Wholesale Customer Segmentation Report

Andrew J. Otis

Description of Dataset¶

There are a total of **8 features with 440 observations each**, the **goal of the report is to identify and describe potential customer segments hidden within the data**. Based on annual spending on diverse product categories (*e.g. Grocery, Frozen, etc.*) of various business types (*i.e. Hotel, Restaurant, and Café grouped as ‘HORECA’ and ‘Other’*) located in various regions.

The variables present can be described as follows

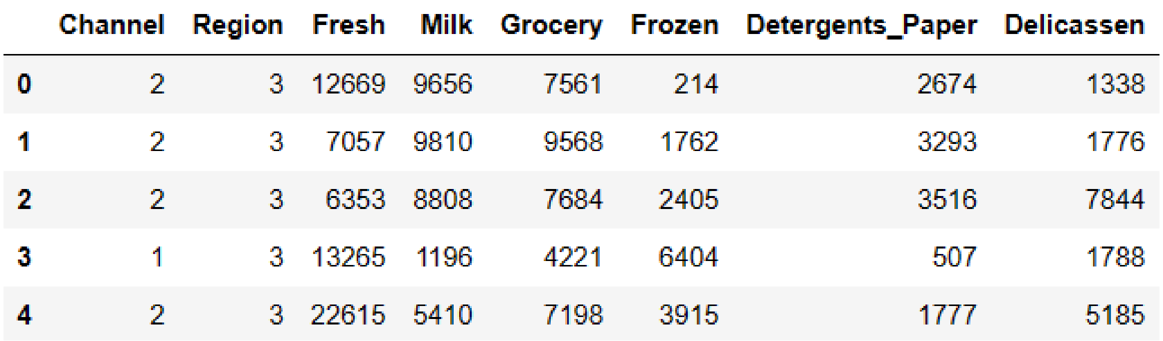
- **REGION:** customer Region (Nominal)

REGION	Frequency
Lisbon	77
Oporto	47
Other Region	316
Total	440

- **CHANNEL:** "Horeca" or "Retail" (Nominal)

CHANNEL	Frequency
Horeca	298
Retail	142
Total	440

- **FRESH:** annual spending on fresh products (Continuous)
- **MILK:** annual spending on milk products (Continuous)
- **GROCERY:** annual spending on grocery products (Continuous)
- **FROZEN:** annual spending on frozen products (Continuous)
- **DETERGENTS_PAPER:** annual spending on detergents and paper products (Continuous)
- **DELICATESSEN:** annual spending on and delicatessen products (Continuous)



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

Table 1: *Data frame of the dataset scraped from the UCI Machine Learning Repository*

Exploratory Data Analysis¶

In machine learning(*i.e. ML*) clustering models, there is no specific “target”/”predictor” variable since the goal is to group or cluster the data based on its inherent structure/similarity or patterns, rather than predicting a target variable (*i.e. unsupervised ML model*). Thus, exploratory analysis is simply getting a better understanding of the data in relation to the goal, rather than making sure it meets data assumptions of a particular clustering model.

Based on count of unique responses for ‘Region’ and ‘Channel’ being cross referenced with the metadata from the original source, the following can be deduced.

Encoded elements for the Region Column:
1 = Lisbon
2 = Oporto
3 = Other Region

Encoded elements for the Channel Column:
1 = HORECA (hotel+restaurant+cafe)
2 = Other (non-HORECA business)

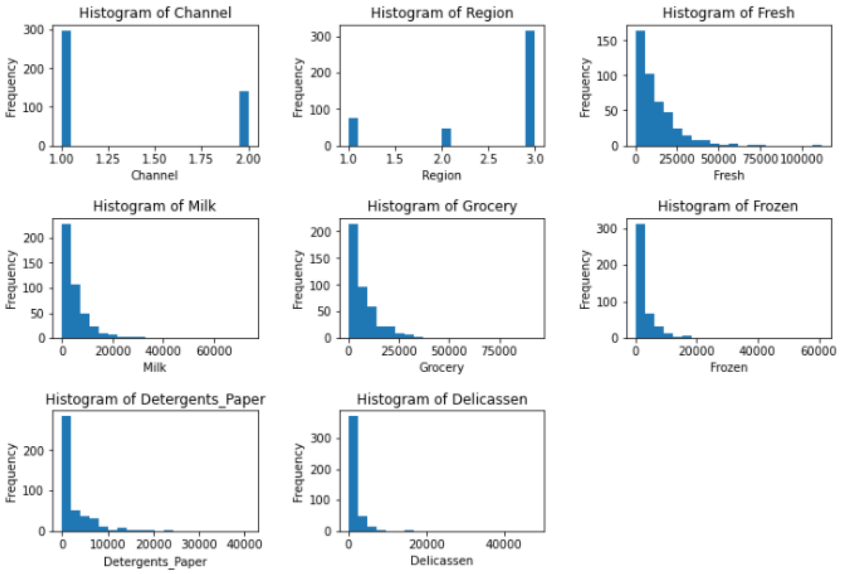


Figure 1: Distribution of all features present, respectively

The distributions of the data all are skewed to the right in common, excluding the categorical features. Supporting these observations are boxplots representing summary statistics, outliers skewing the distribution are clearly visible and present in all numeric features.

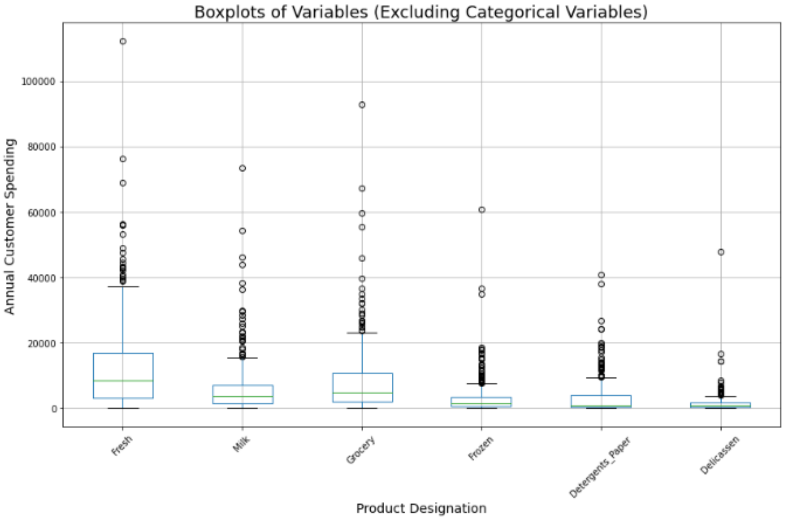


Figure 2: The plot visualizes the summary statistics of the data (e.g. range, variance, etc.)

Data Pre-Processing

The data was uploaded to the original source with the categorical features already encoded, so it was not necessary for me to re-encode them (*i.e. categorical features already prepped for sklearn ML modules*).

For clustering, data only needs to be scaled to create and assign cluster labels to the un-scaled data for cluster visualization.

To address this, data frames are created from the original un-altered data for each clustering. From there, the steps are the same for every clustering algorithm. Scale the data appropriately before being fit to the model, create, and assign clusters to the unscaled data. This was done using both 2 and then 3 Principal Components, respectively.

Model Construction & Cluster Visualizations

K-Means Clustering

The number of clusters needs to be explicitly specified; this is the nature of the k-Means clustering algorithm. This is accomplished using an Elbow Plot.

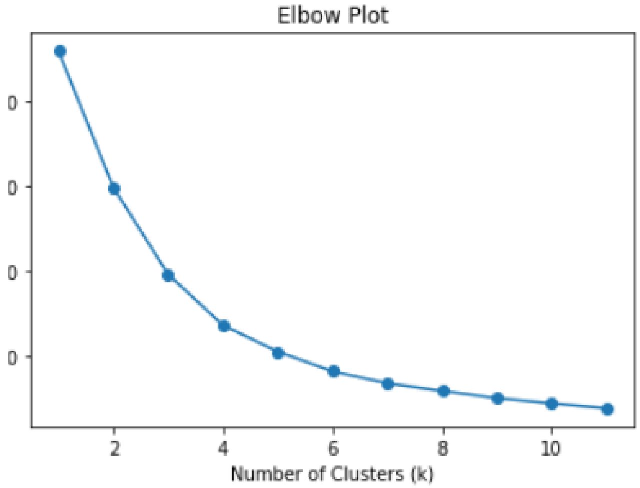


Figure 3: Remember, k-means requires one to explicitly specify the desired number of clusters. In this case, 4 clusters appears to be the optimal number of clusters.

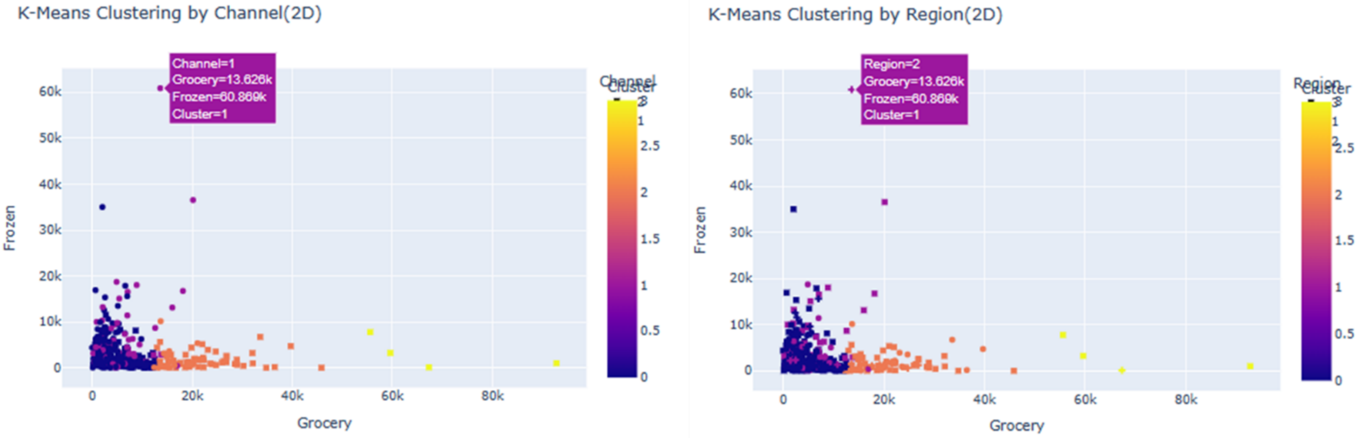


Figure 4: Clusters visualized by Channel and Region, for 2D scatterplot representation, respectively.

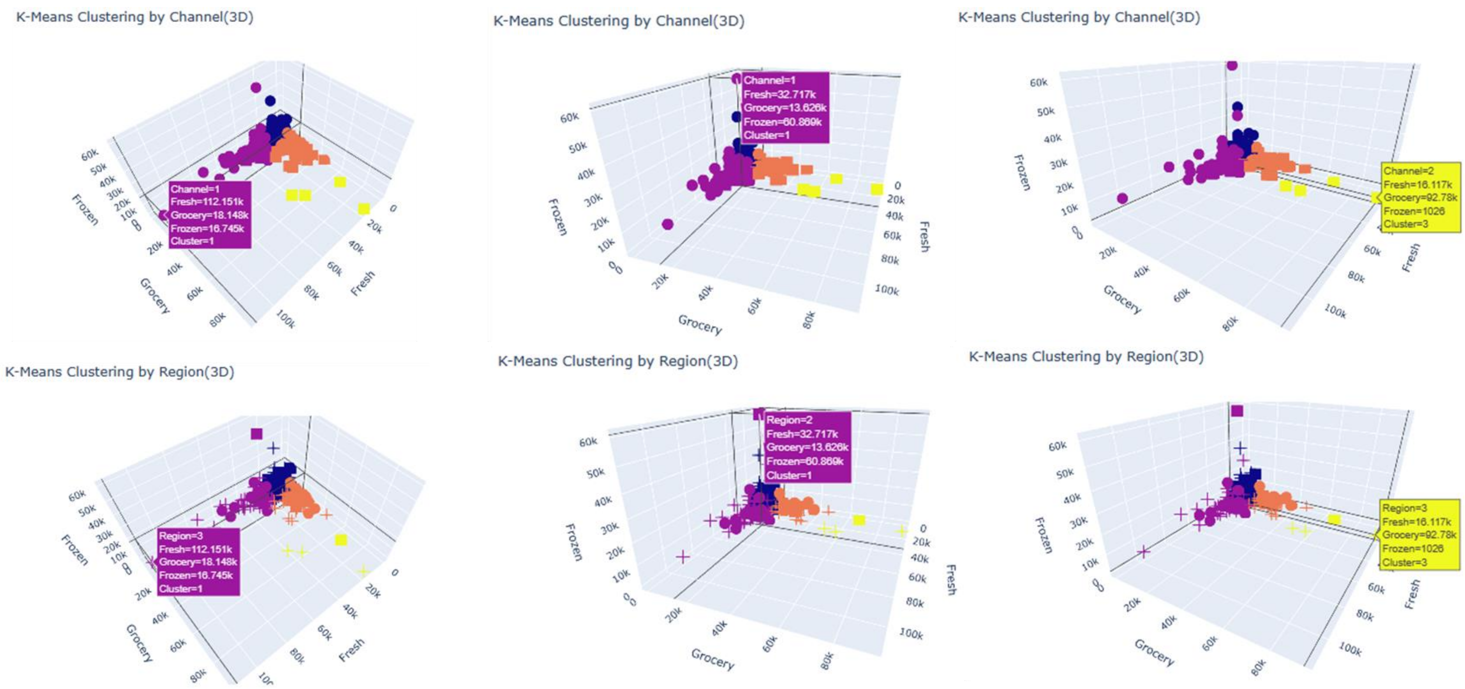


Figure 5: Clusters visualized by Channel and Region, for 3D scatterplot representation, respectively.

Hierarchical Clustering (agglomerative)

Unlike K-means clustering, agglomerative clustering does not require the explicit specification of the desired number of clusters, it will automatically determine that for us.

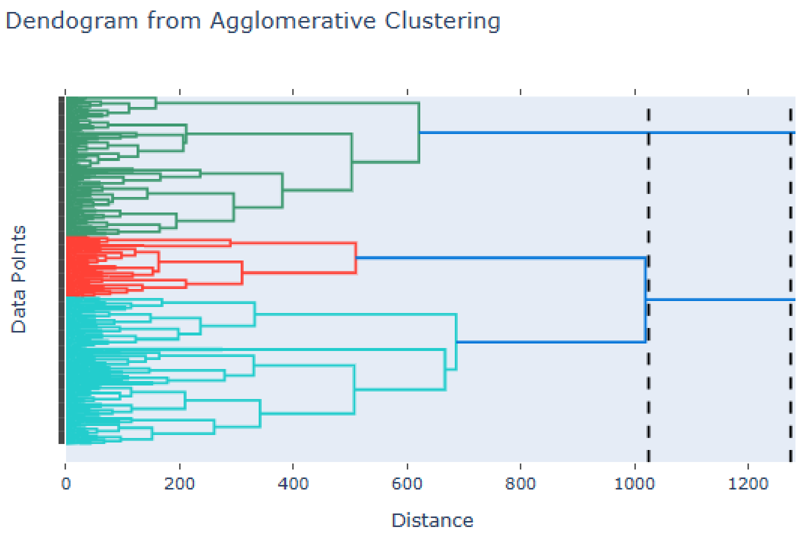


Figure 6: Algorithm automatically clusters the data, no need to explicitly specify the desired number of clusters. In this case, the algorithm determined 2 clusters to be the ideal number. This is supported by the only cluster labels resulting being “0” & “1”

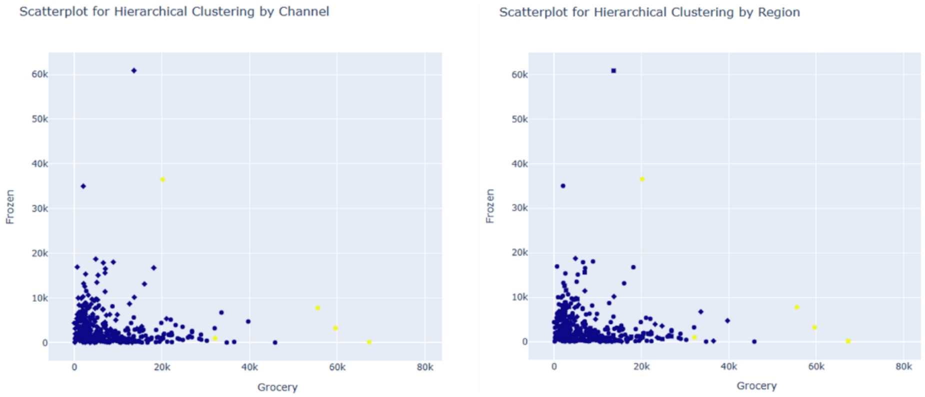


Figure 7: Clusters visualized by Channel and Region, for 2D scatterplot representation

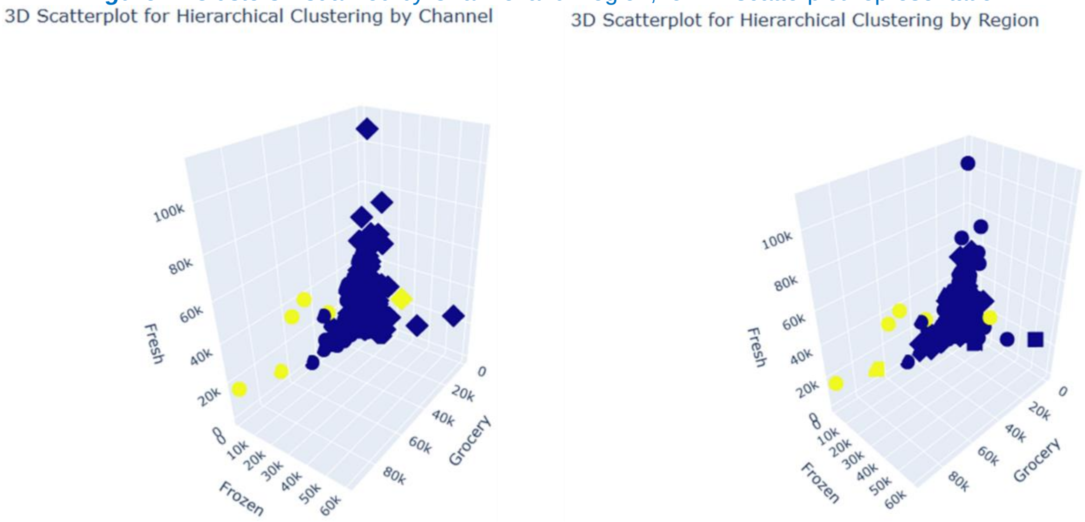


Figure 8: Clusters visualized by Channel and Region, for 3D scatterplot representation.

Model Evaluation & Customer Segmentation

K-Means Clustering

Customer Segment 1:
A HORECA customer in region “Oporto”, from cluster 1 that has most of the annual spending in the ‘Frozen’ category

- The business could be something like an ice cream shop, where a majority of products would likely be labeled under that ‘Frozen’ product designation.

Customer Segment 2:
A HORECA customer in region “Other”, from cluster 1 that has most of the annual spending in the ‘Fresh’ category

- The business could be something like a farmers' market, where most of the products were something sourced recently. Unlike grocery stores or other HORECA businesses, Farmers markets typically occur only once or twice a week. These markets are typically advertised to have the freshest items for sale.

Customer Segment 3:
A non-HORECA customer in region “Other”, from cluster 3 that has most of the annual spending in the ‘Grocery’ category

- The business is probably an actual grocery store (e.g. Costco, King Soopers, etc.) or places like gas station convenience stores.

Hierarchical Clustering (agglomerative)

Customer Segment 1:
A HORECA customer in region “Other”, from cluster 0 that has most of the annual spending in the ‘Fresh’ category

- This could likely be a HORECA business that specializes in Fresh products, possibly a farmers’ market

Customer Segment 2:
A HORECA customer in region “Oporto”, from cluster 0 that has most of the annual spending in the ‘Frozen’ category

- This could likely be a HORECA business that specializes in Frozen products, possibly an ice cream bar

Customer Segment 3:
A non-HORECA customer in region “Other”, from cluster 3 that has most of the annual spending in the ‘Grocery’ category

- A non-HORECA business, likely an actual grocery store (e.g. Costco, King Soopers, etc.)

Discussion

Notice, the customer segments clustered in each type of clustering algorithm are the same data points. The only difference being how the color-coded clusters are distributed, and axis orientation.

So the question now becomes **“which algorithm is more appropriate given the context?”** Based on the task at hand, I believe the **K-means clustering addresses those needs best**.

The issue with only having 2 clusters (*i.e. Hierarchal Clustering results*) is there’s too much overlap between the 2 identified clusters. Additionally, there are data points so different from one another that intuitively one would assume they’d be in different clusters.

As shown, with 4 clusters, the data is much more intuitive to segment and therefore interpret.

In addition to the customer segments described and identified, it can also be observed that 2 clusters present can be considered to be wholesale customers who’s needs are not met by the current product supplier. While not possible with the current dataset, more depth can be added into the analysis if supplemental data containing actual item names and prices can be included. Where the needs not being met by those 2 other clusters (i.e. cluster 0 and 2) could potentially be identified.

References

- Cardoso, M. (2014). Wholesale customers Data Set. Retrieved from UCI Machine Learning Repository.
- Dalton, D. (2023). Data Science Captsone [COMP 4449]. University of Denver.