Data Mining
MSc in Data Science and Business Analytics
NOVA IMS

Paralyzed Veterans of America

Omar JARIR m20201378
Yu Song m20200572
Natalia Castañeda m20200575

Cluster Analysis Paralyzed Veterans of AmericaLapsed Donors

I. INTRODUCTION

Paralyzed Veterans of America (PVA) is a non-profit organization that provides programs and services for US veterans with spinal cord injuries or disease.

Considering a sample from PVA's most recent fundraising which contains 95.412 Lapsed donors, an important group of individuals who made their last donation 13 to 24 months ago, our team has been asked to develop a Customer Segmentation to better understand how these donors behave and identify smaller groups in which these donors share similar qualities, in order to create meaningful marketing and communication strategies to recapture and build trusty and strong relationships with lapsing donors.

Once the data base was received, the team determined that different clustering methods should be tested under three main conditions: professional experience (know-how) with theoretical concepts, the cleanest dataset possible and statistical testing for cluster tendency.

II. BACKGROUND

Preprocessing methods and cluster algorithms discussed and applied during practical classes were tried in search of the best cluster solution. However, our group decided to search statistical tools for assessing clustering tendency before try any algorithm.

The first tool was the algorithm of the visual assessment of cluster tendency VAT, which computes and orders the dissimilarity matrix between the objects in the data set and detects the clustering tendency in a visual form by counting the number of square shaped dark blocks along the diagonal.

The second tool was the Hopkins Statistic which assesses the clustering tendency of a data set by measuring the probability that a given data set is generated by a uniform data distribution. In other words, it tests the spatial randomness of the data[1]. The computation of the Hopkins Statistic follows the next step-by-step:

1. Sample uniformly n points $(p_1, \dots, p_n)$ from D.
2. For each point $p_i \in D$, find the nearest neighbor p_j; then compute the distance betweenp_i and p_j and denote it as $x_i = dist(p_i, p_j)$.
3. Generate a simulated dataset $(random_D)$ drawn from a random uniform distribution with n points $(q_1, \dots, q_n)$ and the same variation as the original real dataset D.
4. For each point $q_i \in random_D$, find its nearest neighbor $q_j \in D$; then compute the distance between q_i and $q_j$and denote it $y_i = dist(q_i, q_j)$.

Hopkins statistic (H) is calculated as the mean nearest neighbor distances in the random dataset divided by the sum of the mean nearest neighbor distances in the real and across the simulated dataset.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

A value of H about 0.5 means that $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i$ are close to each other, and thus the data D is uniformly distributed.

The null and alternative hypothesis are defined as follow:
Null hypothesis: The dataset D is uniformly distributed (i.e., no meaningful clusters), Alternative hypothesis: the dataset D is not uniformly distributed (i.e., contains meaningful clusters).

If the value of the Hopkins statistic is close to zero, then we can reject the null hypothesis and conclude that the dataset D is significantly a clusterable data.

III. METHODOLOGY

Design

Our team adopted a usual approach for defining the most accurate method to cluster, following the steps of feature selection and engineering, visualization, data preprocessing, clustering and characterization.

Instruments

General python libraries were used for importing and exploring data, others were imported for specific purposes such as preprocessing, statistical analysis, clustering and visualization (Table 1).

General	**Visualization**	**Preprocessing, Clustering, Metrics**	**Statistical Analysis**
Pandas	Matplot	Sklearn	pyclustertend
Numpy	Seaborn		
Math	T-sne		
Datetime	Sompy		
Clone, os, re			

Table 1: Libraries imported in Python

Procedure

In the preprocessing stage, numeric and categorical variables were identified, visualized and checked for missing values, data type and formatting. The missing values were filled using k-nearest-neighbors imputer method for numeric variables and mode for categorical ones. Then, outliers were removed applying the interquartile method with 90th and 10th quantiles which allowed us to keep 96% of the data. The numeric variables were normalized and the categorical ones encoded.

The algorithm of the visual assessment (VAT) was applied to the data set to detect the clustering tendency. Afterward, the Hopkins Statistic was calculated in order to identify if the data set is uniformly distributed.

We decided using standard individual algorithms as K-means and Hierarchical, and also compute a combination of them in search for better results. To combine these two algorithms, it was necessary to first get a random large number of clusters using K-means to then use its centroids to compute hierarchical cluster. This process was done in two scenarios, first one, using all variables, and second one, dividing them by demographic and engagement perspectives.

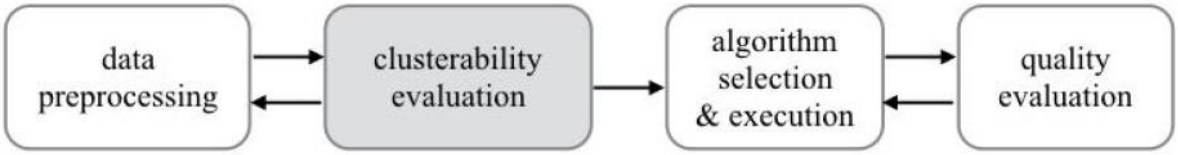


Figure 1: Cluster process approach

After an evaluation of the quality of the results and due to the high dimensionality of the data base, the above process was repeated several times using a different set of variables. Finally, our team selected the following set, based on characteristics such as demographics, interests, size of gift, giving frequency and engagement level.

Criteria	Variable Name	Description
Size of Gift	LOG_AVGGIFT	Logarithm of average dollar amount of gifts to date
	LASTGIFT	Dollar amount of most recent gift
	MINRAMNT	Dollar amount of smallest gift to date
	LOG_RAMNTALL	Logarithm of dollar amount of lifetime gifts to date
Giving Frequency	GIFT_FRECIENCY (RFA_2F)	Number of gifts based on the period of recency (13-24 months ago)
	YEARS_FIRST_GIFT	Number of years from first gift to date
	NGIFTALL	Number of lifetime gifts to date
	TIMELAG	Number of months between first and second gift
Engagement Level	HIT	Indicates total number of known times the donor has responded to a mail order offer other than PVA's
	CARDGIFT	Number of lifetime gifts to card promotions to date
Demographics	STATE	US States
	LOG_AGE	Logarithm of actual age

	INCOME	Household Income 1.0 =<$15,000 2.0 =$15,000 - $24,999 3.0 =$25,000 - 34,999 4.0 = $35,000 - $49,999 5.0 = $50,000 - $74,999 6.0 = $75,000 - $99,999 7.0 = $100,000 - $124,999
	GENDER	M = Male F = Female U = Unknown J = Joint Account, unknown gender C, A and blank spaces were grouped as unknown
	URBANICITY_LEVEL	1st byte form DOMAIN U=Urban C=City S=Suburban T=Town R=Rural Blank spaces were group as unknown

Table 2: Variable selection

IV. RESULTS

1. Data Preprocessing

After carefully reviewing the detailed document which explained each variable, the most relevant variables were selected. The variables DOB (Date of birth) and ODATEDW (Date of donor's first gift) were used to create two new variables: AGE and YEARS_FIRST_GIFT (Table 2).

Also, the logarithm of the numeric variables: AGE, RAMNTALLand AVGGIFT was used instead of its original form, since they seemed to be log normally distributed (Figure 2, 3).

Average Gift Amount Count Plot

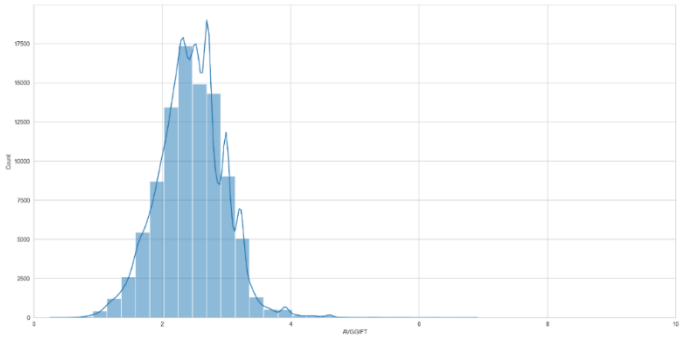


Figure 2: Count Plotof logarithm of average dollar amount of gifts to date

TotalGift Amount Count Plot

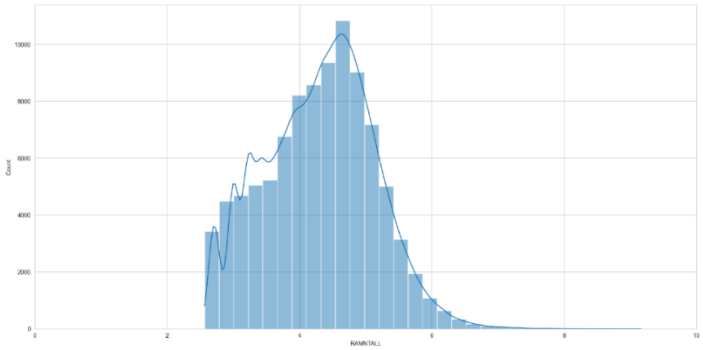


Figure 3: Count Plot of logarithm of dollar amount of lifetime gifts to date

The distribution of AGE variable, before taking its logarithm, by INCOME showed that the lapsed donors with higher household income were concentrated in the individuals around 40 and 70 years old (Figure 4).

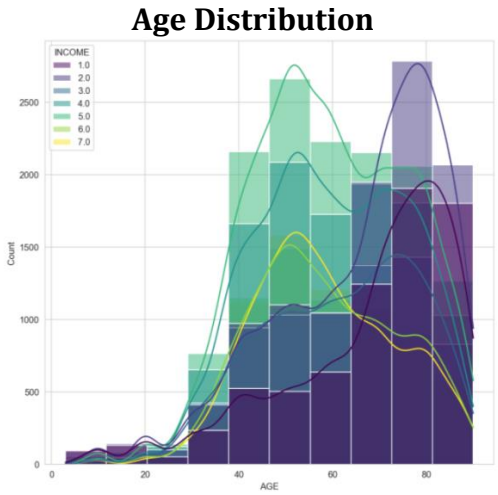


Figure 4: Donor's age distribution by household income

Visualizing the number of lapsed donors by state showed that almost 35% of the individuals live in three states: California (18.17%), Florida (8.77%) and Texas (7.89%). (Figure 5)

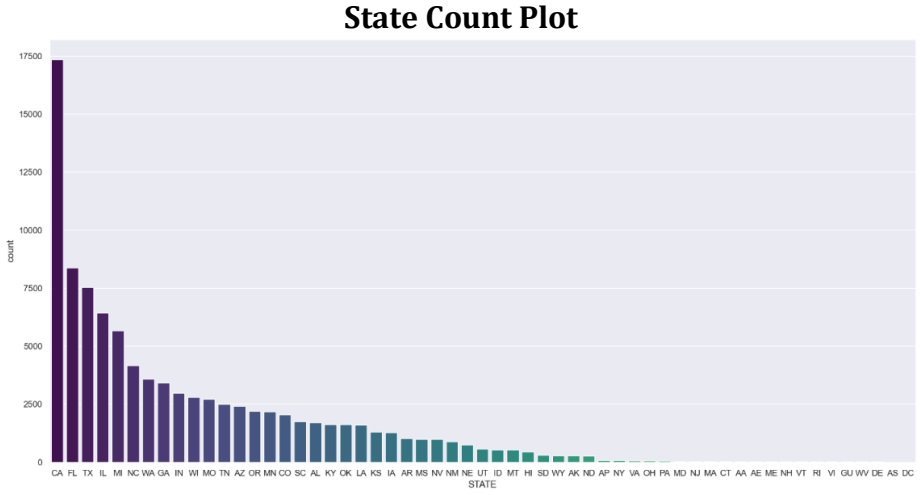


Figure 5: State count plot

In order to visualize the data set in a two-dimensional space, the two principal components, which represent 52% of the variance between the variables, were plotted. As shown in Figure 6 it is not possible to visualize any clear presence of clusters.

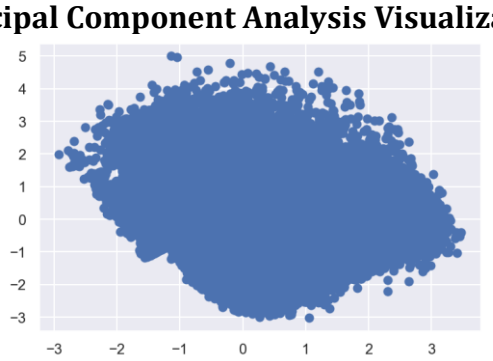


Figure 6: Visualization with two principal components

2. Clusterability Evaluation

On one hand, after computing the VAT algorithm for a represntative sample of PVA's database we can conclude that there are four square shaped dark blocks along the diagonal (Figure 7) which mean that the data clusterable with four clusters.

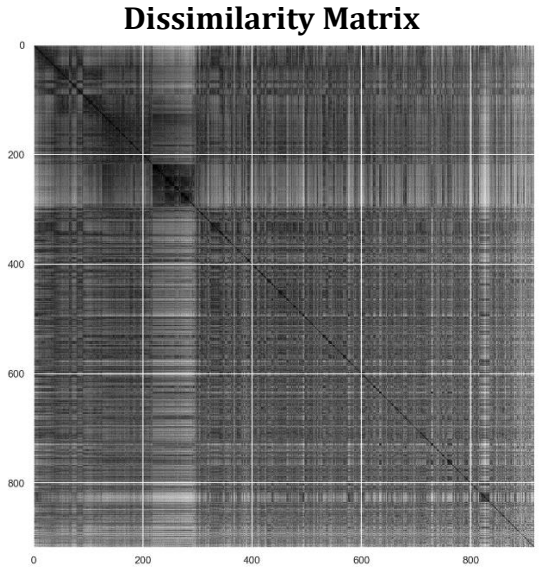


Figure 7: Visual assessment of cluster tendency VAT.

Likewise, after calculating the Hopkins statistic we found H=0.12 which is close to zero meaning that the data is clusterable.

3. Algorithm selection and execution

Defining the optimal cluster solutions for PVA donor's database segmentation (find clusters with distinctive characteristics), was the result of several clustering attempts with different sets of variables and algorithms.

3.1. K-Means

Before computingK-means clustering, the Elbow Method was used to define the number of clusters. The Figure 8suggested four clusters as the ideal number.

Figure 8: Inertia plot for a range of 2 to 10 clusters

After computing K-Means clustering we obtain the following four clusters (Figure 9):

Cluster 0: The individuals in this group donate more frequently than the others, nevertheless the total amount and average gifts arebetween the lowest within clusters.

Cluster 1: This group took more time to donate after their first gift and their frequency of donation is low, however their last donation is one of the highest.

Cluster 2: These individuals' donation frequency is the lowest, but the average total amount and the smallest value of their gifts are the highest. Also, they are the ones with the highest income, and they are less responsive to card promotions.

Cluster 3: These are the ones who have donated more times and amount in total. Also, they have been donating for longer time, and they represent the donors with the lowest income and the oldest in term of age. Furthermore, they are significantly more responsive to card promotions than the others.

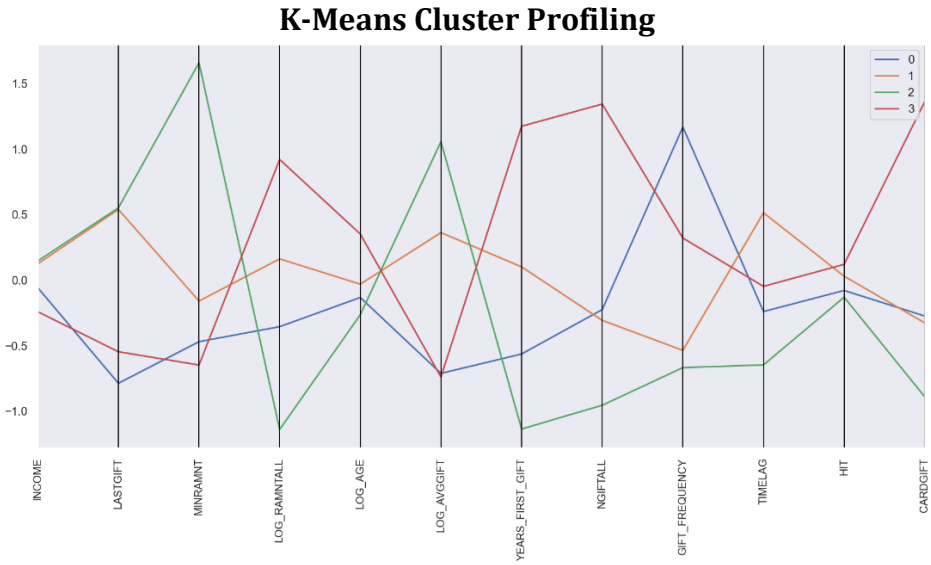


Figure 9: K-Means cluster evaluation. Characterization.

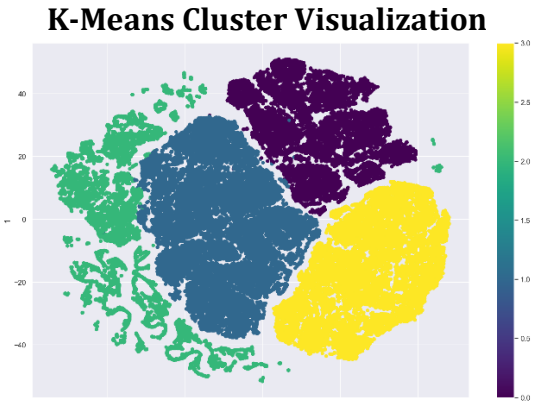


Figure 10: K-Means cluster evaluation. T-Sne visualization

3.2. Hierarchical Clusteringon top K-Means

Due to the high dimensionality of the data set, individual hierarchical clustering was not possible to compute, however, a random large number of clusters was defined to calculate K-Means algorithm to then use its centroids as input data for the hierarchical algorithm. Using this approach, the final number of clusters defined was five as shown in the Figure 11.

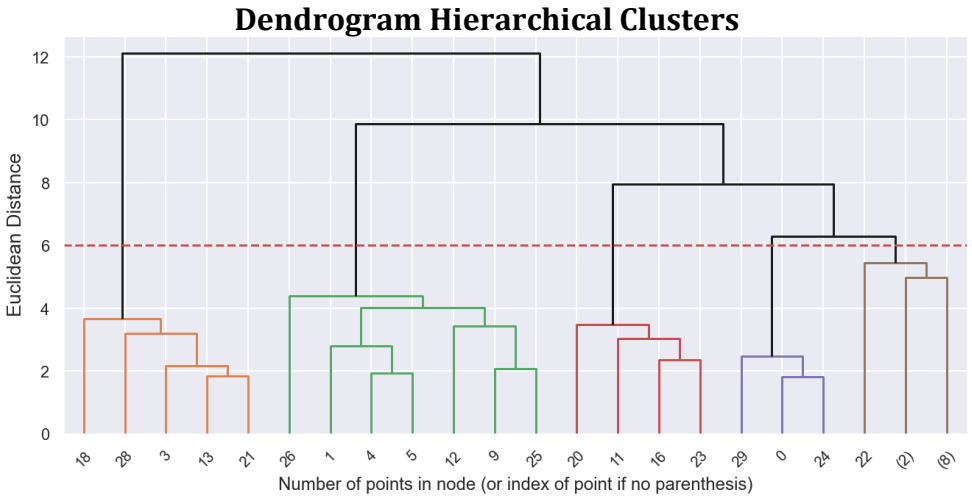


Figure 11: Dendrogram Hierarchical Clusters. K-Means centroids as input data.

After computing Hierarchical clustering on top of K-Means clustering we obtain the following five clusters (Figure 12):

Cluster 0: These individuals took more time to make their second donation than the others.

Cluster 1: These individuals donation frequency is the lowest, but the average total amount and the smallest value of their gifts are the highest. However, the are the newest donators. Also, they are the ones with the highest income, and they are less responsive to card promotions.

Cluster 2: The total amount and number of gifts are the highest in this group. Also, they have donated for longer time and their engagement to PVA promotions is high.

Cluster 3: The age of these individuals is the highest compare with the others. And they are significantly more responsive to mail promotions than the others.

Cluster 4: This group present the highest frequency of gifts; however, their average gift is the lowest. Also, they are one the newest groups of donors and the group with the lowest income.

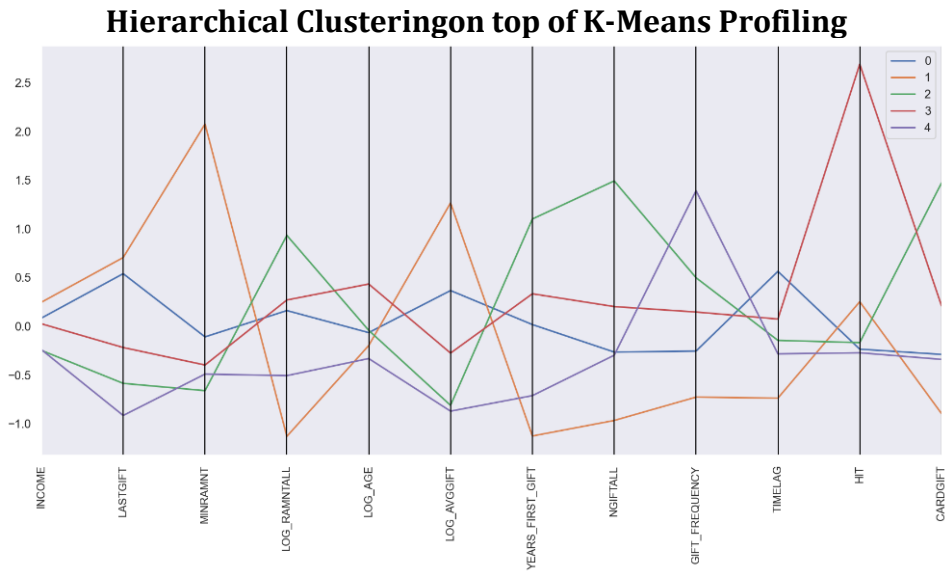


Figure 12: Hierarchical Clustering on top of K-Means evaluation. Characterization.

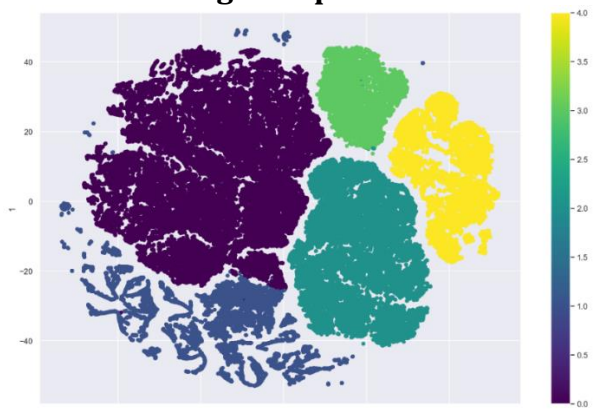


Figure 13: Hierarchical Clustering on top of K-Means evaluation. T-Sne visualization

3.3. Hierarchical Clustering on top of K-Means by Perspective

In order to remove the randomness when defining the number of clusters of K-Means, a second attempt to compute Hierarchical Clustering on top of K-Means approach was made after the data set was split intotwo groups depending on demographic and engagementperspectives of the variables.

The K-mean algorithm was computed to each data set, andtheir number of clusters wasdefined by plotting the R^2 scorefor a range of ten clusters. The optimal number of clusterssuggested by the R^2 score were six and seven fordemographic and engagement, respectively (Figure 14).

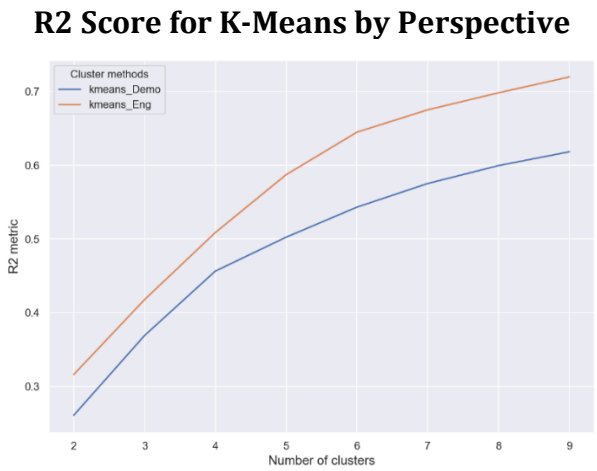


Figure 14: R^2 scores for a range of ten clusters Demographic and Engagement data sets.

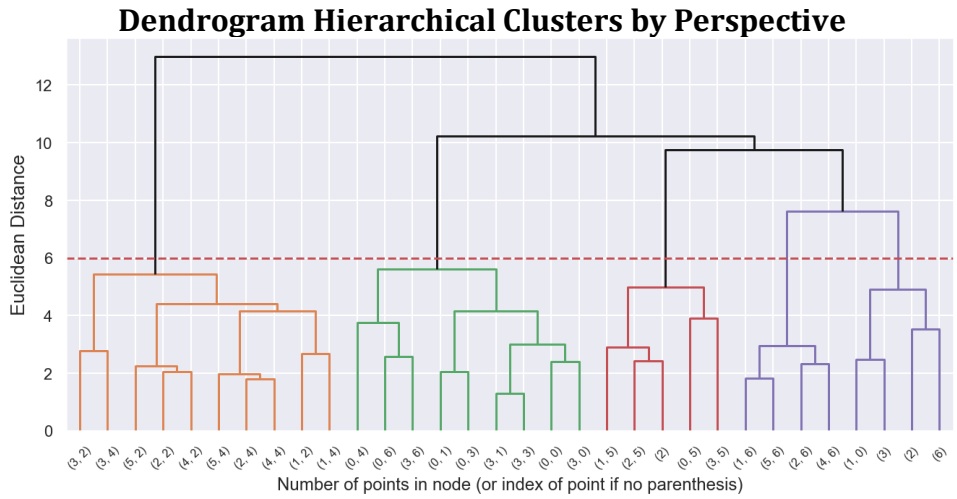


Figure 15: Dendrogram Hierarchical Clusters. K-Means merged centroids by perspective as input data.

After computing Hierarchical clustering on top of K-Means clustering by perspective we obtained the following final clusters (Figure 16):

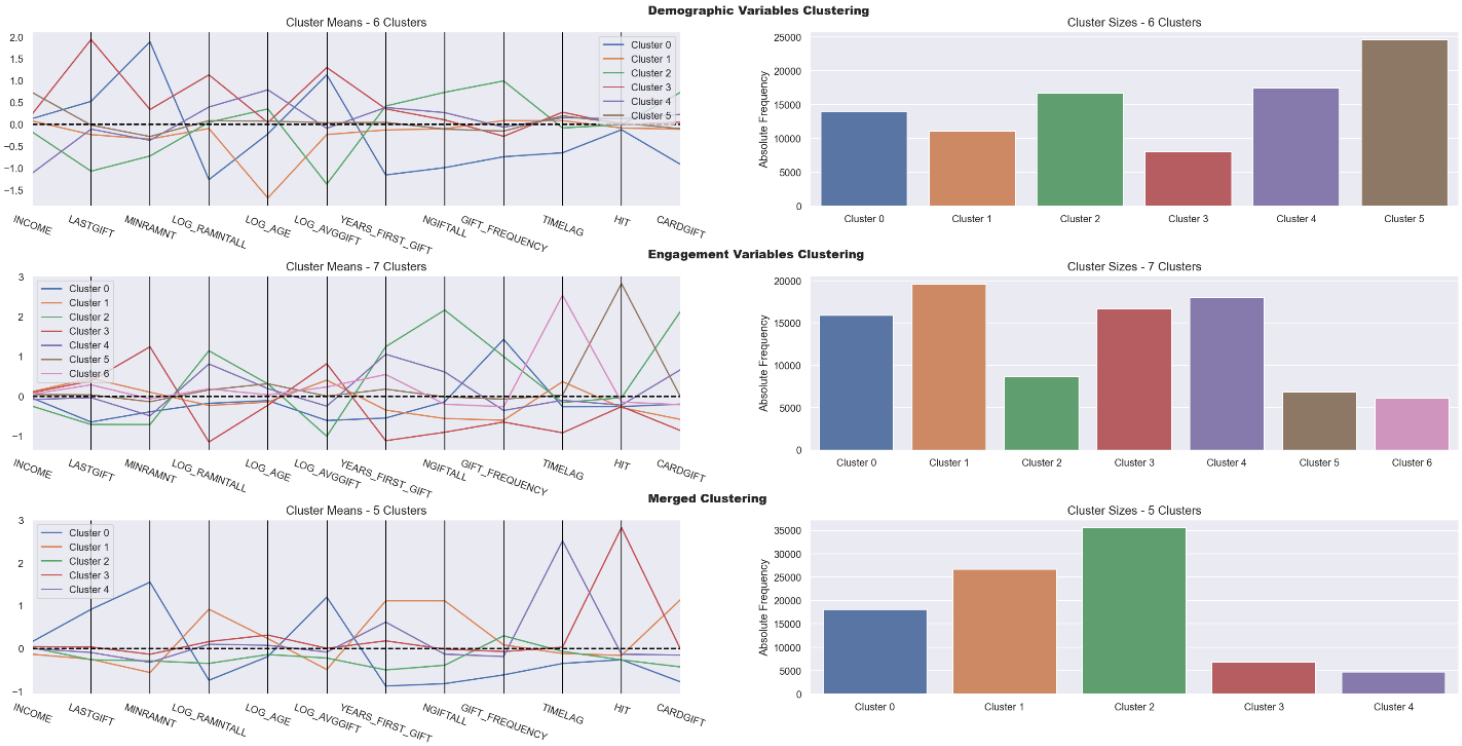
Cluster 0: These individuals donation frequency is the lowest, but their average total gift amount and the minimum value of their gifts are the highest. Also, they are the ones with the highest income and the newest donators.

Cluster 1: They have given the highest amount and number of gifts in total. However, the have the lowest income. Also, they are the oldest donators and highly responsive to card promotions.

Cluster 2: These individuals donate more frequently than the others.

Cluster 3: This group is significantly more responsive to mail promotions. Also, they represent the group with the highest age.

Cluster 4: These individuals took less time to donate for the second time than the others.



Hierarchical Clusteringon top of K-Means Profilingby Perspective

Figure 16: Hierarchical Clustering on top of K-Meansevaluation by perspective. Characterization.

Hierarchical Clustering on top ofK-MeansVisualization by Perspective Merged

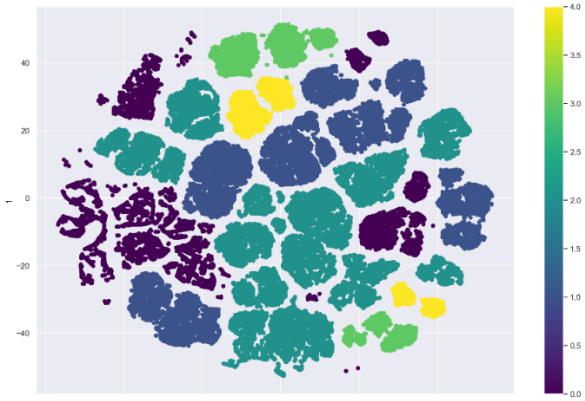


Figure 17: K-Means on top of Hierarchical cluster evaluation by merged perspectives. T-Sne visualization.

V. DISCUSSION AND CONCLUSIONS

Paralyzed Veterans of America Lapsed Donors case was the group's first experience applying unsupervised learning algorithms. Thus, it improved our understanding of the

clustering alternatives, their advantages and limitations, also our ability to address and implement clustering tohigh dimensional databases.

The fact that our dataset had over 400 variables and the lack of domain experience made the feature selection process time consuming. Another difficulty was the inconsistency in the values of some variables, it was necessary to convert them to meaningful ones, for example, the dates were given in string form and some categorical variables had empty inputs.

Due to the high number of missing values, filling them with their mean was not anappropriate option because itreduces the variance of the variable, so a k-nearest-neighbor imputer was considered and applied.
Since the standard interquartile method resulted in removing more than 50% of the observations, we decided to extend the interquartile range, taking the 10^{th} and 90^{th} quantiles.

The structure of the data did not look suitable for density clustering when visualizing using its principal components. Also, computing hierarchical clustering was not possible because of the high dimensionality of the data set. For this reason, only partitional clustering as K-means and a combination with agglomerative clustering as Hierarchical was computed.

After trying different algorithms, K-Means was the optimal clustering solution for PVA's case, because its results were easy to understand and self-explanatory. The final clusters represent well segmented groups for re-engagement marketing purposes such as:

Cluster	Marketing Strategy
Cluster 0: The individuals in this group donate more frequently than the others, nevertheless the total amount and average gifts are between the lowest.	Reaching out these donors again through various communication channels asphone calls, emails and social media, to identify to which communication style they respond better.
Cluster 1: This group took more time to donate after their first gift andtheir frequency of donation is low, however their last donation is one of the highest.	These donors seemed to be emotionallymotivatedto give, in response to well-defined needs. These donors should be re-contacted with information about the positive impact the organization is making, and hopefully, they will thinkabout donating again.
Cluster 2: These individuals donation frequency is the lowest, but the average total amount and the smallest value of their gifts are the highest. Also, they are the ones with the highest income, and they are less responsive to card promotions.	These donors were not happy with the communication strategies of the organization. Re-contacting these donors should not be entirely for asking about donations, instead, these donors should be informed about the positive impact the organization is making and how the donations are spent. Likewise, surveys asking about suggestion and their interest.
Cluster 3: These are the ones who have donated more times and amount in total. Also, they have been donating for longer time, and they represent the donors with	These donors seemed to be happy with the organization, it is possible that a change in the communication style made them stop donating. Sending a survey asking why they stopped donating,

the lowest income and the oldest in term of age. Furthermore, they are significantly more responsive to card promotions than the others.	acknowledge and follow up their feedback is crucial to re-engage these donors.

Table 3: Marketing strategies per cluster

VI. REFFERENCES

Jain, A. (1999). Data clustering: a review.

Kassambara, A. (2017). *Practical Guide To Cluster Analysis in R unsupervided Machine Learning.* STHDA.

Patel, A. A. (2019). *Hands-On Unsupervised Learning Using Python.* O'Reilly Media, Inc.