– Customer segmentation on Wholesale customers data using Cluster Analysis

• Table of Content

1. An abstract

1.customer sugmentation:

It is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age ,gender,interests and spending habits.

2.using cluster analysis:

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning.

Clustering or cluster analysis is an unsupervised technique.

2.Exploratory data analysis

```
#For channel : 1 - Horeca (hotel/restaurant/cafe) ,2 - Retail channel
#For region : 1 - Lisbon, 2 - Oporto , 3 - Other region
```

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	Retail channel	Other	12669	9656	7561	214	2674	1338
1	Retail channel	Other	7057	9810	9568	1762	3293	1776
2	Retail channel	Other	6353	8808	7684	2405	3516	7844
3	Horeca	Other	13265	1196	4221	6404	507	1788
4	Retail channel	Other	22615	5410	7198	3915	1777	5185

Clearly the data set is fit for either classification or clustering. I have chosen clustering.

Categorical statistical analysis of variables:

```
#Adding a column 'NoOfCustomers' so that pivot operations can be done
mydata['NoOfCustomers'] = 1
```

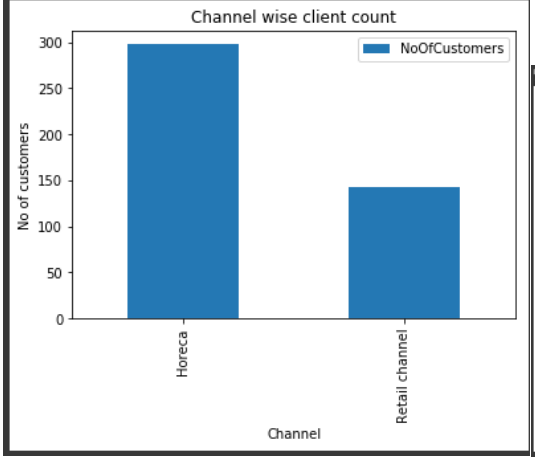
```
mydata.head()
```

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	NoOfCustomers
0	Retail channel	Other	12669	9656	7561	214	2674	1338	1
1	Retail channel	Other	7057	9810	9568	1762	3293	1776	1
2	Retail channel	Other	6353	8808	7684	2405	3516	7844	1
3	Horeca	Other	13265	1196	4221	6404	507	1788	1
4	Retail channel	Other	22615	5410	7198	3915	1777	5185	1

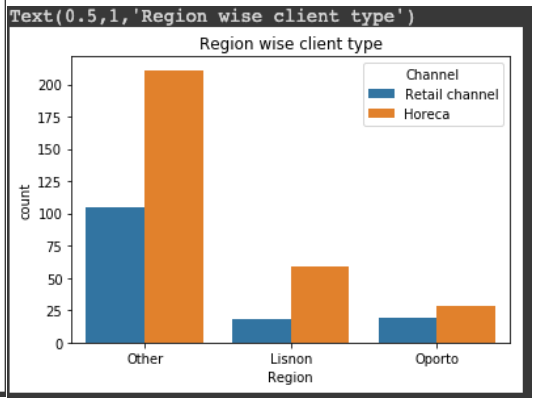
It is then possible to count the number of customers in the different channels, 298 customers belong to channel 1 and 142 customers belong to channel 2.

The largest customer of the entire sales distributor is the hotel/restaurant/cafe owner.

```
                NoOfCustomers
Channel
Horeca                    298
Retail channel            142
Text(0,0.5,'No of customers')
```

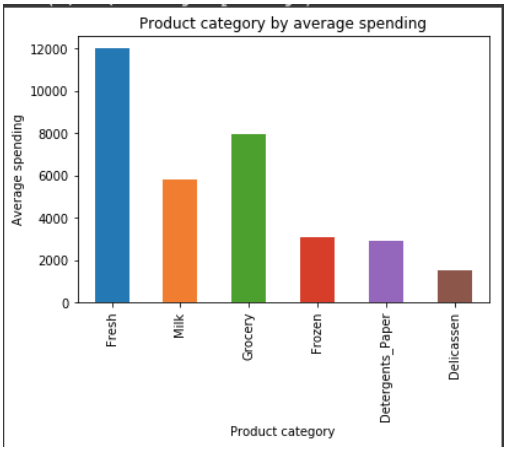



```
Text(0.5,1,'Region wise client type')
```

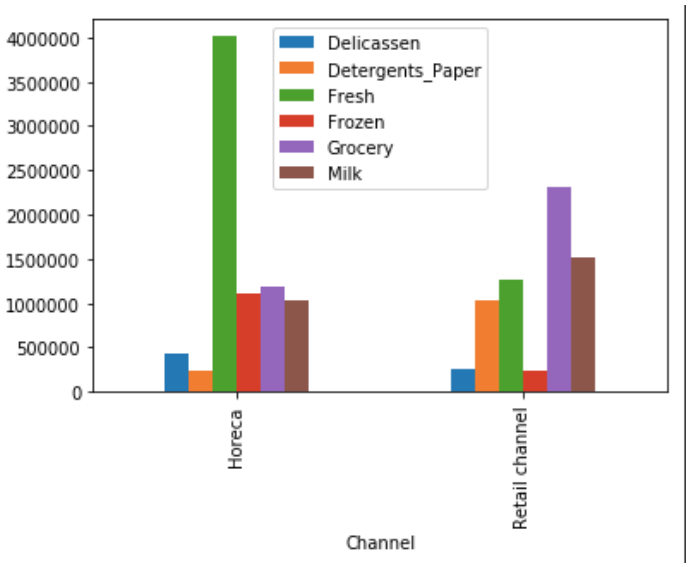


Besides, Across all three regions the clients of type1/channel1(Hotel/restaurant/cafe) are more than the clients of type2/channel2(Retail)

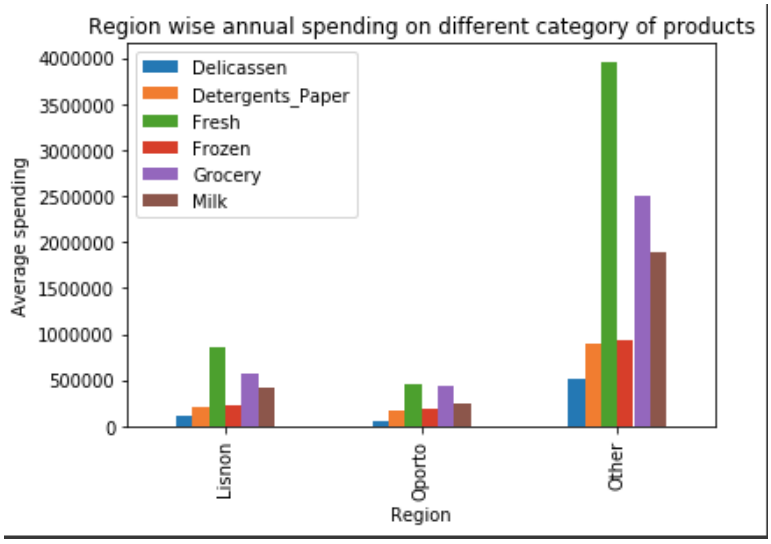
	count	mean	std	min	25%	50%	75%	max
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicassen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0



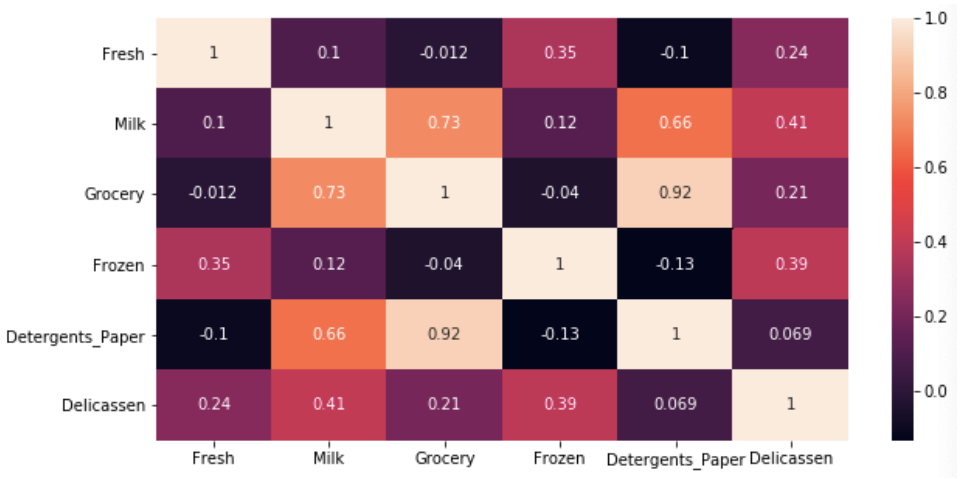
Creating a data frame of all numeric variables, On an average clients spend/wholesale distributor earns the most from Fresh products and least from Delicatessen products.Second most bought/sold category of products being Grocery and third being Milk products.



Channel wise annual spending on different category of products,Clients who are Hotel/Cafe/Restaurant owners spend most on Fresh products and least on Detergents_paper products and retail clients spend most on Grocery products and least on Frozen products.



In the mean time,From Region wise annual spending on different category of products,Across all three regions the annual spending on Fresh products is the highest and the annual spending on Delicatessen products is the lowest.



Final,Checking for correlation among the numeric variables,we can get some strong correlations:

1)Grocery and Milk have a strong positive correlation (0.73) - People who buy grocery also buy milk and visa versa

2)Detergents_Paper and Milk have a strong positive correlation (0.66) - People who buy Detergents_Paper products also buy milk and visa versa

3)Detergents_Paper and Grocery have a strong positive correlation (0.92) - People who buy Detergents_Paper products also buy grocery and visa versa

3.Data Modelling(including methodology)

Cluster Analysis, There are two broad types of cluster analysis/clustering:

1.Non-heirarchical clustering 2.Heirarchical clustering

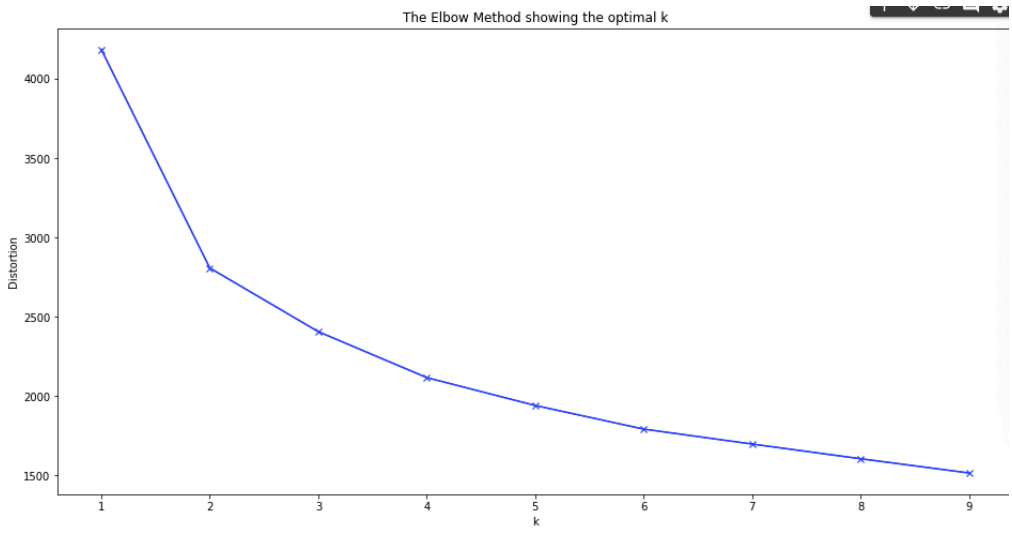
First, using Non-heirarchical clustering - The k-means algorithm[1]:

1) Choose value for K 2) Randomly select K featuresets to start as your centroids 3) Calculate distance(similarity/dissimilarity) of all other feature sets to centroids 4) Classify other featuresets as same as closest centroid 5) Take mean of each class (mean of all feature sets by class), making that mean the new centroid 6) Repeat steps 3-5 until optimized (centroids no longer moving)

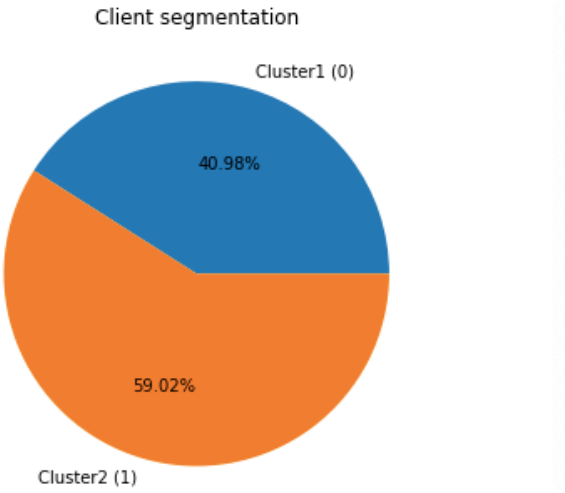
Here, set k=2, Data set={2,3,4,10,11,12,20,25,30}, Then,Eliminating the dummy variables for building the k-means model.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	9.446913	9.175335	8.930759	5.365976	7.891331	7.198931
1	8.861775	9.191158	9.166179	7.474205	8.099554	7.482119
2	8.756682	9.083416	8.946896	7.785305	8.165079	8.967504
3	9.492884	7.086738	8.347827	8.764678	6.228511	7.488853
4	10.026369	8.596004	8.881558	8.272571	7.482682	8.553525

Finding the optimum number of clusters into which the data may be clustered,Using The elbow curve,This method uses WSS(within sum of squares) - It is the is the total distance of data points from their respective cluster centroids.



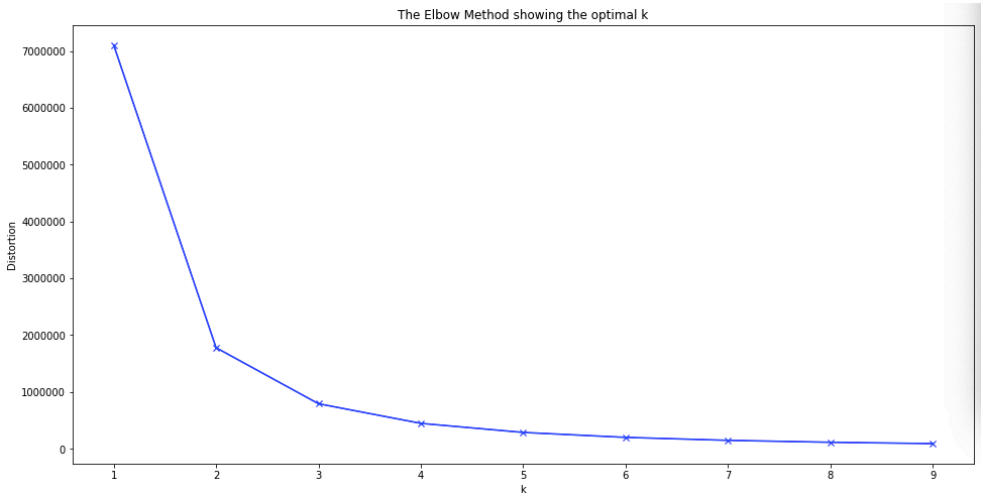
Building the K-means model with 2 clusters.We can see that 2nd cluster has maximum number of samples, while 1st cluster has minimum number of samples. Through output shows which customer belongs to which cluster.



2.Heirarchical clustering - The agglomerative approach

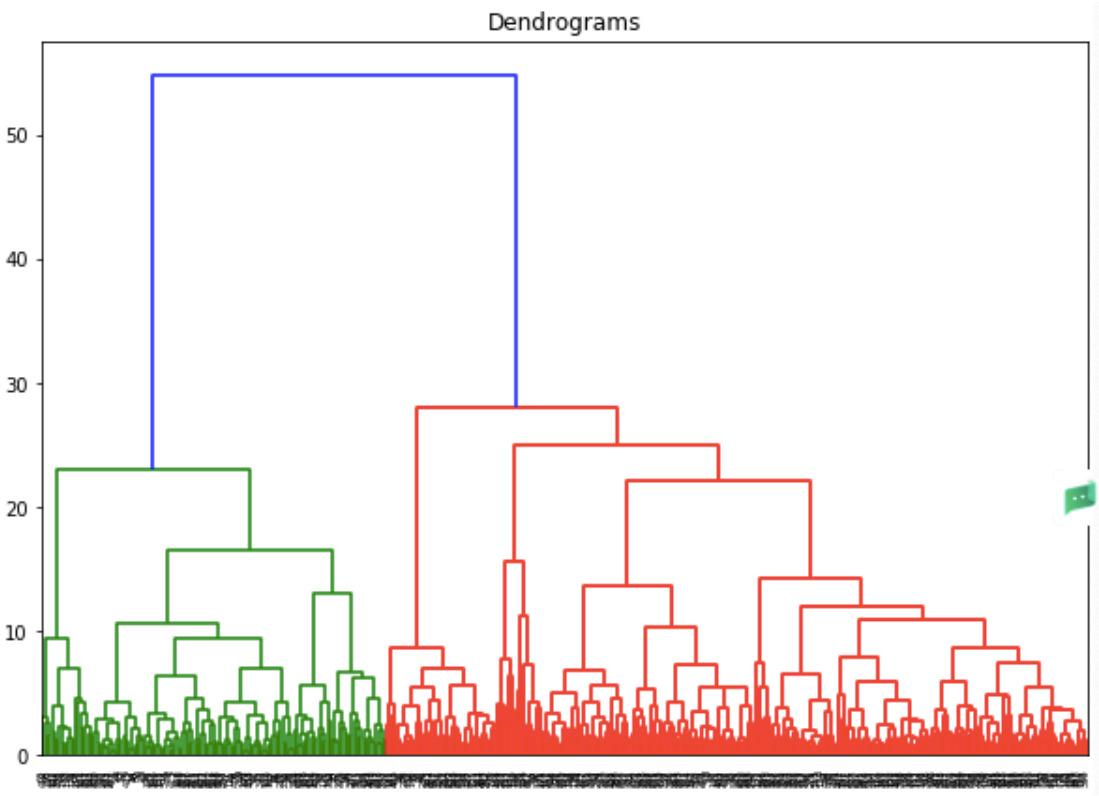
In this approach, all objects start in separate clusters till slowly similar objects are combined and this process is repeated until all objects are in a single cluster. Finally, the optimum number of clusters are chosen from among all options.

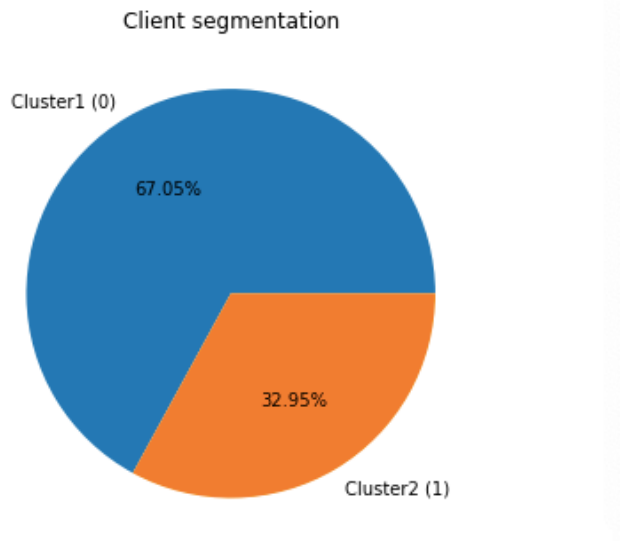
First, using elbow curve to find the optimum number of clusters into which the data may be clustered.



Similarly, the sample was divided into two categories according to clusters, yielding a customer cluster classification diagram.

Besides, A dendrogram[2] is generated from the samples, a threshold distance is set, a horizontal line is drawn and the number of clusters is the number of intersections with the threshold horizontal and vertical lines as seen in the intersections (from the dendrogram we get 2 clusters).





4.Results

K-means algorithm is the simplest, very fast and easy to implement, all you need to do is find the suitable value of K.

5.Discussion

As a business owner, it is always desirable to know what groups of customers you're dealing with, be it in terms of spending patterns, retention, or similar factors. This could help you with tailoring your product or service, branding, marketing and eventually, customer satisfaction.

In this project we will apply unsupervised learning techniques on product spending data collected for customers of a wholesale distributor to identify customer segments hidden in the data. We will first explore the data by selecting a small subset to sample and determine if any product categories highly correlate with one another. Afterwards, we will preprocess the data by scaling each product category and then identifying (and removing) unwanted outliers. With the good, clean customer spending data, we will implement clustering algorithms to segment the transformed customer data. Finally, we will compare the segmentation found with an additional labeling and consider ways this information could assist the wholesale distributor with future service changes.

6.Conclusion

In this project we studied the customer segments of a wholesales supplier. We started by understanding the underlying distribution of the data, and moved on to cleaning outliers, reducing features and finally clustering customers based on their spending profile.

As a business owner, it is always desirable to know what groups of customers you're dealing with, be it in terms of spending patterns, retention, or similar factors. This could help you with tailoring your product or service, branding, marketing and eventually, customer satisfaction. For example, if the wholesale distributor wanted to change its delivery service from 5 days a week to 3 days a week, it could run an AB test on either cluster first to ensure it's suitability for that customer segment. After all, different segments have different needs and could react differently to a business decision.

One thing to note is that we looked at a snapshot of data along the continous

timeline of customer spendings. It would be quite interesting to see how the segments we identified evolve throughout the time. Could they unify, diverge, or even split in holiday season?

Overall, understanding customer segments sheds more light on our blind spots and data-driven decisions that follow, would lead to improved service and higher customer satisfaction.

7.References

[1] Zhang, Q., Couloigner, I. (2005). A New and Efficient K-Medoid Algorithm for Spatial Clustering. In: , et al. Computational Science and Its Applications – ICCSA 2005. ICCSA 2005. Lecture Notes in Computer Science, vol 3482. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11424857_20

[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html