ANALYTICS PROJECT

Marketing Campaign Customer Segmentation

BY RODRIGO COSTA

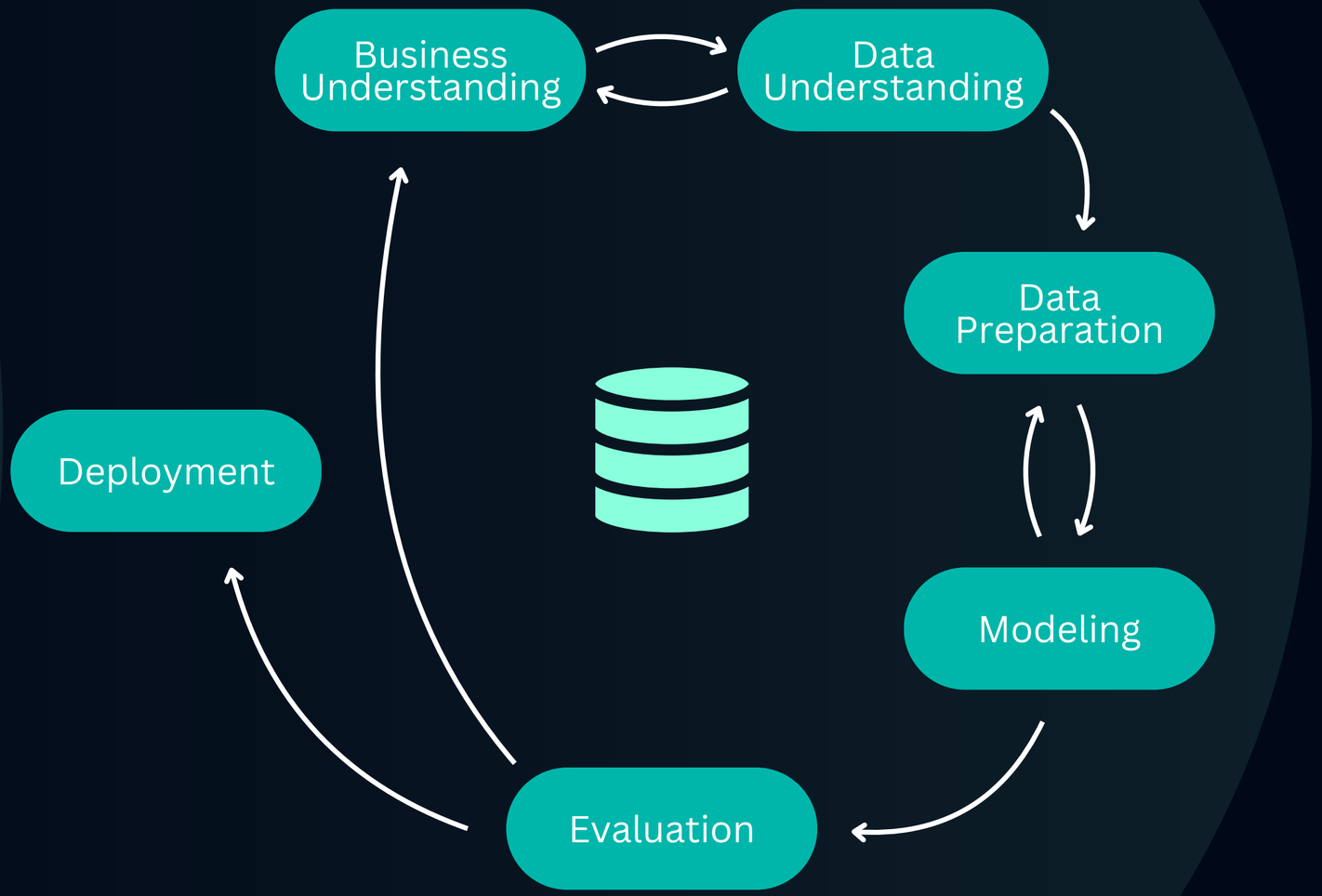
Table of Contents.

1. METHODOLOGY
2. BUSINESS UNDERSTANDING
3. DATA UNDERSTANDING
4. DATA TRANSFORMATION
5. DATA CLEANING
6. DATA PREPARATION
7. CLUSTERING
8. RECOMMENDATIONS
9. CONTACTS

Methodology.

To support the problem solving, a **CRISP-DM** approach was used.

This is the most commonly used approach for data science projects. It is easily **generalizable** to every type of project, since it is very **flexible**, having the **right start** (business understanding), and a strong finish (**deployment**)



Business Understanding.

- Customer segmentation is the process of breaking up your clientele according to **shared traits**, such as behaviors or demographics.
- Segmentation helps organizations know what **new products and services** they might want to invest in, and uncovers ways to improve how the business sells.
- By implementing customer segmentation, organizations can enhance their **marketing effectiveness**, improve **customer satisfaction**, and drive overall **business growth**.

Areas of Implementation

- **Retail & E-Commerce**
- **Healthcare**
- **Financial Services**
- **Travel & Hospitality**

Data Understanding.

2240 observations of clients from a retail store | 29 attributes

Client Identity Data

- ID
- Year_Birth
- Education
- Marital_Status
- Income
- Kidhome
- Teenhome

Client Store Data

- Dt_Customer
- Recency
- #Products
  - Wines
  - Fruits
  - Fish
  - Meat
  - Sweet
  - Gold
- #Purchases
  - Deals
  - Web
  - Store
  - Catalog
- #VisitsMonth

Campaign Related Data

- Accepted_Campaign
  - 1
  - 2
  - 3
  - 4
  - 5
- Complain
- Cost_Contact
- Revenue
- Response

Data Transformation - Feature Engineering.

Marital Status Variable

Married	Together	Single	Divorced	Widow	Alone	Absurd	YOLO
864	580	480	232	77	3	2	2

Partner → Living variable ← Alone

Education Variable

Graduation	PhD	Master	2nd Cycle	Basic
864	580	480	232	77

Graduate | Undergraduate

Creating the Age Variable

Year_birth → Age

Tells how old is the customer

Data Transformation - Feature Engineering.

Creating the Customer_for

Dt_Customer ⟶ Customer_for

Gives the number of days since becoming a customer of the store

Calculating the Total Spending Variable

Spent = Wines + Meat + Fruits + Sweet + Fish + Gold

$$$

Simpler way of knowing how much the client spent on the store until the day of analysis

Creating the Dependents Variable

Dependents = Kidhome + Teenhome

Is_Parent (binary) Variable

Family_Size Variable

- Alone: +1
- Partner: +2
- Dependents: +#Dependents

Data Cleaning.

Pair Plot – Variable Analysis

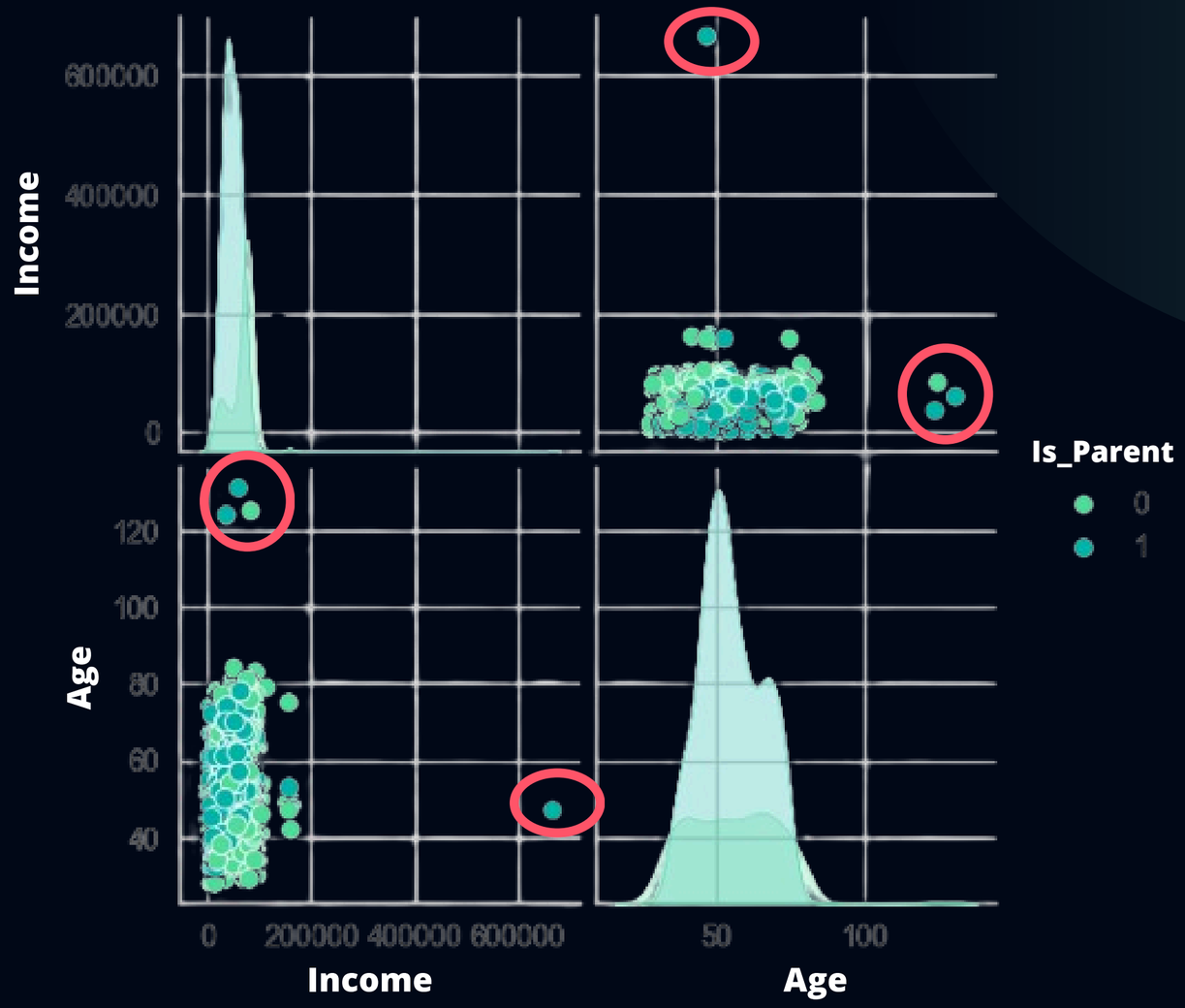


Figure 1: Pairplot graph Income - Age

- 4 **outliers** were found and eliminated, since **no further information** was available in order to correct them
- **No mathematical approaches** were used to identify/remove outliers, since we would have no proofs if those values are actually wrong - real outliers.

Data Preparation.

Missing Values

24 missing values were found on the Income variable → Imputation of values using the K-Nearest Neighbor method, with k=3

Customer Enrollment

Newest Customer	Oldest Customer
2014-06-29	2012-07-30

Marital Status Variable

Married	Together	Single	Divorced	Widow	Alone	Absurd	YOLO
864	580	480	232	77	3	2	2

Education Variable

Graduation	PhD	Master	2nd Cycle	Basic
864	580	480	232	77

Data Preparation.

Correlation Matrix

High correlations Analysis:

- **Income**: Wines, Meat, NumCatalogPurchases, NumStorePurchases, NumWebVisits, Spent
- **Wines**: Meat, NumCatalogPurchases, NumStorePurchases
- **Is_Parent** is highly correlated with the amount **Spent** on the store
- Some conclusions can also be taken when analyzing the **acceptance of previous campaigns.**

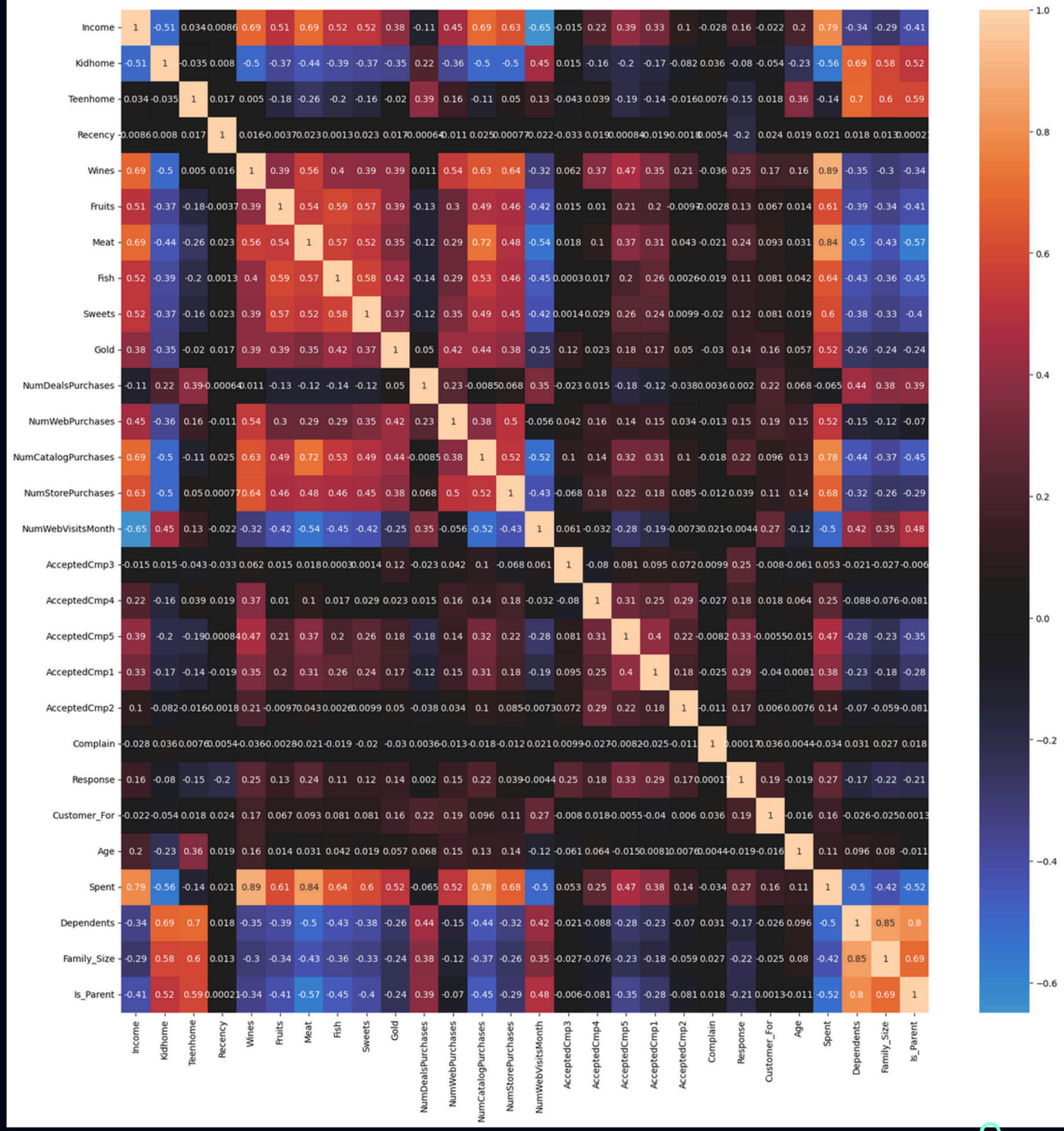


Figure 2: Correlation Matrix

Data Preparation.

Encoding

- Education
  - Undergraduate: 1
  - Graduate: 0
- Living
  - Alone: 0
  - Partner: 1

Normalization

Numeric Variables

$$Z = \frac{x - \mu}{\sigma}$$

Score, Mean, SD

Feature Reduction – PCA

Principal Component Analysis (PCA)

Allows to reduce the number of feature to a specified amount of dimensions (3 in this case)

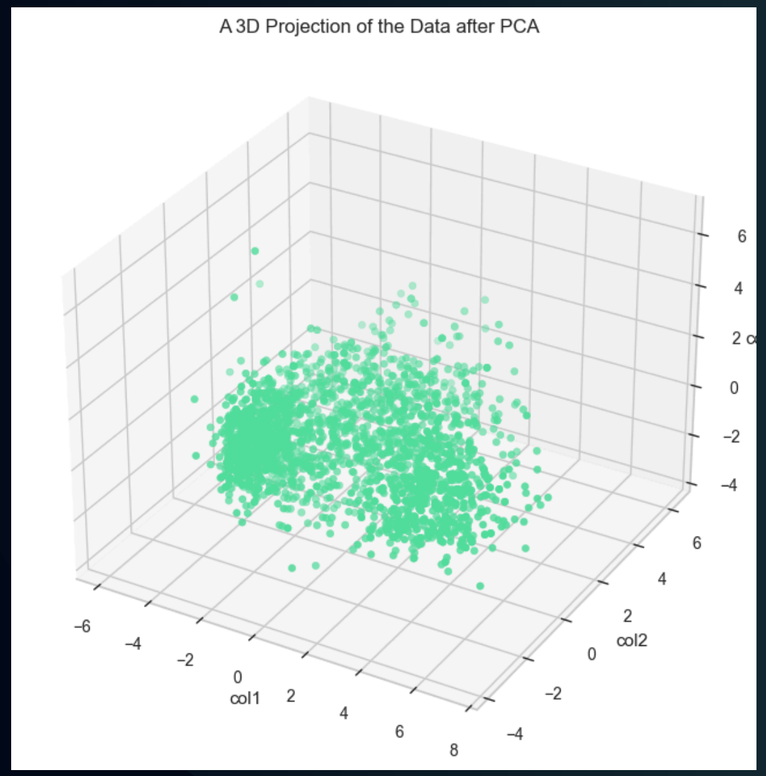


Figure 3: PCA points 3D Projection

Clustering.
Implemented Algorithms

K-Means

- **Simple and easy to implement**;
- **Efficient** for large datasets;
- Works well when clusters are spherical and evenly sized;
- Assumes clusters are of similar size and shape;
- Sensitive to initial seed selection and outliers;
- **Requires specifying the number of clusters in advance**.

CLARANS

- **Effective for large datasets**;
- Can handle noise and outliers better than K-Means;
- **Does not require the number of clusters to be predefined**;
- **Computationally intensive due to random sampling**;
- **Less interpretable** compared to simpler algorithms;
- Performance can be inconsistent depending on the quality of the random samples.

DBSCAN

- Can find **arbitrarily shaped clusters**;
- Handles noise and outliers effectively;
- **Does not require specifying the number of clusters**;
- Performance **depends on the choice of parameters** (epsilon and minPts);
- **Struggles with clusters of varying densities**;
- Can be **inefficient** for very large datasets.

Agglomerative Clustering

- **Does not require the number of clusters to be specified beforehand**;
- Produces a **dendrogram**, providing insight into the data hierarchy;
- Flexible with various linkage criteria (single, complete, average);
- Computationally expensive for large datasets;
- Sensitive to noise and outliers;
- **Can result in imbalanced clusters**.

Clustering.
Dismissed Results

CLARANS

- The computation of this algorithm took some hours to run, revealing not to be efficient on clustering the data.
- In addition, the results gotten didn't seem very interpretable when analyzed

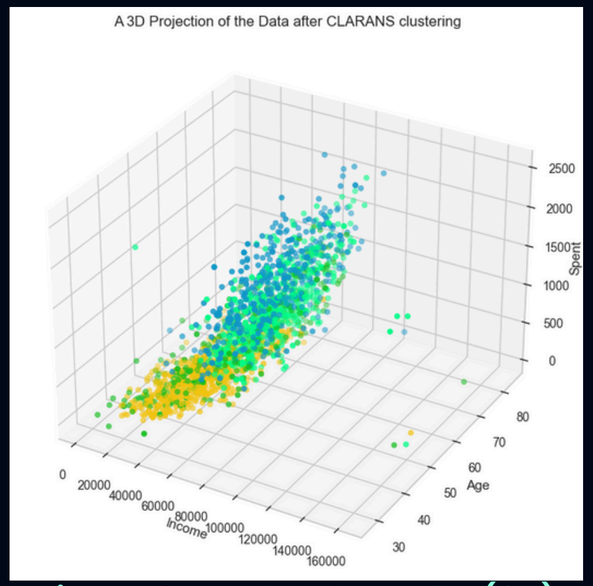


Figure 4: CLARANS results (3D)

DBSCAN

- Results were also not easy to interpret, and the algorithm seemed to struggle in the creation of the clusters.
- This difficulty might be because of the variation in density of points along the dataset.

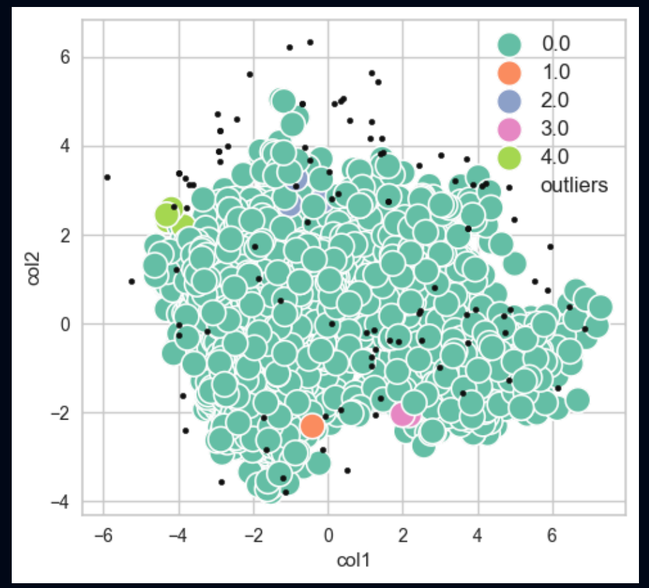


Figure 5: DBSCAN results

Agglomerative Clustering

- Produced good results, but K-Means was preferred due to its easier implementation

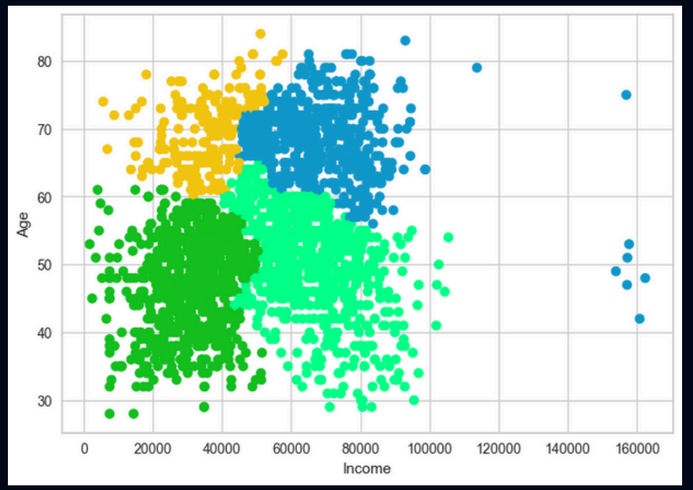


Figure 6: Agglomerative clustering results

Clustering.

Results

The K-Means algorithm was the one chosen to produce the final results

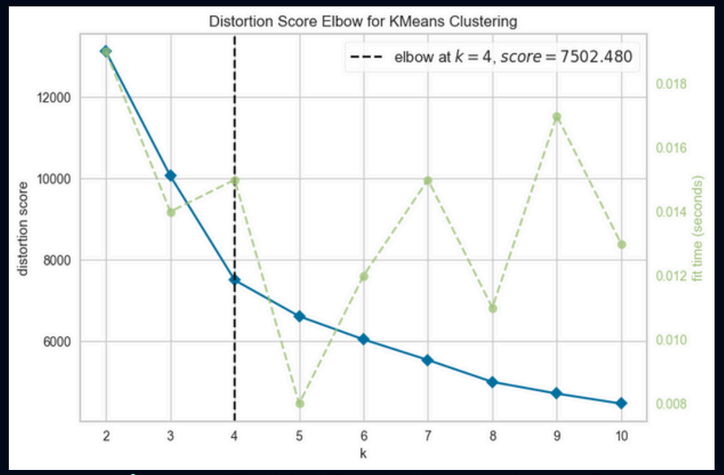


Figure 7: NbClust Score Graph

Using the Nbclust algorithm, it was possible to find the optimal value of k in order to apply the K-means algorithm with the best balance between score and error.

k* = 4 clusters

Figure 8: Age and Income by cluster

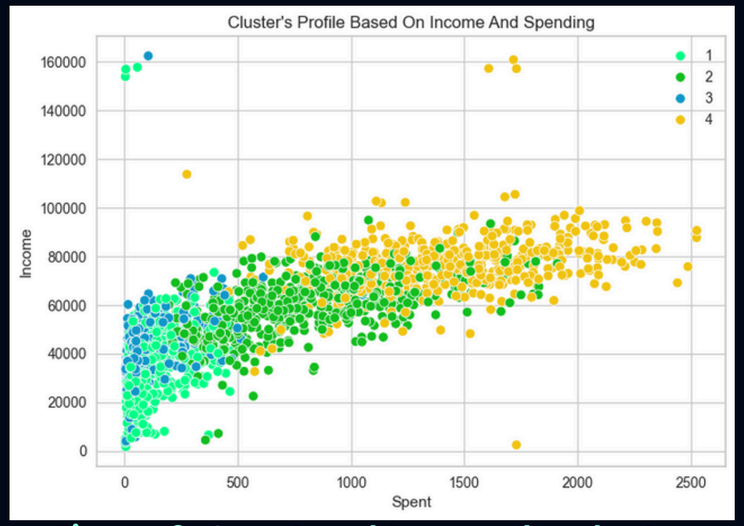


Figure 9: Spent and Income by cluster

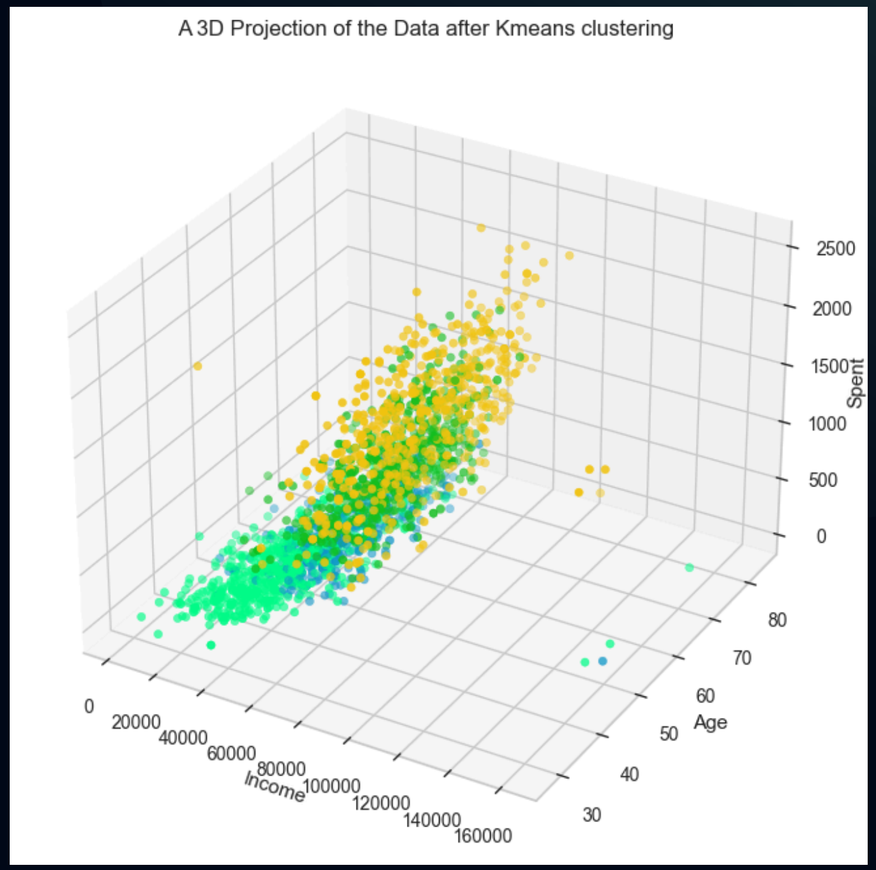


Figure 10: Income, Spent, and Age 3D projection by cluster

Clustering.

Results

Customer characteristics in each cluster

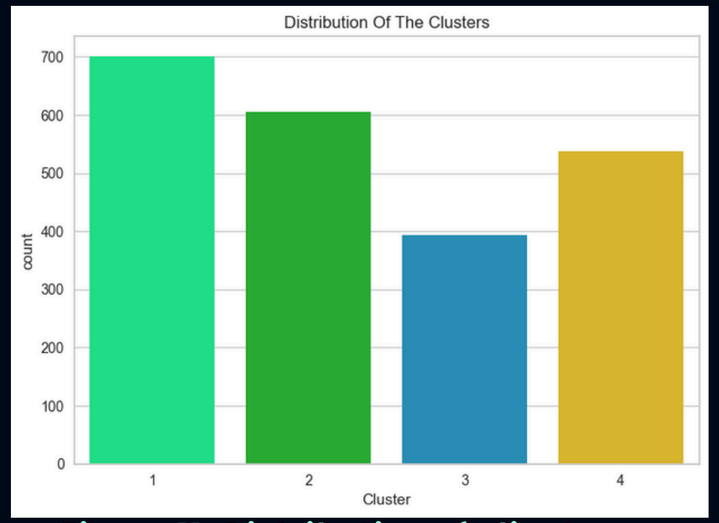


Figure 11: Distribution of clients per cluster

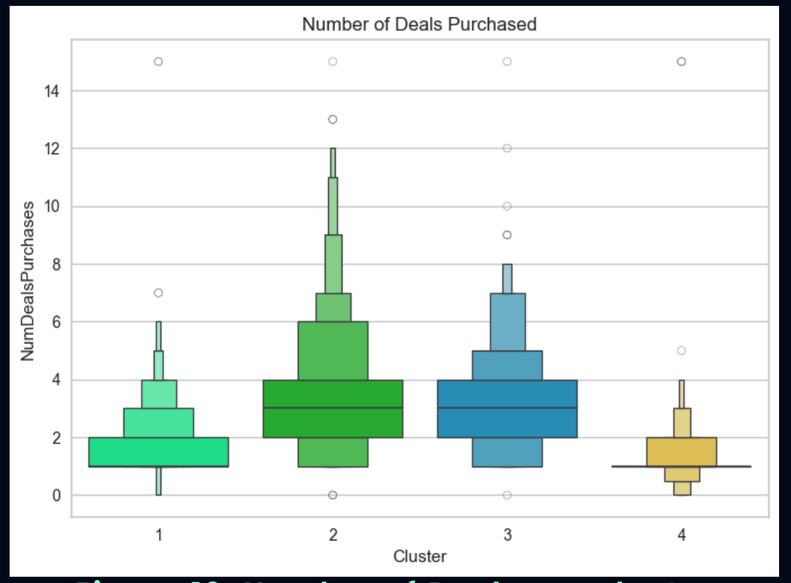


Figure 12: Number of Deals per cluster

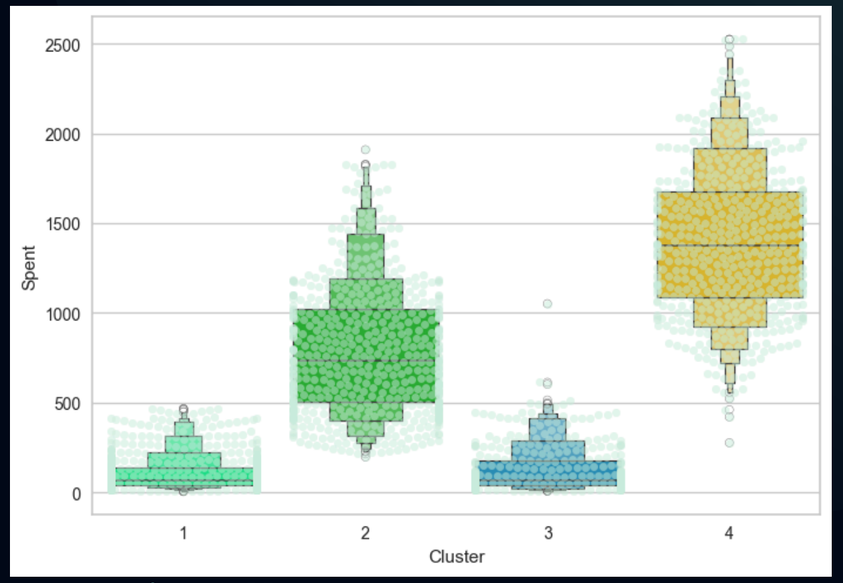


Figure 13: Money Spent per Cluster

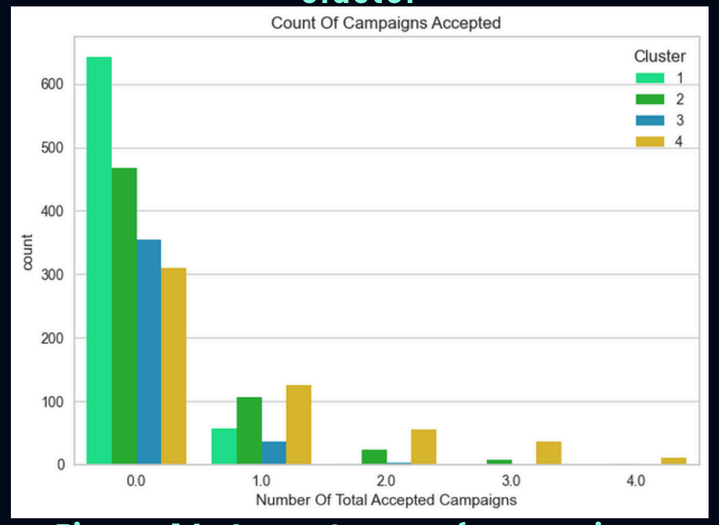


Figure 14: Acceptance of campaigns per cluster

Results Summary

- The clusters seem to be **well distributed** along the dataset (which is a characteristic of K-Means)
- Clients from **group 4** are the ones who **spend the most**, but **purchase little**. The ones from **group 3**, spend way less, but **purchase a reason amount of deals** (on average). The clients from the **2nd cluster** purchase the same as the 3rd, but spend way more than them.
- Clients from the **1st cluster** are the ones who **spend the least**, but accept a bit more deals than the fourth group

Recommendations.

- As seen before, there is a **high correlation between income and the total money spent on the store**. With this said, clients from the **4th group are the ones who receive and spend the most, also accepting the highest number of campaigns**. These clients can be considered the gold, of the store, and the marketing team should keep their strategies to keep them as current clients.
- Following an identical customer profile, we have the **2nd group**. These are also high income and high spendings clients, but who don't accept many campaigns. The marketing team should try to approach these clients in another way, since they are older customers, and probably want different campaigns.
- The **3rd group** reflects the smallest percentage of clients, these seem to represent mid-income, older customers who spend very little on the store. There might be two reasons for this: they either prefer to spend their money on other store, or are just generally not satisfied with the store, which can be justified by the total number of complaints which is twice higher than the 4th group. In my opinion, the marketing team shouldn't focus on advertising for these clients, but, if the resources allow to, increasing their customer satisfaction rates would be very beneficial for the store. An initial step could be to send surveys to know how can the store improve in these clients' opinion.
- The **group 1** simply represents shorter-income, younger clients, who are not that interested in the store and who probably just don't have the time (or money) to spend in the store. These should also not be a big focus of the marketing team.

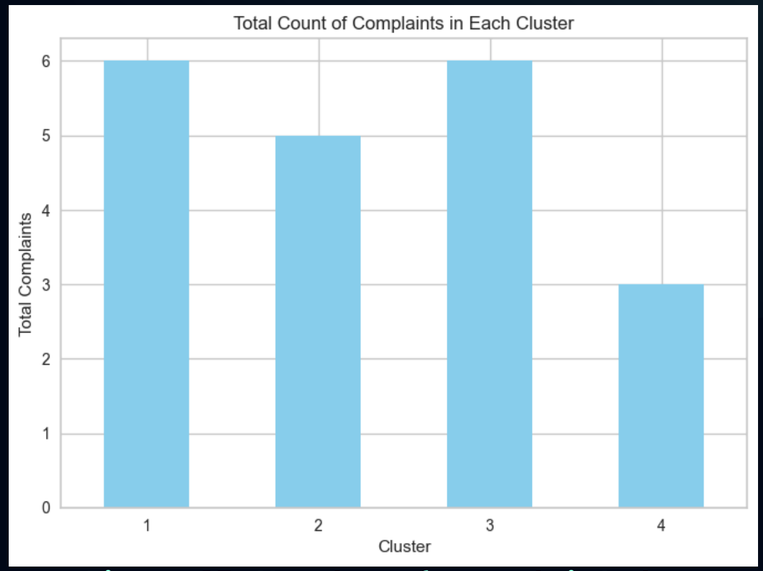
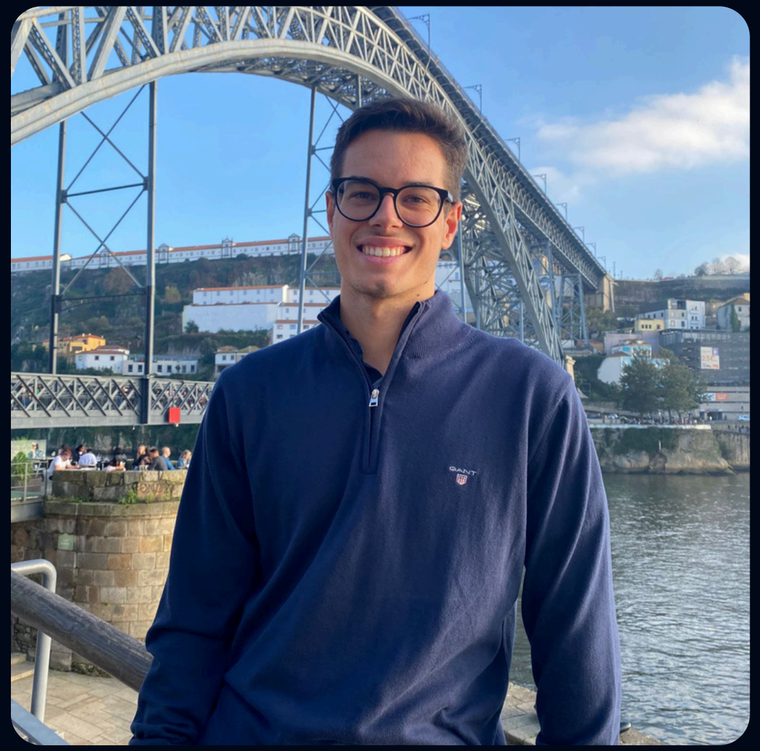


Figure 15: Number of complaints by cluster



Contact me.

@rodrigoco5ta

@rodrigoco5ta

@rodrigoco5ta

rodrigo-costa1@live.com.pt