Analysis on Customer Segmentation in Grocery Store Industry

Yinuo Guo[1]

[1] Boston College, Chestnut Hill MA 02467, USA
guoyo@bc.edu

Abstract. The practice of segmenting a company's customer base into groups based on shared traits like demographics, behavior, preferences, or requirements is known as customer segmentation. It is a crucial tool for businesses because it enables companies to better understand their customers and customize their marketing initiatives to fit the unique demands of each group. This paper goes through how an unsupervised clustering of data on customer records from a grocery store company's database was performed by an author named Karnika. The result shows that Karnika created four clusters and utilized them to profile clients in clusters based on their family configurations, income levels, spending habits, and many other characteristics. This segmentation enables the grocery company to better grasp each customer group's purchasing characteristics and may be applied while creating more effective marketing plans.

Keywords: Customer segmentation, dimensionality reduction, clustering.

1 Introduction

The grocery store industry is an essential part of the global economy, providing food and other essential items to consumers. According to U.S. Department of Agriculture, the total amount of food sold in the United States in 2019 at supermarkets, other grocery shops (apart from convenience stores), warehouse clubs, and supercenters was $653 billion. The 20 biggest food merchants in the country generated $410.4 billion in sales in 2019, or 65.1% of all food sales. Between 1990 and 2019, its proportion climbed by 85.9% (from 35.0% to 65.1%, respectively) [1]. However, McKinsey experts claim that there are many rapid changes in the grocery market. Pandemic shapes this market because once people's lifestyle is affected by the pandemic, grocery market, the necessity provider, will be shaped as well. According to the official data, in December 2019, the share of e-commerce grocery sales was less than 5% [2]; it is predicted to increase rapidly and exceed 20% by 2030. Online groceries' application is useful and convenient so it's reasonable to predict that its use will rise dramatically in the few years worldwide [2]. Although the purchase of groceries is inelastic, people are becoming more price sensitive and product sensitive under nowadays situation. Therefore, it's very important to investigate what customers' preferences are and what kind of customers that grocery company should especially keep focus on.

2 Literature Review

The research on customer classification has always been a topic of concern in academic circles. Smith (1956) was the first to present it, and since then, it has been widely validated by academic research and effectively implemented by several businesses across a range of industries [3]. Customer segmentation is the process of dividing a customer base into distinct subgroups that have similar needs and traits [4]. Many businesses are using customer segmentation today as a core marketing strategy to better understand the traits and demands of their consumers. There are two studies that clearly investigate the importance of customer segmentation [5,6].

The study by Bachtiar (2019) aimed to develop a customer segmentation approach using a two-step mining method based on the RFM model. In order to provide a more descriptive interpretation of the segmented customer, after cluster division through k-Means clustering method, the researchers use RFM model to further illustrate each cluster's importance and application. This study's two-step mining method, which combines k-Means clustering with association analysis, is useful in devising marketing strategies since now people have more comprehensive information about each cluster through applying the RFM model and making the association. The author found that the proposed approach was effective in identifying distinct customer segments with different transactional behavior patterns and preferences. The author also found that the additional variables identified by the decision tree algorithm improved the accuracy of the segmentation results. The study's findings suggest that the proposed approach can be a valuable tool for customer segmentation in online retail contexts. The author suggests that the approach could be used by online retailers to develop more effective targeting and communication strategies that align with customer needs and preferences. They also suggest that future research could explore the effectiveness of alternative segmentation approaches and the impact of customer segmentation on online retail performance [5].

The study by Allaway, D'Souza, Berkowitz, and Kim (2014) aimed to develop a dynamic segmentation approach for analyzing customer behavior in loyalty programs. It's interesting to see that researchers, at this time, turns to evaluate data that is already collected rather than on surveys and other immediate tests. In this way, the researchers find and analyze those statistically predetermined customer cluster. This method provides many benefits for those marketers who have access to historical purchase data because now they can analyze those existing data easily, track their consumers' purchase activities, and cluster consumers based on their business value and previous responses to their products. Additionally, surveys might offer crucial data for marketers because it reflects the current situation and may herald some latent and surprising problems within an industry [6].

There's one study containing the comparison of different clustering methods. The study by Li et al (2022) finds that as opposed to PC-based approaches such as partitional clustering, HC-based methods, such as hierarchical clustering, produce higher quality clusters while being more sophisticated. Consequently, HC-based ensemble clustering techniques can be effective but have issues with computational complexity. This research suggests model selection techniques to reduce the complexity of HC-based ensemble clustering methodologies [7]. The notebook that this study goes through performs HC-based clustering.

However, there are still some weaknesses existing in these two particular studies. The study by Bachtiar (2019) states how effective and successful the RFM model can help to improve customer segmentation analysis, it ignores the bias towards high-frequency buyers: RFM analysis tends to give more weight to customers who make frequent purchases, regardless of their monetary value [5]. This can lead to a bias towards high-frequency buyers, which may not necessarily be the most profitable or valuable customers and put over reliance on monetary value: RFM analysis places a significant emphasis on the monetary value of purchases, which may not always reflect the true value of a customer. For example, a customer who makes infrequent purchases but has a high lifetime value may be overlooked in favor of a high-spending customer with a lower lifetime value. The study by Allaway, D’Souza, Berkowitz, and Kim (2014) clearly state that the fact that it is never clear how many segments should be used to represent the loyalty database's underlying structure is a drawback shared by all existing segmentation methods [6]. Therefore, this paper will further enrich the research in this field by going through the details of customer segmentation on a grocery store. The research in this paper is helpful for readers to better understand the whole process of customer segmentation and realize the importance of customer segmentation.

3 Data and Variables

3.1 Variables

In the notebook, the author performs customer segmentation in order to better understand the business value of each consumer cluster and additionally to allow the firm to address the needs of various consumers based on their characteristics [8].

Firstly, the author imports libraries and loads data, which is public-open data from Kaggle. There are several types of data.

In the consumer information category, it contains: ID, Year_Birth, Education, Marital_Status, Income, Kidhome, Teenhome, Dt_Customer, Recency, and Complain.

In the Products category, it contains MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, and MntGoldProds. These variables all indicate the number of products purchased in the last 2 years.

In Promotion category, it contains NumDealsPurchases, AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, and Response. These variables represent different marketing strategies such as coupon, deals, and sale. It’s important to see whether consumers respond to these deals. If so, the frequencies matter.

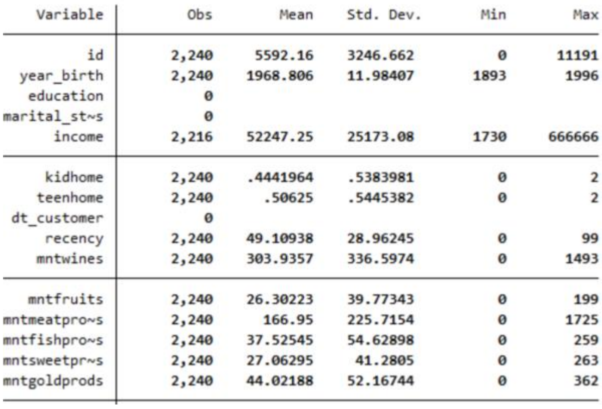
In Place category, in contains NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth. These variables represent how often consumer go to purchase through different ways. This is an important indicator, which tells markers consumer’s business value clearly.

3.2 Data Cleaning

The next step is to clean and format data. Firstly checks how many data points are in this dataset [8]. This study uses Stata to confirm that there are 2240 data points overall (Table 1). Then, the author finds if there are any missing values for each feature: there are missing values in income (Table 1). For the missing values, drops the rows that have missing income values directly [8].

Also, there are some categorical features in the data frame: education, marital_status, and date_customer, which are with zero observation in State and are supposed to be encoded into numeric forms for further analysis to solve this problem.

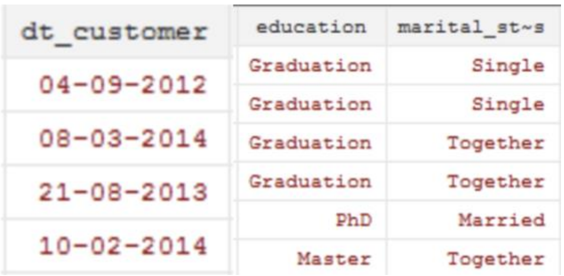
Table 1. Information about Variables in the Dataset.



Variable	Obs	Mean	Std. Dev.	Min	Max
id	2,240	5592.16	3246.662	0	11191
year_birth	2,240	1968.806	11.98407	1893	1996
education	0				
marital_st~s	0				
income	2,216	52247.25	25173.08	1730	666666
kidhome	2,240	.4441964	.5383981	0	2
teenhome	2,240	.50625	.5445382	0	2
dt_customer	0				
recency	2,240	49.10938	28.96245	0	99
mntwines	2,240	303.9357	336.5974	0	1493
mntfruits	2,240	26.30223	39.77343	0	199
mntmeatpro~s	2,240	166.95	225.7154	0	1725
mntfishpro~s	2,240	37.52545	54.62898	0	259
mntsweetpr~s	2,240	27.06295	41.2805	0	263
mntgoldprods	2,240	44.02188	52.16744	0	362

From table 2, the date of the customer's enrollment with the company is presented. The date a consumer entered the database is shown by the feature Dt_Customer, which is not processed as a DateTime. Therefore, the author comes up with a new feature called “Customer For” by setting a fix point (the last day that consumers make a purchase) and making a subtraction between the last day and the enrollment day.

Table 2. Information about Categorical variables.

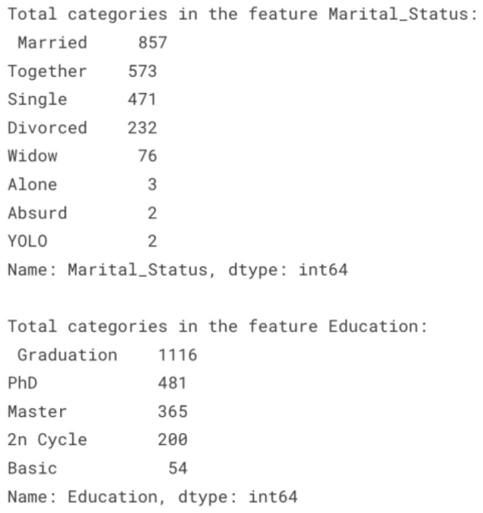


dt_customer
04-09-2012
08-03-2014
21-08-2013
10-02-2014

education	marital_st~s
Graduation	Single
Graduation	Single
Graduation	Together
Graduation	Together
PhD	Married
Master	Together

From table 2, “education” and “marital_status” are presented. Before the author transfers those categorical variables into numerical ones, firstly the author needs to simplify the categories for each feature [8]. For the feature Marital_status, there are eight situations, which is complicated. So, the author simplifies them into two new features: “partner” and “single”. “Married” and “together” can be attributed to the feature “partner”. All else can be attributed to the feature “single”. For the “Education” category, the author simplifies them into three features: “undergraduate”, “postgraduate”, and “graduate”. “2n Cycle” and “Basic” can be sorted into “undergraduate”. “Master” and “PhD” are sorted into “postgraduate”. “Graduation” is sorted into “graduate”.

Table 3. Information about Categorical Variables. (Karnika, 2022).



```
Total categories in the feature Marital_Status:
 Married     857
Together    573
Single      471
Divorced    232
Widow        76
Alone         3
Absurd        2
YOLO          2
Name: Marital_Status, dtype: int64

Total categories in the feature Education:
 Graduation    1116
PhD            481
Master         365
2n Cycle       200
Basic           54
Name: Education, dtype: int64
```

The author used the following actions to create some additional features after sorting the category variables [8]. The author retrieves a customer's age from their "Year Birth" [8]. A new feature called "Spent" is added because it indicates how much a customer spend in the last two year, which can be used in the analysis of business value. Another feature "Living With" is added because it shows how many years a couple live together. The feature "Children" show how many children live in a home, which is essential because when a family have more children, it’s likely to infer that this family have more needs and purchase motivations in grocery industry [8]. Furthermore, adding a feature that indicates "Family Size" will provide more clarity on the home. Finally, removing some outliers is also important.

After data cleaning and feature engineering, the author preprocesses the data in order to do clustering procedures [8]. The data is preprocessed using the procedures below: labeling the category characteristics, using the default scaler to scale the features, and making a subset dataframe to reduce dimensionality.

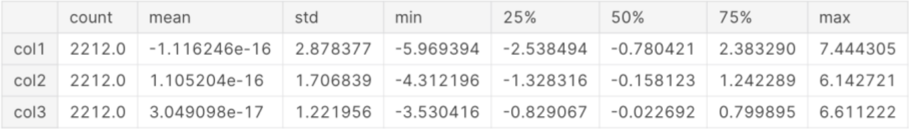
4 Results

4.1 Dimensionality Reduction and Clustering

Those variables representing a customer's trait will determine the result of categorization. If there are more variable in the analysis, it's hard and slow to get a good result. Also, there are several variables conveying the same meaning, so it's necessary to remove them to avoid collinearity. Therefore, the author reduces the dimensionality of the characteristics before subjecting them to a classifier [8]. The author uses PCA to reduce the dimension to 3 and plot the reduced dataframe [8]. PCA play a core role in dimensionality reduction because it will remove those unrelated aspects and keep those explainable and useful information [9]. When useless information removed, the straightforward information can be better and rapidly processed. The study by Alkhayrat et al. claims that dimensionality reduction is an important step because when reduced dimensions apply, they find that the clustering performance was improved dramatically [9].

In table 4, there is statistical information about each dimension.

Table 4. Statistics about Three Dimensions. (Karnika, 2022).



	count	mean	std	min	25%	50%	75%	max
col1	2212.0	-1.116246e-16	2.878377	-5.969394	-2.538494	-0.780421	2.383290	7.444305
col2	2212.0	1.105204e-16	1.706839	-4.312196	-1.328316	-0.158123	1.242289	6.142721
col3	2212.0	3.049098e-17	1.221956	-3.530416	-0.829067	-0.022692	0.799895	6.611222

After reducing the characteristics to three dimensions, the author first uses the Elbow Method to determine how many clusters should be formed and finds that the optimal number of clusters is four [8]. The elbow method is a common technique used in data science to determine the optimal number of clusters to use for clustering algorithms. The distortion score is a measure of how well the data points are grouped around the centroids. It is calculated as the sum of the squared distances between each data point and its assigned centroid, divided by the number of data points. The lower the distortion score, the better the clustering. But it's also important to balance the distortion score and fit time. In Figure 1, it shows that four clusters are effective and optimal. Then, Agglomerative Clustering Model is used to recheck the performance of clustering and make sure that the number of final clusters is what the author expects [8]. Agglomerative clustering is a hierarchical clustering algorithm that recursively merges the closest pairs of clusters based on a distance metric, until a desired number of clusters is reached or all data points are contained in a single cluster. In Figure 2, it shows the plot of the clusters, which are separately converged in the center of the plot.

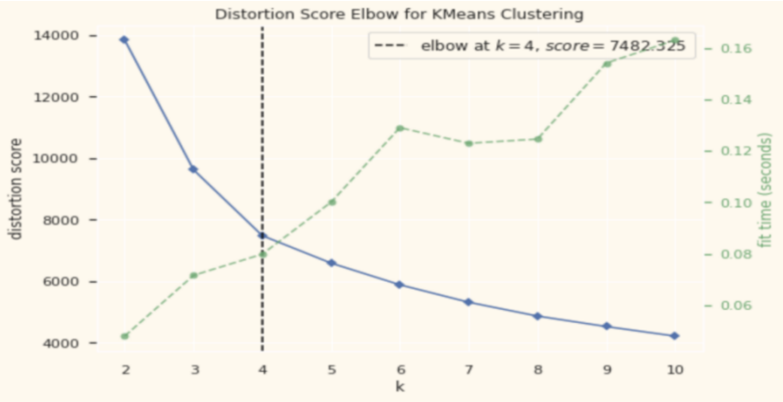


Fig. 1. Distortions Score Elbow for KMeans Clustering. (Karnika, 2022).

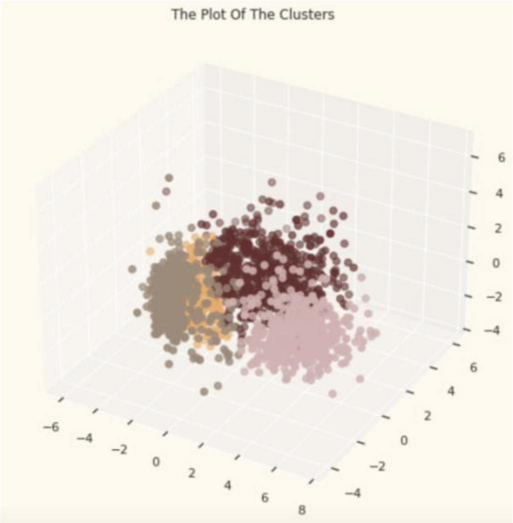


Fig. 2. The Plot of The Clusters. (Karnika, 2022).

4.2 Performance

After the author comes up with the number of clustering, she checks if this cluster dividing is optimal and effective [8]. The author first sees the group distribution of clustering and then figures the Income vs spending plot to see the customer characteristics for each cluster [8]. In Figure 3, the distribution of clusters is not skewed, which is a good sign. In Figure 4, the Income vs spending plot shows the cluster pattern: consumers from group 0 are inclined to have average income and make high transactions; consumers from group 1 are likely to have high income and make high transactions; consumers from group 2 are likely to have low income and make low transactions; consumers from group 3 tends to have low income but make high transactions.

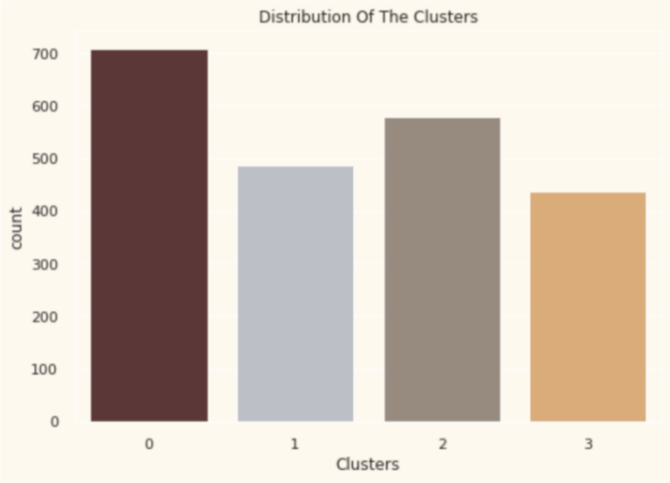


Fig. 3. Distribution of the Clusters. (Karnika, 2022).

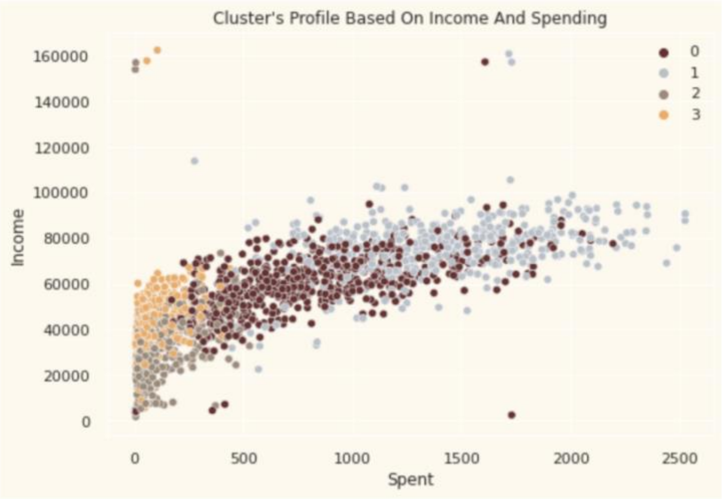


Fig. 4. Cluster’s Profile Based on Income and Spending. (Karnika, 2022).

4.3 Customer characteristics

Studying the patterns in the clusters formed and determining the nature of the clusters' patterns are also important. The author creates graphs that shows how much each cluster spend in each product from a grocery store. These products are wines, fruits, meat, and snacks [8]. Therefore, the author can clearly see each product’s consumption and each cluster’s preference [8]. Devising corresponding marketing strategies will be based on this useful information. Furthermore, the author explores how the campaign ads did in the past and sees the total number of promotions accepted in each customer cluster (Figure 5). The author concludes that consumers don’t make eager response to those different campaign ads and it’s surprising to see that cluster selectively choose one or two of the deals rather than taking all of 5 deals. Perhaps more well-designed and targeted campaigns are needed to increase sales. Though campaigns fail, the deals offered are successful. In Figure 6, the greatest results come from clusters 0 and 3. Cluster 1, one of our top consumers, isn't very interested in the deals, though. Nothing appears to powerfully draw cluster 2 in.

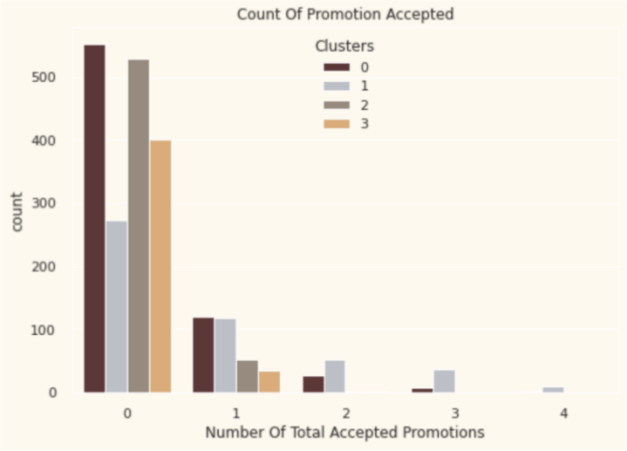


Fig. 5. Count of Promotion Accepted. (Karnika, 2022).

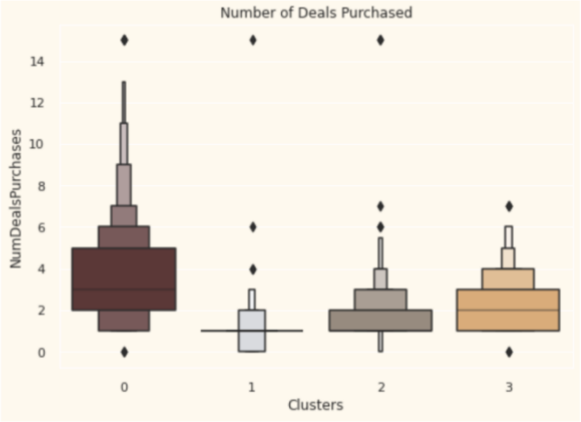
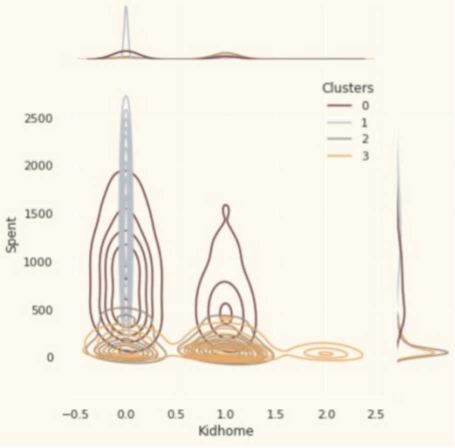


Fig. 6. Number of Deals purchased. (Karnika, 2022).

Business value and purchasing habit can be examined through analyzing each cluster. The author then takes a look at each individual in these clusters [8]. It's important to know which consumer cluster has already being the most valued one and which cluster needs further attention because marketers will devise different methods to handle distinctive consumer clusters [8]. For this reason, the author creates several graphs that represent how much each cluster spend when they are tested with different standards. For instance, in graph 10, the x-axis is age and y-axis is spent. In this case, the author wants to see how each cluster distributed depending on age group and which group contains the most numbers of youngest or of eldest.



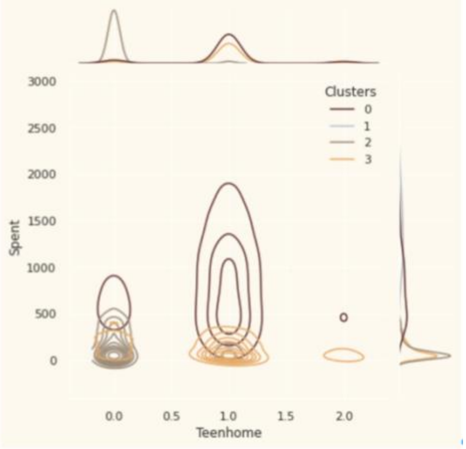
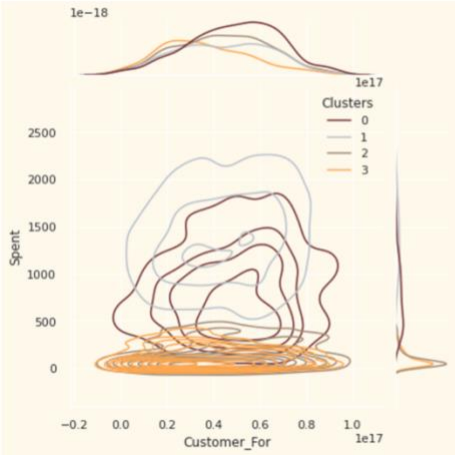


Fig. 7. Spent vs Kidhome. (Karnika, 2022). **Fig. 8.** Spent vs Teenhome. (Karnika, 2022).



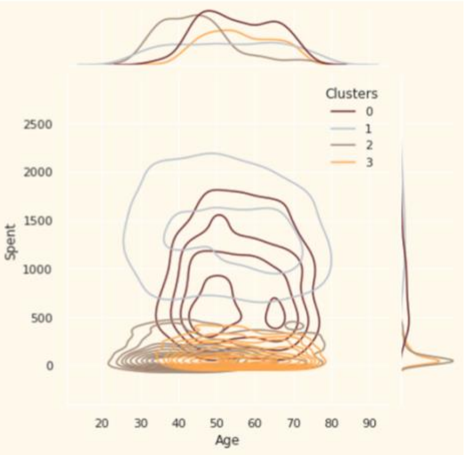
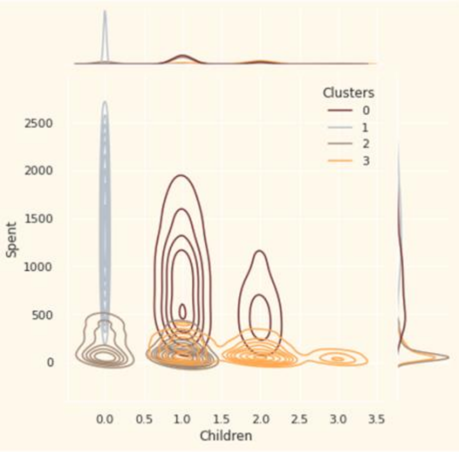


Fig. 9. Spent vs. Customer_For. (Karnika, 2022). **Fig. 10.** Spent vs Age. (Karnika, 2022).



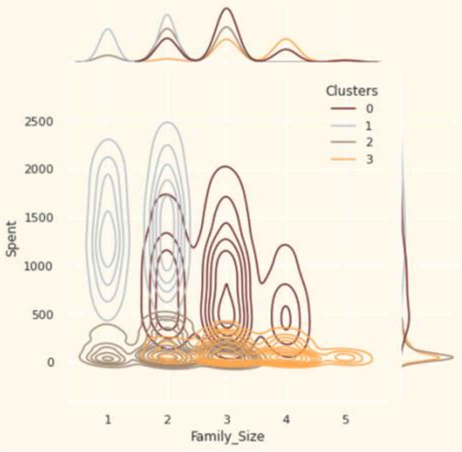
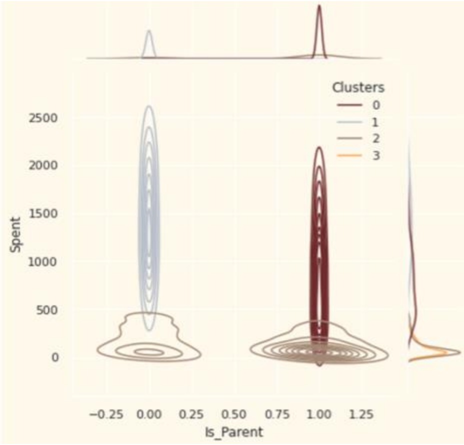


Fig. 11. Spent vs. Children. (Karnika, 2022). **Fig. 12.** Spent vs Family_Size (Karnika, 2022).



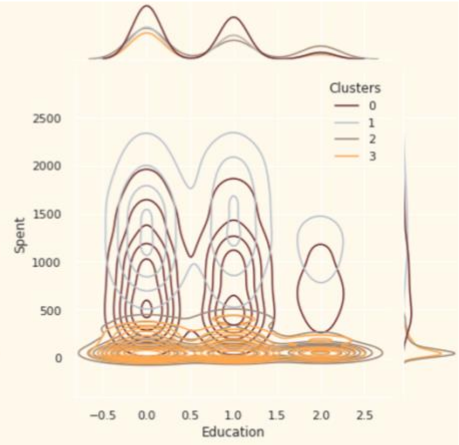


Fig. 13. Spent vs.Is_Parent. (Karnika, 2022). **Fig. 14.** Spent vs Education (Karnika, 2022).

Based on Figure 7-14, the author deduces the following information about the customers in different clusters. For cluster 0, the author infers that most have a teenager at home and are a parent [8]. They are relatively older compared to consumers in other clusters. For cluster 1, they have high income, are less likely a parent, and span all ages; maybe there are no more than 2 members in the family. For cluster 2, most of them may be a parent and have more than 2 family members. Also, they are relatively younger. For cluster 3, with lower income, they are relatively older and majority of them have a teenager at home. These are just possible inference, and individual differences cannot be easily ignored when discussing possible characteristics of each cluster.

5 Conclusion

Based on (Karnika, 2022), this study does the data cleaning first and reduces dimensionality by using PCA. Then clusters can be well divided through the Agglomerative clustering method [8]. After knowing the number of clusters, the author checks the performance and deduces the customer characteristics from each cluster [8].

This paper on customer segmentation has some implications. Customer segmentation is particularly important for retail grocery stores because it can help them optimize their marketing, merchandising, and operations to better serve the diverse needs and preferences of their customers. Management and marketing teams may profit from customer segmentation in two important ways. The first is to make it possible to accurately identify the core customer groups, which contain the most lucrative and devoted clients. Some research projects have used customer segmentation to estimate the earning potential of consumers [10]. With this chance, decision-makers may particularly target these targeted segments with their efforts by focusing on these high-value client groups [10]. The management of businesses is given the opportunity to comprehend consumer behavior and preferences as well as learn about various client groups, which is another significant advantage of customer segmentation [10]. However, this study doesn't put forward a new method and empirically investigating on customer segmentation, it needs to be further discussed in future research.

References

1. USDA ERS - Retail Trends. (n.d.). Retrieved April 17, 2023, from https://www.ers.usda.gov/topics/food-markets-prices/retailing-wholesaling/retail-trends/
2. McKinsey. (n.d.).The state of the grocery retail industry. Retrieved April 17, 2023, from https://www.mckinsey.com/industries/retail/our-insights/the-state-of-grocery-retail-around-the-world
3. Smith, W.R. (1956), "Product differentiation and market segmentation as alternative marketing strategies", The Journal of Marketing, Vol. 21 No. 1, pp. 3-8.
4. McDonald, M. and Dunbar, I. (2004), Market Segmentation: How to Do It, How to Profit from It, Elsevier Butterworth-Heinemann, Oxford.
5. F. A. Bachtiar, "Customer Segmentation Using Two-Step Mining Method Based on RFM Model," 2018 International Conference on Sustainable Information Engineering and Technology (SIET), Malang, Indonesia, 2018, pp. 10-15, doi: 10.1109/SIET.2018.8693173.
6. Allaway, A. W., D'Souza, G., Berkowitz, D., & Kim, K. (Kate). (2014). Dynamic segmentation of loyalty program behavior. Journal of Marketing Analytics, 2(1), 18–32. https://doi.org/10.1057/jma.2014.2
7. Li, T., Rezaeipanah, A., & Tag El Din, E. M. (2022). An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement. Journal of King Saud University - Computer and Information Sciences, 34(6, Part B), 3828–3842. https://doi.org/10.1016/j.jksuci.2022.04.010

8. Customer Segmentation: Clustering. (n.d.). Retrieved March 23, 2023, from https://kaggle.com/code/karnikakapoor/customer-segmentation-clustering
9. Alkhayrat, M., Aljnidi, M., & Aljoumaa, K. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. Journal of Big Data, 7(1), 9. https://doi.org/10.1186/s40537-020-0286-0
10. Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: A case study. Marketing Intelligence & Planning, 35(4), 544–559. https://doi.org/10.1108/MIP-11-2016-0210