Capstone three project proposal: Customer segmentation

Can we identify customer clusters/ groups of a department store that have similar profiles and what characteristics do these customer clusters have?

Problem Identification:

Companies want to adjust their marketing campaigns, product promotions, and other individualized selling strategies. Identifying customer clusters/groups helps them to create targeted marketing and personalized services for the customers. Sometimes this segmentation process can be imprecise, one-dimensional, and not enough to create an effective marketing plan or product improvement strategy. To improve the segmentation we can use machine learning.

Context:

Identifying the need for a customer segmentation process is the first step to implementing a process that aligns with the overall business plan. When businesses don't have an effective customer segmentation process, they might find themselves providing the same service level for all customers and all products without focusing on the top-level customers or products that bring in the best margins.

The customer segmentation project should be continuously refined to help the business better respond to changing market trends. Companies' top-level customers should certainly receive more attention while building customer relationships and drawing retaining plans.

On the product level, without clear differentiation, businesses may invest in products that may not align with current market trends and profitability goals. Differentiation helps businesses to make more strategic investments in customer service and products. (https://blog.arkieva.com/customer-segmentation-strategy-issues/, By Editor, March 21st, 2017, Segmentation, Supply Chain Strategy)

Well-defined processes and technology can help support customer segmentation and the value it brings to a company. How the different segments are defined, and what the data in each customer segment is, is also essential to user /customer adoption and the successful usage of the segments in analytics. (FEBRUARY 6, 2019/BY MIKE WILSON, https://www.ironsidegroup.com/2019/02/06/challenges-customer-segmentation/)

This recognition and definition of each segment was previously a challenging and time-consuming task, that demanded hours of manually poring over different tables and querying the data in hopes of finding ways to group customers.

Machine learning algorithms can help marketing analysts find customer segments that would be very difficult to spot through intuition and manual examination of data.
(https://bdtechtalks.com/2020/12/28/machine-learning-customer-segmentation/)

With customers expecting brands to provide personalization, the brands who don't focus on providing that within a modern customer segmentation strategy will soon be missing out.
A company's data can be used in machine learning and significantly change the way a company perceives its customers, segmentation processes, and marketing or product strategies.
(https://formation.ai/blog/customer-segmentation-models-theres-a-better-approach-for-2022/
April 25, 2022)

Project Goal:

This project's goal is to create a sophisticated, based on machine learning segmentation that will lead to a profit increase.
The customer data will be used to segment and create clusters of customers. The company can use this segmentation to strategize its marketing or product creation.

Method

Steps:

1. Data will be cleaned and checked for inconsistency.
2. Data will be explored to have an understanding of available data, the current distributions, trends, and what features can be successfully incorporated into the model.
3. Data will be preprocessed and prepared for modeling.
4. Unsupervised machine learning in form of clustering algorithms will be used. Several modeling options will be tried. Models will be created to segment the customer data and create clusters.
5. The models will be evaluated for best separated (inter clusters distance) and dense (intra-cluster distance) clustering (Silhouette score). The number of clusters also has to be viable to use for our purpose.
6. The best model will be applied to the segmentation of current data but also can be used on future data of the store.

DATA

Dataset: Customer Personality Analysis (please click on the name) is a publicly available dataset on Kaggle. Acknowledgment: The dataset was provided by Dr. Omar Romero-Hernandez. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis?resource=download
The dataset has 2240 entries that represent customers' interaction with the company's products.
It also has 29 features that can represent this information:

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 the last 2 years
- MntFruits: Amount spent on fruits in the last 2 years
- MntMeatProducts: Amount spent on meat in the last 2 years
- MntFishProducts: Amount spent on fish in the last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in the 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 the 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 the company’s website in the last month

Data wrangling:

The dataset was checked for size and cleaned before we use machine learning to answer our question. In the data wrangling process, we started with 2240 entries and 29 columns.

The two features Z_CostContact and Z_Revenue were uninformative, they had the same value for all entries (probably created for another model use). These columns were dropped.

Checking for the validity of the data, mistakes were found and outliers were detected: for example, customers’ age that didn’t make sense and unusual numbers in income. These few rows were dropped. The marital status that didn’t aline with most groups was classifieds as ‘other’. There were only a few cases, they were not dropped since marital status is not absolutely critical for our clustering model and we don’t want to lose more information(rows).

After checking for duplicates, there were no duplicated rows.

In the income feature, 24 missing values were detected and replaced with a median. Since the max value is very high compared to the mean and the distribution is skewed the missing values were replaced by the median and not the mean.
Preprocessing: As preparation for machine learning, the data were scaled. In the beginning, the few categorical features were dummy encoded. But after final decisions on the features that will be used in the model, the categorical features were not used.

EDA- exploratory data analysis

After identifying all useful features and cleaning the data, there was a need for feature engineering. New features were created for machine learning use. For example, converting dates of becoming a customer to a number of days since becoming a customer'days_customer'. More features: “Total products’- this was the sum spent on all products. This feature was created by summing the separate features of each product. Total spent- this is the number of purchases(not the sum spent).“Any_promo”- is the sum of all promotions ranging from 0 to 5 as there were 5 promos in total. This was created by summing all promo features for each customer.

The next step was visualizing the distributions and the correlations of all features. We discovered some interesting initial understanding of the data:
Income is positively correlated with any kind of purchases, mostly meat wine, in-store, and catalog. This means that higher income is connected to spending more.

The total number of purchases also correlates with how much people spend, slightly negatively with having small kids, and also responding to deals. We still don’t know, which groups of customers are responsible for making parentheses and responding to deals. People who complain don't buy less, meaning that complaints are not predicting what we want to know: purchases, responses to promotions, and spending on products. This feature is not as important as we could have assumed.

What do we know about responding to promotions?
Wines were more correlated with responding to promotions. Previous responses to promotions correlated with recent ones.
Web visits and purchases: Income negatively correlated with web visits. Also, webvisits negatively correlated with buying fresh foods like meat and fish, those purchase mores via catalog, next in store. Web visits are not correlated with web purchases but negatively with catalog or store visits. Other purposes or products.

Demographics: People with no kids make more web purchases. Kids at home, correlated with less this type of purchasing, more web visits, and fewer purchases in all locations.It also looks like store purchases are distributed evenly across ages except for the older and younger ages which makes sense. Also by regressing total products and how many days someone was we could see that customer the longer the customer the more they spend!

This exploration, gave a general idea about the customers' behavior, income, purchases, and some demographics. It looks like it wasn't an easy task to segment these buyers. Except for income and having small children we didn't discover many coherent meaningful groups yet. The next step was selecting the features that will be used in the model.
Since we want to know who are the people that make more purchases, spend more, and respond to promotions and deals, these features were selected for the model: 'NumDealsPurchases', 'Response', 'any_promo', 'Total_products', 'Total_spent'.

Model selection:
Unsupervised learning with clustering algorithms was performed. For the purpose of customer segmentation, only algorithms that can predict new data were selected, because the goal was to create a model that can be used for new unseen data.

The models that were created with hyperparameter tuning, were also evaluated for best separated (inter clusters distance) and dense (intra-cluster distance) clustering by calculating the Silhouette score.
For parameter tuning, grid-search and randomized search were tried, but they didn't optimize the models' parameters as expected. A manual tuning function was used for hyperparameter tuning.
These are the results for all models:

Model	Silhouette score
KMeans(n_clusters=5)	0.459
KMeans(n_clusters=5) with hyperparameters tuning	0.459
AffinityPropagation(n_clusters=65) with hyperparameters tuning	0.458
MeanShift (n_clusters=3)	0.45
MeanShift() with hyperparameters tuning	only one cluster was generated and the score could not be calculated
Birch(n_clusters=4)	0.423

Birch(n_clusters=2) with hyperparameters tuning:
Cluster_0= 2116 entrees
Cluster_1= 121 entrees
	0.482

BisectingKMeans(n_clusters=2)	0.377
BisectingKMeans(n_clusters=6) with hyperparameters tuning	0.429

Model	Silhouette score
GaussianMixture(n_components=7)	0.355
GaussianMixture(n_components=8) with hyperparameters tuning	0.454
KMedoids(n_clusters=7)	0.245
KMedoids(n_clusters=7) with hyperparameters tuning	0.442

Silhouette score visualization?

Medel selection:

The best score was 0.482 for Birch model with hyperparameters tuning. It has shown two clusters: cluster_0= 2116 entrees and cluster_1= 121 entrees.
The second best was KMeans with or without parameter tuning- its score was 0.459 with 5 clusters. I decided to explore both options and decide on the best model to apply. To make the decision I explored the groups, and how informative the generated clusters are for my purpose. I didn't want to select just the highest score, but rather to select the model that best answers this project's main question.
The 2 clusters have shown us the "top" customer group, which is the smallest group with high income, high spending, and was responsive to promotions. The second group seems to include all other customers.
With k means there were 5 groups. One similar "top customers" group but also "second highest spending" and "distinct low income, low purchasing" group and two more.
This Birch model doesn't separate these groups and put them into one.
The Birch model with 2 clusters, in my opinion, oversimplifies and lacks important information to better understand the store's customers. It focuses solely on the "top customer" and assumes that all other customers are the same in terms of buying and promotions.It also doesn't help the store to decide on steps that need to be taken for better marketing and better-customized services. The second-best score with 5 clusters gives the store more information about the other groups, who are not the top customers, but have different characteristics and may need different attention. For example, group 2 has mostly married people with children and has the highest rate of deal purchases, although it doesn't respond to promotions. Customers in this group have also responded best to the fourth promotion compared to other promotions. It is clearly a different group with different characteristics.

Conclusion

After deciding that focusing only on the top customers wasn't the only goal of this project and will only partially answer the question, the Kmeans was selected as the best model, with 5 clusters and a Silhouette score of 0.459.

Information from 5 clusters:

Overview, demographics, income, and spent on products:

The clusters created based on the KMeans model, are 5 groups. The biggest group of customers (cluster 4) is the lower income group, 998 people. The smallest cluster (cluster 1) is the higher spending, higher income group with 164 customers. The second biggest is the second in paying more and the second highest income- 628 customers. The income was different for each group: group 1 has the highest income, then groups 0, 2,3, and 4 was the lower income group. This is parallel to how much they spend on products. The total of purchases(total spent)was the same, highest in group 1. then 0 ,2,3,4. but taking into account the standard deviation, the difference in purchases is probably not significant in most groups except the lower income 4 and possibly 3. Also, the overview shows that all groups have mostly a graduate level of education and they are mostly married, except group three. But, there wasn't a difference in average age, except for group 0 and 2 which has fewer younger customers, born late 80’s, beginning of 90’s. Also no difference in how long they have been customers. It looks like complaints are rare, but they come mostly from groups 0, 3, and 4.

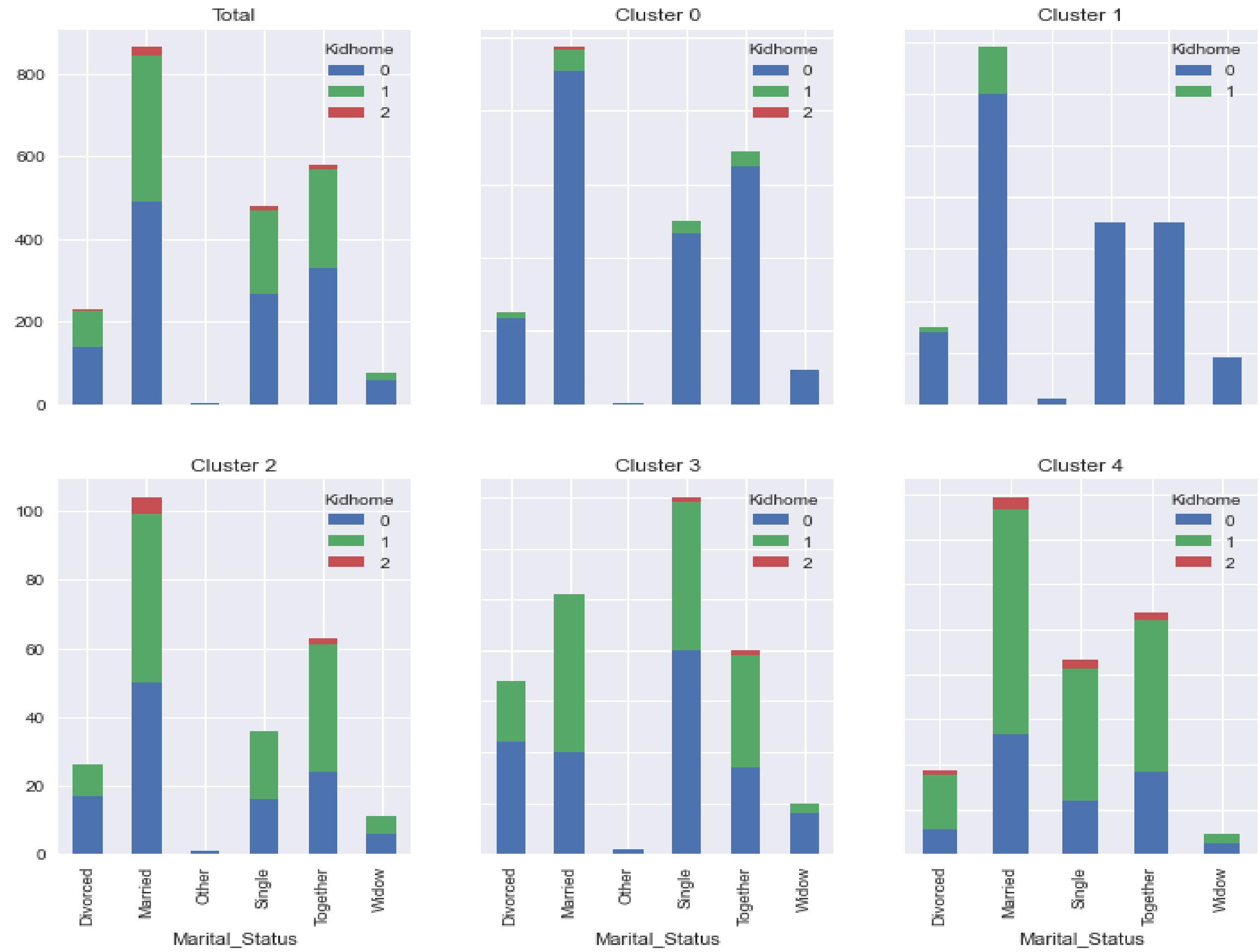
The first group 1 with the highest income and spending. It looks like the people in our top spending customers group mostly don't have kids or have one teen at home. The data also shows that the group with higher income, also spends more on wine, meat and fish products, buys in-store, and accepts more promotions. It looks like we identified the store's "top customer", who will respond well to promotions and will buy more. This group might be already targeted, but now that it's well identified it can receive even more personalized services/promotions.

The second group with higher income, that spends more is group 0. This group buys in-store and doesn't respond much to promotions. This group needs more attention as they are a bigger group with buying ability that usually doesn't respond to promos. There is a big potential for change in promotions to specifically target this group, but also all the others that are not responding to current promos.

Groups 2 and 4 are mostly married or with partners, mostly with children. Only group 3, were mostly single, one-third of the customers in this group, the rest are married or with a partner, also, it has slightly more divorced people than other groups around 16%. Two third of people in the third group have at least one child(different variations of young kids and teens), and a third have two children.

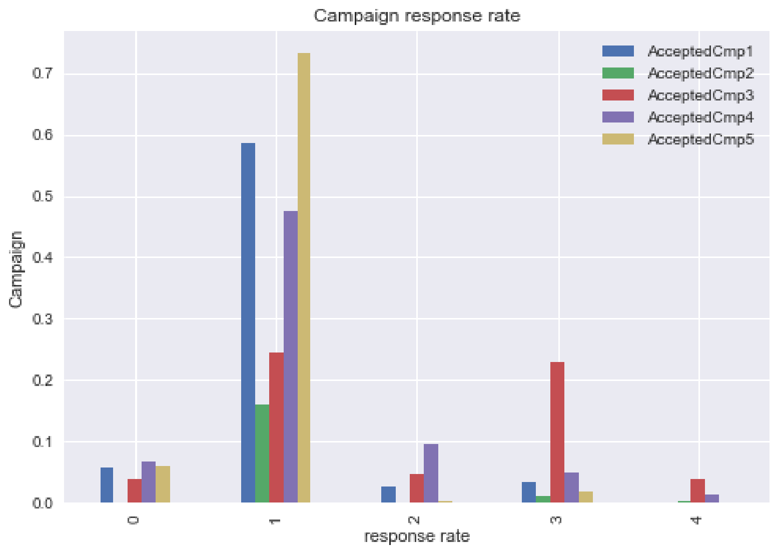
Lower paying, lower income groups visit the website more frequently per month than the top paying customers. Interesting to investigate if the promos are not on the website(more in-store?).

The next gragphs show marital status and having small kids in each cluster:



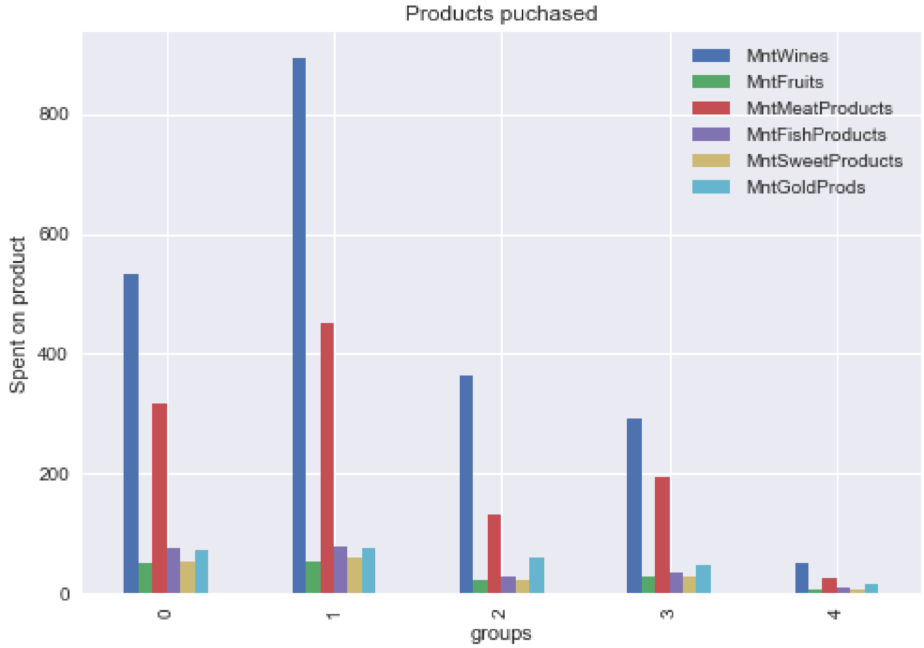
Summary of promotions:

This looks like group 1 was responsive to all promos, mostly the fifth promo. The second promo had the lowest response rate among all groups. Promo 3 was appealing to groups 1 and 3 , having similar response rates. Promo 4 had the highest response from the other groups, not only group 1. It looks like clusters 1 and 3 were most responsive to the last campaign. This means that the last campaign targeted the high-paying customers as usual and missed the second high-paying group, where no one responded to the promo. Everybody in group three responded to the promo- so this can be repeated if the store wants to target again this group. Also, the last promo didn't match the biggest group (with lower income) needs at all. It looks like group 2 has the highest rate of deal purchases, group 3 is the next to make deal purchases. This store might consider these two customer groups as targets of the store's special deals. Group 4 wasn't responsive to other promotions, but was responsive to deals. Considering this is the largest group, it is important to know what type of marketing strategy works for this group.

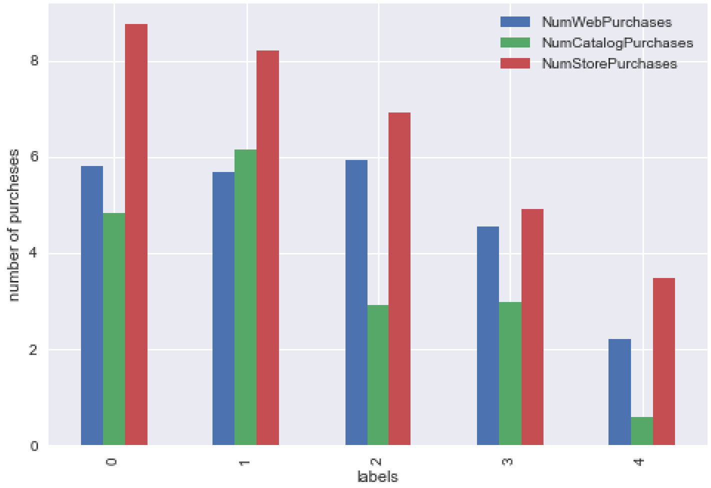


Preferred products by clusters:

The segmentation by clusters and the product groups shows a consistent picture, where groups 0 and 1 spend more in all product categories. In some cases, the difference between these two strong buyer groups is not significant because of the high standard deviation, but they notably spend more than all other groups. It wasn't surprising that people in all groups spent more on expensive products like wine and meat. This could be also the store's strongest/most popular products that interest their customers. All groups spent less on the rest of the products, except for gold. The "top" groups 0 and 1 didn't spend on gold as much as the other groups in proportion. This means gold was in a lower proportion of all of their purchases, they might prefer to buy their gold in a different store.



Preferred location: All groups make more store purchases. Group 1 also makes more catalog purchases, second after in-store. There is a potential to increase some of the web (or catalog) purchases in all groups! Groups 0 ,1 and 3 make most purchases in-store. Groups 0, 2,3, and 4 make most purchases in-store and the next preferred place is online.



Summary of the clusters and recommendations:

Cluster 0: 628 customers

- Second highest income
- Second in purchases and spending on products
- Married or with a partner, mostly
- Have fewer younger people.
- Level of education: 50 % graduate level and almost 40 % higher education.
- Mostly don't have small kids. 50 % have no kids and the rest have mostly one teen.
- Don't respond to most promotions. Somewhat responses to deals (lower response than other groups.
- Buy in-store mostly, next preferred is online. Don't visit website as much.
- Spend on meat and whine and fish, less on gold.

Cluster 1: Top customer group 164

- Highest income
- Highest purchases and spending on products
- Married or with partner
- Level of education:50 % graduate level and 42 % higher education.
- Mostly don't have kids. 20% have one teen at home.
- Respond to most promotions. The relatively low response rate to deals (lower response than other groups.
- Best promotion was:
- Buy in-store mostly, next preferred is catalog. Don't visit the website as much.

- Spend on wine, meat, and fish, less on gold.

Cluster 2: 241 customers

- Third highest income
- Third in purchases and spending on products
- Married or with a partner.
- Have fewer younger people.
- Level of education: around 40 % graduate level and 40 % higher education.
- Mostly have kids. about 60% have one teen, while some have one teen and 1 younger child.
- Don’t respond to most promotions. The highest response rate to deals. The best promotion was no 4.
- Buy in-store mostly, next preferred is online. Visit the website frequently.
- Spend on wine, meat, and the next gold.

Cluster 3: 206 customers

- Forth in income
- Forth in purchases and spending on products
- Single mostly, also a higher rate of divorced individuals.
- Level of education: around 40 % graduate level and 40 % higher education.
- Mostly have kids. Two third of people in the third group have at least one child (different variations of young kids and teens), and a third of all people in the group have two children.
- Don’t respond to most promotions. The best response rate was to promotion no. 3. Does respond to deals second after group 2.
- Buy in-store mostly, the next preferred is online. Visit the website frequently.
- Spend on wine, meat, and the next gold.

Cluster 4: 998 customers

- Lowest in income
- lowest in purchases and spending on products
- Single mostly, also a higher rate of divorced individuals.
- Level of education: around 40 % graduate level and 40 % higher education.
- Mostly have kids or teens.
- Don’t respond to most promotions. Does respond to some deals third after group 2.
- Buy in-store mostly, the next preferred is online. Visit the website frequently.
- Spend on wine, meat, and the next gold.

Recommendations:

1. Top customers, group 1 can have more specific attention. As this is a group with strong buying potential, more time should be spent personalizing service to this group. Fro example, this group can receive a catalog that serves their needs, as they also buy using

the catalog. Promoting popular products of this group like wine and meat will benefit in general but for this group can be especially beneficial. Also, recognition as a prime customer can encourage to buy,   if this group feels prioritized. There is a need to invest more in gathering information about this group to continue promotions that match this group's needs.

2. Group 0 , second in buying ability might be not interested in promotions because of their characteristics, fewer young people, and they have fewer kids, but probably because they are not targeted well enough. Again by gathering even more information about this group, it can be targeted better and respond well to promotions, eventually also buy more.
3. In general invest in products that people spend more money on like wine, meat fish, and gold. Having data about profit from these products can lead to more precise calculations and investment in promoting the selected products.
4. Group 4 is buying less, but it’s the biggest group and a small investment of time finding their preferences can also bring a big profit. Again specifically targeting this group with deals can help, as they relatively responded to deals.
5. Groups 2,3 and 4 visit the website more but still prefer buying in-store. The website can be adjusted to this group's preferences and call for action. In general, there is a place for improvement for all groups.
6. To have more information on each group we can find their preferences by surveys and A/B testing on the website or experimenting in-store.
7. The store can prioritize time investment in specific groups over others.
8. The store can repeat what worked, and check what exactly worked for each group. For example, promo 4 worked for group 2. This group has many teens at home. If this promo targeted products that teens like, this means that promoting this kind of products will be effective in this group.
9. More information needs to be collected: gender, website traffic, what times a day people respond to deals in targeted groups, what time buying in-store online, seasonal buying , Preferred brands, More specific shopping cart descriptions for each group and much more.