NOVA IMS Information Management School

Spring Semester 24/25

BIG DATA ANALYTICS

Customer Segmentation for Marketing Campaign Optimization

GROUP 14
Diogo Duarte 20240525
Rodrigo Sardinha 20211627
Rui Luz 20211628

*Understanding customer diversity is key for designing marketing strategies. Our **motivation** is to explore how unsupervised learning techniques can reveal meaningful customer segments based on real behavioral and demographic data, helping optimize campaign targeting and customer engagement.*

GitHub Repository Link

1. Problem Statement

Business Problem

Companies invest heavily in marketing campaigns but often struggle to reach the right customers. Without understanding customer differences, campaigns can lack impact and efficiency. Our project addresses this by identifying distinct customer segments to improve targeting and overall marketing performance.

Project Goal

To segment customers into meaningful groups using clustering techniques. The objective is to support more personalized marketing efforts by revealing differences in customer profiles and behaviors.

Big Data Justification

Customer segmentation in marketing naturally fits within the scope of Big Data. Companies collect large volumes of customer information across multiple channels, combining demographic, behavioral, and transactional data. This data evolves constantly, with new purchases, website visits, and campaign responses recorded daily, requiring segmentation processes that can adapt with high velocity.

Our approach leverages Spark and scalable clustering techniques prepared for these Big Data challenges. Even when applied to a smaller sample, the process is designed to scale to large and fast-changing customer databases, aligning with the Volume and Velocity characteristics of Big Data.

2. Data Collection & Preprocessing

2.1 Data Sources

The data for our project was obtained from Kaggle through the dataset *"Customer Personality Analysis"*, providing detailed customer information across demographics, purchases, and marketing interactions.

2.2 Data Characteristics

The dataset is structured, with each row representing an individual customer and columns containing numerical and categorical attributes. It includes demographic information (age, education, marital status, household composition), purchase amounts across product categories, online interactions (website visits and purchases), and responses to marketing campaigns. This variety of information made it well-suited for exploring customer behavior through clustering.

2.3 Data Cleaning and Preprocessing Steps

Preparing the dataset for clustering required several steps to ensure that the results would be meaningful. We first reviewed all available features and removed columns that were either irrelevant for segmentation or derived from others, to avoid redundancy. For

example, we created a new "Children" variable by combining the "Kidhome" and "Teenhome" columns, which allowed us to represent the number of children in the household more clearly.

After this initial feature engineering step, we carefully selected the set of variables to be included in the clustering. The goal was to focus on features that describe customer behavior and potential for segmentation, while excluding those that would not add value to this type of analysis, such as identifiers or variables with little variability. We paid particular attention to variables reflecting customer activity and engagement, as well as purchasing patterns, as these were expected to provide useful insights when grouping customers into segments.

Continuous variables, such as income, age, recency, and spending amounts, were scaled to ensure that no variable would dominate the distance calculations in clustering. Categorical variables, such as education and marital status, were converted into numerical representations suitable for clustering. Binary variables, including campaign responses and complaints, were also prepared to be used directly. The dataset was then prepared in a consistent format that supported the application of different clustering algorithms. This step was crucial to avoid mixing unrelated variables and to ensure that each clustering result would be interpretable and focused on a coherent aspect of customer behavior.

3. Methodology & Tools

Machine Learning Techniques

3.1 K-Means Clustering

We used K-Means Clustering to divide the dataset into K distinct clusters based on feature similarity, to identify customer segmentation and then take conclusions marketing-related. Choosing the K (number of clusters) was not an obvious choice, therefore we analyzed the Silhouette Score to decide which would be the best K to fit each clustering application. We implemented 3 different Clustering analyses using the K-Means algorithm.

In the first analysis, customers were clustered based on their Income, Age, Total_Spend and NumTotalPurchases where the goal was to find relations between Income and Spending Behavior. After evaluating different values of K using the silhouette score, we chose K=2. The resulting clusters allowed us to conclude that, cluster 0 represents customers with higher income who are significantly higher spenders and purchase more frequently, also this group of customers is, on average, slightly older than the other cluster and shows slightly higher online engagement. The cluster 1, in contrast, represents customers with lower income, much lower total spend and fewer purchases. As can be seen in *Figure 1*. We believe that it would be beneficial to implement marketing strategies targeting Cluster 0 focusing on loyalty programs or premium offers, while Cluster 1 may be more sensitive to price-driven campaigns.

In the second analysis, customers were clustered by product purchase preferences (MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds), combined with demographics (Education, Marital_Status, Age, Children). As observable in *Figure 2*, the clustering revealed three distinct customer segments: Cluster 0: a high-spending, wine-oriented cluster, composed of slightly older customers with higher education levels and more stable family structures. We could apply marketing strategies for this cluster focusing, for example, in loyalty programs for wine and luxury products.

Cluster 1: This is a low-spend cluster, likely to be composed of younger families with more children and possibly more price-sensitive. Focus on buying essential products and less luxury items. The Education distribution is more varied in this cluster (as seen in *Figure 3*), and the Marital_Status shows a higher proportion of Singles. A recommended marketing strategy could be price promotions and family packs. Cluster 2: This cluster represents diverse product interest with high spending across all categories. Education is relatively high but less extreme than in Cluster 0, and Marital_Status shows a higher proportion of Singles (as seen in *Figure 4*). A suitable marketing strategy could be cross-selling opportunities across diverse product categories.

In the last analysis using K-Means, customers were clustered by purchase behavior (Total_Spend and NumTotalPurchases) in relation to promotion acceptance (AcceptedCmp1 toAcceptedCmp5) and online engagement (Website_Engagement). As shown in *Figure 5*, cluster 0 represents a group of very valuable customers, large spenders and responsive to promotions. Marketing campaigns targeted at this group may be around keeping these customers engaged with high-value promotions. The cluster 1 represents a group of low value customers and not responsive to promotions and also low online engagement. This group shouldn't be a priority for expensive campaigns but instead testing low-cost re-engagement campaigns.

The cluster 2 represents a medium-good value customers group, more online active than the last two groups but, on the other hand, not very responsive to promotions. Marketing actions could be by promoting with website-driven offers, loyalty rewards and/or exclusive online discounts. The last cluster (cluster 3) represents a very loyal, high frequency buyer group, also very active online. The company could implement marketing strategies to reinforce this customer loyalty with, for example, rewarding frequent purchase behavior, encouraging online buying by making discounts or other campaigns.

3.2 Gaussian Mixture Models (GMM)

To enhance our clustering analysis, we used Gaussian Mixture Models (GMM) to explore customer segments from three different perspectives: Monetary value, Engagement, and Response to Promotions. Unlike K-Means, which assigns each customer to a single group, GMM performs soft clustering — meaning each customer can belong to multiple segments with different probabilities. This approach allowed us to capture more complex, overlapping behaviours, which are common in marketing settings where customer traits often span across groups.

Before applying GMM, we made the deliberate decision not to remove outliers from the dataset. In customer segmentation, unusual or extreme behaviour is often meaningful — not noise. For example, a small group of high-spending or low-engagement clients might represent VIP targets or niche strategies. GMM naturally accommodates these cases by assigning low-probability points to small, specialized clusters, allowing the model to adapt to unique behaviours without filtering them out. Unlike density-based methods that classify outliers as noise, GMM treats all customers as valid members of some distribution — a key advantage when trying to understand the full diversity of customer profiles.

We began with a monetary segmentation, based on scaled financial features, which resulted in five distinct clusters (*Figure 6*). These groups showed clear differences in overall spending levels and product preferences. For instance, one cluster included customers who, on average, spent the most on wine — and this same group also had the highest average spending on fish products. This relation helps understand consumer habits and may support targeted marketing strategies that promote wine and fish products together.

For the engagement segmentation, we identified three clusters based on customer activity across digital and offline channels (*Figure 7*). One group demonstrated frequent website visits and online purchases, clearly representing highly engaged users. Another showed minimal interaction and appeared largely disconnected from digital channels. A third cluster revealed mixed behaviours, falling somewhere in between.

The final segmentation focused on response to promotions, using variables such as accepted campaigns, deal-seeking behaviour, and complaints. The analysis indicated that four clusters were optimal in this context (*Figure 8*). These included loyal responders, indifferent customers, frequent deal seekers, and a group marked by high recency but low promotional activity.

One of the most relevant insights emerged when profiling the third cluster: it included the highest proportion of customers who had submitted complaints in the past two years. On closer inspection, we found that this same group also had the lowest average education level and household income (*Figure 9 and 10*). This adds an important socioeconomic dimension to their behavioural profile, reinforcing the value of multi-perspective segmentation.

3.3 Self-Organizing Maps (SOM)

Overview

Self-Organizing Maps (SOM) were used as an exploratory complement to our clustering analyses, providing a visual representation of customer behavior patterns. This helped us identify relationships and distributions that might not be as easily captured through traditional clustering methods like K-Means or GMM.

Implementation

We implemented three SOM analyses, each focusing on a distinct perspective of customer behavior: customer profile, product consumption, and campaign response. This division

helped avoid mixing variables with very different meanings and ensured clearer interpretation. Each SOM was analyzed using a U-Matrix (to visualize cluster boundaries), Component Planes (to assess variable influence), and Hit Maps (to observe customer distribution across the map).

Insights

Customer Profile Perspective: The SOM revealed clear differences in spending and engagement patterns across the customer base. The U-Matrix (*Figure 11*) showed well-separated regions, indicating distinct customer groups. Component Planes highlighted that higher income and total spend align in certain clusters, while online engagement and website visits vary independently. The Hit Map (*Figure 12*) confirmed a good distribution of customers, suggesting that the SOM captured meaningful variability in profiles.

Product Preferences Perspective: The SOM showed that product consumption behaviors vary significantly among customers. The U-Matrix indicated distinct segments, and Component Planes (*Figure 13*) revealed that preferences for wines, meats, and sweets are not uniformly distributed. Some customers show consistently low spending across categories, while others exhibit clear preferences. The Hit Map highlighted a common product profile among most customers, with smaller groups showing distinct patterns.

Campaign Response Perspective: The SOM showed that most customers exhibit similar behaviors regarding marketing campaigns, with a single dense cluster of non-responders. A smaller number of engaged customers were mapped to isolated areas. Component Planes confirmed that responses to individual campaigns were concentrated in specific segments, reflecting the selective nature of campaign responsiveness.

Overall, Self-Organizing Maps provided a complementary view of the customer data, helping to explore behavioral diversity across perspectives. Though not used to define final clusters, they highlighted patterns and relationships that complemented insights from the other clustering methods.

4. Conclusion

Across the different clustering approaches explored in this project, we were able to identify several actionable customer segments with clear relevance for marketing decisions. K-Means provided well-defined groups that can support targeted strategies, such as loyalty programs for high-value customers or price-driven campaigns for more sensitive segments. Gaussian Mixture Models helped capture overlapping behaviors, revealing more subtle patterns such as cross-category buying habits and the relationship between customer complaints and socioeconomic factors. Self-Organizing Maps complemented these insights by providing a visual understanding of how different customer behaviors are distributed, helping to confirm and further explore patterns identified through clustering. Together, these methods contributed to a richer and more nuanced segmentation of the customer base, offering valuable support for future marketing strategies.

ANNEX

- **This is just a gentle reminder that the original member of our group, Filipe Ferreira (20240741), is no longer part of our group and therefore should not be considered in the evaluation of the project.**

	prediction	avg_Income	avg_Age	avg_Total_Spend	avg_NumTotalPurchases	avg_Recency	avg_Website_Engagement
0	0	70,398.76	47.58	1,135.84	21.50	49.31	9.93
1	1	35,990.08	43.03	133.49	8.95	48.75	8.93

Figure 1

	prediction	avg_MntWines	avg_MntFruits	avg_MntMeatProducts	avg_MntFishProducts	avg_MntSweetProducts	avg_MntGoldProds	avg_Age	avg_Children
0	0	652.89	31.58	292.89	44.63	30.12	73.73	47.46	0.66
1	1	94.99	6.44	35.31	9.07	6.41	18.62	44.14	1.24
2	2	521.02	88.23	439.74	127.12	94.51	88.20	45.41	0.34

Figure 2

	prediction	Education	count	total	percentage
12	0	Graduation	258	552	46.74
2	0	PhD	163	552	29.53
3	0	Master	96	552	17.39
4	0	2n Cycle	34	552	6.16
5	0	Basic	1	552	0.18
13	1	Graduation	620	1294	47.91
7	1	PhD	273	1294	21.10
8	1	Master	226	1294	17.47
0	1	2n Cycle	123	1294	9.51
14	1	Basic	52	1294	4.02
10	2	Graduation	238	370	64.32
6	2	PhD	45	370	12.16
9	2	Master	43	370	11.62
11	2	2n Cycle	43	370	11.62
1	2	Basic	1	370	0.27

Figure 3

	prediction	Marital_Status	count	total	percentage
3	0	Married	212	552	38.41
12	0	Together	147	552	26.63
5	0	Single	108	552	19.57
15	0	Divorced	66	552	11.96
10	0	Widow	19	552	3.44
0	1	Married	510	1294	39.41
16	1	Together	334	1294	25.81
13	1	Single	276	1294	21.33
14	1	Divorced	131	1294	10.12
2	1	Widow	38	1294	2.94
7	1	Alone	3	1294	0.23
4	1	YOLO	2	1294	0.15
11	2	Married	135	370	36.49
1	2	Together	92	370	24.86
9	2	Single	87	370	23.51
6	2	Divorced	35	370	9.46
17	2	Widow	19	370	5.14
8	2	Absurd	2	370	0.54

Figure 4

	prediction	avg_Total_Spend	avg_NumTotalPurchases	avg_AcceptedCmp1	avg_AcceptedCmp2	avg_AcceptedCmp3	avg_AcceptedCmp4	avg_AcceptedCmp5	avg_Website_Engagement
0	0	1,589.89	20.39	0.22	0.05	0.09	0.19	0.31	8.52
1	1	80.11	7.29	0.00	0.00	0.07	0.01	0.00	8.26
2	2	531.43	17.02	0.04	0.01	0.06	0.09	0.02	10.49
3	3	967.80	25.65	0.08	0.01	0.08	0.09	0.06	11.96

Figure 5

	Income	Total_Spend	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
1	188901.7	730.4	21.9	3.9	670.3	3.6	27.3	3.4
2	34801.1	36.8	21.9	0.6	8.7	1.1	0.5	4.1
3	58342.0	777.0	394.1	33.7	211.4	48.1	34.5	55.3
4	22860.7	46.0	7.9	4.4	11.3	6.6	4.7	11.1

Figure 6

	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	Website_Engagement	AcceptedCampaigns_Total	Recency
1	2.2	5.1	4.1	6.6	5.0	10.1	1.6	44.1
2	7.6	6.3	9.0	0.2	8.5	14.8	0.1	47.9
3	2.3	3.7	2.1	5.5	5.4	9.1	0.0	50.9

Figure 7

	AcceptedCampaigns_Total	NumDealsPurchases	Complain	Recency
1	2.2	1.2	0.0	53.3
2	0.6	10.0	0.0	35.4
3	0.3	2.3	1.0	53.0
4	0.2	2.3	0.0	48.6

Figure 8

	2n Cycle	Basic	Graduation	Master	PhD
1	5.9%	0.7%	50.9%	15.6%	27.0%
2	4.4%		40.0%	24.4%	31.1%
3	19.0%		66.7%	9.5%	4.8%
4	9.5%	2.8%	50.3%	16.5%	20.8%

Figure 9

	avg_Age	avg_Income	avg_Kidhome	avg_Teenhome	avg_Children
1	44.9	70271.3	0.2	0.2	0.4
2	49	55131.3	0.8	1	1.8
3	48.9	45242.3	0.7	0.5	1.2
4	45.1	49457.6	0.5	0.5	1

Figure 10

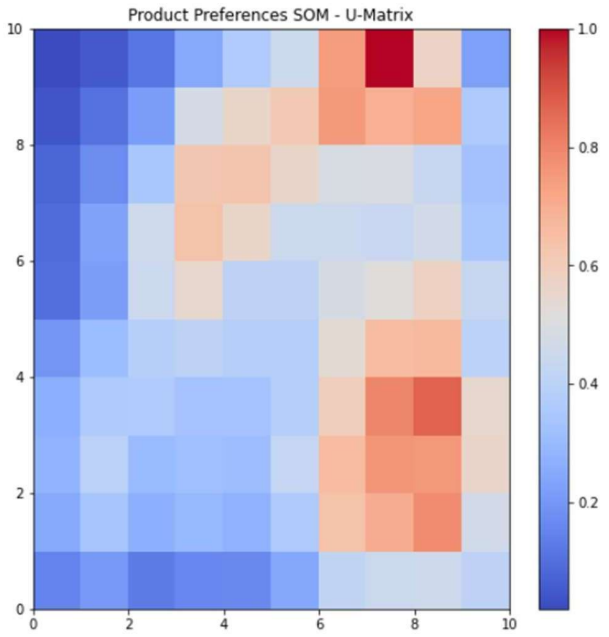

Figure 11

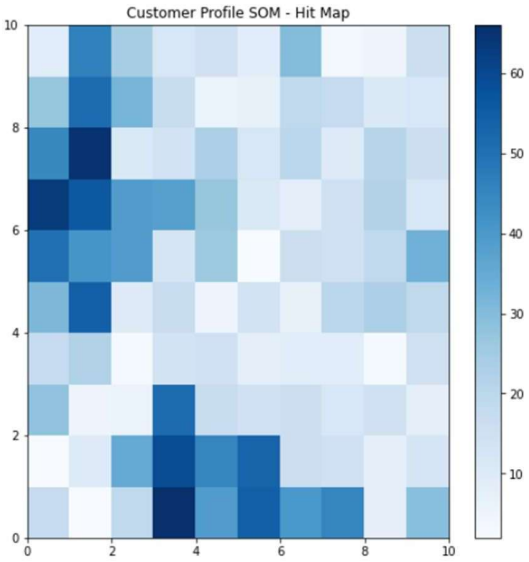

Figure 12

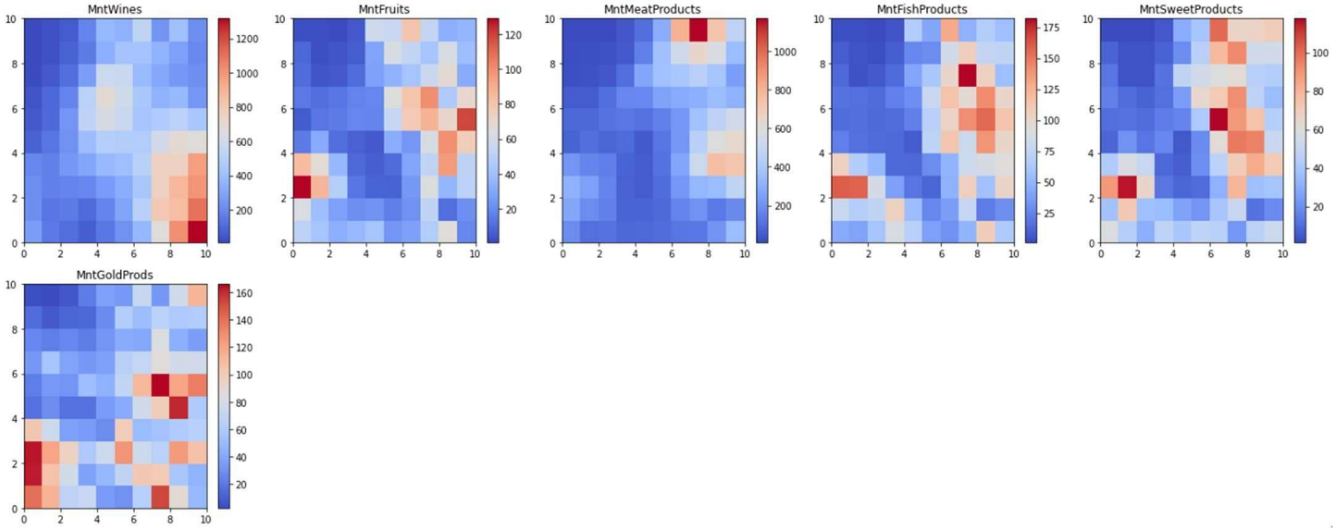

Figure 13