Higher School of Economics

Faculty of computer science

***Project For the Course:***

***Modern Method in Data Analysis***

Presented by:

Fares Ghazzawi

Anwar Ibrahim

2021-2022

1- Problem Statement:

Customer Personality Analysis is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors, and concerns of different types of customers.
Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company's database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment.

2- Dataset summary with basic statistics and respective plots:

***The used dataset Link***: https://www.kaggle.com/imakash3011/customer-personality-analysis.

***Describing the dataset:***

The given dataset has 2240 objects and 29 attributes and those attributes are listed below:

The Attributes:

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
- Teen home: 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

***Please refer to the attached python file for data statistics and plots.***

3- Methodology:

a. Justify the selected ML/DM approach:

***K- mean***

we decided to use K-Means, because we are trying to find groups "clusters", which haven't been explicitly labeled in the data, and it is the least complex method.

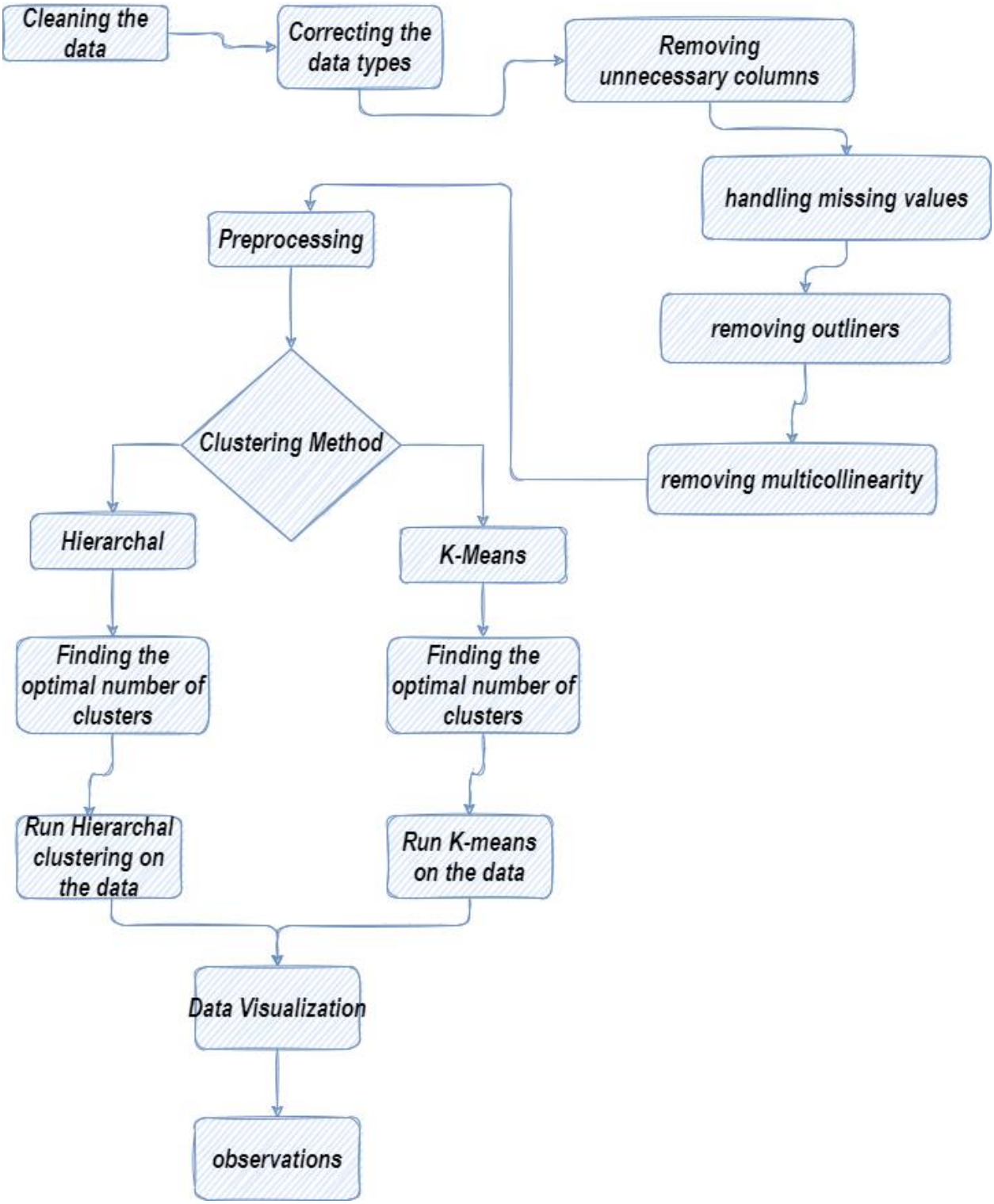
***Hierarchal***

We also used Hierarchal clustering, because it allows us to build tree structures depending on the similarities between data,

4- Experiment:

a. Setup: each step is explained thoroughly in the attached python file.

1- Cleaning the data.
   Correcting the data types.
   Removing the unnecessary columns.
   Handling missing values.
   Removing outliers.
   Removing multicollinearity.
2- Preprocessing.
   Scaling.
   Encoding.
3- Clustering method
   Apply elbow method to determine the optimal number of clusters.
   Run the algorithm on the preprocessed data, with the resulting optimal number of clusters.
4- Data visualization.
5- Observations.




We also used Orange, to study the Hierarchal clustering method, and you will find the results in the attached orange file.

b. Results:

| Please notice that K-Means results are in the attached python file. |
|---|

***Hierarchal results From Orange File:***

Some of the results of the hierarchical clustering.

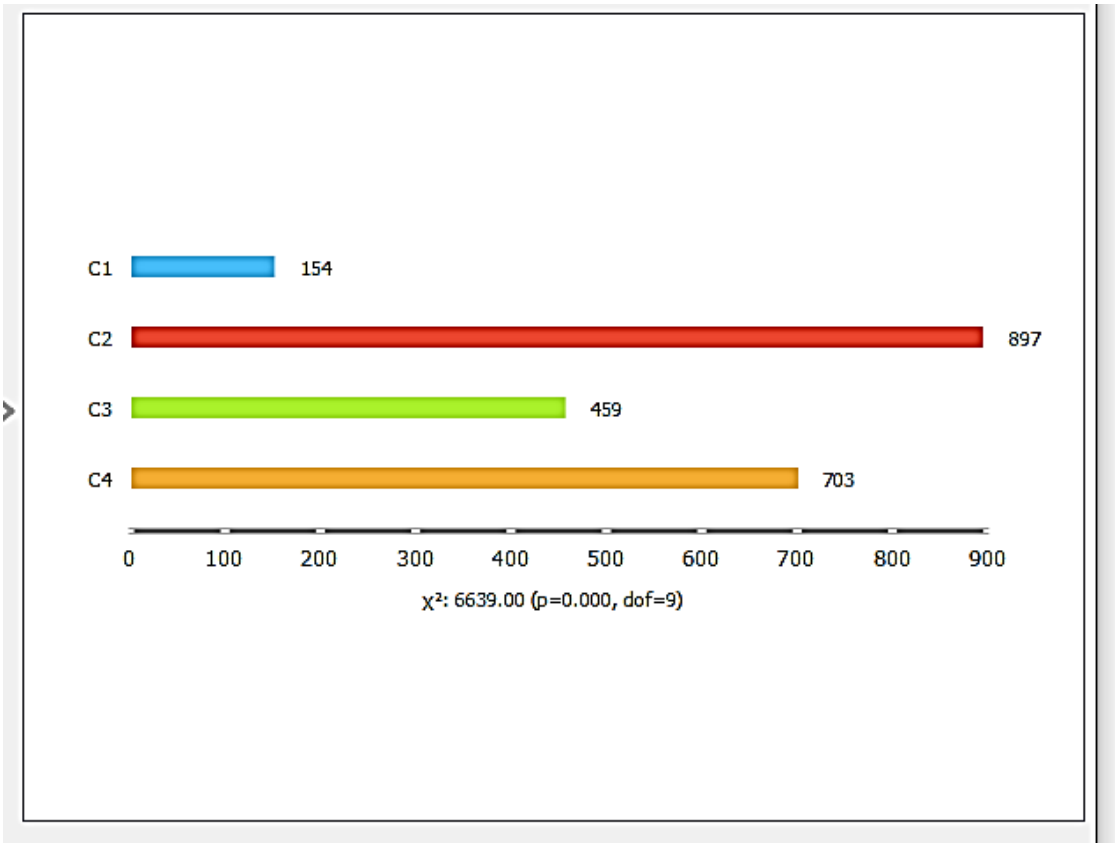



*Figure 1: box plot, Hierarchal clustering number of objects in each cluster*

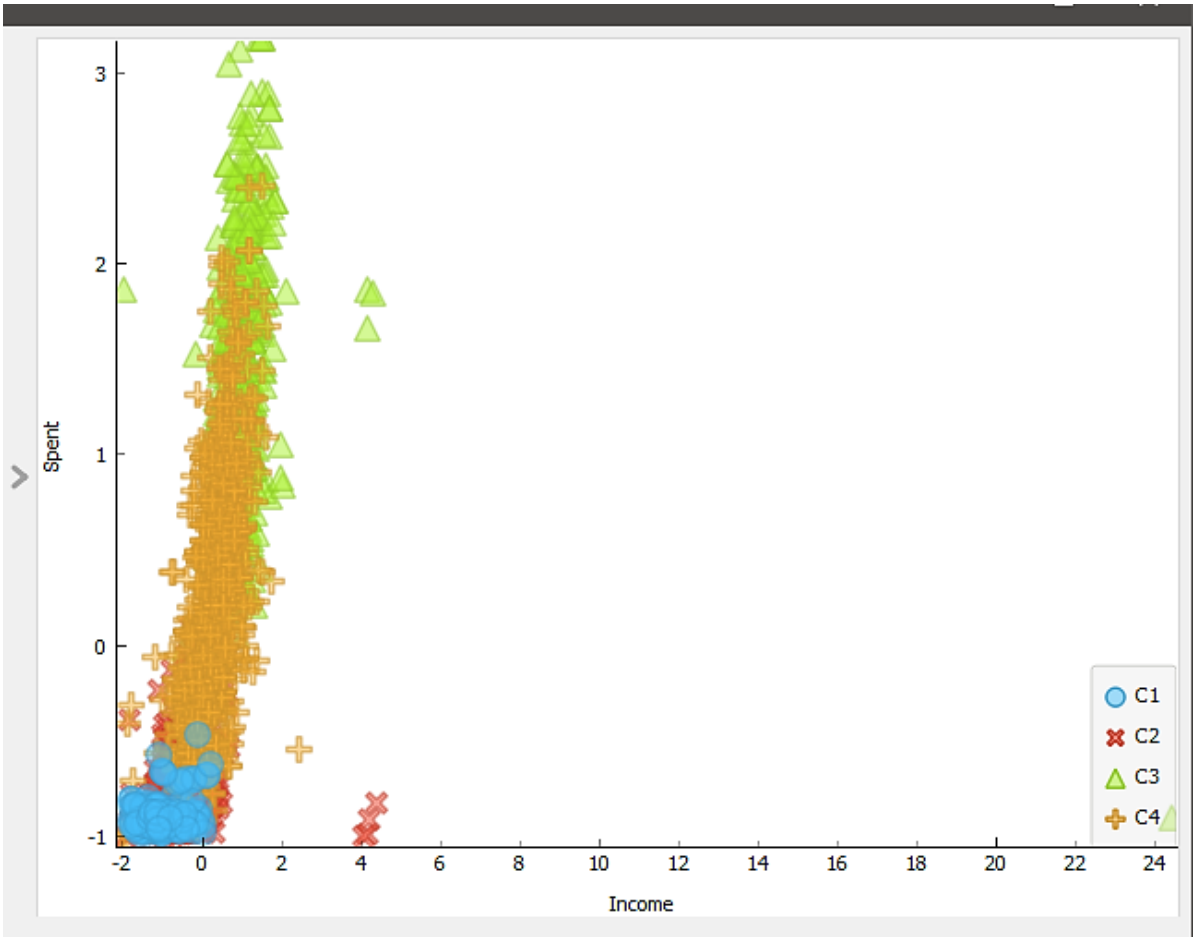



*Figure 2: scatter plot of the spending in terms of income*

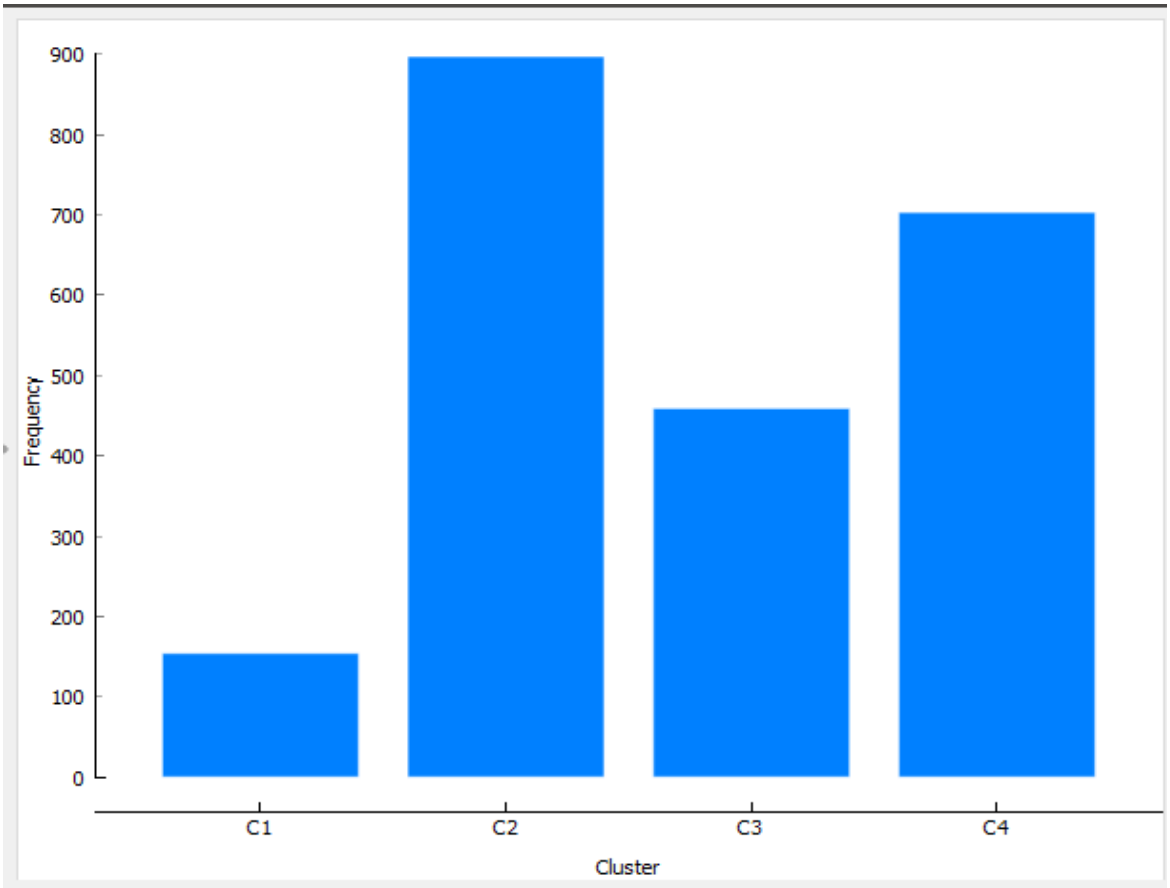



*Figure 3: distribution of clusters*

***Observations and results of hierarchical clustering:***

*Please Notice, results displayed below are obtained by manipulating the plots inside the "Orange" file, not only by the above three figures.*

From the "Orange" file we obtained the following properties for each cluster:

***Cluster 1 "blue":***

- Has the least number of objects. "154"
- They spend the least.
- All of them are undergraduates.
- They are the youngest.
- Most of them have no kids.

***Cluster 2 "red":***

- Has the greatest number of objects. "897"
- Has only postgraduates.
- Don't spend a lot.
- Most of the customers in it have either no kids or two kids.

***Cluster 3 "green":***

- Has 459 objects in it.
- Most of the customers in this cluster have either one kid or no kids.
- They spend the most out of all the clusters.
- Most of them are postgraduates.
- The most active people.

***Cluster 4 "orange":***

- Has 703 objects in it.
- Has both undergraduates and postgraduates.
- Most of them have one kid.
- Most of them spend moderately.

***Conclusion***:

In this project, we have used two clustering methods "K-Means, Hierarchical clustering" to solve the problem of "Analyzing Customer Personality", where we used "K-Means" to explore the finical properties in each cluster such as the customer's reaction to the company's

campaigns and the relation between the income and spending for each customer. We also used hierarchical clustering to further explore the general properties of each cluster. We have come to the conclusion that both these methods “under the same preprocessed data” gave similar results “i.e.: spent in terms of income”. K-Means is better in terms of complexity, however, hierarchical clustering is more flexible in terms of knowing the optimal number of clusters.