Capstone Report - Market Segmentation
Tashlin Reddy
Data Science Intensive - Springboard

1. Problem to be Solved

According to PwC' research *Fragmentation and Simplification* released on 2015 "in both developed and emerging markets, there is a wider variety among consumers now than at any time in the recent past." Their finding reveals that although there is shrinking in the middle, there is growth at the top and bottom of the market. The top of the market comprises the consumers spending more on higher priced items, as well as buying more items in general. In comparison, the bottom of the market includes an increasing amount of customers focusing on the value of goods. As stated before, the variety in the consumer base is at an all-time high. Consumers are changing and becoming more complex. The industry of marketing consumer goods, therefore, needs to know their customers to the slightest detail on order to remain competitive and not to loose presence in the market.. In the past, the middle of the market has played a major role in driving the consumer goods industry. This segment of the market, however, is shrinking and a readjustment in focus should occur. A proactive approach would entail not only having a profound understanding of their consumers but also making this knowledge a competitive advantage. The question therefore arises: How can the consumer base be targeted in the most efficient possible manner?

2. The Client

Coop is a system of Italian consumer cooperatives which operates the largest supermarket chain in Italy. As of 2010, Coop's system operates with 115 consumers' cooperatives of various sizes (9 large, 14 medium, and 92 small), with 1,444 shops, 56,682 employers, more than 7.429.847 members, and an annual revenue of €12.9 billion. Since Coop. has a customer base it needs to be able to target market their customers. This will help maintain customer satisfaction by tailoring their stores according to different segments identified in our research. Such segmentation will also allow Coop to target new customers.

3. Data

The data is a condensed version of customers shopping habits and distances from 5 of Coop Italia's shops.

Import fields

- Data pertaining to each individual shop
    - distance to each shop
    - amount of product purchased
    - unique products purchased
    - amount purchased
    - average purchase
    - average price of items
- Amount of shops used

- The data collected is a reduced dataset, therefore all the fields are quite necessary in forming a sound analysis. Some of the important data is implicitly stated or can be approximated from the fields given.

For example:

- Shopping Frequency can be calculated from (amount_purchased/avg_purchase)
- If they buy the same items multiple times —> (products_purchased / unique_products_purchased)
- Store Variety - take the max of all the customers unique_products_purchased at each shop.

Drawbacks of the Data

- There is no customer history or their social class or income
- There is no information of the geolocation of branches or relative location of one to another.
- There is no information regarding exact items on sale at each store. This makes difficult to provide reliable information on possible overlap of products offered in each store.

Bottom line, the dataset includes 5 arbitrary stores. It seems like all the stores are within a small geographic area, as some customers shop at all the stores. But as COOP Italia is Italy's largest supermarket chain, with over 100 stores across all of Italy, it is difficult to make a statement as to which stores this dataset may include.

Data Wrangling and Cleaning

As mentioned before, it is a reduced dataset so wrangling and cleaning it was relatively simple. Replacing the NaN values was more or less the extent of the situation. Luckily there were an extremely low percentage of NaN values (0.00037%); eliminating such a short number of data rows will not have any adverse affects on the analysis.

Other Datasets

It would be ideal to have detailed information regarding what products are available in each of the five stores. This would allow us to learn more about the customers, their habits, and what they might be looking for in a shop.

4. Approach

Stage 1 - Exploratory analysis
After wrangling and cleaning the data it is necessary to understand the data as much as possible. In this case it meant learning about the customers and more importantly the market. This stage is possibly the most essential part of the analysis.
The goal is to find out which customers do what, and suggest hypotheses and reasonable answers as to why.
The main libraries used in this portion:

- pandas - data handling
- numpy - data handling
- matlibplot - data visualization

Stage 2 - Hypothesis testing
Visualizations help give insight, but do little as to offer statistical proof. In this stage, A/B testing is conducted in order to compare distributions.
Main tests used in this stage:

- Scipy.stats
  - Students T-test, one-sided, for normal distributions
  - Mann–Whitney–Wilcoxon (MWW), as the distributions might not be normal, and its almost as efficient as a standard t-test

Stage 3 - Clustering
Using a machine learning algorithm, k-means clustering, to segment the market. The challenge here is deciding which of the features to segment the market according to, as well as deciding whether to scale the data or not.
Libraries used:

- Sklearn
  - Kmeans
  - cross_validation.train_test_split - for partitioning the data
  - preprocessing - to help scale the data

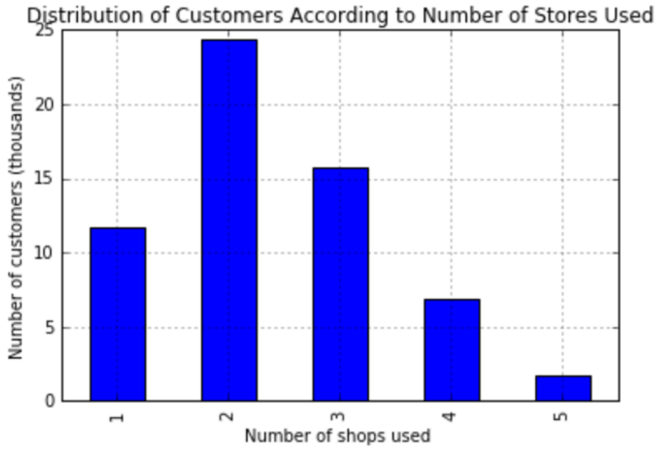
Stage 4 - Understand the market segments
Using the clusters created in the previous stage to understand the trends, and finding out the reason they were grouped together. The stage solidifies our exploratory analysis and is crucial to the next stage.
Libraries used:

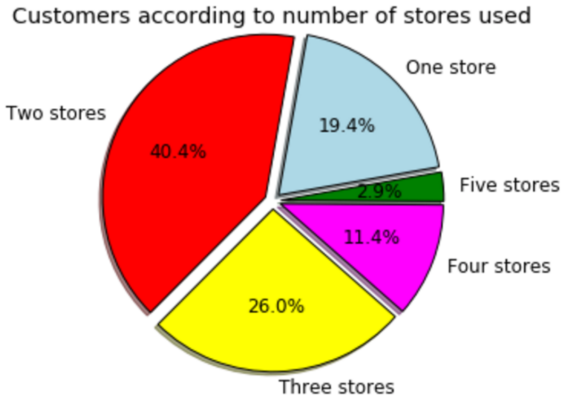
- pandas - data handling
- numpy - data handling
- matlibplot - data visualization

Stage 5 - Target Markets
The final stage is deciding which market segments are the most efficient to target. This could be based on a number of factors. Primarily these would include the segments that are easiest to target, and bring in the most revenue.

5. Results

Stage 1 - Exploratory Analysis

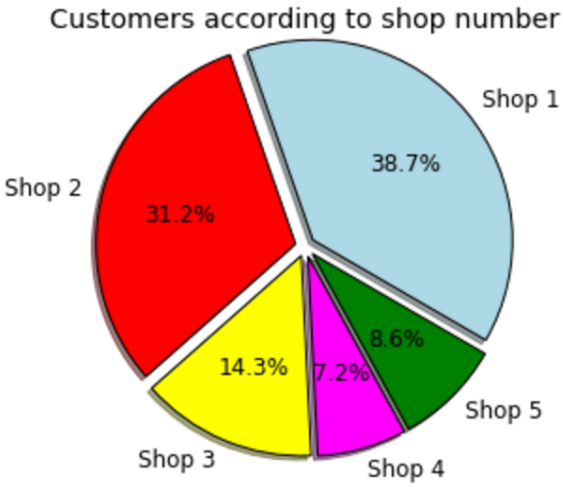


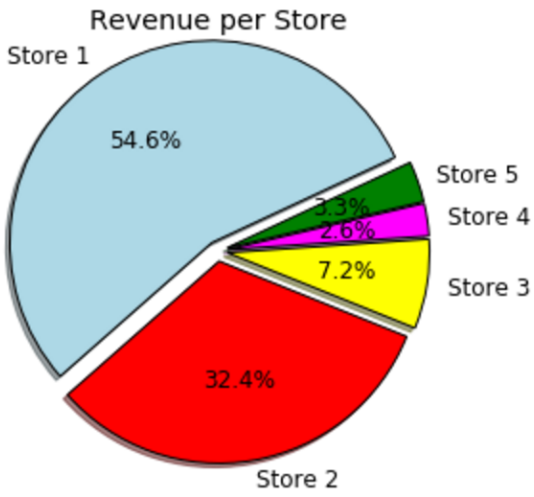


The first thing done, was to check how many of the shops customers used, as well as the distribution of the customers. It was found that most of the customers used 2 stores followed by 3, 1, 4 and finally 5 stores. This offered some insight, suggesting to possibly focus on customers who shop at 3 stores and fewer.

To confirm where to focus the graph above is the distribution of stores used according to revenue. Customers who use only 2 stores account for 40% of the total market revenue. If the client can target customers who use 3 or fewer stores more efficiently, that would bring in about 80% of total revenue. Logic would dictate it's easier to target customers who use fewer stores.

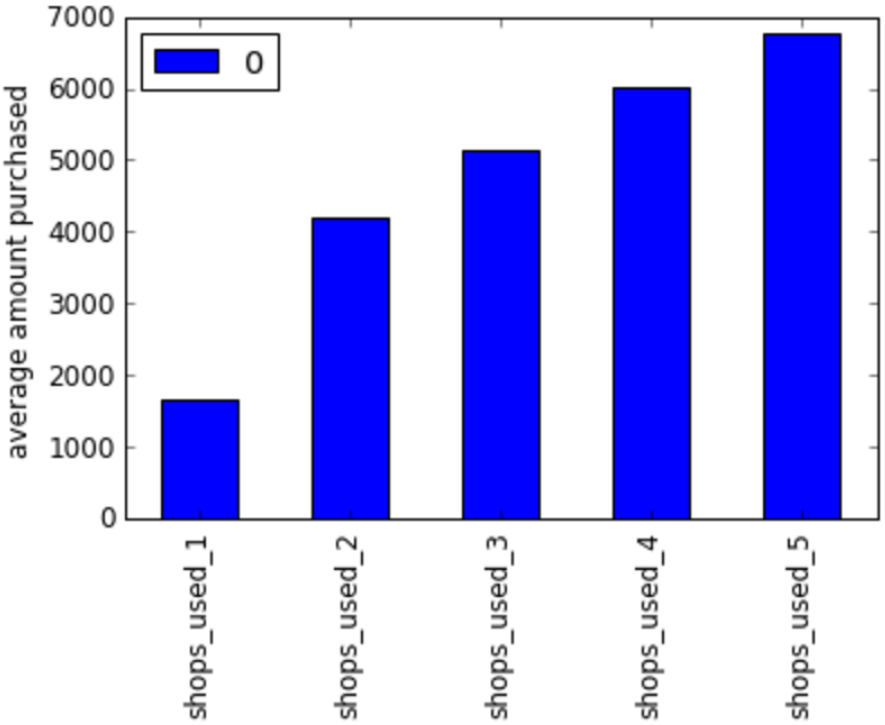
Next, we need to understand which of the 5 stores the customers are most likely to purchase from. The graph above shows that shop 1 brings in quite a lot more revenue than the other stores. Shop 2 is the next highest grossing followed shop 3,4,5. The likely assumption, is that shop 1 has something interesting to offer that the others stores do not. We try to find out what this could be.



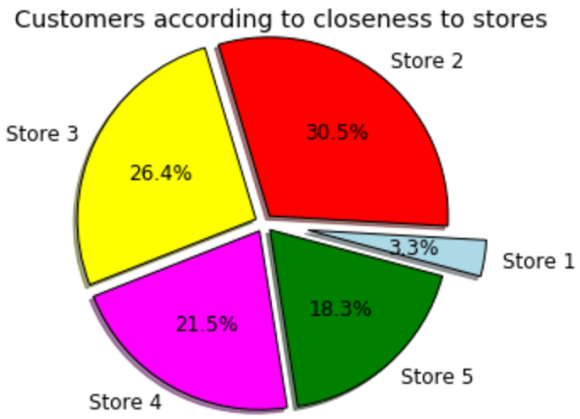


The charts above clearly show that most of the market pertains to Stores 1 and 2. These 2 stores combine to hold 87% of the market revenue. Store 1 alone holds the majority of the market. Let's check if customers visit certain shops more often than others.

Above is a graph showing the estimate of the amount of visits a customers makes at each store. For store 1, it shows that customers who shop at more stores, visit store 1 more frequently. In other words, customers who shop at fewer stores, also visit those stores less. This would suggest these customers enjoy bulk buying. Bulk purchasers have more or less two types. There are those that purchase a lot of unique items, and those that buy similar items but multiple of those, so as not to need to visit the stores more often than needed. Let's find out which group of customers purchase the most per transaction.



Interestingly, even though customers who use more stores tend to visit each store more frequently, it would seem that those customer actually purchase more per visit as well. Therefore the more visits a customer makes, the bigger each purchase. We'll leave this train of thought and come back to it. Next we will focus on distance to stores.



This graph offers quite a bit of insight. We know that store 1 brings in more than 50% of the total revenue, and yet, only 3.3% of the customers live closest to that store. Again, it brings up the question of why store 1 is so successful. Customers are clearly traveling to shop there, meaning they are choosing store 1 over a store that is closer to their home.

To get a more in depth understand, we focus on customers who shop exclusively at one store (~20% of customers). Again, for stores 2-5, the majority of exclusively one store customers, are customers that live closest. And yet, only 3.5% of store 1 customers are customers that live close to it. Two hypotheses come to mind. Either customers closest to store one go elsewhere,

or customer that live closer to another store shop at store one and reduce the percentage shown above. Looks more likely to be the latter.

This graph is similar to the previous one. What it shows is, of the customers that live closest to each store and shop exclusively at one store, what percentage of them, purchase at their closest store. As we can see, our previous hypothesis, is answered. 89% of exclusively one shop customers who live closest to store 1, choose store 1 as their sole store. 60% of similar customers, but closest to store 2, purchase at store 2. On the other hand, the majority of exclusively one shop customers, that live closest to store 3,4, and 5, decide to travel to another store. This again raises the question, why is store 1 and to a lesser extent store 2, so enticing? Before we explore store 1, let's look at which store the exclusively one store buyers decide to shop at, and which store they are closest to.

This distribution shows that most of the only-one-store buyers that shop at store 1 are actually closest to stores 4 and 5. This group accounts for approximately 40% of customers. Maybe the reason store 1 and 2 are so enticing is product variety.

This graph furthers the argument for variety being a driving factor. Store 1 has the most variety, and this seems to be the reason why so many customers are willing to travel to such a store.

We will turn our focus now to customers who shop at exclusively two stores. This is biggest market segment in terms of both size and revenue. We will primarily focus on customers that shop at store 1 and another store.

As we can see of the customers who shop exclusively at two stores, shop 1 being one of them, 80% of them choose 2 as the other. Interestingly 84% of revenue from customers who shop at 2 stores, are customers who shop exclusively at store 1 and 2, not anywhere else. That accounts for 33.5% of total revenue brought in by the market. Since stores 1 and 2, offer the most variety of products, it would seem like variety is the biggest factor when customers decide where to purchase from. We will keep this in mind for clustering purposes. Before clustering we will quickly check why customers choose to shop at multiple stores, more specifically, if item prices are a factor.

This graph would suggest that one-store-only customers pay the most per item, and customers that purchases from more store, purchase less. Especially if we focus of store 1. Let's conduct some A/B testing to confirm or reject this.

Stage 2 - Hypothesis testing

Since our graph on variety in each store was an estimate it would be necessary to test out the difference. Our AB tests executed indicate that only in the pair of shops 3 and 4 the difference in proportion between these two groups may have been due to chance. This may indicate that the driver in customers' decisions in what shop to buy, more than anything else, is variety of products. It would be nice to have more information regarding the specific type of products offered in each store.

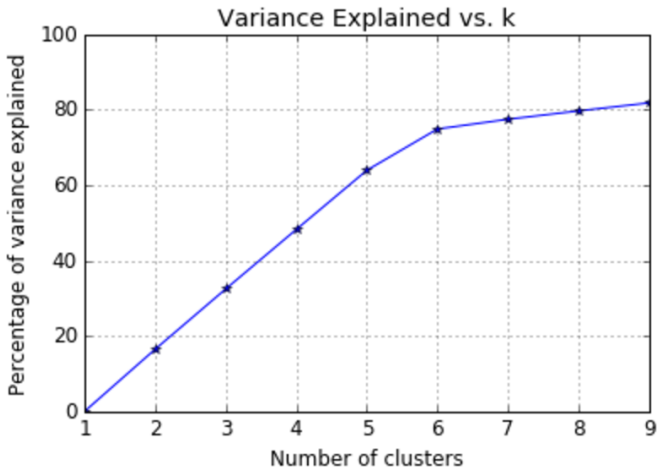
Next we test whether price of items purchased are different across the customers.
Let's remind ourselves that we're trying to find out if, customers who purchase in more stores buy items that are less expensive than customers who purchase in less stores. Maybe shops used has a relation to customers looking for lower prices.

According to these tests and inference from the above graph:
Store 1: Exclusively one shop customers purchase the most expensive items at store 1. 2 shop customers a little less Followed by customers who use 3,4, and 5 shops, all of which, we can group together.
Store 2: Customer who purchase from 1 and 2 shops in total, purchase similar priced items. NOTE: This indirectly shows that of customers who use 2 shops, those stores being exclusively store 1 and 2, they buy cheaper items than exclusively one store customers at store 1. Customers who use 3,4, and 5 shops who purchase at store two purchase similar priced items, significantly cheaper than those who purchase from 1 or 2 stores.
Store 3: Again, Customers who use 3,4, and 5 shops who purchase at store 3, purchase similar priced items, significantly cheaper than those who purchase from 1 or 2 stores. Exclusive store 3 customers buy the most expensive items.

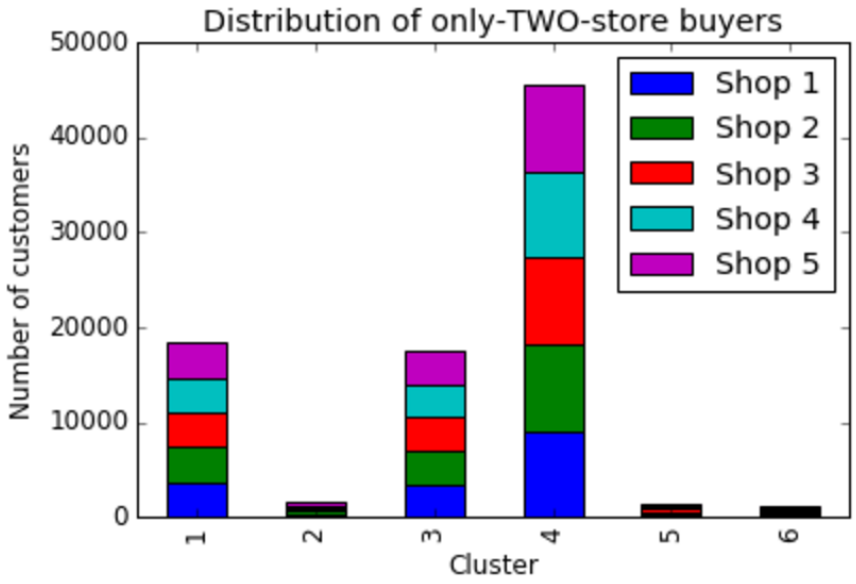
Stage 3 a) - Choosing how to Cluster

From the exploratory analysis stage, we have made a good argument for variety being a big factor. For sake of thoroughness, we will cluster according to variety, as well as distance. Also of importance is the choice to scale or to leave the data as is. Let's find out.

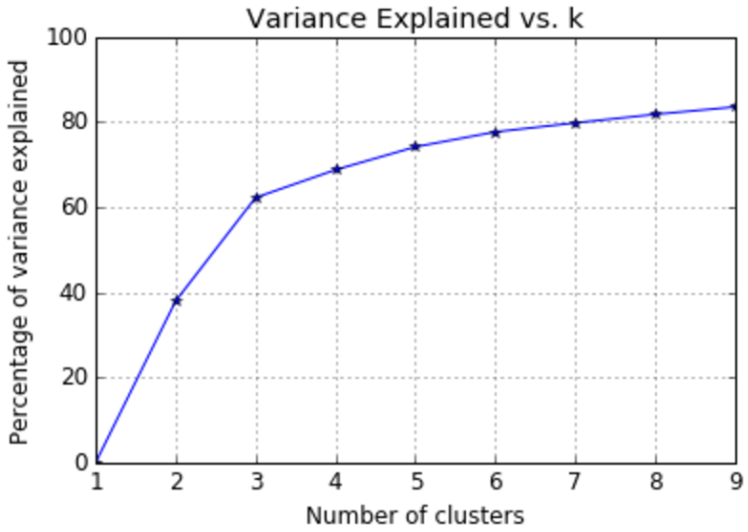
First we cluster according to variety and we scale the data. We have to decide how many clusters to choose so a variance explained graph vs # of clusters would be useful, as shown below.



We use the elbow method, hoping to retain about 70-80% of the variance. In this case we choose 6 clusters.

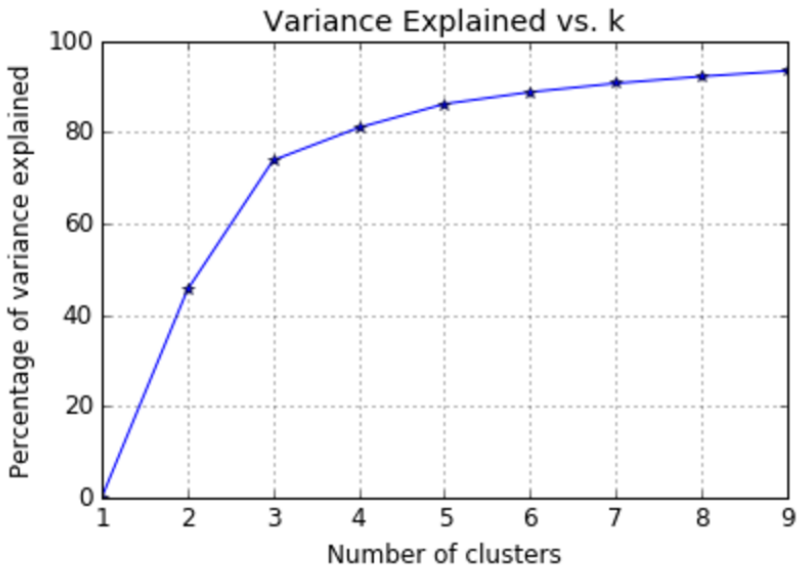


Something does not look right in the above graph. Within each cluster there is an equal amount of customers from each shop. We would expect there to be more customers from shops 1 and 2 as they are the most popular shops. Let's try without scaling.

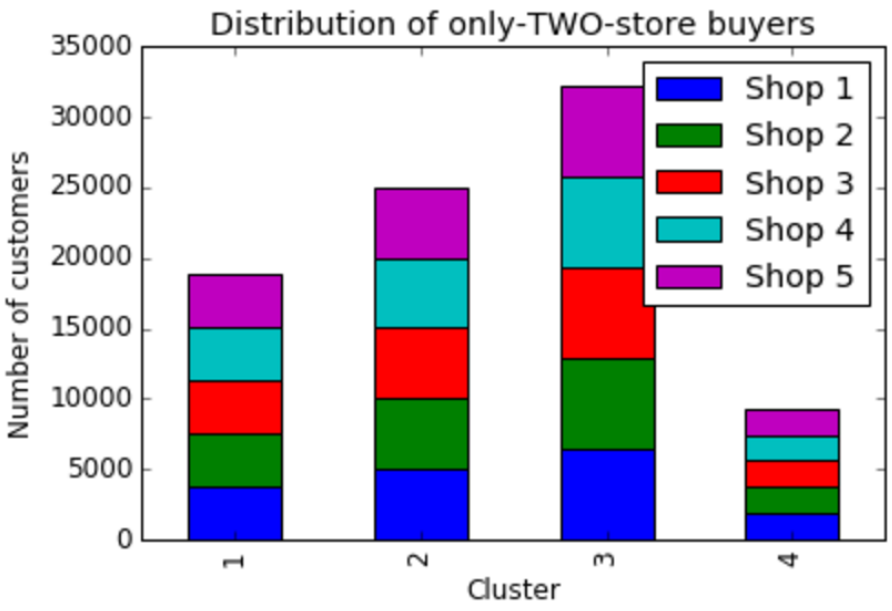


Without scaling the data, we note that variance retained for 6 clusters has increased a couple percent. Lets see what the distribution of the clusters look like.

This graph makes much more sense. For only-two-store buyers, our previous analysis shows that the customers consist primarily of store 1 and 2 purchasers, and this confirms that. Plus, since we are clustering based solely on variety, where the metric between points are equal, and the variance difference is crucial, both are important to preserve. Therefore we will continue without scaling.
Next we will cluster according to distance, where the same argument holds for not scaling, and we will compare the to clustering methods and decide which is better suited to our purposes.



This will be the last variance explained graph shown. It should be understood this method for choosing the number of clusters was maintained throughout. Here we choose 4 clusters which retain about 80% of variance.



This looks very similar to the issue we had with scaling. Within each cluster there are the same amount customers from each store.
Therefore we take the decision to segment the market based on unique products purchased, and according to the number of shops the customers use.

Stage 3 b) - Results of Clustering

The first market chosen to segment are exclusively one store buyers.

One shop buyers were segmented into 4 groups. Cluster 1 has the most customers, and includes all the stores. Where as cluster 2 and 3 have customers who purchase only in store 1. And Cluster 4 has exclusively store 2 customers

The next market will be two store buyers.

The distribution for only-two store buyers shows clusters 2,5,6, and maybe 4, with customer who shop exclusively at stores 1 and 2. Cluster 1 and 3 have a wider variety in their grouping of customers.

Now onto three store customers.

Clusters 1,6 and maybe 3 are solely customers who shop at store 1,2 and 3. Where as cluster 2,4 and 5 include customers show shop at all stores.

Finally we target customers who use 4 stores exclusively.

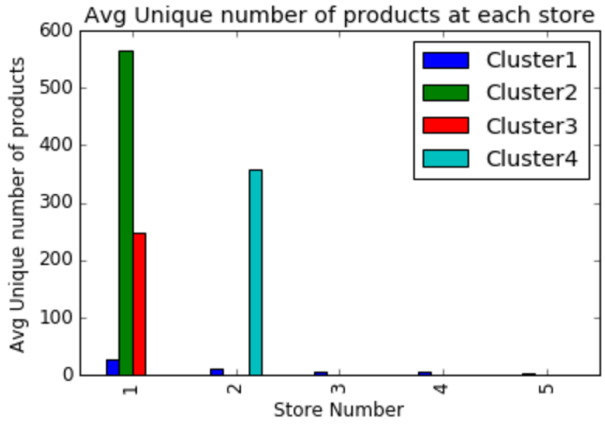
Clusters 3 and 6 are primarily store 1,2,3, and 5. Whilst the other clusters have customers who shop at all 5 stores. In the next section we will analyze the market segments in more detail and get a better understanding of why they are grouped together.

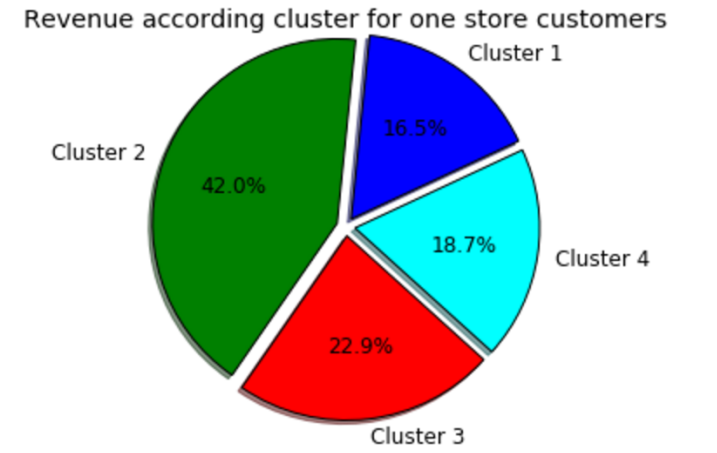
Stage 4 - Understand the market segments

1. One Shop Buyers

A reminder that this group accounts for 7.6% of the total market revenue. Let's take a closer look at the segments within this group.

From the graphs above; on the left is a reminder of how the clustering algorithm segmented the population. The graph on the right shows the relation each cluster has to the amount of unique products they each purchase on average. With the help of the machine learning algorithm, it shows 4 distinct clusters.





Above on the left shows exactly which store each of the clusters are shopping at. On the right, it is showing the revenue of each of the segments within the group only one shop buyers.

Cluster 1
This cluster which accounts for the most amount of customers in this group, but only 16.5% of this segments revenue, is interesting because it's a group, where the customer could shop at any store. The main distinguishing fact is that they are customers who are not looking for a lot of variety.

Cluster 2
This segment is exclusively store 1 shoppers. It accounts for 42% of revenue and they value variety highly. On average they purchase approximately 550 unique products at store 1.

Cluster 3.
Again this segment is exclusively store 1 shoppers. It accounts for 23% of revenue. Variety isn't valued as much as cluster 2, purchasing on average 230 unique products.

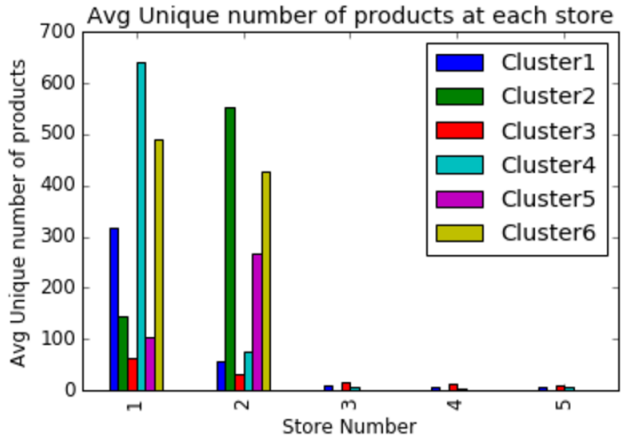
Cluster 4
This segment is exclusively store 2 shoppers. It accounts for 18.7% of the revenue. Variety is important and they purchase on average 360 unique products. The distinguishing feature for this group is that they are exclusive store 2 buyers.

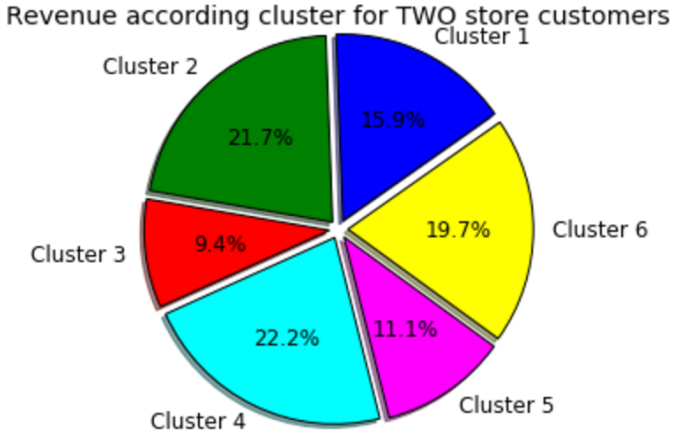
Conclusion
The preliminary conclusion is that cluster 2,3, and 4 are easy to target. Since they are exclusive buyers to one store, and variety is a driving factor. Cluster 1 might be a bit tougher, but since it holds the smallest amount of the market, it might not be worth targeting them

2. Two store buyers

This segment of the market is the biggest, it accounts for 40% of the total market. Therefore it is imperative that there is an in depth understanding.

As done with the previous segment of exclusively one store buyers, here are similar graphs, but for two store buyers. For this segment, it was necessary to use 6 clusters to retain a appropriate amount of variance.





Above, as mentioned before, are familiar graphs of two shop buyers according to the cluster they belong to. Along with the distribution of the market as well.

Cluster 1
Even though it looks like cluster 1 has a customers that shop at all the stores, it would seem that the majority of their purchases occur at store 1. So thats where the focus should be put.

Cluster 2
This segment accounts for 21.7% of the revenue and yet is a relatively small amount of customers. They are primarily purchasers at store 2, where they enjoy lots of variety. And a secondary purchasers at store 1.

Cluster 3
This cluster is the most difficult to understand. It accounts for the most customers by number and yet, the smallest revenue at 9.4%. They enjoy the least amount of variety and shop at any store. There seems to be similarities with this segment and cluster 1 from the one-shop-only buyers. Maybe this will help target market them in the future.

Cluster 4
They account for 22.2% of the market (the most in this segment). They primarily and heavily favour store 1, enjoying the variety it has to offer, as well as making smaller purchases at store 2.

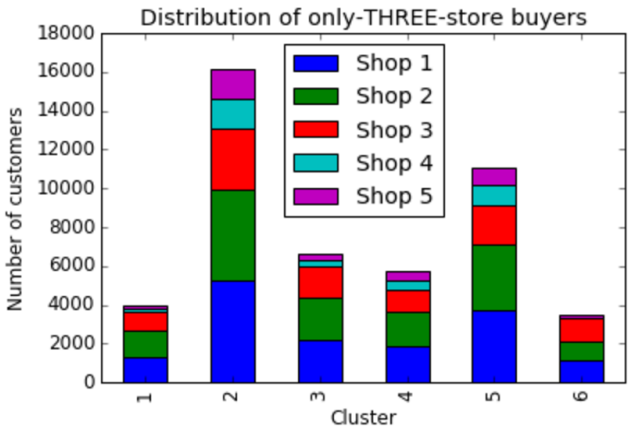
Cluster 5
A small portion of this segment, but quite easy to understand. They shop exclusively at store 1 and 2, and seem to buy a fair but not a lot of unique products. The prefer store 2 slightly over store 1.

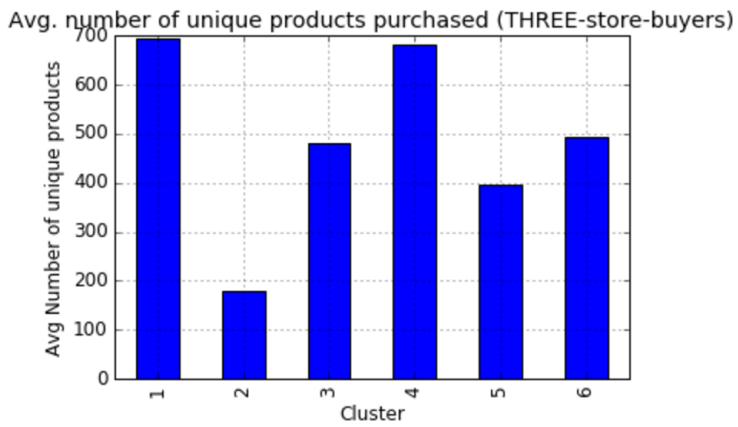
Cluster 6
This group accounts for 20% of the revenue. Are exclusive buyers at shop 1 and 2, which they buy about equal unique products at. The amount of variety they enjoy is relatively a lot, at 650 unique products on average.

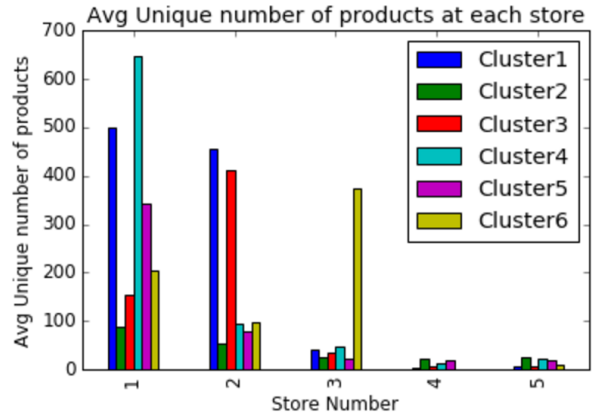
Conclusion
The easiest segments to target are cluster 2, who are primarily store 2 shoppers, cluster 4, heavily favour store 1, with lots of variety, and cluster 6, as they seem to enjoy variety above all else and do not favour any store. Cluster 3 might be the hardest to target, and again it accounts for the least revenue.

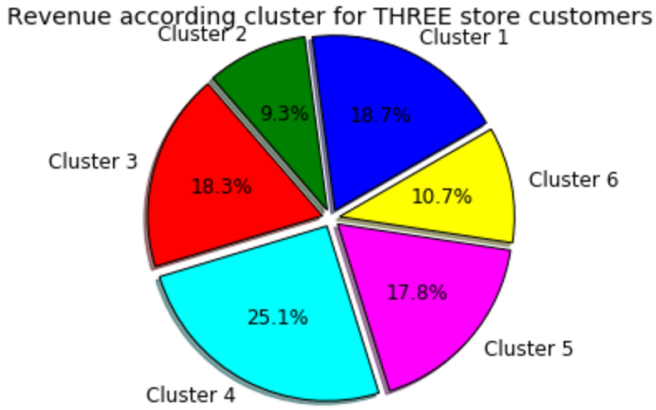
3. Three Shop Buyers
The segment of the market accounts for 31.6% of total revenue.

Below we have the familiar graphs to aide in the analysis of the clustering algorithm.









Cluster 1
The original clustering graph makes it seems like there is even amounts of customers who purchase from shops 1,2 and 3. Upon further inspection, over 900 of the 1000 unique items purchased from this segment came from stores 1 and 2. This would be the most efficient was to categorize them. This cluster also accounts for 18.7% of revenue.

Cluster 2
This cluster has the same trends as cluster 3 in the two-shop buyers, and cluster 1 in the one-shop-buyers. The smallest market share, the least number of unique products purchased, and the most amount of customers.

Cluster 3
Primarily purchasers at store 2, along with store one. Other than that, they only buys small amounts at the other shops. Account for 18.3% of the market.

Cluster 4
Largest group according to market share, 25.1%. Heavily favour variety, close to 700 unique products purchased, and are primarily store 1 buyers, with small purchases at store 2 and 3.

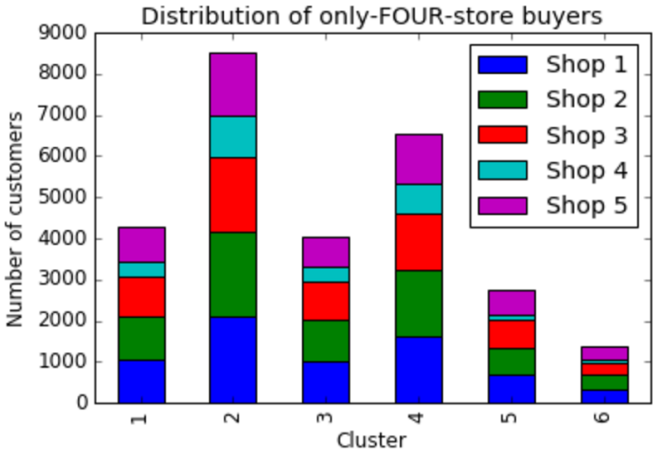
Cluster 5
Account for 17.8% of the market, and enjoy a fair amount of variety (400 unique products) and are primary purchasers at store 1.

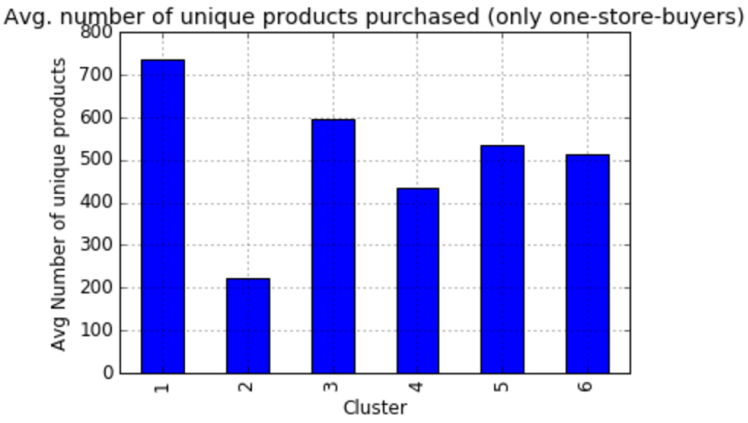
Cluster 6
This group are predominantly and the only major purchasers at store 3. They buy the most of their products there, and account for 10.7% of the market.

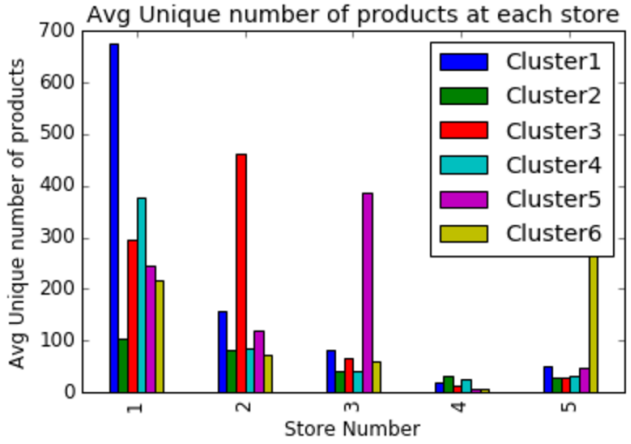
Conclusion
Cluster 6 is an easy group to target market, as they prefer store 3 so heavily. Cluster 4 is as well, again predominantly purchasing at store 1, with preference of a lot of variety. Cluster 1 seems to favour variety over everything else. The do not seem to have a preference as to which store. Cluster 5 favours store 1, but with less variety, along the same lines as cluster 2, except they prefer store 2.

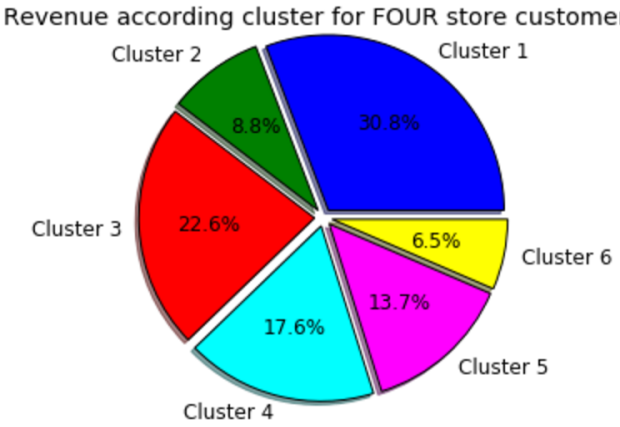
4. Four store buyers
This segment account for 16% of the total revenue.

Below are the familiar graphs in accordance with this market segment.









Cluster 1
This is the largest group according to revenue (30.8%), they primarily shop at store 1, and unique products purchased is a major driving factor. On average above 700 unique products are purchased. This makes this segment of the whole market, the most variety driven.

Cluster 2
This is the cluster that has similar to cluster 2 in the three-shop-buyers, cluster 3 in the two-shop buyers, and cluster 1 in the one-shop-buyers. A small market share, the least number of unique products purchased, and the most amount of customers. T

Cluster 3
This is an interesting cluster. They account for 22.6%, the value lots of variety, and yet they spread their purchases amongst all 5 stores. Still the heavily favour store 2.

Cluster 4
The group is heavily favouring store 1. There's a slight amount of unique purchases at store 2, but store 1 is the driving factor here. The segment accounts fro 17.6% of the market.

Cluster 5
The group is interesting, they are the major purchasers at store 3. Enjoying a fair amount of variety in their purchases

Cluster 6
These customers shop almost exclusive shop 5 buyers. From our early analysis, shop 5 does not offer a lot of variety, but this segment purchases a lot from store 5, then travels to store 1 and 2 for the rest of their products.

Conclusion
Clusters 3, 5, and 6 are good target markets, as they all enjoy the same variety but have different store preferences and heavily favour one of them. The prefer stores 2,3, and 5 respectively. Cluster 1 enjoys the most variety, as well as primarily shopping at store 1. Cluster 4, similar to cluster 1, except they prefer less variety.

Stage 5 - Target Markets and Recommendations

1. One-shop- buyer

19.4% of the total customers are in this group, and account 7.6% of total revenue. Clusters 2,3 and 4 are considered the target markets. These subgroups are defined as exclusively one shop customers who purchase a medium to a large amount of unique products. These subgroups are either store 1 or store 2 shoppers. Targeting these subgroups would secure 6.4% of that 7.6% (86.5%) of the total market and only require targeting 34% of the subgroup population

2. Two- shop-buyers

40.4% of the total customers are in this group, and account for 39.9% of total revenue. Cluster 4 and 2 are the priority, they enjoy lots of variety and shop primarily at shop 1 and shop 2 respectively. Both shop at the other store but not to the same extent. They are exclusive customers to shop 1 and 2. Cluster 6 is another target market. They enjoy a lot of variety in their purchases but are spread evenly between stores 1 and 2, therefore this segment would be easy to sway, since they are solely driven by variety. Cluster 1 would be the other target market, since they prefer variety but not as much as the others, cluster 1 is on the boundary line of the benefit-cost ratio. If we target market cluster 4,2, and 6 that would account for 63.6% of 39.9% the total revenue (20.5%). Where as if cluster 1 is included that total revenue goes up to 25.7%.

3. Three-shop-buyers

26.6% of the total customers are in this group, and account for 31.6% of total revenue. Cluster 1 and 4 are top priority. These two include the customers who value the highest variety possible. Cluster 4 enjoys store 1 primarily, where as cluster 1, does not seem to favour one or the other, as long as the variety is there. Cluster 6 is another target market, as they value a fair amount of variety but are major customers are store 3, making them relatively unique. Clusters 3 and 5 are somewhat similar in terms of the value they put on variety on items. Cluster 5 favours store 1, where as cluster 3 favours store 2. If we target market clusters 1,4, and 6, we will be targeting 17.2% of the total market. If we include cluster 3 and 5, that would increase to 28.6%. The latter would seem like the likely choice.

4. Four-shop-Buyers

11.4% of the total customers are in this group, and account for 16.2% of total revenue. Cluster 1 is the target market here, the enjoy the most variety and primarily shop at store 1. Cluster 3 would be the next target, they also enjoy variety but spread their purchases between mainly store 1 and 2, focusing more on store 2. Cluster 6 is interesting, because it happens to be the only cluster who enjoy variety but who also prefer to shop at store 5, making them an essential target group. Cluster 4 and 5 are quite similar, enjoying a fair amount of variety and primarily spreading their purchases amongst two stores. Cluster 4 being stores 2 and 1, and cluster 5 being store 3 and 1. If we target cluster 1,3, and 5, that would account for 9.7% of the total market. If we include clusters 4 and 5 that would increase to 14.7% of the total market.

Final Conclusion

After an in-depth analysis into COOP Italia's customers for a specific unknown geolocation it can be concluded that product variety is the driving factor for market trends and customer habits. Distance is less of a factor than what might have been initially thought. Customers are willing to travel and choose a farther store over a store closer to them, if that store suits them better in terms of variety. If we segment customers based solely on the unique products purchased, it gives an ample amount of insight, that enables, COOP Italia to better understand, and better provide for their customers.