Report: Marketing Analysis

Alessio Baldini
Antonella Mele
Sofia-Zoi Sotiriou
Florian Derchain

Analysis: Profiling Customers

1 The dataset

Objective:
The aim of our analysis is to optimize marketing strategies by identifying customer segments based on preferences and personal characteristics. This approach seeks to categorize each client into ideal clusters aligned with company products, enabling more precise targeting and efficient budget allocation in marketing efforts.

Approach and Dataset:
Using a dataset of over 2,000 customer profiles, we analyze correlations between static characteristics (demographics, income, etc.) and purchasing behavior. This analysis involves four key segments:

1. People:
    - ID: Customer's unique identifier
    - Year_Birth: Customer's birth year
    - Education: Customer's education level
    - Marital_Status: Customer's marital status
    - Income: Customer's yearly household income
    - Kidhome: Number of children in customer's household
    - Teenhome: Number of teenagers in customer's household
    - Dt_Customer: Date of customer's enrollment with the company
    - Recency: Number of days since customer's last purchase
    - Complain: 1 if the customer complained in the last 2 years, 0 otherwise
2. Products
    - MntWines: Amount spent on wine in last 2 years
    - MntFruits: Amount spent on fruits in last 2 years
    - MntMeatProducts: Amount spent on meat in last 2 years
    - MntFishProducts: Amount spent on fish in last 2 years
    - MntSweetProducts: Amount spent on sweets in last 2 years
    - MntGoldProds: Amount spent on gold in last 2 years
3. Promotion
    - NumDealsPurchases: Number of purchases made with a discount

- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

4. Place

   - NumWebPurchases: Number of purchases made through the company's website
   - NumCatalogPurchases: Number of purchases made using a catalogue
   - NumStorePurchases: Number of purchases made directly in stores
   - NumWebVisitsMonth: Number of visits to company's website in the last month

Goal:
Using a range of analytical techniques, including clustering, correlation analysis, and PCA, we aim to create customer segments with distinctive attributes tied to purchase behaviors, demographic factors, and marketing responses.

Customer Personality Analysis is all about getting to know a company's ideal customers on a deeper level. By understanding who they are, what they like, and how they behave, businesses can better tailor their products to fit the unique needs of different customer groups.

In this project, we dive into this analysis, helping businesses see beyond the numbers. With these insights, companies can adjust their products and marketing strategies to better connect with specific customer segments. For example, instead of trying to sell a new product to everyone, the company can focus on the group most likely to love it, making their efforts more targeted and effective.

2 First exploration

After having charged all the libraries, we charged the dataset chosen called 'marketing.csv' and we made a first exploration using essential functions such as: head, info, describe to get a general view of the subject. We handled null samples and duplicated detecting and deleting them.

Results:
- The dataset consists of 2240 observations across 29 columns.
- There were 24 missing value in the 'income' column, there are no duplicates.
- Most columns are numerical, the categorical are 'marital_status', 'education', 'Dt_customer'.

3 Data Cleaning and Feature Adjusting

We have removed unnecessary variables such as 'ID', 'Z_ConstContact' and 'Z_Revenue' because they do not impact on the analysis.
As 'Dt_Customer' is a date and not a categorical variable, we decided to calculate the number of days that each customer has been with the company as a numerical variable called Days_is_client.

After we examined the categorical variables, we decided to group some categories to simplify the dataset. We opted to group the variable 'Marital_Status' in Partner and Single:

- Partner includes: Married and Together
- Single includes: Single, Divorced, Widow, Alone, Absurd and YOLO.

We manipulated Education too, regrouping it into Postgraduate, Graduate and Undergraduate:

- Postgraduate includes: PhD and Master
- Graduate includes: 2n Cycle and Graduate
- Undergraduate includes: Basic

Other considerations we made are about sum variables under a bigger one.

- Kids is the sum of the variables Kidhome and Teenhome.
- Expenses is the sum of MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds.
- TotalAcceptedCmp is the sum of AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5
- TotalNumPurchases is the sum of NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumDealsPurchases.

About outliers: we detected outliers applying the method of the Z-score on the numerical variables. This step is really important, since individuals with very high incomes have consumption habits that are generally very different and not comparable to those of the rest of the observed population.

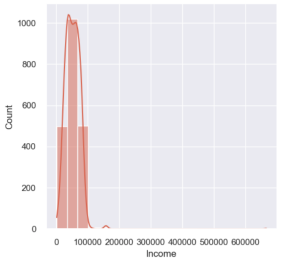
What is Z score? The Z-score tells how many standard deviations away a data point is far from the mean. The process of transforming a feature to its Z-score is the 'Standardization'. The formula is:

$$Z - score = \frac{X - \mu}{standar\_dev} \tag{1}$$

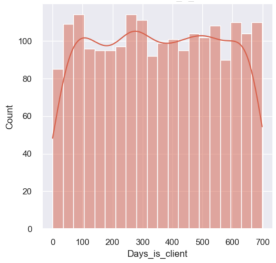
If the Z-score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Z-score can be both positive and negative. The farther away from 0, higher the chance of a given data point being an outlier. Typically, Z-score greater than 3 is considered extreme.
And to conclude we considered just the significative variable: Education, Marital_Status, Income, Kids, Days_is_client, Recency, Expenses, TotalNumPurchases, TotalAcceptedCmp, Complain, Response.

4 Exploratory Data Analysis

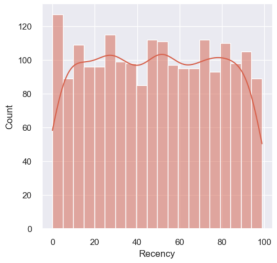
We started the exploration plotting some Histograms of the distribution of the following variables: Income, Days_is_client, Recency, Expenses, TotalNumPurchases.



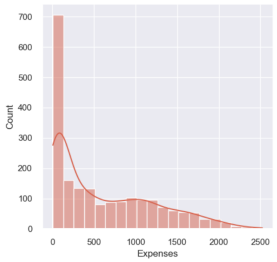
(a) Income



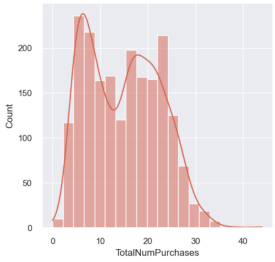
(b) Day_is_client



(c) Recency



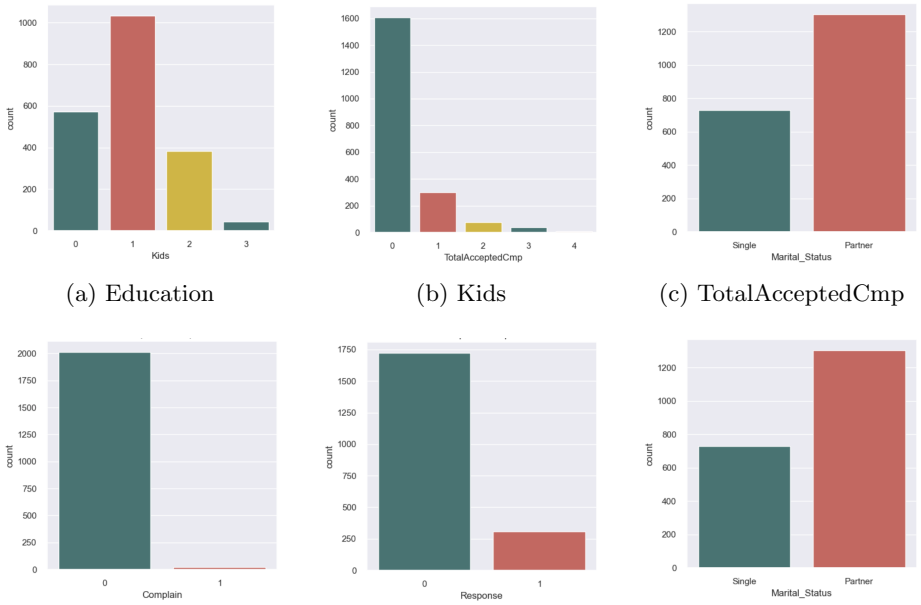
(d) Expenses



(e) TotalNumPurchases

Figure 1: Distributions

In the next step we made the barplot of Education, Kids, TotalAcceptedCmp, Marital_Status, Complain and Response

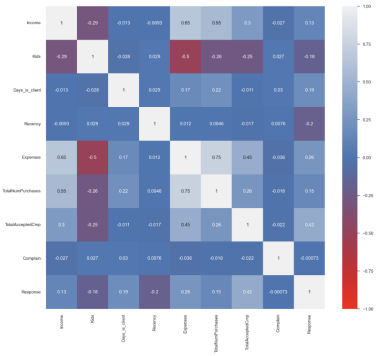


(a) Education (b) Kids (c) TotalAcceptedCmp

(d) Marital_Status (e) Complain (f) Response

Figure 2: Barplots

And to have a comprehension about the correlation between variables we generated the correlation matrix.



Results:

- The set of histograms displays the distributions of numerical variables like Income, Expenses, and TotalNumPurchases. For example, Income is right-skewed with most values below 100,000, while Expenses peak around 500.

- Categorical count plots (e.g., Education, Kids, Response) reveal imbalances. For instance, most customers have 0 kids, and over 1,200 have graduate education, highlighting dominant groups.
- The heatmap of correlations indicates strong relationships between Expenses and Income (0.65) and between Expenses and TotalNumPurchases (0.75), while variables like Kids and Income show a negative correlation (-0.29).
- Minimal correlation is seen for variables such as Recency and Days_is_client (near 0), indicating independence.
- These plots effectively summarize key trends and relationships in the data, with some distributions and correlations guiding potential predictive features.

5 Data Pre-Processing

Before performing PCA, we applied pre-processing to the dataset to prepare variables for dimensionality reduction. First, categorical variables such as "Education" and "Marital_Status" were encoded using a 'LabelEncoder' to convert their values into numerical format. We also removed certain columns (''TotalAcceptedCmp'', ''Complain'', and ''Response'') that were unnecessary for the analysis that is about selling products to specific targets. Following this, the dataset was standardized using 'StandardScaler' to scale all features (mean = 0 and standard deviation = 1), ensuring that variables with larger ranges did not dominate the PCA results.

6 PCA and 3D visualization

The we applied the Principal Component Analysis (only to numerical variables) to reduce the number of features to 3.

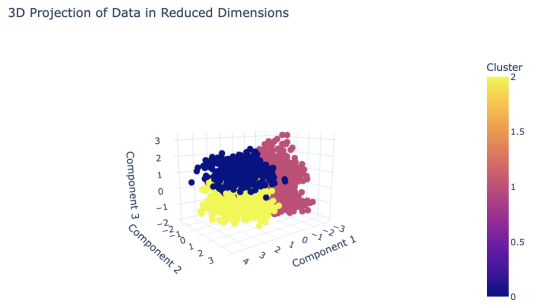
The PCA reduced the dataset to three components.

Principal Component	**Standard Devitation 1**	**Standard Deviation 2**
Component 1	1.67	2.53
Component 2	1.02	1.41
Component 3	1.00	1.09

Table 1: Standard Deviation of the first three principal components

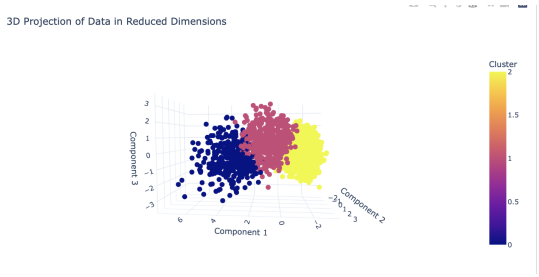
The dimensionality reduction simplifies the data, retaining most of the information for clustering. The decreasing variance across components suggests an effective compression of key features into fewer dimensions.

To analyze the structure of the data, we applied k-means clustering on the scaled dataset, specifying 3 clusters as an initial assumption. The clustering results were added to the PCA-reduced dataset to visualize how the data is grouped in the reduced dimensional space.



The 3D scatter plot shows the clustering results in the PCA-reduced space, with each axis representing a principal component. Data points are grouped into three clusters, which exhibit moderate separation with some overlap, indicating that the clusters are distinct but not perfectly isolated. Cluster 0 (pink) appears compact, while Clusters 1 (blue) and 2 (yellow) are more dispersed, suggesting higher internal variability. The clear spread along all three axes highlights the relevance of the selected components in capturing variance.

We have also applied the PCA to the same dataset but with Expenses not assembled only under the same variable and the result was the following.



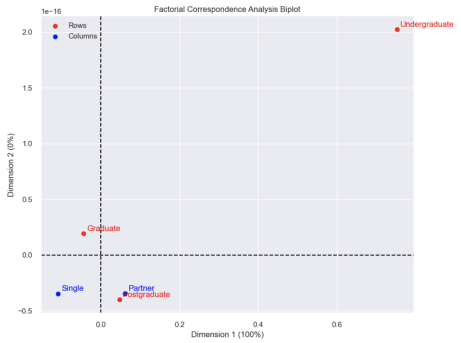
The difference lies in how expense variables are handled. When expenses are combined into one variable, the PCA captures overall spending behavior, resulting in more compact and distinct clusters. Separating expenses into multiple variables increases dimensionality and variance, causing clusters to appear more dispersed. Aggregation simplifies patterns, while separation highlights finer details but reduces cluster clarity.

We decided to proceed with the analysis using expenses assembled, as this approach provides better-defined customer profiles and more easily interpretable results.

7 FCA

We implemented the Factorial Correspondence Analysis (FCA) to explore and visualize the relationship between categorical variables in our dataset. This method helps reduce the complexity of the contingency table, making it easier to identify patterns or associations between the categories of Marital_Status and Education.

	Partner	Single
Graduate	218	135
Postgraduate	162	83
Undergraduate	5	0



Contingency Table and FCA Biplot

In this biplot we can observe how the categories of Marital_Status (blue points) and Education (red points) are related. The first dimension (Dimension 1) explains 100% of the variance, with categories such as Undergraduate and Graduate positioned distinctly along this axis. The Single and Partner categories from Marital_Status are placed closer together on the negative side of Dimension 1, while Postgraduate appears near Graduate on the positive side. The biplot shows that individuals with higher education levels (e.g., Graduate and Postgraduate) tend to be associated with different marital statuses compared to those with lower education levels (e.g., Undergraduate), indicating a possible relationship between education and marital status in this dataset.

8 Clustering

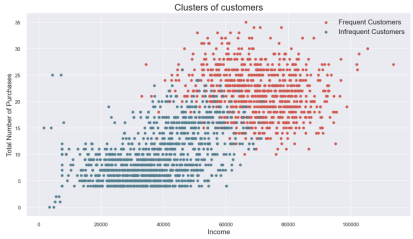
How does K-Mean Clustering work?

- Initialization: The algorithm starts by randomly selecting a predefined number of cluster centers (centroids)
- Assignment: Each data point is assigned to the nearest centroid, forming initial clusters
- Update: The centroids are recalculated as the average position of all points within each cluster
- Iteration: The process of assignment and updating continues until the centroids stabilize and no longer change significantly

- Final output: The result is a set of clusters where data points in the same cluster are more similar to each other than to those in other clusters

Before applying the KMeans clustering algorithm, we first performed some pre-processing on the dataset. We identified the categorical columns and we dropped them off. Then, we scaled the encoded data using the StandardScaler to standardize the features, ensuring they all have similar ranges and distributions. This scaling helps prevent any single feature from dominating the clustering process. Finally, we prepared the dataset by transforming the data back to its original scale for visualization after performing the clustering.

(a) Graph Income-Expenses



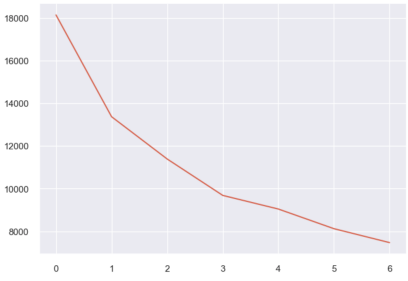
(b) Graph Income-TotNumbPurchases

The first plot shows two clusters: high-value customers with higher income and expenses, and low-value customers with lower values. This segmentation highlights distinct spending behaviors.

The second plot clusters customers by income and number of purchases. Red represents frequent, higher-income buyers, while blue shows infrequent, lower-income ones. This helps target strategies for different customer groups.

8.1 The Elbow Curve and Silhouette Analysis

The elbow curve is a graphical method used to determine the optimal number of clusters in clustering algorithms like k-means. It plots the Sum of Squared Distances (SSD) for different cluster counts. The "elbow" point, where the rate of decrease sharply changes, indicates the best k, balancing simplicity and accuracy.



The Silhouette analysis evaluates the quality of clustering by measuring how

close each data point is to points within its own cluster compared to points in other clusters. A silhouette score (ranging from -1 to 1) is calculated for each point: values close to 1 indicate good clustering, near 0 suggest cluster overlap, and negative values mean incorrect assignment. It helps identify the optimal number of clusters by comparing average scores.

n_clusters	**Silhouette Score**
2	0.2041
3	0.2241
4	0.1568
5	0.1773
6	0.1710
7	0.1733
8	0.1705
9	0.1775
10	0.1842

Table 2: Silhouette scores for different values of *n_clusters*

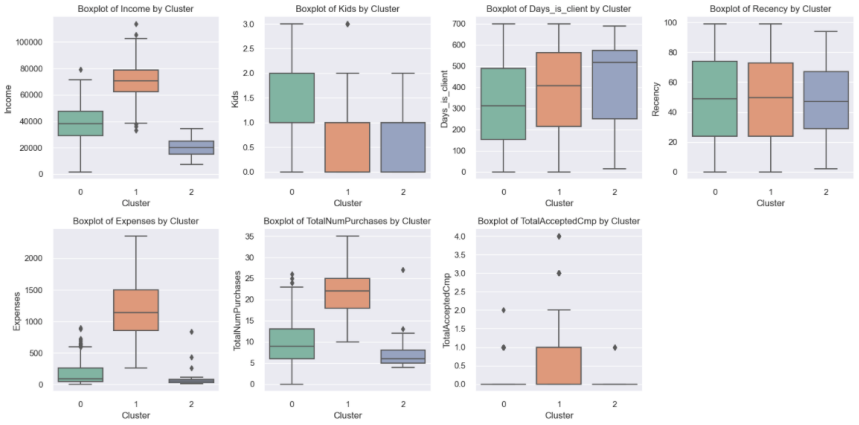
Considerations: The Elbow Curve and the Silhouette Analysis both suggest that 3 clusters provide the best solution.

9 K Mean with the chosen number of clusters

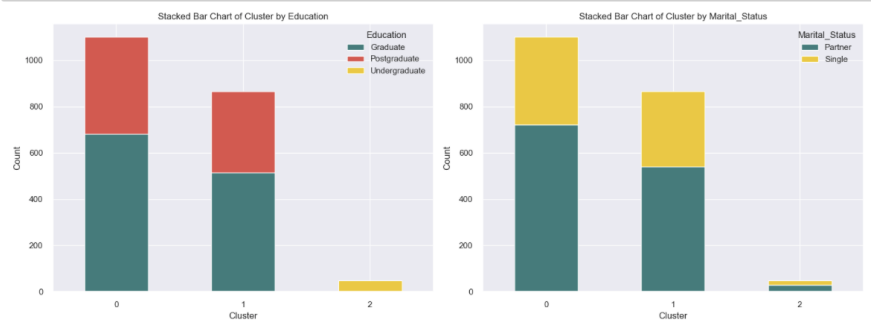
This scatter plot illustrates customer segmentation by income (x-axis) and expenses (y-axis). Three clusters are visible: Cluster 2 (yellow) represents low-income, low-expense customers; Cluster 0 (green) indicates moderate income and expense levels; Cluster 1 (red) consists of high-income, high-expense customers. The clear separation reflects distinct spending and earning patterns across groups.

10 Profiling Customers

After data analysis, customer profiling involves segmenting customers into homogeneous groups based on common characteristics such as demographics, purchasing behaviors, and preferences. These profiles help personalize marketing strategies, improve customer satisfaction, and optimize offers.



These boxplots compare key features (e.g., income, expenses, purchases, etc.) across clusters. Cluster 1 has the highest income, expenses, and purchases, while Cluster 2 has the lowest values for most metrics. Cluster 0 shows intermediate characteristics. The boxplots effectively highlight variability and outliers within and between clusters.



These charts analyze clusters by education level and marital status. Cluster 0 contains a higher proportion of postgraduates and graduate and partnered individuals, Cluster 1 contains similarly to Cluster 0, while Clusters 2 include more undergraduates. This breakdown offers deeper demographic insights for each cluster.

Feature	Cluster 0	Cluster 1	Cluster 2
Income	Moderate	Highest	Lowest
Education	Graduate and PostG.	Graduate and PostG.	Mostly undergraduates
Marital Status	Mostly partnered	Mostly partnered	Partnered and Single
Expenses & Purchases	Moderate spending & purchases	High spending & purchases	Low spending & purchases
Promotions	Occasionally responsive	Most responsive	Least responsive

Table 3: Compact Cluster Analysis Summary

To sum up Cluster 0 represents the Mid Value Customers, Cluster 1 the High Value Customers and Cluster 2 the Low Value Customers.