# IEOR 4523 Data Analytics Project: Customer Analysis

Asmae Bouzid, Chang Liu, Sukun Wang, Xinyi Wu, Yanchen Liu

December 9, 2021

**Abstract**

Customer analytic has proven to be a powerful tool. In fact, a good understanding of the customers results in many benefits for the company. To name a few, customer analytic helps increase sales and improve profit margins, it also helps with the customer retention and the marketing efficiency. In this project, we explore a dataset of 2240 customer to understand their consuming behaviours in order to predict the most adequate campaign for each of them. To achieve this, we start by processing our data before running a clustering model to it. Our k-means model yields clusters of customers which we analyze and visualize through PCA-visualization.

## 1 Data Processing

The dataset of our project is a customers dataset that can be found on the platform Kaggle. It consists of 2240 rows; each row corresponds to a customer. The dataset contains 29 features including 3 attributes:

- People: ID, Year of Birth, Education, Marital Status, Income, Kids.
- Consuming Behaviours: Product preference (Wines, Fruits, Meat, Fish, Sweet, Gold), Consuming frequency, Shopping place.
- Feedback: Campaign preference(Campaign 1 to 5),Complaint/Response

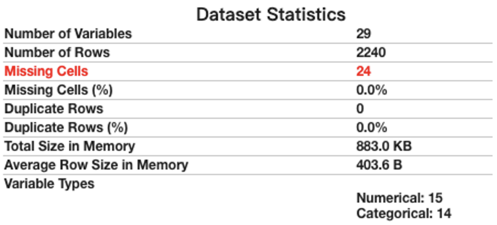
In this part, we will show the different steps that we have undergone in the processing of our data. First of all, We will start with displaying the problems that we identified in the dataset. Secondly, we will explain what measures we have taken in order to clean our data. Finally, we will go through the different data processing steps that we followed from columns manipulation to data encoding and standardization.

### 1.1 Some problems with the dataset

Our dataset does not suffer from major abnormalities. However, in order to get the best performance out of our clustering model, we aimed to have a data as clean as possible. Here are some problems that we identified in our data:

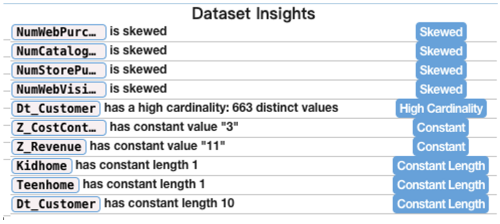
- Missing values: In our dataset, we have overall 24 missing cells. Fortunately the percentage of missing data is not significant; less then 1 percent of the total rows. Therefore dropping these rows would not impact the performance of our clustering model.
- Consuming Behaviours: Product preference (Wines, Fruits, Meat, Fish, Sweet, Gold), Consuming frequency, Shopping place.
- Feedback: Campaign preference(Campaign 1 to 5), Complaint, Response

Figures 1 and 2 show some statistics and insights of our data:



| Dataset Statistics | |
|---|---|
| Number of Variables | 29 |
| Number of Rows | 2240 |
| Missing Cells | 24 |
| Missing Cells (%) | 0.0% |
| Duplicate Rows | 0 |
| Duplicate Rows (%) | 0.0% |
| Total Size in Memory | 883.0 KB |
| Average Row Size in Memory | 403.6 B |
| Variable Types | Numerical: 15<br>Categorical: 14 |

Figure 1: Some statistics of the dataset



| Dataset Insights | |
|---|---|
| NumWebPurc… is skewed | Skewed |
| NumCatalog… is skewed | Skewed |
| NumStorePu… is skewed | Skewed |
| NumWebVisi… is skewed | Skewed |
| Dt_Customer has a high cardinality: 663 distinct values | High Cardinality |
| Z_CostCont… has constant value "3" | Constant |
| Z_Revenue has constant value "11" | Constant |
| Kidhome has constant length 1 | Constant Length |
| Teenhome has constant length 1 | Constant Length |
| Dt_Customer has constant length 10 | Constant Length |

Figure 2: Dataset insights

## 1.2 Data Cleaning and Columns Manipulation

Our data cleaning process consisted of dropping the following items:

- Missing values: As mentioned above, we only had 1 percent of rows with missing data. Dropping these rows would harm the analysis.
- Some irrelevant features: namely CostContact and Revenue.
- One outlier data point within the Income column.

Beside data cleaning, we did some columns manipulation. In fact, our data contains some categorical information that needed to be addressed. These columns are: Marital Status and Education. For the Marital Status column, it originally contained 8 unique values as shown in figure 3. We turned this column into a feature with only two possible values: Together and Alone. The graph of distribution of these two values is shown in figure 4. Moreover, we encoded these categorical values into numerical ones as shown in figure 5.

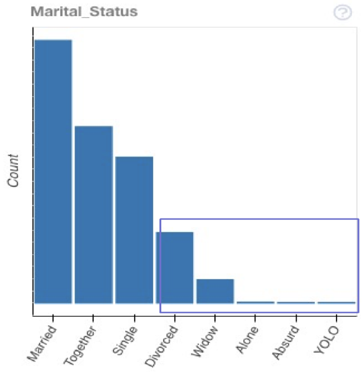


Figure 3: Marital Status before the processing

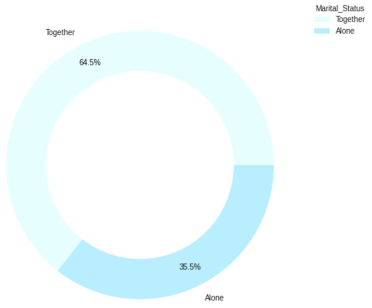


Figure 4: Marital Status values after the processing

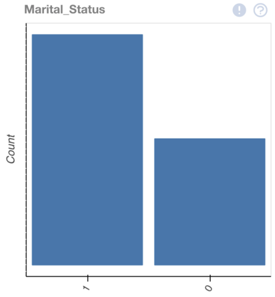


Figure 5: Marital Status values numerically encoded

Regarding the Education column, it originally contained 5 unique values that we combined into three values: Highly educated, Average educated and Less educated (Figure 6). These three values were numerically encoded as well, as shown in figure (Figure 7).

Furthermore, we converted some consumption metrics that had a one sided frequency distribution (figure 8) into percentages; Times of each kind of product consumption/Total number of consumption (figure 9). The final step of our data processing was the standardization of features.

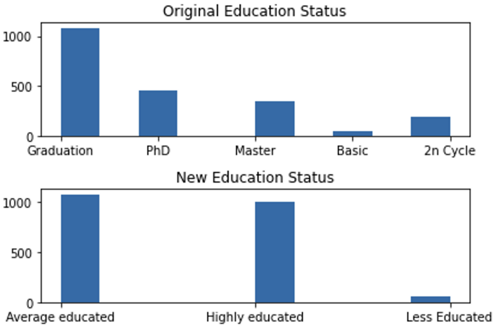


Figure 6: Education column processing

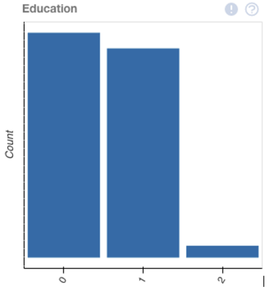


Figure 7: Education column encoded

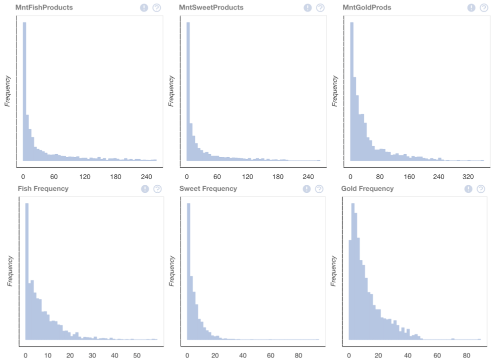


Figure 8: Consumption metrics before conversion

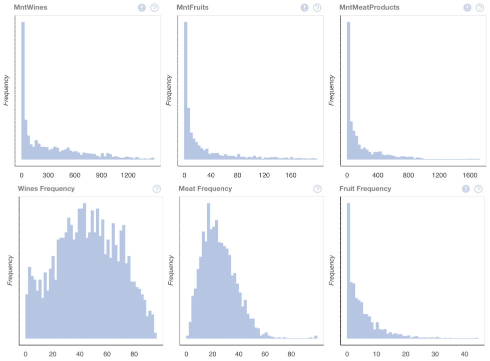


Figure 9: Consumption metrics after conversion

# 2 Clustering

## 2.1 KMeans model

After undergoing all the necessary data processing steps, we moved to building the clustering model. We made the decision to make our clustering using the KMeans model from the sklearn.cluster library. In order for us to pick the best number of clusters k, we combined two methods; the Elbow method using inertia and the Silhouette analysis. The Elbow method we used consists of plotting inertia as a function of the number of clusters k. Where inertia is computed as the sum, across each cluster, of the square distances between each data point and the centroid of that cluster. The best k corresponds to the point where the decrease of inertia begins to slow. In our case, the visualization of this plot (figure 10) gave us two best possible values for k; 4 and 5. To choose between these two candidate values for k, we used the Silhouette visualizer from the yellowbrick.cluster library. This method consists of plotting silhouette coefficients per cluster for different values of k. The silhouette coefficient is measured by taking the average of intra-cluster distance and nearest-cluster distance for each sample normalized by the maximum value. atahe resulting score lies between -1 and +1, where a score close to +1 is obtained when clusters are highly separated and a score close to -1 is obtained when the samples are potentially being assigned to the wrong cluster. Beside the silhouette score, the thickness of the clusters is another criterion for choosing the best k. In fact, a homogeneity of thickness reflect a good balancing among clusters. Taking into account the criteria mentioned above, our plots (figures 11 and 12) made us choose k = 5 for our clustering model.

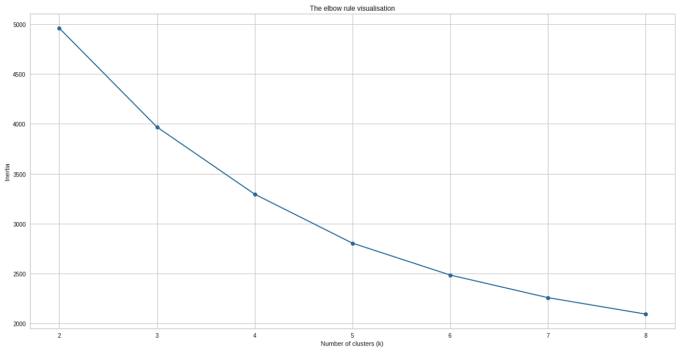


Figure 10: Elbow method visualization

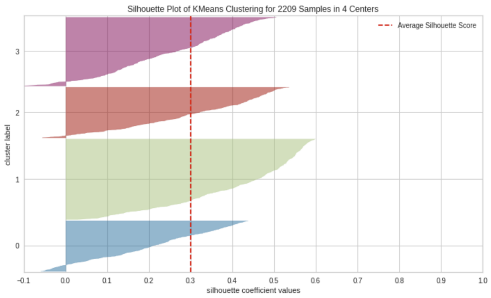


Figure 11: Silhouette visualization for k=4

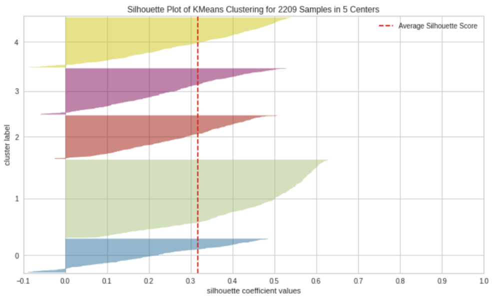


Figure 12: Silhouette visualization for k=5

## 2.2 Clustering performance

As explained in the previous subsection, we chose to apply a KMeans model for clustering. Our choice for the number of clusters was well studied. In fact, we have combined two methods to make this choice: the Elbow method using inertia and the Silhouette visualization using the silhouette coefficients and the thickness of clusters as criteria. This yielded a nicely averaged and separated 5 clusters as we can see in the figure 13). We can also see how separated the clusters are in the 3D figure 14). In this 3D plot, the clusters are displayed in terms of the three following features: Age, Income and AvgCheck. However, the separation of clusters is not perfectly demonstrated in this 3D plot. Therefore we decided to attempt a visualization of cluster using the PCA method to reduce the dimensions to 2D. The result of this PCA visualization is shown in figure 15).

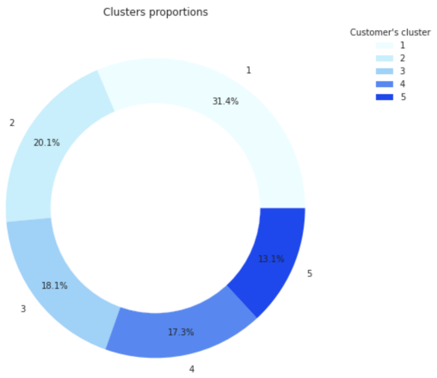


Figure 13: Clusters proportions

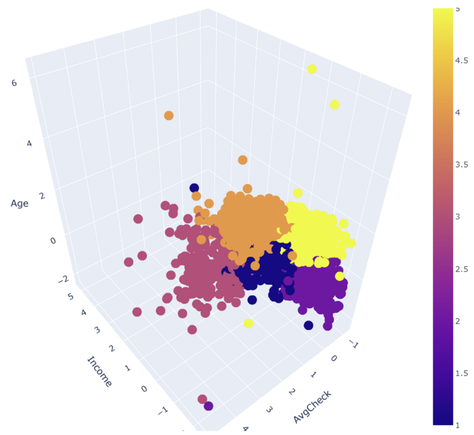


Figure 14: 3D visualization of clusters

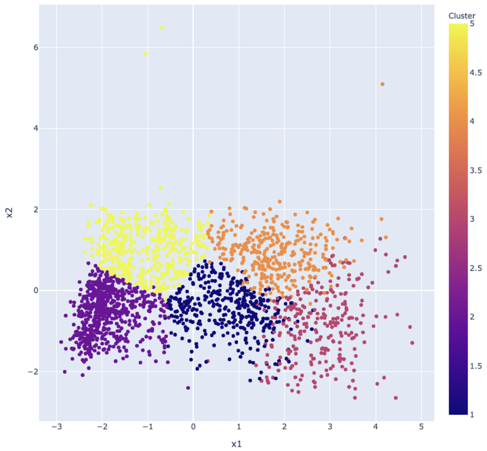


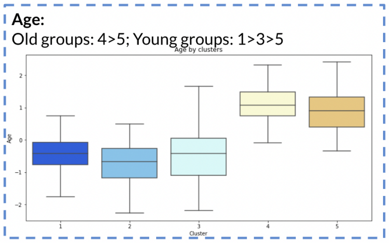
Figure 15: PCA visualization of clusters

# 3 Clustering Analysis

## 3.1 Clusters features

The KMeans model that we applied yielded 5 clusters. In this subsection we are going to highlight how some important features are distributed within these clusters.

- Age: By plotting the distribution of age within each of the clusters, we can see that clusters 4 and 5 regroup the ”oldest” customers, whereas the population of clusters 1, 2 and 3 is younger (figure 16).
- Income: The plot of how Income is distributed per cluster (figure 17) shows that clusters can be ranked in the following order from the highest to the lowest income: 3, 4, 1, 5, 2.
- Education and Marital Status: Figure (figure 18) shows how the features Education and Marital Status are distributed per cluster.



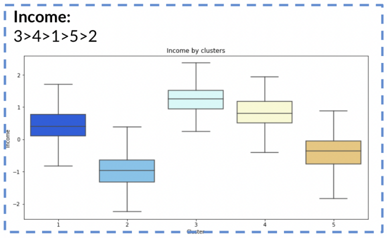


Figure 16: Distribution of Age per cluster

Figure 17: Distribution of Income per cluster

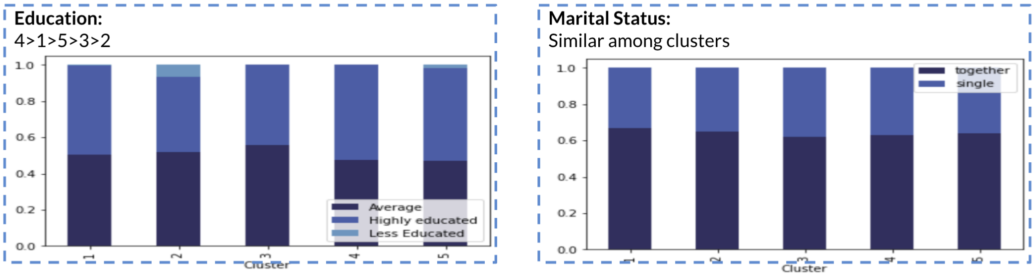


Figure 18: Distribution of Education and Marital Status per cluster

## 3.2 Consuming behaviours

In this subsection we are going to focus on comparing the consuming behaviours among our 5 clusters. Plots in figure 19 show that there is a shared pattern among all clusters regarding their consumption of Wine, Meat, Fruit, Fish and Sweets. In fact, all clusters tend to spend the most on Wine, and the least on Sweets. All clusters tend to have approximately the same gold consumption, except for cluster 2 that has a slightly higher gold consumption than the other clusters. Figures 20 and 21 respectively show the distribution of Online shopping frequency and Amount of purchase within each of the clusters. Lastly, figure 22 displays the percent of complaint per cluster.

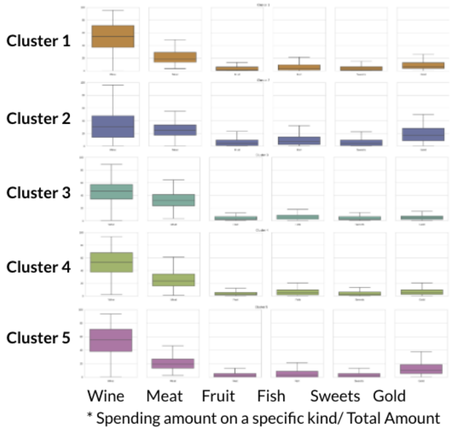


Figure 19: Spending amount on specific products per cluster

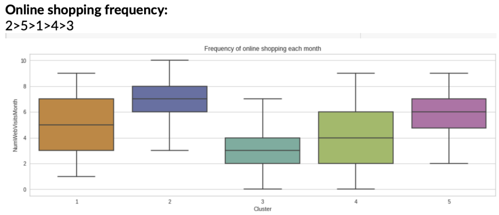


Figure 20: Distribution of online shopping frequency per cluster

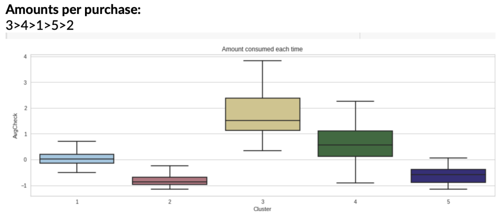


Figure 21: Distribution of Amount per purchase per cluster

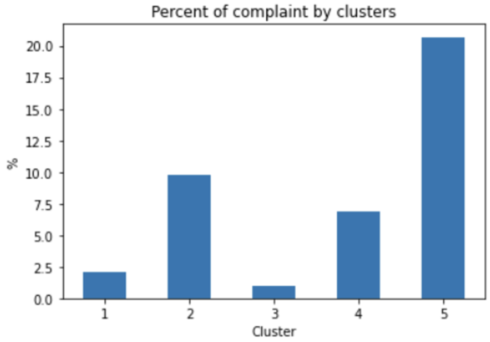


Figure 22: Percent of complaint per cluster

## 3.3 Campaigns recommendations and conclusion

Based on the 5 clusters that we obtained through our KMeans model, we were able identify the most adequate campaign for each cluster. In fact, as shown in 23 Cluster 1 (plots in dark blue) has almost only one campaign that is most adequate, which is campaign 3. Although for the other clusters their is not a single campaign that stands out as significantly as campaign 3 for cluster 1, we can see that the most adequate campaigns are respectively 5 and 1 for clusters 2 and 3 and campaign 4 for both clusters 4 and 5. We were also able to draw the conclusions synthesised in figure 24.

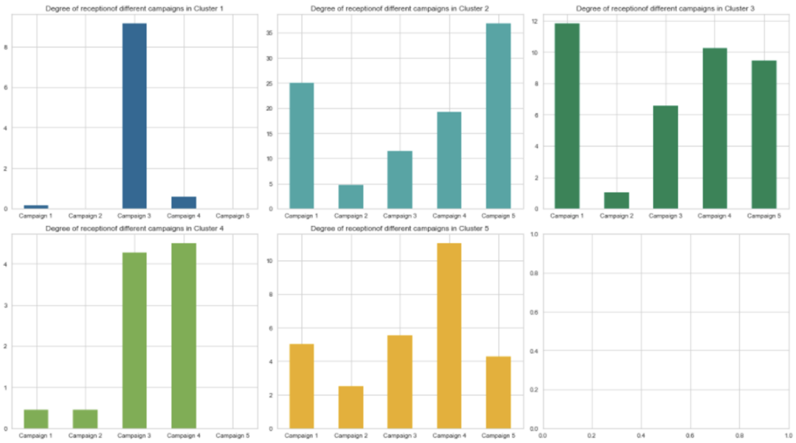
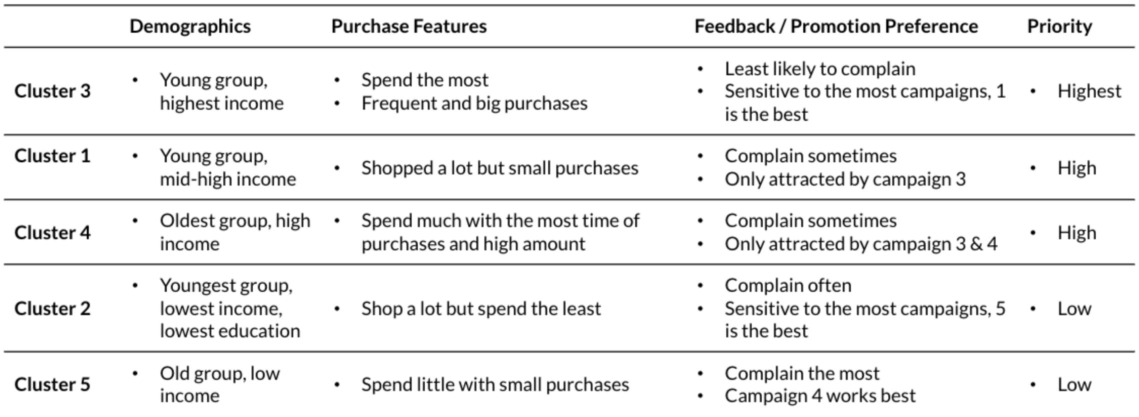


Figure 23: Adequacy of campaigns per cluster



| | Demographics | Purchase Features | Feedback / Promotion Preference | Priority |
|---|---|---|---|---|
| **Cluster 3** | • Young group, highest income | • Spend the most<br>• Frequent and big purchases | • Least likely to complain<br>• Sensitive to the most campaigns, 1 is the best | • Highest |
| **Cluster 1** | • Young group, mid-high income | • Shopped a lot but small purchases | • Complain sometimes<br>• Only attracted by campaign 3 | • High |
| **Cluster 4** | • Oldest group, high income | • Spend much with the most time of purchases and high amount | • Complain sometimes<br>• Only attracted by campaign 3 & 4 | • High |
| **Cluster 2** | • Youngest group, lowest income, lowest education | • Shop a lot but spend the least | • Complain often<br>• Sensitive to the most campaigns, 5 is the best | • Low |
| **Cluster 5** | • Old group, low income | • Spend little with small purchases | • Complain the most<br>• Campaign 4 works best | • Low |

Figure 24: Conclusions drawn per cluster