# Capstone Project- Marketing Campaign Analysis

A Sample Report by Kaitlin Bennett

## Introduction

Grouping like things together can be as simple as sorting apples from oranges, or as complex as sorting cancer cells and healthy cells under a microscope. In either example the success of the sorting is the determining factor in the success of the outcome. Customer segmentation dwells in the same realm in terms of how vital the clustering method being implemented is for the success of the business.

This report focusses on our clients dataset of customers and their demographic and behavioral attributes over a two year period, collected in the year 2016. Our client is a business selling various food and beverages through three distinct avenues: store, website and their catalog. From income to family size, they see a wide variety of customers purchasing their products but customer engagement with their previous marketing campaigns is low. Our client is looking to increase sales and productivity by effectively targeting their customers according to their profiling segments.

## Executive Summary

By implementing unsupervised learning methods such as dimensionality reduction and clustering we will be better able to identify patterns in our clients dataset of customers.
To better understand our customers and increase ROI we will be using customer segmentation that will group our clients into 'like' clusters. This will help identify the customers' preferred method of purchasing and maximize the client's marketing potential.

To achieve this goal we preformed data analysis and dimensionality reduction that gave us insight into the customers' defining attributes. When we began exploring our clustering methods the algorithm's success was judged based on the rank of their silhouette scores. We chose the method with the highest silhouette score to ensure that our clusters would be grouped based on their cohesion compared to other clusters. Our most successful clustering method that gave insight into purchasing patterns while also providing diverse customer profiles was K-Means with k equal to 5 which gave us a silhouette score of 0.24. This method gave us five distinct clusters that were first segmented according to whether their income was high, middle, or low and then separated those classes further according to age, family size and buying habits. By implementing this clustering method we were able to identify patterns in the customer dataset that allowed for more effective marketing campaigns; thus increasing productivity and sales.

At this point we need to consider that there are other methods that may be good at forming distinct clusters. While we may want to consider these methods in the future, for the time being they are too computationally expensive and therefore K-Means is our best option. Our stakeholders should be made aware of the value that using K-Means brings to the table without taking too much time, money and resources to run.

We can confidently proceed with the implementation of our model, knowing that the clusters that were created are giving us the best possible outcome based on our carefully chosen segmentation attributes and the precise method we used to procure them. We are able to draw viable conclusions about our clusters in regards to buying habits and how receptive each group will be to various types of advertising.

# Problem Summary

In the world of business, customer segmentation creates invaluable results for how best to tailor marketing and sales strategies within different groups. When done well, customer segmentation can lead to increase in ROI and sales because it allows businesses to better understand their customers' behaviors and inducements. Having a better perception on our customer dataset will also make it easier to target potential customers in the future.

The central purpose of this project is to build a clustering model that will effectively group our dataset into groups of similar customers so that we can customize our marketing strategies. The clustering methods being proposed will divide our customers into groups according to their buying habits and preferred method of purchasing and campaign interaction.

Our client is currently not seeing many of their customers accepting the offers that they have been promoting through their advertising campaigns which is a poor use of their marketing division as seen in **Figure 1**. Out of their 2,240 customers, 50% have not accepted an ad from one of their 5 campaigns. The mean for total accepted campaigns is 0.45 and Q3 is 1 leaving a lot of room for improvement.

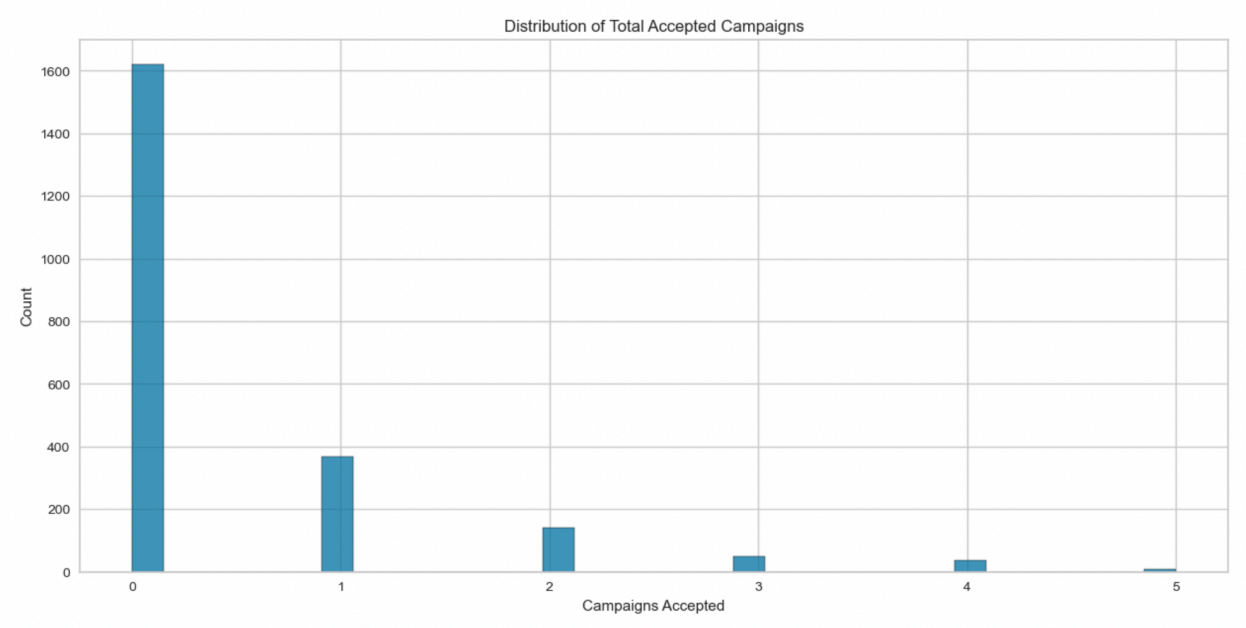



Figure 1: Distribution of Total Accepted Campaigns

# Solution Design

**Overview**

Many different methods were applied to our dataset to determine the best number of clusters and the best clustering method to use. We had to pay special attention to which method gave us more than just basic insight into our customer profiles as well as what method would be best to recommend from a business perspective. **Our recommendation with this in mind was to go forward with K-Means, k=5 which gave us a silhouette score of 0.24**. The clusters made from this method can be seen in **Figure 2** below. These clusters were divided not only by income level but also by family size and age as we will see later in our cluster profiles.

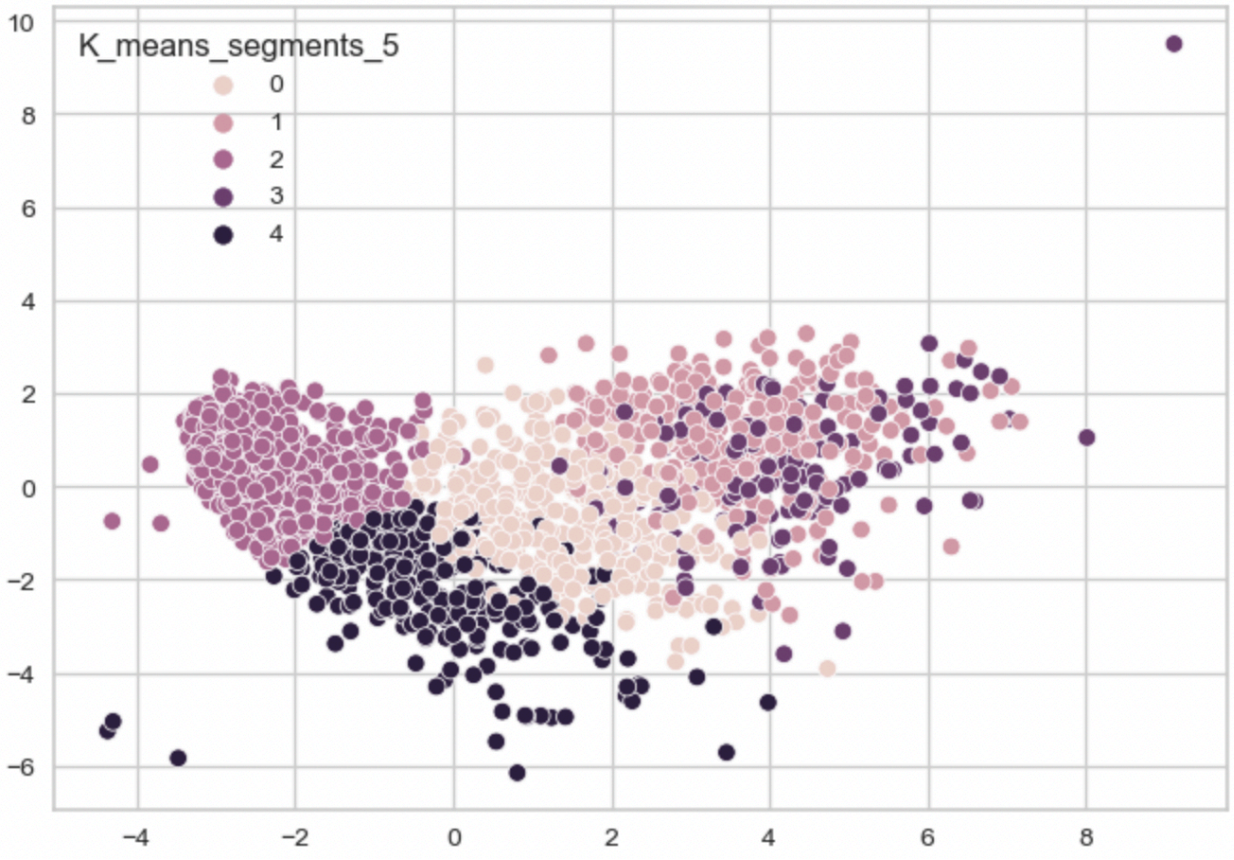


Figure 2: K-Means Clusters with k=5

**Segmentation Preparation**

Before building our models we had to prepare our data for segmentation. We did this by dropping the columns that were demographic in nature, using them later for our profiling. We then used our behavioral data (amount spent, purchasing method ect.) for segmentation. By doing this we limited the number of variables that would influence our clusters down to 17.

**Feature Scaling**

We needed to scale our data so that all our features would have a mean of 0 and a variance of 1. We then applied t-SNE and PCA to the data to visualize its distribution in 2 dimensions and chose to go with PCA because our variables were highly correlated and PCA helps to reduce multicollinearity.

**Applying Clustering Algorithms**

As we sought to apply the best clustering method for our dataset we did a thorough exploration of the following algorithms:

- Elbow method to determine best number of clusters
- Silhouette score was used to determine best number of clusters
- K-Means clustering
- K-Medoids clustering
- Hierarchical Clustering- Cophenetic Correlation
  - Citiblock
  - Chebyshev
  - Mahalanobis
  - Euclidean
- DBSCAN clustering
- Gaussian Mixture Model clustering

In building our models we first started with the most simple model, K-means. **Figure 3** shows the elbow plot we made to see what the ideal number of clusters would be and saw that 3 or 5 would make an optimal value for k.

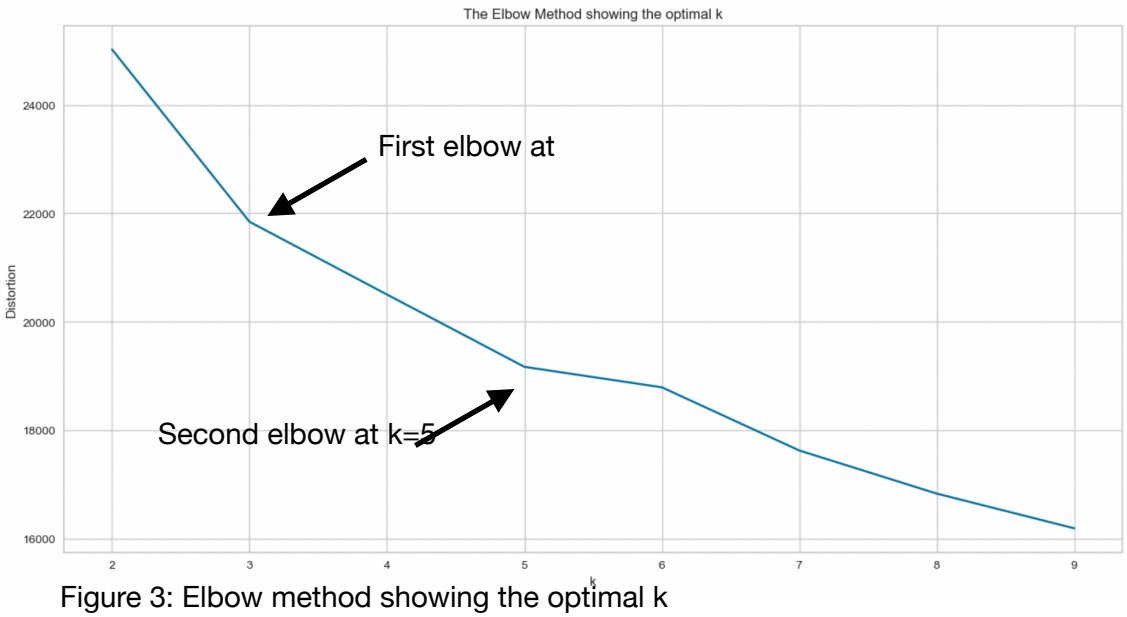



Figure 3: Elbow method showing the optimal k

We executed different clustering algorithms with k equal to both 3 and 5 and found that k=5 gave us more insight into our customers' purchasing behavior and their marketing preferences. We then went on to calculate silhouette score for all of our clustering methods beginning with K-means where our silhouette score for 3 clusters was 0.27 and our silhouette score for 5 clusters was 0.24. The clusters for K-Means with k=3 that gave us our best silhouette score can be seen in **Figure 4**. While this method was a good option that was computationally inexpensive it only showed us basic clusters that lacked complexity.

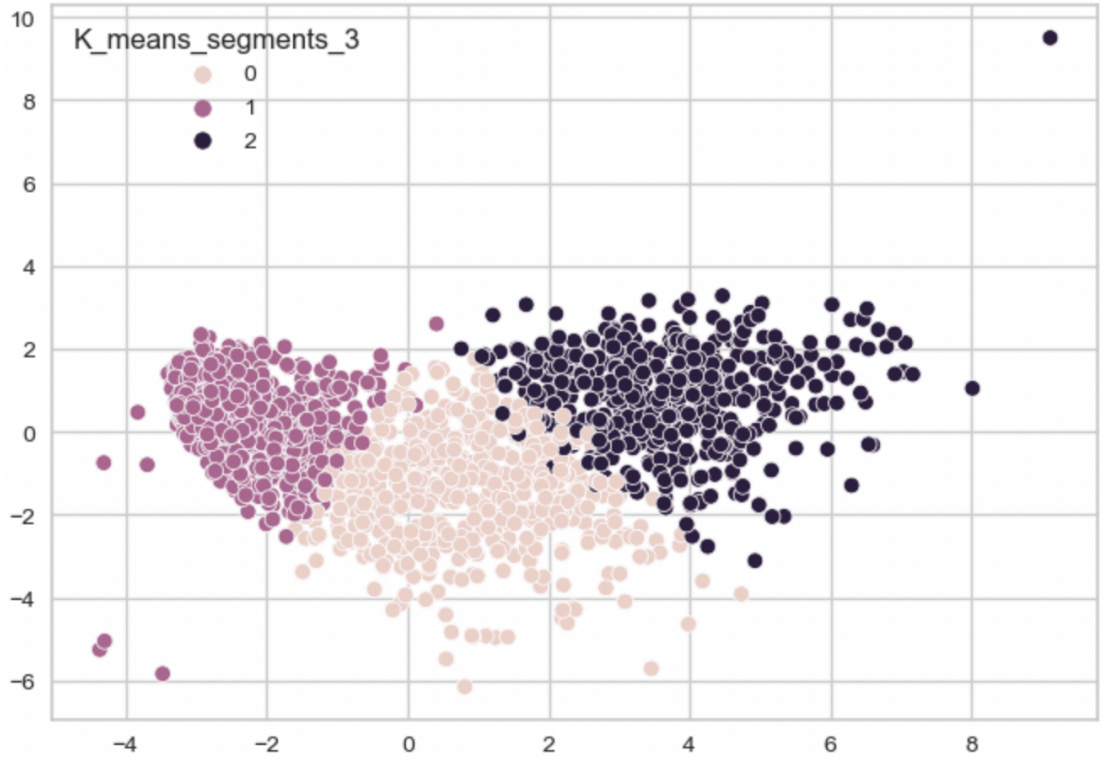



Figure 4: K-Means with k=3

Second we preformed K-medoids clustering to see if our data responded better. This method is good at reducing noise and outliers and it gave us a silhouette score of 0.11. The clusters for this method were similar to that of K-means but with a lower silhouette score.

Hierarchical clustering led us to find the cophenetic correlation for different distances using different linkage methods. Our highest cophenetic correlation is 0.86, which is obtained with Cityblock distance and average linkage. This method gave us three clusters but 2 have only one customer in them showing this method to not work well with our data.

We preformed our hierarchal clustering method using euclidean as our distance and ward as our linkage which gave us a silhouette score of 0.25. While this silhouette score was high it only gave us three clusters because our dendrograms did not indicate that 5 was advisable.

When we preformed DBSCAN we found our best silhouette score for eps value =3 and min sample =20. The silhoutte score for this method was 0.34, our best so far, however it divided our data into just 2 clusters which would not give us enough information about our customers and this was also the reason why the silhouette score was so high, see **Figure 5**. The issue with using DBSCAN is that there is overlap in our clusters and so multiple clusters got grouped together into one large cluster. We changed the hyperparameters to see if we could get better results without success.

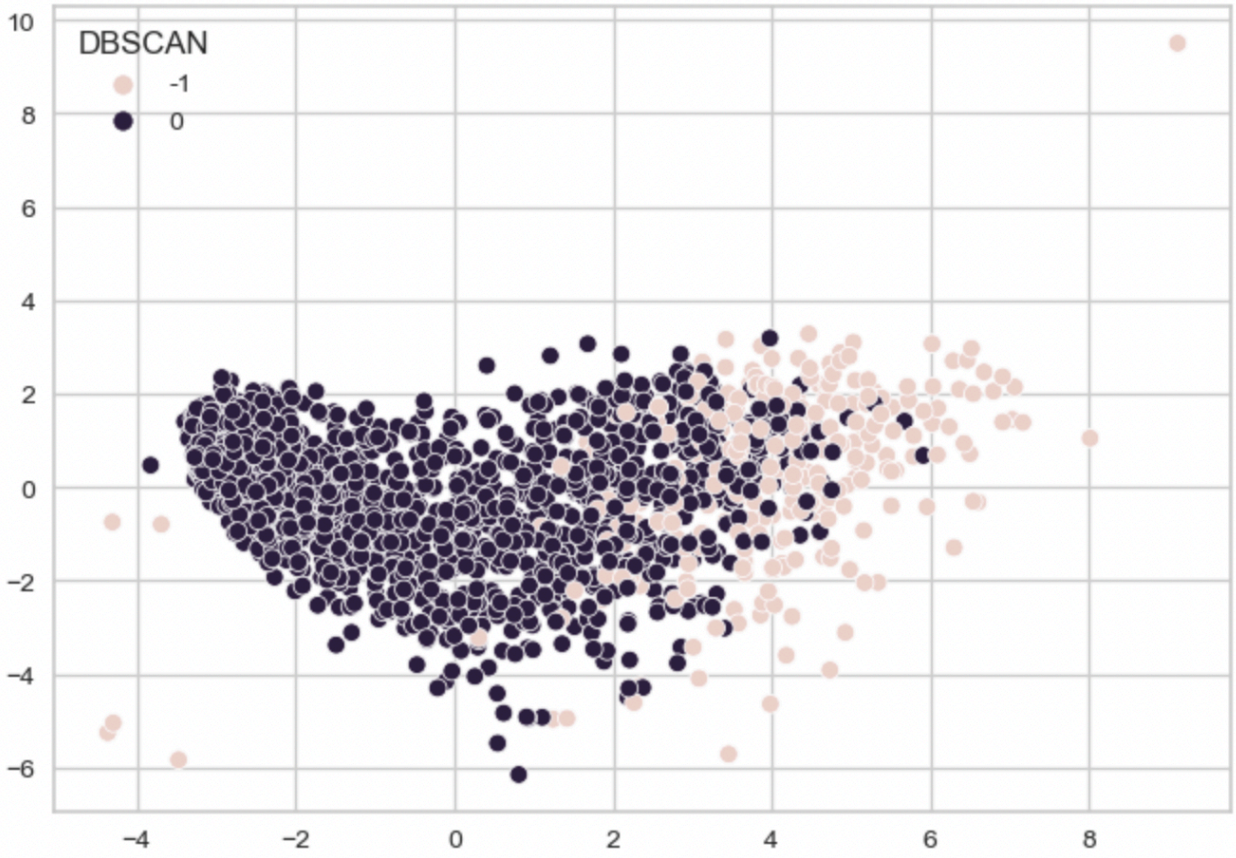


Figure 5: DBSCAN with eps=3 and min sample=20

Lastly we used the gaussian mixture model with k=5 and got back a silhouette score of 0.14. With k equal to 5 this gave us the best silhouette score after K-means, out performing K-medoids.

Of the clustering methods that divided our data into 5 clusters K-means preformed the best. We discovered early on that 5 clusters would give us the most insight into our customers' buying

habits and this remained true throughout the analysis. Below in **Figure 6** is the table summarizing our findings from each clustering method.

### Comparison of Techniques and their Outcomes

| Clustering Algorithm | Best Method | Silhouette Score | Key Points |
|---|---|---|---|
| K-Means | k=3 | 0.27 | Most obvious elbow visible at k=3 and best silhouette score |
| K-Means | k=5 | 0.24 | Elbow visible at k=5, these clusters gave more insight |
| K-Medoids | k=5 | 0.11 | We observed similar clusters to K-means. |
| Hierarchical Clustering | Euclidean distance ward linkage k=3 | 0.25 | High silhouette score with k=3 but computationally expensive. |
| DBSCAN | eps=3 and min=20 | 0.34 | There was overlap in our clusters so multiple clusters got grouped together into one large cluster. |
| GMM | k=5 | 0.14 | Lower silhouette score but this method is better able to handle overlapping. Similar to K-means |

Figure 6: Comparison of Techniques and their Outcomes

# Analysis and Key Insights

In order to make valuable conclusions about the results of our clusters we needed to make sure that we had a good understanding of our data first. To prepare our data we first established that all our variables fell into one of three categories: demographic data, purchase data, and engagement data. In examining the data we found that there were no missing values in any of the columns except for income which was missing 24 values or 1.07%.

### Data Exploration

This data collected in 2016 contains information on customers and their spending habits and interaction with the business. We are not only given demographic data but behavioral data as well. Our data is in the form of both numerical and categorical data.

Our demographic data contains the following categories:

- ID: Unique ID of each customer
- Year_Birth: Customer’s year of birth
- Education: Customer's level of education
- Marital_Status: Customer's marital status
- Kidhome: Number of small children in customer's household
- Teenhome: Number of teenagers in customer's household
- Income: Customer's yearly household income in USD
- Recency: Number of days since the last purchase
- Dt_Customer: Date of customer's enrollment with the company

Our consumer purchase data contains the following categories:

- MntFishProducts: The amount spent on fish products in the last 2 years
- MntMeatProducts: The amount spent on meat products in the last 2 years
- MntFruits: The amount spent on fruits products in the last 2 years
- MntSweetProducts: Amount spent on sweet products in the last 2 years
- MntWines: The amount spent on wine products in the last 2 years
- MntGoldProds: The amount spent on gold products in the last 2 years
- NumDealsPurchases: Number of purchases made with discount
- NumCatalogPurchases: Number of purchases made using a catalog (buying goods to be shipped through the mail)
- NumStorePurchases: Number of purchases made directly in stores
- NumWebPurchases: Number of purchases made through the company's website
- NumWebVisitsMonth: Number of visits to the company's website in the last month

Our engagement data contains the following categories:

- AcceptedCmp1: 1 if customer accepted the offer in the first campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the second campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the third campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the fourth campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the fifth campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise
- Complain: 1 If the customer complained in the last 2 years, 0 otherwise

**Observations and Insights**

We preformed basic statistics to gain insights from the data. We found the average, min, max, Q1, Q2, and Q3 of all our columns. It was here in our basic summary statistics that were able to see that there were low interactions between customers and the previously ran campaigns. At least one person interacted with each campaign but 1 was the max and Q3 was 0.

It was in this data treatment and preprocessing phase that we were able to identify columns that didn't contribute any value to our analysis as well as misleading data and outliers that would negatively affect our clustering algorithms. Listed below are some of the modifications that we made that were the most important and had the greatest impact on the successful implementation of our clusters.

**Data Treatments/Pre-Processing**

- Birth year was showing us a minimum of the year 1893. We discovered that there were 3 individuals who's birth year put them over the age of 115 so we dropped these.
- For income the third quartile was 68,522 while the maximum value was 666,666. We found that 99.5% percentile value for the income variable was 102,145.75 and so we dropped the 8 rows who's outliers were affecting the data.
- In looking at the summary statistics for our categorical data we discovered that we could simplify the sub-categories for Education, Marital_Status, Kidhome and Teenhome.
- Over half the customers, 57.5%, have 0 kids in their home and 51.6% have 0 teens in their home. For those who do not have children, their income is higher than those who do.
- At least one customer accepted each campaign. Campaign 2 was accepted the least and campaign 4 was the most accepted.
- We preformed bivariate analysis and found positive correlation between income and purchases of wine (.73), meat (.70), store purchases (.69) and catalog purchases (.71).
- We also found positive correlation between catalog purchases and wine (.67) and meat purchases (.64).

- People with higher income are making more pricey purchases via the catalog in the form of meat and wine purchases. They are also purchasing more items from all the food categories over all. More income equals higher budget to spend on various food items.
- We saw negative correlation between the number of web visits and income (-.65), meat purchases (-.54) and catalog purchases (-.53).
- Number of deal purchases is most positively correlated with number of web visits (.37), web purchases (.25), and very interestingly having teens in the home (.40).
- We created a new feature ‘Expenses’ by combining the total amount spend on MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', and ‘MntGoldProds’.
- We created a new feature ‘NumTotalPurchases’ by combining the 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', and ‘NumStorePurchases’.
- We created a new feature ‘Engaged_in_days’ by taking our threshold date, 1-1-2015 minus ‘Dt_Customer’.
- We created a new feature ‘TotalAcceptedCmp’ by adding ‘AcceptedCmp1’, ‘AcceptedCmp2’, ‘AcceptedCmp3’, ‘AcceptedCmp4’, ‘AcceptedCmp5’, and ‘Response’.
- Finally we analyzed our new feature ‘Expenses’ with our new ‘Income’ variable and found a positive exponential relationship that will not fit well with a linear model.

The following are the insights that were gathered on our data based on our preferred clustering method K-Means with k=5. These insights are made while keeping the above observations in mind.

**Meaningful Insights for K-means- k=5**

**Cluster 0:** Middle Income with Kids and Oldest
- This group is the second largest cluster made up of 403 of the oldest customers, the average age being 50. The have the medium income of all the groups at $63,397.92 and have more teens in their home with a family size of 2.5. They spend the medium amount on all food purchases and have the second lowest recency, shopping more frequently than 3 other groups. This group has the highest frequency for in store purchases, web purchases, and total purchases which come to 21.53. They also have the highest number of complaints. Despite making the most purchases they are the medium group for expenses (797.87) and total amount per purchase (37.40). They responded worst to campaign 2 and best to campaign 4 and have been customers for 544 days.

**Cluster 1:** High Income with no Kids and Older
- This group is the third largest cluster made up of 357 customers whose average income is $74,565.29 which ranks them second highest for income. The average age in this group is 48 and they have the least number of children in their homes, average family size being 1.8. This group has the highest recency, making purchases less frequently but they have the highest dollar amount spent on fruit, fish, sweets, and gold products. They are second highest for catalog and store purchases and second lowest for web purchases. They are also lowest for web visits which makes their low web purchases make sense. They have the second highest expenses (1257.18) and amount per purchase (63.25) which checks out because they are not spending the most for meat and wine which are the most expensive items. Their higher purchase amount may account for their higher recency. In spending more per purchase they are needing shop less. They responded best to campaigns 1 and 5 and worst to campaign 2 making them the medium group for total accepted campaigns and they have been customers for 531 days.

**Cluster 2:** Low Income with Kids and Youngest
- This group is the largest group made up of 994 customer with the lowest income of $34,929.67 and the highest number of young kids in the house, family size being 2.9. The average age in this group is 44 making them the youngest and they have the second highest recency and the lowest numbers for purchases of all the different food categories. This fits because if they are not shopping very often they will not be making as many purchases. Also lower income

customers will be spending less overall than higher income customers. They are the middle group for deals purchases and the lowest for store, web and catalog purchases- again this makes sense. They are second highest for the number of website visits per month and this coupled with the low income category and higher deals purchases indicates that they are mostly shopping when they have a deal and are regularly checking for deals. The are the newest customers (491 days) with the lowest number of expenses (77.13), and purchases (8.97). They were most receptive to campaign 3 and least receptive to campaign 5.

**Cluster 3:** High Income with no Kids and Young

- This group is the smallest group made up of 165 customers with the highest income of $79,500.80. They have no children, average age of 46, and have spent the most on wine and meat of all the groups by a large margin. Their wine purchase are almost double that of the next closest group. They have the highest number of catalog purchases and are the middle group for web purchases and store purchases. All of this information adds up because they have the highest income and in not having children they are able to spend more on non essential food items like alcohol. They are lowest for the number of deals purchases that they have made, indicating that the prices of the items that they are buying are not an issue for them. This group has accepted more of the campaigns than any other group with campaign 5 being the best and campaign 2 being the worst. They have the highest expenses (1601.02) and the second highest total number of purchases (21). They spend the most per purchase (87.93) and have been customers for 596 days, the second longest of all the groups. **These are our key customers and utilizing marketing strategies to get them to bring in more business is vital**.
-

**Cluster 4:** Low Income with Kids and Older

- This group is the second smallest group made up of 308 customers with a low income to middle income of $48,841.00. The average age is 49 making them the second oldest group and they have the largest family size of 3 with the highest number of teens in the house. They are second lowest in all purchases but are ranked highest for making purchases with deals. This checks out as they have a lower income they are waiting for good deals before making purchases. They are second highest for website purchases and second lowest for catalog purchases and in store purchases. This adds up due to the fact that they have the largest family size, thus making purchases on the website is most likely more convenient. They have the highest number of website visits per month which can also be attributed to them shopping for the best deal. They are the middle group for total accepted campaigns with 4 being their best and 5 being their worst. The have been customers the longest at 658 days which may also account for them being the highest users of deals. In being more familiar with the company they may have been knowledge of what goes on sale. They are the second lowest for expenses (455.26, total purchases (19), and amount per purchase (22.53).

# Recommendations

These clusters are clearly defined according to income, family size, age, purchasing habits, and campaign interactions. We can also make some valuable suggestions based on the trends in the clusters as to how to treat each group going forward.

**Cluster Recommendations**

**Cluster 0:** Middle Income with Kids and Oldest

- As this group is highest for web and store purchases with the highest number of total purchases we can conclude that they are doing a lot of shopping on a frequent basis. Recommendation systems can be utilized for this group to send them ads through email and to recommend products for them in checkout on the website. Knowing that they have teens in the house will help to customize their ads toward foods that teens consume more of. Explore campaign 4 and why the responded best to it and mimic those conclusions.

**Cluster 1:** High Income with no Kids and Older
- This group does not utilize the website very much and does most of their shopping in store and through the catalog. Keep sending them the catalog and utilize coupons through the mail that are time sensitive to help get them in the store more as they have the highest recency. Mimic campaigns 1 and 5 as they responded best to these campaigns.

**Cluster 2:** Low Income with Kids and Youngest
- This group is the highest for deals purchases and second highest for web visits. Email marketing to urge them to the website would be beneficial with deals targeted toward foods fit for families, specifically young kids. Campaign 3 was the campaign that they responded best to so having the marketing department explore deeper into proving a similar campaign to them should be beneficial.

**Cluster 3:** High Income with no Kids and Young
- As the highest income group with the highest expenses and campaign interaction this group represent our **key high value customers**. Success with this group spells success for similar potential customers as well. At this time we would do well to provide incentives for them to promote their purchases through social media or other platforms and get them talking about the products as much as possible. We can use the catalog to advertise new higher priced items to them that the business may want to start selling. By using predictive models we can anticipate what this group will be purchasing seasonally and target them for specific items as well. Explore campaign 5 which they responded best to and go forward from there.

**Cluster 4:** Low Income with Kids and Older
- This group has the highest number of teens in the house and the highest for deal purchases. The visit the website the most of all the groups so push deals through email marketing and use recommendations systems for them similar to group 0. We need to keep in mind the high number of teens in this group and the food purchases that are on trend with this family type. Campaign 4 was the one in which they responded to the best and we can base future campaigns for them off of that correlation.

**Implementing Recommendations**

The analysis that we preformed provided us with key insights into our customer dataset. We can make a few recommendations that our client can implement in the future to grow their customer base and manage their dataset and algorithms being used.

Going forward we would like to see more data gathered for younger groups of customers. All of our groups were in their mid to late 40's and the information we could gain from having some younger or older groups might be highly beneficial. This would perhaps open the door for more marketing through social media and the internet which is much less expensive to run ads through than print ads and the catalog.

The easiest way for us to gather new data would be by incentivizing our customers to use an app or the website. If they are creating profiles in order to access coupons we will easily gain information on new customers. In creating profiles we will also be able to ascertain that they plan on being repeat users and are motivated by ad campaigns in the first place. We can also obtain new data through surveys, competitions, transaction history, web-tracking, and promotions. On average companies spend between 5-20% of their annual revenue on consumer research. Given the age and limits of our current dataset we can say that they should aim on the high side for future data procurement. Customer market research projects can range from $20,000-$50,000.

**Maintenance Recommendations**

Cloud services would allow us to worry less about monitoring our changing data and using K-Means would give us some relief on costs versus other methods. Cloud services have flexible spending options that allow you to only pay for what you need. An example would be AWS which does not charge for inbound data transfers. Costs for running our clustering algorithm on the cloud depend on the speed we want, as well as the time it takes to run the algorithm itself. Amazon has SageMaker which offers on-demand pricing and charges per second as it applies to 12 different features. Quotes can be easily obtained through the AWS website.

Having an in house scrapper is another option that would allow us to monitor the incoming data that we would be getting from new customers as well as watch how the changes in our data might affect our clustering method going forward. In house scrappers charge roughly $55 an hour however and that does not include software and hardware costs.

# Risks

There are always risks associated with the implementation of any findings from our data analysis as it applies to marketing decisions and advertising. Producing different marketing campaigns for all of our clusters can be expensive on a mass scale. When considering various needs, behaviors, and preferences, actual business application can get expensive and time consuming.

Changes in the market are also something that need to be taken into consideration. Some of our groups were not made up of a large amount of customers and so crafting specific campaigns for these groups is a risk. It is important to note that these risks can pay off, for instance our high value customers made up the smallest group but they also are the most engaged and spend the most money with the company.

Characteristics in the market can change as well. For instance when we start adding customers with more age variation but similar spending habits, their preferences and habits may be different from what we are seeing now. We risk lumping them into a cluster that does not actually speak to their interests.

# Further Exploration

- We took a closer look to see if changing the hyperparameters would increase the performance of DBSCAN but the problem with overlap of the clusters remained and it could not separate them well enough. We would have liked to spend more time exploring this.

- We also changed k=3 for GMM and was surprised by the results. Our silhouette score only increased to 0.17 but our clusters were very evenly distributed. The clusters themselves were well separated but the profiling lacked some division. GMM might be a good clustering method to explore because some customers might be susceptible to multiple types of campaigns, not just one. People have more than one interest and GMM can account for this. This is a clustering method that should be explored further.

- We could also find Bayesian information criterion for GMM to find the ideal number of clusters without having too many. It would be interesting to see if BIC would give the same number of cluster as silhouette score did.

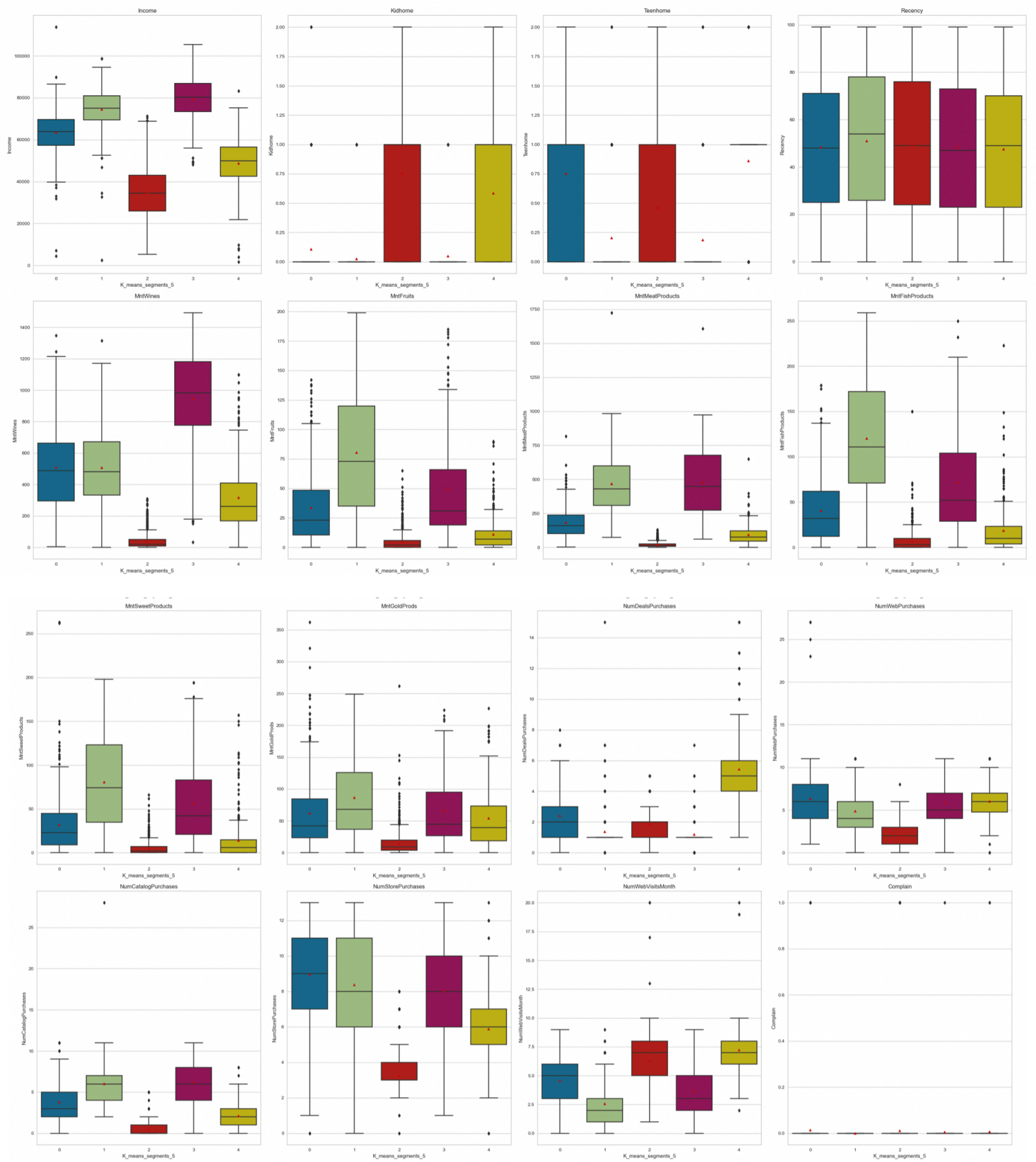
- I would have liked to look at Dunes Index and see if it does better than using silhouette score.

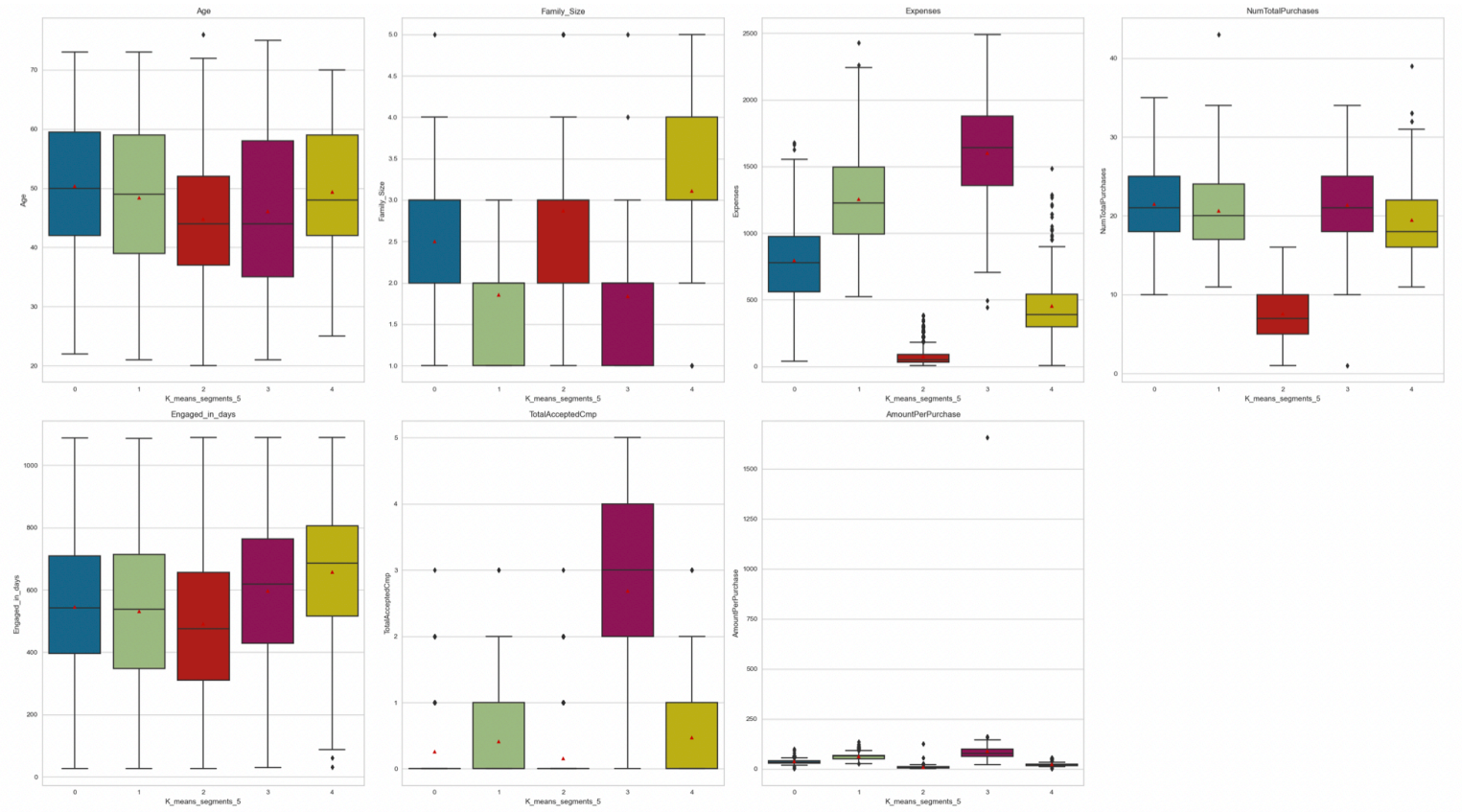
- Another interesting thing to see would be to look at the eigenvalue plot to determine the best number of variables and which ones we want to keep.

- Since we know that our data was overlapping I would like to take a look at OKM (overlapping k-means).

# Appendix

Box plots for K-Means with k=5:






Scatter plot showing positive exponential relationship between income and expenses: