CUSTOMER SEGMENTATION

Inés Avello Solís
Data Science @The Bridge
December 2nd, 2022

Content

Introduction

Customer segmentation allows companies to separate their set of customers into different groups or clusters that share traits. Once the clusters are created, the company can tailor their marketing campaigns to the specific groups, improving the campaign's impact or conversion rate.

Dataset

Name: Company's Ideal Customers | Marketing Strategy
Source: Kaggle
URL:
https://www.kaggle.com/datasets/whenamancodes/customer-personality-analysis
Published by: Aman Chauhan

Attributes in dataset:
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

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

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

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

Preliminar feature generation and reduction

The dataset consisted of 2240 rows and 29 columns.

Created features:

- Age: User age, to replace user Year_Birth
- Purchases: the addition of NumDealsPurchases, NumWebPurchases, NumCatalogPurchases and NumStorePurchases.
- AcceptedCmps: the total number of campaigns accepted by the user. Sum of AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5.
- MntProducts: the total amount spent on products
- Kids_Home: Number of kids in the household
- Family_Size: Number of people in the family

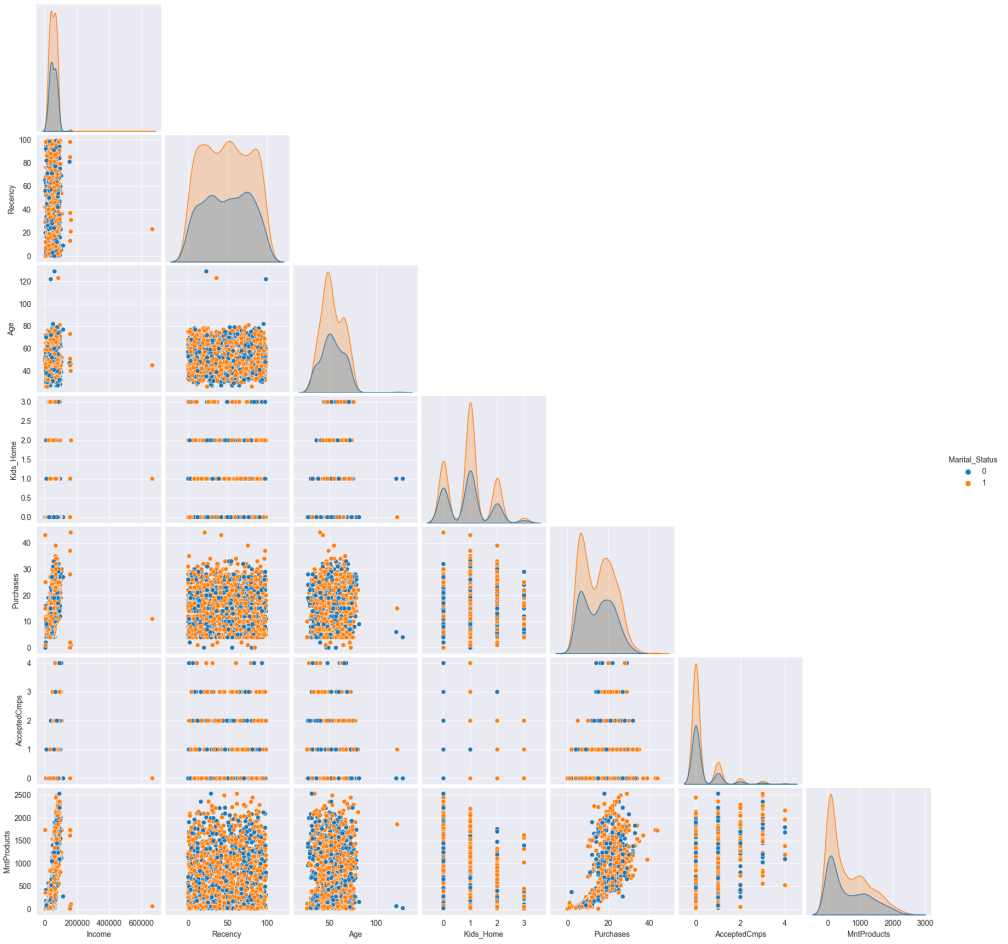
Eliminated features:

- ID, Year_Birth, Complain, Z_CostContact, Z_Revenue, and Response: these features were unique identifiers or had no explanation from the creator of the dataset.

Adjusted features:

- The categorical features Marital_Status and Education were encoded.
- Education: Basic = 0, Graduation = 1, 2n Cycle = 2, Master: 2, PhD = 3
- Marital status: Single = 0, Together = 1, Married = 1, Divorced = 0, Widow = 0, Alone = 0, Absurd = 0, YOLO = 0

The following correlation matrix between the features was obtained using Phik's correlation coefficient.



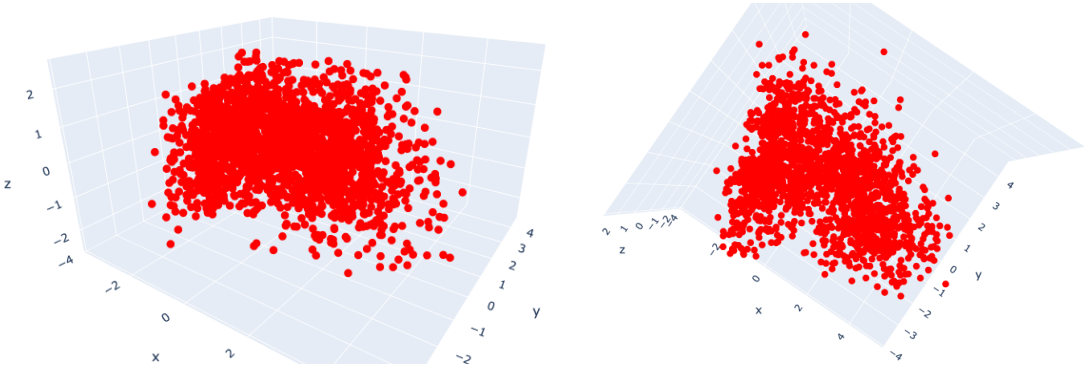
Finally, the features chosen for the analysis were: Education, Income, Age, Kids_Home, Purchases, AcceptedCmps, MntProducts, Family_Size.

Data cleaning

The rows with null values in 'Income' were dropped (2 rows). Moreover, the dataset was scaled and a DBScan (epsilon=2) model was applied to identify the outliers. A total of 37 outliers were eliminated, adjusting the dataset from 2216 rows to 2179.

Data visualization and dimensionality reduction

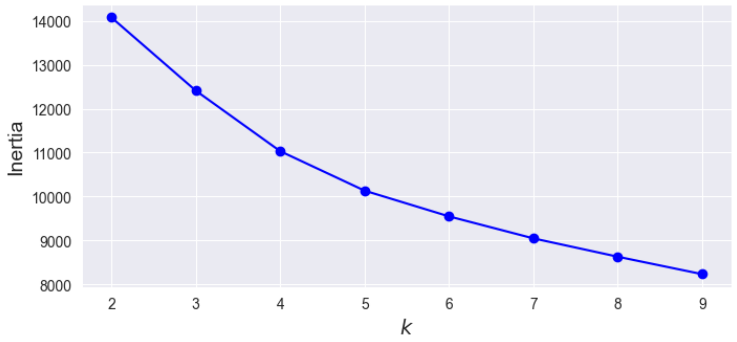
A Principal Component Analysis (PCA) was carried out to visualize the data. The dimensions were reduced to 3, obtaining an explained variance of 65.3%.

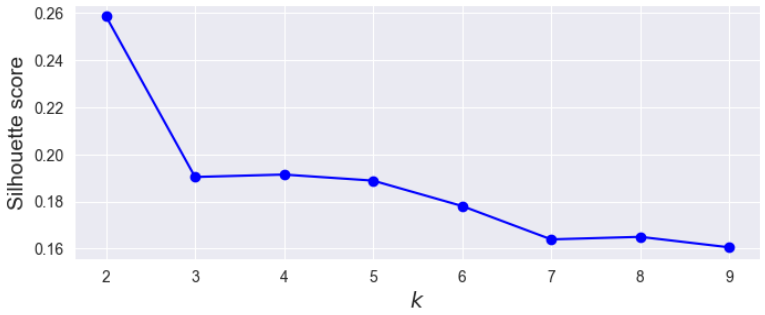


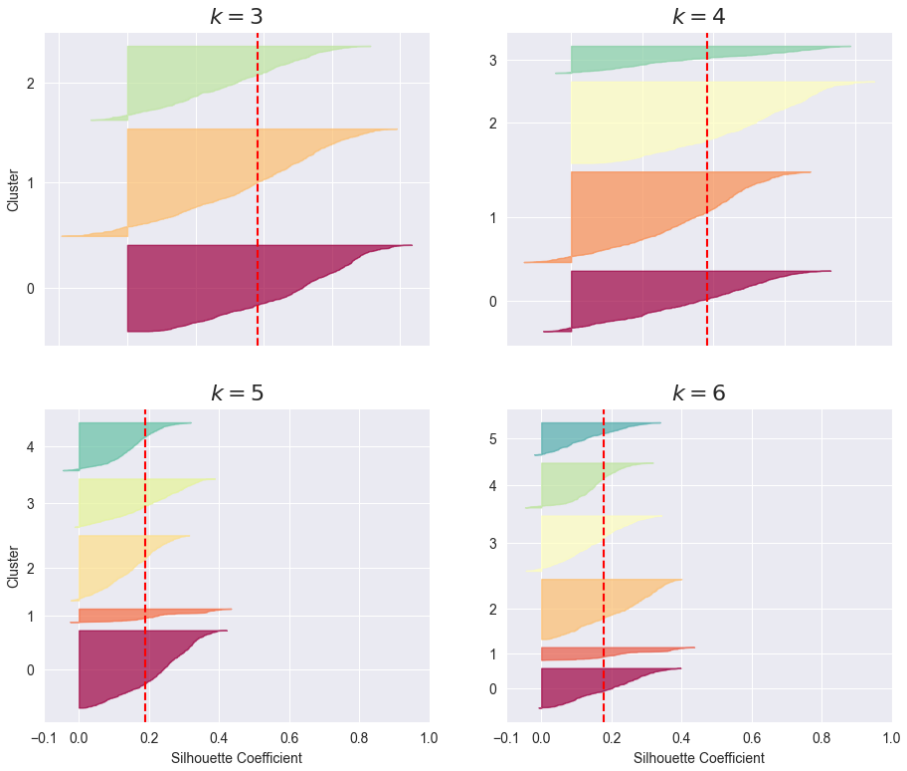
Data clustering

1. KMeans

The data was preprocessed with StandardScaler and several KMeans models were applied.



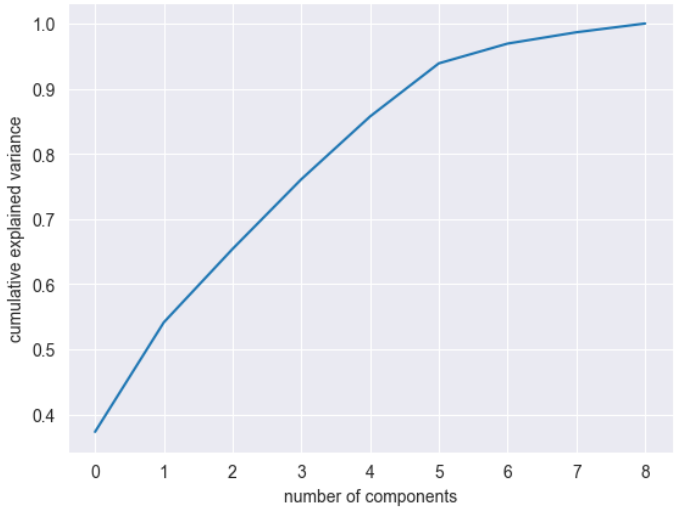




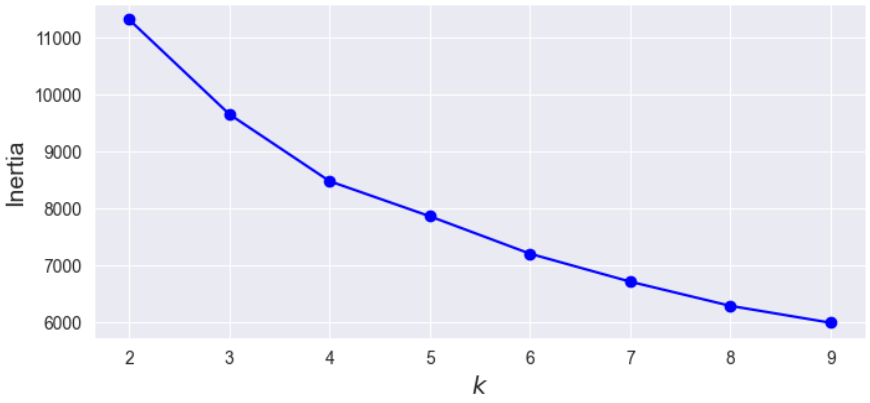
Considering that it is desirable to have low inertia and a silhouette score close to 1, as well as balanced clusters, the KMeans model with 3 clusters would be the better choice.

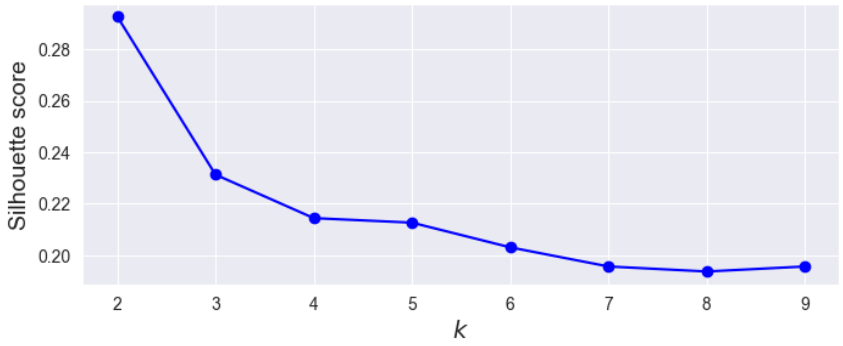
2. PCA and KMeans

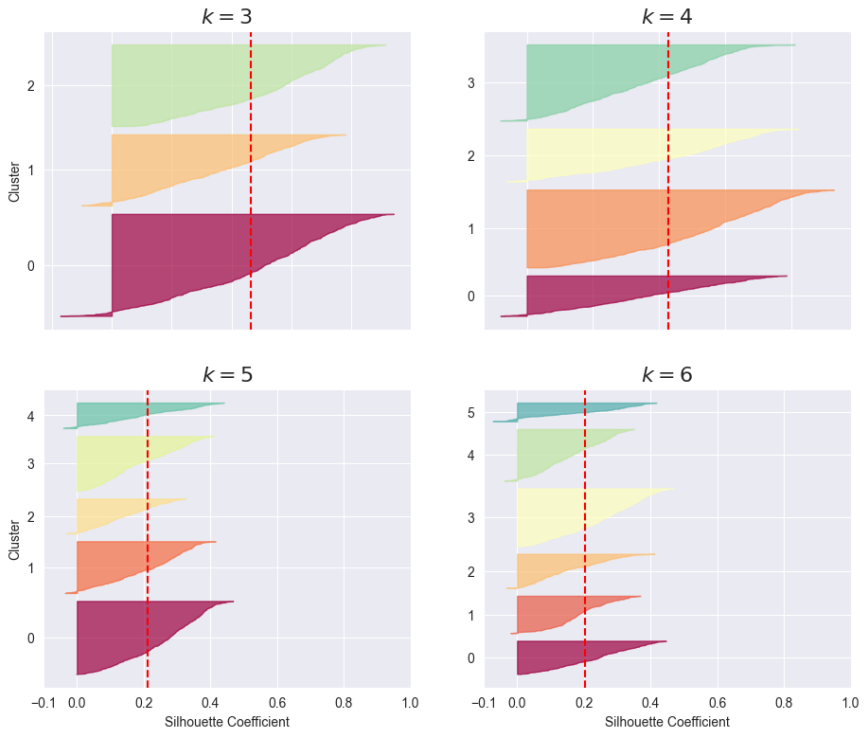
Carrying out a Principal Component Analysis, it was possible to reduce the dimensions of the dataset to 5 Principal Components while maintaining a 93.88% cumulative explained variance from the previous dataset.



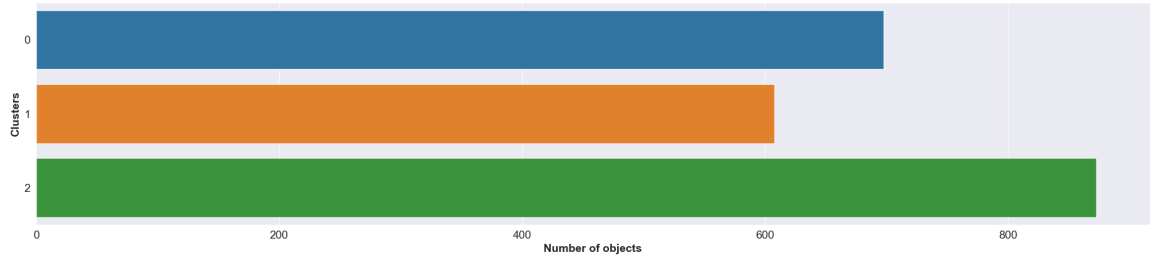
The different KMeans models were applied, obtaining new inertias and silhouette scores.



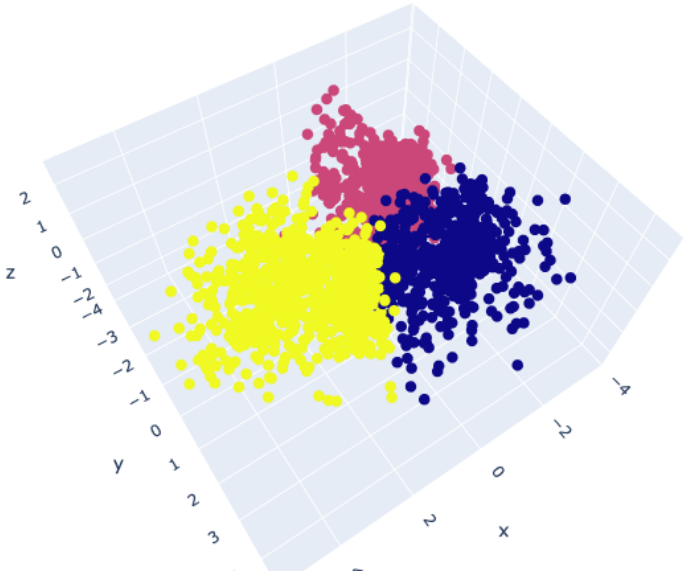




Combining both PCA and KMeans, the inertias decreased, and the silhouette scores increased. The KMeans model with 3 clusters was kept due to its inertia, silhouette score and balanced number of observations.



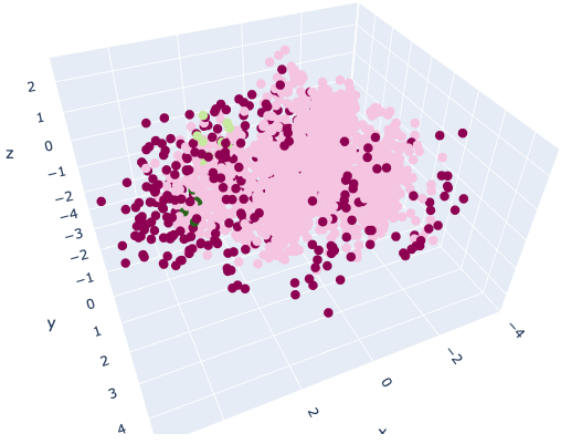
To visualize the clusters, the 3 principal components in the PCA model were used, find below the 3D scatterplot:



3. PCA and DBScan

DBScan was also implemented. An epsilon of 1.15 and minimum number of samples of 15 resulted in 3 main clusters and a 4th cluster for the outliers. This model was considered unsuccessful due to the unbalance in the number of observations per cluster.

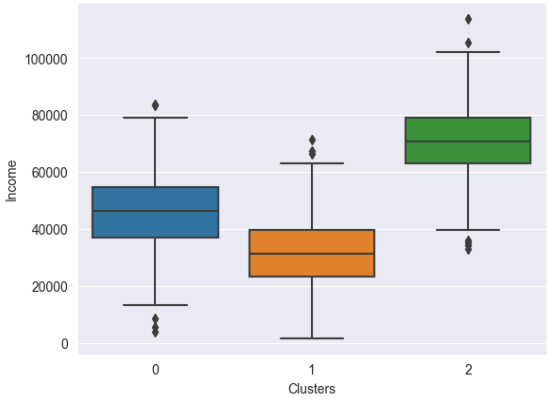
Clusters	P1	P2	P3	P4	P5
-1	348	348	348	348	348
0	1814	1814	1814	1814	1814
1	10	10	10	10	10
2	7	7	7	7	7

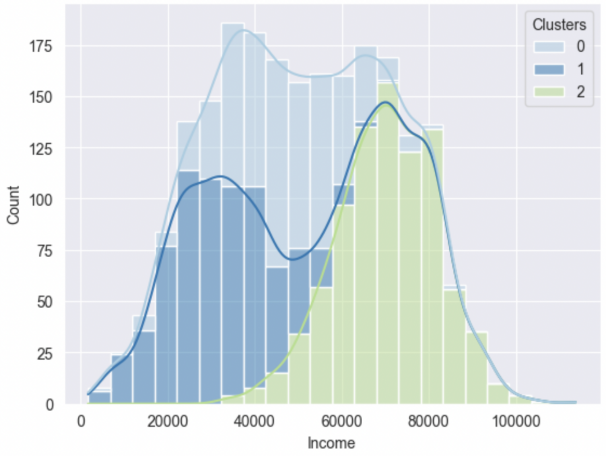


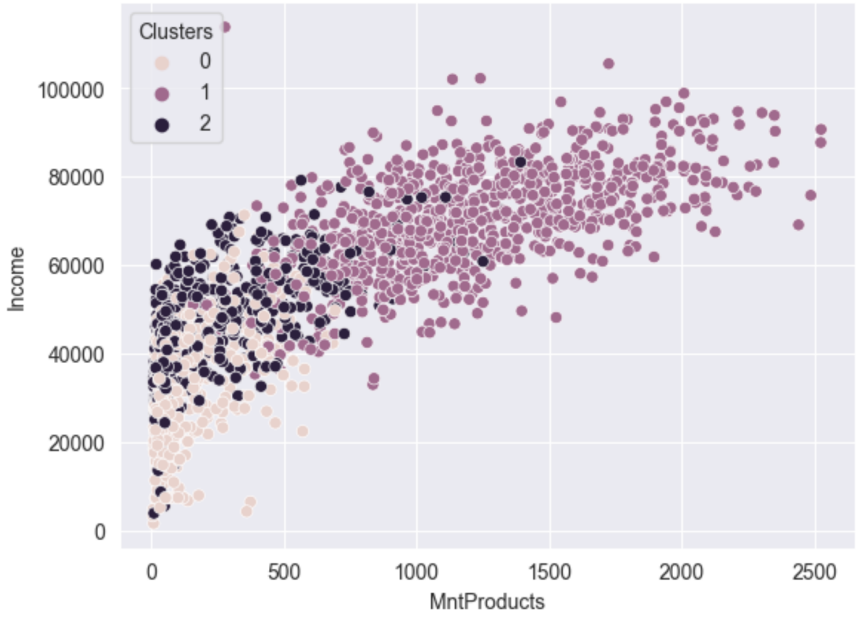
Cluster Analysis

The following analysis applies to the clusters obtained with PCA (number of components = 5) and KMeans (number of clusters = 3).

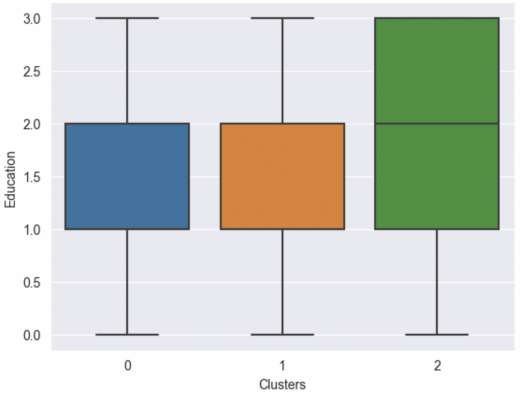
1. Income

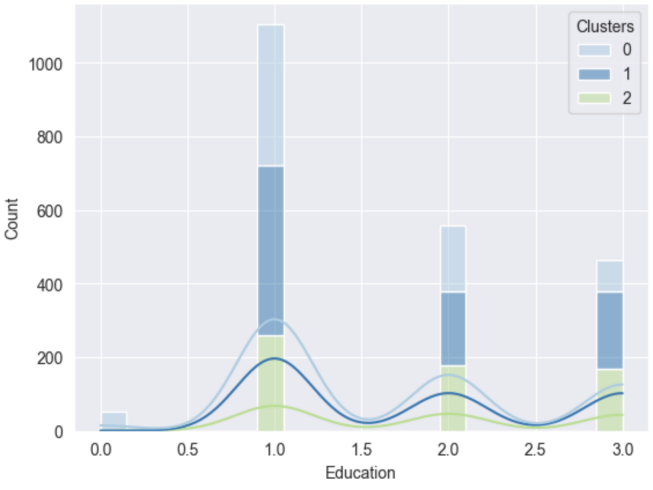




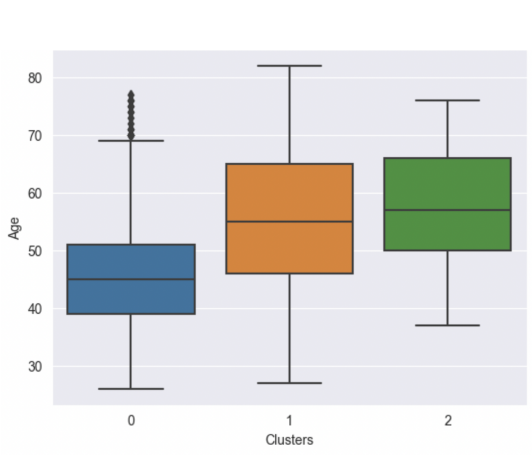


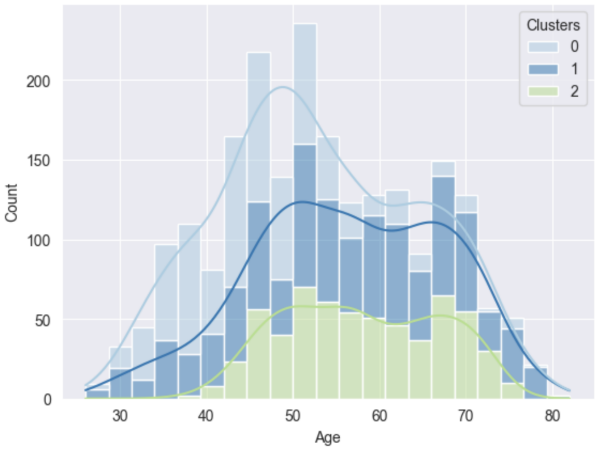
2. Education



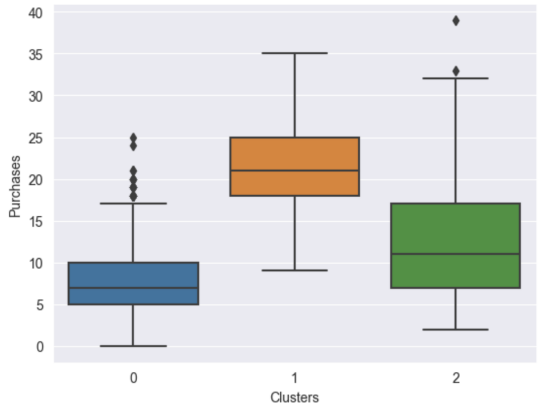


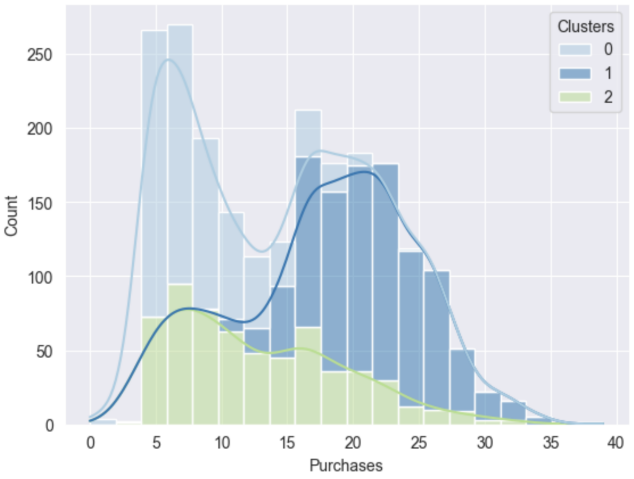
3. Age



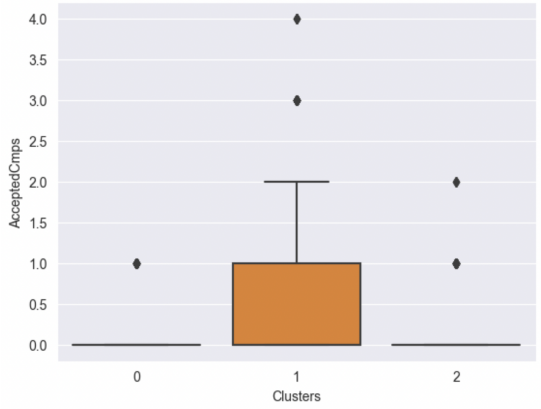


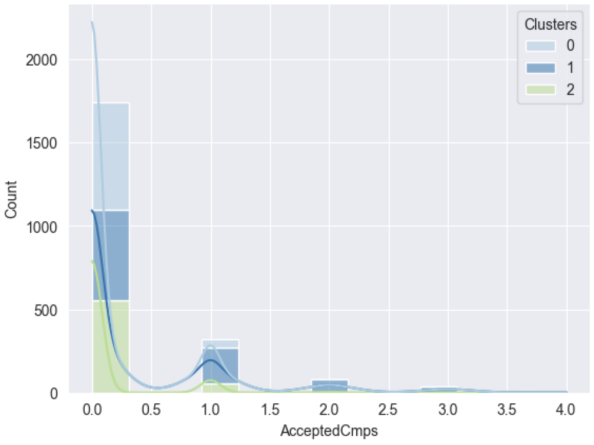
4. Purchases



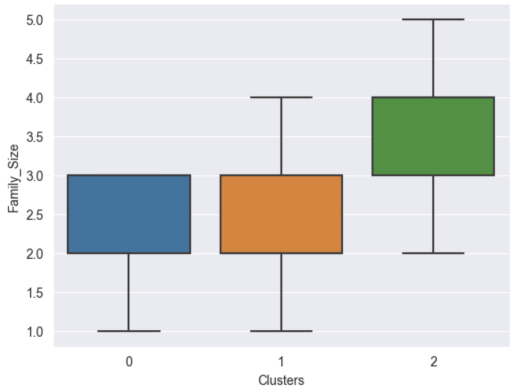


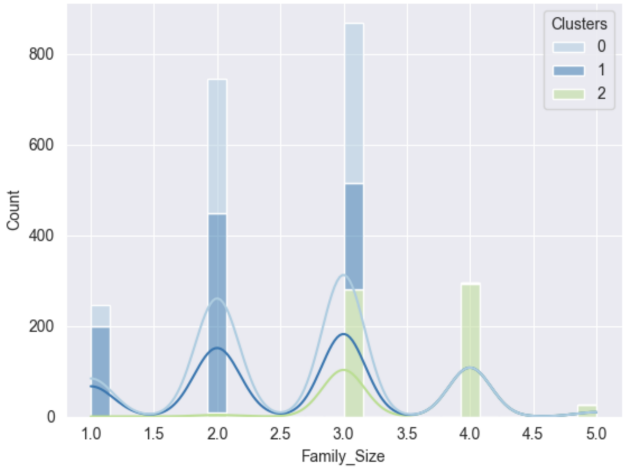
5. Accepted Campaigns

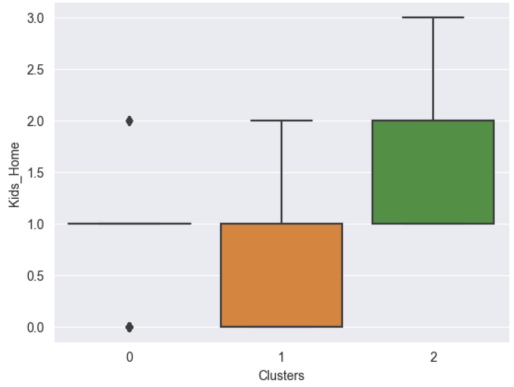


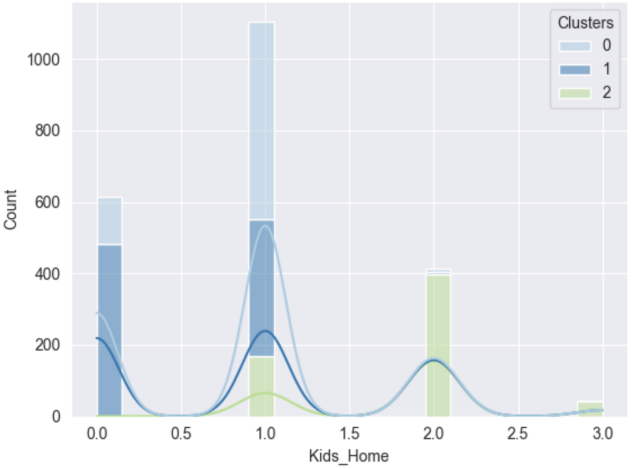


6. Family Size / Number of kids

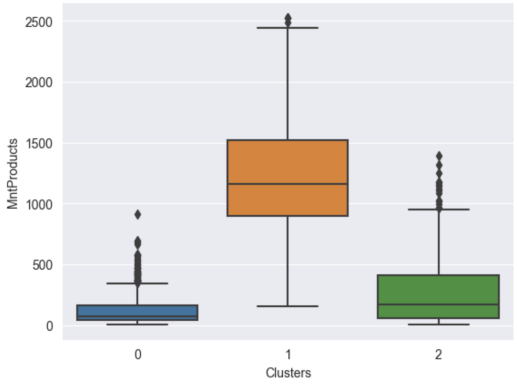


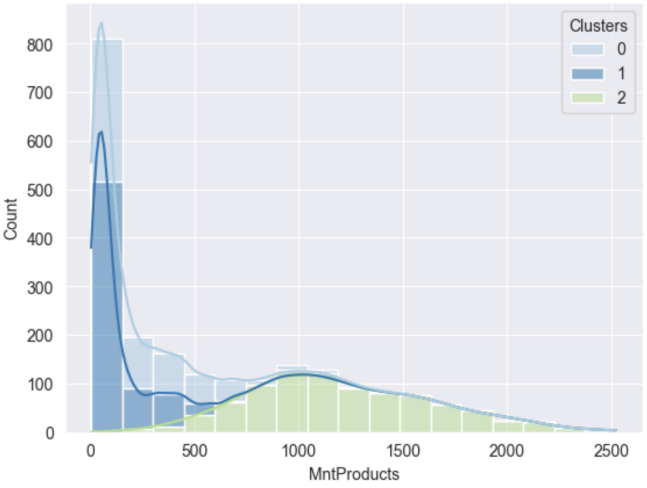






7. Amount spent on products





Cluster traits found:

Cluster 0

- Income between 10000 and 80000
- Low amount spent on products, below 700

- Predominant number of customers with a university degree but no higher education
- Age between 25 and 70
- Lowest number of purchases mostly between 0 and 17
- Campaign acceptance is practically non existent
- Customers with no kids or 1 kid

Cluster 1

- Income mostly between 20000 and 80000
- Wide range of amount spent on products, from 0 to 2500
- Predominant number of customers with a university degree but no higher education
- Age between 27 and 85
- Number of purchases between 10 and 35
- Highest campaign acceptance rate, mostly between 0 and 2 per customer
- Customers with 1 or 2 kids

Cluster 2

- Group with highest income range. Mostly from 40000 to 100000
- Customers with highest amount spent on products, mostly from 500 to 2500
- All kinds of education but also the only group with Master's or PhD
- 3 to 30 purchases
- Campaign acceptance is rare
- Customers with 1 to 3 kids