Clustering Donors of Paralyzed Veterans of America

Group E:

Mario Rodríguez Ibáñez – M20200668

Tiago Jose Isidoro Ramos – M20200613

Abstract:

A Client Segmentation has been requested by the non-profit organization Paralyzed Veterans of American, which provided a dataset containing over 400 features related to the giving history of the donors, promotions they have received and their answer to them, information of the neighborhood where they reside, individual demographic data, interests, etc. For the purpose of this project a CRISP-DM process was followed. An extensive data preparation was needed, including data cleansing, feature selection, feature engineering, among other necessary steps. In a later modelling stage an RFM analysis was performed. K-means and Hierarchical clustering were also applied in combination on two subsets of features: giving history and census data. This implied the difficulties of analyzing separated clustering results and the advantage of having clearer donor profiles.

Interpretable clusters were obtained and some marketing approaches were suggested based on them.

Introduction:

Motivation

The team was asked by Paralyzed Veterans of America (PVA) to segmentate their donors in groups, in such a way that the organization could better understand their behavior, and later apply targeted marketing approaches.

To this purpose, the team was supplied with a sample of PVA's database of lapsed donors - a group of special interest to the organization-. This dataset contains 95412 records, and 475 features.

Due to the high variety in the data entries of the PVA's database, the first step was to select and clean the features to be used. Then, new variables were engineered. The data preparation process finished with a selection of original and engineered features and a proper scalation of them. For the purpose of Client Segmentation, diverse clustering methods were tested, such as RFM analysis, K-means, Hierarchical Clustering, or combinations of all.

Interpretable clusters were found from two perspectives: giving history and census data. Using a contingency table, it was possible to combine the findings in both of them to the suggest different marketing approaches.

Background

The project followed a process based on CRISP DM (Cross Industry Standard Process for Data Mining) (Azevedo et al, 2008). The process is based on six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This framework expects an iterative back and forth between these steps, as knowledge regarding each stage merits a revisit of another, allowing for more robust results.

Dython is a set of data analysis tools in Python. The Associations tool was used for this project, due to its capabilities to create visualizations of measures of association between mixed data types. For this purpose, it applies and computes different measures between each variable type pair: *"Pearson's R for continuous-continuous cases, Correlation Ratio for categorical-continuous cases, Cramer's V or Theil's U for categorical-categorical cases"* (Theil U in the context of this project).

Methodology:

Business Understanding. Given the aim of the project, information regarding the operation of the organization was collected and stored. In addition to the information provided in the project briefing, online resources were used to complement and enrich the team's domain knowledge.

Data Understanding. The provided dataset consisted of 95412 records and 475 different features. Given the high number of variables an initial feature selection was made based on the available information (documentation and Data Frame visualizations) and the outputs of the business understanding phase.

Variables were grouped according to their nature, potential importance, and subject matter. Each group was then further explored and analyzed with a range of different inspection and statistical techniques (e.g., visualizing descriptive statistics, heatmaps, correlation matrices). When relevant, different preprocessing techniques were applied to correct missing or invalid datapoints. Once all variables were analyzed, a narrower selection was done. The criteria for said selection was based on the presence of outliers, missing / invalid values, relevance to the subject matter, and correlation with other features from the same variable group.

This partitioned analysis generated a filtered variable group, which was re-analyzed in conjunction with engineered features, utilizing a similar range of inspection techniques.

Data Preparation. Some preprocessing steps were done in tandem with data exploration, as necessitated for a meaningful analysis of the respective subset (specifically, cases of missing values, incorrectly specified data, and / or formatting errors).

Due to the inconsistency regarding the presence of missing / invalid values in the dataset, heuristics were applied on a group-by-group basis to identify these cases (i.e., depending on the feature, missing values could present as 0, not-a-number (NaN), or blank spaces).

To correct these cases, several different techniques were applied, depending on the severity and impact of the issue: filling data with a highly correlated variable or proxy variable, predicting missing values, using a central tendency measure, heuristics, or a combination of these. Features that still suffered from severe quality issues were discarded from further analyses.

Feature Engineering. Four features were engineered from the available data:

- RECENCY: Defined as the number of days since the last recorded donation. The feature was engineered by computing the timedelta between the defined current year (2018) and the values in LASTDATE feature. While the dataset provides an explicit Recency

set of features, these are categorical and dependent on another categorical feature, hence the need to engineer a more reliable and metric feature.

- AGE: Age is engineered based on DOB feature (date of birth), computing the time difference with the current year.
- ANTIQUITY: Defined as the amount of time someone has been a donor for. According to the documentation, two features could be used to engineer this feature, both representing the date of the first donation, ODATEDW or FISTDATE. However, data between them is not consistent, and FISTDATE was chosen as the base. Further reasoning for this decision can be found in appendix A.
- PROM_TIMELAG: Defined as the average amount of time between sent promotions, in days. The feature was engineered by taking the dates from the feature set ADATE_ and computing the average timedelta from that set for every record.

Scaling. Three different methods were applied and tested for scaling metric features: MinMax Scaling, Robust Scaling, and Standard Scaling. Because most variable distributions were non-normal, very skewed with long tails and / or kurtosis, and in some cases presented extreme ranges of values, Robust scaler was preferred. This method subtracts the median and then divides by the IQR, in such a way that the final range of values is not defined a priori. Because it uses the median instead of the mean, it is more flexible and robust against outliers and not normal distributions.

Outliers. Given the observed distributions of the variables, the approach to outlier removal was very conservative, opting to apply heuristic and manually defined filters to the data (a description of the outlier filters applied can be find in the appendix B). The IQR method for outlier removal was tested, but it is ill-suited for this dataset, as it assumes the underlying distributions to be normal or close-to-normal.

Feature sets. Techniques were applied to the filtered feature set, allowing us to draw further insight from its variables and their relationships. An Associations table was constructed for all types of variables, and a correlation matrix was computed for metric features. Additionally, we applied Principal Component Analysis and Self-organizing maps to reduce the dimensionality of the data, and their behavior regarding the dataset and each other.

Two smaller feature sets were then constructed based on these findings to apply clustering solutions: one regarding giving behavior and another regarding census data.

As the census data pertains to neighborhood metrics from 2010, its usefulness could be limited by the volatility of some features (a lot can change in 7 years), and any interpretations made

with the data must take both these factors into consideration. These points notwithstanding, some relevant insights can still be derived, because:

1. characteristics of a neighborhood are reflective of their household constituents;
2. considering PVA's chosen communication channel (mailing), and assuming a donor's behavior is like its neighborhood in this regard, being able to establish meaningful links between a donor and its neighborhood has tremendous informative power.

Modeling. Given the subject matter and the focus on giving history, an RFM analysis was applied. It is a low-cost high potential analysis segmentation strategy, that permits an a priori segmentation of customers. It is commonly used in database and direct marketing, which suits PVA's marketing objectives.

As the interest was on lifetime metrics (because there was special concern in analyzing lapsed donors), the three pillars of RFM were defined as follows:

- RECENCY: the time since the last donation. Given by the engineered RECENCY.
- FREQUENCY: the amount of donations. Given by NGIFTALL.
- MONETARY: the average amount per donation. Given by AVGGIFT.

For ease of interpretability and consistency, independent binning was used (i.e., the three areas are binned independently of each other). Bins were created by quantiles, ensuring all held roughly the same number of records. Thresholds and bin edges were derived from these constraints, resulting in 3 bins per feature, for a total of 27 categories.

Following this, K-means and Hierarchical clustering algorithms were tested in parallel, with the different feature sets defined previously. To assess the correct configuration of the methods several approaches were taken. First the R^2 scores of K-means and Hierarchical with 4 kinds of linkage (complete, average, single, and ward) were compared for two to ten clusters. K-means consistently obtained the best score, followed by Hierarchical with ward linkage, for any cluster number (appendix C). Then the resulting dendrogram from the Hierarchical clustering was plotted to graphically decide a convenient number of clusters (appendix D). After that, the inertia of the K-means in a range of cluster number was plotted to visualize a feasible number of clusters (appendix E). The results of the inertia plot and the dendrogram were then compared to assess the more correct cluster number. Finally, with that configuration the K-means algorithm was run to cluster the records. Additionally, the silhouette scores were computed and analyzed (appendix F).

For this clustering two feature sets were utilized: giving history and census data. The advantage of clustering with two separated feature sets is that clear profiles can be identified more easily than with a single feature set. The counter point is that there is no direct connection between the results. To solve this issue a frequency table of both clustering sets was utilized, being then possible to link some clusters together.

Evaluation. The cluster solutions were evaluated by their interpretability, as visualized by comparison line charts of their centroids. Silhouette scores were also computed to address suitable number of clusters for K-means. A TSNE visualization was implemented to show the final clustering solutions.

Deployment. Combining the results of the different clustering methods some marketing approaches were suggested.

Results and Discussion:

Business Understanding. The collected knowledge regarding both the organization and the aim of the project suggested a greater emphasis on identifying homogeneous clusters of datapoints, as well as preferring to do so with features closely related to donor's lifetime giving behavior. This knowledge further motivated the team to analyze potential links between these clusters and other included categorical features.

Data Understanding and Preparation. As iteration between both stages was extensive, the most relevant and impactful findings of these are presented together.

The following groups of variables were considered, explored, and analyzed: Datetime, RFAs, giving history, census, household, major donors, and mailings.

Engineered Features. For Age, about 25% of records (23883) were missing date of birth, and records were found of children donors.

Also, 19 people under 6 years old donated. This is likely not invalid data per se; maybe their parents donated in their name. But it does mean their usefulness for clustering is marginal at best. Therefore, consideration must be had regarding including Age because of a) its poor correlation with other variables, b) its lack of interpretability and attributability to giving behavior, c) the unpredictability of its missing values. Some supervised algorithms were applied, with extremely poor results.

After the analysis and consideration of several groups of variables, a group of about 39 variables remained. These were further divided into metric and categorical features. In Fig. 1 it is possible to visualize the computed correlation matrix, after outlier removal:

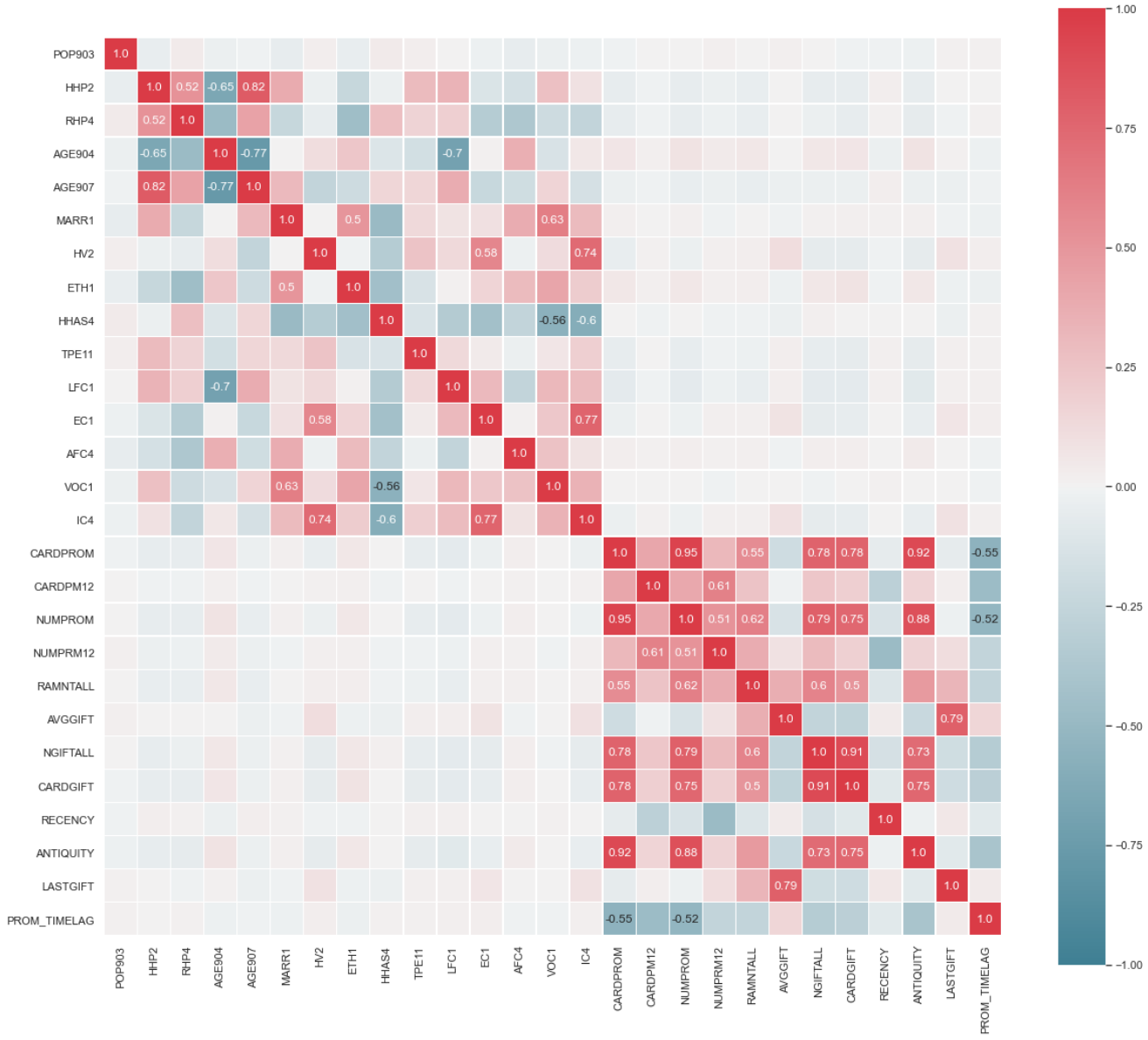

Fig 1: Correlation matrix of selected features

Several observations could be made regarding the selected variables:

- The features formed two groups of correlated variables, which did not interact strongly between one another – one group was composed of census variables, the other of giving and promotion history related variables.

- Several features showed no significant correlation to any other variable (e.g., POP903, TPE11, AFC4).
- In contrast, several highly correlated features were also identified. (e.g., CARD variables and NUM variables; ANTIQUITY and NUMPROMS.).

Based on these observations, domain knowledge, and previous exploration, the feature set was further narrowed. POP903, AFC4 and TPE11 were dropped – in addition to poor correlation, their distributions or nature relayed poor information (for example, AFC4 was almost uniformly 0). AVGGIFT and RECENCY were kept, given their relevance to the subject matter. NUMPRM12 and CARDPM12 were dropped because PVA was concerned with lifetime metrics, and because the respective impact is non-measurable in this dataset (i.e., all records are lapsed donors for at least 334 days; from there it follows that the amount of promotions sent in the previous 12 months to these donors had no effect on their present giving behavior). NUM variables were also kept instead of CARD variables, being that 0.9 of all promotions were card promotions, and no specific impact in giving behavior was detected when distinguishing between the two formats.

The feature sets to be clustered were defined as thus:

Giving History Set		**Census (Neighborhood) Set**	
Name	*Description*	*Name*	*Description*
RECENCY	Number of days since most recent donation.	HHP2	Average person per household.
AVGGIFT	The average donation amount, in dollars.	RHP4	Average number of persons per room.
RAMNTALL	The lifetime donation amount, in dollars.	AGE904	Average age of population.
NGIFTALL	The number of lifetime donations made.	AGE907	Percent of population under the age of 18.
NUMPROM	Number of mailing promotions sent to this record.	MARR1	Percent of married people.
LASTGIFT	Amount of most recent donation, in dollars.	LFC1	Percent of adults in labor force.
PROM_TIMELAG	Average period between promotions, in days.	EC1	Median years of education completed by adults over 25 years old.
		HV2	Average home value in hundreds of dollars.
		IC4	Average family income in hundreds of dollars.
		ETH1	Percent of people of white ethnicity.
		HHAS4	Percent of people below poverty level.

The two sets reflect the relationship and nature of features identified, capturing the most significant information for interpretation and characterization to be meaningful and actionable

to PVA. Furthermore, it allows for a clear separation of individual household / donor behavior and their respective neighborhoods.

At this point, it is relevant to address the lack of individual demographic variables (such as AGE, GENDER, or INCOME) as a feature set. While an effort was made to include metric features capturing this information in the clustering (including engineering AGE), the quality issues of these features prohibited its use. A further exploration of this issue can be found in appendix G.

RFM. Binning of the three features used produced the following quantiles:

RECENCY			**FREQUENCY**			**MONETARY**		
Quantiles	*Bin Range*	*Records*	*Quantiles*	*Bin Range*	*Records*	*Quantiles*	*Bin Range*	*Records*
1	[1037,792[	29150	1	[1,4[	31853	1	[1.28,9.44[	31358
2	[792,731[	27954	2	[4,11[	33186	2	[9.44,14.50[	31392
3	[731,334]	36955	3	[11,50]	29020	3	[14.50,100[	31309

A quick glance of this table corroborates some of the findings from the data understanding phase, regarding the skewness of the data and some extreme values. To evenly divide the number of records, some bin ranges were more compact (Recency 2 and Recency 1 being the starkest example), while others were excessively larger (Monetary 3). It must be noted that these values originated from the filtered dataset, after the outlier removal step – the values would be even more extreme otherwise. Two more notes must be made:

First, the highest number of days in recency (close to 3 years) was surprising given that the dataset was composed of only lapsed donors and ANTIQUITY's 3rd quantile is 9 years. If PVA was not automatically deleting records which had not donated in more than 1037 days, that implies an average donation expectation of at least once every three years. (A visualization of this effect in relation to these clusters can be seen in the appendix H.)

Second, the donation from a quantile 3 donor could potentially match the amount of a thousand donors from quantile 1 or be more than hundred times more frequent.

Following RFM practice, the Fig. 2 shows the segment counts:

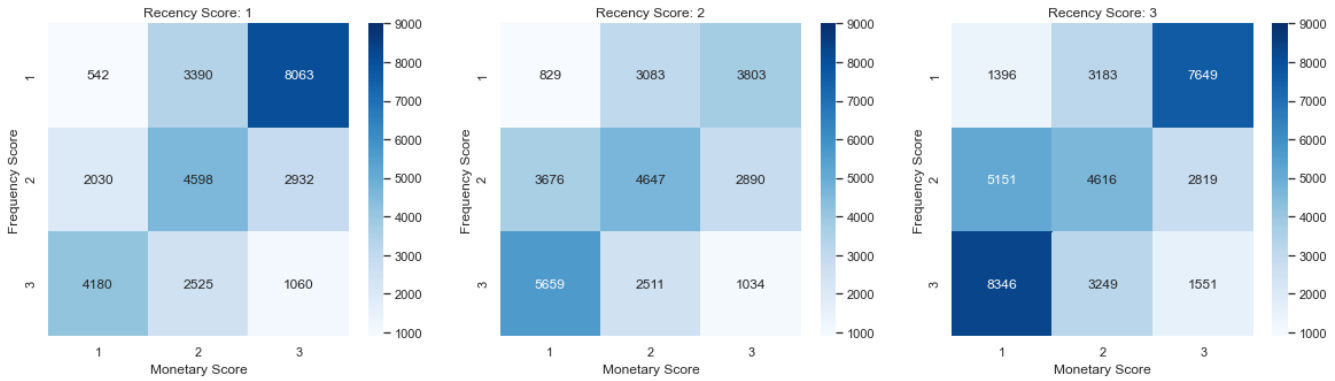

Fig 2: Count of RFM combined categories

The edges of the squares were generally the most populated ones, which indicated a skew towards either the highest level of frequency or the highest level of monetary value. The lower right edge, (best monetary and frequency value), had roughly half the number of donors that its neighbors had. The worst score was the most sparsely populated.

This visualization indicates that donors tend to either donate frequently or donate high amounts. Also, for all levels of recency, infrequent donors that donate more were higher in count, and therefore there is potential in targeting these donors according to some specific defined windows.

Clustering by Giving History:

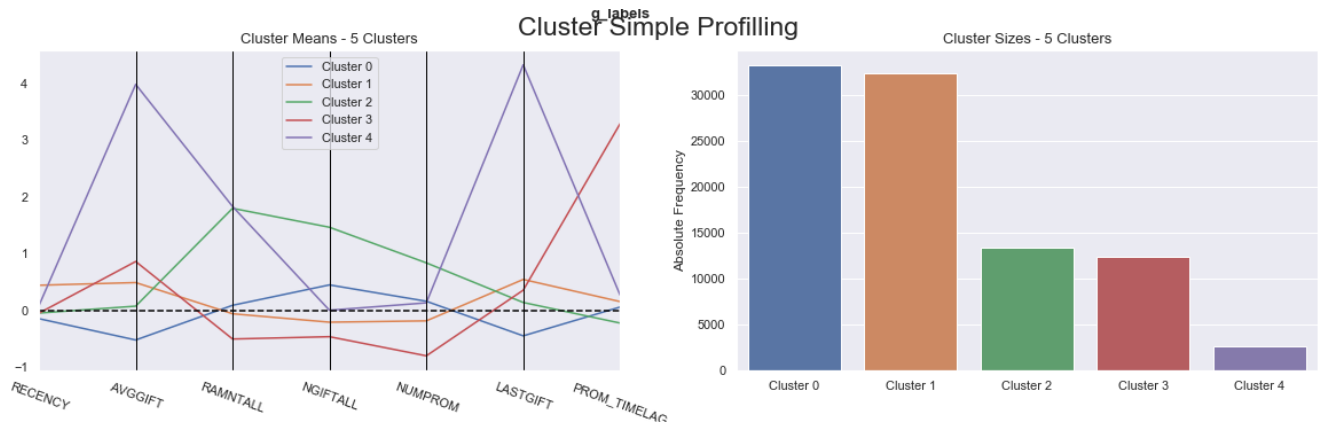

Fig 3: Cluster Profiling on Giving History features

Cluster 0 (*Frequent Low Dollar donors*): Represents approximately one third of the records. It has very regular (close to the median) values. Even though their total number of gifts is higher than the median, their donations are smaller in comparison with other donors.

Cluster 1 (*Infrequent Core donors*): Also represents approximately one third of the total records. The donors in this cluster are very close to the median in almost all giving history features. These donors give more money per donation, but they have not donated or received promotions as many times as the others. Also, they donated more recently than other donors.

Cluster 0 and 1 are the most similar, representing about one third of the records each. Their profiles are mirrored: donors of cluster 0 donate more frequently but donate less per donation; donors of cluster 1 donate less frequently, but each donation is of a higher amount. Both profiles have the same lifetime value and similar number of promotions received.

Cluster 2 (*Prolific donors*): Representing approximately a ninth of lapsed donors, this cluster bolsters both the highest average number of donations and promotions received. They match the sum of their lifetime donations with Top Donors (Cluster 4), but the average amount of donation matches the median. This can be the kind of loyal donor that donates in almost every promotion, but their individual donations will never go past a certain cap.

Cluster 3 (*One-off donors*): This cluster represents donors who have received the least promotions, which were also very spaced apart. The frequency of their donations matches these low values, but in contrast the average donation amount is the second highest of all clusters. Despite this, they have the lowest average lifetime donation value. While this signals donors who prefer to make one or two larger lifetime donations, it may also signal a group of untapped donors who are responsive to promotions, but that PVA has not been able to consistently reach through their mailings.

Cluster 4 (*Top* donors): The smallest cluster represents donors with the highest last gift and giving average. These donors had not received more promotions, donated more times, or more often than others. However, when they donate, it is an extreme amount. These are extremely high value donors who reveal relatively low responsiveness to the number of promotions; given their worth, more specific strategies could be implemented for the cluster.

Clustering by Census Data:

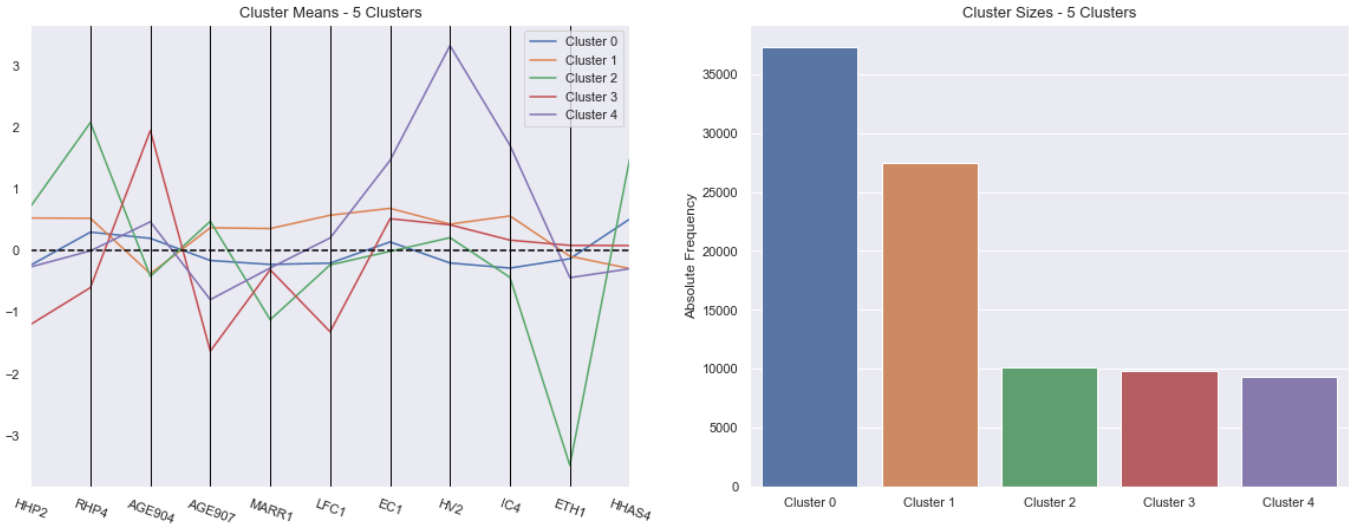

Fig 4: Cluster Profiling on Census data features

Cluster 0 (*Median neighborhoods*): Represent almost half of the records. These donors do not stand out for their behavior in any of the features. The size of this cluster is not surprising given that in preprocessing census data, many missing values had to be filled with median values.

Cluster 1 (*Middle-class familiar neighborhoods*): The donors in this cluster are characterized by living in neighborhoods with a median young population, with a high percentage of married people, and persons per household ratio. This could indicate the presence of families. In the areas of this donors the education level, and income is above the median, as well as the percentage of people in labor force.

Cluster 2 (*Lower-class* neighborhoods): The donors that belong to this cluster are characterized by living in neighborhood with an income below the median. Other features of this cluster are related to this fact, such as a high percentage of people below poverty level, and a high average number of people per room. Additionally, these neighborhoods are distinguished from all other clusters by the majority of non-white and young population, and the below median percentage of married people.

Cluster 3 (*Elder neighborhoods*): This cluster represents donors that live in neighborhoods with a high average age, a low percentage of non-adults, low people per household ratio, and a percentage of adults not working on labor force below the median.

Cluster 4 (*Upper-class neighborhoods*): The donors that belong to this cluster live in neighborhoods with high average income, very valuable houses, and which population has more years of education than the median, and a percentage of children below the median.

Combining the two cluster sets.

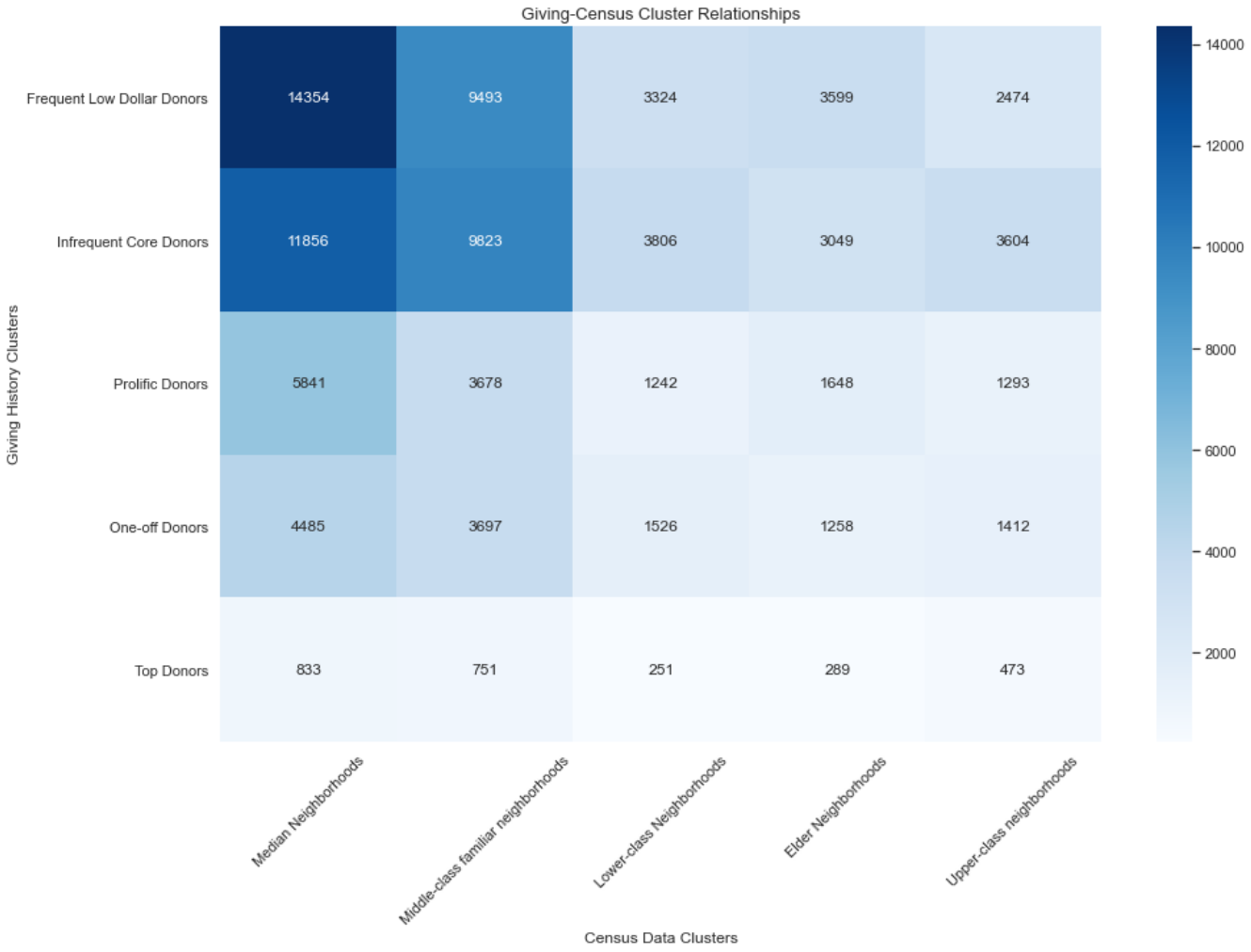

Fig. 5: Frequency table of Giving History clusters and Census Data clusters.

The Fig. 5 points towards *Top Donors* being relatively more frequent in *Upper-Class Neighborhoods,* despite most being part of *Median Neighborhoods.*

Median Neighborhoods is the most populated census cluster. It does not represent a concrete kind of neighborhood, in the sense it does not have any distinguishing features. Given this nature, it can be looked at as a benchmark, and no specific marketing approach needs to be tailored to this cluster.

Middle-class familiar neighborhoods, is a relatively smaller and distinctive cluster – they tend to be *Infrequent Core* or *One-off Donors*. They are characterized by uncommon but relatively generous donations. A reccommended marketing approach for this cluster is to focus on *One-off Donors,* by increasing the number of promotions sent to them.

Lower-Class Neighborhoods share similar proportions to those observed in *Middle-class familiar neighborhoods* but tend to a higher count of *One-off donors.* The marketing approach can be similar to the previous cluster, with special care to note these neighborhoods host the lowest amount of lifetime donation value.

Elder Neighborhoods follow different trends: a higher count of *Frequent Low Dollar Donors* can be found in these neighborhoods compared to others, as well as *Prolific Donors.* The common theme for these neighborhoods is the frequency of their donations. There is no evidence that indicates any specific strategy will increase the monetary amount, but rather that these neighborhoods are more responsive to sent promotions, and so the recommendation is to focus on more frequent campaigns in these neighborhoods.

Donors from *Upper-class neighborhoods* are mostly *Infrequent Core donors* but they also consititute a big part of the *Top donors,* so it would be convenient to transform them from *Infrequent Core donors* to *Top donors* by incrementing the number of promotions they receive, when possible. The number of *One-off donors* that live in these neighborhoods is also high, and data shows they are more likely to become *Prolific* or *Core donors.*

Conclusion:

Real life databases suppose a challenge obtaining clear interpretable results. A good data pre-processing is crucial to this end, focusing on the feature selection and data cleansing. Combinations of models are a reasonable approach to reach a fine tuning. At the end of the process donor profiles can be segmented and marketing decisions can be targeted to them.

Based on our findings and discussion, the marketing approaches suggested to retrieve these lapsed donors consist of augmenting the promotions sent to targeted clusters: *Upper-class neighborhoods*, *One-off donors,* as well as adjusting the frequency of promotions for donors identified as *Infrequent Core* or *Top Donors,* seeing as those are not particularly reactive to it (specificlly for top donors, given their extreme worth, it is reccommended to consider other targeted and individual approaches). In terms of frequency, there is value in experimenting with

more frequent mailing campaigns for *Elder Neighborhoods,* as these donors tend to be part of the more '*Frequent'* type clusters.

Further steps on this matter could include a density-based clustering for outlier detection, or a larger feature selection with a deeper insight on dimensionality reduction.

In the future, it is the team's suggestion that the data collection and warehousing be improved, ensuring the quality and availabiliy of the data generates more useful and actionable insights, specifically in what concerns individual demographic data.

References:

Azevedo, A. and Santos, M. F. (2008); KDD, SEMMA and CRISP-DM: a parallel overview. In Proceedings of the IADIS European Conference on Data Mining 2008, pp 182–185

Appendices

A. Engineering ANTIQUITY.

The following reasons justify the choice of FISTDATE over ODATEDW for the base:

- Month is defined for every record, unlike ODATEDW.
- It comes from the giving history file, and is consistent with the data there, which is used extensively.
- ODATEDW's earliest recorded gift is 2003, whilst FISTDATE is 1969. There is discrepancy between many of the dates. An assumption was made: ODATEDW is a more recent variable that did not inherit the values from FISTDATE, maybe originating in the implementation of a new system, and registered the date of the first gift after that moment in time. If that assumption holds, then ODATEDW does not capture a complete picture of the lifetime of a donor and is therefore misleading.

B. Outlier Removal Filters.

The following table presents the variables chosen for outlier removal, and the respective filter applied:

Name	*Threshold for Removal*	*Name*	*Threshold for Removal*
RAMNTALL	Above 1000	LASTGIFT	Above 200
NGIFTALL	Above 50	CARDPM12	Above 10
MAXRAMNT	Above 500	NUMPROM	Above 118
AVGGIFT	Above 100	ANTIQUITY	Above 15
MINRAMNT	Above 200	PROM_TIMELAG	Above 365

C. R^2 score of Clustering Methods

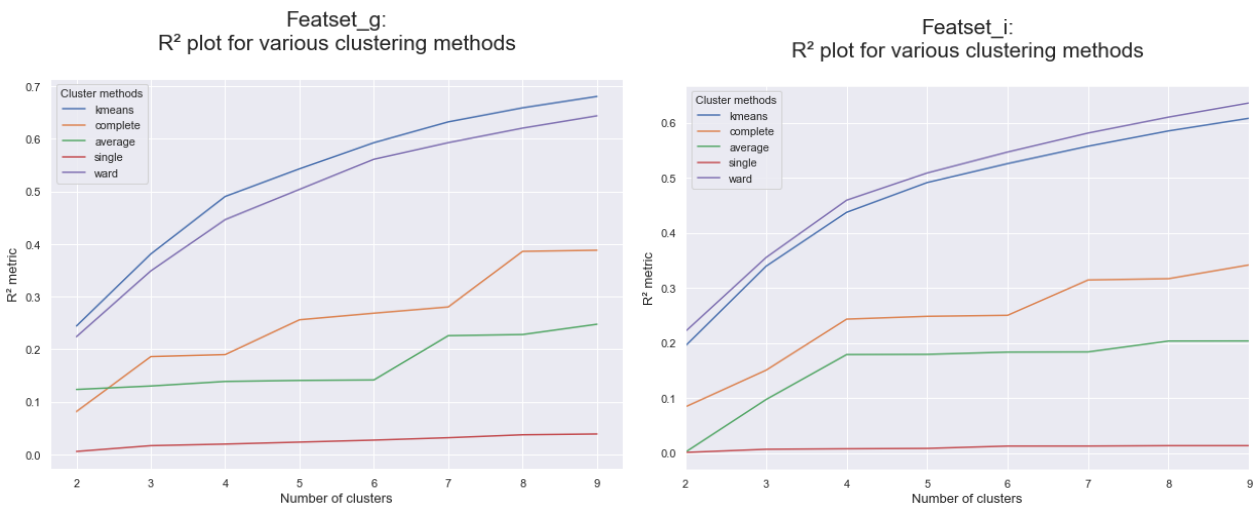

Fig. c.1 and c2: R^2 scores for featset_g (giving history) and featset_i (census variables) clusterings.

D. Dendrograms of Hierarchical Clustering Methods.

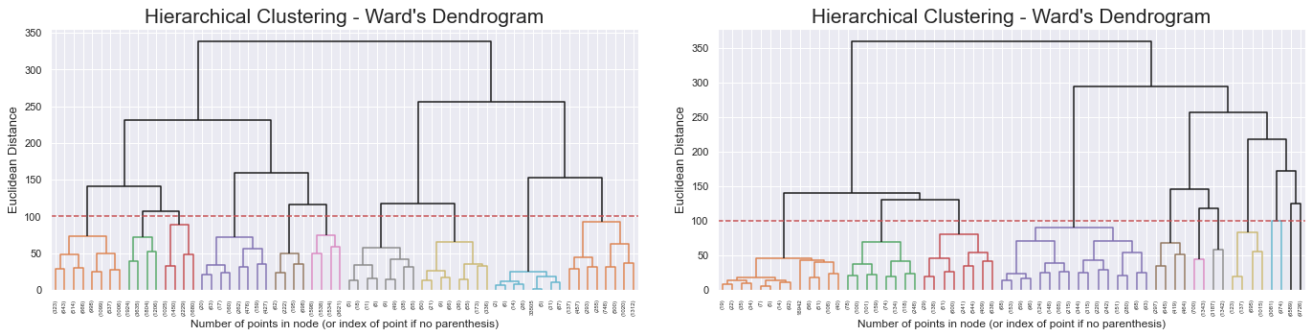

Fig. d.1 and d.2: Plotted Dendrograms for featset_g (giving history) and featset_i (census variables).

For giving history feature set, 4 to 5 clusters are indicated.

For census data feature set, 3 to 5 clusters are indicated. It is clear a sizeable cluster will be found for any number of clusters.

E. Inertia Plots of K-Means Clustering Methods.

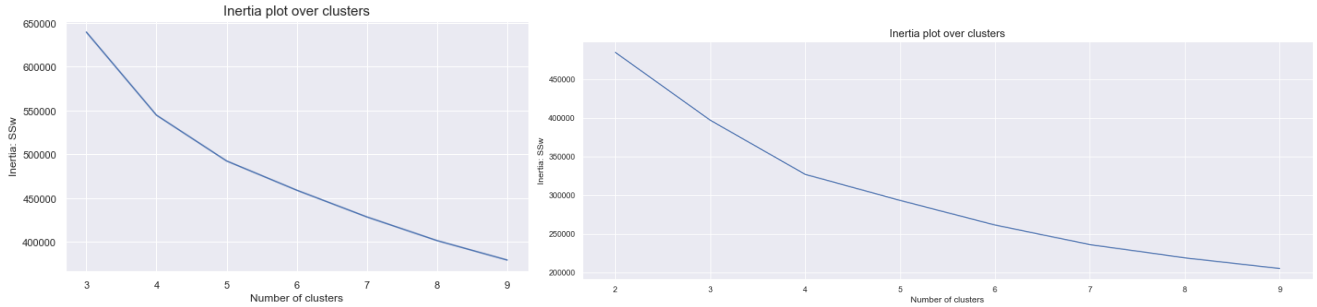

Fig. e.1 and e.2: K-means inertia plots for featset_g (giving history) and featset_i (census variables).

Following the Elbow Method criteria for selecting the number of clusters, within the range from 4 to 5 clusters is the appropriate amount for both sets.

F. Silhouette scores for the considered feature sets.

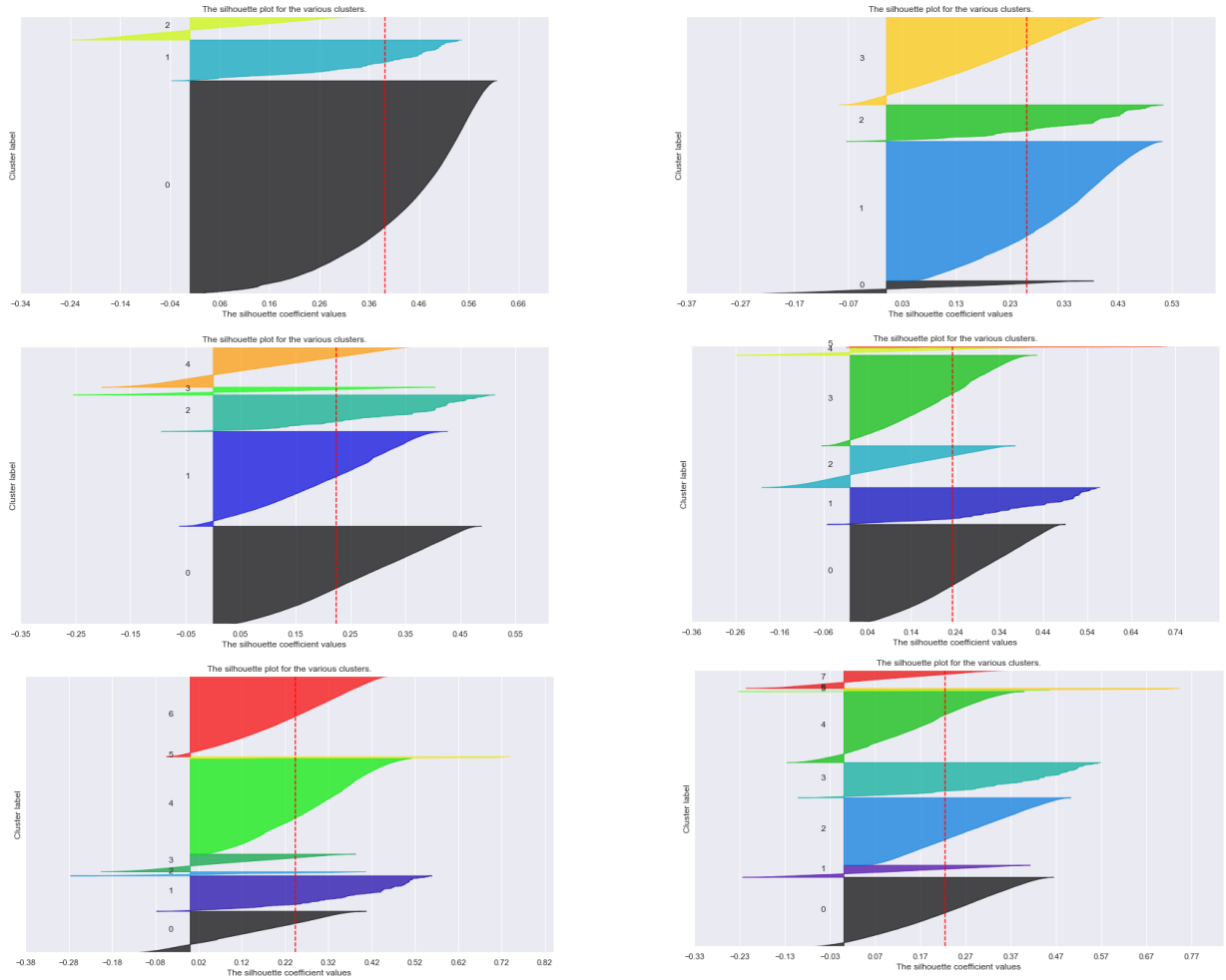

Fig. f.1 to 6: Silhouette plots for featset_g (giving history), for 3-8 K-means clusters.

An analysis of the silhouette scores for this feature sets indicates 5 clusters is a balanced number to use. Even at the lowest amount of clusters, small-sized groups are unavoidable, as well as these groups having datapoints that are “fuzzy”. Increasing the number of clusters increases the sparseness of already sparse clusters, which is not desirable.

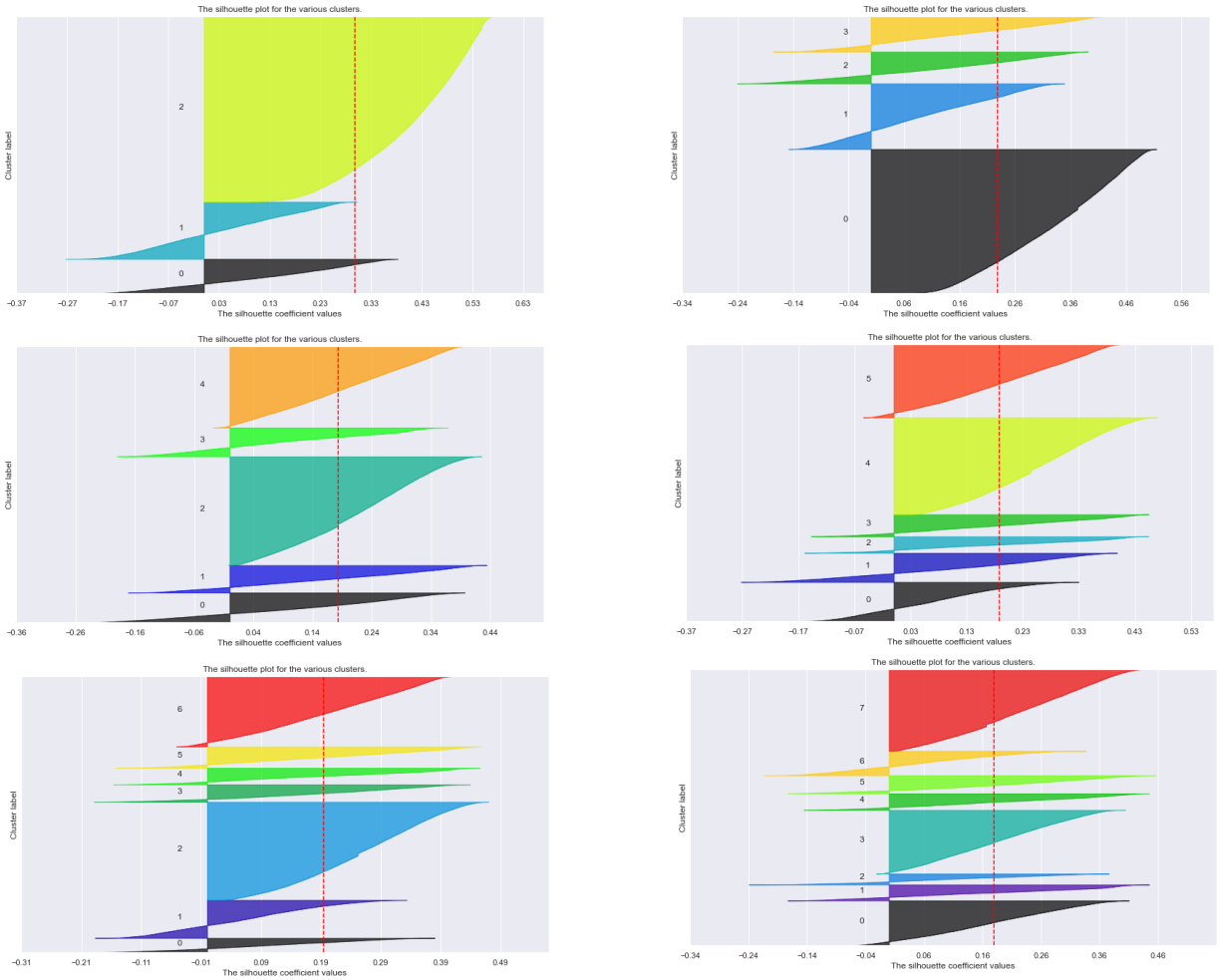

Fig. f.7 to 12: Silhouette plots for featset_i (census data), for 3-8 clusters.

In featset_i, the trend observed in the dendogram is visible – roughly half the data is clearly belonging to one sizeable cluster, which is fairly well represented, while other clusters are more sparsely populated. 5 is the number at which a separation of this all-encompassing cluster is split into two more meaningful but still clearly distinguishable clusters. No number of clusters significantly reduces the amount of data points which are not well represented by their assigned cluster.

G. Individual demographic (household) variables.

Below a summary of the exploration and preprocessing of individual demographic (household) variables can be found.

Name	*Description*	*Missing Values (%)*	*Additional Issues and Actions Taken*
PVASTATE	Categorical variable describing the presence of a PVA chapter.	98.47	Had two categories, one of these had only 5 records. Excluded.
HOMEOWNR	Categorical variable describing the state of knowledge regarding the donor's ownership of residence.	23.30	Two categories, not mutually exclusive. As such, information from this feature is unreliable. Excluded.
GENDER	Categorical variable, describing gender of the donor or identifying cases where the donor is not a single person.	3.10	Did not describe the gender of a single person. Missing values were transformed to Unknown gender. Considered for analysis.
NUMCHLD	Metric feature describing the number of children present in the household.	87.02	Minimum value was 1, indicating zeroes (which would be the vast majority) had been coded as NaNs. Proxy variables indicating the presence of children were used to fill some of these missing values; the rest was filled with 0. This preprocessing did not improve the overall distribution of the variable, with less than 15% of data being different than zero. No relation was later identified with any other variable, so NUMCHLD was dropped.
INCOME	Categorical feature describing the general level of income of the record.	22.31	Documentation provided did not clarify the magnitude of the differences between categories. The most frequent category was the highest level of income / wealth which is not consistent with the nature of the variable. As such, these were combined with other relevant variables to create a general measure of richness for the records in the dataset, such as WEALTH2, DOMAIN_S, and IC variables from census data, and an applied Decision Tree and KNN imputer to predict missing values.
WEALTH1	Categorical feature describing the general level of available wealth of the record.	46.88	
DOMAIN	Categorical composite feature describing the urbanicity and richness level of neighborhood.	2.43	Second byte had different ranges of values based on the first. Variable was decomposed into two categorical variables.

Additionally, age was considered as a demographic variable. However, it had severe quality issues, as described, and no correlation was identified with any giving history feature. It was therefore dropped.

H. RFM Analysis.

Antiquity.

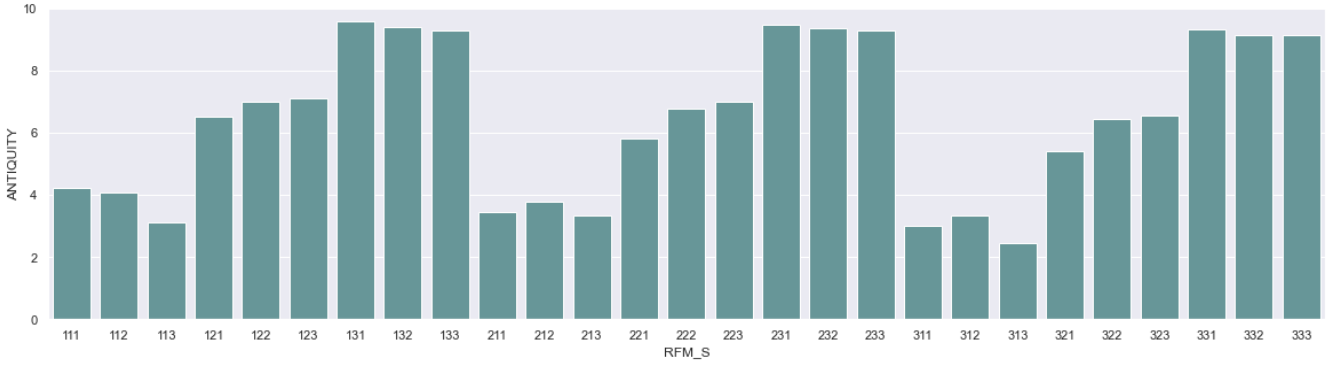

Fig. h.1: Average antiquity by RFM category

There is a clear relation between years since first donation and the frequency tiers of donations (i.e., number of lifetime donations). There is strong evidence lapsed donors return eventually, implying donor retention should be a secondary concern to acquiring new donors.

Sent promotions in RFM analysis.

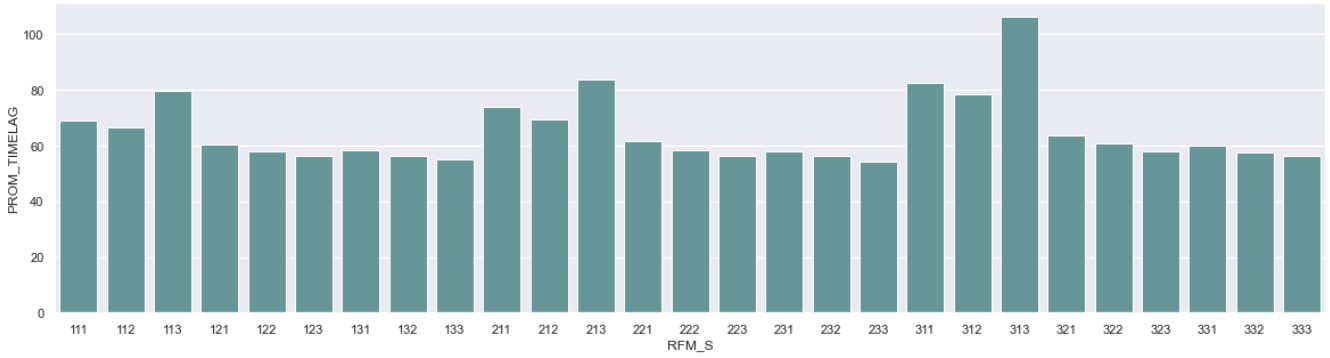

Fig. h.2: Average mean time between promotions by RFM category

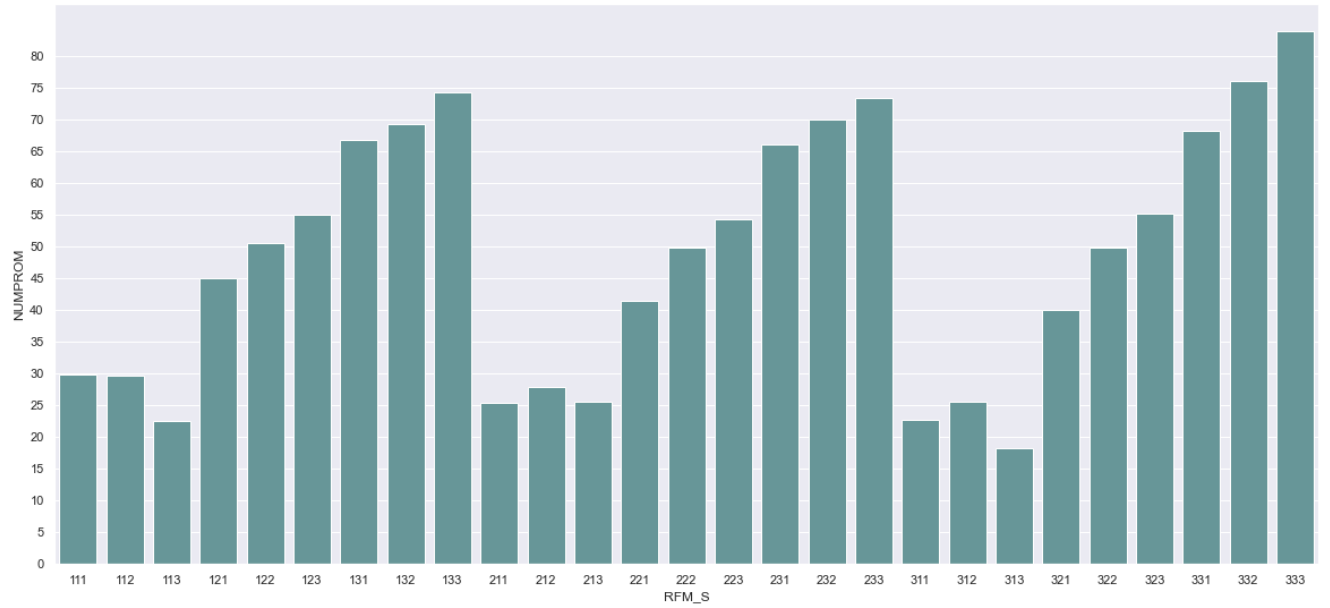

Fig. h.3: Average number of promotions by RFM category

Two trends can be found from this analysis: one, time between promotions decreases as Monetary Score increases, albeit slightly. Whether this is due to PVA knowingly targeting these donors for more promotions or not is unclear. A second trend can be found but only for the Frequency 1 donors: the monetary value increases with the time between sent promotions, as well as number of promotions sent. While interpretation of this behavior is illusive, it does

nonetheless indicate that there might be value in spacing promotions out for known infrequent donors, especially for those with high monetary value.