Data Mining Project Report

Segmentation of the Paralyzed Veterans of America Dataset

Data Mining

Nova IMS

January 2020

GROUP AE

Students: Berfin Sakallioglu, 20200545

Emil Ahmadov, 20201004

Ema Mandura, 20200647

Instructor: Fernando Lucas Bação

Introduction

Paralyzed Veterans of America (PVA) is a non-profit organization focused on helping war veterans with diseases or injuries by making direct mail fundraisers. Their database holds 13 million donors, describing each donor with many different attributes. As a client, the PVA has made a request for a data mining approach that will separate their donors according to their characteristics and behavior, so they get a better understanding of the patterns within their database.

PVA's most important fundraising strategy is re-engaging former donors. A group of special interest amongst those donors are the one who donated recently, within the last 13 to 24 months, so called lapsed donors. Therefore, the result of this customer segmentation should provide segments within lapsed donors such that an appropriate marketing strategy could be applied for each segment in order to get them to donate again.

Within this project, a suggested data mining approach will be provided in form of a project report and code.

Data Exploration

Before starting to manipulate the data in any way, it is crucial to first understand it. After loading the dataset into a data frame in the Jupyter notebook and checking the shape, it became apparent that the dataset in question had 95412 rows and 476 columns. Considering its large number of columns, the dataset was challenging to explore and understand. In addition, column names were not self-explanatory, but declared as codes, so it was necessary to look at the metadata in order to get any understanding.

To start off the process, a pandas profiling report was first generated. Parallelly looking at the report and checking the metadata, a few conclusions were made.

First, it became apparent that some of the information provided in the metadata did not match the actual data. Some values of a variable explained in the metadata were represented differently in the dataset.

Further on, it was noticeable that a lot of the variables have very high percentages of missing values, which means that there is a need for either filling missing values or dropping variables.

Further exploring the data also lead to the conclusion that some of the missing values represent the actual absence of a value for the row in question, rather than it not being inserted into the dataset. While in some different examples, it seemed that the value '0' represents a missing value, rather than an actual value of zero.

Looking at this dataset and considering the good data storing practices taught in class, it seems that this dataset was not designed very well for computer analysis and some data preparation and preprocessing is needed to perform clustering on it.

While looking at all the columns one by one, the team considered which of them could be of importance and suitable for doing the clustering. In this process, nineteen variables were chosen as potential variables to perform clustering with.

In order to deal with the issue of dimensionality, the number of variables had to be reduced. The team decided to look at the correlation between variables and set the threshold at 0.9. This way, with two variables that have a correlation above 0.9, only one of them will be kept. During this process, the number of columns was reduced to 403.

Data Preprocessing

In this part, preprocessing steps which were done in order to have more reliable data will be explained in detail. To start with, since clustering is a method which relies on calculation of distances, and non-metric variables cannot be used to calculate distances, only metric variables were selected. Non-metric variables are stored in a list in order to be used for further analysis after clustering will be done.

Correlated Variables

One of the problems related to the given dataset was that some of the variables had high correlation between them in the dataset as mentioned above. In order to clean data, correlations between all metric variables in the dataset were calculated. After finding the correlation between variables, tuples that had more than 90 percent of correlation between them were analyzed in detail and the variables which contained more information and was more appropriate for the aim of clustering analysis were kept. Also, not only correlation value but also intuition was used for deciding which variables to keep. It was found that there were 73 variables of this type and these variables were dropped from dataset.

HPHONE_D	**HPHONE_D**	1.000000e+00
RAMNT_17	**RAMNT_5**	9.998189e-01
RAMNT_5	**RAMNT_16**	9.982933e-01
HHAGE3	**HHAGE1**	9.939968e-01
HV2	**HV1**	9.934116e-01

Figure 1

Missing Values

Second problem was related with the missing values in the dataset. During the exploration it was found that some of the variables have a high percentage of missing values. For identifying those variables, missing values percentages for each column was analyzed. While checking for percentages of missing values in each feature, it was seen that while some of the features have 23% and lower missing values, on the other side, other variables have 50% or more missing values. It was decided to drop variables with more than 30 percent of missing values, since filling this much of missing values would create redundancy because these variables would be highly correlated with other variables since imputing methods use other variables to impute. There were 37 variables which had missing values more than 30 percent. After dropping the variables with more than 30% of missing values, remaining features were used to impute in order to fill the missing values. K nearest neighbors algorithm was used for this purpose with 15 neighbors.

Feature Engineering

Some of the variables, which were considered important while checking variables, in the dataset had to be engineered in order to be used for clustering purposes. First variable transformed was ODATEDW (date of first gift of donor). This variable was given in the format of date and for clustering purposes it was transformed into recency variable, which indicates the days that have passed after the first gift of donor. In order to find the recency and for accuracy reasons, time of this analysis had to be found. It was assumed that latest date in that column is the date this analysis was supposed to be done. Time difference between latest date and origin date was calculated in terms of days. LASTDATE, MAXADATE and FISTDATE variables are also transformed using same approach. Other than these variables, DOB (date of birth) was also transformed into age variable. In addition to the variables above, SOLIH and SOLP3 variables were transformed. These variables indicate the solicitation limit for a donor. However, they were given in a format of two digits. Each digit was given with a zero before them; for example, if the limit for a donor is 1, it was given in a format of '01' and 2 as '02' etc. In order to be able to use these variables in clustering, zeros were cleaned from the beginning of the values.

Feature Selection

In order to decide which variables to perform the clustering on, all 476 original columns had to be considered. In the initial data exploration, both the metadata and pandas profiling about the dataset were checked in order to choose variables that are numerical and seemingly relevant for clustering. In this process, the following columns were chosen to be further observed as clustering candidates: ODATEDW, INCOME, LASTDATE, MAXADATE, FIRSTDATE, SOLP3, SOLIH, DOB, INCOME, POP901, IC5, NUMPROM, RAMNTALL, NGIFTALL, LASTGIFT, TIMELAG, AVGGIFT, RFA_2F, CARDPM12, NUMPRM12, MALEMILI, MALEVET, VIETVETS, WWIIVETS.

The number of variables was then reduced through elimination. For example, the DOB column, converted into age, had to many missing values (25%) to be properly used, therefore it was not

used for clustering, but it was still used for the analysis of the clusters in the end. The INCOME variable was very informational but could not be used as it was an ordinal variable. Some variables got dropped in the data preprocessing, due to high correlation.

The issue that with this dataset is also the lack of individual data for the donors. First of all, most of the columns hold data related to the donor neighborhoods, rather than to the donor's themselves. The data actually describing the donor's individually is mostly categorical, and hence cannot be used for clustering. The remaining personal data, that is numerical, is mostly incomplete and inconsistent, and would not be suitable for clustering. For these reasons, an assumption had to be made that a donor is accurately described by its neighborhood.

Eventually, the following variables were chosen for clustering: IC5 – per capita income, used as a substitute for income; POP901 – number of persons in a neighborhood; AVGGIFT – average dollar amount of gifts in dollars, and MALEVET, VIETVETS, WWIIVETS – percentage of male, Vietnam, and WWII veterans.

These features were selected as final clustering variables after trying different clustering algorithms with different combinations of the candidate variables, as it was impossible to choose the right columns from such a large number of columns upfront.

Outlier Handling

Another potential problem related to the data is that it might contain outliers. After choosing the important features on which, clustering is done, boxplots were plotted for checking potential outliers.

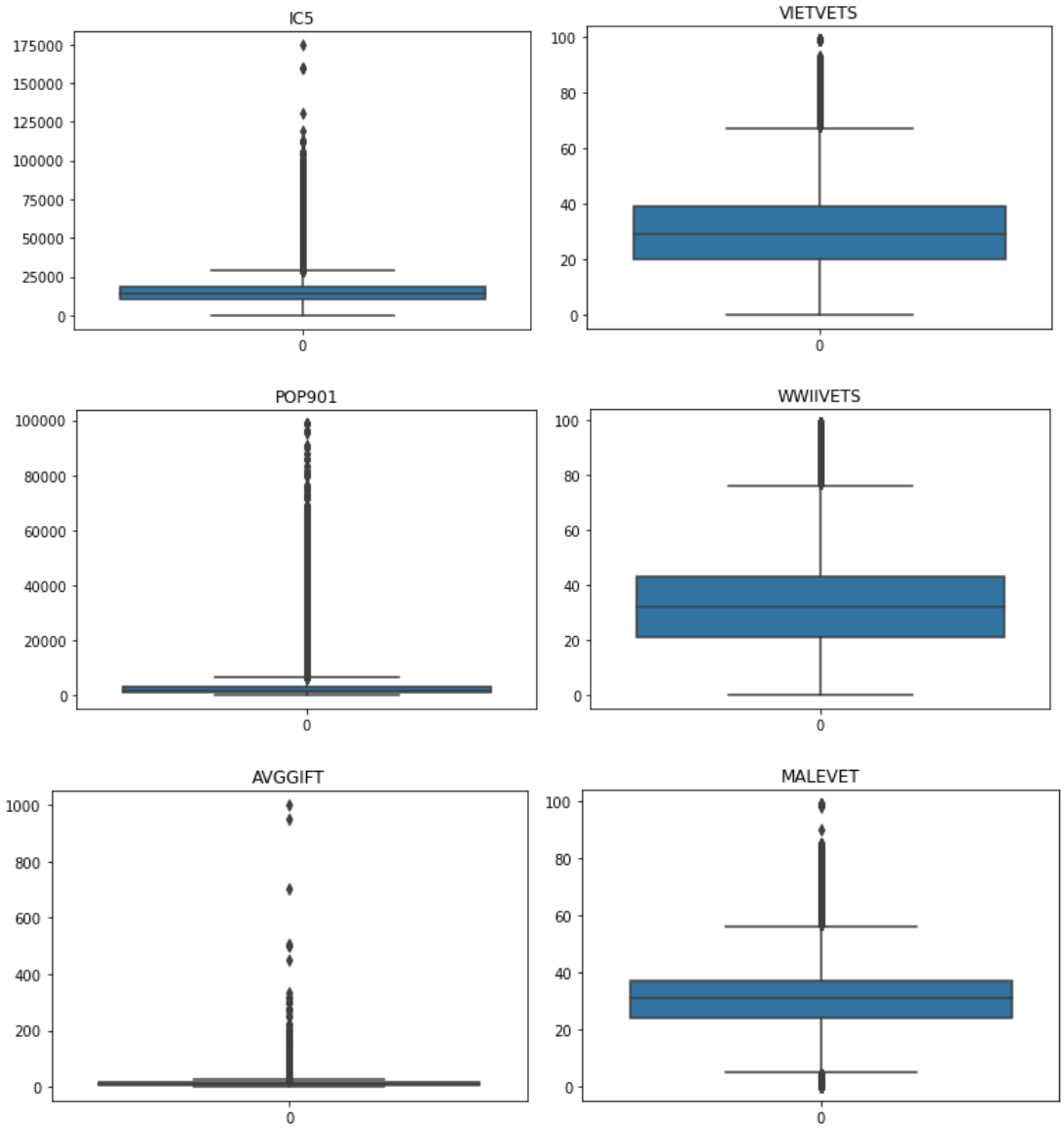

Figure 2

Boxplots are showing some values out of the whiskers, which might be potential outliers. However, when these values are analyzed closely, it can be concluded that these values are not really outliers, they are different customer group. For example, having average income per capita of 175000 for a neighborhood is a possible value. Closer analysis was done on these cases and it was concluded that they present real people and neighborhood. For demonstration purpose, it can be seen from the figure below that it is a neighborhood from California and specific person has a reasonable age value.

	STATE	IC5	ZIP	AGE
82814	CA	174523	90077	84.0

Figure 3

There are neighborhoods where rich people live. Also, for the number of people in a neighborhood can be more than 50000, there might be some neighborhoods which are overcrowded.

Also, for variable AVGGIFT, which is the variable for average gift of a donor till now, donors above 100 cannot be treated as outliers. They represent specific type of donor group which donate more than others and this type of donors are of interest for the purpose of this project.

Remaining three variables are representing percentages and all of these variables are in the range of 0 and 100. It cannot be concluded that if a value is higher than a specific value, then it is an outlier.

Each of the variables have the different scales and this can cause some problems during distance calculations. For eliminating this problem, data had to be scaled. Minmax Scaler was used for this purpose.

Clustering Methods

Self-organizing maps

The first clustering algorithm used on the selected features are self-organizing maps. This approach was chosen as it is good for visualization and detecting possible patterns and segments in the data.

First, component planes were plotted for the selected variables, in order to see if they are suitable for clustering. The result shown in Figure 4 confirms that the variables have a wide enough range of values and variation for different segments to be recognized.

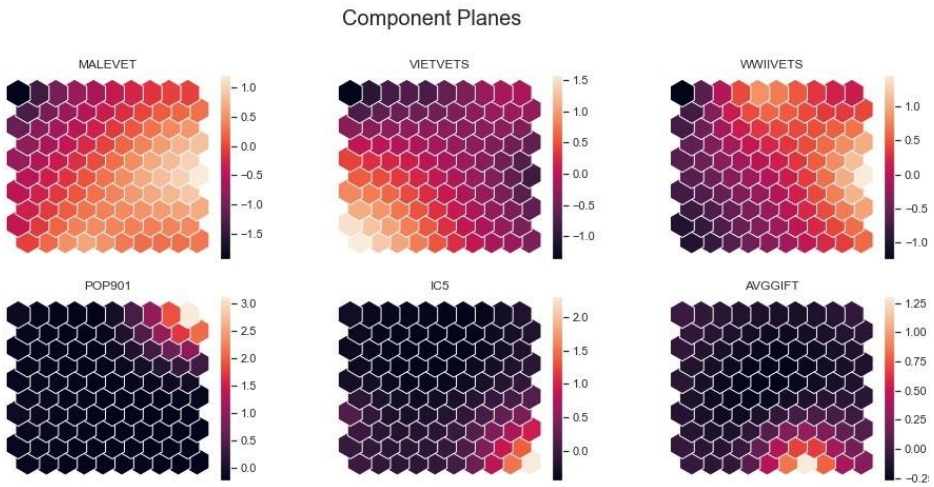

Figure 4

Then, a U-matrix view was created in order to visualize the isomorphic curves that can show the clusters present in the data. Looking at Figure 5, it can be concluded that most of the observations in the dataset belong to the same cluster and that there are no obvious patterns in the data.

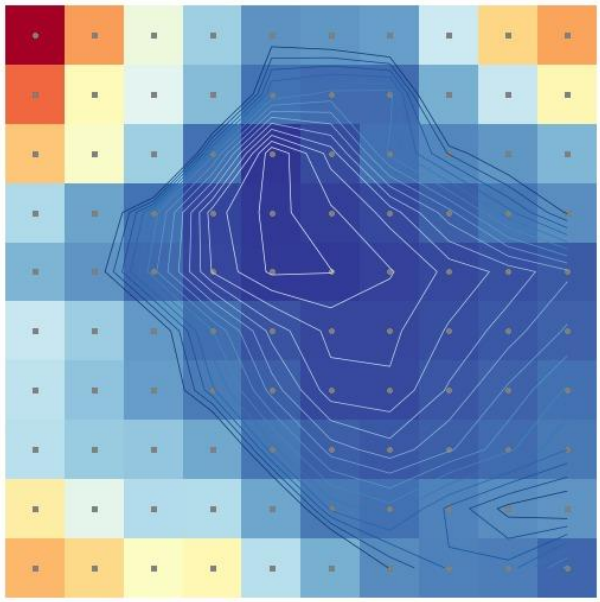

Figure 5

Taking this into consideration, it was concluded that self-organizing maps are not a good approach for the dataset in question. However, this approach showed that the observations in the dataset are quite similar, and data segmentation is not obvious.

Density Based Clustering – DBSCAN

DBSCAN was also implemented on the dataset, mainly for its ability to identify outliers. Because of the conclusion that was made about outliers in this project, DBSCAN was used to either confirm or deny that the outliers identified with outlier detection methods are not really outliers, but rather specific cases.

Choosing an eps value for DBSCAN was difficult to define, because of many columns and observations the 'elbow' plot did not give usable results. For this reason, the default value of eps was used, to predict the number of clusters.

With this eps value, and different variations for the minimum number of points, the DBSCAN algorithm came up with two distinct clusters. However, with DBSCAN identifying all outliers as one cluster, this means that the result of DBSCAN is just one cluster, with the rest of the data being outliers. This supports the assumption that the outliers in the dataset should be taken into consideration, given that they make up a big portion of the data, and do not seem to be actual outliers. Also, with this result, DBSCAN also proved not to be the right choice of algorithm for this project.

K-Means with Hierarchical Clustering

The final clustering approach considered for this problem is k-means clustering. However, with a dataset of this dimension, it is difficult to decide about the number of clusters for k-means. In order to pre-define the number of clusters, the team has decided to use hierarchical clustering dendrogram.

With hierarchical clustering being a bottom-up algorithm and the dataset having 95412 observations, it would be too time-consuming to run it. Instead, k-means was performed first in order to create 100 clusters that are then going to be treated as individual observations. Those 100 observations will then be input for hierarchical clustering.

After performing k-means and creating the 100 clusters, the data frame is then grouped by the cluster labels, and the means of the columns used for clustering are saved in a new data frame. This data frame is then used for hierarchical clustering without a predetermined number of clusters, and then a dendrogram was plotted, as seen in Figure 6.

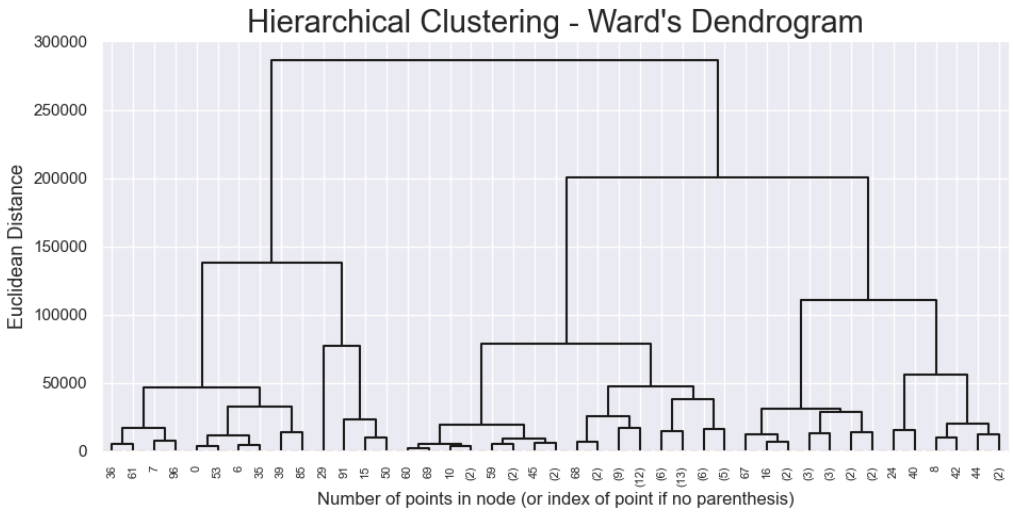

Figure 6

Looking at the dendrogram, three distinct clusters can be spotted. Now, k-means clustering can be done on the original data frame for clustering, with the number of clusters set to 3.

The resulting clusters were then observed based on the means of the columns used for clustering. As seen in Figure 7, there are some noticeable differences between the clusters. Given that the previous solutions offered no clusters at all, the team decided that these clusters were satisfactory.

labels	MALEVET	VIETVETS	WWIIVETS	POP901	IC5	AVGGIFT
0	30.367244	30.102149	32.031173	2426.371964	13421.209472	13.008354
1	27.958008	29.258138	31.037760	28973.649740	14316.350260	14.262936
2	31.718875	26.866218	37.575719	2207.205861	33155.791682	15.604986

Figure 7

Defining and Analyzing the Clusters

In these cluster definitions, both metric and non-metric variables are observed. These variables involve some personal information about individuals like age, donor interests, and economic situation as well as information about their neighborhoods like education level, job sector, economic situation of the neighborhood.

For the non-metric variables, the observation has been done with the mod of the clusters while for the metric variables it has been done with both median and mod values.

For reference, the column on which the observation was made is indicated in parentheses at the end of the sentences.

The modes of some variables can be observed in Figure 8. The metric columns which were used to define and analyze clusters in this part can be observed in Figure 9.

More detailed observations and all the results can be found in the Jupyter notebook.

labels	DOMAIN
0	R2
1	T2
2	S1

labels	INCOME
0	5.0
1	5.0
2	7.0

labels	WEALTH1
0	7.0
1	8.0
2	9.0

Figure 8

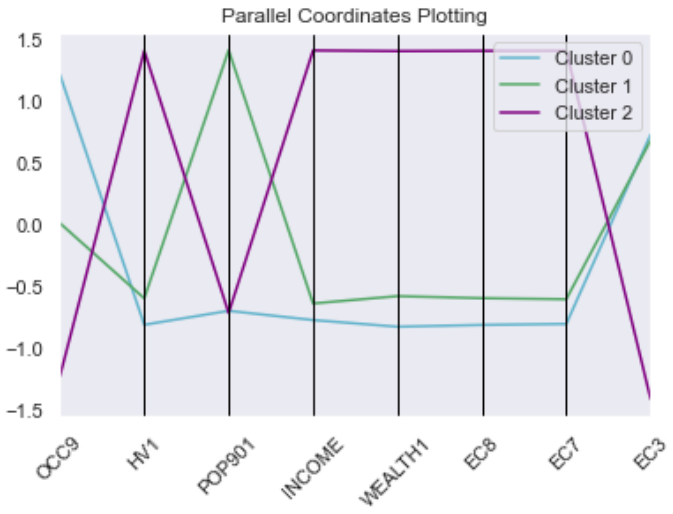

Figure 9

- Cluster 0

The majority of this cluster is living in a rural area with average socioeconomic status (DOMAIN).

This cluster has a higher percentage of people who have gardening hobbies (GARDENIN).

It can be also observed that, in terms of profession, percent “farmers” and “craftsmen, precision, repair” are higher in the neighborhood where this cluster mostly lives in compared to the other clusters (OCC9, OCC10).

In terms of hobbies, this cluster is more interested in the bible than the other clusters (BIBLE).

It was also observed that even though this cluster is not the cluster with the highest income or savings, they had more tendency to donate high amounts of money (RAMNTALL).

- Cluster 1

The majority of this cluster is living in town with a high socioeconomic status (DOMAIN).

This can be also supported with the observation of the number of people, the number of families, and the number of households. These values are significantly higher when compared to other clusters (POP901, POP902, POP903).

This cluster is a younger cluster compared to other clusters (AGE).

It can also be observed in that the neighborhood of this cluster has almost 10% Spanish speaking people while this number is not that high for other clusters (LSC2).

The most significant differences for this cluster can be observed in Figure 10.

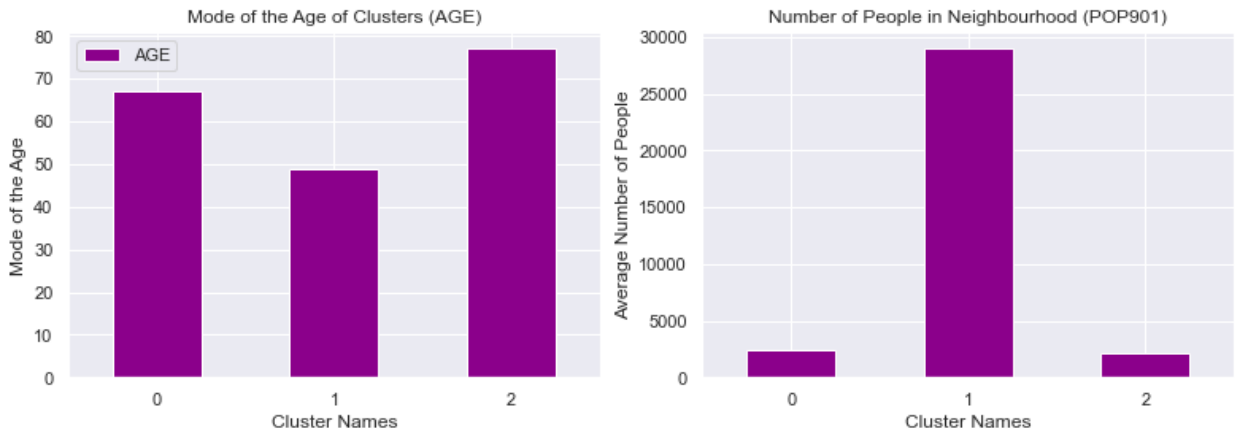

Figure 10

- Cluster 2

The majority of this cluster is living in a suburban area with average socioeconomic status (DOMAIN).

This cluster is a relatively old cluster (AGE). It makes sense that this cluster has a lower interest in kid stuff compared to other clusters (KIDSTUFF).

From an economic point of view, it can be observed that this cluster has the highest level of household income (INCOME). This can be also supported with the wealth rating (WEALTH1).

Besides, these donors are from neighborhoods that have a very low value of percent households with income less than $15,000 compared to other cluster's neighborhoods (IC6). These neighborhoods also have significantly higher percent home value higher than $300,000 compared to other cluster's neighborhoods (HVP6).

In the neighborhoods where this cluster lives mostly, it can be observed that the percent adults (25 years old or higher) with graduate degree and bachelor's degree are significantly higher than the other cluster's neighborhood while the percentage with only high school degree is very low (EC8, EC7, EC3).

This cluster has more percentage of people working from home compared to other clusters (HOMEE). It also makes sense that this cluster has more percentage of people with personal computers (PCOWNERS).

In addition, this cluster has more percentage of CD player owners compared to other clusters (CDPLAY).

The most significant differences for this cluster can be observed in Figure 11.

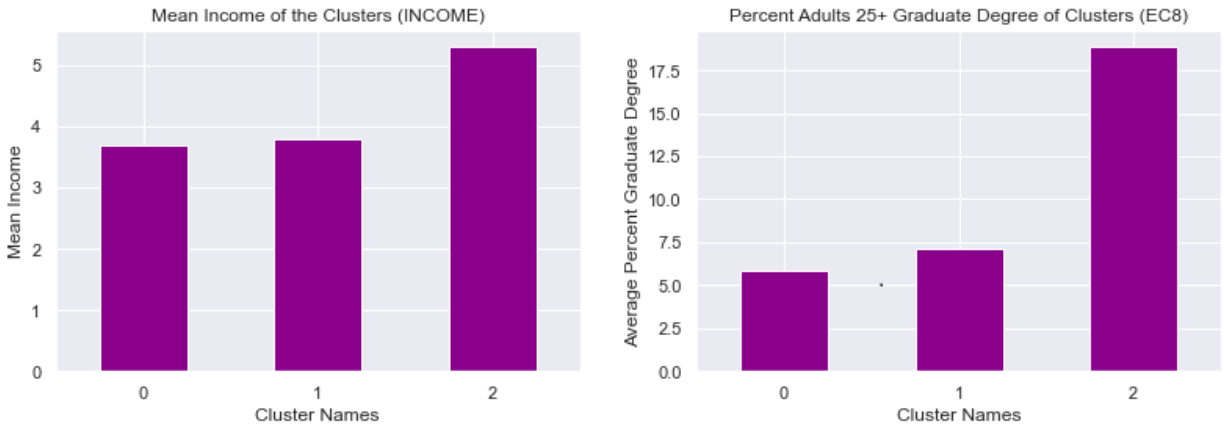

Figure 11

Marketing Approach

Looking at the facts from the cluster analysis part, some interpretations can be done about clusters to create our marketing approach.

In the previous part, facts are given; in this part, some subjective comments will be added to them and marketing approaches will be presented.

All comments will represent a generalization and they will be made as a comparison between clusters.

It can be stated that ***Cluster 0*** has a majority of people living in a rural area with lower income, lower education level, higher percentage of farmers (also more into gardening), repairmen, etc., with a higher interest in religion. This cluster also seems to be less wealthy than the others, but

more prone to donating larger amounts of money, which can mean that they are interested in supporting veterans due to patriotic reasons. A marketing approach that might appeal to this group could be related to patriotic and religious values, or to life in the countryside and nature.

Cluster 1 can be defined as relatively young people living in town with a higher population and with relatively higher Spanish speaking people. One possible marketing approach for this group could be sending offers in Spanish as well as English.

Cluster 2 can be defined as relatively older people living in suburbs (where house prices are higher) with higher wealth, higher education level, and a relatively higher number of people who have PC and working from home. The marketing approach for this group could be focused more on e-mail and digital marketing, as well as presented from a more intellectual angle.

Reference

[1] “DMDSAA, Notebook Solutions”. [Online]. Available: https://github.com/DavidSilva98/DMDSAA/tree/master/notebooks_solutions [Accessed: 04- Jan- 2021].