Use Case: Customer Segmentation

White Papers

1. OBJECTIVE

The goal of this project was to build machine learning clustering algorithms for a consumer goods manufacturing company (hereby referred to as "the company"), with the aim of segmenting its customers into meaningful clusters. The company's ultimate objective from this project was to use the available data to create clusters, such that the company can tailor its sales and marketing efforts to maximize the future value gained from their customers. It is also important to note that the company did not have a specific number of clusters in mind; instead, they allowed the algorithms to determine the optimal number of clusters.

Before this project, the company segmented its customers based on unsophisticated models that were primarily driven by assumptions. The idea of this project was to leverage data, business domain expertise, sophisticated algorithms, and market experience to effectively and productively segment the company's customers.

2. DATA

Since the company did not have a well-developed data infrastructure, it had not collected relevant customer-specific data in the past. However, it collected transaction data (invoice data), which we had to work with to engineer variables that we could use to execute our clustering algorithms. The company's infrastructure allowed us to extract a maximum of 541,909 rows of transaction data and 8 explanatory variables. Figure 1 below shows a sample of the data frame used (please note that the data has been changed for privacy reasons).

Figure 1

Invoice No.	Stock Code	Description	Quantity	Invoice Date	Unit Price	Customer ID	Country
534222	8111A	BLACK HANGING T-LIGHT HOLDER	8	01-12-2019 08:33	2.45	17840	United States of America
534222	8222E	STEEL LANTERN	6	01-12-2019 08:33	3.40	17840	United States of America
534222	72010	CREAM COAT HANGER	6	01-12-2019 08:33	2.75	17840	United States of America
534222	90000	KNITTED COLD WATER BOTTLE	12	01-12-2019 08:33	3.49	17840	United States of America
534233	84029F	RED WOOLLY WHITE HEART	2	01-12-2019 08:47	3.59	17842	United States of America

3. PREPROCESSING

In order to convert the available dataset into one that was suitable for our algorithms, we had to carry out several preprocessing and cleaning steps. The first step was to handle missing values and negative values in such a way that it did not affect the outcome of our clustering models.

The second step was to perform feature engineering, which is the process of creating variables from the available data that can be used for customer segmentation. The goal of this step was to extract meaningful customer-specific information from the transaction data and to transform

the structure of the data such that each row represented a customer and not a line item of an invoice. From the available data, we were able to engineer numerous variables, some of which are highlighted below:

1. **Monetary**: Total revenue obtained from each customer
2. **Recency**: Number of days since last purchase for each customer
3. **Frequency**: Frequency of purchases (number of transactions) for each customer
4. **Variety**: Number of products purchased (unique number of products) by each customer
5. **Average Purchase Amount**: Average invoice value of each customer
6. **Longevity**: Number of days since first purchase for each customer (how long they have been customers)
7. **Geography**: Country each customer is from

From the above list of variables, one can infer that we have converted unsuitable transaction data into variables that have the potential to create successful clusters. These variables (and the others that have not been listed) contain information on customers that could be used to segment them based on their value to the company. The number of rows decreased from 541,909 line items to 4,372 customers.

It is important to note that these variables intrinsically incorporate the concept of RFM (Recency, Frequency, Monetary Value) that is popularly used to segment customers. Figure 2 contains an example of the dataset after this feature engineering step (please note that the data has been changed for privacy reasons).

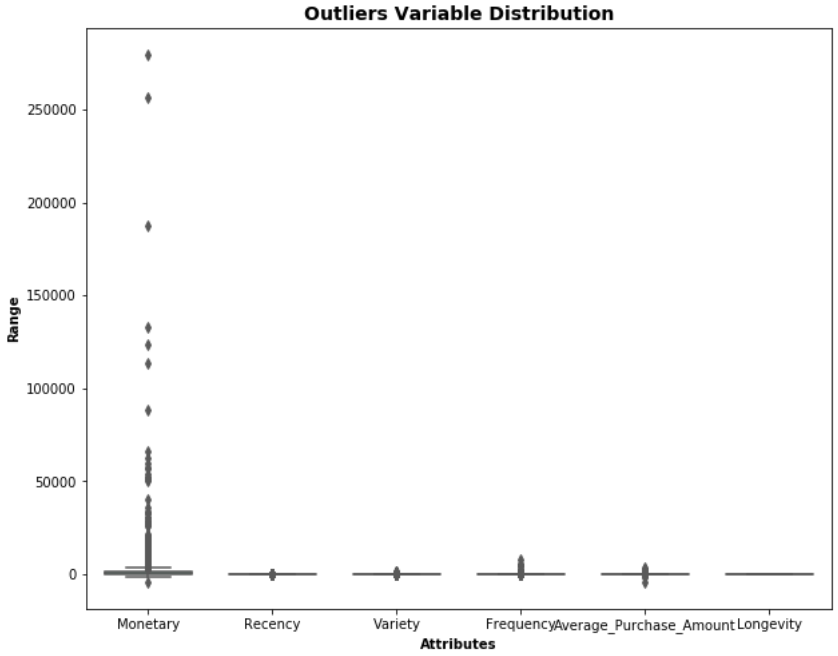
Figure 2

Customer ID	Monetary	Recency	Frequency	Variety	Average Purchase Amount	Longevity	Geography
11350	21.00	325	2	1	10.50	325	United States of America
11351	4310.00	1	182	103	23.68	366	Canada
11352	1797.24	74	31	22	57.98	357	United States of America
11353	1757.55	18	73	73	24.08	18	Mexico
11354	334.40	309	17	17	19.67	309	United Kingdom

The third step was to duplicate the current data frame and remove the outliers. The reason for this step is that clustering algorithms tend to perform poorly with outliers in the data, as they distort the resulting clusters. Out of the 4,372 customers, the number of outlier customers removed were around 150. The outliers will be individually assigned to the resulting clusters at the end of the project to maximize the accuracy and value of the clustering project. As shown in Figure 3 below, most of the outliers seem to be clients with unusually large values in the monetary variable (left-most boxplot), which can be easily assigned to the high-worth cluster at the end of the project.

It is essential to note that we attempted to perform clustering algorithms on the dataset that included the outliers as well as the dataset that excluded the outliers. Our goal was to pick the algorithm that performed the best at segmenting customers.

Figure 3



The fourth step was to standardize (scale) the numerical variables, such that each variable was on the same scale. This is necessary to ensure that no single variable dominates the clustering process by virtue of its numerical scale.

Furthermore, before moving on to the modeling section of our project, we performed a holistic exploratory data analysis to understand the characteristics and nature of the data.

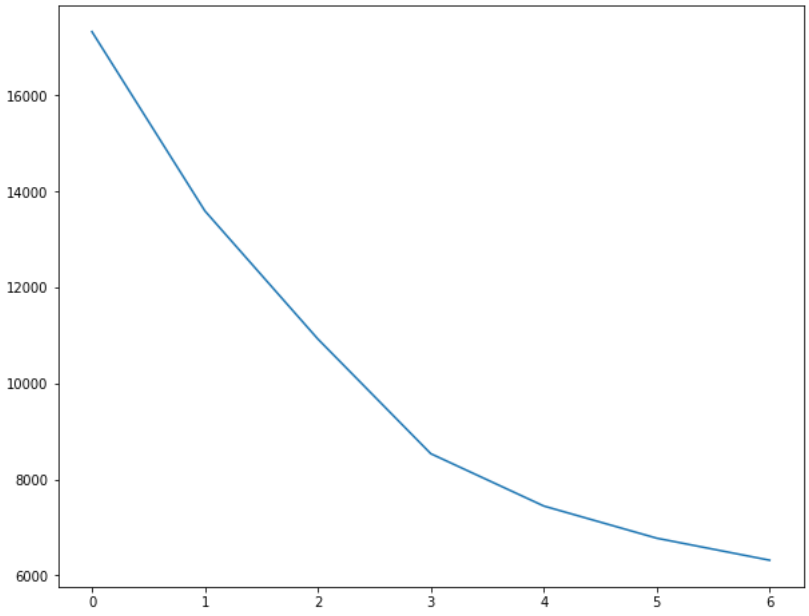
4. MODEL BUILDING #1: ALL FEATURES (WITHOUT OUTLIERS)

After cleaning the data, we moved on to the modelling stage of the project. In the first model building experiment, we used the cleaned dataset that excluded the outliers. Moreover, in this experiment, we used every single variable that we previously engineered. In other words, for these benchmark clustering models, we did not select a subset of the variables based on their clustering power.

4.1: Selecting Optimal Number of Clusters

Since the company did not specify a particular number of clusters, we had to determine the optimal number of clusters based on the outputs from the clustering algorithms. To do this, we used two different methods with the K-Means Clustering Algorithm: the Elbow Method and Silhouette Scores. The results of the Elbow Method are shown in Figure 4 below. Both methods indicated that the optimal number of clusters for this dataset was 3. Therefore, for this model building experiment, we customized the algorithms to produce an output of three distinct clusters.

Figure 4



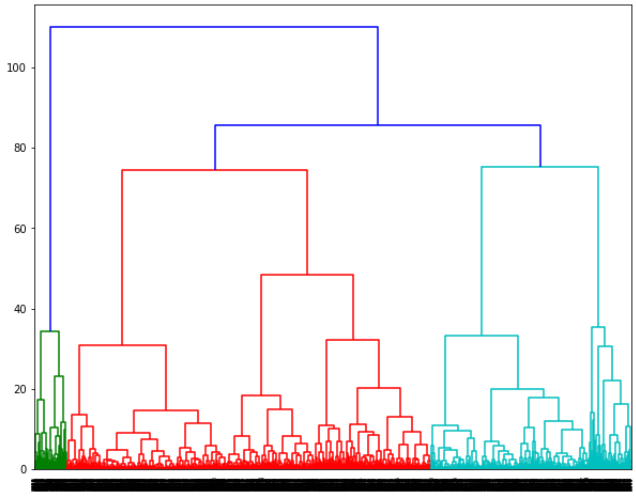
4.2: K-Means Clustering

Next, we developed and executed the K-Means Clustering machine learning algorithm on the cleaned dataset to obtain 3 clusters. We tested different combinations of model hyperparameters to improve the clustering process.

4.3: Hierarchical (Agglomerative) Clustering

After developing the K-Means Clustering algorithm, which is a non-hierarchical clustering method, we developed and executed the Agglomerative Clustering algorithm on the cleaned dataset to obtain 3 clusters. In this step, we tested different linkage methods (Single Linkage, Complete Linkage, Average Linkage, and Ward), and determined that the Ward method produced the most meaningful clusters. Therefore, in the final iteration of this algorithm, we created the Agglomerative Clustering model with the Ward Linkage parameter. Figure 5 contains the Dendogram of the Ward Linkage Agglomerative Clustering model.

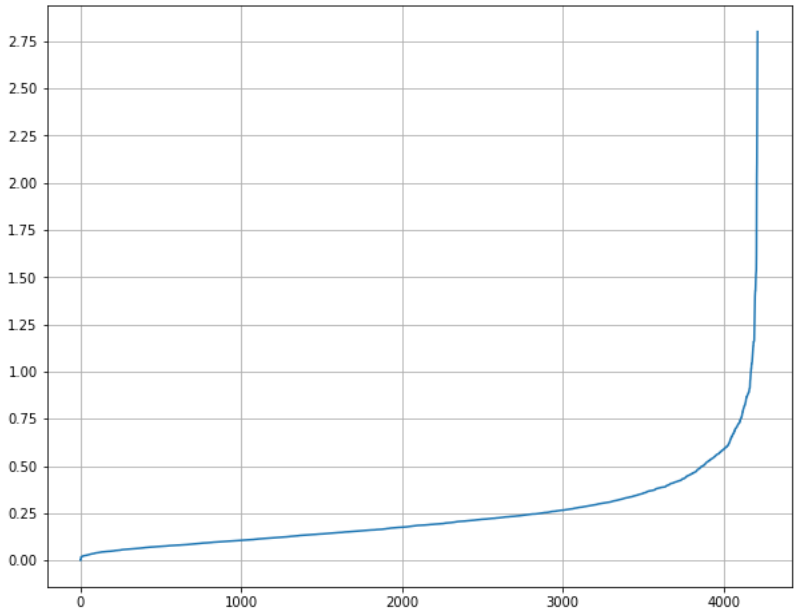
Figure 5



4.4: DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Next, we created the DBSCAN model on the cleaned dataset. During this process, we had to determine the optimal 'epsilon' and 'min_samples' parameters. We optimized epsilon by calculating the distance to the closest n points for each data point and plotting the results. We picked the epsilon value that yielded the most pronounced change. From Figure 6 below, one can conclude that an epsilon value of approximately 0.60 yields the most pronounced change.

Figure 6



Using an epsilon value of 0.60, we tried minimum samples of 5 and 50, and we concluded that min_samples = 50 resulted in more meaningful clusters. Using these parameters, we developed the final iteration of this algorithm.

4.5: OPTICS (Ordering Points to Identify the Clustering Structure)

Next, we created an OPTICS model on the cleaned dataset, which overcomes one of DBSCAN's biggest shortcomings: the problem of detecting meaningful clusters in data of varying density. In addition, since an OPTICS model does not require the Epsilon parameter to be tuned, we created multiple models on different min_samples, and picked the one that resulted in the best clusters. Relatively speaking, this model did not perform as well as the previous algorithms.

4.6: Mean Shift Algorithm

The final machine learning algorithm that we used on the dataset was the Mean Shift clustering algorithm. We experimented with different combinations of model hyperparameters to improve the clustering process. It is important to note that the Mean Shift algorithm detects the optimal number of clusters on its own; in other words, we could not force the algorithm to segment the data into a specific number of clusters (3 in our case). As such, this algorithm picked 6 as the optimal number of clusters. However, the results of this algorithm were not as meaningful as compared to the first few algorithms.

4.7: Results of Model Building #1

Out of all the models built in this first experiment, the clusters formed by the K-Means Clustering algorithm and the Agglomerative Clustering algorithm seem to be the most meaningful, in terms of splitting the customers into distinct clusters. The insights/results of only the best clustering model are shown at the end of this white paper.

5. MODEL BUILDING #2: ALL FEATURES (WITH OUTLIERS)

In the second model building experiment, we used the cleaned dataset that included the outliers. In this experiment, we used every single variable that we previously engineered. In other words, we did not select a subset of the variables based on their clustering power.

We followed the same process in this model building stage. First, we used the elbow graph and silhouette score to determine that the optimal number of clusters for this dataset was 2. With that in mind, we developed and executed the following clustering algorithms: (1) K-Means Clustering, (2) Agglomerative Clustering, (3) DBSCAN, (4) OPTICS, and (5) Mean Shift. We tuned and optimized the relevant hyperparameters, wherever applicable (for eg: epsilon in DBSCAN and linkages in Agglomerative Clustering).

Finally, we compared the results of the different models and concluded that all the algorithms performed similarly in clustering the customers. However, an important point to note is that the resulting clusters from the second model building experiment were not as meaningful or significant as the resulting clusters from the first model building experiment. This conclusion is consistent with our initial assumption that outliers distort the clustering power of machine learning algorithms.

6. MODEL BUILDING #3: MICROSOFT AZURE'S MACHINE LEARNING STUDIO

In the third model building phase, we replicated a part of the first model building phase on Microsoft Azure's Machine Learning Studio. More specifically, we created the K-Means Clustering algorithm using Azure ML with the cleaned dataset that excluded outliers and included all the variables.

In this experiment, we segmented the customers into 3 distinct clusters, as we previously determined that the optimal number of clusters using this dataset is 3.

While evaluating the results of this clustering algorithm, we concluded that this algorithm performed well, but it did not outperform the custom models that we created in the first model building phase. Since the first model building experiment produced the best results, we decided to iteratively improve its performance in the next phase.

7. MODEL BUILDING #4: FEATURE SELECTION

In the final model building experiment, we improved the performance of the previous best models by implementing a feature selection process. The objective of this process was to use Principal Feature Analysis (PFA)[1], which is a technique that is fundamentally based on

[1] Read about PFA at http://venom.cs.utsa.edu/dmz/techrep/2007/CS-TR-2007-011.pdf

Principal Component Analysis (PCA), to determine which variables had the highest clustering power. The ultimate goal was to eliminate the insignificant variables (noise) that did not add value to or negatively affected the clustering process.

The feature selection process helped reduce the dimensionality of the dataset to only include powerful variables such as Monetary, Recency, Frequency, Average Purchase Amount, and so forth.

Subsequently, we built the following clustering algorithms using the selected features with high clustering power: (1) K-Means Clustering, (2) Agglomerative Clustering, (3) DBSCAN, (4) OPTICS, and (5) Mean Shift. We tuned and optimized the relevant hyperparameters, wherever applicable (for eg: epsilon in DBSCAN and linkages in Agglomerative Clustering).

8. MODEL SELECTION

After the fourth model building experiment, we compared the resulting clusters of each model building experiment. We concluded that the clustering algorithms of the fourth model building phase produced the best results. More specifically, the feature selection process successfully selected significant variables that improved the clustering algorithms.

Within the fourth model building experiment, the K-Means algorithm yielded the best clusters. The resulting clusters had data points whose characteristics were similar to other data points within its own cluster but were different from the characteristics of the data points in other clusters.

At this stage, since we used the dataset without outliers to build the clusters, we had to assign the outliers into the most appropriate clusters.

9. RESULTING CLUSTERS AND INSIGHTS

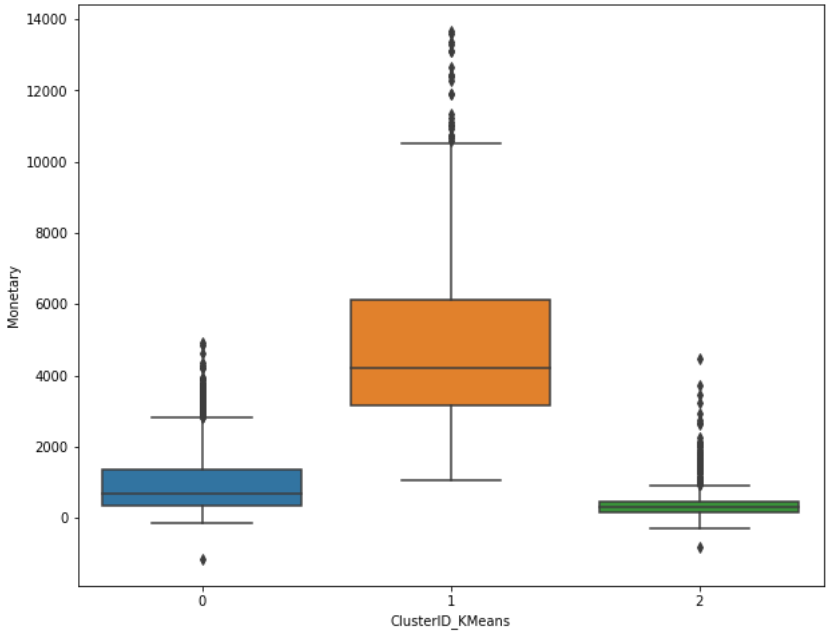
After creating the final clusters, we created an output dataset that assigned each customer into a specific cluster. The customers were segmented into three different clusters: Cluster 0, Cluster 1, and Cluster 2. Figure 7 below shows a sample of the output dataset with a few variables and the assignment of clusters (please note that the data has been changed for privacy reasons).

Figure 7

Customer ID	Monetary	Recency	Frequency	Variety	Average Purchase Amount	Longevity	Geography	Cluster
11350	21.00	325	2	1	10.50	325	United States of America	2
11351	4310.00	1	182	103	23.68	366	Canada	1
11352	1797.24	74	31	22	57.98	357	United States of America	0
11353	1757.55	18	73	73	24.08	18	Mexico	0
11354	334.40	309	17	17	19.67	309	United Kingdom	2

The following visualizations demonstrate the characteristics of the different clusters. These visualizations showcase that the clustering algorithm was successful in distinguishing customers into meaningful segments.

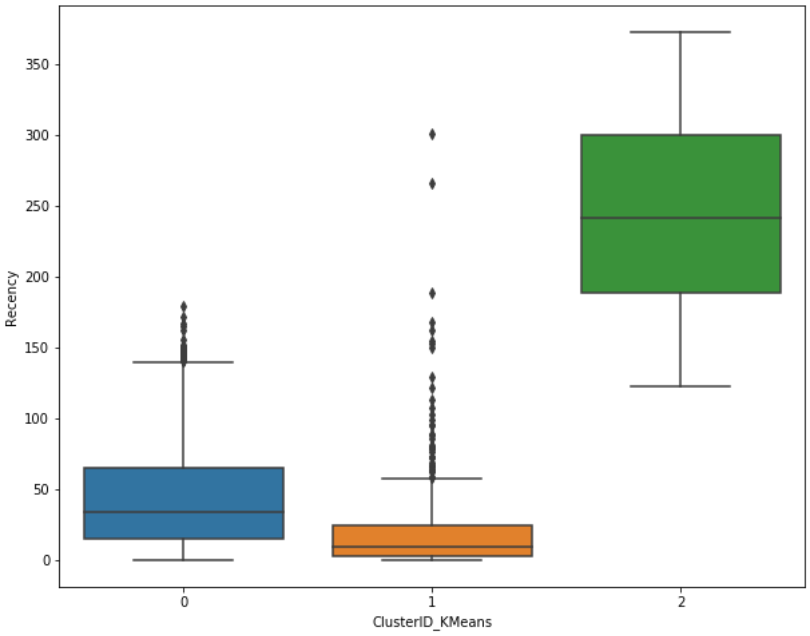
Figure 8



The insights from Figure 8 are as follows:

- Customers in Cluster 1 tend to be high spending customers
- Customers in Cluster 2 tend to be low spending customers
- Customers in Cluster 0 tend to be low-to-medium spending customers

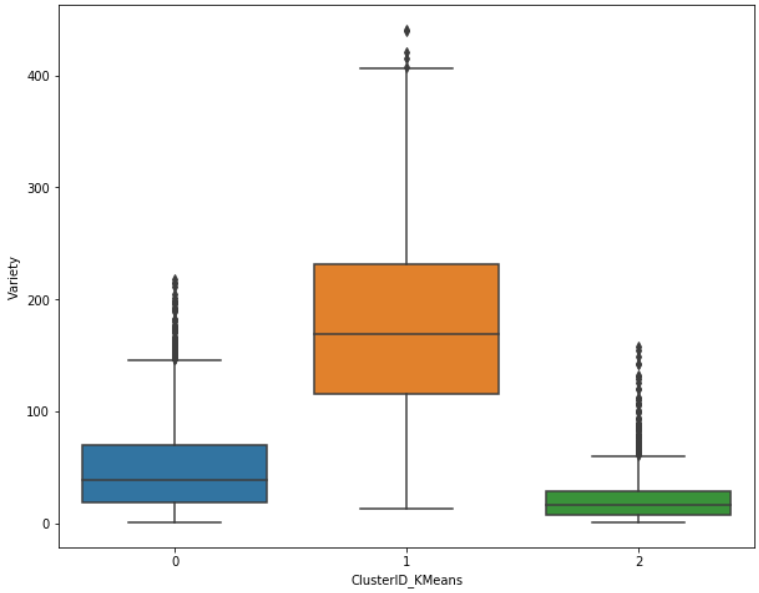
Figure 9



The insights from Figure 9 are as follows:

- Customers in Cluster 1 have a low number of days since their last purchase (recent customers)
- Customers in Cluster 2 have a high number of days since their last purchase (non-recent customers)
- Customers in Cluster 2 have a low-to-medium number of days since their last purchase (semi-recent customers)

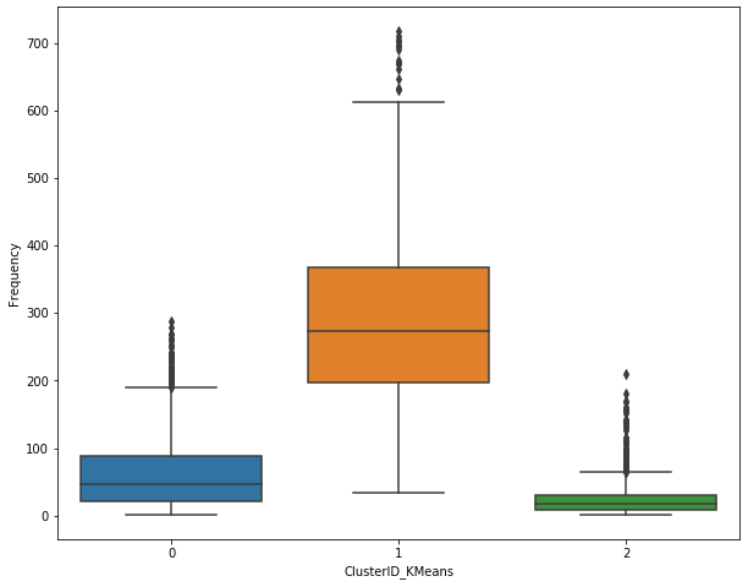
Figure 10



The insights from Figure 10 are as follows:

- Customers in Cluster 1 purchase a high variety of products
- Customers in Cluster 2 purchase a low variety of products
- Customers in Cluster 0 purchase a low-to-medium variety of products

Figure 11



The insights from Figure 11 are as follows:

- Customers in Cluster 1 are frequent buyers (high frequency of purchases)
- Customers in Cluster 2 are infrequent buyers (low frequency of purchases)
- Customers in Cluster 0 are relatively infrequent buyers (low-to-medium frequency of purchases)

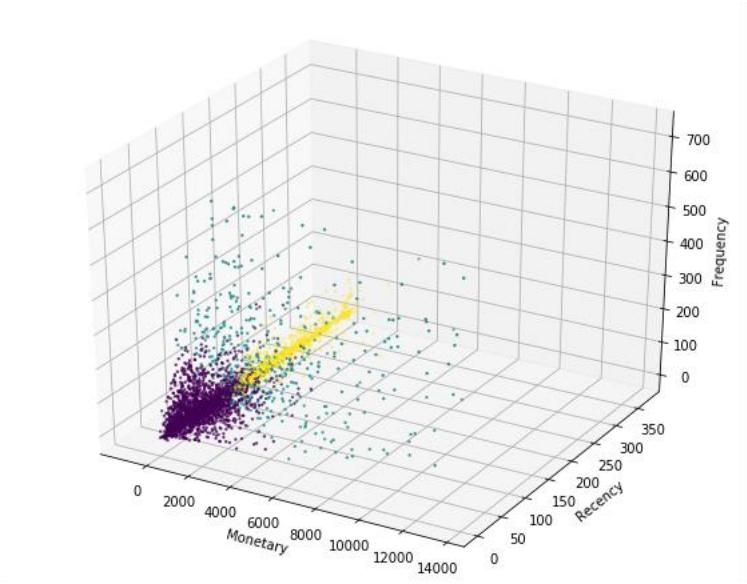
A summary of the resulting clusters' characteristics is given in Figure 12 below:

Figure 12

Characteristic	**Cluster 0**	**Cluster 1**	**Cluster 2**
Monetary	Low-Medium Spending	High Spending	Low Spending
Recency	Semi-Recent Customers	Recent Customers	Non-Recent Customers
Variety	Low-Medium Variety	High Variety	Low Variety
Frequency	Relatively Infrequent Customers	Frequent Customers	Infrequent Customers
Longevity	Short-to-Medium Term Customers	Long Term Customers	Short-to-Medium Term Customers

Another visualization that showcases the distinction between the three clusters is given in Figure 13 below. Since we are only able to view data in three-dimensions at the most, we decided to pick three variables (RFM) to visualize how the customers in different clusters vary from each other.

Figure 13



This Figure showcases the three different clusters:

- **Cluster 0 (Purple):** Low-to-Medium Monetary, Low-to-Medium Recency (Semi-Recent), and Low-to-Medium Frequency (Relatively infrequent)

- **Cluster 1 (Blue):** High Monetary, Low Recency (Recent), and High Frequency (Frequent)
- **Cluster 2 (Yellow):** Low Monetary, High Recency (Non-Recent), and Low Frequency (Infrequent)

10. BUSINESS INTELLIGENCE DASHBOARD

After segmenting the customers into different clusters, we also provided the company with a dashboarding and reporting solution (made in PowerBI) that visualized key characteristics of each company and each cluster. This interactive dashboard allowed the company to ingest valuable information and answer key questions about specific customers and specific clusters.

11. CONCLUSION

In essence, we helped the company segment its customers into meaningful clusters using merely transactional data. A future step with this company will be to revisit the clustering process once they collect customer-specific data. We have seen that the companies in Cluster 1 are high-priority customers, whereas the companies in Cluster 2 are low-priority customers.