

CURRICULAR UNIT: DATA MINING

GROUP MEMBERS: DANIEL CORREIA (20200665),  
JOANA RAFAEL (20200588),  
RICARDO SANTOS (20200620)



## PROJECT REPORT

# Donor Segmentation and Marketing Strategy for the Lapsed Donors of Paralyzed Veterans of America

THIS DOCUMENT IS ACCOMPANIED BY THE FOLLOWING JUPYTER NOTEBOOKS:

1. DATA\_CLEAN\_EXPLOR
2. KMEANS\_HC
3. GAUSSIAN\_MIXTURE
4. SOM\_HC
5. CLUSTER\_CHARACTERIZATION

(HIGHLIGHTED ONES ARE SUFFICIENT AND NECESSARY FOR THE FINAL SOLUTION OF THE PRESENT REPORT)

FIND THEM ON GITHUB: [HTTPS://GITHUB.COM/RICARDOSANTOS0/PROJ-DM](https://github.com/RicardoSantos0/PROJ-DM)

---

# Abstract

In this work, we apply diverse unsupervised algorithms to segment and cluster lapsed donors from a dataset provided by the organization Paralyzed Veterans of America (PVA). We explored diverse methods of feature selection in an attempt to select the most promising features for the segmentation and evaluated the efficacy of 4 clustering algorithms (Kmeans, Kmeans + ward's hierarchical clustering, Gaussian mixture model clustering, and self-organizing maps + ward's hierarchical clustering). We assessed the adequacy of our clustering solution by resorting to 3 different metrics:  $R^2$ , Silhouette score, and the Calinski-Harabász score. From the selected models, the solution presented by self-organizing maps yielded the best results.

On the final stage of the project, we created 3 different representative personas (one for each cluster), made some remarks over the persona's distinguishable characteristics, and outlined the foundation of custom-built targeted marketing strategies to each cluster, to maximize the chances of getting these donors back to donate.

## I. Introduction

Paralyzed Veterans of America is a non-profit organization that relies on donations of concerned citizens to be able to programs and services for paralyzed US army veterans. Throughout time, PVA has managed to rely on mail fundraisers to remain active.

The evolution of technology has allowed for faster and more efficient ways to communicate and, more importantly, allowed PVA to explore more meaningful interactions with its donors while in the process, expand its donor base. PVA contacted our firm to request our assistance in the development of effective marketing strategies to keep donors engaged. To that effect, we were provided with a dataset of 95,412 so-called lapsed donors, which are donors that have not donated for some time but may still be recoverable. The first step of the project was then to get patterns that could translate into real marketing strategies. In essence, a data mining problem.

To address the problem, we relied on the Cross-Industry Standard Process for data mining (CRISP-DM) framework. CRISP-DM is an iterative and standardized process whose 6 fundamental steps cover the requirements of a successful supervised or unsupervised ML project (Chapman et al., 2000). This report showcases the results of our application of a CRISP-DM approach in i) the identification of representative lapsed donors and ii) the development of targeted marketing strategies designed to keep donors engaged with PVA.

## II. Background

In this section, we will briefly outline the theoretical background behind the techniques, algorithms, and assumptions that we used throughout the project. We will start by making a brief overview of the CRISP-DM general framework and then transition into some of the specifics of feature selection, clustering, and visualization.

---

## II.1. CRISP-DM

CRISP-DM is a dynamic and fluid approach that encompasses 6 logical steps that range a project's life from inception to deployment:

i) Business Understanding

Business understanding is the first step: the analysts look to understand a customer's needs and translate them into a solvable problem. After identifying the actual problem, teams then design problem-solving strategies and develop the success criteria by which the success of a strategy is measured.

ii) Data Understanding

In this step, the analysts perform exploratory data analysis by collecting, exploring, and assessing the quality of the available data. Data exploration is the identification of key features, how they are distributed, or how they relate to others. Data quality assessment is identifying data completeness and correctness. By the end of this step, teams should have a clear idea of which features to analyze and what manipulations need to be performed on the data.

iii) Data Preparation

Data preparation is the manifestation of the decisions planned in the data understanding phase. It is at this step that processes like data cleaning or feature selection take place. It is possible (and it was the case in our project) for the data preparation and data understanding steps to have some overlap (e.g. to perform feature selection/exclusion on a particular set of features before analyzing a different set of features). The output of the step is a cleaned, transformed, and duly formatted dataset that is ready for analysis.

iv) Modeling

In the modeling step teams either experiment with the different possible models that are capable of addressing the business question or perform different tests on an a priori selected model. Different alternatives are tested and benchmarked against one another according to the previously defined success metrics. The best performing model(s) will then go through to the evaluation step where the model's fitness to address the initial problem will be under scrutiny.

v) Evaluation

The evaluation step measures the adequacy of the model to address the initial questions and problems. It is at this step that the entire project is reviewed, the flaws in the approach are identified and discussed, and what are the possible and most adequate future steps concerning the project.

vi) Deployment

Should the evaluation step determine that the model is ready to leave the crib, then it is up to the team to outline a deployment strategy (including medium to long term monitoring) and ensure that the model is deployed correctly. It is also at the end of the project that the team writes a final report where the project is summarized, and the main results are presented. The current report is a part of the deployment step of our PVA DM project.

CRISP-DM is a dynamic and fluid approach in which projects go back and forth from step to step. It is not uncommon for projects on more advanced steps (e.g. Modeling) to be sent to a previous step (e.g. Data Preparation) for additional transformations.

## II.2. Feature Selection

Feature selection is a technique that is performed to reduce dimensionality during data preprocessing by selecting a subset of features based on their relevance or redundancy to address the identified problem (Yu and Liu, 2004). Feature selection models and algorithms can be grouped into 3 distinct categories:

- i) Filter methods, in which the *best features* are determined and selected according to one or more statistical measures like Pearson correlation or Chi-Square. These are faster and computationally more efficient but tend to ignore unobserved latent patterns between features.
- ii) Wrapper methods are black box solutions where inductive algorithms select subsets of features, and later assess whether to add or remove features from the subset. These methods tend to be more accurate than filter methods at the expense of being slower and more demanding. E.g.: the sci-kit learn (SKlearn) Select K-Best method (which we've used for feature selection of the census data) is a *wrapper* that calculates a score and returns the k features with the highest scores.
- iii) Embedded methods perform feature selection during model training phases which allows these methods to be as accurate as wrapper methods, while simultaneously lighter than them. However, reviews of the literature note some concerns with the suitability of these methods to handle high dimensional data (Venkatesh and Anuradha, 2019).

## II.3 Clustering: algorithms used and success metrics

Unsupervised machine learning is used to solve problems where the goal is to discern patterns from data. When faced with a set of observations with no clear labeled relationship between them, unsupervised learners look to learn the underlying data structure given a certain set of prior assumptions or heuristics. In particular, clustering is a method that partitions the unlabeled observations into different groups according to an underlying rule that may be stated as: *members of a group are considered similar, members of the outgroup are considered different* (Saxena et al., 2017).

Different clustering algorithms look to solve the clustering problem in different ways. During this project's modeling phase, we tested several different alternative clustering algorithms and different alternative configurations of the same clustering algorithm to find the best possible solution:

- i) K-means algorithm (Yadav and Sharma, 2013)

K-means is a clustering algorithm whose purpose in life is, when provided with the intended number of groups that is determined by the practitioner, to calculate and return the group configuration that minimizes the sum of squared distances (the similarity criterion of k-means which, in our case, is the Minkowski distance with  $p = 2$ , also known as Euclidean distance) between

the data points and their group's center of mass (the centroid). The algorithm iterates over the following two steps until convergence is reached or a threshold of iterations is passed:

- a) An assignment step where, for every data point, k-means calculates the distance between each data point and each centroid and assigns the data point to the closest centroid.
- b) An update step where the algorithm calculates the coordinates of the centroid that would minimize the group's sum of squared errors. Set the centroid's coordinates to be equal to the result.

K-means is a very popular algorithm due to its simplicity, fast computation, and relatively easy scalability. However, the algorithm also has several relevant drawbacks: k-means assumes that the data is metric (which means that there is a computable distance involved), it can only find local optimums (by definition cannot find a globally optimal solution), is sensitive to both outliers and initiation and (due to it being a hard clustering algorithm) is unable to properly divide the samples in edge cases.

These drawbacks recommend that analysis of results of k-means should be complemented with other algorithms (like Hierarchical Clustering).

- ii) Expectation-Maximization (EM) algorithm in Gaussian Mixture Models (Melnykov and Maitra, 2010; Sugiyama, 2016)

As previously discussed, k-means is not ideal to distinguish between edge cases. In some cases, there may be some overlap and uncertainty on whether a data point belongs to a group or another, whether it is because the distance to different centroids is very similar or simply because other factors shape the distribution of the group members on the multidimensional space. Approaches like weighted k-means try to bring some uncertainty into the clustering method by weighing the distance calculations of each data point to each centroid but, overall, k-means lacks the theoretical foundation that can address issues unrelated to distance (Kerdprasop et al., 2005).

Mixture models are generative models that take the vector of weights used in weighted k-means (which we'll call  $\Phi$ ) and claim that the weights represent a family of probability distribution functions (Gaussian distributions in the case of Gaussian mixtures models (GMMs)). In essence, GMMs are weighted combinations of simpler Gaussian distributions: each group (cluster) has its unknown Gaussian distribution (with its own set parameters) and each data point was generated from its own group's distribution.

Clustering by Gaussian Mixtures can, then, be seen as a reverse engineering problem. We assume that k prior gaussian distributions generated our data and we infer the distribution and parameters that originated each set of data points. Inference in ML is generally performed by Maximum-likelihood estimation (MLE), which is generally complex and difficult to solve. An elegant solution is to solve MLE by Expectation-Maximization (EM), an algorithm whose function assumes that the difficulty of the MLE computation is justified by the fact that we do not observe the whole picture and that there are unobserved, latent properties (e.g. group assignments) that would complete the observations and maximize the likelihood more easily. Like k-means, the EM

algorithm is also an iterative two-step algorithm that repeats until convergence is reached or a threshold of iterations is passed:

- a) An expectation step, where the model estimates the expected value of the group assignment,
- b) A maximization step, where the parameters of the distribution are maximized by MLE. Set the parameters to be equal to the result.

Despite being a generative model that does not need the existence of computable distances and is well suited to handle different configurations of data, GMMs still share a significant number of drawbacks with k-means. K-means can be regarded as a special case of GMMs.

iii) Hierarchical Clustering (Murtagh and Contreras, 2012; Sharma et al., 2017)

The interpretation of what constitutes a group depends on the considered level of analysis. At the lowest possible resolution, the entire dataset is considered to be a group. On the highest possible resolution, the differences between members are so significant that each individual is considered his own group. Hierarchical clustering looks to find a middle ground through a top-down (divisive) or a bottom-up (agglomerative) approach to the analysis of group similarities. The output of this method is a dendrogram that groups clusters according to their degree of similarity or *linkage*.

iv) Self-Organizing Maps (SOM) (Miljkovic, 2017)

SOM is an unsupervised neural network approach that intends to capture the essential relationships in the data while outputting low dimensional visual representations of higher-dimensional data. SOMs organize themselves in grids. After initialization, the training process of a SOM iteratively follows the next steps until convergence is reached or a threshold of iterations is passed:

- a) Competitive process, where all points are matched with their corresponding best matching most similar neurons (the winners or best matching units - BMUs) according to some metric (with Euclidean distance between the neuron and the vector being the most commonly used).
- b) Cooperative process, where the previously determined most similar neuron looks to interact with a neighborhood of excited neurons via a Gaussian neighborhood function.

After completing training, the most important statistical characteristics of the input space are presented in lower dimension visualizations that, in theory, retain the essential patterns (i.e. the topological differences between neurons remain) of the high dimensional data.

v) Success metrics

Regardless of the considered clustering algorithm, it is imperative to note that, even though all of the aforementioned algorithms look and perform grouping with a set of data they are provided with, they do not assess whether the grouping performed has any meaning beyond the instructions they need to fulfill their objective functions.

To assess the quality of our clusters, we relied on three different scoring metrics that are commonly used for this purpose:

i)  $R^2$ 

$R^2$  is an estimate of the proportion of the sample variance explained by the target variables that is used and accepted as a goodness of fit metric. Its value can vary between 0 and 1 (zero meaning no sample variance is explained by the target and 1 meaning all sample variance is explained by the targets). (Miles, 2005). Despite its widespread use, a high  $R^2$  does not mean high quality of the model and may even be misleading (Shalizi, 2015).

ii) Silhouette score

The Sk-learn built-in Silhouette score accounts for the silhouette coefficient varies between -1 and 1, where the minimum value represents *incorrect clustering* (meaning it is likely that there are mislabeled data points) and 1 is a highly dense, well-separated cluster (Rousseeuw, 1987).

iii) Calinski-Harabász score

The Calinski-Harabász score is another sk-learn built-in method whose unique utility is to perform a comparative analysis between different models, or the same model with different numbers of clusters, to assess which method performs better at obtaining dense and well-separated clusters. The evaluation criteria for this score is simple: even though there is no standard for what a good value is, the higher the score's value the better (Calinski and Harabasz, 1974).

## III. Methodology

The following section will focus on the methods and transformations that allowed us to explore, clean the necessary data, select the features ultimately used for clustering, employ the used clustering algorithms, and to characterize the final groups of donors. It will be split into the CRISP-DM framework's 6 main steps. Even though we did experiment with different combinations in our project, we will, in the interest of clarity and brevity, only highlight the ones that made it to the final selection process.

### III.1. Business Understanding

Donor segmentation is an important strategic process. It allows organizations to develop targeted marketing strategies custom built to specific groups of donors as a way to incentive them to donate. This process is particularly important for groups of consistent donors that stopped donating recently. Capturing these donors back is of paramount importance for PVA. If they can maintain their donor base as engaged and as best as possible, they can ensure their subsistence. Not only that, studying donor behaviors can potentially allow PVA to capture new donors.

### III.2. Data Understanding

A preliminary overview of the data's properties was performed. We started by dividing the dataset into 4 subsets of data: *Personal/Demographic Information*, *Personal Interests*, *Promotion/Giving History File*, and *Data from Census*. All categorical variables were immediately discarded for clustering purposes, as the techniques we employed are difficult to maneuver with this sort of feature. Please note that some of these variables were still cleaned and preprocessed, as we intended to use them for the final analysis.

We observed that not all donors were misclassified as lapsed (i.e. the difference between the dates of the last two donations was inferior to 13 months or superior to 24 months). These incorrect entries were removed (~11.7%) alongside one-time donors (~10.5%). These manipulations left us with ~77.9% of the initial dataset.

We relied on appropriate methods from the *pandas* library to identify missing values, potential outliers, and uncover the statistical properties and attribute behavior of each feature. Given the vast extension of our dataset, we performed data cleaning and transformation on selected variables only, as will be further discussed.

### III.3. Data preparation

The following manipulations of the original dataset were made as a consequence of the immediate observations made in the previous phases. Both data exploration and data cleaning were performed simultaneously. Some of the following manipulations were performed before the application of clustering algorithms, but do not result from immediate or self-evident observation. They instead resulted from the adoption of heuristics:

- i) Entries with missing values, e.g. “ ” or nonsense characters for the variables in question, were replaced by np.nan;
- ii) Variables with a high amount of missing values were rejected and not used for subsequent analysis. Missing values of the remaining variables were filled using the median (for the census subset) or the mode (for the remaining subsets), except for some anecdotal situations that will be discussed in the following sections;
- iii) Some feature engineering was performed, as a way to create and/or modify new variables that we considered important for the analysis. Relevant examples include:
  - a) AGE, calculated from the date of birth (DOB) and the date of the last promotion (ADATE\_2) (we assumed the last update on the dataset was made on this date);
  - b) LONGEVITY, calculated as the delta between FIRSDATE and LASTDATE;
  - c) WEALTH1 and WEALTH2 showed a high number of missing values (~50%); furthermore, they showed to be highly correlated. We joined them into a single variable WEALTH (WEALTH1 with missing values filled with the value of WEALTH2); remaining missing values were filled with the mode;
  - d) Categorical binary variables with alphabetical characters were converted to binary variables (0/1). Categorical non-binary variables were transformed to numerical variables (if they weren't already);
  - e) The variable PROP\_RESP\_CARDPROM was obtained by dividing the number of gifts donated in response to card promotions by the number of card promotions received to date, as a way of getting an insight into the incentive of the donors to donate in response to received card promotions.
- iv) Features with a Pearson correlation superior to 0.8 were rejected (the determination of which of the correlates was rejected depends on a diverse set of factors ranging from the number of missing



values to correlations with other variables). Correlations of variables between subsets of data were also checked, and the same logic was applied.

v) Variables deemed as irrelevant for the problem in hands were dropped at this stage. The discard of these variables relied upon the adoption of heuristics. We are aware of the potential dangers of this approach: it can lead to the loss of potentially important information. We tried to drop variables that we considered flat-out useless for the challenge and/or that represented redundant information (please refer to the *DATA\_CLEAN\_EXPLOR.ipynb* file). Our approach to the Census data relied on the employment of other feature selection techniques as further discussed in point vii of this subsection;

vi) Outliers were removed using the 1.5 Inter Quartile Range (IQR) rule, which accounted for ~13.9% of the working data. At this stage, we were left with ~68% of the initial dataset to perform clustering. A data frame containing the donor's data of these outliers was exported and kept for cluster analysis.

vii) Feature selection for the census variables combined different selection methods. In a first approach, we used a wrapper method test (k-Best with linear regression) and tested several targets (RAMNTAL, NGIFTALL, and AVGGIFT). We picked the top 30 scoring variables and verified the correlation between them. Again, correlations greater than 0.80 led to the discard of one of the correlates. Criteria for exclusion included the degree of association with other features and the K-Best score.

viii) To reduce the dimensionality we performed additional selection by analyzing the behavior of the features in the Component Plates in SOM. Finally, we relied on domain knowledge to finalize our selection process: 7 features proceeded to the next phase: LONGEVITY, RAMNTALL, NGIFTALL, PROP\_RESP\_CARDPRM, HVP4, IC5, EC4.

ix) All numerical variables were subject to sk-learn's StandardScaler and normalize methods.

#### III.4. Modeling

The modeling phase followed different approaches: the single method and the hybrid method clustering. The hybrid method has a primary layer that applies a base cluster to aggregate similar data points. The next layer aimed to aggregate the formed sets of clusters. To choose the ideal number of clusters, we used the Elbow method and the Euclidian distance dendrogram.

Based on the previous assumptions, we tested the following models: K-means with 6 clusters, K-means and Hierarchical with 2 clusters, GMMs with 6 clusters, GMMs with 3 clusters, SOM50 (2500 nodes) and Hierarchical with 3 clusters, and SOM100 (10000 nodes) and Hierarchical with 3 clusters.

#### III.5. Evaluation

We assessed model adequacy and selected the clustering algorithm to use through different metrics, namely R-squared, silhouette score, and the Calinski Harabasz criterion. At this stage, editions to the selected variables were performed iteratively and continuously assessed. Editions were validated if and only if they resulted in improved clustering model performance.

#### III.6. Deployment

At the final stage of the CRISP-DM framework, we performed cluster characterization, to identify patterns of behaviors intrinsic to each obtained cluster. Donors previously considered as outliers were integrated into the respective cluster using a classification tree trained with the clustered data.

## IV. Results and Discussion

### IV.1. Model selection

During the process of selecting models for clustering, we tested two types of models: the single method with K-means and GMMs; and the hybrid method: K-means with Hierarchical clustering and Self-Organizing maps with Hierarchical clustering.

Table I: Summary of scores obtained with "R2" (Coefficient of determination), "Silhouette score" and Calinski-Hárabasz

	R2	Silhouette	Calinski-Harabasz
<b>K-means (6c)</b>	0.59	0.21	18 952
<b>K-means + Hier (2c)</b>	0.29	0.28	26 362
<b>Gaussian (6c)</b>	0.53	0.15	14 556
<b>Gaussian (3c)</b>	0.39	0.22	21 112
<b>SOM (50) + Hier (3c)</b>	0.39	0.53	108 090
<b>SOM (100) + Hier (3c)</b>	0.41	0.53	115 804

Table I summarizes the  $R^2$ , Silhouette, and Calinski-Harabasz scores obtained from the cluster models. In the context of this problem, as expected, the SOM cluster obtained, for the most part, higher and more consistent scores than the others. We expect this to be due to its ability to map a large input in a lower dimension grid while preserving the topological order present in the input space (Miljkovic, 2017). It also has the fundamental advantage of being able to automatically generate data overviews, specifically a visualization of the grid structure, and quality measures (Sacha et al., 2018).

To choose which of the SOM models brought the most advantages to the study, we took into account the behavior of the SOM with 2,500 nodes (50x50 matrix) and with 10,000 nodes (100 x100 matrix). For the SOM algorithm taken into account the problem context, we used the same parameters in SOM50 and SOM100: random initialization, batch training, hexa lattice, 100 rough train and 100 finetune train.

Computationally, both models showed differences: SOM 100 is much more computationally demanding, taking almost 7 times longer than SOM50 to run. Another aspect to take into consideration is the number of nodes, which has a great effect on the result (Ponmalai and Kamath, 2019). We observed that with SOM100 we were able to obtain more clearness in the results.

Taking into account these factors, and the fact that the average distance from each data point to the nearest node - quantization error - was smaller in SOM100, we decided to move on to the next phase with the SOM100 model.

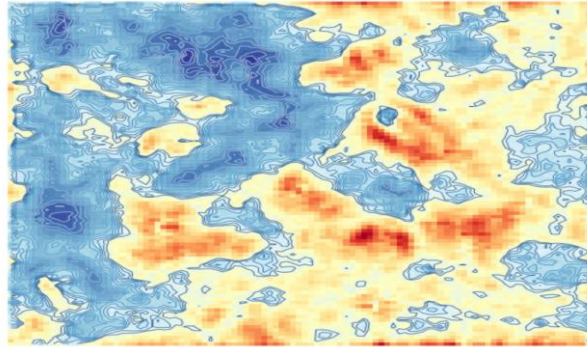


Figure 1: Topological graph of the SOM100 model. The blue zones represent a higher density of nodes.

Figure 1 represents the density of nodes. It is possible to observe two denser stains on the left side of the image, and several small stains scattered throughout the image. To aggregate these dispersed sets, we used a second layer of clustering: Hierarchical Clustering.

Hierarchical clustering aims to aggregate nodes so that it is possible to generalize the groups even more. We decided to use ward linkage as a form of aggregation (Murtagh and Contreras, 2012). To define what is the best approach to take, regarding the number of clusters to use, we used a dendrogram to visualize the aggregation strata taking into account the Euclidean distance between groups. As shown in Figure 2, we used a threshold based on the averaged euclidean distance to define the number of clusters used to segment our dataset of donors.

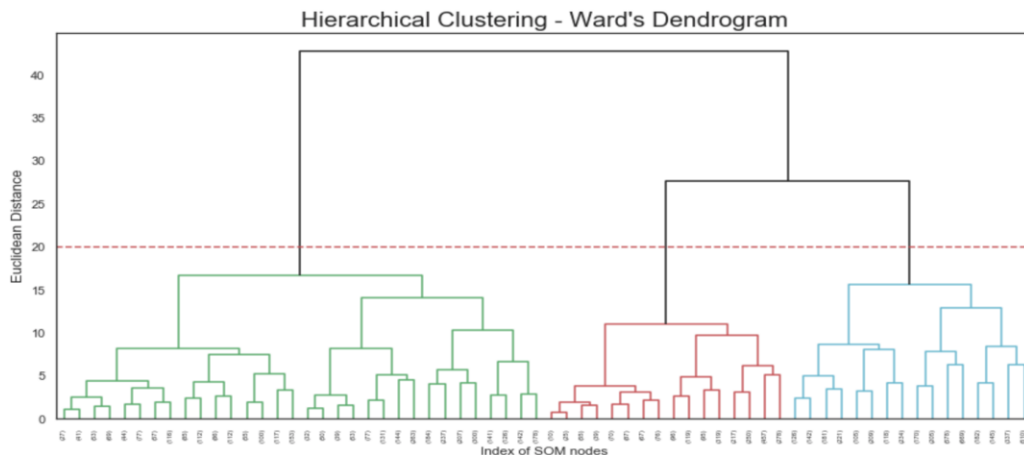


Figure 2: Euclidean distance dendrogram based on the ward linkage between the nodes of the SOM100 model.

We further used the KNN supervised learning algorithm to predict the new labels of each of the 3 clusters. Moreover, at this point, we re-introduced the outliers (previously removed from our dataset) in our data. We predicted the label of the outliers using KNN (K=5).






Figure 3: Cluster division of the SOM100 model. Each color represents a different cluster.

#### IV.2. Donor Segmentation and Characterization

Our clustering algorithms returned 3 different clusters whose representative persona is shown in Table II. Each persona represents the most noticeable distinguishing features of our population groups. Our donors are in their 50s to early 70s. PVA can use these personas as placeholders for targeted marketing campaigns.

Table II: Representative personas of each of our clusters. In the second row, we present a small biography, the third row briefly outlines the main identified characteristics of the group, and the fourth row proposes a marketing strategy to apply to this population.

	CLUSTER 1:	CLUSTER 2:	CLUSTER 3:
Persona			
	SHEILA	KAREN	BILL

<b>Biography</b>	<i>Sheila has a long family history of veterans that felt the consequences of the war. She lost her grandfather who fought bravely for the country. Hence, since young she has been sensitive to the cause and had been contributing with generous donations to PVA, an organization she truly admires.</i>	<i>Karen has been lucky enough to be able to afford a comfortable lifestyle. She lives in a wealthy neighborhood in California and makes sure to maintain an active social life. She is proud of her country and sensitive to the support that PVA concedes to the disabled veterans of America, so she donates to PVA (if reminded).</i>	<i>Bill lives in a rural area in the interior of the country. He has friends who have felt directly the consequences of the war. Although he doesn't have the means to contribute monetarily as often as he would like, he tries to help as much as he can.</i>
<b>Population Characteristics</b>	Sheila has been a loyal PVA donor for years (high longevity). She has the highest average number of gifts (high NGIFTALL) and has donated more than any other person (high RAMNTALL). Sheila also responds regularly to card promotions (highest PROP_RESP_CARDPROM). Sheilas exist in all socio-economic strata, with no relevant incidence in any particular class except a low incidence in the tend to belong to the higher socioeconomic status.	Karen is characterized by her high social-economic status. She lives in neighborhoods with a larger average per capita income - IC5). This observation was strongly supported by several features from the census data, as discussed. About 1/3 of our sample's Karens live in the state of California. This is an important observation as it might be useful to target them with marketing strategies based on their location.	Bills are present all around the USA. We noticed some nuances (further discussed below) that allowed us to distinguish them from the remaining his peers: fewer years spent studying (also seen from the variable EC4 used for clustering) and lower economical possibilities. Indeed, the lower values in the clustering variable HVP4 hints that Bills live tend to live in more modest neighborhoods and rural areas. This assumption is further confirmed by some additional variables from the census data, further discussed below.
<b>Marketing Strategy</b>	We suggest targeted marketing campaigns (possibly through card promotions since they seem to be effective with Sheila in comparison to her peers) thanking her for her continued help. It would be interesting to let Sheila know the veterans that are being assisted with her previous donations. It would give her a sense of pride for helping and remind her how important her contributions were and continue to be.	A useful approach would be <i>on the ground</i> marketing campaigns with flyers, street signs publicity, and public charity fundraising/auction events. It'd be interesting to further study if these people live in specific neighborhoods so that this task could be made easier and cheaper.	Bills live in rural areas all around the country with more preponderance in the States of Texas, Florida, and Michigan. They represent the forgotten men and women who are proud to be American and are not afraid to say so. To reach these citizens, PVA should focus on contacting state TV and Radio channels and create advertisement campaigns through those channels. This could also help to raise new donors from these regions to donate to PVA.

The construction of our personas considered discernable characteristics between clusters. We characterized the groups according to the features used for clustering (Figure 4) and relevant demographic and social-economical features (Figure 5). We also looked into all available census data for other distinguishing features in the Census data. We found the clusters to be mostly distinguishable by frequency, amount history of donations, social-economic status, and/or area and neighborhood of residence.

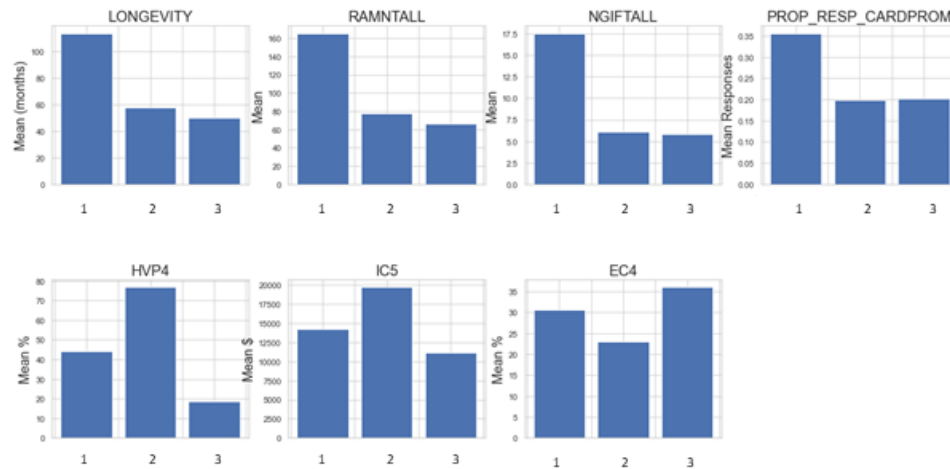


Figure 4: Characterization of Clusters based on the features used. LONGEVITY: number of months passed between his first and the last PVA promotion. RAMNTALL: Total amount of money donated to date. NGIFTALL: Number of donations made to date; PROP\_RESP\_CARDPROM: the proportion of donations made to card promotions received. HVP4: the percentage of homes valued at \$75,000 or more in this donor's neighborhood. IC5: *per capita* income of neighborhood; EC4: percentage of adults with high school as highest academic achievement in the neighborhood.

The wealth gap between Karen and her peers is extensively supported by the census data. E.g. Karens have, or at least tend to live in areas with higher income per family/household (refer to boxplots of features IC\_ in Jupyter notebook 5); they also live in areas with more expensive houses and higher rents (boxplots of features HV\_, HVP\_ and RP\_ plotted Jupyter notebook 5). Moreover, Karens showed to live in neighborhoods with a higher percentage of people employed in sectors typically well paid (e.g. finance, insurance, health services, see boxplots from feature EIC9 and EIC13 plotted in Jupyter notebook 5), while Bill's neighbors tend to be employed in sectors from rural environments (e.g. farmers, craftsmen, precision, repair operatives; refer to boxplots OCC8-13 in Jupyter notebook 5).

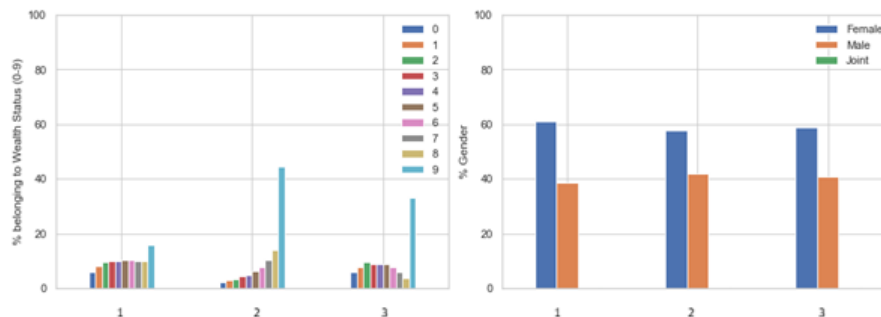


Figure 5: Left: Characterization of clusters based on wealth status (sum of WEALTH1 and WEALTH2); right: distribution of people by gender.

It is worth mentioning that all donor groups were dominated by females. Also, most donors live in 3 major states (Figure 6): California, Texas, and Florida, which is justified by the fact that these are the most populated states in the country. It would be interesting to analyze this data from a population-adjusted lens. Nevertheless, this doesn't change the fact that Karens overwhelmingly live in California, the most populated and one of the wealthiest states.



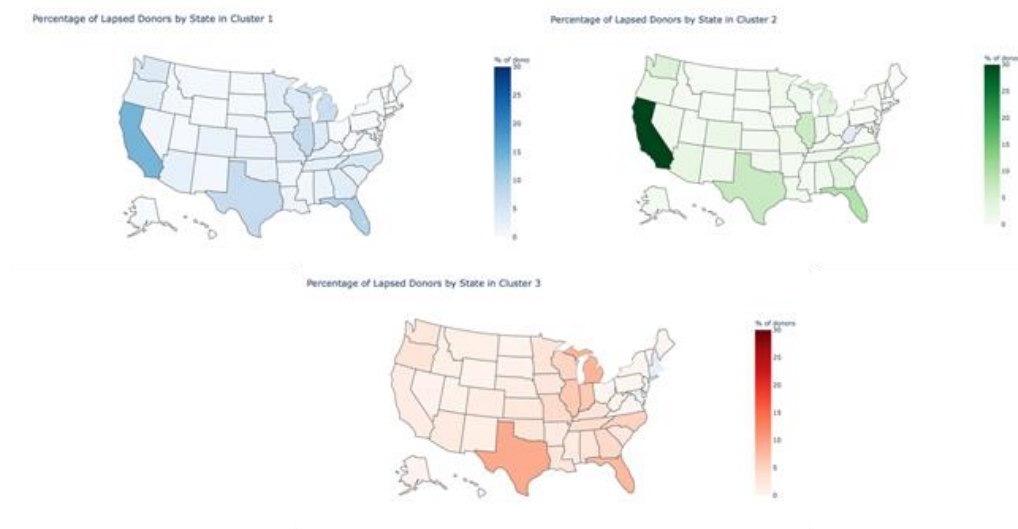


Figure 6: US heatmap with the % representation of each cluster by the state of residence (scale varies between 0 and 30%)

We also used the data related to the interests of the donors to look for strong potential targeting content. However, we did not find any significant intra-cluster interest or response pattern to specific types of emails that would allow us to make a coherent and incisive campaign (see supplementary figures S1 and S2). Nevertheless, this data could be used to target very specific donors individually. In particular, donors marked with any type of special status (Major Donors or PEP Star RFA Status) seem to be especially engaged, which would be a good approach to capture these few important donors back.

## V. Conclusion

In this study, we use self-organizing maps and ward's hierarchical clustering to segment and cluster donors from a dataset from PVA, in an attempt to aid the non-profit develop targeted marketing campaigns that boost donations from their previous lapsed donors. Our approach allowed us to partition the dataset into 3 main groups, which were then characterized and given a representative personality.

It is worth mentioning that this was a very challenging dataset to work with and that it took a lot of trial and error to find our footing in this project: the number of features made, at times, the task seem overwhelming and we feel that the aggregation of 3 different datasets (personal data, census, and data collected by external companies) led to some unnecessary dilution of information.

In the future, we would like to make less heuristic decisions and focus more feature selection techniques to look for potential patterns that we may have missed.

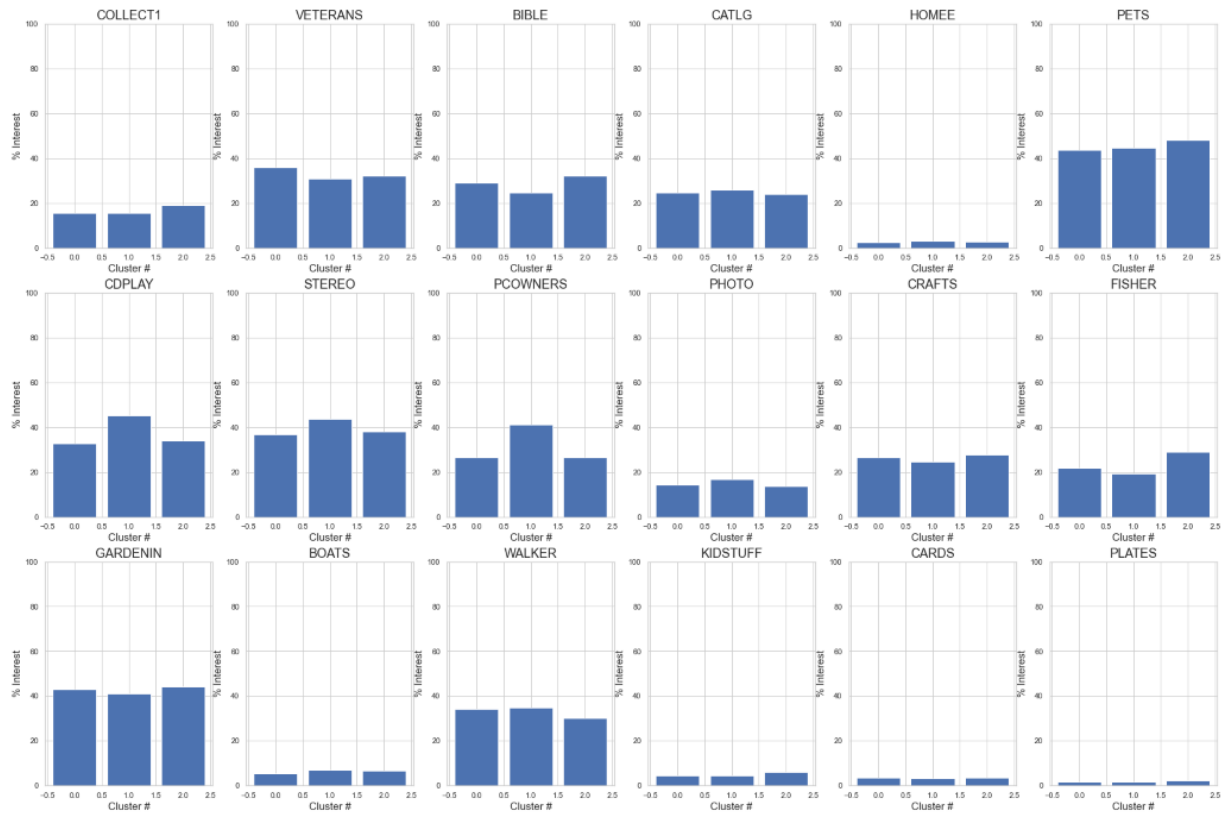
As a final note, we would also like to note that we do realize that patterns used to distinguish clusters taken from the census data lack some form of hypothesis testing and statistical confirmation. For what is worth, we tried to be careful with our conclusions and make assertions that withstood the eye test.

## VI. References

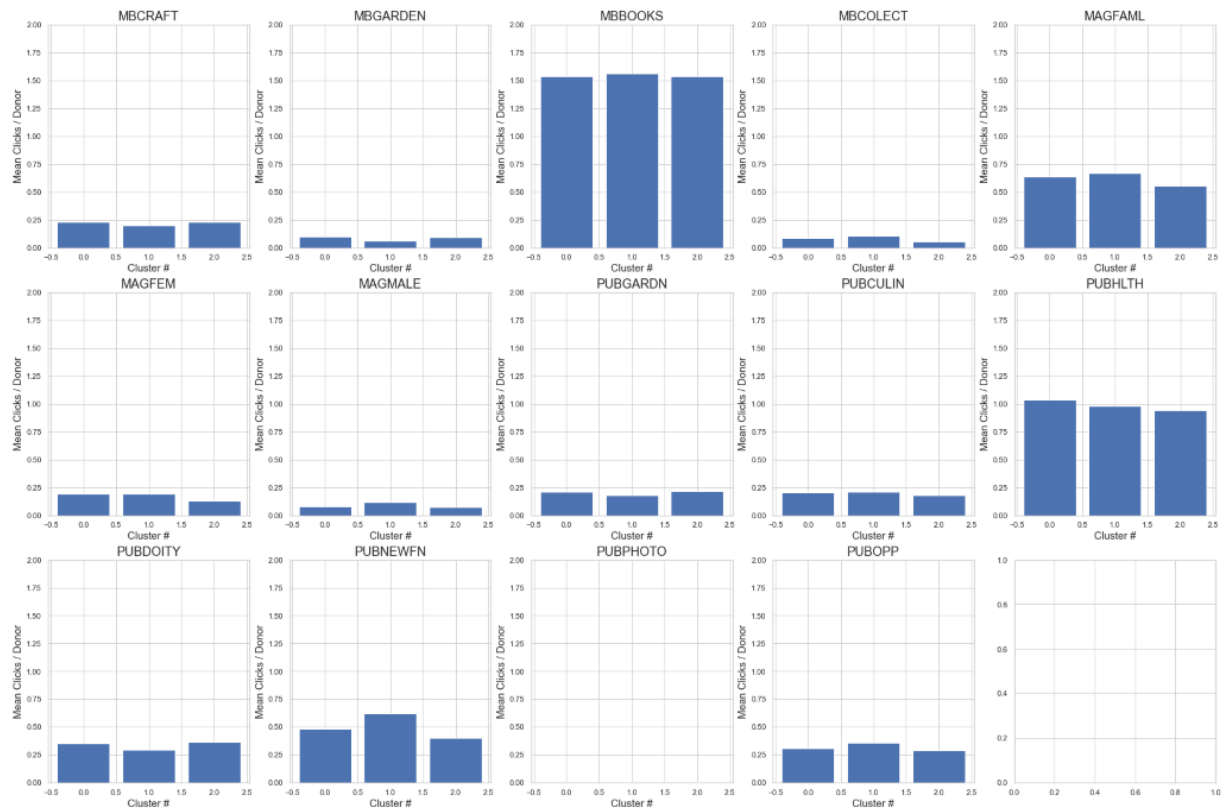
- Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. *Commun. Stat. - Theory Methods* 3, 1–27. <https://doi.org/10.1080/03610927408827101>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. Step-by-step data mining guide 76.
- Kerdprasop, K., Kerdprasop, N., Sattayatham, P., 2005. Weighted K-Means for Density-Biased Clustering, in: Tjoa, A.M., Trujillo, J. (Eds.), *Data Warehousing and Knowledge Discovery*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 488–497.
- Melnykov, V., Maitra, R., 2010. Finite mixture models and model-based clustering. *Stat. Surv.* 4, 80–116. <https://doi.org/10.1214/09-SS053>
- Miles, J., 2005. R Squared, Adjusted R Squared 3.
- Miljkovic, D., 2017. Brief review of self-organizing maps, in: 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). Presented at the 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, Opatija, Croatia, pp. 1061–1066. <https://doi.org/10.23919/MIPRO.2017.7973581>
- Murtagh, F., Contreras, P., 2012. Algorithms for hierarchical clustering: an overview. *WIREs Data Min. Knowl. Discov.* 2, 86–97. <https://doi.org/10.1002/widm.53>
- Ponmalai, R., Kamath, C., 2019. Self-organizing maps and their applications to data analysis. Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States).
- Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Sacha, D., Kraus, M., Bernard, J., Behrisch, M., Schreck, T., Asano, Y., Keim, D.A., 2018. SOMFlow: Guided Exploratory Cluster Analysis with Self-Organizing Maps and Analytic Provenance. *IEEE Trans. Vis. Comput. Graph.* 24, 120–130. <https://doi.org/10.1109/TVCG.2017.2744805>
- Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W., Lin, C.-T., 2017. A review of clustering techniques and developments. *Neurocomputing* 267, 664–681. <https://doi.org/10.1016/j.neucom.2017.06.053>
- Shalizi, C., 2015. Lecture 10: F-Tests, R<sup>2</sup>, and other distractions." *Modern Regression* (2015).
- Sharma, A., López, Y., Tsunoda, T., 2017. Divisive hierarchical maximum likelihood clustering. *BMC Bioinformatics* 18, 546. <https://doi.org/10.1186/s12859-017-1965-5>
- Sugiyama, M., 2016. Maximum Likelihood Estimation for Gaussian Mixture Model, in: *Introduction to Statistical Machine Learning*. Elsevier, pp. 157–168. <https://doi.org/10.1016/B978-0-12-802121-7.00026-1>
- Venkatesh, B., Anuradha, J., 2019. A Review of Feature Selection and Its Methods. *Cybern. Inf. Technol.* 19, 3–26. <https://doi.org/10.2478/cait-2019-0001>
- Yadav, J., Sharma, M., 2013. A Review of K-mean Algorithm. *Int. J. Eng. Trends Technol.* 4, 5.
- Yu, L., Liu, H., 2004. Efficient Feature Selection via Analysis of Relevance and Redundancy. *J Mach Learn Res* Vol. 5, 1205–1224.



## VII. Supplements



Supplementary Figure S 1: Supplementary Figure S1: Mean percentage donors from each cluster that reflect interest in a given field (from donors who showed any interest in the data, as we assumed that no interests at all meant missing data)



Supplementary Figure S 2: Mean percentage of donors that have responded/clicked in other times of email offers (from donors who showed any responses in the data, as we assumed that no responses at all meant missing data).