CUSTOMER SEGMENTATION

Foundation of Data Science

GROUP 8

Made by-

Manvendra Sharma –AM.EN.U4CSE19034

Mayank Kumar Shakya – AM.EN.U4CSE19035

Sourav Chindarmony – AM.EN.U4CSE19053

Challa Venkata Saikrishna – AM.EN.U4CSE19064

Abstract

In this project, we will implement customer segmentation in R. Whenever you need to find your best customer, customer segmentation is the ideal methodology.

In this project, DataFlair will provide us the background of customer segmentation. Then we will explore the data upon which we will be building our segmentation model. Also, in this data science project, we will see the descriptive analysis of our data and then implement several versions of the K-means algorithm.

Introduction

In this project, I will be performing an unsupervised clustering of data on the customer's records from a groceries firm's database. Customer segmentation is the practice of separating customers into groups that reflect similarities among customers in each cluster. I will divide customers into segments to optimize the significance of each customer to the business. To modify products according to distinct needs and behaviours of the customers. It also helps the business to cater to the concerns of different types of customers.

Dataset

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.

ID Customer's unique identifier	Year_Birth Customer's birth year	Education Education Qualification of customer	Marital_Status Marital Status of customer	Income Customer's yearly household income	Kidhor Number c customer
0 total values	2240 total values	[null] 100%	[null] 100%	2240 total values	to
5524	1957	Graduation	Single	58138	0
2174	1954	Graduation	Single	46344	1
4141	1965	Graduation	Together	71613	0
6182	1984	Graduation	Together	26646	1
5324	1981	PhD	Married	58293	1
7446	1967	Master	Together	62513	0
965	1971	Graduation	Divorced	55635	0

The dataset has the following features

ID – Unique ID

Year_Birth

Education

Marital_Status

Income

Kidhome

Teenhome

Dt_Customer – enrollment date of the customer

Recency – Cumtomer's last purchase

A Naïve Overview on dataset

Pandas provides some specialized functions to get a bird’s eye view of the dataset involved before delving deeper into the dataset itself.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2212 entries, 0 to 2239
Data columns (total 30 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Education            2212 non-null   int32
 1   Income               2212 non-null   float64
 2   Kidhome              2212 non-null   int64
 3   Teenhome             2212 non-null   int64
 4   Recency              2212 non-null   int64
 5   Wines                2212 non-null   int64
 6   Fruits               2212 non-null   int64
 7   Meat                 2212 non-null   int64
 8   Fish                 2212 non-null   int64
 9   Sweets               2212 non-null   int64
 10  Gold                 2212 non-null   int64
 11  NumDealsPurchases    2212 non-null   int64
 12  NumWebPurchases      2212 non-null   int64
 13  NumCatalogPurchases  2212 non-null   int64
 14  NumStorePurchases    2212 non-null   int64
 15  NumWebVisitsMonth    2212 non-null   int64
 16  AcceptedCmp3         2212 non-null   int64
 17  AcceptedCmp4         2212 non-null   int64
 18  AcceptedCmp5         2212 non-null   int64
 19  AcceptedCmp1         2212 non-null   int64

show more (open the raw output data in a text editor) ...

 27  Children             2212 non-null   int64
 28  Family_Size          2212 non-null   int64
 29  Is_Parent            2212 non-null   int32
dtypes: float64(1), int32(3), int64(26)
memory usage: 509.8 KB
```

Using the describe method to get measures like mean, median, std etc.

```
data.describe()
```

✓ 0.1s Python

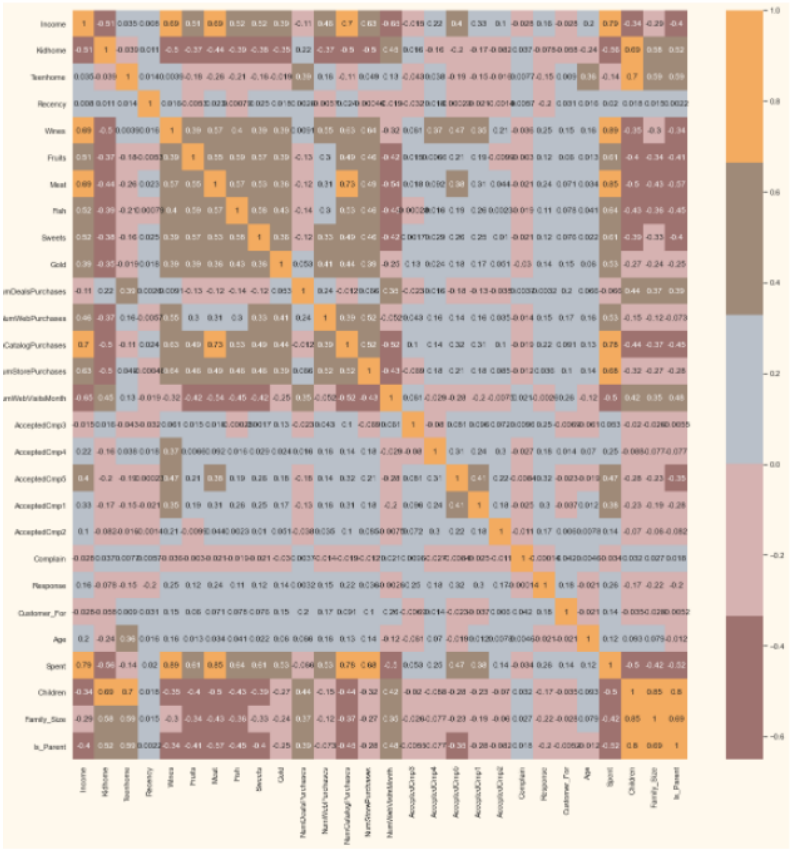
	Income	Kidhome	Teenhome	Recency	Wines	Fruits	Meat	Fish	Sweets	Gold	...	AcceptedCmp1	Accepte
count	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	...	2216.000000	2216.
mean	52247.251354	0.441787	0.505415	49.012635	305.091606	26.356047	166.995939	37.637635	27.028881	43.965253	...	0.064079	0.
std	25173.076661	0.536896	0.544181	28.948352	337.327920	39.793917	224.283273	54.752082	41.072046	51.815414	...	0.244950	0.
min	1730.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.
25%	35303.000000	0.000000	0.000000	24.000000	24.000000	2.000000	16.000000	3.000000	1.000000	9.000000	...	0.000000	0.
50%	51381.500000	0.000000	0.000000	49.000000	174.500000	8.000000	68.000000	12.000000	8.000000	24.500000	...	0.000000	0.
75%	68522.000000	1.000000	1.000000	74.000000	505.000000	33.000000	232.250000	50.000000	33.000000	56.000000	...	0.000000	0.
max	666666.000000	2.000000	2.000000	99.000000	1493.000000	199.000000	1725.000000	259.000000	262.000000	321.000000	...	1.000000	1.

8 rows × 28 columns

The describe method is really useful in learning about the distribution of the dataset. It tells us about the mean, median and standard deviation of the dataset involved. It also gives us information top 25,50,75 percentile

Correlation

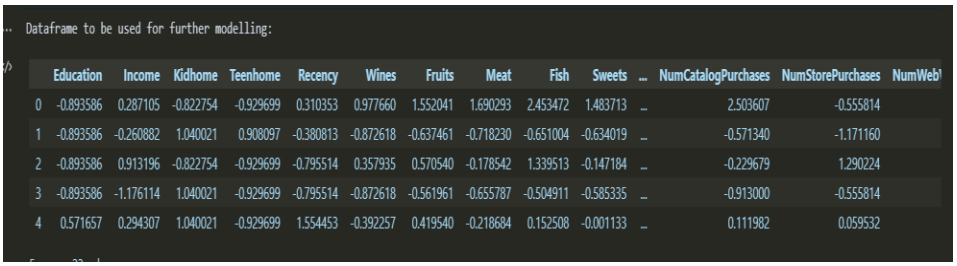
Correlation map helps us analyze the correlation that exists between variables we might be trying to predict. This is the correlation matrix for our dataset. Notice how values are in a range of -1 to 1 which signifies direct or inverse relation between variables



DATA PREPROCESSING

The following steps are applied to preprocess the data:

- Label encoding the categorical features
- Scaling the features using the standard scaler
- Creating a subset dataframe for dimensionality reduction



Dataframe to be used for further modelling:

	Education	Income	Kidhome	Teenhome	Recency	Wines	Fruits	Meat	Fish	Sweets	...	NumCatalogPurchases	NumStorePurchases	NumWeb
0	-0.893586	0.287105	-0.822754	-0.929699	0.310353	0.977660	1.552041	1.690293	2.453472	1.483713	...	2.503607	-0.555814	
1	-0.893586	-0.260882	1.040021	0.908097	-0.380813	-0.872618	-0.637461	-0.718230	-0.651004	-0.634019	...	-0.571340	-1.171160	
2	-0.893586	0.913196	-0.822754	-0.929699	-0.795514	0.357935	0.570540	-0.178542	1.339513	-0.147184	...	-0.229679	1.290224	
3	-0.893586	-1.176114	1.040021	-0.929699	-0.795514	-0.872618	-0.561961	-0.655787	-0.504911	-0.585335	...	-0.913000	-0.555814	
4	0.571657	0.294307	1.040021	-0.929699	1.554453	-0.392257	0.419540	-0.218684	0.152508	-0.001133	...	0.111982	0.059532	

DIMENSIONALITY REDUCTION

There are many factors on the basis of which the final classification will be done. These factors are basically attributes or features. The higher the number of features, the harder it is to work with it. Many of these features are correlated, and hence redundant. This is why we will be performing dimensionality reduction on the selected features before putting them through a classifier.

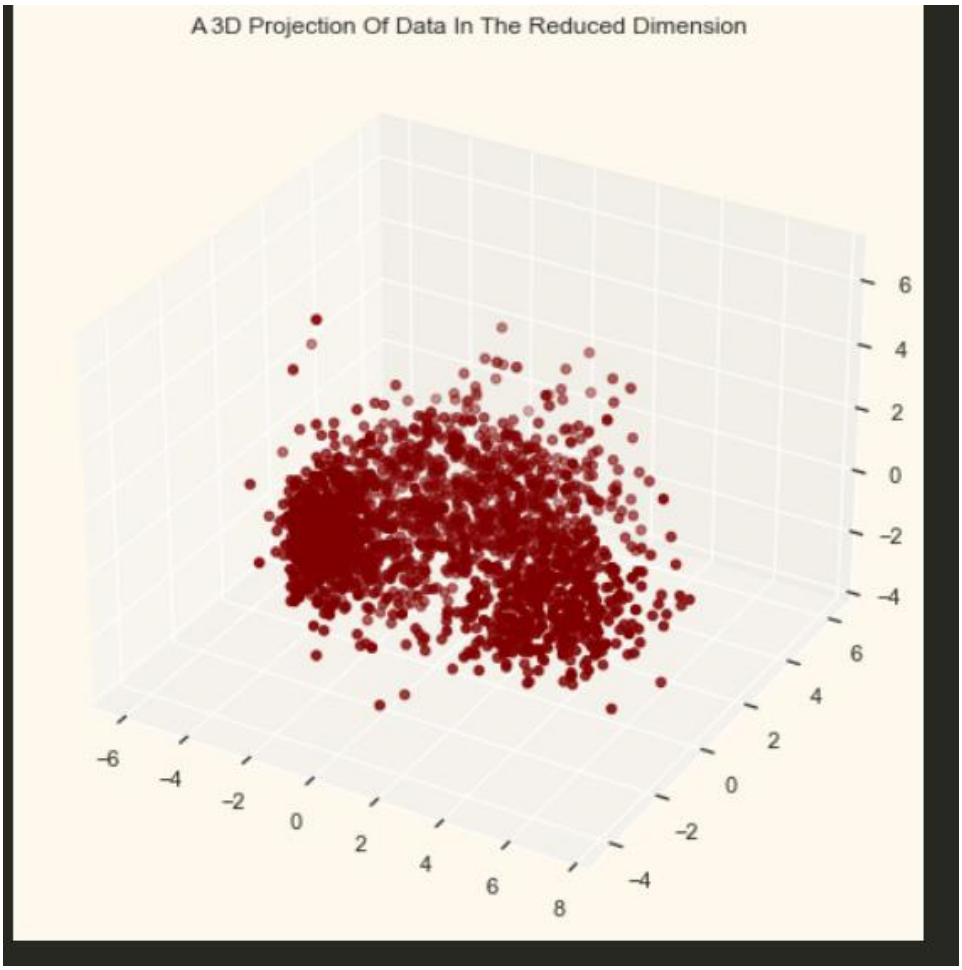
Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.

Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss.

Dimensionality reduction using PCA

	count	mean	std	min	25%	50%	75%	max
col1	2212.0	-1.116246e-16	2.878377	-5.969394	-2.538494	-0.780421	2.383290	7.444305
col2	2212.0	1.105204e-16	1.706839	-4.312196	-1.328316	-0.158123	1.242289	6.142721
col3	2212.0	3.049098e-17	1.221956	-3.530416	-0.829067	-0.022692	0.799895	6.611222

3D Projection Of Data In The Reduced Dimension



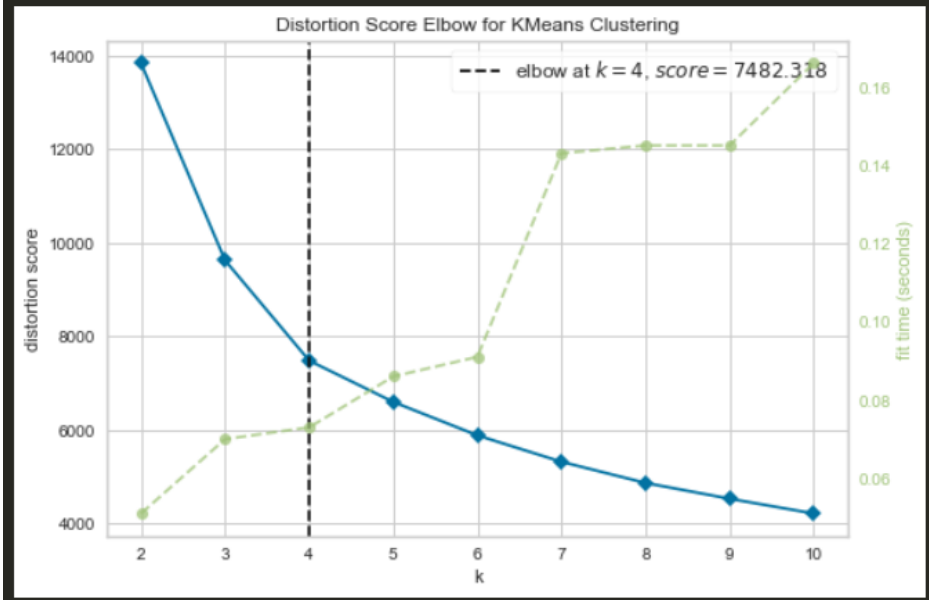
CLUSTERING

Now that we have reduced the attributes to three dimensions, I will be performing clustering via Agglomerative clustering. Agglomerative clustering is a hierarchical clustering method. It involves merging examples until the desired number of clusters is achieved.

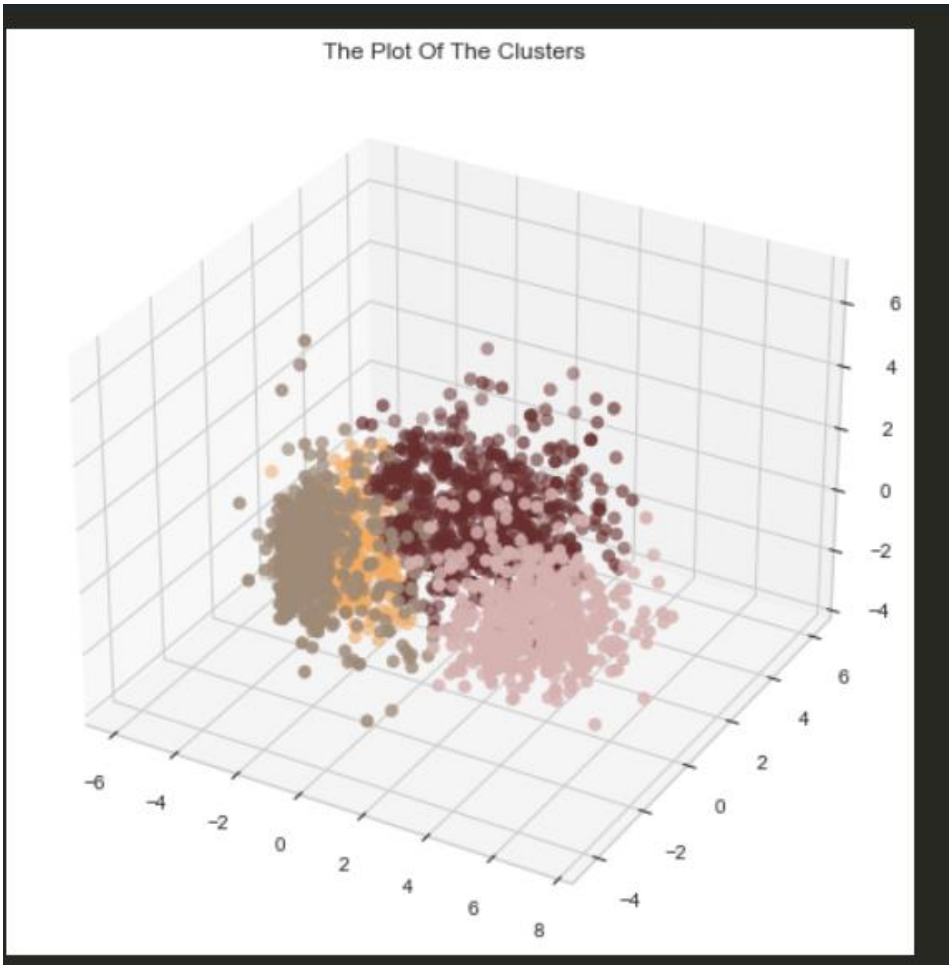
Steps involved in the Clustering

- Elbow Method to determine the number of clusters to be formed
- Clustering via Agglomerative Clustering
- Examining the clusters formed via scatter plot

Elbow Method to determine the number of clusters to be formed:



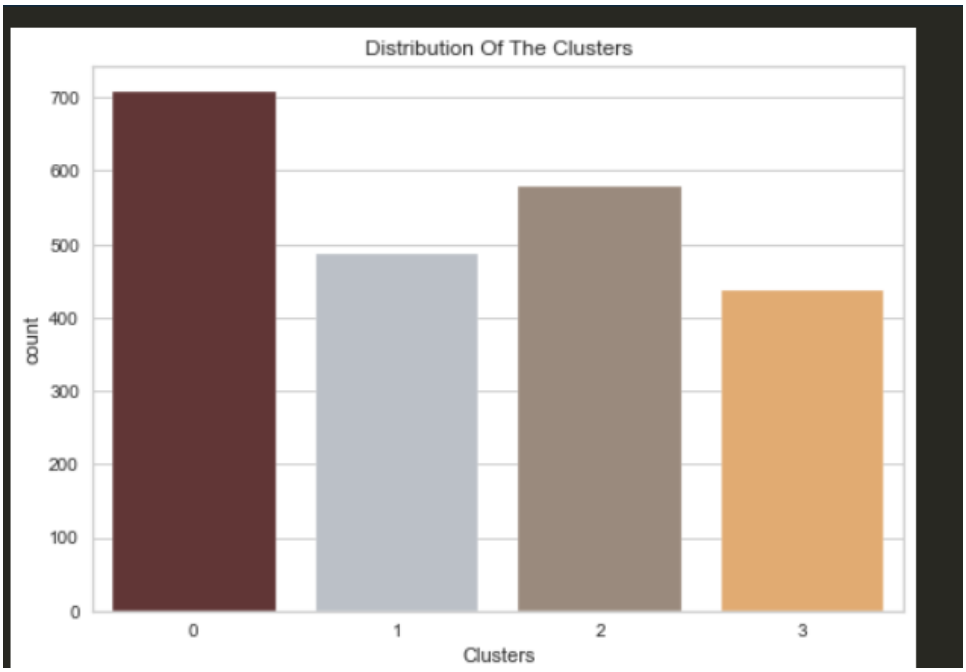
The above cell indicates that four will be an optimal number of clusters for this data. Next, we will be fitting the Agglomerative Clustering Model to get the final clusters.

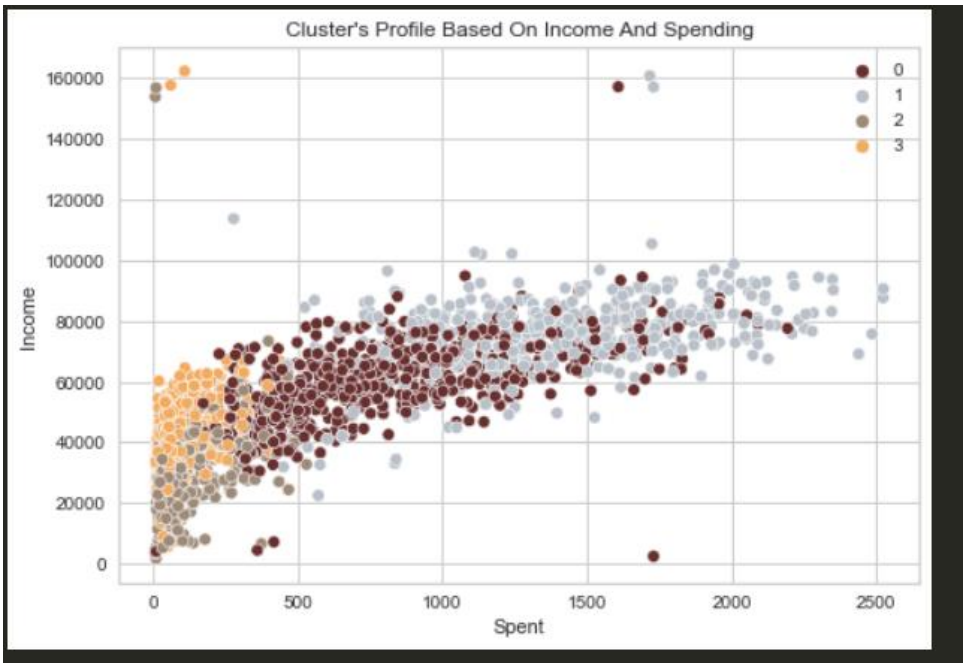


EVALUATING MODELS

Since this is an unsupervised clustering. We do not have a tagged feature to evaluate or score our model. The purpose of this section is to study the patterns in the clusters formed and determine the nature of the clusters' patterns.

For that, we will be having a look at the data in light of clusters via exploratory data analysis and drawing conclusions.

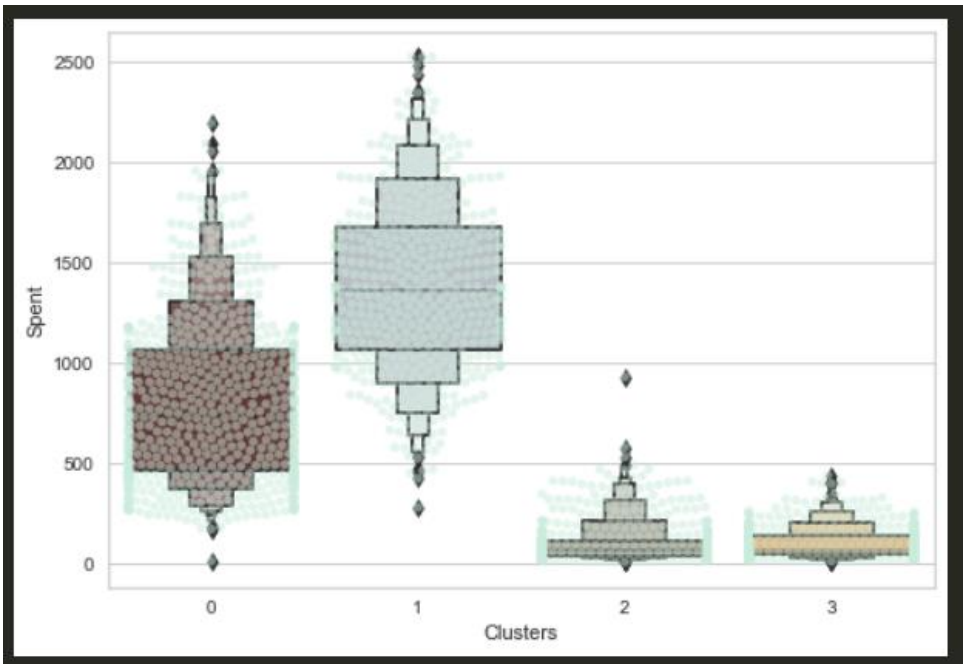




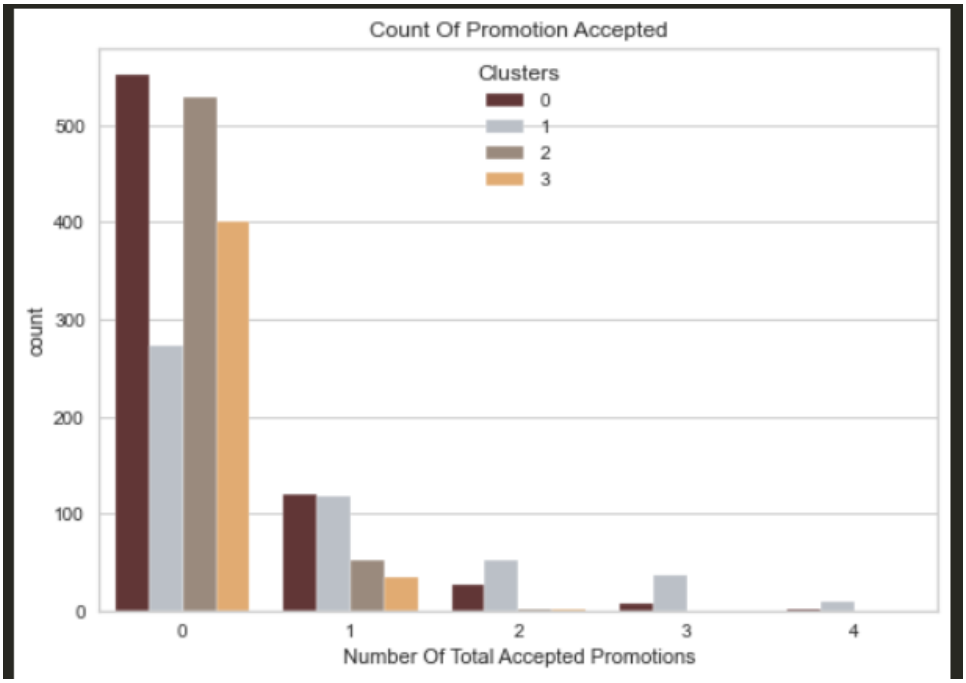
Income vs spending plot shows the clusters pattern

- group 0: high spending & average income
- group 1: high spending & high income
- group 2: low spending & low income
- group 3: high spending & low income

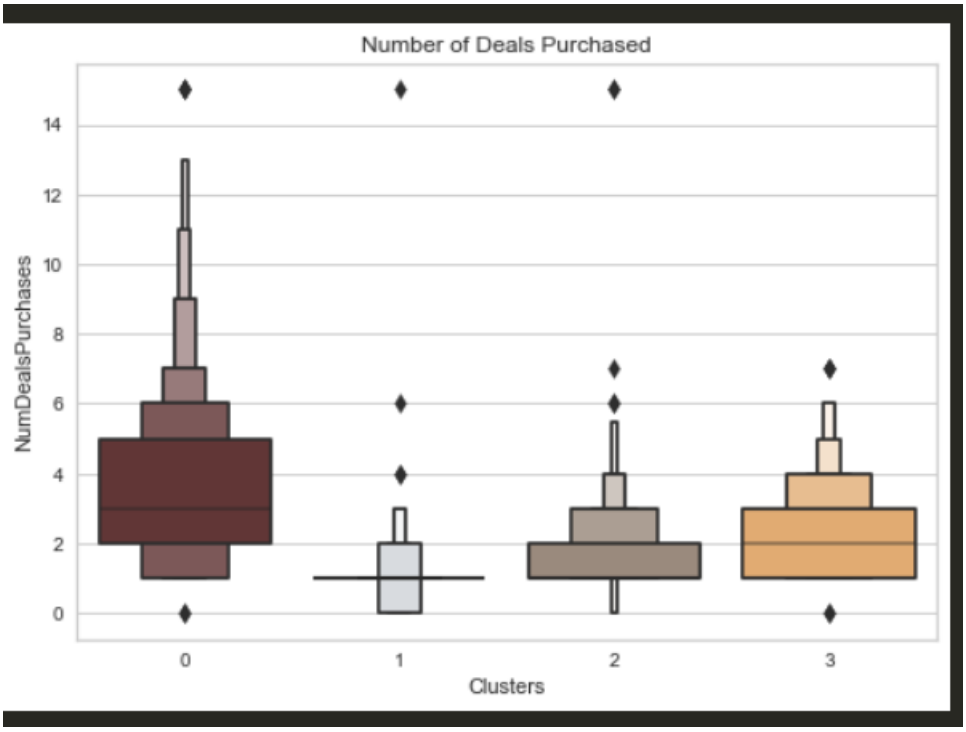
Next, we will be looking at the detailed distribution of clusters as per the various products in the data. Namely: Wines, Fruits, Meat, Fish, Sweets and Gold



From the above plot, it can be clearly seen that cluster 1 is our biggest set of customers closely followed by cluster 0. We can explore what each cluster is spending on for the targeted marketing strategies.



There has not been an overwhelming response to the campaigns so far. Very few participants overall. Moreover, no one part take in all 5 of them. Perhaps better-targeted and well-planned campaigns are required to boost sales.



Conclusion

CONCLUSION

In this project, we performed unsupervised clustering. I did use dimensionality reduction followed by agglomerative clustering. We came up with 4 clusters and further used them in profiling customers in clusters according to their family structures and income/spending. This can be used in planning better marketing strategies

References

Dataset:

https://www.kaggle.com/karnikakapoor/customer-segmentation-clustering/data

Scikit-Learn:

https://scikit-learn.org/

K-means Clustering:

https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

Seaborn:

https://www.section.io/engineering-education/seaborn-tutorial/

Kneed:

https://kneed.readthedocs.io/en/stable/

Python Libraries used:

- Scikit-learn
- Scipy
- NumPy
- Pandas
- Matplotlib
- Seaborn
- Plotly
- Kneed
- Mpltoolkits