Wholesale Customers Clustering

Contents

I. INTRODUCTION

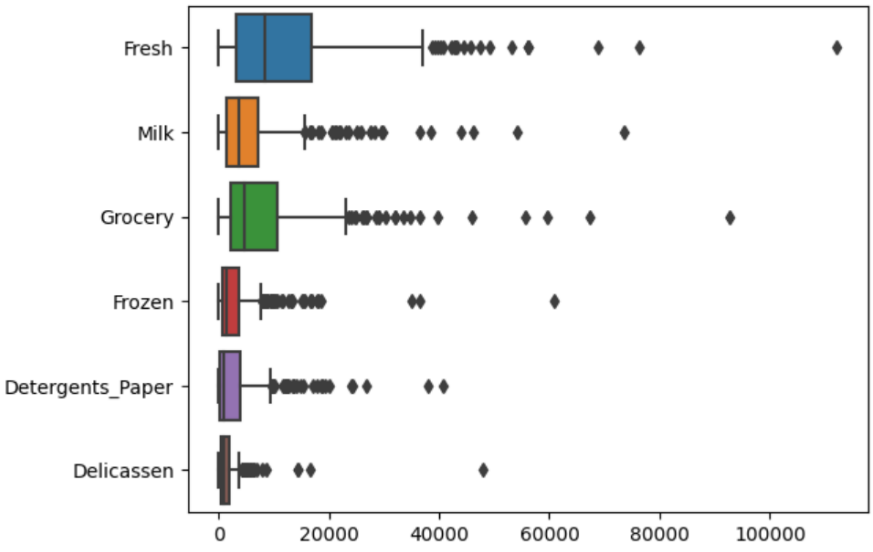
Clustering analysis is one of the most popular tasks in Machine Learning. Clustering algorithms are considered unsupervised, which do not require data label, or dependent variable, to train the model. The purpose of clustering algorithms is to separate a given dataset into smaller groups, each group is expected to contain observations with similar characteristics. To measure such similarity, common algorithms such as KMeans Clustering, Hierarchial Clustering, and DBSCAN are normally used. These models rely on the distances between observations and the level of observations' density to perform grouping. One example, which can easily be related to, is that people living in the same country are expected to have similar skin color, speak the same language, and grow up with the same culture, while people living in distant countries may show extremely different identities. Cluster analysis is very important and widely applied to solve complex real-life problems such as market segmentation, social network analysis, search result grouping, medical imaging, image segmentation, and anomaly detection ("What Is Clustering?," n.d.).

II. DATASET DESCRIPTION

In this paper, the data set Wholesale Customers, (Cardoso, 2014), is used to demonstrate the application of clustering analysis in solving customer segmentation problem. The data set contains 440 instances with the information of 8 variables recorded for each. In which, 6 variables illustrate the annual spending of each customer in different product categories, which are fresh, milk, grocery, frozen goods, delicatessen, and detergents and paper. The other two variables, region and channel, respectively indicate the location of customers and the type of their business. The following table and graph summarize the variables' attributes:

Variable	Data Type	Description	No of entries
Fresh	Numeric	Annual spending on fresh products	440
Milk	Numeric	Annual spending on milk products	440

Grocery	Numeric	Annual spending on grocery products	440
Frozen	Numeric	Annual spending on frozen products	440
Detergents_paper	Numeric	Annual spending on detergents and paper products	440
Delicatessen	Numeric	Annual spending on delicatessen products	440
Channel	Categorical	Horeca (Hotel, Restaurant, and Café) or Retail channel	440
Region	Categorical	Lisbon, Porto, or Other	440



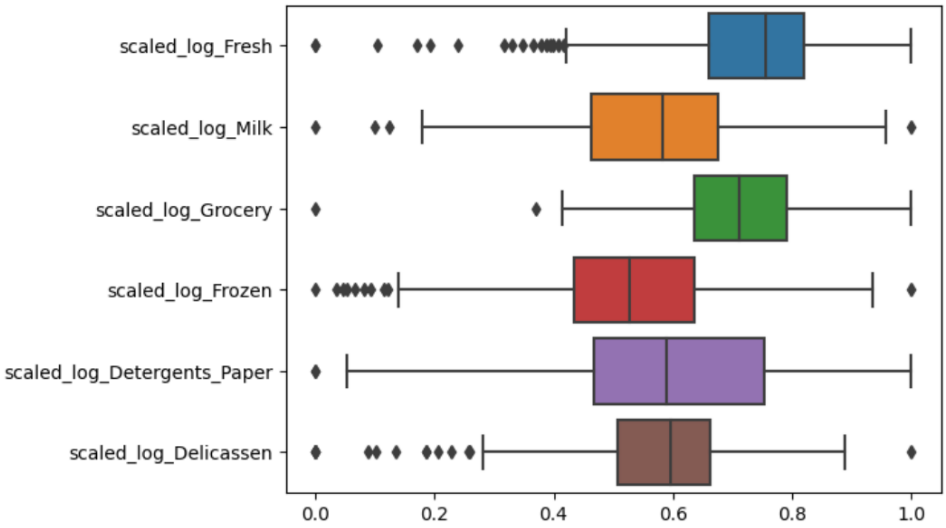
There are no missing values in 8 variables. However, the boxplots discover extreme outliers in 6 numeric variables, which need to be addressed in the next steps.

III. METHODOLOGY

To develop models, which segment wholesale customers, the following steps are conducted: (1) data preprocessing, (2) model fitting, and (3) cluster evaluation.

- Data preprocessing:

  The process contains three sub-steps: variable selection, outlier treatment, and data normalization. In the first sub-step, two categorical variables, Channel and Region, are excluded. Using the Tukey-HSD test, the mean differences in expenditure by categories between regions are insignificant, hence, removing this feature helps reduce noise in our dataset. For Channel variable, significant mean differences in expenditure between channels are found. However, Channel seems to be labels, which are derived from customers' spending behaviors, hence, this feature is removed to prevent the model from being biased. The later clustering results will take Channel into the discussion. In the second sub-step, 6 remaining numerical variables are log transformed to diminish the effect of outliers. And in the third sub-step, the MinMax Scaling is used to transform the values of 6 numerical variables into the same scale from 0 to 1. The following chart illustrates the distribution of 6 numerical variables after being transformed:



- Model fitting:

  Three selected algorithms to cluster the dataset are KMeans Clustering, Hierarchical Clustering, and DBSCAN.

  KMeans Clustering:

The model initiates K centroids, where K is the number of clusters defined by users, (Tan et al., 2018). Each observation is assigned to the closest centroid, and K groups, containing similar observations in terms of distance, are formed. The centroid within each group is then updated by averaging the coordinates of all observation members and the grouping process is conducted again until the centroids no longer change. One issue of KMeans is the initial centroids are randomly selected, and clusters may be formed differently from the optimal clusters, the ones that show the lowest SSE. Therefore, K-means++ is applied to initiate centroids, which are in optimal distances. The process of initiating centroids can be repeated multiple times to find the best output and in this paper, the model is configured to repeat 10 times. The random state is also set to ensure the results remain the same after re-runs. Other than that, different numbers of clusters (from 2 to 20) are experimented.

Hierarchical Clustering:

The model is displayed as a tree graph with a single top node as a peak of the tree and root nodes are single observations, the maximum of clusters is the number of observations. The model merges the two closest clusters into a group, the model performs the grouping until a full link is formed connecting the root nodes to the peak. The Hierarchical Clustering provides multiple methods, which are single link, complete link, group average, and ward, to define the proximity between clusters. In this paper, the complete-link method is used. In which, the proximity between clusters is defined as the maximum distance between any two points in two different clusters, (Tan et al., 2018). Other than that, different numbers of clusters (from 2 to 10) are experimented.

DBSCAN:

The model is density-based clustering that identifies different regions of high-density to form separate clusters. Such regions are determined based on the distance between core points, border points, and noise points. A core point is the point surrounded by at least $MinPts$ points within a distance of Eps, (Tan et al.,

2018). $MinPts$ and Eps are defined parameters by users. Border points are the ones, which have the distance to the core point less than Eps. Noise points are points, which are neither core nor border points. DBSCAN first determines core points, and core points located within a distance of Eps are grouped together. The border points of grouped core points are also in the same cluster. Meanwhile, noise points are left aside. As the the value range of each variable is from 0 to 1, hence, the maximum value of Eps is around 2.45 ($\sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2}$). In this paper, Eps values from 0.05 to 2.45 and $MinPts$ from 5 to 100 are experimented.

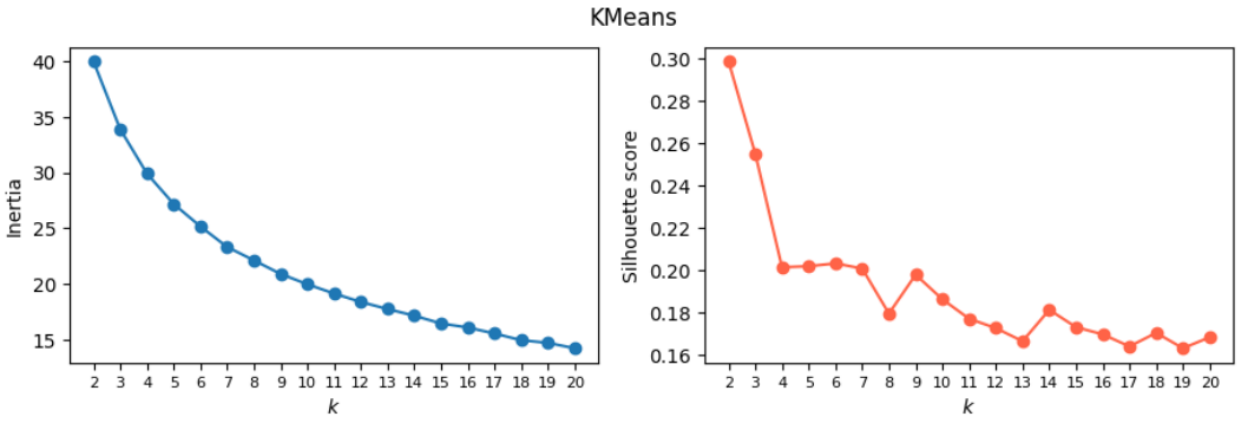
- Cluster evaluation:

  The following metrics are used to select the appropriate number of clusters:

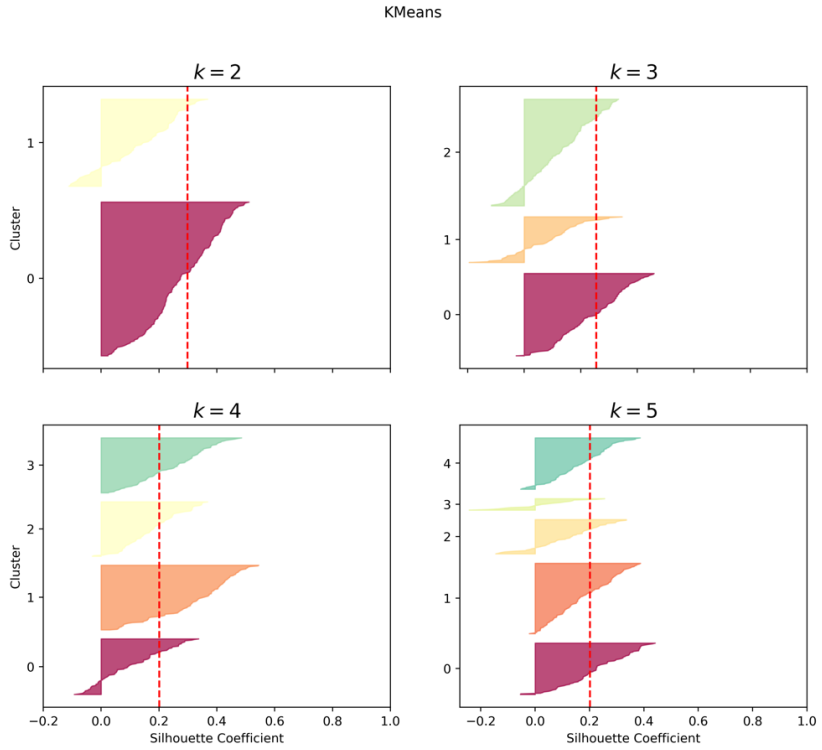
  - Within-cluster sum of squares (WCSS), or Inertia: The sum of squared distance between each point within the cluster to the cluster's centroid. The smaller value of this metric tends to indicate better clustering, however, as the specified number of clusters increases, the value of WCSS decreases. Therefore, the normally selected number of clusters is at the elbow point.
  - Silhouette coefficient: The silhouette coefficient of i^{th} object is determined as $s_i = (b_i - a_i)/\max(a_i, b_i)$. a_i is the average distance of the i^{th} object to other objects in the same cluster. And, b_i is the minimum distance of the i^{th} object to any other object, which is not in the same cluster. The value of the silhouette coefficient falls into the range of -1 to 1. A value close to 1 means that the dataset is well clustered, while a value close to -1 means that the dataset is badly clustered.
  - Number of observations per cluster: The number threshold of observations per group is 44, which is 10% of total observations. The number of clusters is expected to be less than 10.

IV. RESULTS

For KMeans Clustering, different numbers of clusters (from 2 to 20) were experimented and their inertias and silhouette coefficients are shown in the following chart:

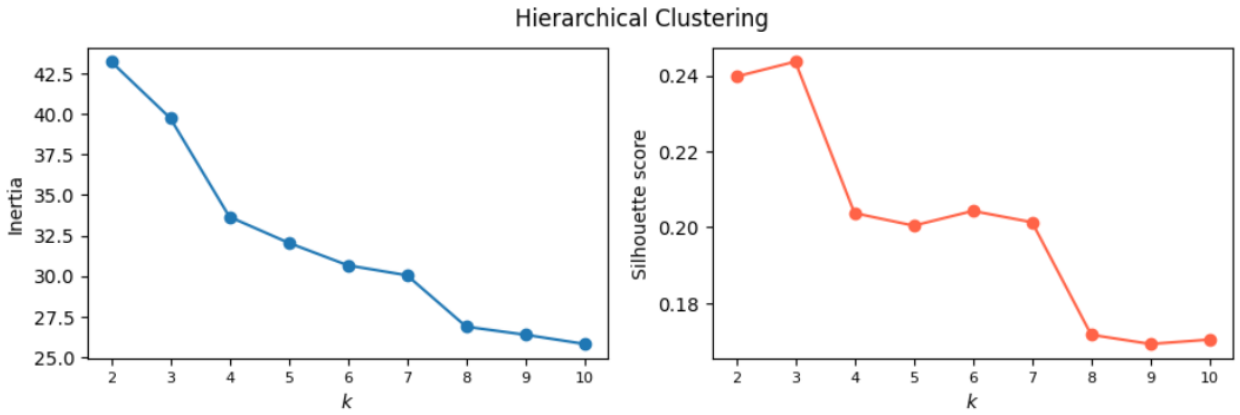


The appropriate number of clusters should be 7, because the number is at the elbow point and the silhouette score is not too low, at around 0.2. However, with 7 clusters, the minimum number of observations per cluster (44) was not met, and only clusters number of 2, 3, 4, and 5 are qualified, therefore, these clusters numbers are examnined further.



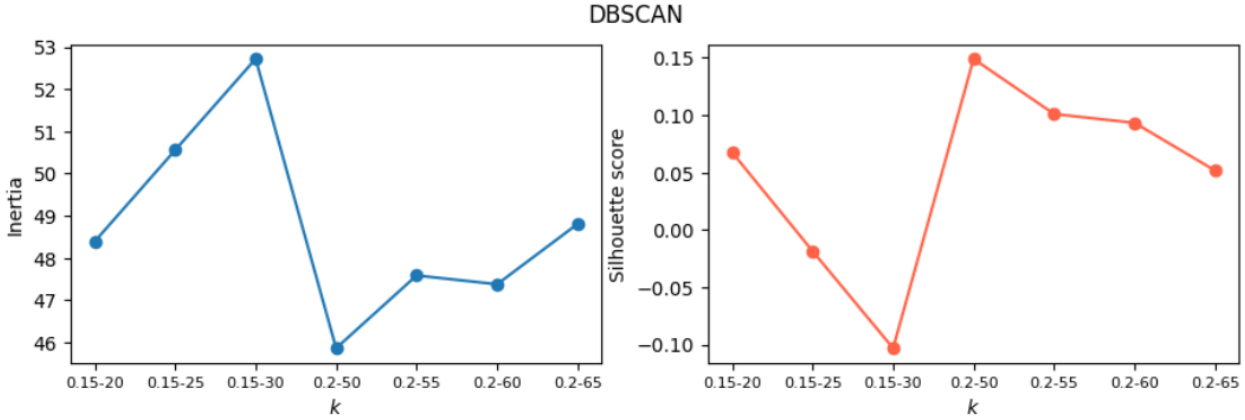
The clusters number of 4 is selected for KMeans Clustering, because each cluster in this option has a significant portion of the top exceeding the red line (the average silhouette coefficient). While, clusters number of 5 seems to misclassify many in cluster 2 and 3. In comparison with clusters number, 2 and 3, clusters number of 4 shows a lower value of Inertia.

For Hierarchical Clustering, different numbers of clusters (from 2 to 10) using the complete-link method were experimented and their inertias and silhouette coefficients are shown in the following chart:



The appropriate number of clusters should be 4, because the number is at the elbow point and the silhouette score is not too low, at around 0.2. However, with 4 clusters, the minimum number of observations per cluster (44) was not met. And, only clusters number of 2 was qualified, hence, clusters number of 2 is selected for Hierarchical Clustering.

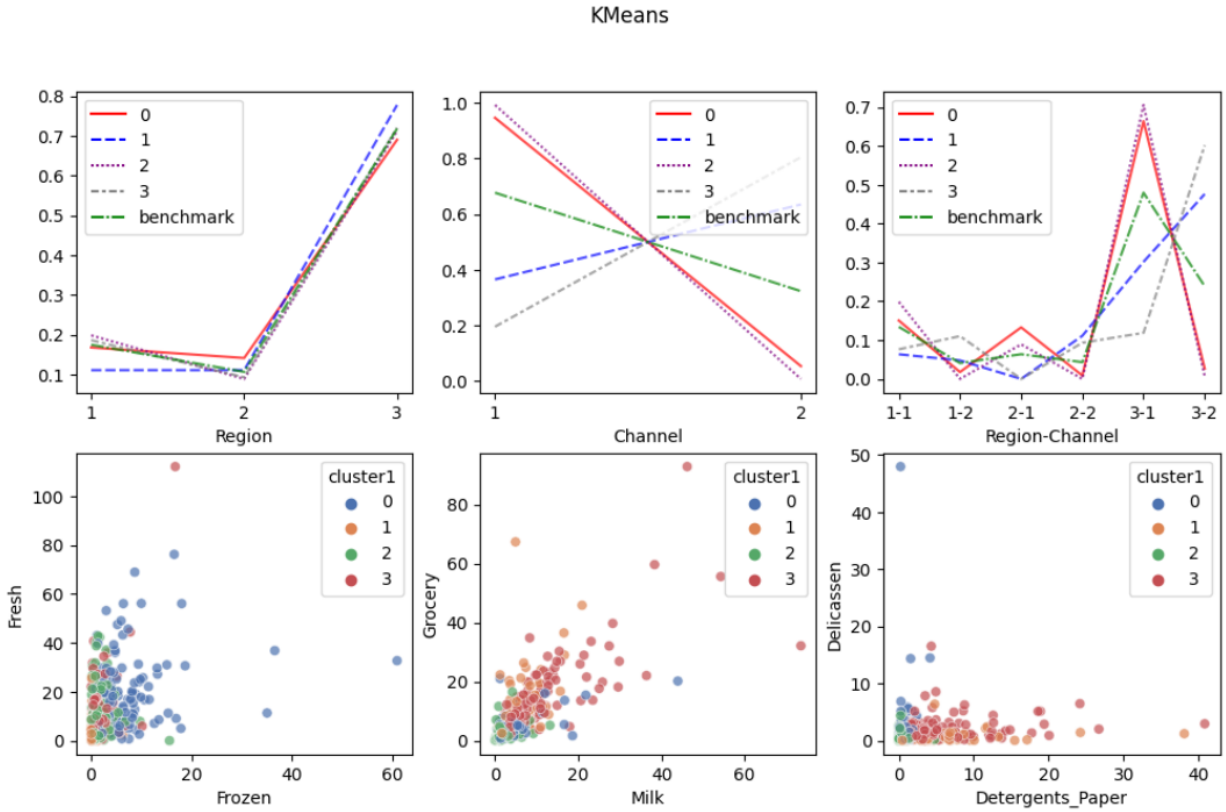
For DBSCAN, different sets of Eps (from 0.05 to 2.45) and $MinPts$ (from 5 to 100) were experimented. The inertias and silhouette coefficients for the sets, which meets the minimum number of observations per cluster and minimum clusters is 3 (because DBSCAN always leaves a group for noise points, and remaining groups are for actual clustering normal observations), are shown in the following chart:



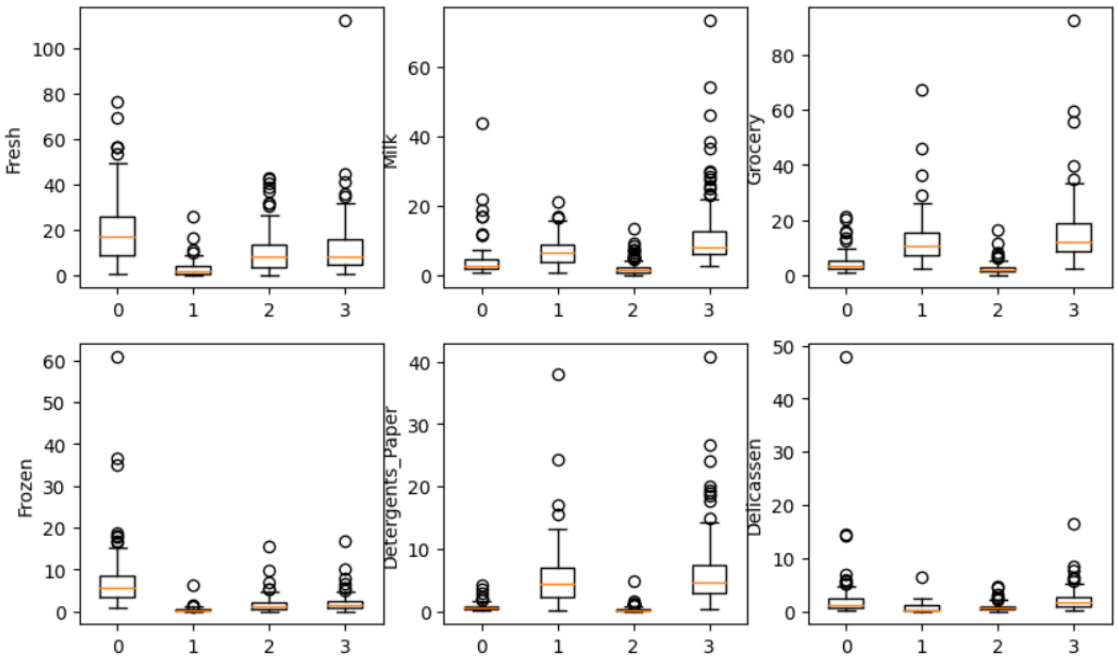
The Eps of 0.2 and $MinPts$ of 50 are selected because this option provides the lowest value of Inertia and the highest value of silhouette score.

V. DISCUSSION

Looking at 4 clusters by KMeans Clustering:

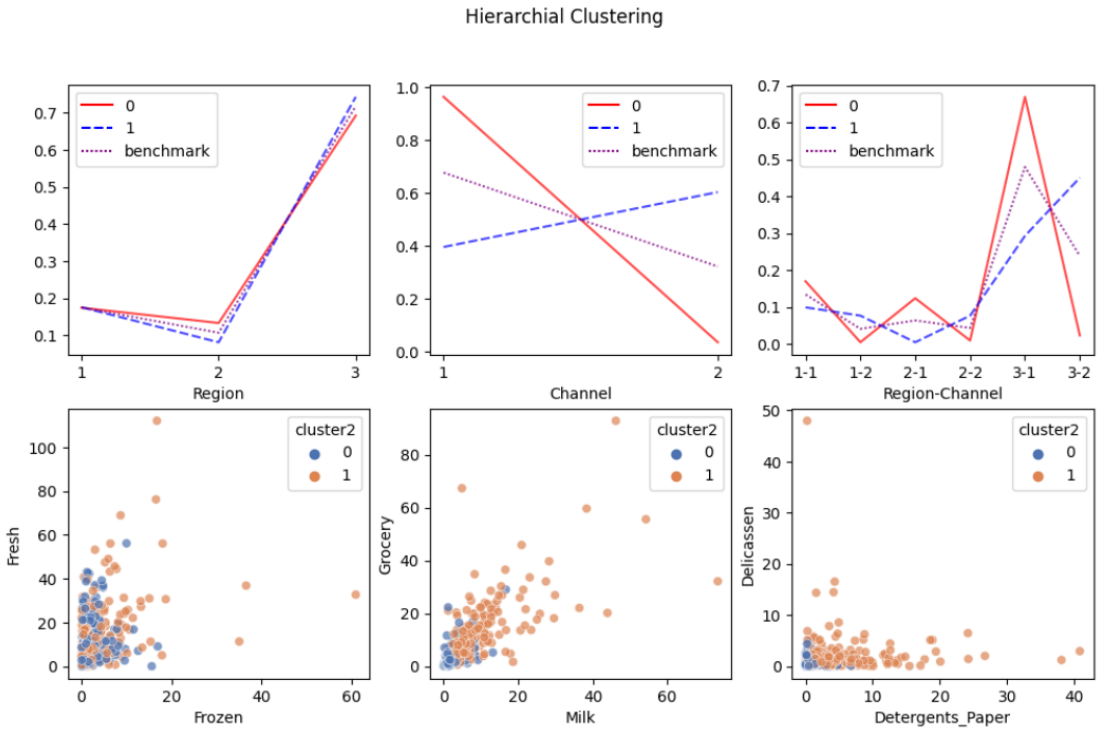


The cluster 0 (113 observations) and 2 (146 observations) mostly contains Horeca customers, while the cluster 1 (63 obs) and 3 (118 obs) have the higher portion of Retail customers. In addition, the two clusters 0 and 2 tend to spend more on Fresh and Frozen products, while the other two tend to spend more on Milk, Grocery, and Detergents and paper products. All these 4 clusters have the location ratios similar to the overall dataset, no significant differences between 4 clusters.



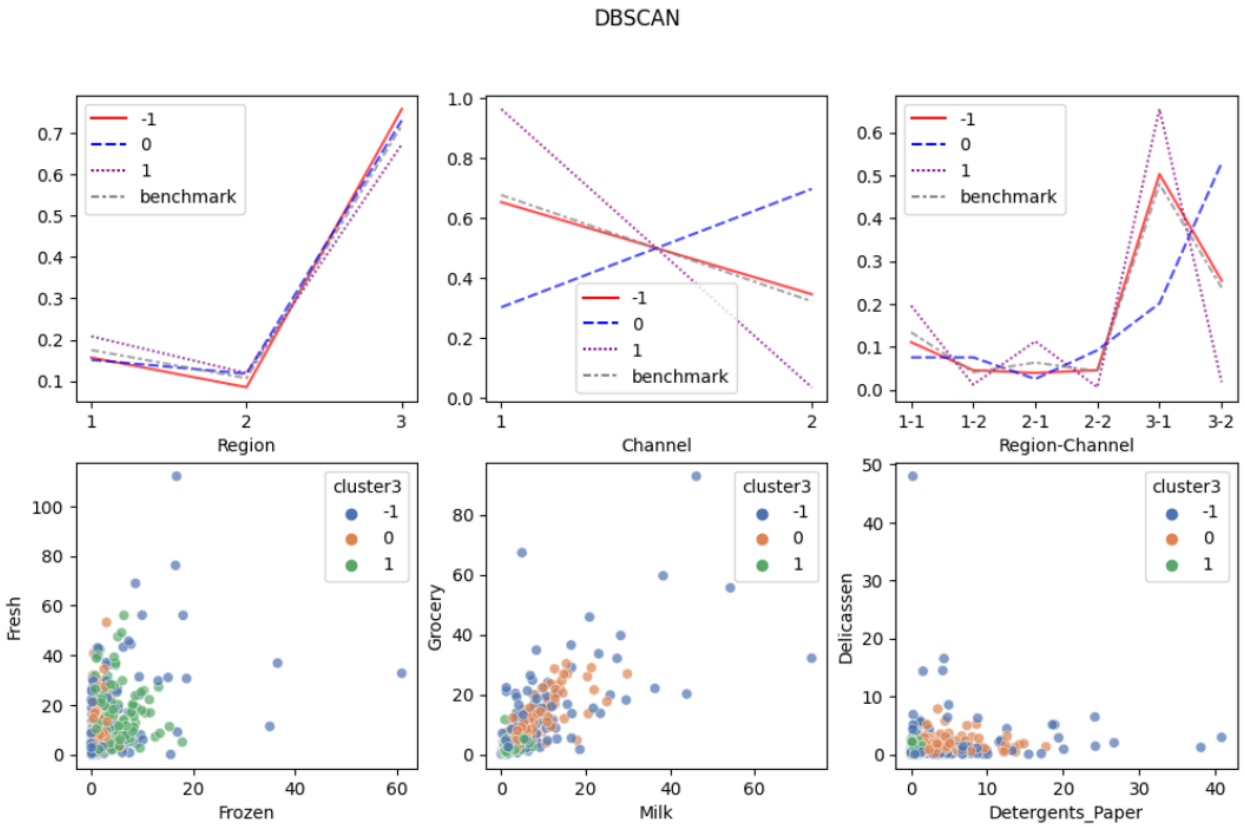
For more details, the cluster 0 contains observations having larger amount of expenditure on Fresh and Frozen products than observations in the cluster 2. And, the cluster 3 tends to consume more Fresh products than the cluster 1.

Looking at 2 clusters by Hierarchical Clustering:



The cluster 0 (218 observations) includes only Heroca customers, while the cluster 1 (222 observations) has a larger portion of Retail customers. The observations in cluster 1 tend to spend more in Milk, Grocery and Detergents and paper than the others in cluster 0. The 2 clusters have the location ratios similar to the overall dataset, no significant differences were observed.

Looking at 3 clusters by DBSCAN:



The cluster -1 (153 observations) is considered to contain noise points, or outliers. The cluster 0 (119 observations) shows a greater amount of expenditure spent on Milk, Grocery and Detergents and paper products, while the cluster 1 (168 observations) consumes more Fresh and Frozen products. In addition, the cluster 1 mostly contains Heroca customers, while the cluster 0 has Retail customers as the majority.

Overall, the three algorithms are able to identify two main groups, one has a larger portion of Heroca customers, and the other one has a larger portion of Retail customers. In which,

the former group tends to spend more in Fresh and Frozen products, while the latter group consumes more Milk, Grocery, and Detergents and paper products. Moreover, KMeans Clustering seems to be more versatile in clustering the dataset into quantity-balanced groups. While, DBSCAN highlights the work on finding the noise points.

VI. CONCLUSION

The three clustering models, which include KMeans Clustering, Hierarchical Clustering, and DBSCAN, are demonstrated through analysing the Wholesale Customers dataset. The three models have a common finding of strongly separating Heroca and Retail customers. However, no models generate the same results as others. This illustrates a limitation of clustering analysis that there are no specific metrics to help decides what set of clusters is the best, the clustering result is normally decided by expert judgment. Other than that, most of the clustering algorithms may be time-consuming as the number of variables increases because such algorithms normally rely on comparing distances, which are calculated from all variables' values, between observations. Therefore, complex clustering problems regularly require expertise in both algorithm optimization and domain knowledge, which involves different human roles in real-life applications. In certain cases, domain knowledge is more helpful and helps quickly cluster the dataset and outperforms complex algorithms. For example, a quick solution for clustering the Wholesale Customers dataset is using the Channel variable.

References

What is Clustering? (n.d.). *Google for Developers*. https://developers.google.com/machine-learning/clustering/overview

Cardoso, Margarida. (2014). Wholesale customers. UCI Machine Learning Repository. https://doi.org/10.24432/C5030X.

Tan, P., Steinbach, M., Karpatne, A., & Kumar, V. (2018). Introduction to Data Mining (2nd Edition). *Introduction to Data Mining*. https://dl.acm.org/citation.cfm?id=3208440