Customer Segmentation with PCA and K-means

This project simulates a real-world problem where a company wants to analyze the market and identify meaningful customer profiles.

We will summarize the information gathered from the history of sold/selling products using a few easily comprehensible concepts to assist managers in making future decisions.

We will perform dimensionality reduction. In such situations, the number of principal components (PCs) is often kept low, as reducing dimensionality is more important than preserving all the information.

We assume the dataset is the result of a data collection procedure performed by the company I am working for.

- I will summarize the available data to make it more interpretable, aiming for at most five features while preserving 40% of the information.
- I will identify “customer profiles” based on these summarized features.

DATASET:

1. **Year_Birth**: Customer’s birth year
2. **Education**: Customer’s education level
3. **Marital_Status**: Customer’s marital status
4. **Income**: Customer’s yearly household income
5. **Kidhome**: Number of children in customer’s household
6. **Teenhome**: Number of teenagers in customer’s household
7. **Dt_Customer**: Date of customer’s enrollment with the company
8. **MntWines**: Amount spent on wine in last 2 years
9. **MntFruits**: Amount spent on fruits in last 2 years

10. **MntMeatProducts**: Amount spent on meat in last 2 years
11. **MntFishProducts**: Amount spent on fish in last 2 years
12. **MntSweetProducts**: Amount spent on sweets in last 2 years
13. **MntGoldProds**: Amount spent on gold in last 2 years
14. **NumWebPurchases**: Number of purchases made through the company's website
15. **NumCatalogPurchases**: Number of purchases made using a catalogue
16. **NumStorePurchases**: Number of purchases made directly in stores
17. **NumWebVisitsMonth**: Number of visits to company's website in the last month
18. **NumDealsPurchases**: Number of purchases made with a discount
19. **AcceptedCmp1**: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
20. **AcceptedCmp2**: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
21. **AcceptedCmp3**: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
22. **AcceptedCmp4**: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
23. **AcceptedCmp5**: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
24. **Response**: 1 if customer accepted the offer in the last campaign, 0 otherwise
25. **Complain**: 1 if the customer complained in the last 2 years, 0 otherwise
26. **Recency**: Number of days since customer's last purchase
27. **ID**: Customer's unique identifier
28. **Z_CostContact:** not specified
29. **Z_Revenue:** not specified

PROCEDURE STEPS:

1. Data wrangling and cleaning: here I should study the dataset to see if it has missing values and stuff
2. Data exploration
3. Perform a PCA analysis
4. Apply K-MEANS to do customer segmentation.

1. Data wrangling and cleaning

After importing the necessary libraries and loading the dataset, the dataset was analyzed for:

- **NaN values**
    - 24 rows presented a null value for the feature “Income”. Since the dataset had a total of 2240 elements I have decided to proceed removing these rows, remaining with 2216 rows.
- **Duplicates**
    - There were a total of 182 duplicated rows. After removing them we end up with 2034 rows

We proceed with feature engineering where we:

- For the feature “**Marital_Status**” the following change was performed:

Married 788 Together 514 Single 439 Divorced 216 Widow 70 Alone 3 Absurd 2 YOLO 2	⟶	Married-Together 1302 Single 442 Divorced 216 Widow 70

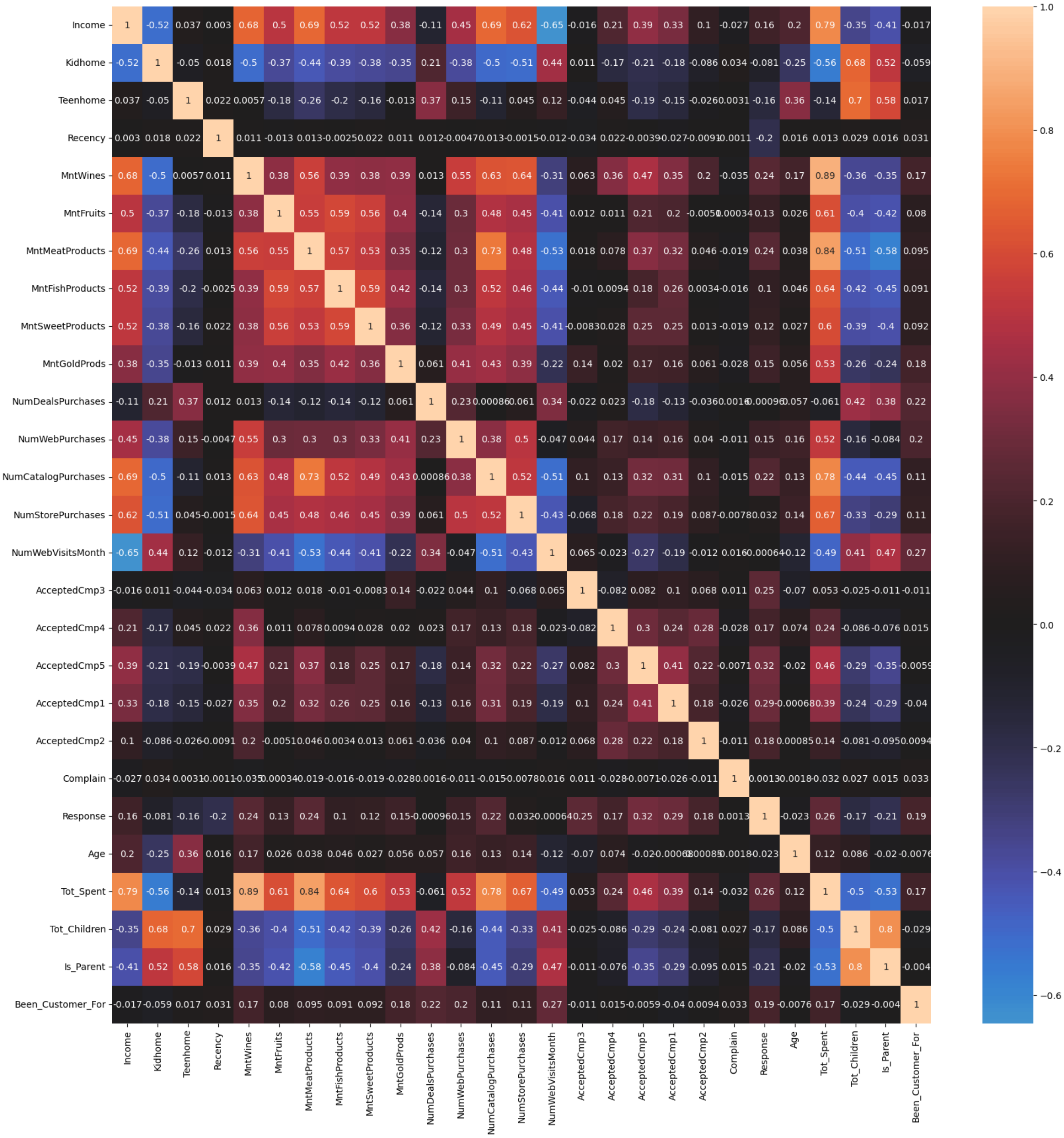
    - “Alone” can be interpreted as being “Single” so it was merged with the later category while “Absurd” and “YOLO” do not have a specific meaning and where removed

- The “**Year_Birth**” was used to compute the “**Age**” at this time (2024). The “Year_Birth” feature was then removed
- The feature “**Tot_Spend**” for total spendings was added by summing all of the partial spendings.
- The feature “**Tot_Children**” for total number of children at home was added..
- The feature “**Is_Parent**” for indicating if the customer is a parent or not was added.
- The feature “**Dt_Customer**” indicating for how long a person has been a customer was transformed into days (to the date 23-12-2024).

The next step was identifying outliers:

- Some persons with more than 100 year were identified and removed.
- Some customer with really high salaries where found. Only the one having an “Income” of 666666.00 was removed as the others may be rich customers.

A correlation matrix was then displayed to see the correlation among all features:



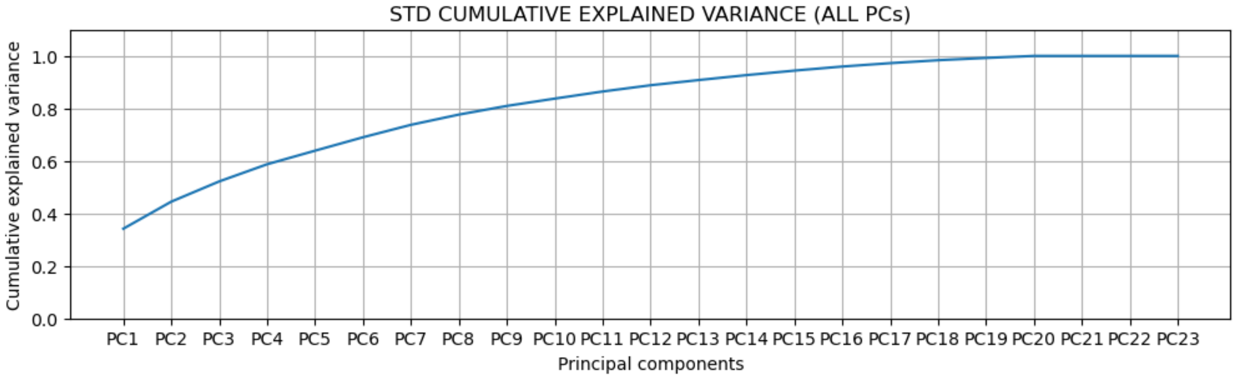
2. Data pre-processing

Since PCA was performed, the categorical features “Education” and “Marital_Status” were converted into numerical features. “Education” was transformed using label encoding, as it has an implicit ranking, while one-hot encoding was applied to “Marital_Status.”

Next, the data was standardized.

3. PCA

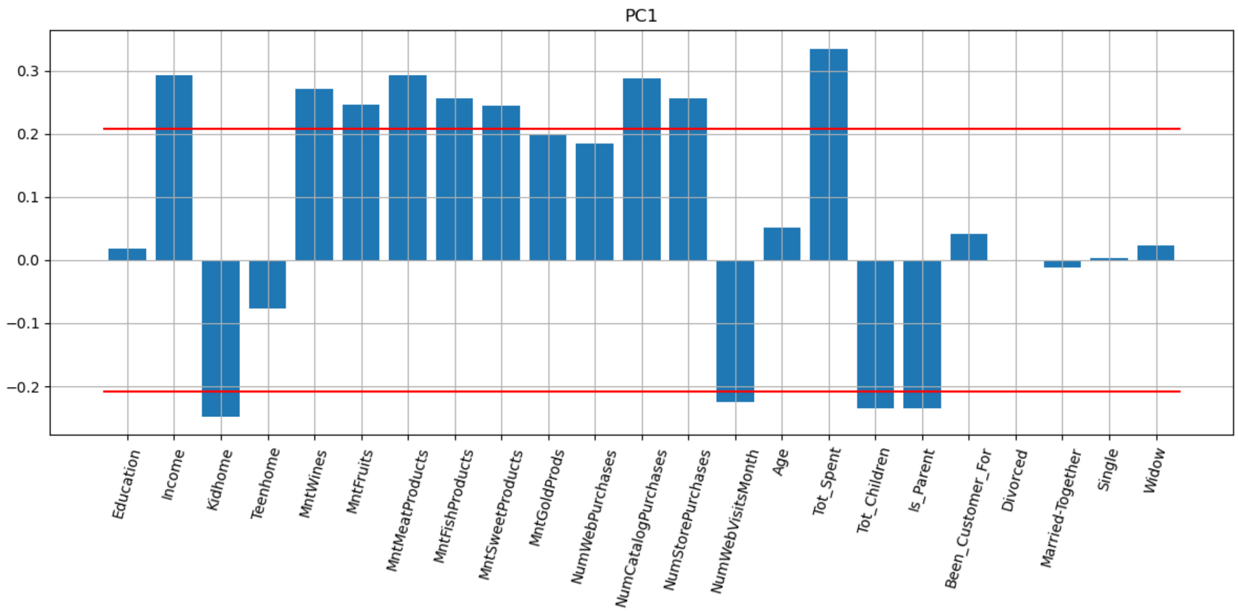
The first step was to perform a PCA to observe how much explained variance each PCs explained.

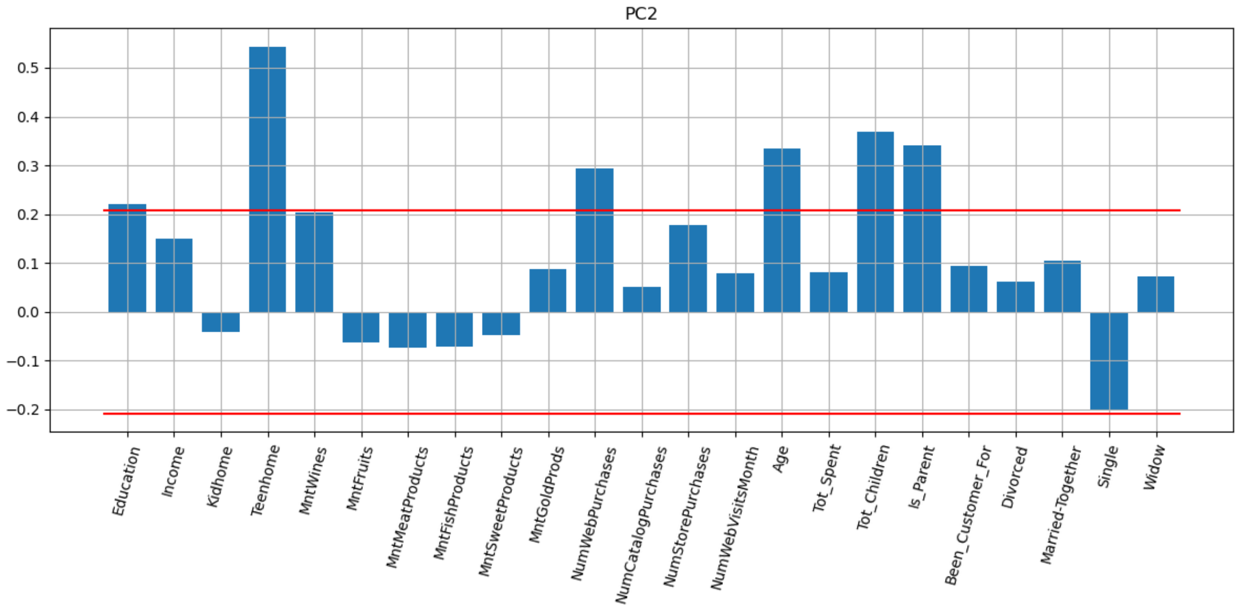


Since, in this simulation, the required information preserved should be 40%, only 2 PCs where selected.

- PC1 explained circa 34% of the total variance
- PC2 explained circa 10% of the total variance

The feature which are most explained by the PCs are the following:





The threshold was selected with an heuristic.

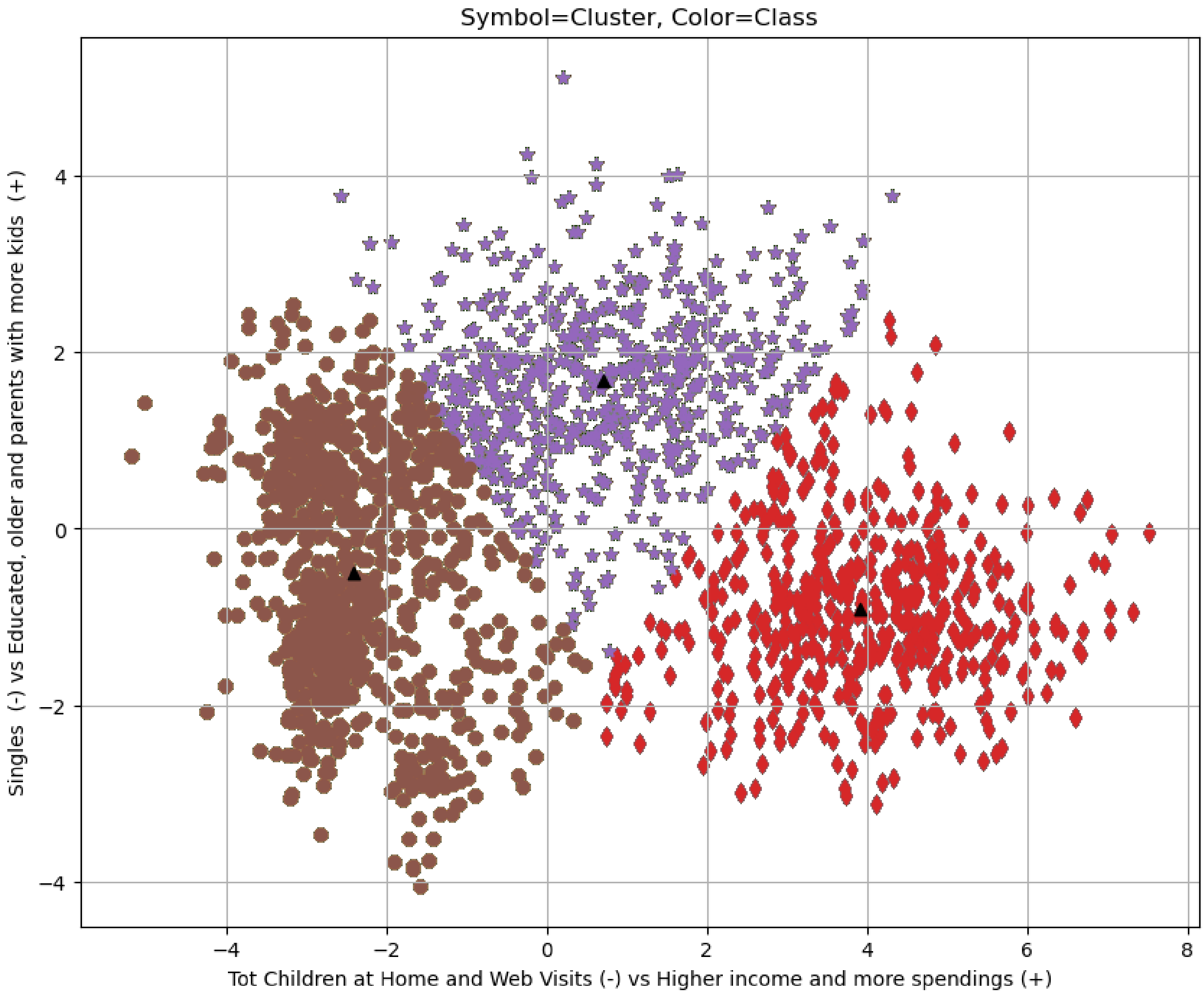
This allow us to interpret each point in the new space (the space where each PC is an axe of the cartesian plan):

- In PC1 we have:
  - **Higher PC1 values** correspond to customers with higher income, who spend more overall, and tend to shop more frequently in physical stores.

    - **Lower PC1 values** are associated with customers who have lower income, spend less overall, but have more children at home (are parents). These customers tend to spend more time browsing online, possibly looking for deals before deciding whether to make a purchase in-store.
- In PC2 we have:
    - **Negative PC2 values** are associated with customers who are more likely to be singles.
    - **Positive PC2 values** correspond to customers who tend to have higher levels of education, are older, are parents with an increasing number of children at home, and make a higher number of purchases online.

4. K-Means Clustering

K-Means clustering, with different numbers of clusters, was performed and the one having the highest silhouette score is selected. The best number of cluster is 3. The following groups where found:



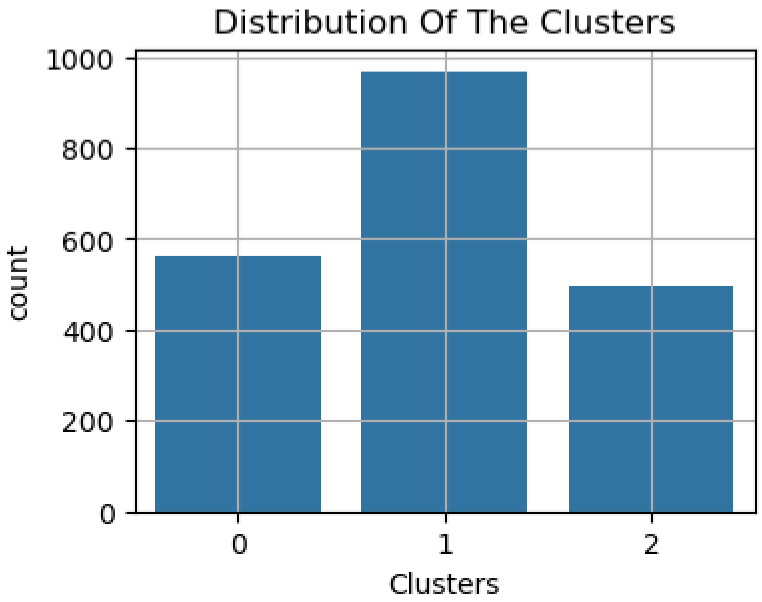
This helps in profiling each customer cluster:

- Cluster 0 (Purple):
  - Are older
  - Are parents
  - Higher education
  - Medium to high income and spenders
- Cluster 1 (Brown):
  - Most of them are single
  - Younger

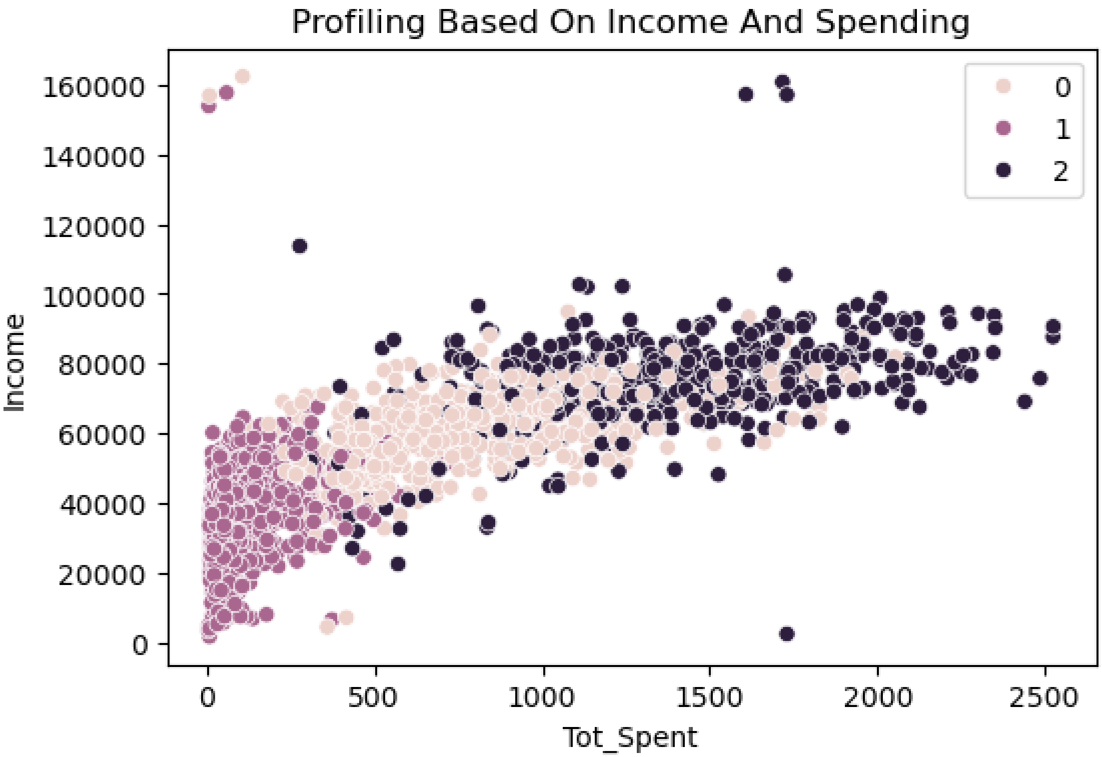
    - Have children at home
    - Don't have much free time to go to the store and prefer to go on the website
- Cluster 2 (Red):
    - Most of them are single
    - Younger
    - High income
    - High spendings

5. Patterns

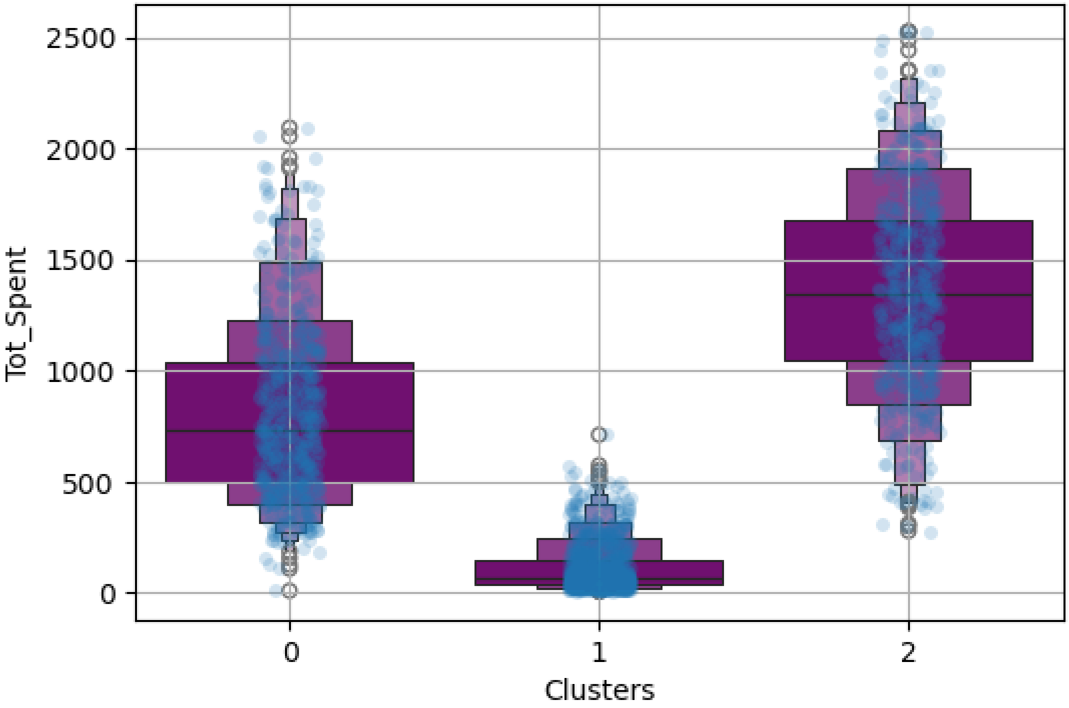
There are the following number of elements per cluster:



The following plot shows how the groups are distributed based on spending and income.

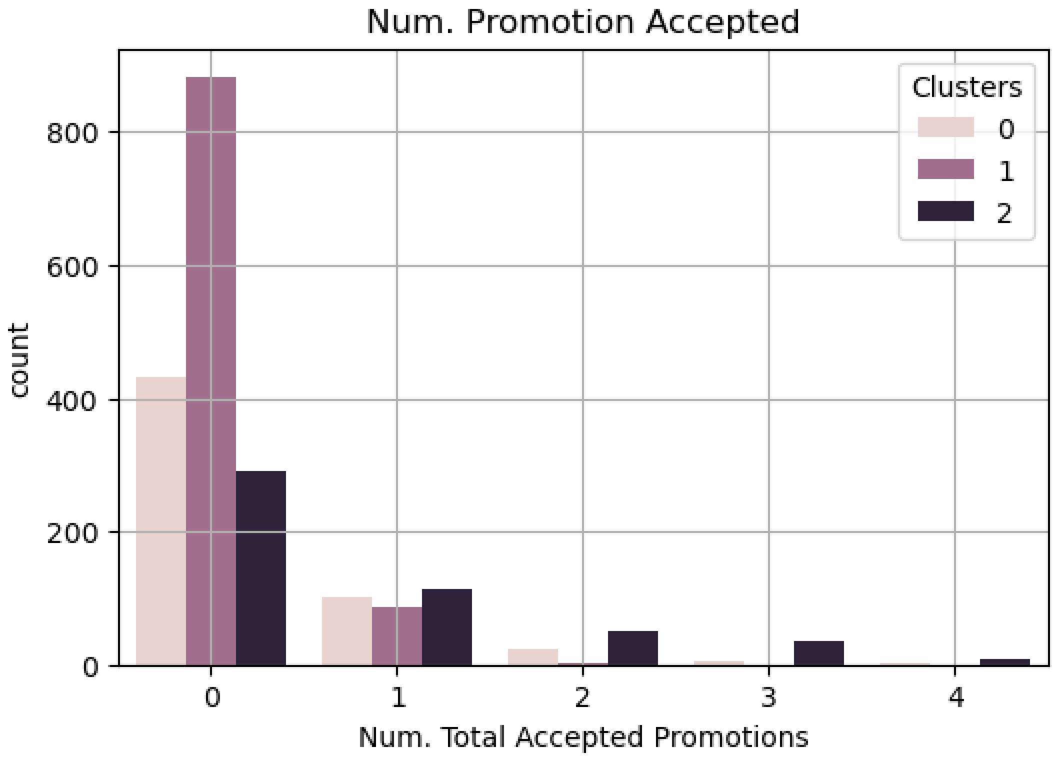


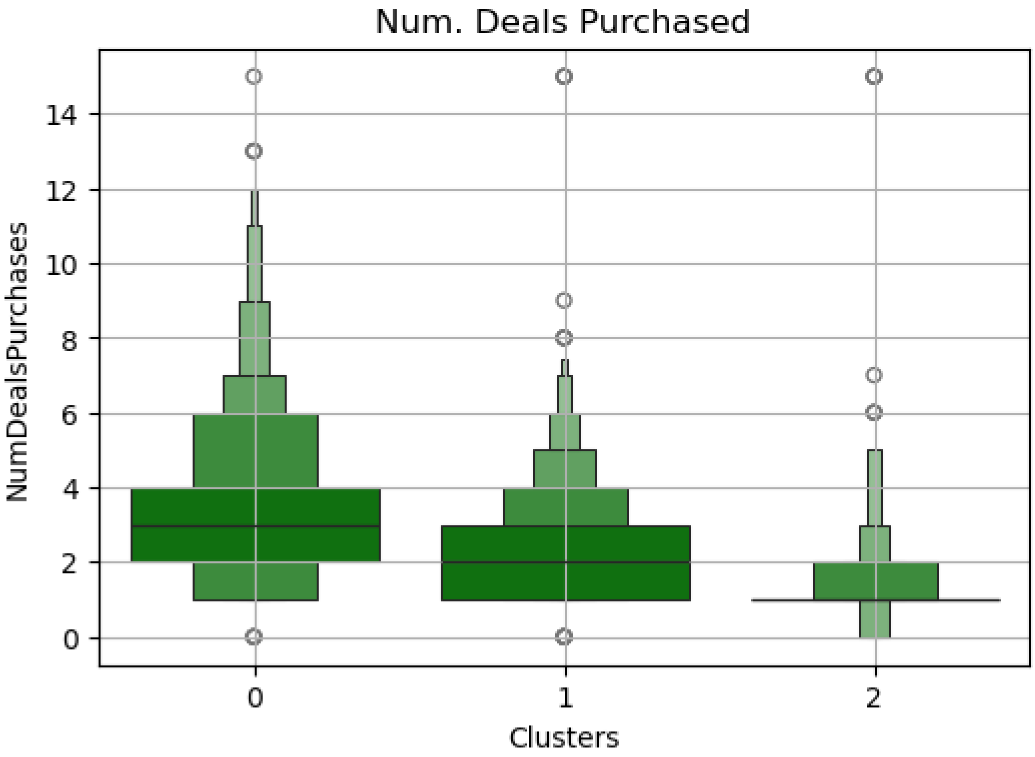
How much each group spend in general:



We can notice that the group 2 has the highest spending habits.

The following two plots show how many promotion campaigns were accepted and how many deals were purchased.





This indicates that promotion campaigns were not very effective and could be improved, while deals were considered appealing across all groups. A targeted deal marketing strategy focused on Group 2, which has higher spending habits, could benefit the company.