NOVA IMS Information Management School

Paralyzed Veterans of America

DATA MINING PROJECT REPORT

Segmentation of PVA's "Lapsed" donors

Group BK:

Ana Marta da Silva (m20200971@novaims.unl.pt)

Beatriz Pereira (m20200674@novaims.unl.pt)

Nadine Aldesouky (m20202568@novaims.unl.pt)

Table of Contents

1 Introduction

Paralyzed Veterans of America is a non-profit organization which provides services for US veterans with spinal cord injuries. Additionally, it is one of the largest direct mail fundraisers in America. The organization has recently released a fundraising appeal to 3.5 million donors in 2017. A representative sample of these donors was extracted for further analysis to help PVA improve and better target its fundraising efforts. As a team of Data Mining consultants, we were entrusted with this sample of 95412 lapsed donor records to profile and recommend future marketing campaigns.

Indeed, the purpose of this project is to explore and segment the PVA dataset into different groups of donors. This is in order to understand how donors behave and how the organization can target them to motivate more donations. Specifically, the end goal is for PVA to be able to recuperate these lapsed donors which have not donated for one to two years. As a result, we will begin by pre-processing the data and then grouping the donors into clusters to analyze rigorously for marketing approaches.

The code for this project can be found in this link: https://github.com/AnaMartaSilva/Data-Mining-Project/tree/master.

2 Data Analysis and Preprocessing

2.1 Initial Feature Selection

To decide on a correct clustering approach and the features to use, we started by analysing and understanding the dataset in order to obtain higher quality data. The initial dataset had 95412 records and 474 variables which required processing and cleaning. Since there is an immense number of variables, some of which are incoherent or lacking information, we began eliminating them using logical justification and having the donation problem in mind.

For this, we went through the metadata parallel with pandas profiling to ensure that whichever variables we choose to drop through logical reasoning will not affect our final analysis and lead to losing relevant information. In short, using both these tools, looking at some statistics, distribution graphics, missing values and their definition, 108 variables were dropped. The logical justification for each of the dropped variables is provided in the appendix.

2.2 Feature Extraction and Engineering

In order to reproduce a better analysis, we had to transform the raw data into new features that better represent our business problem. This procedure included numerous transformations and various methods to fill the missing values of each variable. This was not a linear process but rather interactive and iterative. As a result, we will present our processing steps and operations in chronological order:

1. Transform variables from 'X's and blank spaces or 'Y's and blank spaces to 0s and 1s in order to have an integer instead of a string.
2. Drop donors with a bad address, i.e. **MAILCODE** equal to "B", correspondent to 1399 donors (around 1.47% of the data). Then, drop the column since it no longer adds information to the data.
3. Drop donors with variables **POP90C4**, **POP90C5**, **POP901** and **AGE904** commonly equal to zero, which are clearly incorrect data, correspondent to 771 donors (around 0.84% of the data).
4. Drop donors with variables **POP90C1**, **POP90C2** and **POP90C3** commonly equal to zero since they are all complementary, i.e., the sum of these variables should be 100%. Although some records have a sum of 99%

and 101%, this is normal because it was probably caused by the rounding. However, records with 0% are not normal as they do not define or characterize the neighborhood area. These records correspond to 840 donors of the original dataset but only to 34 more donors from the previous step 3. Thus, in this step we will only be dropping an additional 0.036% of the data.

5. Transform previous variables into dummies given their distribution which is mostly values of 0% or 100%, defining the total area of the neighborhood. Note: we chose to drop **POP90C2** because it presents redundant information and transformed the other variables (POP90C1 & POP90C3) into 0's and 1's putting the threshold at 50%.
6. Drop donors with variables **MINRAMNT** or **LASTGIFT** equal to zero, assuming that donations of 0 dollars are incorrect data since PVA clarified that every record represents a donor who has donated at least once. These observations correspond to 584 donors from the original dataset. Hence, in this step we will drop an additional 0.61% of the reduced data resulting from the previous steps. In total, we have now dropped 2.72% of the initial dataset.
7. Fill **GENDER** missing values:
    a. We first tried to do it with **TCODE**, as it contained some titles that allow us to distinguish the donors' gender. Indeed, there were some incoherencies (e.g. TCODE of "Mr." or "Father" with GENDER as "F" - female) and since we do not know the origin of the error, we opted to not do this association given the assumptions we needed to do. Furthermore, the **TCODE** contained a lot of codes with no definition in metadata so this variable was dropped as it is unreliable.
    b. The solution was to replace **GENDER** missing values with "U" (unknown gender). Likewise, we replaced the undefined categories ("C", "A" and "J") with "U".
8. Treat **STATE**'s high cardinality by keeping the eight most frequent states and grouping the rest of the states in a category – "Other".
9. Drop **ZIP** due to its high cardinality and because the geographic location of the donors can already be found in **STATE** and some neighborhood variables.
10. Create **AGE** variable by subtracting DOB (year format) from 2020.
11. Use the **CHILDXX** variables to update the values of **NUMCHILD**. Compare the sum of the values in the CHILDXX variables (with M, F being at least 1 child and B at least 2 children) with the value of the NUMCHLD variable, and choose the maximum value to be the updated value of the NUMCHLD column. This was necessary because these variables present the same information but were inconsistent.
    → Output example:

CONTROLN	CHILD03	CHILD07	CHILD12	CHILD18	NUMCHLD	NEW_NUMCHLD
29552	0	1	1	0	0.0	2.0
51379	0	1	1	0	0.0	2.0
154274	0	0	1	1	0.0	2.0
32798	0	1	0	1	0.0	2.0
163081	0	0	2	0	0.0	2.0
...	...	...	...	...	...	...
176162	0	0	1	1	0.0	2.0
66127	0	1	0	1	0.0	2.0
81332	0	0	0	2	0.0	2.0
102279	0	2	1	0	1.0	3.0
168461	1	0	0	1	0.0	2.0

    With this association, ~97% of the data remained the same but the incoherencies between CHILDXX and NUMCHLD were treated.

12. The variables **CHILD**XX were dropped because of the demonstrated uncertainty of the number of children by each of the age ranges.
13. Fill **SOLP3** and **SOLIH** blanks with two different approaches:

a. If they have a value of X in RECP3 (or RECIH), it would mean they have given in for the program and so we can assume that PVA would send them the maximum number of mails which is 12.
b. If they have a blank in RECP3 (or RECIH), it would mean that they have not given in for the program and so PVA cannot send them any mails so we fill in the value as 0.

14. Associate variables indicating the number of times the donor has responded to a **Mail Offer** with the donors' **Interests** variables. Since the mail offers responses have a lot of missing data, the idea was to use it to fill in the Interests variables then drop it. We assumed that if a donor responds to a mail with a certain topic, he/she is interested in that topic. Thus, we linked the following variables by topic. If the left variables were filled with a value greater than 0 (donor has responded), the right ones would be 1 obligatorily, admitting that the donor has an interest in that topic.

Mail Offers	Interests
MBCRAFT	CRAFTS
MBGARDEN	GARDENIN
PUBCARDN	GARDENIN
MBCOLECT	COLLECT1
MBCOLET	PLATES
PUBPHOTO	PHOTO
MAGFAML	KIDSTUFF

15. Treat dates' missing values from promotion and gifts history and extract useful information from these dates.
    a. **Missing values:**
    FISTDATE had 2 missing values with very high values of **TIMELAG** (Number of months between first and second gift) dropped in the outlier analysis section.

    NEXTDATE and **TIMELAG** had 9790 missing values in common, corresponding to donors who only donated once, i.e, with **FISTDATE** equal to **LASTDATE**. There were also 52 more donors with the same dates for these variables and the same LASTDATE with a TIMELAG of 0. In total, the data had clearly 9842 one-time donors that needed to have the TIMELAG value replaced (with 0 it would not be a good solution for the algorithm since 0 would be considered the lowest time lag and hence the best value to define an active donor, which is opposite what we want). So the solution was to replace it with the maximum value of TIMELAG which was chosen after doing the outlier analysis in section 2.5.

    b. **Feature extraction:**
    1TIME_DONOR binary variable to define donors who only donated once to PVA. Represented by 1 if **FISTDATE** is equal to **LASTDATE** and with 0 otherwise.

    TENURE variable by subtracting **FISTDATE** from 2020, this gives us how long ago the donor started to donate to PVA.

    R_PERIOD (period of recency) variable by subtracting **ADATE_2** (most recent promotion – 17NK) from **LASTDATE**. Given the definition of lapsed donors, it was expected to have these values between 13 and 24 months but the data had around 11.87% of donors out of these boundaries, having the following differences in months:

```
-26    4293
-25    2353
-4      792
-6      732
-5      718
-27     483
-7      413
-8      390
-9      270
-10     196
-12     137
-11     133
Name: R_PERIOD, dtype: int64
```

This variable ended up not being used because of the inconsistency with PVA's lapsed definition, assuming that the variable **RFA_2R** is the correct one (over **LASTDATE**).

c. Drop all dates variables after extracting the relevant information to define the donors' behaviour.

16. RFA (recency/frequency/monetary) fields analysis:
    a. Drop **RFA_2R** since it has equal information for all the donors – "L" – representing lapsed donors.
    b. Transform **MDMAUD** and **RFA_2** frequency (F) and monetary (A) to type integer while staying true to the ordinal meaning described in the metadata (descending order).
17. Treat **DOMAIN** variable by splitting its bytes into **socio_econ_neighbourhood** and **urbanicity_level_neighbourhood**.
    a. Replace blanks with NaN's.
    b. Transform variables into type float.
    c. Drop DOMAIN.
18. Fill **MSA** and **DMA** common missing values – around 13% of data – with mode (most frequent value of each variable), 0 for MSA and 803 for DMA.
19. Treat high cardinality from previous variables, keeping the 3 most frequent values and grouping the rest as a category – "Other".
20. Fill **AGE** missing values based on **NUMCHLD**, calculating the median by number of children, i.e., replacing the missing values according to NUMCHLD values following the next dictionary.

```
NUMCHLD
0.0    66.0
1.0    51.0
2.0    47.0
3.0    45.0
4.0    44.0
5.0    44.0
6.0    47.0
7.0    41.0
Name: AGE, dtype: float64
```

21. Drop **WEALTH1** since it is highly correlated with **WEALTH2** and has a bigger percentage of missing values (47% *vs* 46%), which are all in common with the **INCOME** variable, contrary to WEALTH2 which has values for 51% of the donors without an income defined.
22. Use of the KNN imputer with the 4 nearest neighbors to fill the **INCOME** and **DOMAIN** (socio_econ_neighbourhood and urbanicity_level_neighbourhood) variables. In this step, we selected the most relevant variables to the economic topic: WEALTH2, RFA_2A, MDMAUD_A, AVGGIFT, MAXRAMT, MINRAMNT, IC variables and HVP variables. Since the filled variables are categorical, we needed to round the values to get an integer output representing the categories of each variable.
23. Drop **WEALTH2**, given the number of missing values – 46.09%, that was only kept to help fill some of the INCOME values.

2.3 Feature Transformations

To ensure consistency and to avoid biasing the algorithm, we proceed to normalize the variables and spread them on a common scale using the following methods.

2.3.1 Encoding

Use **One Hot Encoding** to encode the categorical features as dummy variables for STATE, MDMAUD_R, DATASRCE, GENDER, urbanicity_level_neighbourhood, MSA and DMA. The rest of the nonmetric variables which are ordinal, and hence comparable, were transformed into numeric labels (found in the procedure above in section 2.2).

2.3.2 Scaling

Use **Robust Scaler** to scale all the metric features. This scaler was chosen as it does not require a normal distribution of the features and given its advantage of using statistical metrics that are robust to outliers i.e. median and IQR.

2.4 Second Feature Selection

Following this, we did a correlation analysis of the variables in order to determine which variables were redundant and/or irrelevant. The threshold we defined was 90%, meaning that all variables that had a correlation of more than 90% were up for examination. We went through this list of highly correlated variables and using the metadata as well as the business objective in mind, we performed a critical analysis to decide which variables to keep. This is because, although some variables were highly correlated, they were not necessarily redundant as they provided different valuable and relevant information. A detailed table providing reasoning for dropping and keeping specific Census variables that might or might not be correlated can be found in the appendix.

2.5 Outliers Analysis

Previously in section 2.2 (steps 2,3,4, and 5), a small proportion (2.72%) of the data was detected as outliers and consequently dropped. Nonetheless, this percentage will not be considered in this section since they are clearly wrong data, so the following percentages shown are related to the **currently reduced dataset** and not the initial one.

Two types of outliers where analyzed: **univariate** – only regarding to one specific feature, and **multivariate** – dependent on 2 or more features meaning that it can identify abnormal records that appeared to be normal when analyzed separately by looking at only one feature (e.g. having one child would seem normal but having one child with the donor's age as 4 years old would be abnormal).

2.5.1 Univariate Outliers

For the outlier removal by feature, we started applying the **Inter Quartile Range** method on the entire dataset but it resulted in losing too much information, even after increasing the multiplier to 5. Nonetheless, we were able to apply the IQR method on specific columns instead but mostly we opted to perform this filtering manually, based on some visualizations such as histograms and boxplots.

- **AGE** < 16 – only represented 414 donors (around 0.44 % of the data). These records are probably parents donating in their children's name, which can bias the clustering.
- **HIT** > 31 – 997 donors (around 1.07%) and cumulative percentage of 1.50% (including AGE outliers). In this case, we checked the IQR and the distribution through a boxplot and a histogram. Manually, we were able to decide the cut-off value at 31 replies. This means that records with more than 31 replies would be considered as outliers.

- **TIMELAG** > 45 – only 65 donors and cumulative percentage of 1.57% (including all outliers from previous steps). Likewise, we decided the cut-off value for TIMELAG to be 45 months as it is an extreme value of more than 4 years. This is also the minimum value with a frequency of 4 donors so dropping it would not create a bias. Regarding the step 15 in section 2.2, the time lag of a one-time donor will be replaced by 45 months as the maximum value.
- **AGEC** outliers, choosing the filtering manually by their histograms:
  AGEC1> 66%
  AGEC2> 56%
  AGEC3> 48%
  AGEC4> 34%
  AGEC7> 61%
  Only representing 0.27% of the data, and cumulative percentage (including all outliers from previous steps) of 1.83%.

2.5.2 Multivariate Outliers

For the detection of these type of outliers we used the **Extended Isolation Forest** algorithm, a method which in principle is similar to the well-known Random Forest, a tree ensemble method built on the basis of decision trees. This algorithm utilizes the fact that anomalous observations are few and significantly different from "normal" observations (they lie further away from the regular observations in the feature space), that is why by using a random partitioning they should be identified closer to the root of the tree (shorter average path length, i.e., the number of edges an observation must pass in the tree going from the root to the terminal node), with fewer splits necessary, see [3][4][5].
With 300 trees in the forest, the algorithm detected 1336 abnormal records that will be dropped, corresponding to 1.43% of the data, which accumulated to the data dropped from the previous section gives 3.26%.

2.6 Results

Finally, after rigorous analysis and exploration, we have reduced the dataset to 156 **variables** (around 32.8% of the original variables), including the dummies. As for the number of **records**, summing the dropped data from the incorrect values with the previous outliers removal, we have a total of approximately 5.21% data deleted from the original one.

Database	Original	Original 2 (after dropping incorrect records)	Final Data (after removing the outliers)
# of records	95412	93189	89778

3 Clustering

3.1 Perspectives

In this segment we will describe the process of clustering that allowed us to achieve our segmentation objective. Firstly, we decided to research about which factors would influence donations and match those with the variables available in our dataset. From this endeavor we came up with two perspectives to describe PVA's donors. They are as follows:

1. **Donation Behaviour**

This meant focusing on the behaviour of the donors in terms of when they donated, how much they donated as well as their interests/preferences. This would allow us to better understand their attitudes and relation towards PVA as an organization which supports veterans. This approach is more target oriented, meaning we can better recognize lapsed donors and donors with higher potential for reactivation. On the contrary, this method could be too focused on the behaviour of the donors rather than the donors themselves. In other words, we would understand how the donors spend but not who they are as personas.

As a result, these were the selected **metric** variables: HIT which accounts for campaign emails answered and RAMNTALL, NGIFTALL, MINRAMNT, MAXRAMNT, LASTGIFT, TIMELAG and AVGGIFT. As for **nonmetric** variables, these were chosen: RFA_2F, RFA_2A, MDMAUD_F, MDMAUD_A, MDMAUD_R, 1TIMEDONOR as indicators of donors' recency, frequency and amount per donation. Next, we selected the variables indicating the donors' interests: COLLECT1, BIBLE, CATLG, HOMEE, PETS, CDPLAY, STEREO, PCOWNERS, VETERANS, PHOTO, CRAFTS, FISHER, GARDENIN, BOATS, WALKER, KIDSTUFF, CARDS, PLATES. Finally, were the SOLP3, SOLIH variables which describe the donor preferences regarding receiving mails.

2. **Socio-demographic**

This perspective focuses on understanding who the donors are, socially and demographically. Unfortunately, we do not have enough variables that describe the donors themselves, therefore, we will focus on the census variables of 2010 referring to the donors' neighborhood. This is because we assume that where a donor lives indicates who they are and what influences their actions towards donating to PVA.

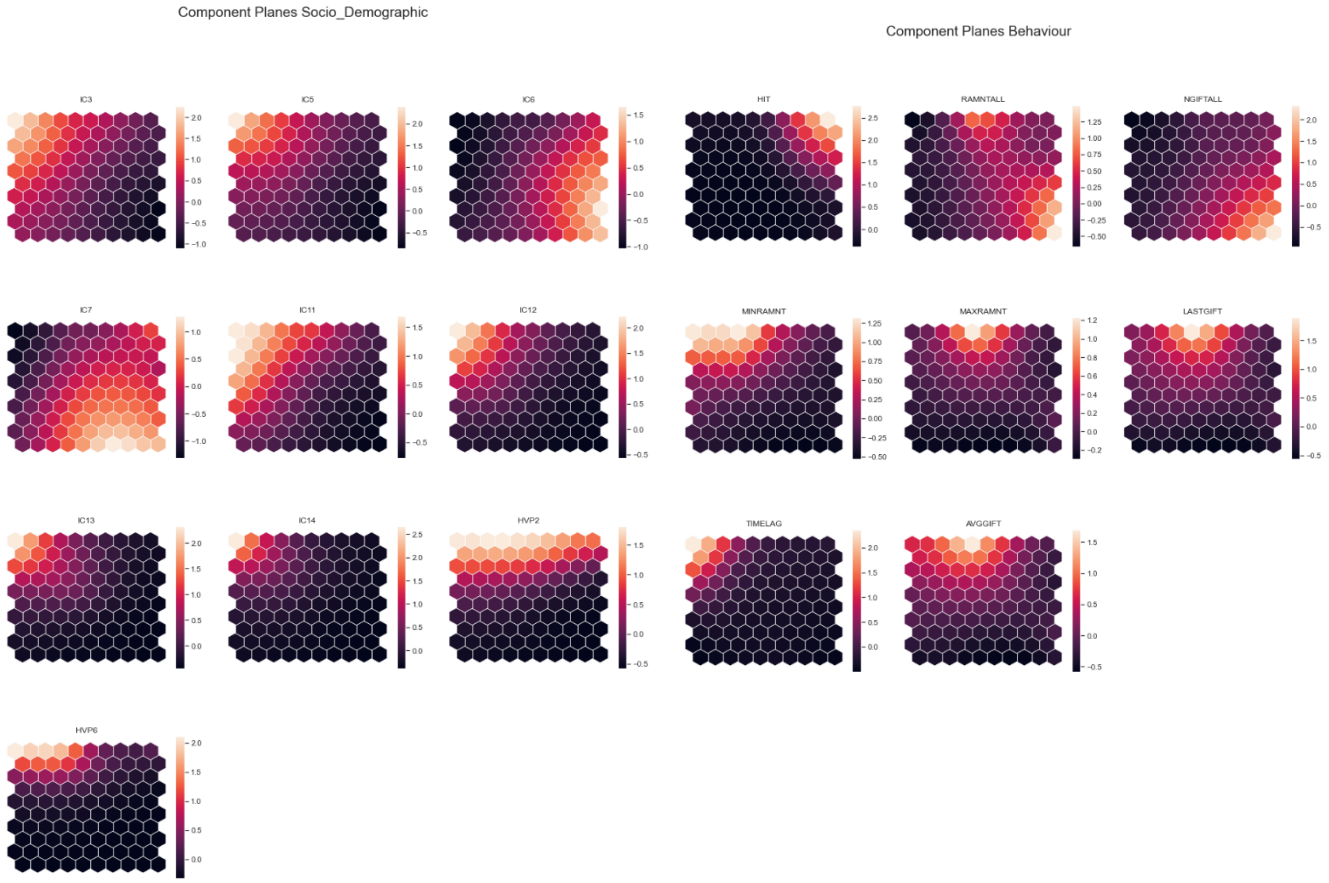
This perspective provides us with more tangible background about the donors, it helps us find out reasons why they may or may not be able to donate. However, this approach tends to be more concentrated on the donors' financial standing which does not fully describe the donor as a person.

The decision for this perspective was supported through the literature review by Chang which uncovered that age, income, gender, education, family loading, and marital status are the top extrinsic factors affecting donation behaviour [1]. This was confirmed by another study by Grohs which emphasized that gender, age, education and income are the most important factors [2].

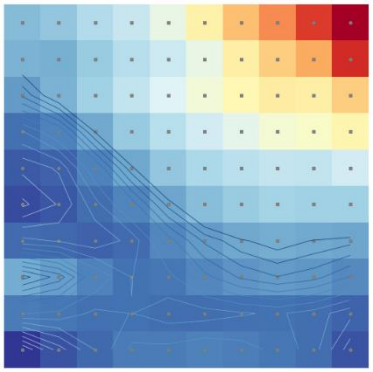
As a result, the **metric** variables selected for this perspective were: AGE, NUMCHLD, variables indicating the percentage of households with a certain income: IC3, IC5, IC6, IC7, IC11, IC12, IC13, C14 and the variables indicating the percentage of homes with a certain value HVP1, HVP2, HVP6. Regarding **nonmetric** features, we selected indicators of location MSA, DMA, STATE, POP90C1, POP90C3, urbanicity_level_neighbourhood and indicators of income rank INCOME, socio_econ_neighbourhood, plus GENDER.

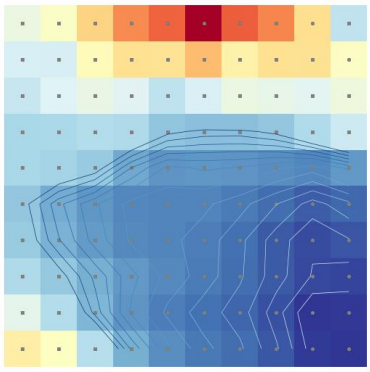
After the definition of variables, we applied the dimensionality reduction technique of PCA to better understand which variables in the socio-demographic perspective were most relevant considering the explained variance. Consequently, we eliminated the variables NUMCHLD and AGE. Then we inspected for high correlation (>= 0.9) between the variables which led us to delete the variable HVP1. This is because it was very correlated to HVP2 and its meaning in terms of range of house value made us prefer the latter.

Following, we proceeded in applying the Self-organizing Maps to visually analyze the importance of our selected features for each perspective. Looking at the **component planes** output below, it is clear that all the variables were discriminatory. However, this method also alerted us to the presence of some extreme values that could bias our clustering. Regarding this, we considered that they were intrinsic to the dataset, to a very diverse donor pool, therefore they should not be eliminated.



While still using the outputs of SOM, we obtained the **U-matrices** which allowed us to see the high-dimensional data in a 2D picture and the potential clusters. The image on the left corresponds to the socio-demographic perspective and the second one to the donation behaviour.





3.2 Techniques

We applied numerous clustering techniques including hierarchical, partition, and density-based clustering. Mean Shift performed well with the socio-demographic option. However, the performance of this technique was very low for the donation behaviour option generating a very low R^2 score. Finally, we attempted to apply K- prototype [6] technique which is a method that capitalizes on the advantages of K-means and K-modes while dealing with mixed data (i.e. numerical and categorical features). We obtained very good results in terms of scores but this method required enormous computing power which was not appropriate for this project as it required iterative adjustments.

We used multiple metrics to compare the performance of the clustering techniques including the aforementioned R^2 score. We also used the Davies-Bouldin index which represents the average similarity between clusters, meaning the average distance between clusters compared with their size. Consequently, results closer to zero indicate better separation between clusters. Finally, we used the Calinski- Harabasz index which corresponds to the ratio of the sum of inter-clusters dispersion and intra-cluster dispersion for all clusters (i.e. dispersion measured by the sum of squared distances). Higher values of this index indicate better defined clusters [7]. Although the K-means method did not have the best scores, we chose it because of its fast implementation which was convenient given the limited resources for this project.

	r2_score	calinski_harabasz_score	davies_bouldin_score
kmeans_socio_demographic	0.337195	22836.003630	1.557394
mean_shift_socio_demographic	0.560782	120211.321397	1.058149
k-prototype_socio_demographic	0.899683	402569.200090	0.496856
kmeans_behaviour	0.261942	7965.266079	1.376607
mean_shift_behaviour	0.017866	7897.784185	1.229242
k-prototype_behaviour	0.562993	28913.467149	0.822276

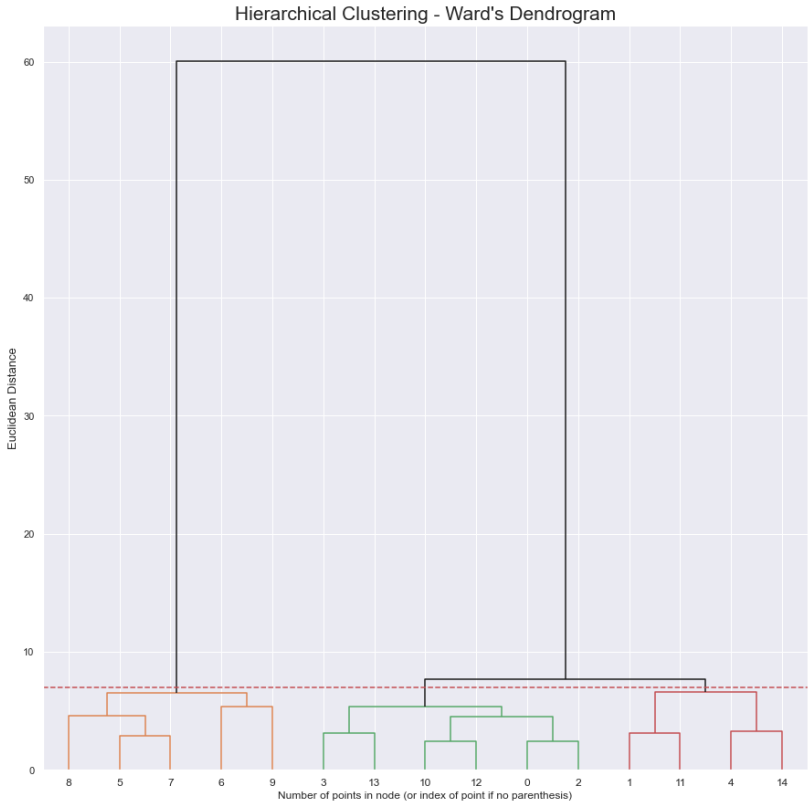
3.3 Final Solution

We finally opted for a combination of the Self Organizing Maps (SOM), K-means, and Hierarchical clustering techniques. The reasoning behind this decision was that SOM allowed for powerful visualization of the clusters. Additionally, K-means had lower computation time and was applied on top of SOM nodes. Finally, the Hierarchical clustering was used to merge the cluster labels from the two perspectives.

Regarding the steps involving our solution, we applied K-means clustering on top of the SOM units for each of the perspectives. In order to decide on the number of clusters, we applied the elbow method based on the inertia concept. From this analysis, we decided on three clusters for the socio-demographic perspective and five clusters for the donation behaviour perspective. We then applied the initialization method of k-means ++ which chooses the initial seeds which are the furthest apart, thereby increasing the probability that the initial centroids belong to different clusters. The result of this clustering method is presented in the table below which shows the number of donors assigned to each cluster by perspective.

behavior_labels	0	1	2	3	4
socio_labels					
0	12769	7943	10662	11878	8100
1	4412	4879	4132	5151	3937
2	3509	2760	3038	3887	2721

Next, we combined both perspectives to have more content rich clusters with higher predictive power. We applied Hierarchical clustering on the centroids resulting from the K-means procedure in order to merge the labels. The algorithm was tuned to join the pairs of clusters that minimize the variance of the clusters being merged (which corresponds to the linkage criterion ward). The distance threshold parameter was set to 0 in order to refrain from limiting the extent to which the clusters can be merged. Later, we created a dendrogram which allowed us to decide on the number of merged clusters to keep. In the end, we decided on three clusters to which we applied the Hierarchical clustering once more to obtain the final cluster labels corresponding to the dataset.



The three final clusters obtained from merging have different sizes. To demonstrate, the third cluster is double the size of each of the other clusters.

	merged_labels	Size
0	[(0, 1), (0, 4), (2, 1), (2, 4)]	21524
1	[(1, 0), (1, 1), (1, 2), (1, 3), (1, 4)]	22511
2	[(0, 0), (0, 2), (0, 3), (2, 0), (2, 2), (2, 3)]	45743

In terms of mean values per metric feature, the obtained clusters present the following results for socio-demographic features:

Cluster	IC3($)	IC5 ($)	IC6(%)	IC7(%)	IC11(%)	IC12(%)	IC13(%)	IC14 (%)	HVP2 (%)	HVP6 (%)
1	334.576	12 880.475	24.817	18.907	3.793	1.101	0.352	0.575	6.077	0.201
2	545.545	23 410.049	12.233	11.556	11.105	5.272	2.444	4.903	64.326	22.145
3	331.107	12 825.806	25.074	19.160	3.629	1.051	0.348	0.568	5.293	0.172
Donors Total Population	385.687	15 492.810	21.792	17.193	5.543	2.121	0.875	1.657	20.283	5.688

Consequently, in terms of socio-demographic characteristics it is possible to conclude that **Cluster 1** and **3** define neighborhoods with very similar income distributions. They are both poorer than Cluster 2 and the average donor in the dataset. However, **Cluster 3** represents a neighborhood slightly poorer as it has a percentage of households with income lower than 24 000$ (IC5 and IC7), higher than the other clusters (approximately 44% in the Cluster 3 compared to 42% in the Cluster 1). On the other hand, on average, clusters 1 and 3 have a percentage of high-class households (with income higher than 125 000$ - IC13 and IC14) lower than 1%.

On contrary, **Cluster 2** is clearly a rich neighborhood as it comprises approximately 7.5% of high-class households, while the percentage of households with income lower that 24 000$ is only around 23%. Besides, this can also be concluded from the percentage of houses with a value above 300 000$ (HVP6) which in this cluster amounts to 22% while in the other clusters it does not reach 0.5%. Indeed, Cluster 2 accounts for a percentage much higher than that of the population which is 5.7%.

Regarding the donation behaviour metrics, the average results per cluster are as follows:

Cluster	HIT	RAMNTALL ($)	NGIFTALL	MINRAMNT ($)	MAXRAMNT ($)	LASTGIFT ($)	TIMELAG	AVGGIT ($)
1	1.034	92.669	4.793	13.628	27.321	24.477	21.740	19.612
2	2.907	106.116	8.844	8.660	21.243	18.561	12.459	14.417
3	3.553	105.524	12.159	4.775	15.352	12.971	6.965	9.518
Donors Total Population	2.787	102.590	9.562	7.872	19.699	17.131	11.885	13.167

We can infer that **Cluster 1** represents donors that donate high amounts to PVA, 19.6$ on average (AVGGIFT), which is much larger than the average of the total donors (13.2$). In terms of frequency (NGIFTALL) they are the ones who have donated the least number of times, on average 4.8 times, which is very low compared to the 9.6 times characterising the average donor in PVA's database.

Contrarily, **Cluster 3** represents donors who donate the most frequently with an average of 12 times. This value is very distinct from the other clusters and the general population, who present lower frequencies of donation. Regarding the amount given per donation, Cluster 3 represents the donors who donate the smallest amounts, with an average donation of 9.5$ (much lower than the average donor).

Cluster 2 represents donors with average donation behaviour which is visible when comparing its values to the last row of the table (aka total donors' population).

To sum up, this merged perspective of clustering allowed for more clear and distinct clusters which allow more actionable and effective marketing efforts.

3.4 Marketing Approach

In this section we will describe each cluster through a SWOT analysis. This is through examining the metric features in the figures above as well as the visualizations of the non-metric features provided in the appendix. Moreover, we will present a list of actionable items for different marketing approaches.

To begin with, we have three clusters which represent different proportions of the dataset. Moreover, in all three clusters, female donors are more prevalent than male donors. Cluster 1 is the smallest but represents our most valuable donors since they gift the most expensive donations. Cluster 3 is our most valuable one since it is the largest. Additionally, this cluster represents the most loyal donors as they gift the largest number of donations. Finally, Cluster 2 speaks for PVA's average donors. There is room for improvement in this cluster in terms of PVA's fundraising efforts. While the donors in all the clusters are geographically dispersed, Cluster 2 seems to have evident presence in Los Angeles and San Francisco-Oak-San Jose.

Final Cluster 1

STRENGTHS	**WEAKNESSES**
• Highest monetary value per donation • Identification with PVA's purpose	• Longest time lag • Least number of gifts • Do not reply to mails • A considerable number of donors in this cluster only donated once
OPPORTUNITIES	**THREATS**
• Potential to donate more times (goal to improve frequency) • Leverage ratio = 0.903	• Lower-middle class neighbourhood

Actions:

- Sending gifting reminders through means other than email.
- Run charity events in order to keep PVA relevant and in the minds of these donors while increasing their 'feel-good' feeling from donating.
- Develop campaigns with more publicity which increase the visibility of the donors (e.g. Wall of Fame in HQ offices of best donors). This is to help donors feel more rewarded by their generosity.
- Create a loyalty program that awards donors points for each donation to finally win a place on the Wall of Fame.

Final Cluster 2

STRENGTHS	**WEAKNESSES**
• Identification with the PVA purpose • Answer mails	• Mails may not be the preferred way of communication

OPPORTUNITIES	THREATS
• Higher class neighbourhood (goal to increase monetary value of gift) • Highest income • Higher amount per donation • Mostly in urbanized areas so easier to reach donors, create events, and spread the word • Leverage ratio = 1.034	• Competition with other charity organizations • Risk of losing interest in PVA due to lack of relevancy and priority

Actions:

- Improve their donation potential in terms of value per donation by appealing to their emotions to entice more impulsive donations.
- Social media campaigns to raise awareness and engagement (for example: Humans of PVA which would tell the stories of the veterans PVA supports) by making PVA more transparent and accessible in the eyes of the donors.
- Refraining from spending too much in targeting these donors as they are already good donors.
- Focus the marketing efforts in the West region (specifically Texas, San Francisco, and Los Angeles) of the US as a considerable number of donors from this cluster reside there.

Final Cluster 3

STRENGTHS	WEAKNESSES
• Lowest time lag • Highest number of donations • High loyalty • Strong identification with the cause • Answer to mails	• Lowest monetary value per donation
OPPORTUNITIES	**THREATS**
• Interest in the topic of Veterans • Potential ambassadors of PVA • Leverage ratio = 1.029	• Poorest neighbourhood • Lower income donors

Actions:

- Organize Meet & Greet events between the veterans and the donors where the donors are encouraged to bring guests. This will be like a fair with presentations and workshops to raise awareness about PVA's mission and achievements.
- Send themed newsletters to remind donors to gift PVA, focus on topics of gardening, children, pets, and crafts since they represent the donors' highest interests.

Priority (time and spend)

i. Cluster **3:** largest cluster and most frequent donors (most loyal) so easiest to recuperate.
ii. Cluster **1:** already attached to PVA's purpose so just need to increase their frequency.
iii. Custer **2:** PVA needs to give them more reasons to donate more money and more frequently.

4 Conclusion

In conclusion, we were able to group PVA's donors into three groups of different sizes. While keeping PVA's objective in mind which is to get back its lapsed donors, we can prioritize the groups in the following ways. Cluster 3 represents the highest priority in terms of urgency and marketing spend as they represent the highest number of donors. Additionally, they have the highest frequency of donations meaning that they are very loyal and connected to the cause so they would be the easiest to recuperate. Next would be Cluster 1 as they are also very attached to the PVA mission and donate the most valuable gifts. This cluster only needs more exposure to PVA which would act as a reminder or a nudge to push them to donate more frequently. In the end is Cluster 2 as they represent the average donor so there is no urgency attached to them. Certainly, they are not as loyal and so they are less prone to coming back to PVA. Moreover, the marketing efforts applied to the other two clusters could have an indirect effect on this group too. The goal here would be any type of improvement or maintenance. Surely, PVA needs to provide more reasons for this group to donate more frequently and more importantly to donate in larger amounts since these specific donors can afford it.

Later, when examining the different clustering views, the donation behaviour outlook seems to provide higher quality clustering. It splits the donors into five distinct and homogenous groups. On the contrary, the socio-demographic perspective generated three clusters which were not very well defined. Thus, when analyzing these clusters through their means we are interpreting biased results because the means themselves are not truthful and quite biased. This results from the fact that the donors are highly spread out. In short, the variables in this perspective have limited discriminatory power which decreases the quality of the clusters. As a result, we decided to combine both perspectives to add more information and descriptive depth to the clusters. Indeed, the more detail about the different donor groups, the more actionable the marketing strategies could be.

Nonetheless, after exhaustive preprocessing, the dataset is still not of the best quality. It includes many observations with extreme values and is very imbalanced. This negatively affects our algorithm as it becomes harder to find coherent patterns within the data.

In the future, PVA would need to invest in collecting higher quality data to extract higher quality knowledge, which is more precise, accurate and actionable.

The code for this project can be found in this link: https://github.com/AnaMartaSilva/Data-Mining-Project/tree/master.

5 References

[1] Chang Chun-Tuan, “Intrinsic or Extrinsic? Determinants Affecting Donation Behaviors” for *Springer Link* (https://link.springer.com/article/10.1057/ijea.2008.2)

[2] Grohs Reinhard, “Increasing Fundraising Efficiency by Segmenting Donors”, for *citeseerx* (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.194.4261&rep=rep1&type=pdf)

[3] Isolation Forest in Python - https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

[4] Lewinson Eryk, “Outlier Detection with Isolation Forest” for *towards data science* (https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e)

[5] Lewinson Eryk, “Outlier Detection with Extended Isolation Forest”, for *towards data science* (https://towardsdatascience.com/outlier-detection-with-extended-isolation-forest-1e248a3fe97b)

[6] Ruberts Antonio, “K-Prototypes - Customer Clustering with Mixed Data Types”, for *Data Science for Marketing* (https://antonsruberts.github.io/kproto-audience/)

[7] Clustering - 2.3. Clustering — scikit-learn 0.24.0 documentation (scikit-learn.org)

6 Appendix

1. Justification of logically dropped variables.

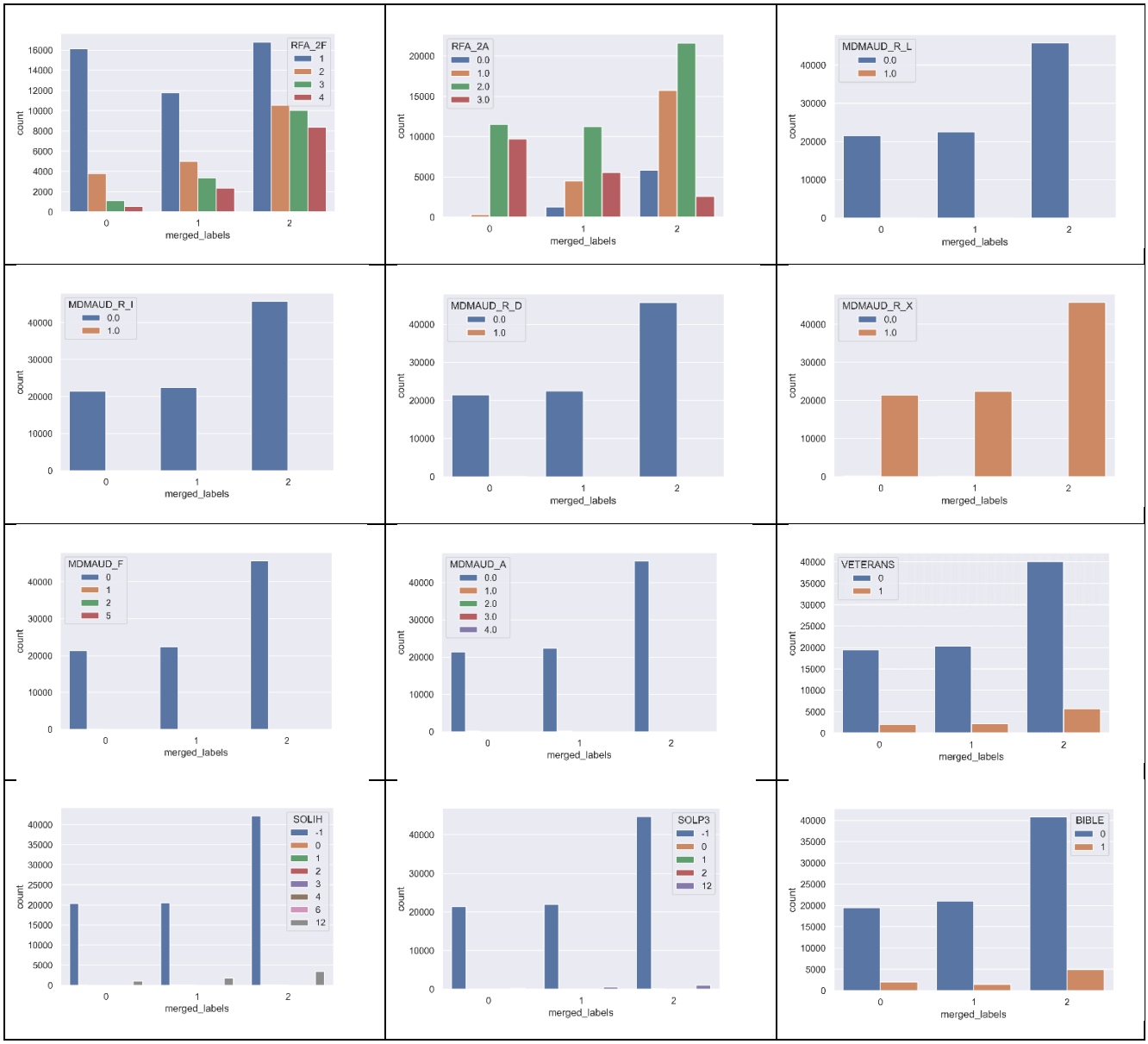
Variables	Foundation	Variables	Foundation
OSOURCE	High cardinality, no predominant mailing list → will have no power to cluster the donors. Not enough information to understand the variable and DATASRCE can substitute for it.	MALEMILI, MALEVET, VIETVETS, WWIIVETS, LOCALGOV, STATEGOV, FEDGOV	Not enough information to interpret its meaning (% of what?) Redundant with census data.
HOMEOWNR	Lots of unknown data.	LIFESRC	Irrelevant because we are interested in the donors' interests so which donor is interested in what not where the interests come from (mailing company).
GEOCODE, GEOCODE2	84% of records do not have a code. STATE and ZIP already provide geographic information so these variables are redundant and unnecessary.	PVASTATE	98.5% missing values.
HPHONE_D	Irrelevant to potential for donating or not.	NOEXCH	Description in metadata not coherent with values and not enough information to interpret the meaning of the values.
MAJOR	Redundant with MDMAUD.	RFA_2, MDMAUD	Both variables already split by bytes R, F and A.
ADATE, RFA, RDATE, RAMNT (3 -24)	Outdated data. Too far back to characterize the same donor who can already have a different behaviour.	CARDGIFT	Redundant with NGIFTALL (and NGIFTALL is more informative since it includes all types of promotions).
CARDPROM	Redundant with NUMPROM (and NUMPROM is more informative since it includes all types of promotions). We don't want to lose the information about the donors who did not receive a card promotions.		

2. Justification of census dropped variables.

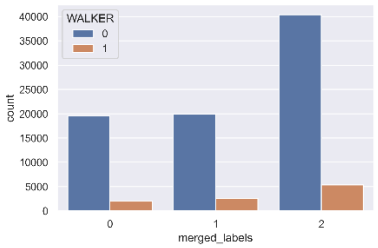
Variables	Foundation	Variables	Foundation
POP902, POP903	Redundant with POP901.	HC(s)	Drop all HC except HC15 (solar energy are environmentally friendly...aka culture of giving back).
ETHC(s)	Irrelevant since the ages of the races of people in the neighbourhood will not affect willingness of donor to donate, also we already have the info on how ethnicity of people living in the neighbourhood.	MHUC1, MHUC2	Lacking unit information.
ETH13, ETH14, ETH15, ETH16, LSC (s)	All the language variables are irrelevant.	EC1	Description in metadata not coherent with values and not enough information to make assumptions to better understand it.
AGE901, AGE902, AGE903, AGE905, AGE906	AGEC variables are more informative and disaggregated so we keep them instead.	AFC(s), VC(s)	Drop all AFCs and VCs except AFC1 and AFC4 because a neighbourhood with veterans & active military can be more prone to donating to PVA.
HHD1	Redundant with AGE907.	HHAS(s)	Redundant with IC6. (demonstrate a 'poor' neighbourhood which is also represented in IC6)
CHIL(s)	Drop the CHIL variables since AGE907 sums them up.	HU1, HU3, HU4	Drop all HU variables except HU2 (correlated with HU1) because it can tell us the probability that the donor is a renter or not (if he/she is a renter means they are mobile/unstable (lower family loading) but also with more bills to pay) Keep HU5 to know if the neighbourhood is only a holiday one so only occasionally inhabited.
HV(s)	Outdated because it is a monetary value and the census was created in 2010 which means the money will not have the same value today. HVP variables are a better substitute.	IC(s)	Drop all IC variables except IC3, IC5 because average, household and per capita are more meaningful than median (especially when they have the same distributions) and families.
RP(s)	Unnecessary because HVP variables tell us more about the donor.	IC15-IC23	Since we will focus on households rather than families' income.
TPE(s), LFC(s), OCC(s), EIC(s), VC(s), HUR(s), ADI(s), MC(s), CHILC(s), HHAGE(s), MARR(s), DW(s), HUPA(s), HHD(s), HHN(s),	Irrelevant to the donors' decision to donate.	SEC3, SEC4, SEC5	Keep SEC1 and SEC2 because there is a cultural distinction between people attending public *vs* private schools, which can affect their moral values leading to donation decision. Can drop the rest of SEC since it does not provide enough information.

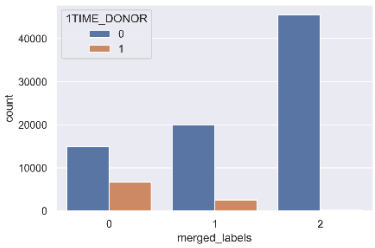
RHP(s), HHP1, HHP2			
HVP3, HVP4, HVP5	Correlated with the other HVP. These three do not define the neighbourhood richness so well, including mixed neighbourhoods with very different home values.	POP90C4	Redundant with female percentage (POP90C5). Together they make 100%.

3. Distribution of non-metric features for Donation Behaviour perspective by cluster.

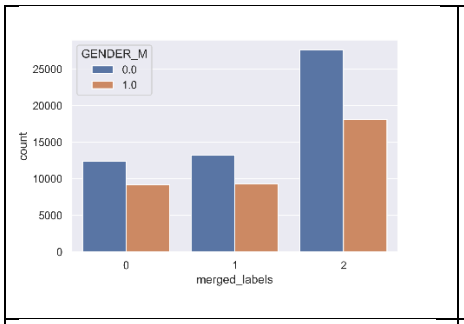


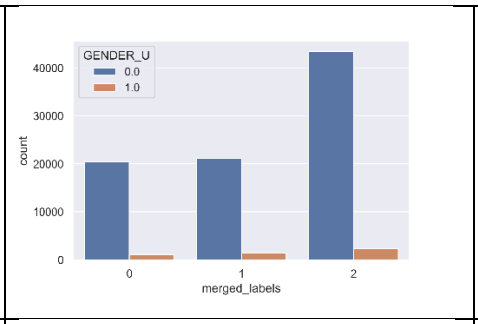

BOATS
CARDS
CATLG
CDPLAY
COLLECT1
CRAFTS
FISHER
GARDENIN
HOMEE
KIDSTUFF
PCOWNERS
PETS
PHOTO
PLATES
STEREO
0
1
count
merged_labels

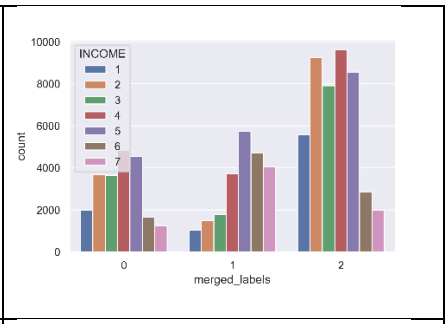


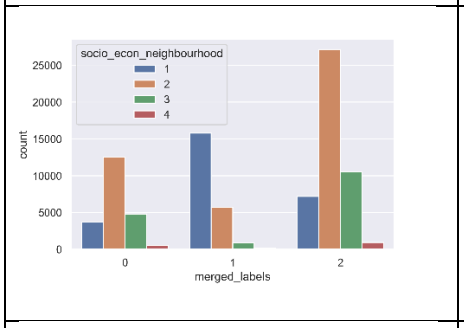


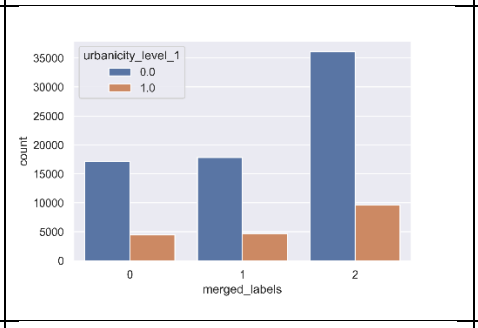
4. Distribution of non-metric features for Socio-demographic perspective by cluster.

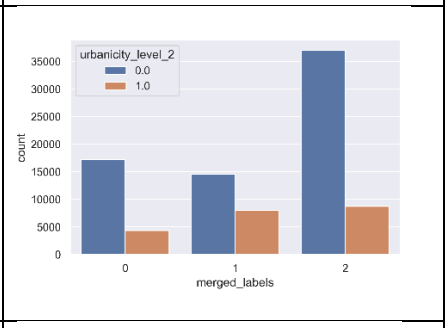


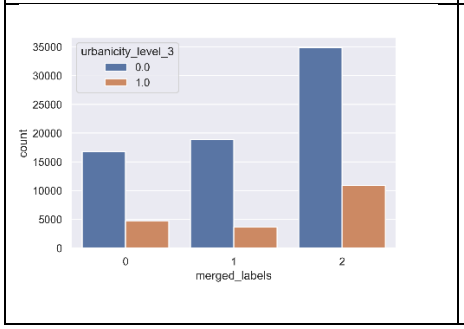


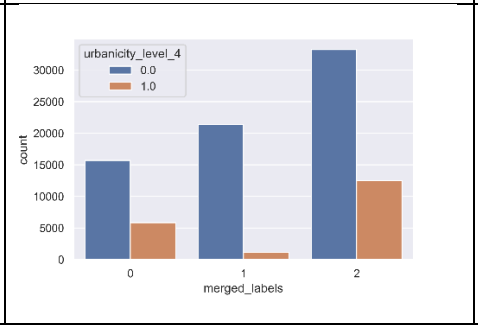


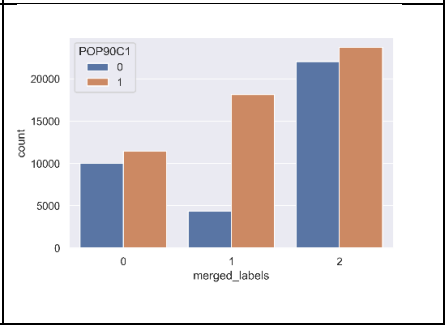


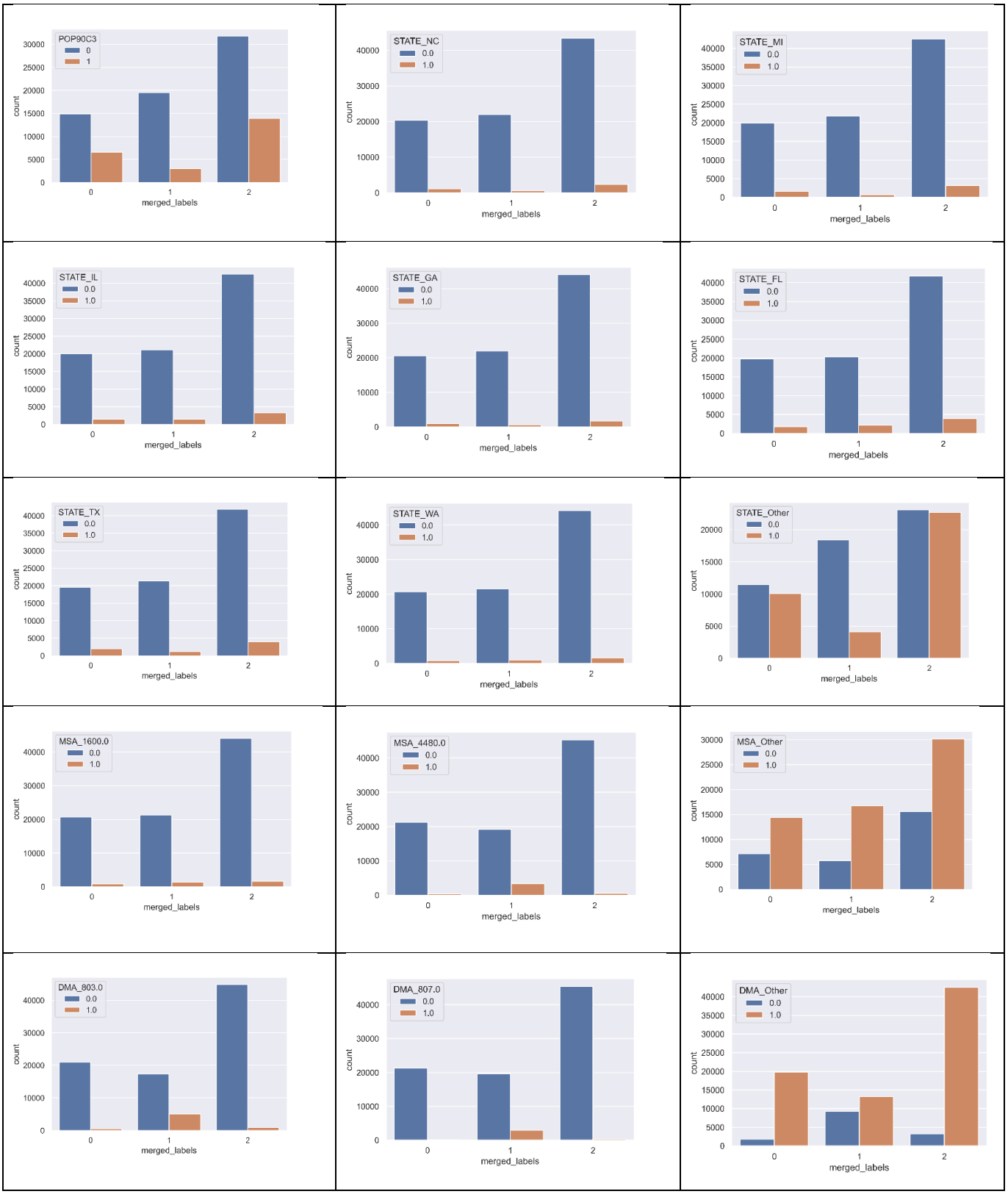











POP90C3
0
1
STATE_NC
0.0
1.0
STATE_MI
0.0
1.0
STATE_IL
0.0
1.0
STATE_GA
0.0
1.0
STATE_FL
0.0
1.0
STATE_TX
0.0
1.0
STATE_WA
0.0
1.0
STATE_Other
0.0
1.0
MSA_1600.0
0.0
1.0
MSA_4480.0
0.0
1.0
MSA_Other
0.0
1.0
DMA_803.0
0.0
1.0
DMA_807.0
0.0
1.0
DMA_Other
0.0
1.0
count
merged_labels
0
1
2