Unsupervised Learning Capstone Project Final Submission

By Marcus Hendricks

Executive Summary

For this capstone project, we are tasked to identify the best possible customer segments using the given dataset. Customer segmentation is the process of dividing a dataset of customers into groups of similar customers based on certain common characteristics, usually for the purpose of understanding the population dataset in a better fashion. Customer segmentation is important because it has a vital role to play in optimizing Return on Investment and make organizations more efficient in terms of utilizing their money, time, and other critical resources in custom marketing strategies for different groups of customers based on their unique needs and motivations.

For example, it has been understood from various research that customer segmentation often has a huge impact on people's email engagement. Segmented campaigns often see over 100% more clicks than non-segmented campaigns, and email marketers who have segmented their audience before campaigning have reported a 6-7 times growth in their overall revenue. It has also been observed in various contexts that in today's world, individual customers prefer personalized communications and offerings that cater to their particular interests.

Good customer segmentation allows marketers/advertisers to engage with each customer in the most effective way. Typically, the variables of interest are customer profiles, campaign conversion rates, and information associated with various marketing channels. With more directed advertising campaigns they will have more success with customers and better conversion rates. Based on these features, I will try to create the best possible customer segments.

Problem Summary

The intended goal of this project is to use Unsupervised Learning ideas such as Dimensionality Reduction and Clustering, and to come up with the best possible customer segments using the given customer dataset. Going into this project there were a few questions that I needed to consider:

- What features/variables have the most correlation?
- What variables have the most variance in the customers?
- What are the most important variables to consider when segmenting customers?
- What is an appropriate number of clusters for segmenting customers?
- What are the cluster profiles, and how are they different?
- How can the cluster profiles be used to target customers for growth and increased revenue?
- How can campaigns be customized based on cluster profiles for a higher acceptance rate?

During this project I used different techniques to help interpret the data and find answers to these questions. In this course and in this project our overall goal is to learn and use data science. And in this project, I will be trying to find the best segmentation of the clusters using Unsupervised learning, dimensionality reduction and clustering.

Solution Summary

A number of different Unsupervised Learning methods and algorithms were explored as part of the solution design, including K-Means, K-Mediods, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Models (GMM). The final proposed solution is the clustering created by using **K-Mediods with K=3**. This algorithm was chosen because it is not affected by outliers in the data, it had one of the highest Silhouette scores, and it seemed to create the best cluster profiles. In this section, I will explain how I came to these results.

The first thing I did was data exploration and clean up. The data I was given contains information on customers from the 2 years prior to collection year of 2016 for a company. The data includes information on income, spending habits, education, marital statuses, complaints, and conversion rates, amongst others. The original dataset consists of 27 columns and 2240 data entries. This data consists of numerical and a few object columns. The only column that was missing information was the Income column (later we added in the median income to those nulls). One of the first major insights was that 27 seems a lot of variables and that would need to be reduced to be able to make sense of all the data. After initially exploring the data, I began doing univariate analysis of the data, identified outliers in the data, and removed those outliers. I also created a few more variables that were easier to manipulate or combined variables to show more insight.

I then began bivariate analysis of the data and see if there were any significant correlations in the data. MntWines and MntMeatProduct had high correlation with Expenses near or above 0.85. These variables could skew the data a little bit. When the variables used in clustering are highly correlated, it causes multicollinearity, which affects the clustering method and results in poor cluster profiling (or biased toward a few variables). To help reduce the multicollinearity between the variables, I will need to use PCA to reduce variables. I then scaled the data to avoid the problem of one feature dominating over others because the unsupervised learning algorithm uses distance to find the similarity between data points.

Once all of that was done, I was then able to begin my analysis of the data using the different clustering methods and algorithms of PCA, t-SNE, K-Means, K-Mediods, Hierarchical, Density Based (DBSCAN), and Gaussian Mixture Model (GMM). First thing I did was try to reduce the dimensionality of the data. I first looked at the data using t-SNE, but it did not give much insight. I then began applying PCA to reduce the dimensionality

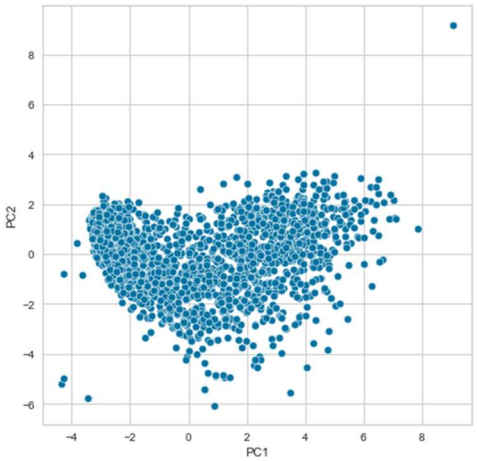


Figure 1 - PCA Plot

came to result a blob of data that did not have too much insight, but it gives me the foundation to build the Clustering techniques off (See Figure 1).

The first technique I tested was K-Means. I used a K-Means Elbow plot to determine the best number of clusters to make out of the data. The Elbow plot showed that K=3 and K=5 seemed to be the best results. Between the two, K=3 had the best Silhouette score of 0.2691, so I then tested K=3 and was given this scatterplot (See Figure 2).

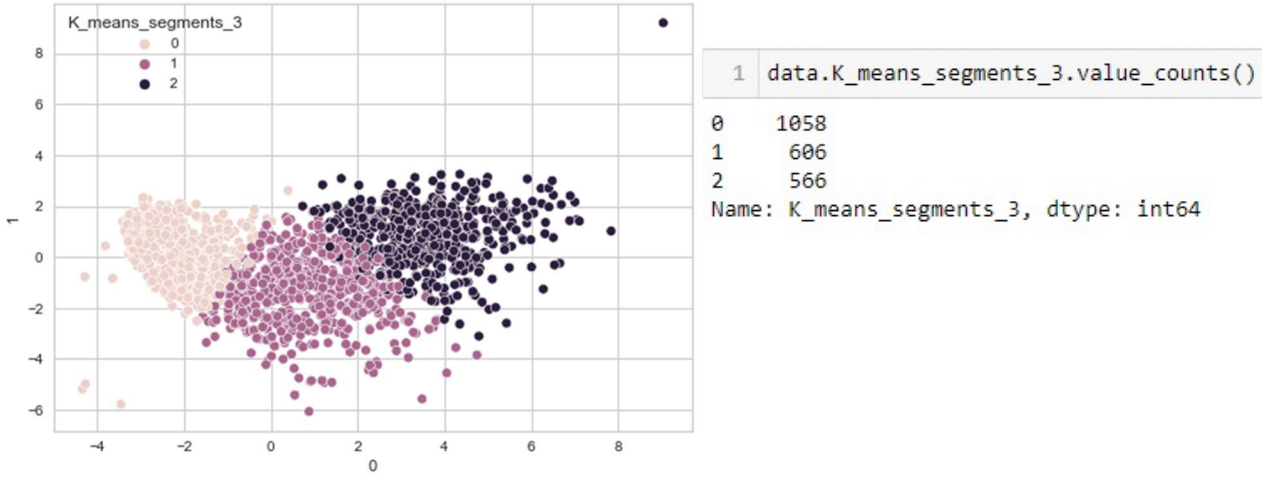
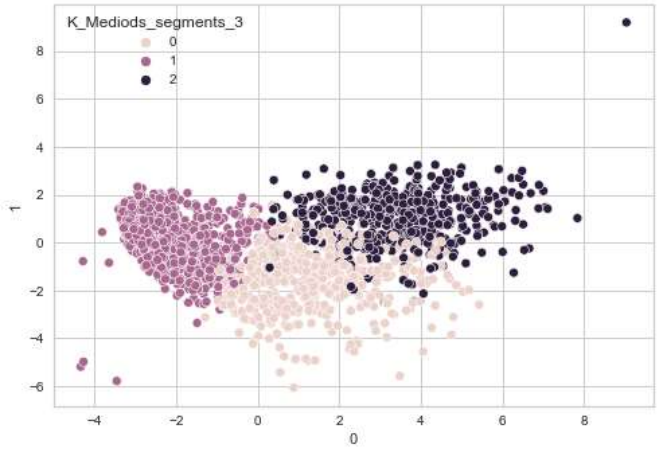


Figure 2 - K-Means K=3 PCA Plot

This provides decent clustering and limited overlapping but let's see how the profiles of the clusters came out. Cluster 0 consists of the low-income group. They are not able make as many purchases, typically have a larger family size of 2-3, and most have not accepted any campaigns. They do seem to visit the web often, maybe to find deals, however, they are the lowest in web purchases. So, they may be very selective. Cluster 1 consists of the middle-income group. They are able make more purchases, typically have a family size of 2-3 as well, about 40% of customers have accepted campaigns. They like to make deal purchases and web purchases. Cluster 2 consists of the higher income group. They can make a lot more purchases, typically have the smallest family size from 1-2, and most have at least 1 campaign. They like to make catalog purchases. This were good initial insights but not the best. Although this scatterplot gave me decent initial insights, it should not be used for this data set due to the outliers in the data and how K-Means is affected by outliers.

The next clustering method I tested was K-Mediods. I test both K=3 and K=5, and K=3 again had the highest Silhouette score of 0.2899. So then went into further exploration of that method and made PCA plot (See Figure 3). This clustering seemed to overlap more than K-Means did. Cluster 0 consists of mid to high income customers, who are the oldest age group, and typically have family size between 2-3, they have high web purchases and store purchases, spend a lot on wine similarly to cluster 2, they take advantage of the most deal purchases, have the most total purchases, and most have accepted 1 campaign. Cluster 1 consists of the lowest income group, they are the youngest age group and typically have the largest family size near 3, they have the least number of purchases across the board and typically accept 0 campaigns.

Cluster 2 consists of the high-income group, who are middle age group in late 40s, who have a smaller family typically of 2 with 0 kids or teens at home, spend the most on wine and other products, they make lots of purchases via the catalog and store, they have the most expenses, highest amount per purchase, and accept the most campaigns on average around 1. These profiles are very distinct from one another, seems logical, cluster counts are evenly distributed, and it gives a lot more insight than the K-Means did.




```
1  data.K_Mediods_segments_3.value_counts()

1    1152
0     540
2     538
Name: K_Mediods_segments_3, dtype: int64
```

Figure 3 - K-Mediods K=3 PCA Plot

Next, we tested the Hierarchical clustering techniques. We went through the Cityblock, Chebyshev, Mahalanobis, and Euclidean distance metrics with single, complete, average, and Ward linking methods. After going through all the hierarchical techniques there was only one dendogram that would be able to clearly cut and create distinguishable clusters. That dendogram was with Euclidean distance and Ward Linkage (See Figure 4).

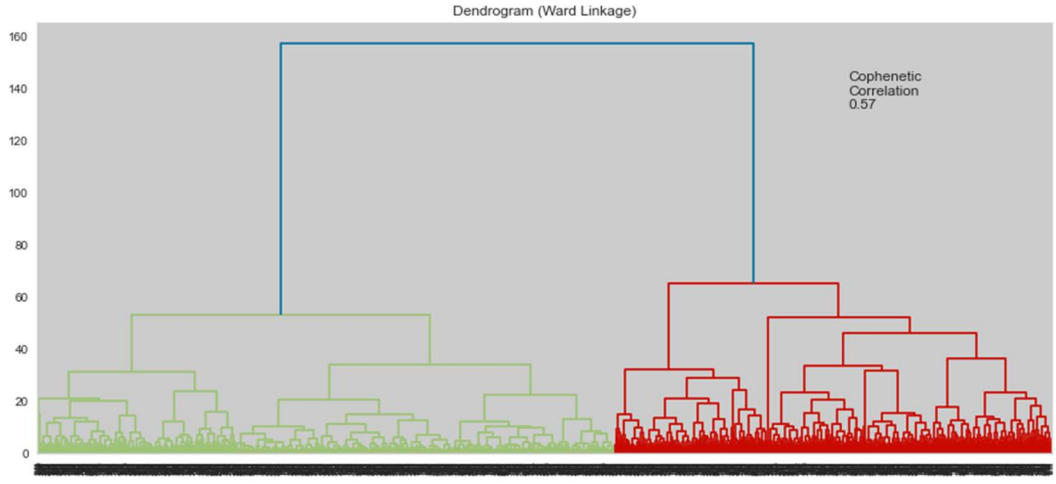
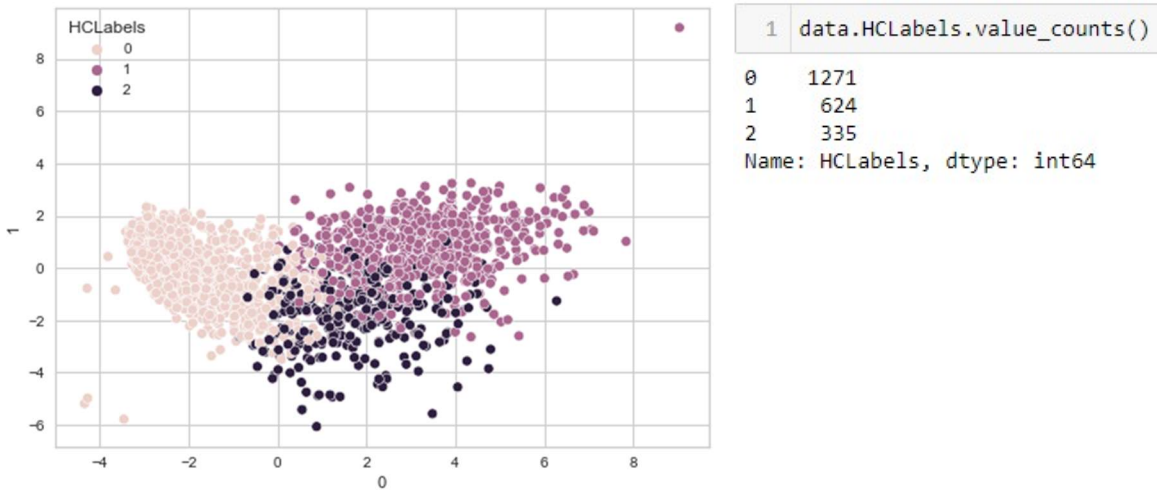


Figure 4 - Euclidean distance and Ward Linkage dendogram



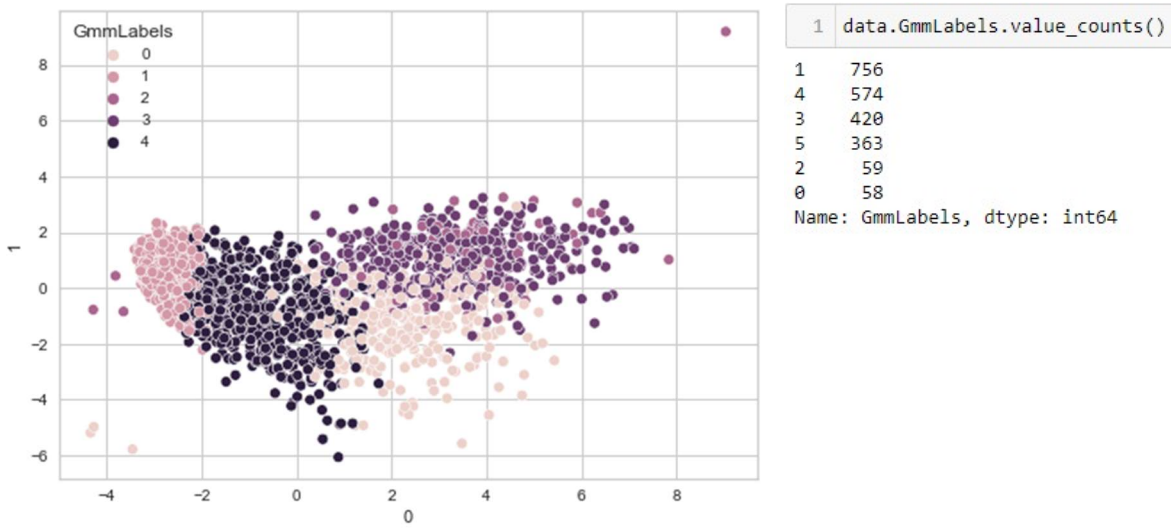

```
1  data.HCLabels.value_counts()
```

```
0    1271
1     624
2     335
Name: HCLabels, dtype: int64
```

Figure 5 - Hierarchical clustering K=3 PCA plot

I tested both K=3 and K=5 once again, and K=3 had the highest Silhouette score of 0.3149. Using this Clustering technique with K=3, I was able to come up with this clustering (See figure 5). Cluster 0 is the low-income customer, youngest age group, biggest family at almost 3, lowest total expenses and purchases, 0 campaigns accepted, and shortest engaged days. This is the young low-income family. Cluster 1 is the high income, small family (2 or less), middle aged group, who spend the most in total, especially on Fruits, Meat, Fish, and sweets, fewest deal purchases, most catalog purchases, and have accepted the most campaigns, near 1, accepted on average. Cluster 2 is the middle-income group, with 2nd largest family average near 3, and oldest age group, who spend the most on wine and gold, by the most deals and on the web purchases, they have the most total purchases, have been engaged with the company the longest, and have a decent chance of accepting a campaign. Although these profiles are very similar to the profiles from K-Mediods and it has a higher Silhouette score, the cluster counts are not as evenly distributed as the K-Mediods clustering. This may cause groups not to be as effectively targeted due to their oversize or limited size.

Next, I tried DBSCAN to see if it could come up with distinguished clusters, but it was not. It resulted in 2 clusters that did not seem to show very good profiles. Even changing values did not improve it much. However, it did have the highest Silhouette score of 0.3441, but that doesn't matter if you can't make anything out of its clustering.




```
1  data.GmmLabels.value_counts()

1    756
4    574
3    420
5    363
2     59
0     58
Name: GmmLabels, dtype: int64
```

Figure 6 - Gaussian Mixture Models (GMM) K=5 PCA Plot

I then moved onto Gaussian Mixture Models (GMM). I tested both K=3 and K=5 and K=3 had a higher Silhouette score of 0.1667 but decided to test K=5 because they were fairly close and produced the PCA plot (See Figure ???). Cluster 0 consists of mid to high income customers, who are the oldest age group, and typically have family size between 2-3, they have high web purchases and store purchases, have the most total purchases, and accept 0-1 campaigns. Cluster 1 consists of the lowest income group, they are the youngest age group and typically have the largest family size near 3, they have the least number of purchases across the board and accept 0 campaigns. Cluster 2 consists of one of the high-income groups, who are middle age group, have a smaller family typically of 2, spend the most on wine and other products, they have the most expenses, highest amount per purchase, and accept the most campaigns average 2. Cluster 3 consists of the other high-income group, who are close to the oldest but have the smallest family below 2, they buy the most fish, are high on other products as well, and typically accept 1 campaign. Cluster 4 consists of a middle-income group, who are close to the oldest group, typically have a larger family size with 1 teen and family size close to 3, they are low to middle on most purchases and do not accept campaigns. These profiles gave some insight but was not quite as clear some of the other clustering methods. Some of the key variable did not show much variance between the different clusters.

Algorithms	K Value	Silhouette Score
K-Means	3	0.2691
K-Mediods	3	0.2889
Heirarchial Clustering (Ward Linkage and Euclidean distance)	3	0.3149
DBSCAN	3	0.3441
Gaussian Mixture Model	3	0.1667

After trying each of the Clustering techniques, there were a few that created good profiles from the clustering. But based off the insights that I was able extract from the different clustering, I believe the K-Mediods ended up having the most insightful profiles while keeping the amount of cluster to a minimum of 3. K-Mediods K=3 also had one of the highest Silhouette scores, meaning that the clusters were classified better than the others. It also had a fairly even distribution counts of customers across the different clusters. I recall one of the mentors in the course mentioning that sometimes simpler is better and if I can get very close to same insights as a more complicated model, I will go with the simpler model. This information will allow the marketing team to target the different customer segments identified more efficiently.

Recommendations for Implementation

Going forward I recommend the marketing team to do targeted marketed towards the different customers segements identified using K-Mediods K=3. The next steps would be to reasscoiate the IDs to each of the customers data and have the K-Mediod labels applied or vise versa and give the listed IDs in each segment to the marketing team. They should then identify the charachersistics of each of the profiles and come up with some sort of label**. Cluster 0, could be simply the mid-income group,** consists of mid to high income customers, who are the oldest age group, and typically have family size between 2-3, they have high web purchases and store purchases, spend a lot on wine similarly to cluster 2, they take advantage of the most deal purchases, have the most total purchases, and most have accepted 1 campaign. This segment could be targeted with web and store deals, most of them are willing to accept campaigns, tend to make lots of smaller purchases, and could be very interested in wine deals. **Cluster 1, low income group,** consists of the lowest income group, they are the youngest age group and typically have the largest family size near 3, they have the least number of purchases across the board and typically accept 0 campaigns. This segment may be harded to target due to their limited discretionary spending. However, they may be more willing to purchase deals or maybe products for their family since they typically have the largest family. **Cluster 2, the high income group,** consists of the high-income group, who are middle age group in late 40s, who have a smaller family typically of 2 with 0 kids or teens at home, spend the most on wine and other products, they make lots of purchases via the catalog and store, they have the most expenses, highest amount per purchase, and accept the most campaigns on average around 1. This group can be easily targeted by sending them catalogs for any products, especially wine, or have propmotions in the store. They are also the most willing group to accept campaigns, so make sure to include them with those.

Benefits going forward from this customer segmentation are expected to be better Return on Investment on different advertisements because they will be able to target the different customer segemnts more effectively. Another benefit would be a more tailored experience for the different customers causing them to be more willing spend money on the company's products and improve customer retention. It will also give the company more awareness of it's own brand.

Some risks with this customer segmentation would be that there may be customers who are missclasified or belong to multiple segments. These customers may feel misaligned or alienated from the brand because they advertising they are getting does not fit to their needs.

This may happen because the Silhouette Score was not the highest for the K-Mediods method at 0.2889. The closer to 1 the more separatred and defined each of the clusters are and less likely to have misclassified customers. The customer segments could also be too narrow or too wide and campaigns may not quite capture the customer effectively.

Further analysis that can be done is to simply get more data so that segments can be more clearly defined. Another thing that could be done is to more strictly define each segment with parameters, so that you can put new customers into the different segments as soon as they become a customer.

All in all, with data science and the different methodologies you can solve many problems in today's world and or make things more efficient. Just about everything we do in today's world produces data and a lot can be done and interpreted with that data. This proect is just another step into the data field and I am excited to continue this journey and hopefully a career in data!

Appendix

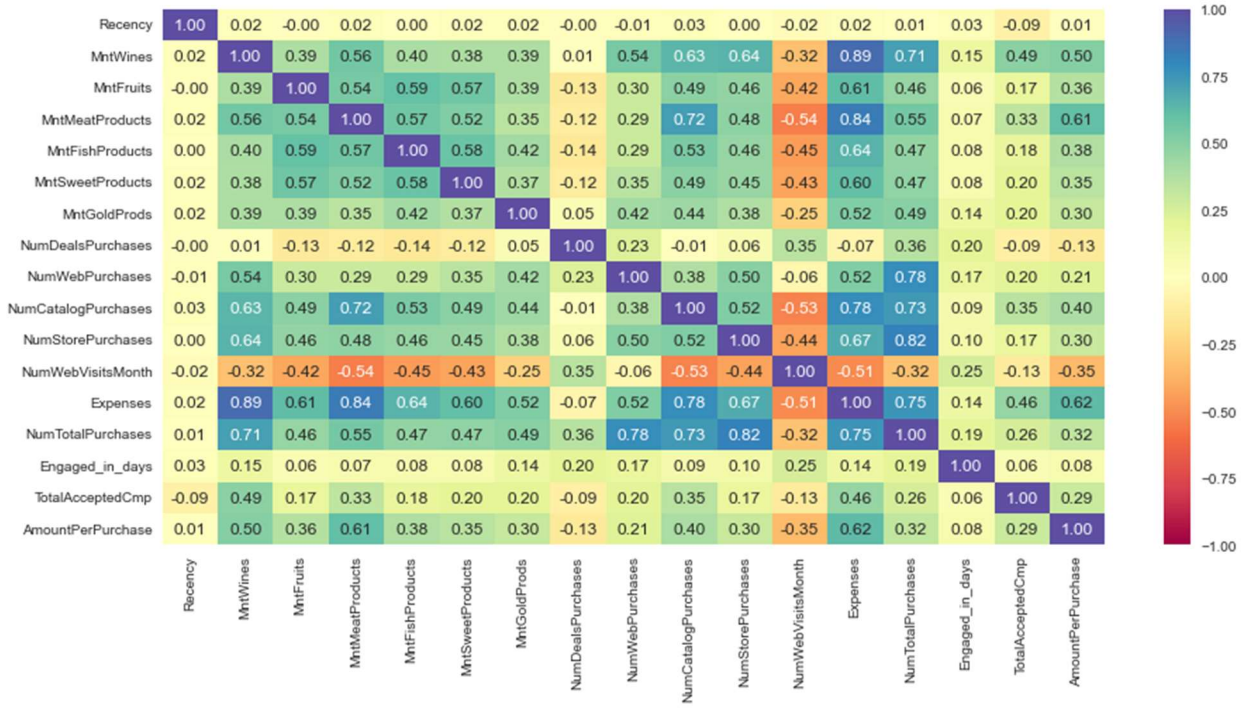


Figure 7 - Correlation Plot

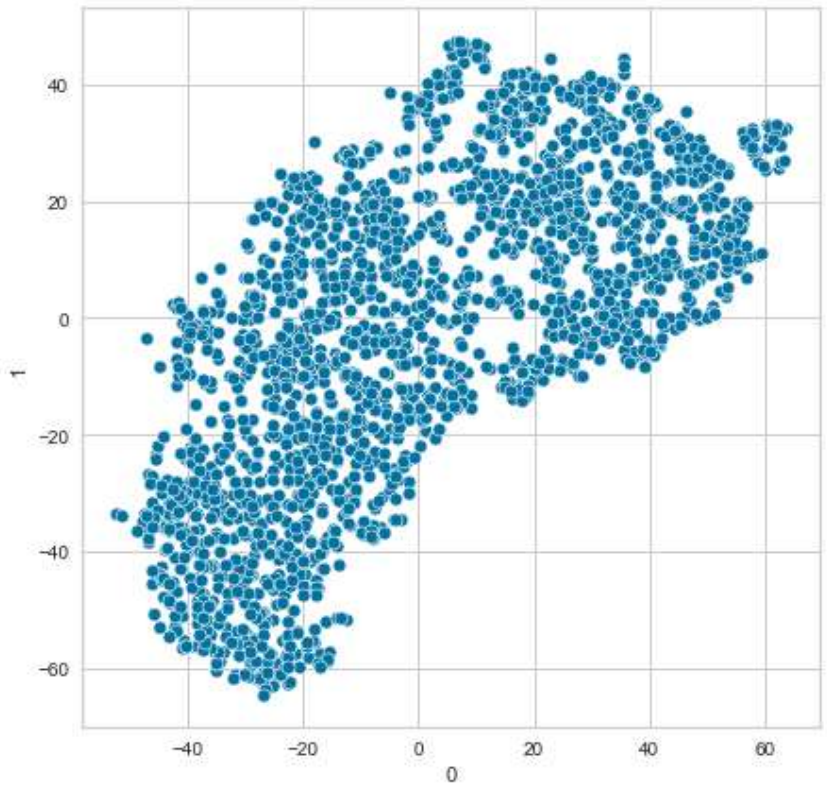


Figure 8 - t-SNE Scatterplot