**Milestone Report - Market Segmentation**

1. **Problem to be Solved**

PWC released a Consumer Goods Trends article in 2015. It was titled Fragmentation and Simplification. According to them, “in both developed and emerging markets, there is a wider variety among consumers now than at any time in the recent past.” PWC goes on to explain 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 of 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; therefore, the industry of marketing consumer goods needs to change as well. In the past, the middle of the market has played a major role in driving the consumer goods industry; however, as this segment of the market is shrinking, a readjustment in focus should occur. A proactive approach, would entail having a comprehensive understanding of the consumer to give clients an advantage. The question therefore arises: How can the consumer base be targeted in the most efficient manner possible?

**2. The Client**

The clients in this case are COOP Italia, and more specifically, their marketing division. They have a customer base, and want to be able to target market their customers. This will help maintain customer satisfaction, giving them what they desire, as well as COOP Italia, which would be able to tailor their stores to the customers that frequent them. Secondly, understanding their customers will allow them to efficiently target new customers. This can be done based on what segment the new customers falls into. Since, all marketing divisions have a limited amount of resources, efficiently using them is essential.

**3. Data**

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

**Import fields**

* distance to each shop
* product purchased
* unique products purchased
* amount purchased
* avg purchase
* 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 mean of all the customers unique\_products\_purchased at each shop.

**Drawbacks of the Data**

* There is no customer history or their social class ie income
* We do not know how far each store is relative to another
* The exact items on sale at each store. Hard to tell if there's overlap with each store inventory.

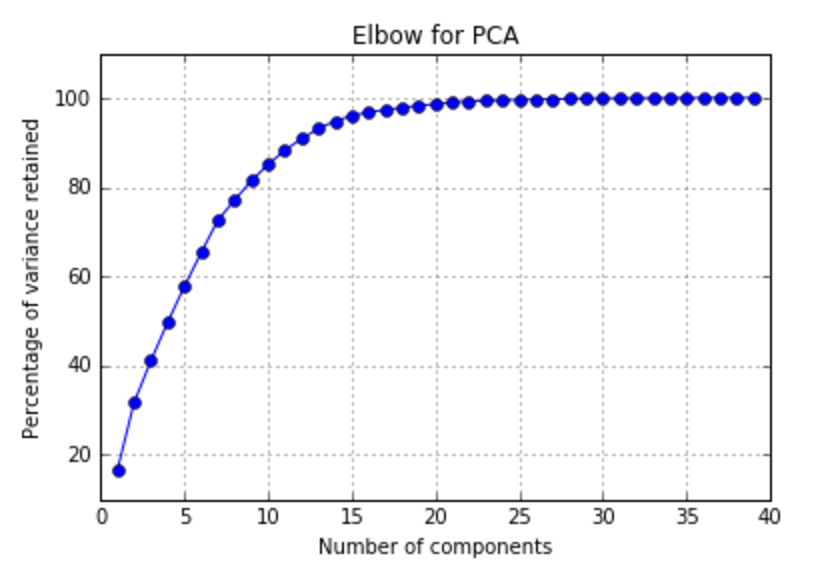
Bottom line, the dataset includes 5 arbitrary stores. It seems like all the stores are within a small geographic area, as some customers shop at all the stores. But as COOP Italia is Italy's largest supermarket chain, with over 100 stores across all of Italy, it is difficult to make a statement as to 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, so it should not have any adverse affects on the analysis.

**Other Datasets**

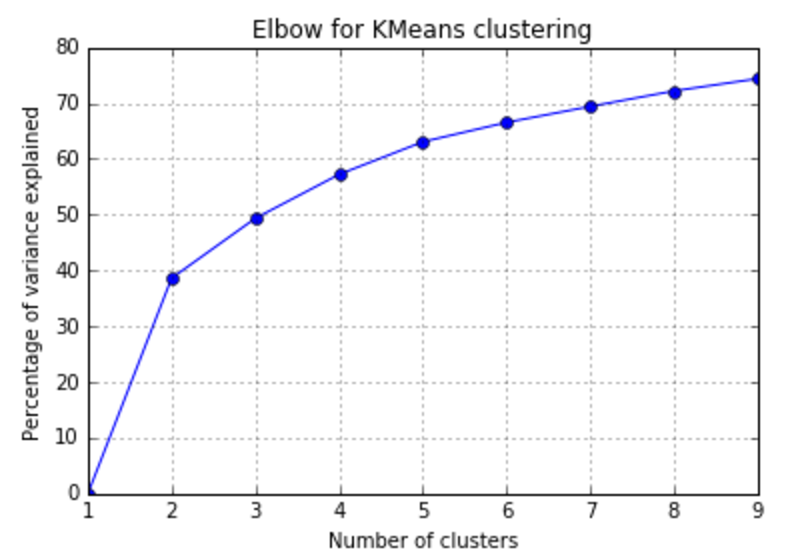
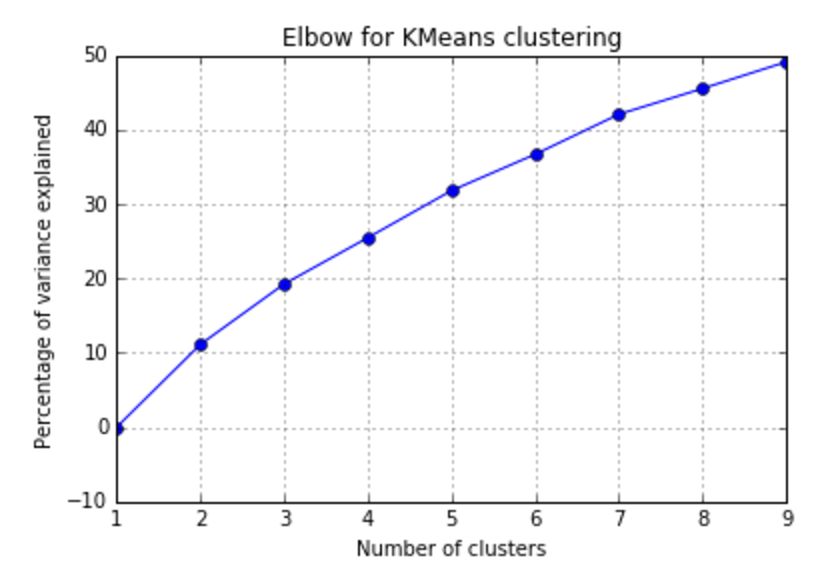
The full/original dataset of this reduced dataset, would be quite helpful. It should allow for more insight into the market. This would include learning more about the customers, their habits, and what they might be looking for in a shop. Unfortunately, that is not possible and the dataset on hand is more than adequate for my purposes.

*Any preliminary exploration you’ve performed and your initial findings. Test the hypotheses one at a time. Often, the data story emerges as a result of a sequence of testing hypothesis e.g. You first tested if X was true, and because it wasn't, you tried Y, which turned out to be true.*

**PCA**

At first I wanted to carry out principal component analysis (pca) to better understand the data. Since the dataset had 40 features, I thought reducing the dimensions would help make sense of it. Initially I wanted to reduce the dimensions from 40 to 2/3, ideally because it would ease visualizing the data. Unfortunately 3 principal components only retained about 40% variance, where as to get to the intrinsic standard threshold of approximately 80%, I would need 9 principal components. This did not seem like it was worth it.

**Kmeans**

Next I attempted market segmentation. The most widespread method used, seems to be Kmeans clustering. So I attempted segmenting the market with this algorithm. After running it over the whole dataset, both scaled and non-scaled version, I noticed again that clustering wasn't giving me the explained variance I wanted (ideally 80%). With 5 clusters the explained variance for the scaled and unscaled data was approx 32% and 62%, respectively.

*Figure: On the left, Variance explained vs # of clusters of a scaled dataset*

*On the right, Variance explained vs # of clusters of a non-scaled dataset*

**Scaled vs Non-scaled**

From this, I had to decide whether scaling the data was beneficial or detrimental to the outcome. I am still testing both. I understand that non-scaled data causes the algorithms weight to be shifted. For example in kmeans clustering, the feature "amount\_purchased" has a large number value. So the algorithm should in theory favour segmenting the market according to a feature such as this. But this is possible how I want to segment the market. More will be done to figure this out.

**Partitioning**

The point of Kmeans was to segment the market, such that if need be, target marketing to each of these segment would be efficient. But which segment of the market brings in the most revenue? A simple graph shows that the top 30% of customers account for 75% of the market in terms of revenue. So targeting this segment of the market is of utmost importance. Focusing on the other 70% of customers, which only bring in 25% of revenue, does not seem like an efficient way to allocate resources. On a side note, I would assume that some of the target marketing done to the top side of the market would overlap with the interests of the bottom of the market.

*Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?*

**Approach Taken**

**1. Partition the market**

* Top 30% including ~75% of total revenue

**2. Market Segmentation**

* Run Kmeans to cluster/segment the population
* Within each segment, understand why each cluster is grouped together
* which shops do they shop at?
* shopping frequency?
* do they buy the same items multiple times?
* market share, ie which segment brings in the most revenue
* is distance a major factor for this cluster?
* do they enjoy stores with variety, ie bigger stores?
* do they enjoy smaller stores? maybe a more personal shopping experience?

**3. Scaled vs Non-scaled**

Do this for both scaled and non-scaled data to see which makes more sense.

- Are the clusters more easily discernible?

- if yes, then they are easier to target when we consider marketing.

**4. PCA**

PCA is a useful tool for visualizing the dataset, this may become useful later in the project.

**Conclusion**

There are two major goals to be achieved. One is to understand why the variance explained is so low, and if its possible to keep the number of clusters small but increase the variance explained. The next goal is to do with market segmentation and understanding each of the segments. I need to find a robust test that can be used to compare the clusters to each other, and to the original dataset. Once these to goals are accomplished, the results will shortly follow.