Customer Segmentation for Optimizing Product Marketing

Rajvi Abhangi, Prajwal Chinchmalatpure, Samuel Steiner

Abstract

Customer segmentation is the process of categorizing consumers into groups based on their shared characteristics. Many businesses use customer segmentation and the analysis of those clusters to aid them in adjusting their products based on their target clients from various customer categories. This process enables a company to better understand its consumers and makes it simpler to change which goods to advertise to meet the individual wants, habits, and concerns of various sorts of customers. Finding and understanding these customer segmentations will help businesses save money by allowing them to market products to the groups which are the most likely to purchase those products, and thus use this information to boost sales. In this paper, we will be looking at data provided by a grocery store which has information on all customers including purchasing habits.

Customer data is perfect for an unsupervised clustering model. Before clustering, analysis of the data which describes its shape as well as other summary statistics will be done to give a better understanding of the data we are working with. The data will then have to be cleaned and preprocessed in order to get it ready for clustering. We will also perform dimensionality reduction to help make the dataset more manageable. To select the number of clusters, a knee point detection will be done to select the optimal amount of clusters to use for the dataset. Then clustering can occur via agglomerative clustering, k-means ++ and spectral clustering. After clustering, we can evaluate the results through exploratory data analysis. We compare and contrast the results across the three methods and identify any difference, as well as evaluate each method based on the extractable knowledge from the clusters. Finally, we conduct some analysis to understand the information extracted and create profiles which build the knowledge to be used to make business decisions.

The results of the process outlined created four customer personas, we found that there was no meaningful difference between the clustering methods and that the clusters were more or less the same. We looked at several statistics for each cluster to create the persona and specifically designed them around the customers' acceptance of the campaigns already run by the grocery store. We also discuss further works which can be done to augment the results of this paper, such as comparing methods of selecting the number of clusters or the numbers of features to use during

dimensionality reduction. Furthermore, We discuss conclusions which can be drawn from the results of this paper.

Introduction

Customer segmentation is the process of dividing your consumers into groups based on similar criteria such as demographics or habits in order to better market to them. It is widely used within the industry and helps businesses make decisions on which products to sell, advertise in campaigns and which items to put deals on. When performing customer segmentation, each cluster can be referred to as a marketing persona. By definition, a marketing “persona” is a personification of a customer segment, and it's not commonplace for organizations to construct multiple personas to correspond to their various customer segments. This allows them to market effectively to different groups rather than zeroing into one group.

The consumer segmentation groupings generated can also be used to start conversations about the marketing personas. Because customer segmentation is commonly used to inform a brand's messaging, positioning, and sales processes, marketing personas must be closely matched to those client categories in order to be effective. However, in order for the knowledge extracted to be useful, a company must have a strong collection of customer segments to work with. Which is the crux of the problem when it comes to customer segmentation.

A collected dataset from an unspecified grocery store will be used to examine different methods of clustering. The dataset contains information pertaining to customer information based on demographics such as marital status, income, and the customer’s purchasing habits. The dataset also includes information on whether the customer responded to one of several marketing campaigns, as well as what mode of purchase the customer used. The dataset has records for approximately two thousand customers for this unspecified grocery store. The purpose of the customer segmentation could give a company like this insight into their customer base and allow them to make business decisions which ultimately provide an increase in revenue.

Creating the customer segmentations and therefore the personas, we want to make sure we are accurately and meaningfully making the clusters which they are based on. To this effect, we can test different segmentation methods and see if there are some which would work better with the data provided versus others. In this paper we look at three different methods of clustering (kmeans++, agglomerative, and spectral) which will be further explained in the methodology section. These 3 methods were selected due to the nature of the problem and their ability to be adapted to this type of data.

The current literature showcases the versatility of customer segmentation with the use cases from commercial banking [1] to telecom customer retention [2] while these use cases use different clustering methods, the overall process of segmentation

is similar. Segmentation allows for the data scientist who is doing the analysis to extract different knowledge based on different methods and data used. A paper written by Shaw, et al. [3] states, “Current emphasis on customer relationship management makes the marketing function, an ideal application area to greatly benefit from the use of data mining tools for decision support.” This showcases a scaled version of our problem and shows how techniques such as the ones presented in this paper can be used in a decision support system to help businesses. Other papers such as one written by Nandapala and Jayasena [4] show and explore similar questions as this paper.

Methodology

For this paper we are exploring the customer data through exploratory analysis, this is useful to provide insight into the data we are working with. We then perform some cleaning of the data as well as preprocessing in order to shape the data in a more reasonable way to prepare it for following steps of the segmentation process. Following the preprocessing, we will perform principal component analysis as a method of dimensionality reduction. We then select the number of clusters to create, and create our clusters. We can then finalize the process through analysis of the segments made , and the personas created from those segments. The following section explores each of these steps in more depth, as well as an explanation of the reasons behind decisions made.

Exploratory Analysis

To start we perform exploratory analysis looking at the shape as well as the summary statistics of the data to find any anomalies in the data so we can deal with them prior to doing any work with the dataset.

Data Cleaning and Preprocessing

Data cleaning of any unnecessary data is done, as well as removal of any entries which have missing fields or outliers. Once the data has been cleaned, it is ready for preprocessing. For preprocessing, we turn categorical data into numerical values. We transform dates into a discrete number to make them easier to work with. We also create a few new features of the data, which are created by joining and modifying existing features. Finally, all the values are scaled using the standard scalar, which gets the data centered around 0 with a standard deviation of 1.

$$\frac{x - mean\,(X)}{stdev\,(X)}$$

Standard Scalar

Dimensionality Reduction

Dimensionality reduction is the projection of data to a lower dimensional space. The number of traits can be quite vast at times, and many of them may be connected and so considered redundant. Principal Component Analysis (PCA) is a widely used statistical technique that relies on the connection between variables to get a perspective from a big amount of numerical data [5]. The first goal of PCA is to reduce dimensionality for the purpose of data compression. Second, having a visual perception of the data can reveal information that was previously concealed from human eyes. According to the optimal demands, we determine the number of dimensions onto which to project the data.

Choosing The Number of Clusters

The first step in clustering is to decide on the number of clusters we wish to use. To that effect we employ a method called the elbow or knee method which helps us find the optimal amount of clusters for the data we are working with. The elbow method accepts a maximum number of clusters (k) in an input. The sum of squared error (SSE) of the distance within clusters is calculated. As the value of K increases towards our selected maximum the SSE will decrease and at a certain value of K where there is the biggest decline in the SSE is the elbow. This is selected as the optimal amount of clusters for the segmentation process [6, 7].

Clustering

In this paper we look at three different methods of clustering which are KMeans ++, agglomerative, and spectral clustering.

KMeans++: The simplest clustering algorithm based on the partitioning principle. The algorithm is sensitive to the initialization of the centroids position; after calculating K centroids in terms of Euclidean distance, data points are assigned to the closest centroid, forming the cluster; after the cluster is formed, the new centroids are calculated using the cluster's means; and this process is repeated until convergence. [8, 9] Kmeans ++ is chosen due to the highly dependent nature of the base kmeans algorithm on the initialization of the centroids. KMeans++ initializes the centroids to be distant from each other, this leads to probably better results than random initialization. [10]

Agglomerative: This method is based on the formation of a dendrogram-based hierarchy. The dendrogram serves as a memory for the algorithm, allowing it to tell how clusters are generated. The clustering process begins with the formation of N clusters for N data points, followed by the merging of the nearest data points in each step, so that the current step has one fewer cluster than the previous one.

Spectral: This method performs a low-dimensional embedding of the affinity matrix between samples, followed by clustering of the components of the eigenvectors in the low-dimensional space (Kmeans). It's especially cost-effective if the affinity is high.

Analysis

Each of the results from the three clustering methods will be analyzed through a second round of exploratory data analysis, but from the resulting clusters rather than the dataset as a whole. This will allow us to find any conclusions which could be drawn from the segmentation process.

Code

For the code of this project, we used a Jupyter notebook in order to have access to an interactive workspace for the data. We set a seed so that the results can be reproduced (120). This project also makes use of several common packages for data visualization as well as the actual clustering which was done. Packages like Scikit-learn, NumPy, pandas, and seaborn were heavily relied on and the basis for the work done in this paper. The code is included with this report and has some details explaining the sections of code.

Results and Discussions

The Dataset

This paper uses a dataset created by Dr. Omar Romero-Hernandez specifically for exploring customer segmentation projects such as this one. The following explains each of the features of the dataset and what they represent.

People

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company
- Recency: Number of days since customer's last purchase
- Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Products

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Promotion

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place

- NumWebPurchases: Number of purchases made through the company's website
- NumCatalogPurchases: Number of purchases made using a catalog
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company's website in the last month

Other

- Z_CostContact: No description provided
- Z_Revenue: No Description provided

Data Cleaning and Exploratory Analysis

The data was cleaned using a listwise deletion of records with missing data; this resulted in the new dataset having 2216 entries down from the initial 2240. List wise deletion was chosen for the ease of implementing and the relatively small amount of deletions needed to complete the operation.

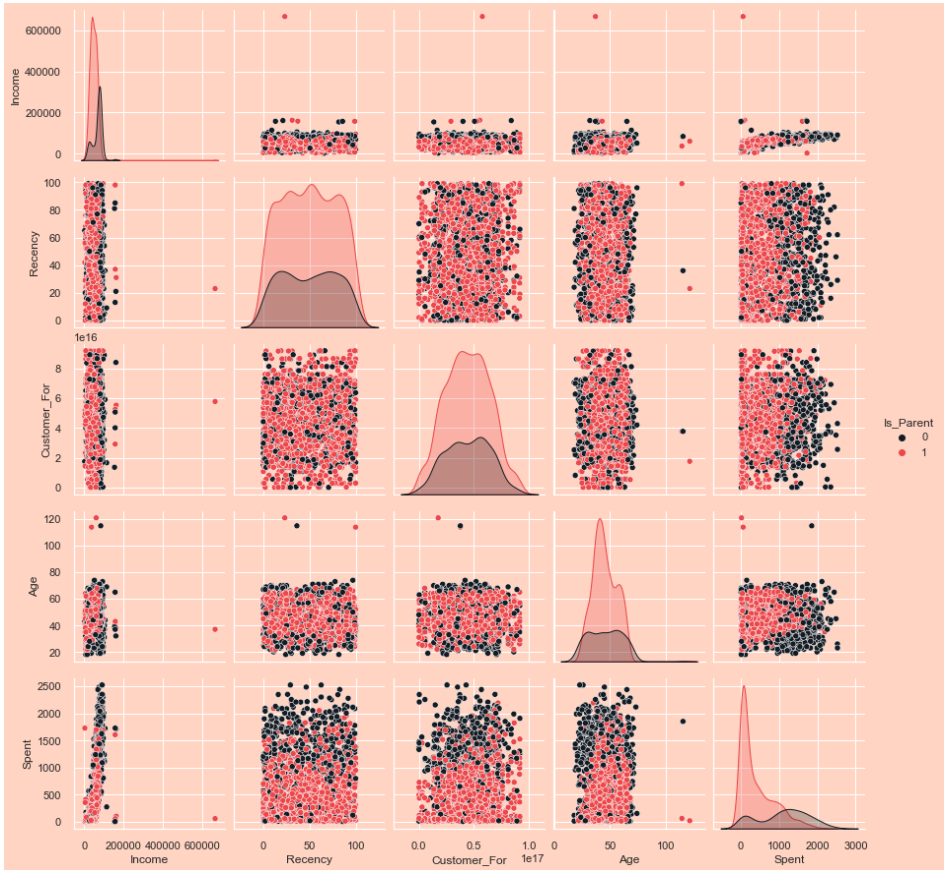
The exploratory analysis revealed a that the initial dataset had 3 columns, which were unnecessary for the segmentation/clustering problem. Those columns being ID (A random unique Identifier), Z_CostContact, Z_Revenue are two features which have no difference in the values of all entries have the same data, so there is no need to keep them. After deletion of the features which are unnecessary, the summary statistics for the relevant columns which will be used during the dimensionality reduction are in the table below.

	Year_Birth	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth
count	2216.00	2216.00	2216.00	2216.00	2216.00	2216.00	2216.00	2216.00	2216.00	2216.00	2216.00	2216.00	2216.00	2216.00	2216.00	2216.00
mean	1968.82	52247.25	0.44	0.51	49.01	305.09	26.36	167.00	37.64	27.03	43.97	2.32	4.09	2.67	5.80	5.32
std	11.99	25173.08	0.54	0.54	28.95	337.33	39.79	224.28	54.75	41.07	51.82	1.92	2.74	2.93	3.25	2.43
min	1893.00	1730.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	1959.00	35303.00	0.00	0.00	24.00	24.00	2.00	16.00	3.00	1.00	9.00	1.00	2.00	0.00	3.00	3.00
50%	1970.00	51381.50	0.00	0.00	49.00	174.50	8.00	68.00	12.00	8.00	24.50	2.00	4.00	2.00	5.00	6.00
75%	1977.00	68522.00	1.00	1.00	74.00	505.00	33.00	232.25	50.00	33.00	56.00	3.00	6.00	4.00	8.00	7.00
max	1996.00	666666.00	2.00	2.00	99.00	1493.00	199.00	1725.00	259.00	262.00	321.00	15.00	27.00	28.00	13.00	20.00

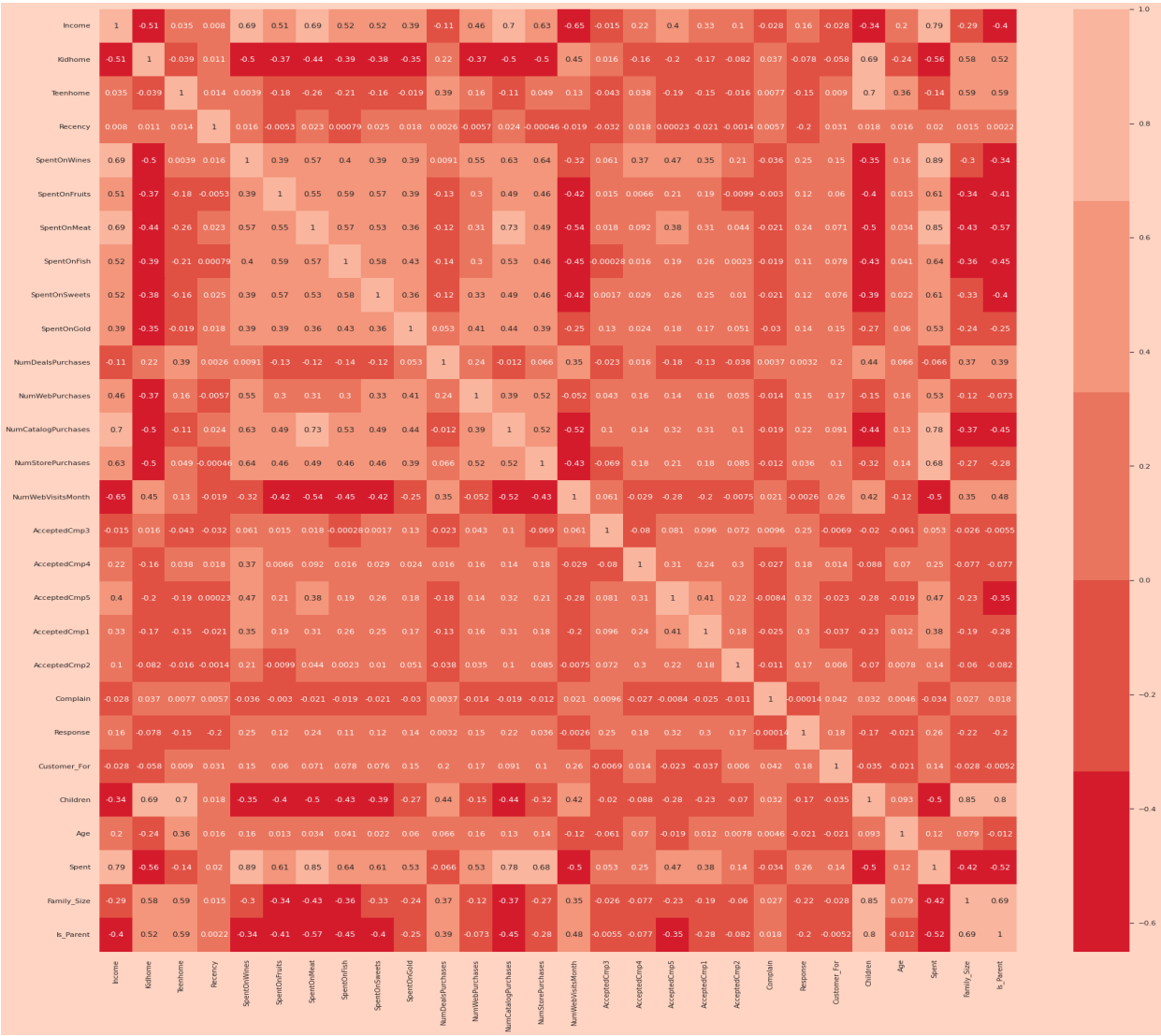
Data Preprocessing

The data preprocessing step in this paper turned any non-numerical data within the dataset into workable values. This step is the largest part of the process, we cleaned up the categorical features of marital status and education. We changed the names of the columns for how much each customer was spending on specific products, as well as created a new feature which was the sum of these columns. Another feature created for was the number of people who lived in the household, this is easily calculated based off the values of the number of kids and teenagers at home plus marital status.

To detect outliers, we graph the data comparatively, we select features which will be later used in the dimensionality reduction to make sure we aren't removing information which we want to use for analysis after segmentation. The graphic below shows the graphs made, this graphic is also available with the code of this paper.



The graphs show that there seems to be some outliers with age and income, it is easy to deal with these, and they are just removed from the dataset using listwise deletion before moving on. Then we can look at the correlation of the features to see if there are any redundant features that are unnecessary. The heatmap shows no features have an extreme correlation, we would expect our augmented features to have correlation with the features that they are created from. This demonstrates we have no redundant features.



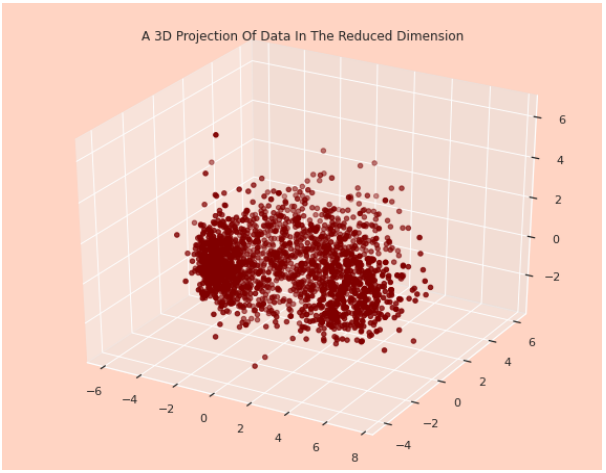
We employ the usual scalar approach, which is discussed in the methodology part of the report, to scale the data as the final stage of data preprocessing.We can now go on to dimensionality reduction and clustering, which will allow us to perform our analysis of the clustering methods as well as construct consumer personas, now that we have our data set up and scaled.

Dimensionality Reduction

For this paper, we used Principal Component Analysis (PCA), which is described in the methodology section of the paper. The summary statistics for the dataset after are in the following table.

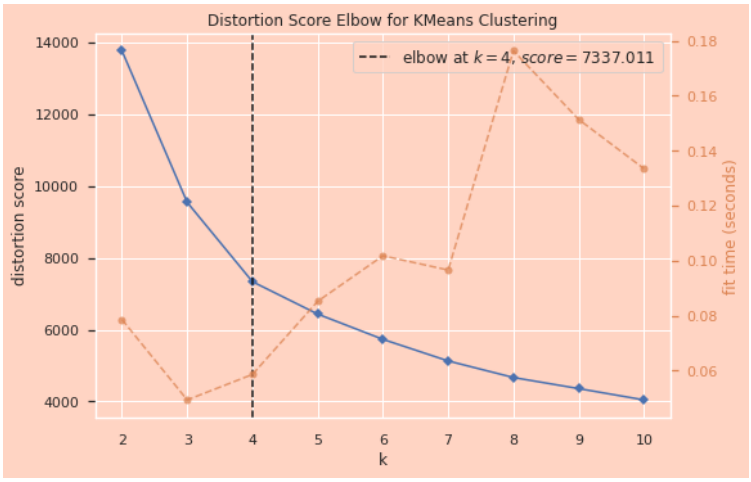
	count	mean	std	min	25%	50%	75%	max
col1	2212	-0.00	2.88	-5.96	-2.54	-0.78	2.37	7.46
col2	2212	0.00	1.70	-4.26	-1.33	-0.17	1.25	6.09
col3	2212	0.00	1.21	-3.33	-0.87	-0.00	0.81	6.35

We can also visually represent the dataset with the following graph.



Clustering and Segmentation Analysis

For the selection of the number of clusters we used the elbow method (described in the methodology section) which selected 4 as the optimal amount of clusters as seen by the graph below.

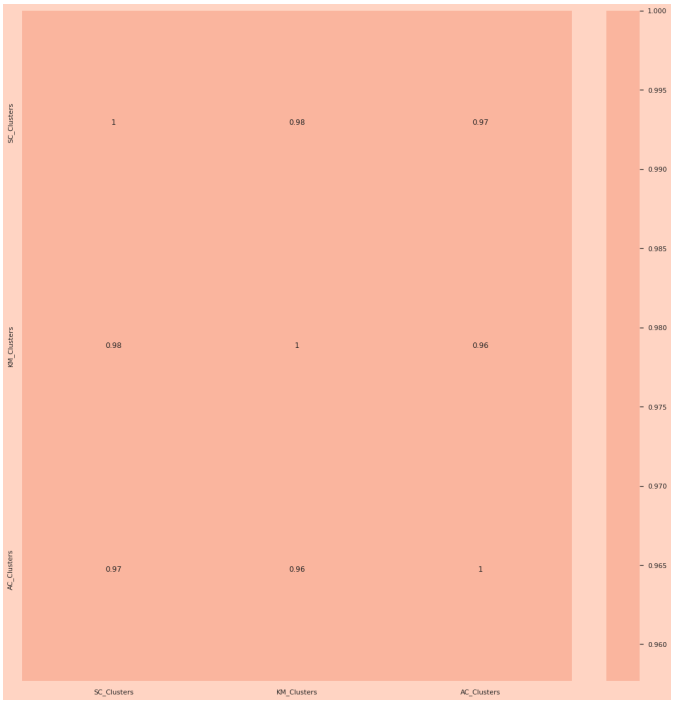


After the clustering has been completed, we can examine the summary of the clusters to determine whether there are any differences between the clustering methods. The data for each of the clustering methods is listed in the table below.

Count of Customers in Each Cluster Based on Clustering Method

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Spectral	699	616	415	482
KMeans ++	439	617	674	482
Agglomerative	729	612	403	468

At first glance, there appears to be a distinction between the approaches, however closer inspection of the clusters reveals that Kmeans++ method cluster 2 is comparable to cluster 0 for the other two methods. We can use a heatmap to see how highly connected the clusters are once they've been swapped. You can also further observe that the differences between the clusters based on the approach are minimal by looking at the supplementary graphs given with the code.



Customer Personas

There are many ways to look at the clusters and make the personas. With the code given, there are many graphs which showcase the details of each cluster. Here are the key points of each persona:

Cluster 0:

- Definitely a parent
- At least 2 people at home at max 5
- Most of them between 39 and 57
- Most have a teen at home

Cluster 1:

- Majority are a parent
- Most between the age of 28-48
- Most have at least 2 people living at home
- 3 people living at home at most
- Almost none of them have teenagers

Cluster 2:

- Most of them are parents
- Maximum household size is 4
- Majority of them between the age of 39 - 57 years old
- Majority salary between 54 thousand and 77 thousand

Cluster 3:

- Not a parent
- Majority of them between the age of 31-59
- Majority of salary between 67.7 thousand to 88.5 thousand
- A majority of the customer households have 2 people in it

With these, we can try and find better marketing strategies to target each group and see curate future campaigns to them.

Future Work

This research can be used in a variety of ways in future projects. There are various other steps in the segmentation process that can be tweaked and tested in order to conduct an analysis like the one presented in this research. Two main choices for this type of analysis are the manner by which the numbers cluster are chosen, and the number of dimensions projected onto during dimensionality reduction

Other future research might look at the segmentation's predictive potential, with the question being whether the segmentations can forecast future purchase behavior or not. This project would necessitate time series data that looked into clients' future purchases following segmentation. This analysis was not possible with the dataset we utilized.

A more expansive experiment using multiple datasets could also be done for future research. This would allow for a more definitive conclusion to be reached on the way in which the clustering method affects the segmentation process.

Conclusion

While no differences could be found between the methods of clustering in this research, valuable follow-up questions produced can be used in future research. This paper is not conclusive since it did the experiment only using one dataset and further research would be needed to make any definitive claims that during customer segmentation the clustering methods have no bearing on the outcome. This paper also showcased the exploratory nature of questions, which allow for experimentation in the segmentation process.

References

[1] H. Su-li, "The customer segmentation of commercial banks based on unascertained clustering," *2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM)*, 2010, pp. 297-300

[2] L. Xie and D. Pan, "On customer segmentation and retention of telecom broadband in Pearl River Delta," *Proceedings of the 29th Chinese Control Conference*, 2010, pp. 5564-5568.

[3] C.-H. Cheng and Y.-S. Chen, “Classifying the segmentation of customer value via RFM model and RS theory,” *Expert Systems with Applications*, vol. 36, no. 3, 2009, pp. 4176–418

[4] E. Y. L. Nandapala and K. P. N. Jayasena, "The practical approach in Customers segmentation by using the K-Means Algorithm," *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, 2020, pp. 344-349

[5] H. Abdi and L. J. Williams, Principal Component Analysis, John Wiley & Sons, Inc., vol. 2, July/August 2010.

[6] C. Yuan, "Research on K-Value Selection Method of K-Means Clustering Algorithm", 2019.

[7] T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, "Customer Segmentation using K-means Clustering", *2018 International Conference on Computational Techniques Electronics and Mechanical Systems (CTEMS)*, pp. 135-139, 2018

[8] Tanupriya Choudhury, Vivek Kumar and Darshika Nigam, "Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering*, 2015.

[9] Tanupriya Choudhury, Vivek Kumar and Darshika Nigam, "An Innovative and Automatic Lung and Oral Cancer Classification Using Soft Computing Techniques", *International Journal of Computer Science & Mobile Computing*, 2015.

[10] “2.3. clustering,” *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#k-means. [Accessed: 01-Apr-2022].