BAN 620 Course Project

Clustering for Targeted Marketing Campaigns

Team Members:

- Sukhman Legha
- Adeel Nasir
- Alex Prasadi
- Ashish Tammana
- Krushi Teja Reddy Padamati

PROJECT SUMMARY

The purpose of this project is to create group customers for a store to have targeted marketing campaigns, using data retrieved from Kaggle. We have used Cluster Analysis, Hierarchical Dendrograms, and K-means Clustering models to categorize the customers into groups. Dropping ‘Recency’ variable to reduce the number of variable columns which have little to no impact on the model efficiency. Customers are uniformly distributed across income ranges and each range has a group of around 224 customers.

This project can be utilized in supermarkets, corporates, and every organization to carry on targeted marketing practices. a clustering algorithm can discover groups of objects where the average distances between the members/data points of each cluster are closer than to members/data points in other clusters.

INTRODUCTION

Targeted marketing can help you streamline your processes by identifying areas where you may be wasting time or resources. They can offer suggestions for automating tasks or using technology to improve efficiency.

Data mining technique called clustering involves assembling related data points based on their attributes or characteristics. Finding patterns in data that might not be immediately obvious or visible to the naked eye is the aim of clustering.

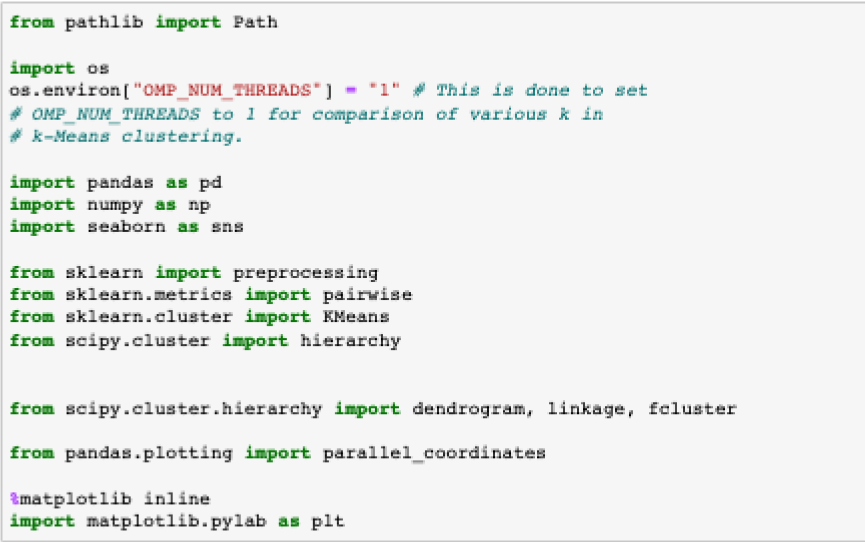
Numerous applications, including market segmentation, customer profiling, anomaly detection, and image analysis, can benefit from clustering. Clustering can assist in revealing hidden patterns and insights in the data that can be used to improve business decisions by identifying groups of related data points.

Clustering is only one of a variety of methods used in data mining and machine learning, and the precise method chosen will depend on the particular issue being resolved and the characteristics of the data being examined.

1. Understanding
   - Our project goal is to help supermarkets develop a targeted marketing strategy based on the previous purchases stored in a CSV file.
   - The ultimate result of this project is to personalize recommendations based on the purchasing history or pattern.
   - This is an ongoing process, with more real-time data available to collect from every purchase. We believe that this will be able to improve our accuracy as it progresses.
2. Obtain Data for Analysis
   - The marketing_data.csv file is retrieved from the kaggle database. The .ipynb file uses the data mining models to create groups for better results.
   - The dataset contains the following columns,
     - Recency: Number of days since customer's last purchase.
     - MntWines: Amount spent on wine in the last 2 years.
     - MntFruits: Amount spent on fruits in last 2 years.
     - MntMeatProducts: Amount spent on meat in the last 2 years.
     - MntFishProducts: Amount spent on fish in the last 2 years.
     - MntSweetProducts: Amount spent on sweets in the last 2 years.

- MntGoldProds: Amount spent on gold in the last 2 years.
- NumDealsPurchases: Number of purchases made with a discount.
- NumWebPurchases: Number of purchases made through the company's website.
- NumCatalogPurchases: Number of purchases made using a catalog.
- NumStorePurchases: Number of purchases made directly in stores.
- NumWebVisitsMonth: Number of visits to the company's website in the last month.

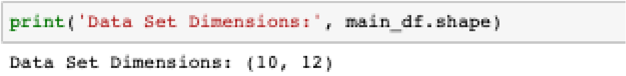
3.1. Download and install the required libraries and packages,



```
from pathlib import Path

import os
os.environ["OMP_NUM_THREADS"] = "1" # This is done to set
# OMP_NUM_THREADS to 1 for comparison of various k in
# k-Means clustering.

import pandas as pd
import numpy as np
import seaborn as sns

from sklearn import preprocessing
from sklearn.metrics import pairwise
from sklearn.cluster import KMeans
from scipy.cluster import hierarchy


from scipy.cluster.hierarchy import dendrogram, linkage, fcluster

from pandas.plotting import parallel_coordinates

%matplotlib inline
import matplotlib.pylab as plt
```

3.2. Explore, Clean and Preprocess Data; Reduce Data Dimension

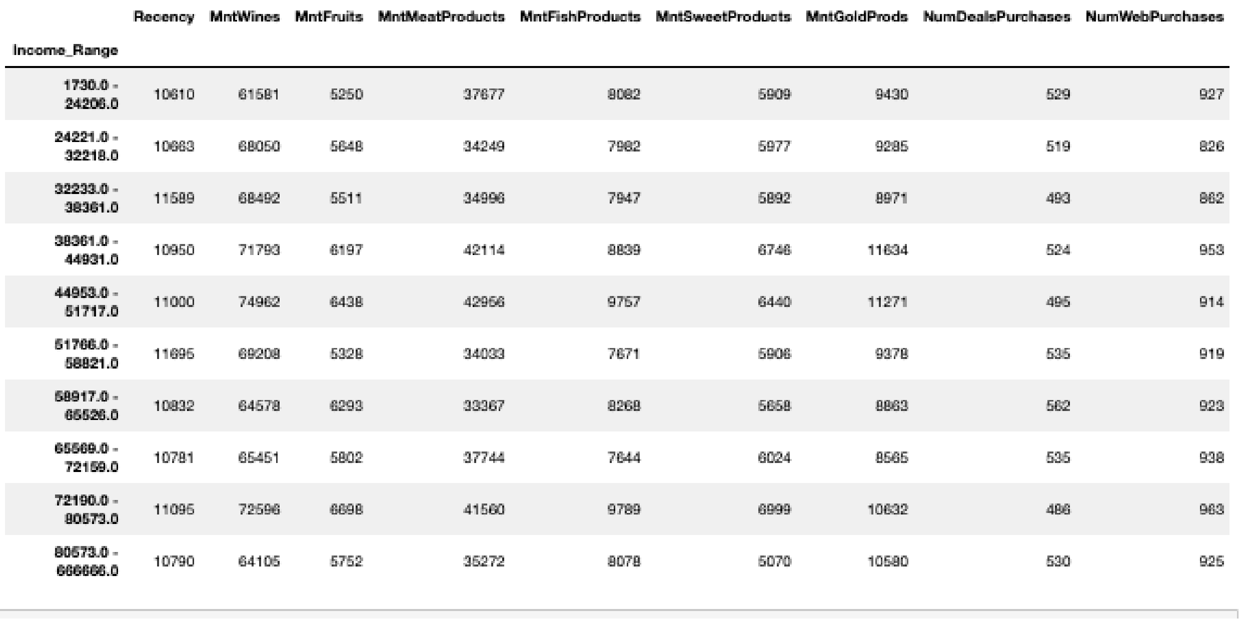
- The data used for this project has 10 rows and 12 columns.



```
print('Data Set Dimensions:', main_df.shape)
```

```
Data Set Dimensions: (10, 12)
```

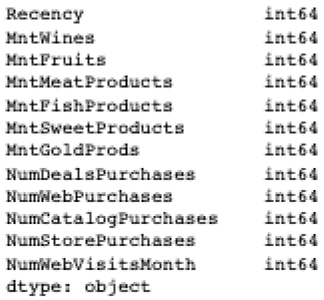
- The following are the columns (Predictors),



	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases
Income_Range									
1730.0 - 24206.0	10610	61581	5250	37677	8082	5909	9430	529	927
24221.0 - 32218.0	10683	68050	5648	34249	7982	5977	9285	519	826
32233.0 - 38361.0	11589	68492	5511	34996	7947	5892	8971	493	862
38361.0 - 44931.0	10950	71793	6197	42114	8839	6746	11634	524	953
44953.0 - 51717.0	11000	74962	6438	42956	9757	6440	11271	495	914
51766.0 - 58821.0	11695	69208	5328	34033	7671	5906	9378	535	919
58917.0 - 65526.0	10832	64578	6293	33367	8268	5658	8863	562	923
65569.0 - 72159.0	10781	65451	5802	37744	7644	6024	8565	535	938
72190.0 - 80573.0	11095	72596	6698	41560	9789	6999	10632	486	963
80573.0 - 666666.0	10790	64105	5752	35272	8078	5070	10580	530	926

  - These are the 12 columns and 10 rows dataset that we will use in this project.

- Here are the datatypes for the data used in this project,

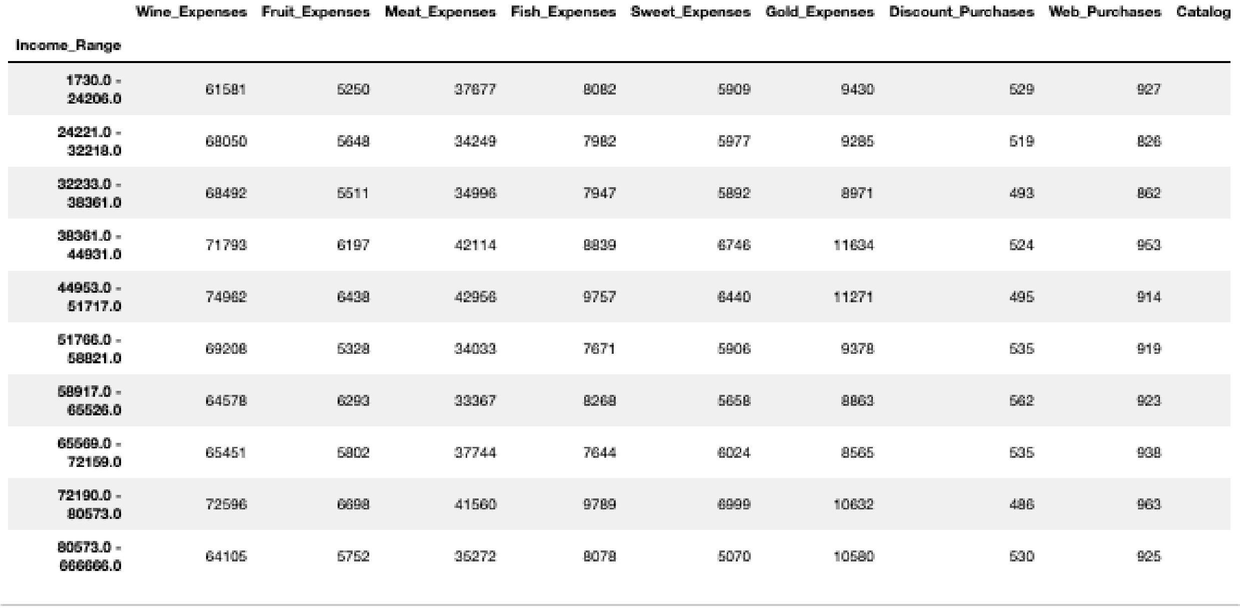


```
Recency                 int64
MntWines                int64
MntFruits               int64
MntMeatProducts         int64
MntFishProducts         int64
MntSweetProducts        int64
MntGoldProds            int64
NumDealsPurchases       int64
NumWebPurchases         int64
NumCatalogPurchases     int64
NumStorePurchases       int64
NumWebVisitsMonth       int64
dtype: object
```

  - As we can see all the variable data types are int64.

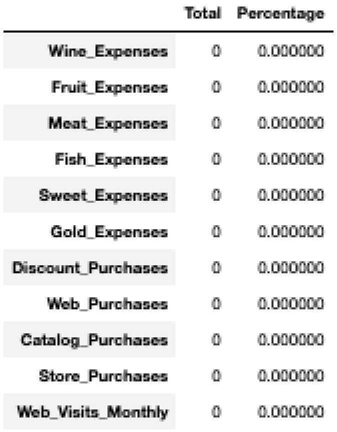
- After column names conversion

  - By doing the column name conversion we have simplified the dataset for understanding and future use to add more data points.



	Wine_Expenses	Fruit_Expenses	Meat_Expenses	Fish_Expenses	Sweet_Expenses	Gold_Expenses	Discount_Purchases	Web_Purchases	Catalog
Income_Range									
1730.0 - 24206.0	61581	5250	37677	8082	5909	9430	529	927	
24221.0 - 32218.0	68050	5648	34249	7982	5977	9285	519	826	
32233.0 - 38361.0	68492	5511	34996	7947	5892	8971	493	862	
38361.0 - 44931.0	71793	6197	42114	8839	6746	11634	524	953	
44953.0 - 51717.0	74962	6438	42956	9757	6440	11271	495	914	
51766.0 - 58821.0	69208	5328	34033	7671	5906	9378	535	919	
58917.0 - 65526.0	64578	6293	33367	8268	5658	8863	562	923	
65569.0 - 72159.0	65451	5802	37744	7644	6024	8565	535	938	
72190.0 - 80573.0	72596	6698	41560	9789	6999	10632	486	963	
80573.0 - 666666.0	64105	5752	35272	8078	5070	10580	530	925	

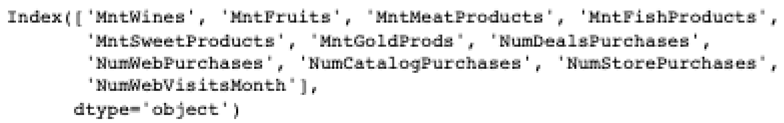
- Cleaning the data by checking for missing data values,



	Total	Percentage
Wine_Expenses	0	0.000000
Fruit_Expenses	0	0.000000
Meat_Expenses	0	0.000000
Fish_Expenses	0	0.000000
Sweet_Expenses	0	0.000000
Gold_Expenses	0	0.000000
Discount_Purchases	0	0.000000
Web_Purchases	0	0.000000
Catalog_Purchases	0	0.000000
Store_Purchases	0	0.000000
Web_Visits_Monthly	0	0.000000

    - To eliminate any errors in the calculation or future confusions it is important to clean the data of any missing columns or rows as this might have a negative impact in the future.

- Remove ‘Recency’ variables from the main_df data frame,

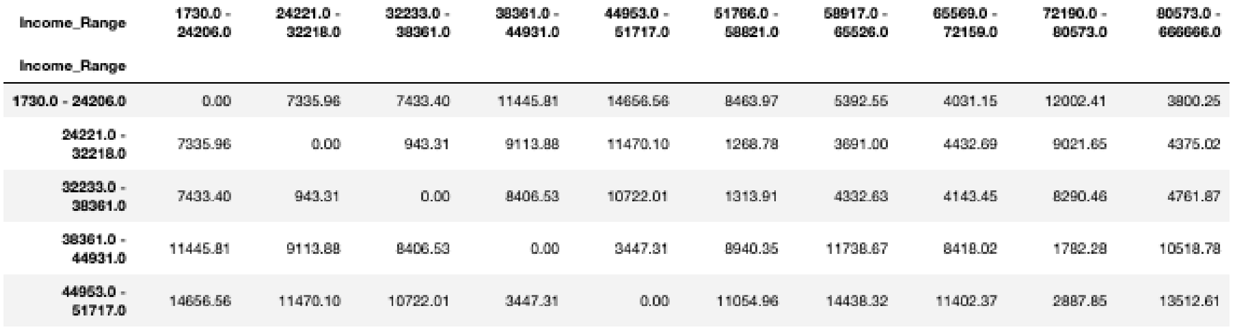


```
Index(['MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts',
       'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases',
       'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases',
       'NumWebVisitsMonth'],
      dtype='object')
```

- We remove the 'Recency' variable from the dataset because the categorical variable in our dataset(the income range) is for a group of customers (224 each) and other variables shown in the dataset are aggregate of these values. An aggregate of recency for a group of customers doesn't make sense.

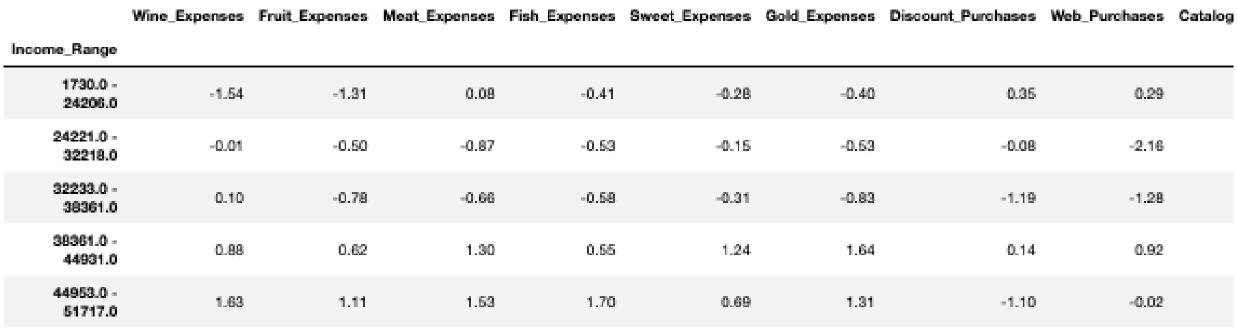
3.2. Determine the Data Mining Tasks

- Firstly, compute the distance matrix
  - A distance matrix is a nonnegative, square, symmetric matrix with elements corresponding to estimates of some pairwise distance between the sequences in a set.
  - Distance matrices became heavily dependent and utilized in cluster analysis since similarity can be measured with a distance metric.



Income_Range	1730.0 - 24206.0	24221.0 - 32218.0	32233.0 - 38361.0	38361.0 - 44931.0	44953.0 - 51717.0	51766.0 - 58821.0	58917.0 - 65526.0	65569.0 - 72159.0	72190.0 - 80573.0	80573.0 - 666666.0
Income_Range										
1730.0 - 24206.0	0.00	7335.96	7433.40	11445.81	14656.56	8463.97	5392.55	4031.15	12002.41	3800.25
24221.0 - 32218.0	7335.96	0.00	943.31	9113.88	11470.10	1268.78	3691.00	4432.69	9021.65	4375.02
32233.0 - 38361.0	7433.40	943.31	0.00	8406.53	10722.01	1313.91	4332.63	4143.45	8290.46	4761.87
38361.0 - 44931.0	11445.81	9113.88	8406.53	0.00	3447.31	8940.35	11738.67	8418.02	1782.28	10518.78
44953.0 - 51717.0	14656.56	11470.10	10722.01	3447.31	0.00	11054.96	14438.32	11402.37	2887.85	13512.61

- Normalize the input variables and compute the distance matrix again.



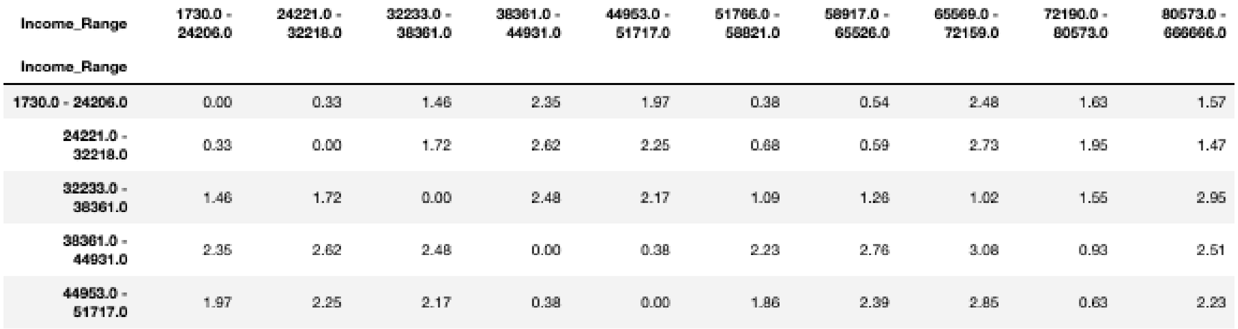
	Wine_Expenses	Fruit_Expenses	Meat_Expenses	Fish_Expenses	Sweet_Expenses	Gold_Expenses	Discount_Purchases	Web_Purchases	Catalog
Income_Range									
1730.0 - 24206.0	-1.54	-1.31	0.08	-0.41	-0.28	-0.40	0.35	0.29	
24221.0 - 32218.0	-0.01	-0.50	-0.87	-0.53	-0.15	-0.53	-0.08	-2.16	
32233.0 - 38361.0	0.10	-0.78	-0.66	-0.58	-0.31	-0.83	-1.19	-1.28	
38361.0 - 44931.0	0.88	0.62	1.30	0.55	1.24	1.64	0.14	0.92	
44953.0 - 51717.0	1.63	1.11	1.53	1.70	0.69	1.31	-1.10	-0.02	

- We normalize the data to eliminate various scales of data in the dataset.
- Having different scales can affect the output values.
- Hence, normalizing the column values will standardize the data and eliminate the effects of the different scale values.
- The normalized values are calculated as Z-score values.
- Z-score is the number of standard deviations of a column value from the mean.
- For a numeric value *xj* in a column *j* in the training partition, the standard score or scaled value *Zj* is calculated as:

$$Zj = (xj - Uj)/Sj$$

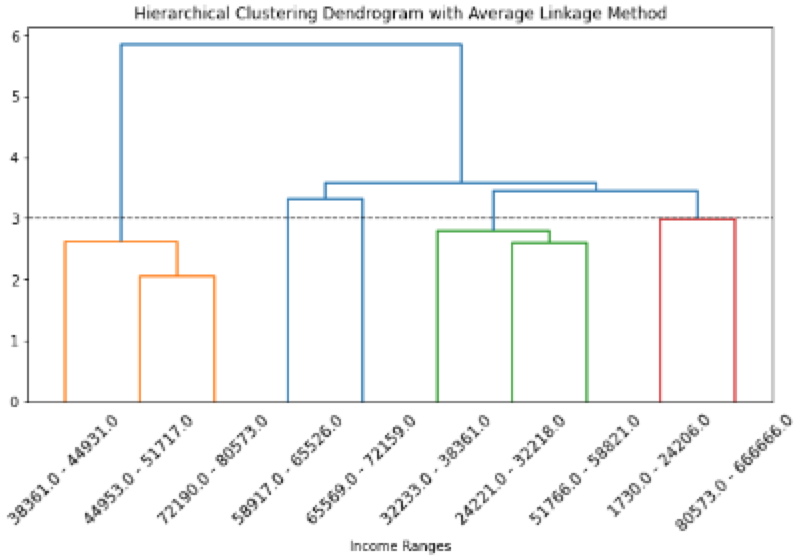
- Where: *Uj* = mean of values in column *j* of training partition,
- *Sj* = standard deviation of values in column j of training partition.

- Normalized distance matrix between income ranges.
    - Here we have computed the normalized distance matrix for income ranges based on 'Gold_Expenses' and 'Store_Purchases' variables.
    - This matrix is developed to understand the relation between ‘Gold_Expenses' and 'Store_Purchases' variables.



Income_Range	1730.0 - 24206.0	24221.0 - 32218.0	32233.0 - 38361.0	38361.0 - 44931.0	44953.0 - 51717.0	51766.0 - 58821.0	58917.0 - 65526.0	65569.0 - 72159.0	72190.0 - 80573.0	80573.0 - 666666.0
Income_Range										
1730.0 - 24206.0	0.00	0.33	1.46	2.35	1.97	0.38	0.54	2.48	1.63	1.57
24221.0 - 32218.0	0.33	0.00	1.72	2.62	2.25	0.68	0.59	2.73	1.95	1.47
32233.0 - 38361.0	1.46	1.72	0.00	2.48	2.17	1.09	1.26	1.02	1.55	2.95
38361.0 - 44931.0	2.35	2.62	2.48	0.00	0.38	2.23	2.76	3.08	0.93	2.51
44953.0 - 51717.0	1.97	2.25	2.17	0.38	0.00	1.86	2.39	2.85	0.63	2.23

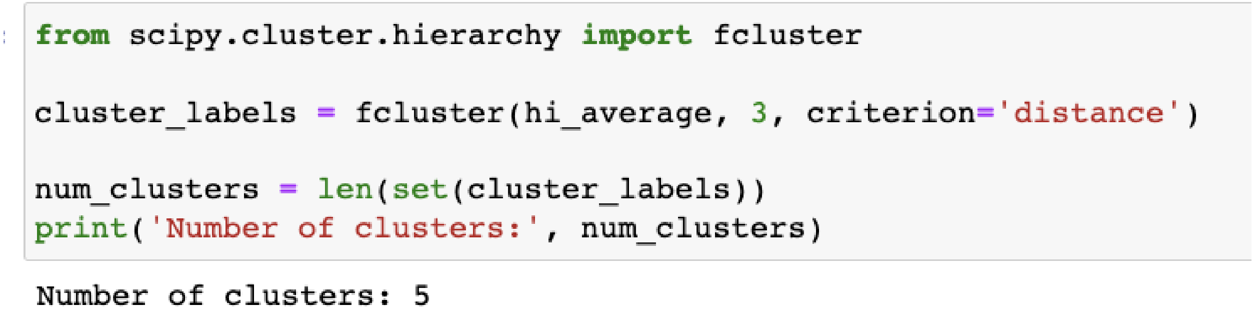
- Plotting hierarchical dendrograms for different measures of distance between clusters.



- Creating clusters based on the distance matrix with a color threshold of 3 which is indicated by the dashed line in the plot.
- The `color_threshold` is a setting used to color the links and nodes in a dendrogram, which is a type of tree diagram used in hierarchical clustering. The purpose of this setting is to help simplify the visualization of the dendrogram by grouping clusters that are similar in distance.

- When the `color_threshold` is set, all the links below that threshold will be colored the same, which means they belong to the same group. Links above the threshold will be colored with a default color, indicating that they belong to different groups. If the `color_threshold` is not specified, it will be set automatically to a default value of 0.7 times the maximum distance between clusters. If the `color_threshold` is set to a value less than or equal to zero, all the nodes and links will be colored with the default color.

- Number of clusters in the above dendrogram

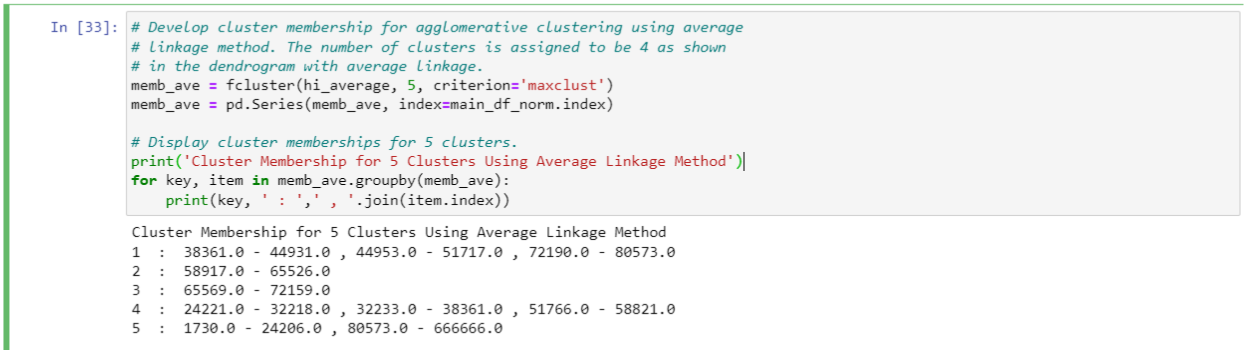


```
from scipy.cluster.hierarchy import fcluster

cluster_labels = fcluster(hi_average, 3, criterion='distance')

num_clusters = len(set(cluster_labels))
print('Number of clusters:', num_clusters)
```

```
Number of clusters: 5
```

- According to the dendrogram we have created 5 clusters for this dataset.

- Creating Cluster membership with average linkage,

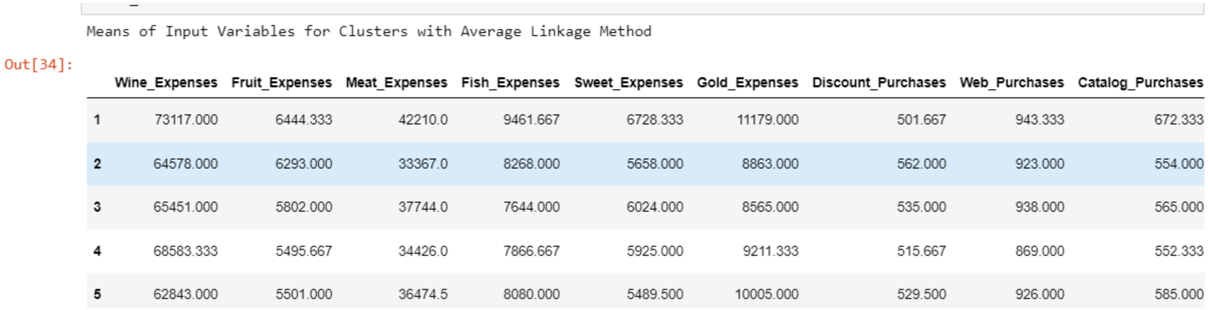


```
In [33]: # Develop cluster membership for agglomerative clustering using average
         # linkage method. The number of clusters is assigned to be 4 as shown
         # in the dendrogram with average linkage.
         memb_ave = fcluster(hi_average, 5, criterion='maxclust')
         memb_ave = pd.Series(memb_ave, index=main_df_norm.index)

         # Display cluster memberships for 5 clusters.
         print('Cluster Membership for 5 Clusters Using Average Linkage Method')
         for key, item in memb_ave.groupby(memb_ave):
             print(key, ' : ',' , '.join(item.index))
```

```
Cluster Membership for 5 Clusters Using Average Linkage Method
1  :  38361.0 - 44931.0 , 44953.0 - 51717.0 , 72190.0 - 80573.0
2  :  58917.0 - 65526.0
3  :  65569.0 - 72159.0
4  :  24221.0 - 32218.0 , 32233.0 - 38361.0 , 51766.0 - 58821.0
5  :  1730.0 - 24206.0 , 80573.0 - 666666.0
```

- Cluster membership refers to the grouping of data points based on their similarities. The process involves assigning each data point to a specific cluster based on its characteristics and proximity to other data points in that cluster.
- It helps to organize complex data sets and provides insights into the characteristics of each cluster. By understanding which data points belong to which group, it becomes easier to identify patterns, compare different clustering algorithms, and make more informed decisions based on the data.

● Means of Input Variables for Clusters with Average Linkage Method



```
Means of Input Variables for Clusters with Average Linkage Method
```

Out[34]:

	Wine_Expenses	Fruit_Expenses	Meat_Expenses	Fish_Expenses	Sweet_Expenses	Gold_Expenses	Discount_Purchases	Web_Purchases	Catalog_Purchases
1	73117.000	6444.333	42210.0	9461.667	6728.333	11179.000	501.667	943.333	672.333
2	64578.000	6293.000	33367.0	8268.000	5658.000	8863.000	562.000	923.000	554.000
3	65451.000	5802.000	37744.0	7644.000	6024.000	8565.000	535.000	938.000	565.000
4	68583.333	5495.667	34426.0	7866.667	5925.000	9211.333	515.667	869.000	552.333
5	62843.000	5501.000	36474.5	8080.000	5489.500	10005.000	529.500	926.000	585.000

1. Cluster 1: In comparison to the other clusters, this cluster has the highest average values for Wine_Expenses, Meat_Expenses, Fish_Expenses, Sweet_Expenses, Gold_Expenses, Discount_Purchases, Web_Purchases, Catalog_Purchases, Store_Purchases, and Web_Visits_Monthly. This shows that customers in this cluster are more likely to spend more across all categories and to make purchases through a variety of channels (web, catalog, store). Additionally, they appear to be more likely to purchase expensive goods like wine and gold.
2. Cluster 2: For Wine_Expenses, Meat_Expenses, Fish_Expenses, Sweet_Expenses, and Gold_Expenses, this cluster has the second-highest average values. In comparison to Clusters 3 and 5, they also frequently have higher values for Discount_Purchases,

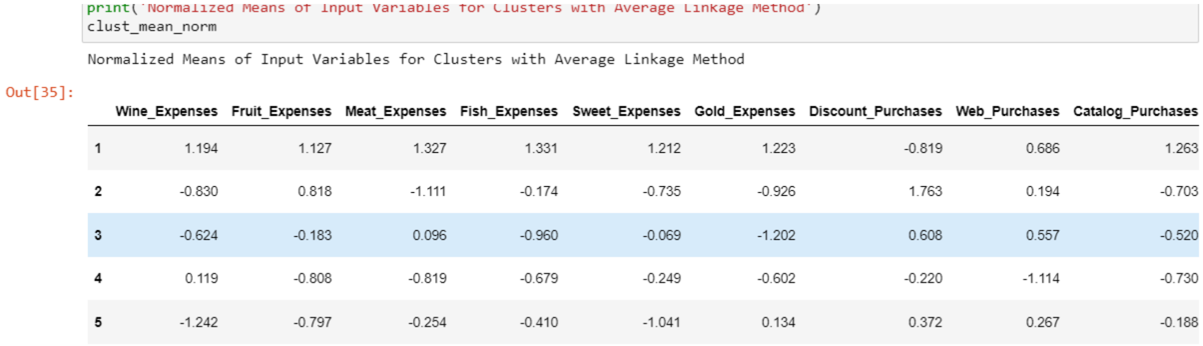
Catalog_Purchases, and Store_Purchases, but lower values for Cluster 1. This suggests that although they do spend more on luxury goods than Cluster 1, it is not as much. They use several channels for purchases as well, but not as many as Cluster 1.When compared to the other clusters,

3. Cluster 3 has the second-lowest average values for Wine_Expenses, Meat_Expenses, Fish_Expenses, Sweet_Expenses, and Gold_Expenses. Additionally, they frequently have the lowest Discount_Purchases, Catalog_Purchases, and Store_Purchases values. This suggests that they generally spend less and are less likely to buy expensive items or through multiple channels.
4. Cluster 4: In comparison to the other clusters, this cluster has the highest average value for Wine_Expenses. In comparison to Cluster 3, they also frequently have higher Meat_Expenses, Fish_Expenses, and Gold_Expenses values. In contrast to Clusters 1 and 2, they have lower values for Discount_Purchases, Catalog_Purchases, and Store_Purchases. This suggests that they may not shop through multiple channels as frequently because they tend to spend more on expensive wines and other items.
5. In comparison to the other clusters, Cluster 5 has the lowest average values for Wine_Expenses, Meat_Expenses, Fish_Expenses, Sweet_Expenses, and Gold_Expenses. In comparison to Clusters 1 and 2, they frequently have lower values for Discount_Purchases, Catalog_Purchases, and Store_Purchases but higher values for Cluster 3. This shows that they may spend more frequently than Cluster 3 but tend to spend less overall and are less likely to buy expensive items or through multiple channels.

Overall, it appears that the clusters are organized according to spending and shopping habits. While Clusters 3, 4, and 5 exhibit lower spending and less varied purchase behavior, Clusters 1 and 2 exhibit higher spending and more varied purchase behavior. While Cluster 5 stands out for making more frequent purchases despite having lower overall spending, Cluster 4 stands out for having high wine expenses, they are more likely to respond to any wine marketing related campaigns as well as clusters 3,4 and 5 less likely to respond to discounts offered by the store . All advertisements are likely to be targeted towards Clusters 1 and 2 as they have the highest likelihood of spending larger amounts on luxury goods.

 Since these two clusters (Clusters 1 and 2) are also actively responding to the discounts offered by the store,more deals can be thrown on these clusters to have better purchases.

- Normalized Means of Input Variables for Clusters with Average Linkage Method



```
print('Normalized Means of Input Variables for Clusters with Average Linkage Method')
clust_mean_norm
```

```
Normalized Means of Input Variables for Clusters with Average Linkage Method
```

Out[35]:

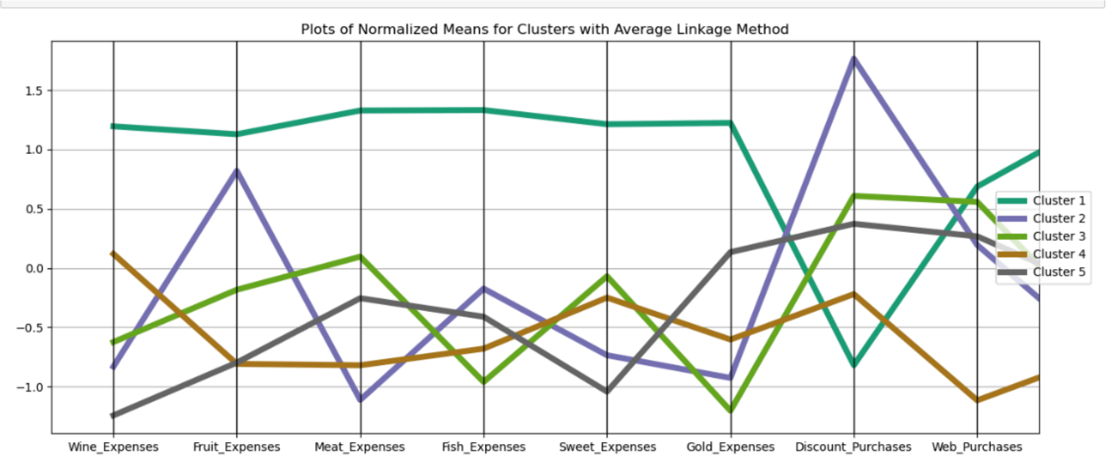
	Wine_Expenses	Fruit_Expenses	Meat_Expenses	Fish_Expenses	Sweet_Expenses	Gold_Expenses	Discount_Purchases	Web_Purchases	Catalog_Purchases
1	1.194	1.127	1.327	1.331	1.212	1.223	-0.819	0.686	1.263
2	-0.830	0.818	-1.111	-0.174	-0.735	-0.926	1.763	0.194	-0.703
3	-0.624	-0.183	0.096	-0.960	-0.069	-1.202	0.608	0.557	-0.520
4	0.119	-0.808	-0.819	-0.679	-0.249	-0.602	-0.220	-1.114	-0.730
5	-1.242	-0.797	-0.254	-0.410	-1.041	0.134	0.372	0.267	-0.188

We can see that: based on the normalized means of the input variables.

- The most money is spent in the first cluster (Cluster 1) across all categories, especially on wine and fish. This cluster also has a propensity for online and catalog shopping, but not as much for in-store visits.

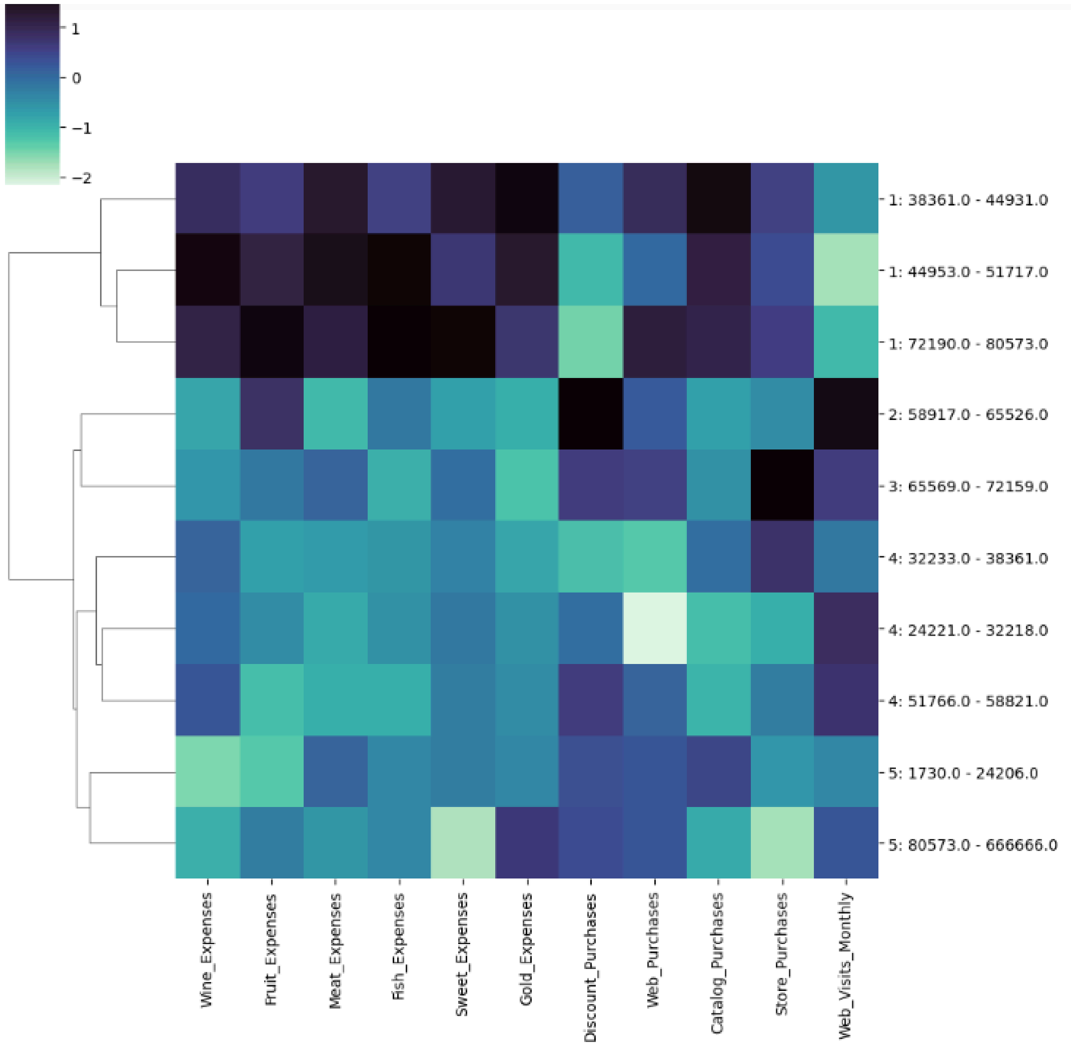
- Cluster 2 has the most purchases made through the web channel and spends the most on gold expenses. Additionally, they tend to make fewer in-store visits and buy less meat and fish.
- The least money is spent on fruit and sweets by Cluster 3, but they tend to use the store channel more frequently.
- Comparatively to the other clusters, Cluster 4 spends less money overall. They tend to visit stores less frequently and make more purchases online and through catalogs,campaigns.
- While spending on wine and sweets is generally lower in Cluster 5, spending on gold is generally higher. Additionally, they frequent stores more frequently and make fewer online purchases.

● Plots of Normalized Means for Clusters with Average Linkage Method,



- This plot is useful to visually understand if there are any relations between the clusters.
- From the Dendrogram We label the clusters

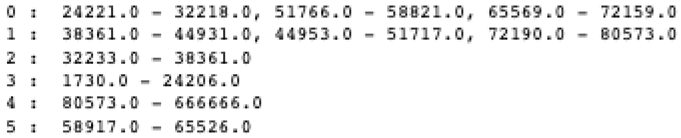
● Heatmap for Marketing Hierarchical Clustering with Average Linkage Method,



- In a heatmap, each row and column represents a variable, and the cells represent the values of that variable. The colors of the cells represent the magnitude of the values, with higher values typically represented by brighter or darker colors.
- To sum it up:

- Cluster 1 has Spending which is highest overall, with a focus on wine and fish. Online and catalog shopping is more common, but store visits are declining.Cluster 2 shows The majority of purchases are made online, and gold expenses are significant. less money spent on meat and fish, as well as fewer store visits. From Cluster 3 we can interpret that it spends low on fruit and sweets spending but high store visits. Cluster 4 has Lower overall spending, fewer store visits, and more online and catalog purchases.Cluster 5 shows Less money spent on alcohol and sweets, but more money spent on gold. Frequently visiting stores, less online shopping.

● Creating k-Means clustering of Utilities records into 6 clusters.



```
0 :  24221.0 - 32218.0, 51766.0 - 58821.0, 65569.0 - 72159.0
1 :  38361.0 - 44931.0, 44953.0 - 51717.0, 72190.0 - 80573.0
2 :  32233.0 - 38361.0
3 :  1730.0 - 24206.0
4 :  80573.0 - 666666.0
5 :  58917.0 - 65526.0
```

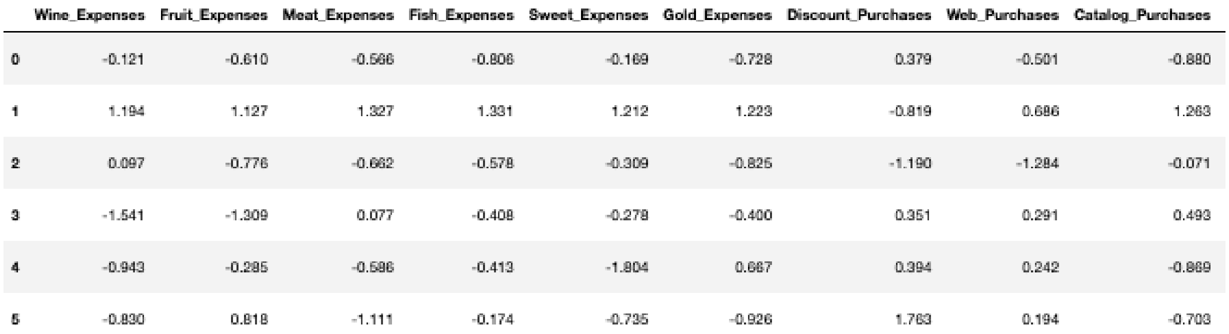
- The two main differences between the hierarchical and k-means clustering are,

- In k-means clustering we need to specify the number of clusters before running the algorithm, but in hierarchical clustering the number of clusters is not specified.

- K-means clustering uses cluster centroids to measure the distance between clusters.

- Hierarchical clustering can use different measures to calculate the distance between clusters like single, average, etc.

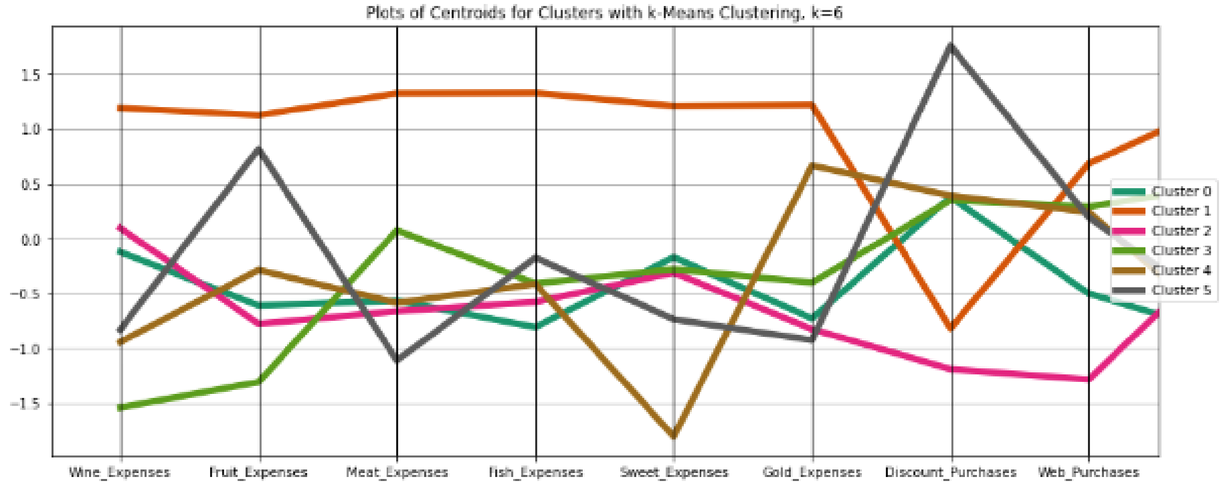
- Cluster centroids



	Wine_Expenses	Fruit_Expenses	Meat_Expenses	Fish_Expenses	Sweet_Expenses	Gold_Expenses	Discount_Purchases	Web_Purchases	Catalog_Purchases
0	-0.121	-0.610	-0.566	-0.806	-0.169	-0.728	0.379	-0.501	-0.880
1	1.194	1.127	1.327	1.331	1.212	1.223	-0.819	0.686	1.263
2	0.097	-0.776	-0.662	-0.578	-0.309	-0.825	-1.190	-1.284	-0.071
3	-1.541	-1.309	0.077	-0.408	-0.278	-0.400	0.351	0.291	0.493
4	-0.943	-0.285	-0.586	-0.413	-1.804	0.667	0.394	0.242	-0.869
5	-0.830	0.818	-1.111	-0.174	-0.735	-0.926	1.763	0.194	-0.703

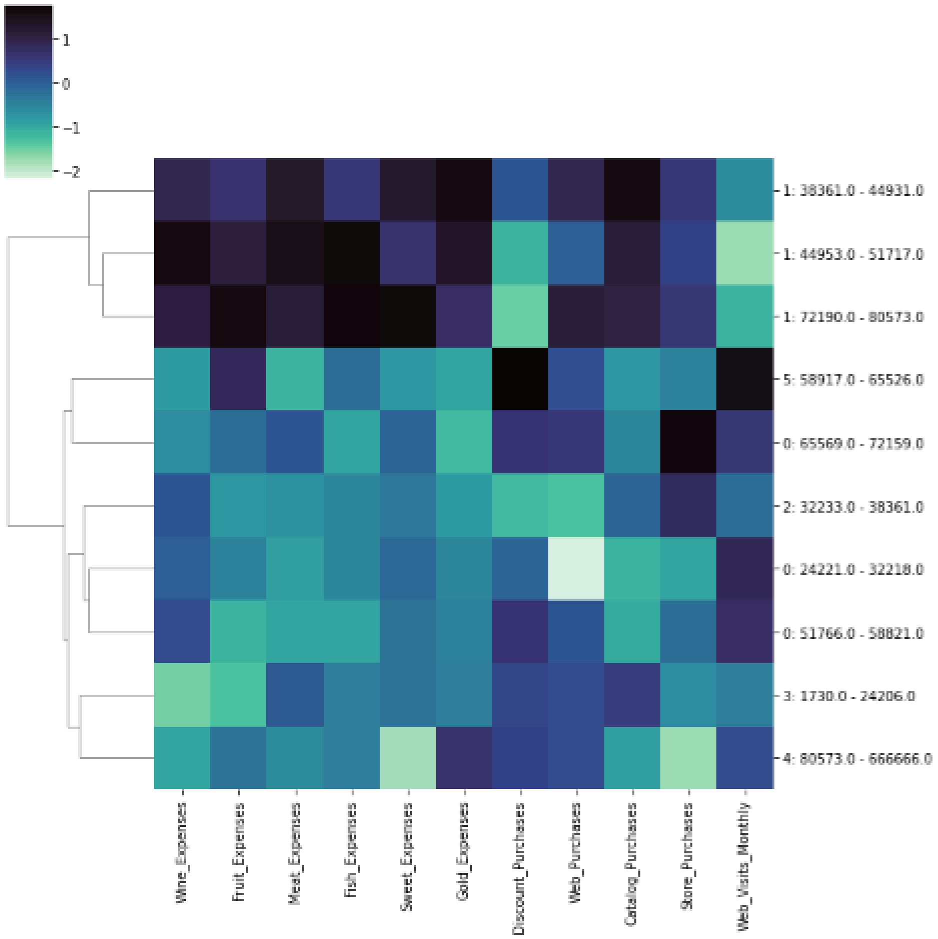
  - Cluster centroids are center points of clusters in clustering analysis, representing the average of all data points in the cluster. They are used to characterize clusters, identify outliers, and detect new data points belonging to the same cluster.

- Profile plots of centroids



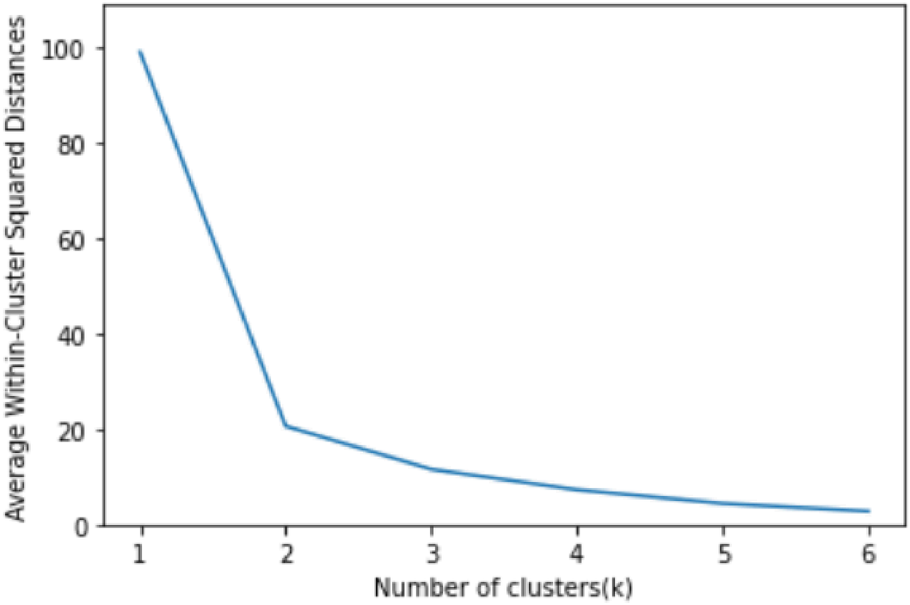
- This is the plot diagram for k-means based on the centroids for clusters with k = 6.
- As shown in the profile plot we can interpret that Cluster 1 is more likely to spend purchases in the store including luxury products like wine and Gold and they are less responsive to the discounts offered by the store.
- It can be seen in Cluster 5 that they will be more likely to spend money on fruits and might avoid spending money on luxury items like wine and gold. Also, this cluster is most responsive to discounts offered by the store and are likely to make purchases when a reduced price is offered.
- Cluster 4 is highly responsive to the marketing campaigns for Gold and less likely to spend money on sweets. They are also fairly responsive to store discounts.
- The remaining three clusters (0, 2 and 3), have similar purchase patterns and are moderately responsive when discounts are offered.

Heatmap for centroids for clusters with k = 6.



Each cluster appears to have distinct spending habits and receptivity to promotional offers and discounts. High-end consumers in Cluster 1 appear to be less price-sensitive and more likely to make in-store purchases. Cluster 5 is more cost- and discount-conscious and concerned with their health. Luxury consumers in Cluster 4 appear to be open to marketing initiatives. Clusters 0, 2, and 3 have comparable buying habits and a moderate sensitivity to discounts.

Elbow Chart



- An elbow chart is a visual tool used in clustering to identify the optimal number of clusters. It plots the number of clusters against the within-cluster sum of squares, and the elbow point represents the optimal number of clusters. This helps to balance minimizing within-cluster sum of squares and avoiding overfitting.
- K=5 would be appropriate because there is less improvement to cluster homogeneity. As it is almost a flat line after 4 so it will be easier to explain and interpret, as the level of homogeneity is good and that will be the number of clusters that we may consider any number greater than k=5 can a better results which would be appropriate number of clusters in k-means clustering of the marketing data.

4. Conclusion

Hierarchical Clustering

Using Python, we were able to arrange the data into a tree-like structure using the hierarchical clustering approach, and new clusters are created by joining or dividing old ones. By observing the heights at which each branch connects, the dendrogram gives us a visual representation of the clustering process and allows for the ideal number of clusters to be calculated.

The average linkage method was used to perform hierarchical clustering on the input data. Cluster 1, Cluster 2, Cluster 3, Cluster 4, and Cluster 5 were the names given to the resulting clusters. Spending patterns, preferred channels, and receptivity to discounts vary depending on the cluster. For instance, Cluster 1 spends the most overall and utilizes a variety of channels, whereas Cluster 3 spends less overall and favors in-store purchases.

K-means clustering

K-means clustering works by iteratively assigning data points to the cluster whose centroid is closest to them in order to divide the data into k clusters. The centroids for each cluster are displayed in the plot diagram for the k-means clustering that was done on the inputs with k = 6. The profile plot offers information about each cluster's buying habits and discount receptivity. For instance, while Cluster 5 concentrates more on fruit purchases and is highly responsive to discounts, Cluster 1 tends to spend more on luxury items like wine and gold and is less responsive to discounts.

Based on these conclusions, the final 5 cluster recommendations would be:

1. **Luxury Enthusiasts:** Target this cluster (Cluster 1) with high-end products like wine and gold. Focus on exclusivity and quality rather than offering discounts.
2. **Online Luxury Shoppers:** Engage with this cluster (Cluster 2) through online channels and emphasize their interest in luxury items, particularly gold. Provide convenient online shopping experiences and showcase the uniqueness of the products.
3. **Store Shoppers:** This cluster (Cluster 3) prefers the in-store shopping experience, although their spending patterns may not be as distinct. Consider further analysis or refining the clustering approach to better understand their preferences and needs.
4. **Gold Campaign Responders:** Target this cluster (Cluster 4) with marketing campaigns for gold products, highlighting their responsiveness. Offer discounts and promotions to incentivize purchases. Also, consider cross-selling opportunities with related luxury items.
5. **Discount-Driven Shoppers:** Focus on this cluster (Cluster 5) by offering discounts and reduced prices, especially on fruits. Emphasize the value proposition and savings they can achieve through store purchases.

Bibliography

Patel, A. (2021, August 22). *Customer Personality Analysis: Analysis of company's ideal customers*. Kaggle. Accessed May 1, 2023https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

Kapoor, K. (2021, October 8). Customer segmentation: Clustering . Kaggle. https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering/notebook Accessed May 1, 2023