	PCA	Cumulative
AGCE1:AGCE7;AGEC907	Population age	3	78.89%
HVP1:HVP6	Home value (%)	2	92.08%
RP1:RP4	Renters paying (%)	1	84.55%
IC6:IC14	Household income(%)	4	81.83%
IC15:IC23	Families income(%)	4	79.62%
TPE1,TPE2,TPE3,TPE5,TPE6,TPE7,TP	Transportation(%)	6	85.08%
OCC1:OCC13	Occupation (%)	8	80.04%
EC1,EC2,EC3,EC5,EC6,EC7	Educatin level (%)	5	79.03%
'HC1':'HC10'	Owners occuppied period (%)	3	80.56%
'HC11':'HC21'	Occupied houses heated (%)	5	85.43%

2.5 - DATA NORMALIZATION

In order to prepare the data for the clustering, the next step of feature engineering was to normalize the data. We transformed the categorical values in binary values applying One hot encoding. We also applied Standard scale in numeric features.

Many machine learning algorithms perform better when numerical input variables are scaled to a standard range.This includes algorithms that use a weighted sum of the input, like linear regression, and algorithms that use distance measures, like k-nearest neighbors.[3]

The Standard scale normalizes the data with zero mean and one of standard deviation. The algorithms assume the features have the same variance. *If a feature has a variance that is orders of magnitude larger that others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected. [4]*

At the end, our data has 90.828 observations with 283 attributes.

3 - RESULTS

3.1 - CLUSTER SOLUTION

Clustering is an unsupervised learning problem, that means the algorithms will classify the observation into groups without previous knowledge of the class labels. The groups are generated according to the similarity of the observations.

There are some cluster techniques: Hierarchical algorithms, partitional algorithms, density algorithms and others.

Our problem is about segregating donors according to their similarity. So we need to use algorithms that deal better with spherical-shaped clusters. Because of that, we don't choose to use DBSCAN to perform the clustering.

The meanshift algorithm and the hierarchical algorithm aren't good with big datas. We tried to apply it, but it took too much time and we couldn't run the R2, in order to compare their performance.

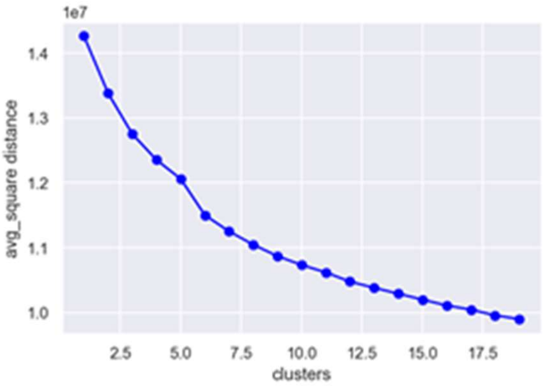
The next two tentatives were SOM and Kmeans. SOM is a neural network clustering. We decided not to follow it, because it is too complex to interpret.

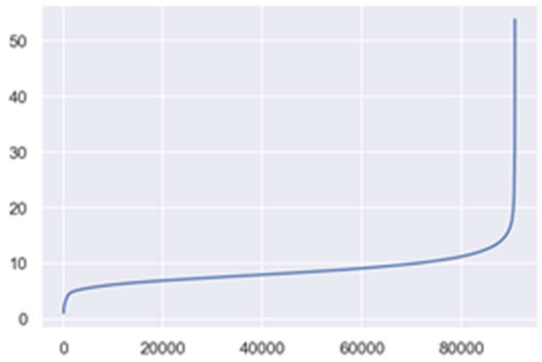
So, our final decision was to apply K means.

3.2 - OUTLIER

One problem to most of the clustering techniques is dealing with outliers. We already removed some of them, but we still had the census part that we didn't apply outlier removal. Also, we still had some skewed observations. In order to get better performance in our cluster algorithm, we applied DBSCAN to segregate the outliers.

We first found the neighbors number by elbow curve which gave us 6. With that result was possible to find the eps number by the k-distance graph. So, we set the eps as 15, number of samples of 20 to apply the DBSCAN.





The DBSCAN detected 1249 outliers in our data. After we had the final cluster solution, we estimated the cluster of these observations.

3.3 - FEATURE SELECTION

Our dataset has 283 attributes, with 157 metric features, 41 principal components and 85 categorical features. To run the cluster, we have to use only numerical data.

We decided to use the features more directly related to the donors: Donors demographic information and historical data of donations. The data about donors preferences is not normally distributed and the cluster algorithm doesn't perform well with this kind of distribution. The census data has a lot of information that doesn't influence a marketing approach, so, instead of giving a better view of the donor's behaviours, adding the census data in the cluster didn't help in the donors segregation.

We selected the follow variables:

1) Promotion variables

- Number of promotions received;
- Number of card promotions received;
- Number of promotions received in the last 12 months
- Number of card promotions received in the last 12 months;
- Total amount of donations;
- Number of donations;
- Number of donations related to a promotion;
- Amount of smallest donation;
- Amount of highest donation;
- Dollar amount of most recent donation;
- Average amount of donations
- Frequency rate of response to a promotion;
- Timing difference from "now" to the last donation;
- Average timing difference between donations.

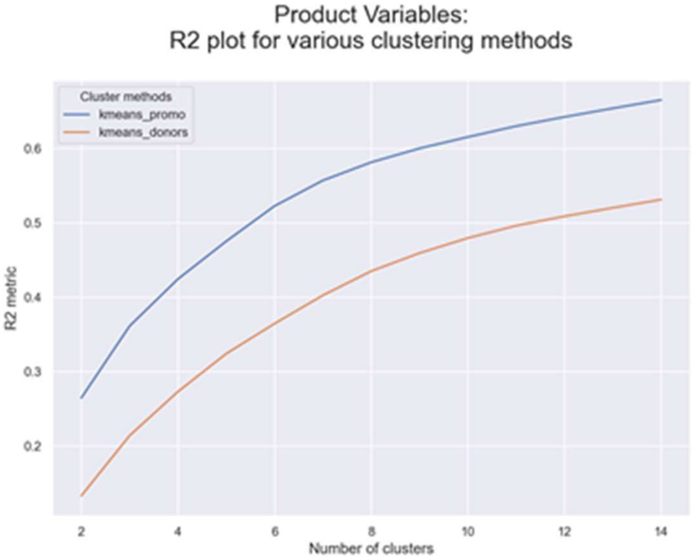
2) Donor characteristics

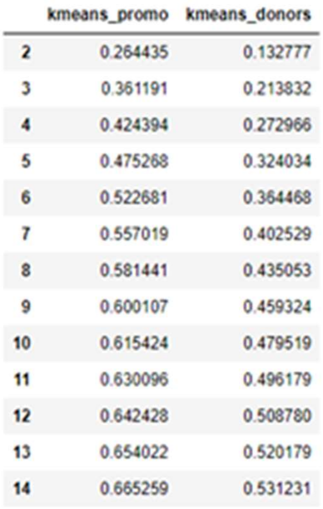
- Income
- Wealth rating;
- Males active in the Military
- Males Veterans
- Vietnam Vets
- WWII Vets
- Employed by Local Gov
- Employed by State Gov
- Employed by Fed Gov
- Age
- Socio-Economic status of the neighborhood

3.4 - KMEANS

K means is the most popular clustering algorithm. It's efficient in terms of time, easy to understand and easy to implement. But its major problem is to define the number of clusters and its very sensitive to the initial position of seeds. *A major problem with this algorithm is that it is sensitive to the selection of the initial partition and may converge to a local minimum of the criterion function value if the initial partition is not properly chosen.[5]*

To choose the number of clusters, we applied the R2 to compare the performance.

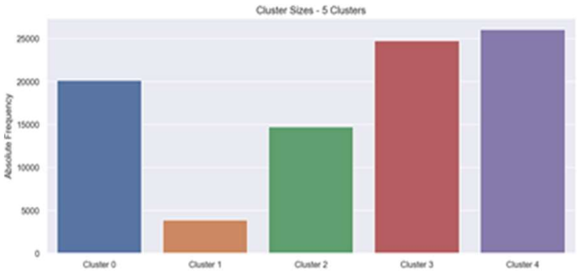


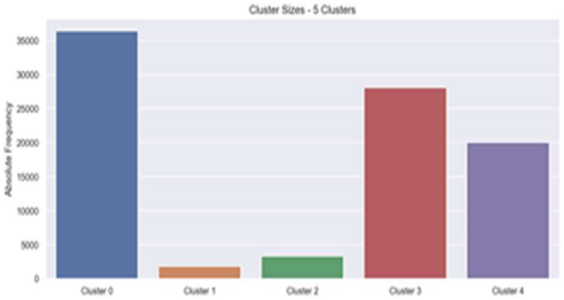


	kmeans_promo	kmeans_donors
2	0.264435	0.132777
3	0.361191	0.213832
4	0.424394	0.272966
5	0.475268	0.324034
6	0.522681	0.364468
7	0.557019	0.402529
8	0.581441	0.435053
9	0.600107	0.459324
10	0.615424	0.479519
11	0.630096	0.496179
12	0.642428	0.508780
13	0.654022	0.520179
14	0.665259	0.531231

We defined 5 clusters to both subsets. We set the ‘k-means++’ as the initialization method (init parameter) with 20 number of iterations to choose the best centroid in terms of inertia (n_init parameter).

In the graphics below, we have the number of observations by cluster.



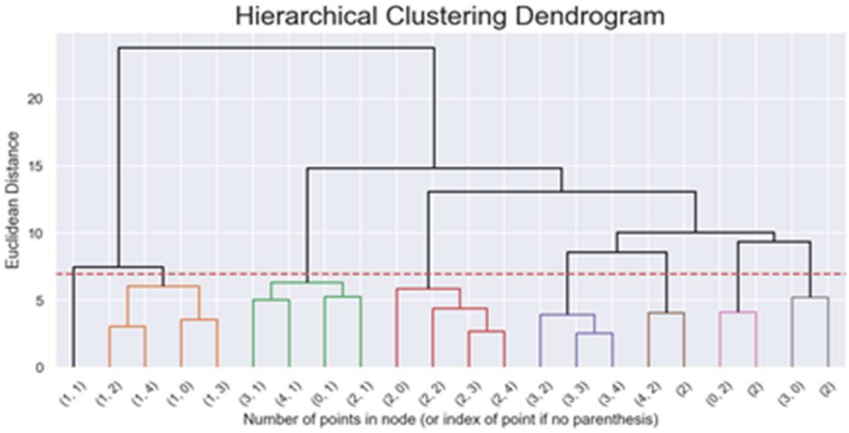


3.5 - MERGE CLUSTERS

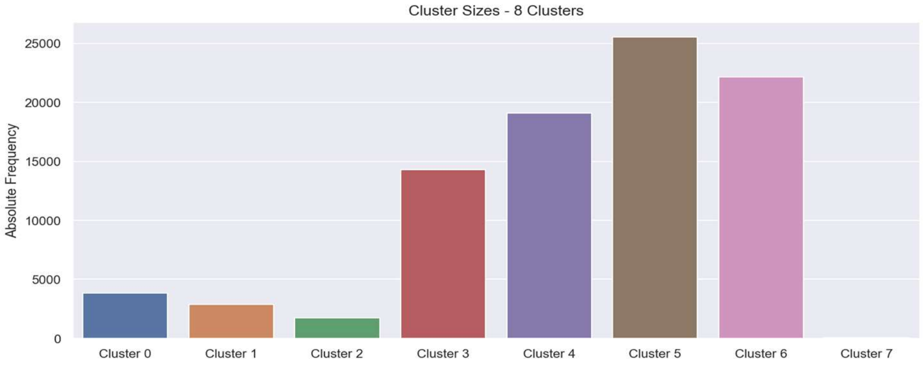
We applied the hierarchical cluster to perform the merge between the both clusters generated by K-means. After we had the centroids , we applied the agglomerative clustering algorithm.

The parameters set were: linkage: ward, euclidean distance and threshold 7.

As a result, we got this dendrogram below.

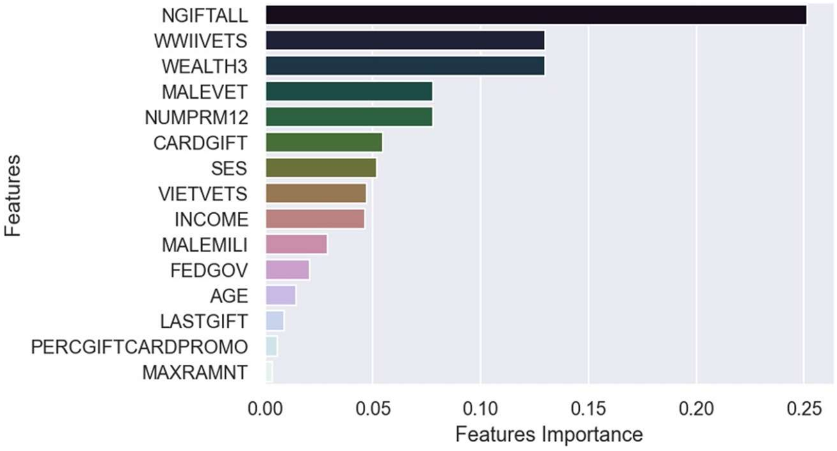


As a final result, we had 8 clusters. The graphic below shows the distribution of observations by clusters.



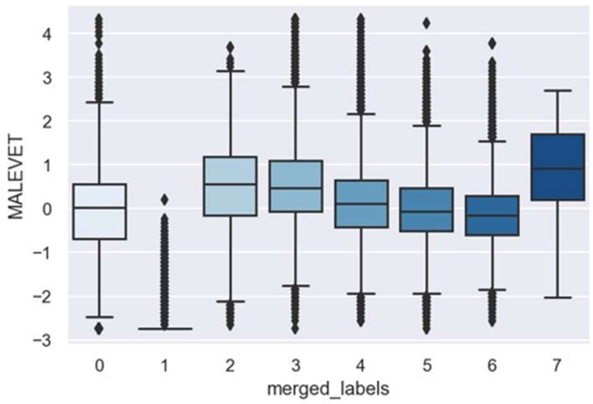
In order to test our solution, we applied a decision tree classifier after splitting the data into training and test. It was able to predict 88,8% of the donors correctly

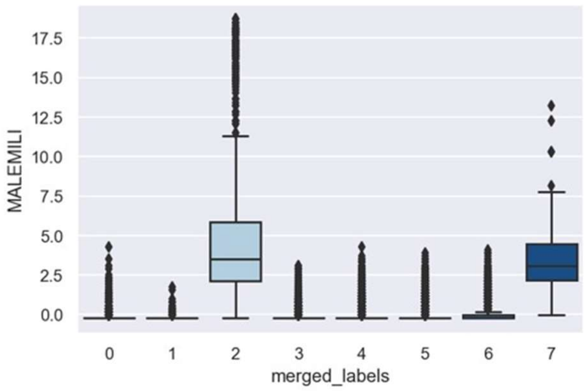
We also got the feature importance of each variable in predicting the cluster. The graphic below shows the 15 more important features:

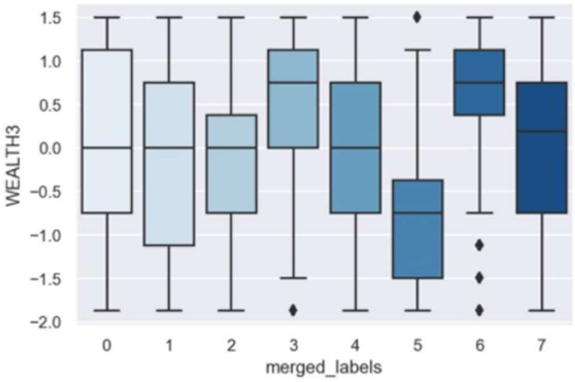


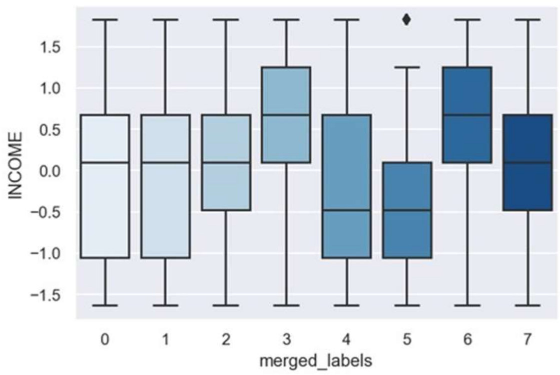
3.6 - DESCRIBING CLUSTERS

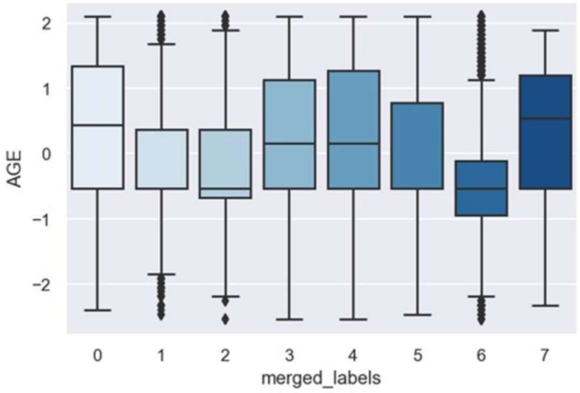
We plot some boxplots to see how the clusters were distributed.

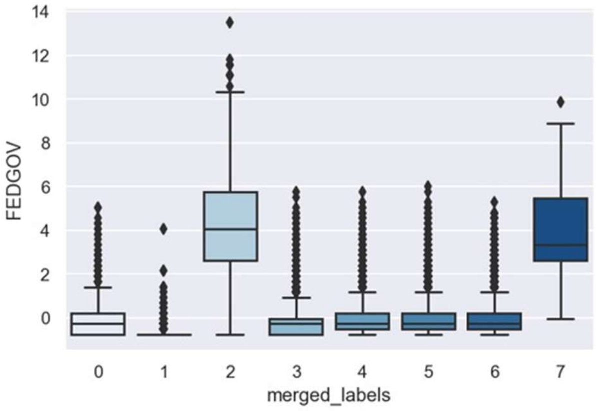


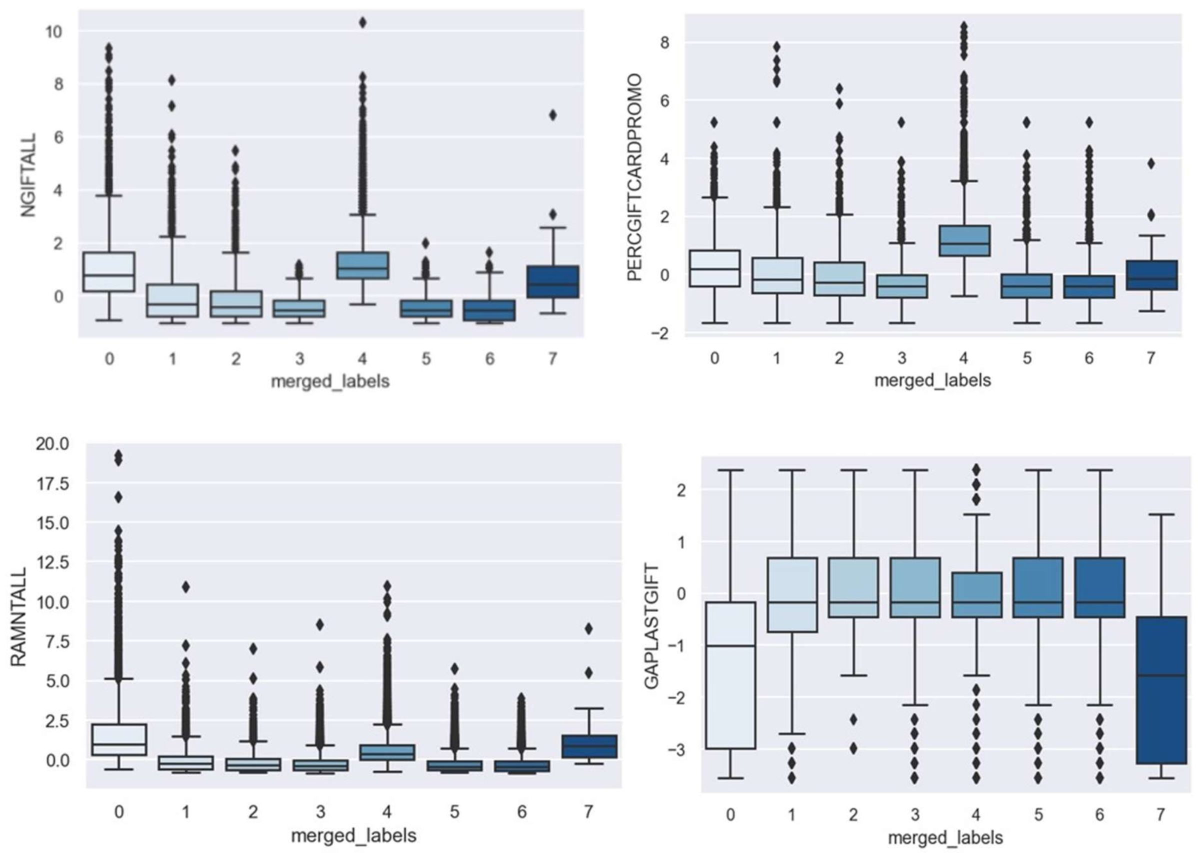












As we can see, the clusters numbers 7 and 2 are the ones with high number of militars.

The cluster number 7 has higher frequency of donations (NGIFTALL) and this group also has a high income and wealth rating. To that cluster, our approach would be inviting them to become a frequent donor membership, which means that they will have a personal account that in certain periods (can be per month, trimester or semester) the nonprofit will discount from their bank account the donation.

The cluster number 02, they have a good number of militaries that participated in the Vietnam war. We suggested to contact them about the cause and try to bring them back and remember how it's important.

To cluster number 03, we suggested a similar approach of the cluster number 2, but with foccus in the Second war. This group has a good number of militaries that participated in this war. They also have income and wealth higher than averaged when compared with the other groups.

The cluster number 4 has a high response rate (PERCGIFTCARD) and low income. Instead of inviting to become a membership, we decided to contact them regularly and also share a link so they can invite others to know the organization and per each new donor that gives a gift by their link, the donors from cluster number four will receive a small gift based on their preference.

The cluster number 6 is the younger group, and they have higher income and wealth rates compared with the others. Considering their age, our approach would be inviting them to be more involved in the organization’s activities. So, they can help on spreading the foundation with volunteer work as an example. It is proven that those who work for a nonprofit organization also donate frequently and invite more people to know their work.

The cluster number 0 is the group with the highest amount of donations, and they have a high frequency of donation. For then, we would suggest the same approach of cluster number 7. They can become a frequent donor membership.

In cluster number 1, there are no militaries. In this case, we suggest sending emails offering some products. They could investigate a little bit more about this group to know which type of product they like more, or their hobbies (sports, books). The cluster number 5, we couldn’t detect any interesting information to suggest a different marketing approach. So, our suggestion is keep sending promotions and investigate better these donors.

4 - CONCLUSION

As stated in the introduction, the main objective of this project was to provide marketing approaches in order to reestablish the donors from the nonprofit organization.

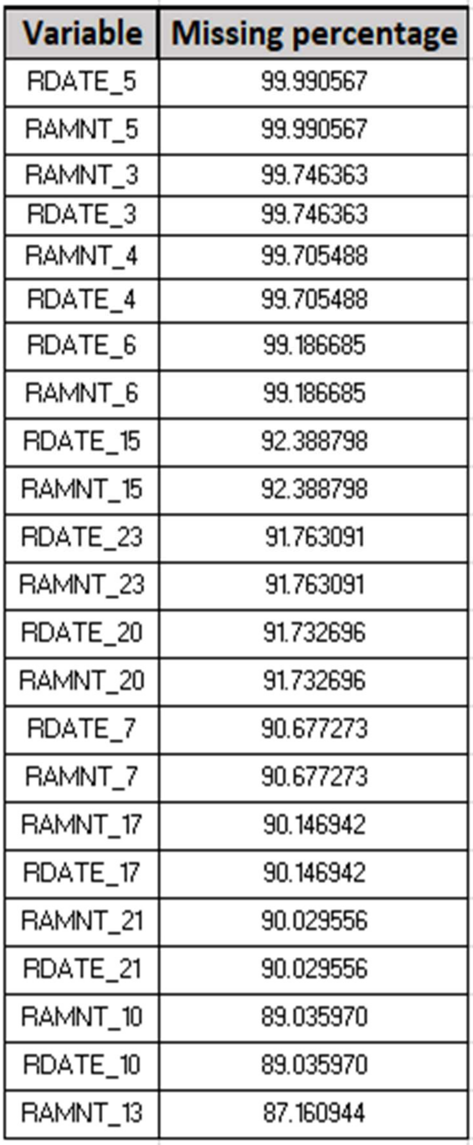
In order to be able to have a better overview of the donors, we decided to cluster them based on demographic information and historical data of donations.

Our main steps before clustering were analyzing the data, transforming the data using techniques such as correlation matrix, PCA and dbscan to detect outliers.

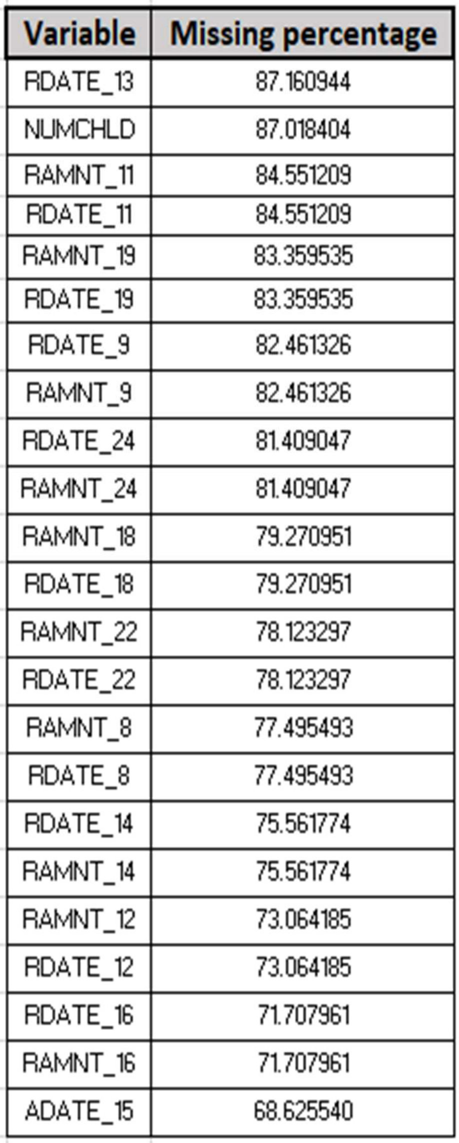
Finally, for clustering, we tried different cluster techniques, but the one that best fit to our dataset was kmeans. The final result gave us 8 clusters, and as a kmeans property the donors were grouped by their similarities.

Applying the algorithm and grouping the donors, we finally developed different marketing approaches for each cluster, considering their principal specificities.

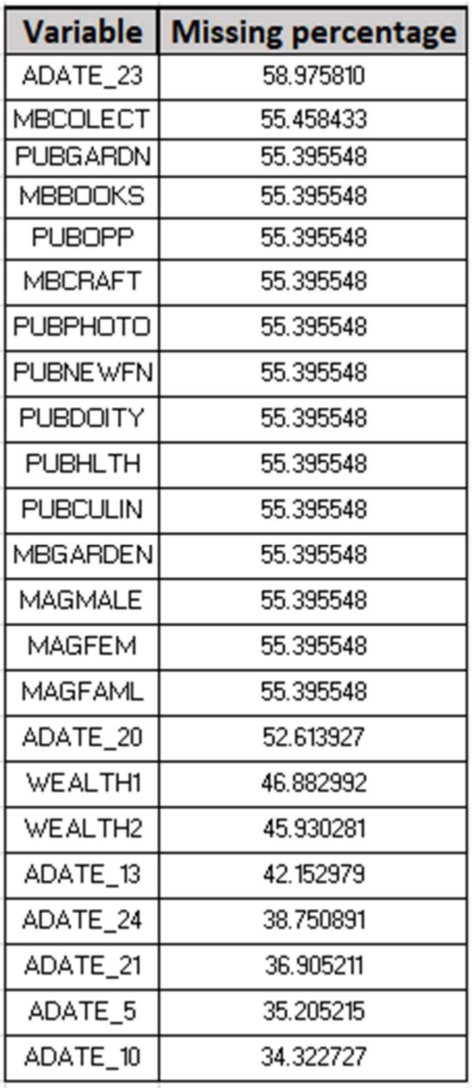
5 - APPENDIX



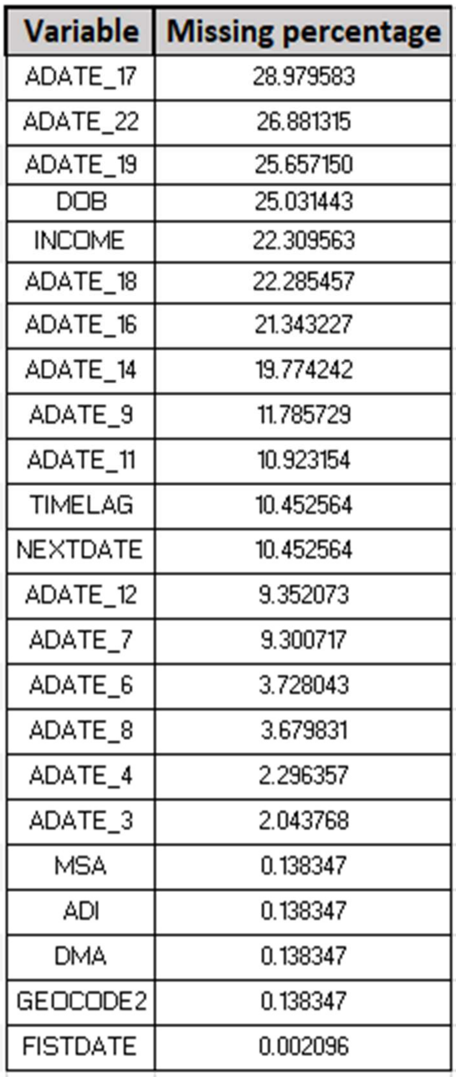
Variable	Missing percentage
RDATE_5	99.990567
RAMNT_5	99.990567
RAMNT_3	99.746363
RDATE_3	99.746363
RAMNT_4	99.705488
RDATE_4	99.705488
RDATE_6	99.186685
RAMNT_6	99.186685
RDATE_15	92.388798
RAMNT_15	92.388798
RDATE_23	91.763091
RAMNT_23	91.763091
RDATE_20	91.732696
RAMNT_20	91.732696
RDATE_7	90.677273
RAMNT_7	90.677273
RAMNT_17	90.146942
RDATE_17	90.146942
RAMNT_21	90.029556
RDATE_21	90.029556
RAMNT_10	89.035970
RDATE_10	89.035970
RAMNT_13	87.160944



Variable	Missing percentage
RDATE_13	87.160944
NUMCHLD	87.018404
RAMNT_11	84.551209
RDATE_11	84.551209
RAMNT_19	83.359535
RDATE_19	83.359535
RDATE_9	82.461326
RAMNT_9	82.461326
RDATE_24	81.409047
RAMNT_24	81.409047
RAMNT_18	79.270951
RDATE_18	79.270951
RAMNT_22	78.123297
RDATE_22	78.123297
RAMNT_8	77.495493
RDATE_8	77.495493
RDATE_14	75.561774
RAMNT_14	75.561774
RAMNT_12	73.064185
RDATE_12	73.064185
RDATE_16	71.707961
RAMNT_16	71.707961
ADATE_15	68.625540



Variable	Missing percentage
ADATE_23	58.975810
MBCOLECT	55.458433
PUBGARDN	55.395548
MBBOOKS	55.395548
PUBOPP	55.395548
MBCRAFT	55.395548
PUBPHOTO	55.395548
PUBNEWFN	55.395548
PUBDOITY	55.395548
PUBHLTH	55.395548
PUBCULIN	55.395548
MBGARDEN	55.395548
MAGMALE	55.395548
MAGFEM	55.395548
MAGFAML	55.395548
ADATE_20	52.613927
WEALTH1	46.882992
WEALTH2	45.930281
ADATE_13	42.152979
ADATE_24	38.750891
ADATE_21	36.905211
ADATE_5	35.205215
ADATE_10	34.322727



Variable	Missing percentage
ADATE_17	28.979583
ADATE_22	26.881315
ADATE_19	25.657150
DOB	25.031443
INCOME	22.309563
ADATE_18	22.285457
ADATE_16	21.343227
ADATE_14	19.774242
ADATE_9	11.785729
ADATE_11	10.923154
TIMELAG	10.452564
NEXTDATE	10.452564
ADATE_12	9.352073
ADATE_7	9.300717
ADATE_6	3.728043
ADATE_8	3.679831
ADATE_4	2.296357
ADATE_3	2.043768
MSA	0.138347
ADI	0.138347
DMA	0.138347
GEOCODE2	0.138347
FISTDATE	0.002096

6 - REFERENCES

[1] Manohar Swamynathan. Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python

[2] Fernando Bação, Victor Lobo, Marco Painho. Clustering census data: comparing the performance of self-organising maps and k-means algorithms

[3] https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/. Accessed in 04/01/2021

[4] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. Accessed in 04/01/2021

[5] A.K. Jain, M.N. Murty, P.J. Flynn. Data Clustering: A Review.