Wholesale Customer Segmentation

Andrew J. Otis

Task to Accomplish

To identify and describe potential customer segments hidden within the data…

The Data

Encoded 'Region' Column
1 = Lisbon
2 = Oporto
3 = Other Region

Encoded 'Channel' Column
1 = HORECA
2 = Other Region

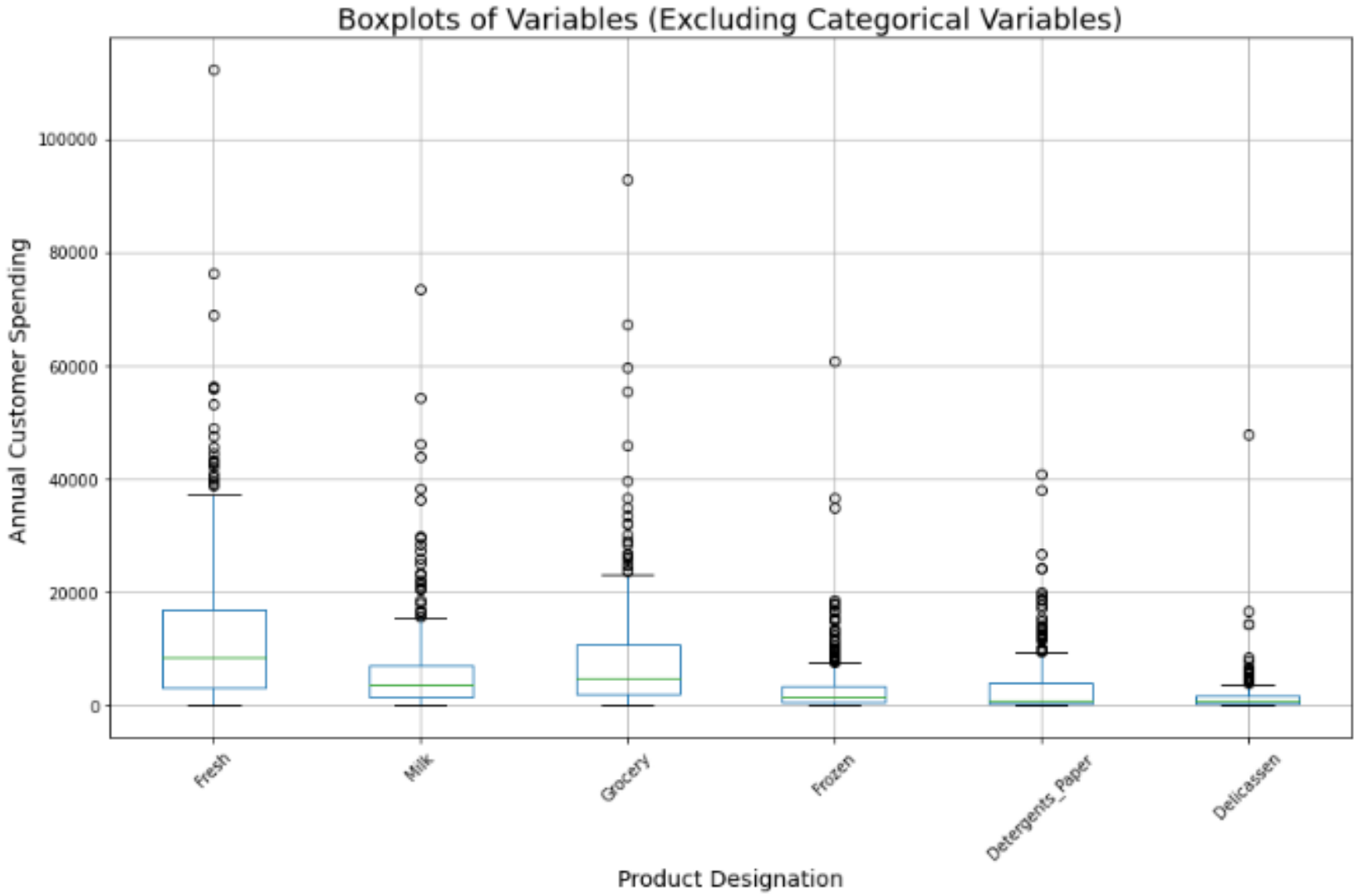
… based on annual spending on diverse product categories (*e.g. Grocery, Frozen, etc.*) of various business types (*i.e. Hotel, Restaurant, and Café grouped as 'HORECA' and 'Other'*) located in various regions.

	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

Summary Stats (boxplot)

Figure 1: The plot visualizes the summary statistics of the data (e.g. range, variance, etc.)



Feature Distributions

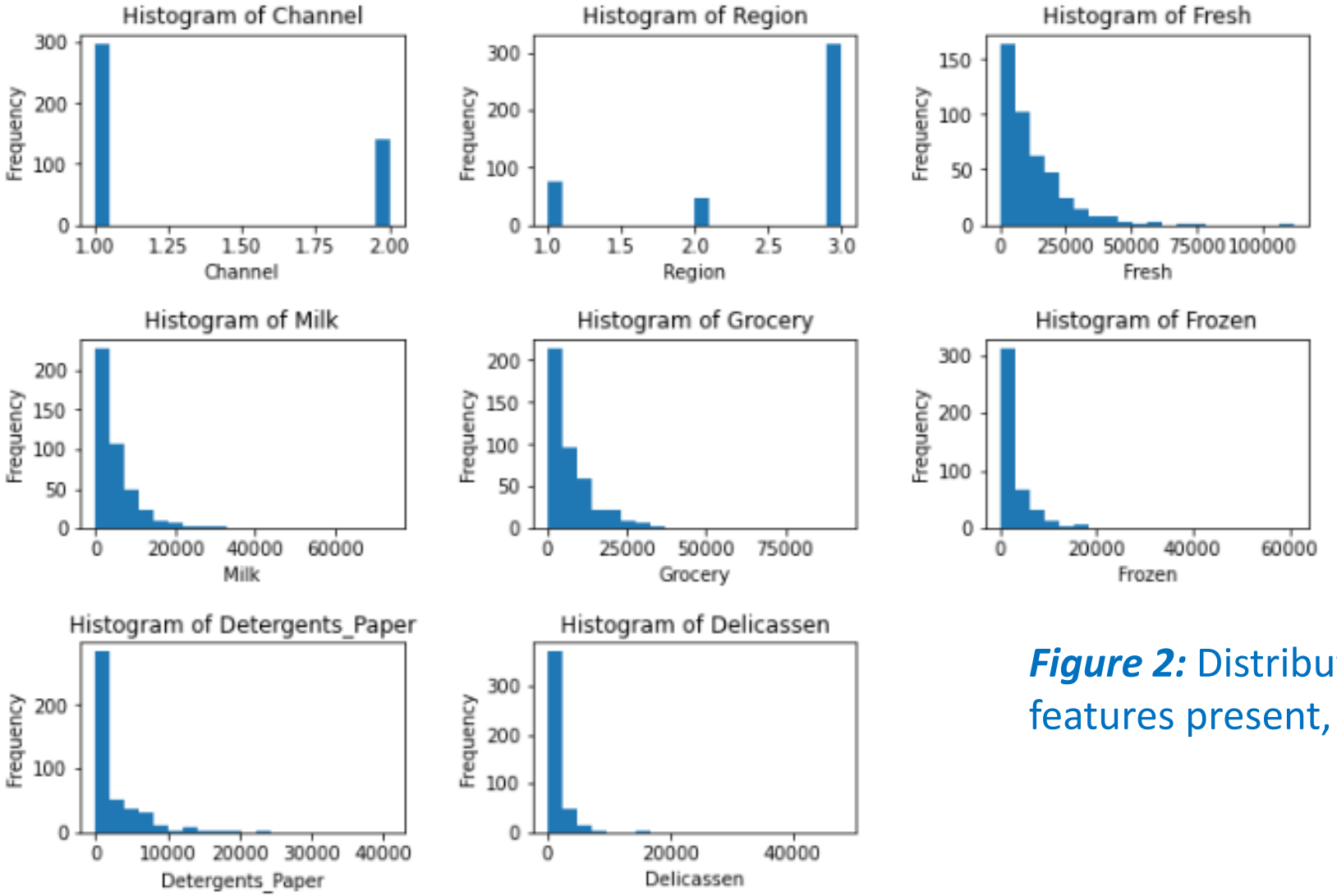


Figure 2: Distribution of all features present, respectively.

Principal Component Analysis

The purpose is to reduce dimensions of the data to the most important features (*i.e. principal components*) to be inputs for our clustering models.

For models that require specifying clusters manually(*e.g. K-means*)

- **A model with 1 categorical & 2 numeric features :**

The candidate features for clustering are the product categories 'Grocery' and 'Frozen'

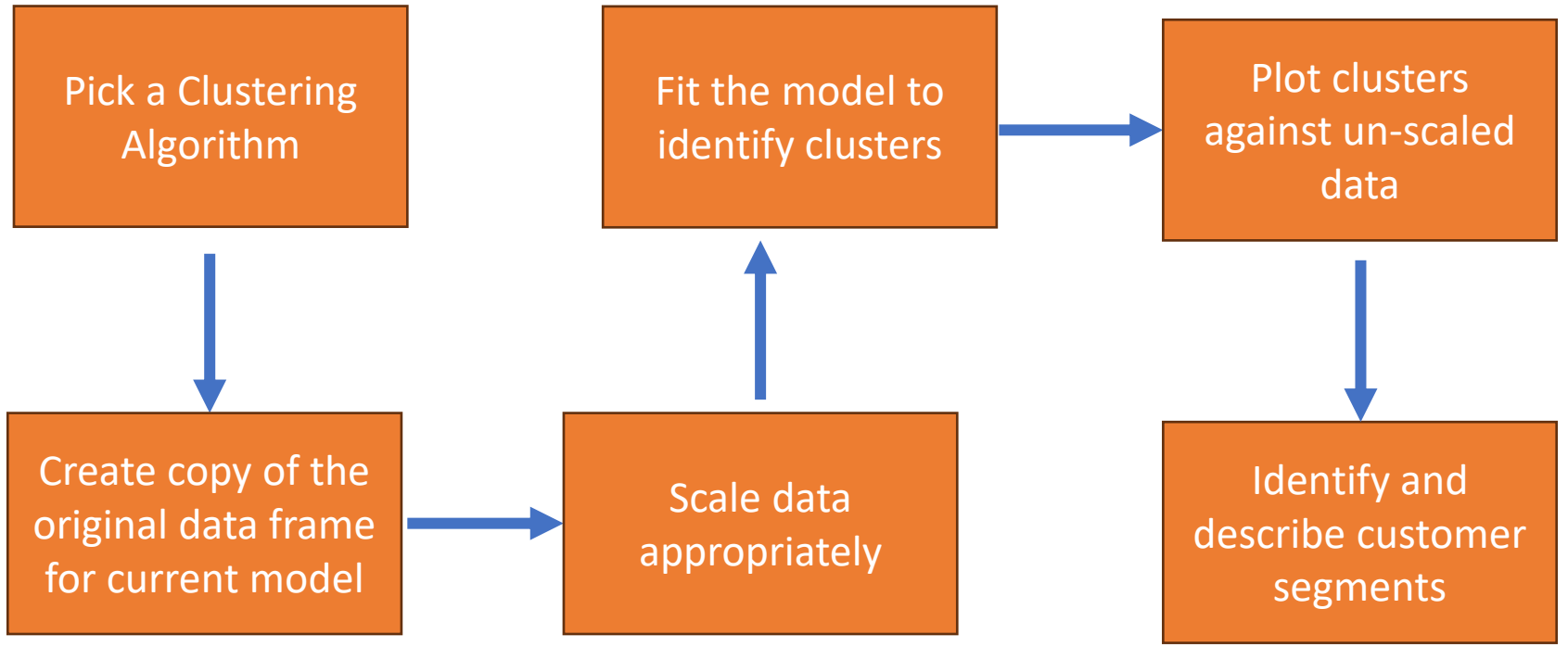
- **A model with 1 categorical & 3 numeric features:**

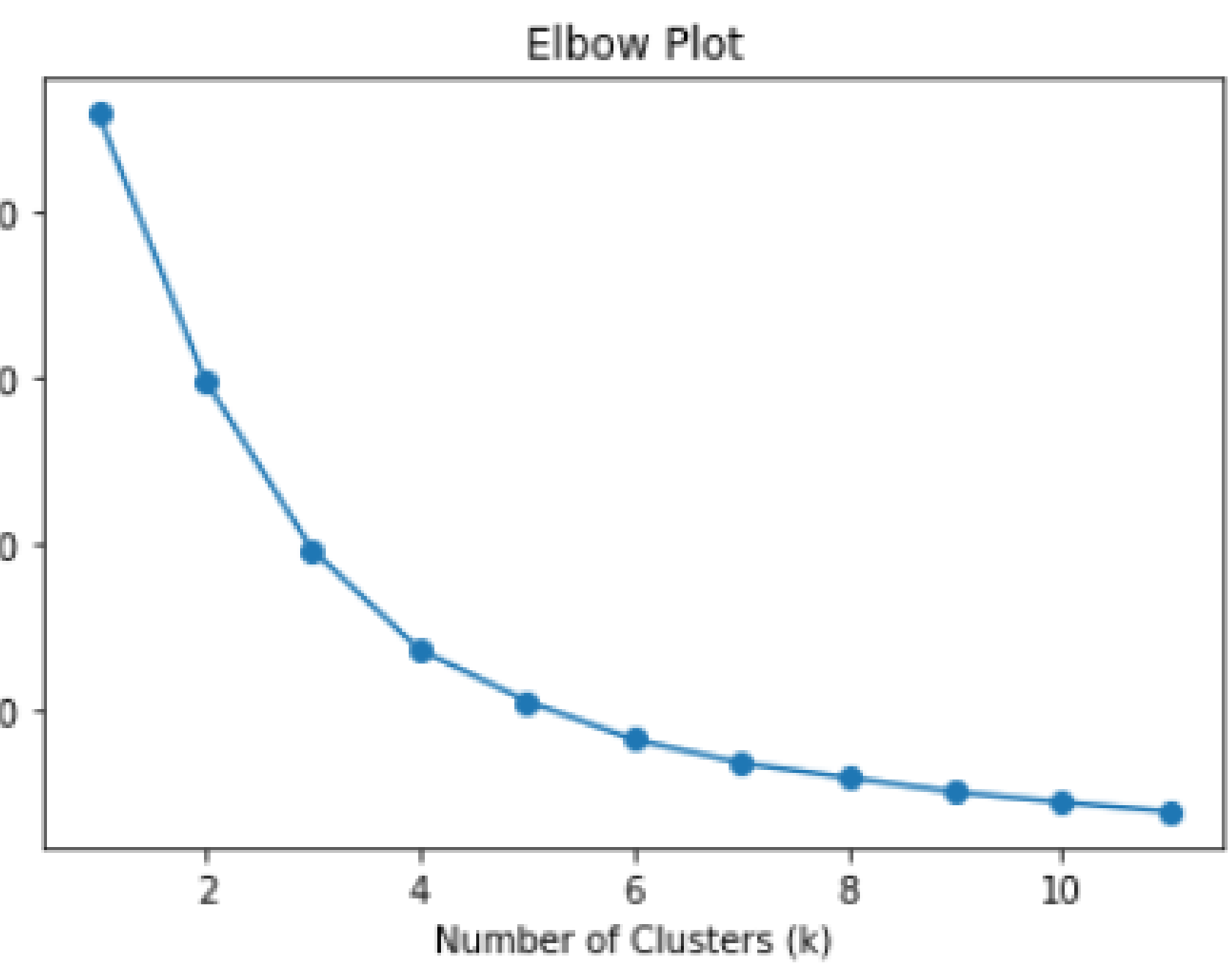
The candidate features for clustering are the product categories 'Grocery', 'Frozen', and 'Fresh'

Clustering Algorithms and Strategy

1. **K-Means Clustering**
2. **Hierarchal Clustering (agglomerative)**

Unlike supervised ML models, clustering only attempts to identify patterns in data, and thus has no “target” variable.



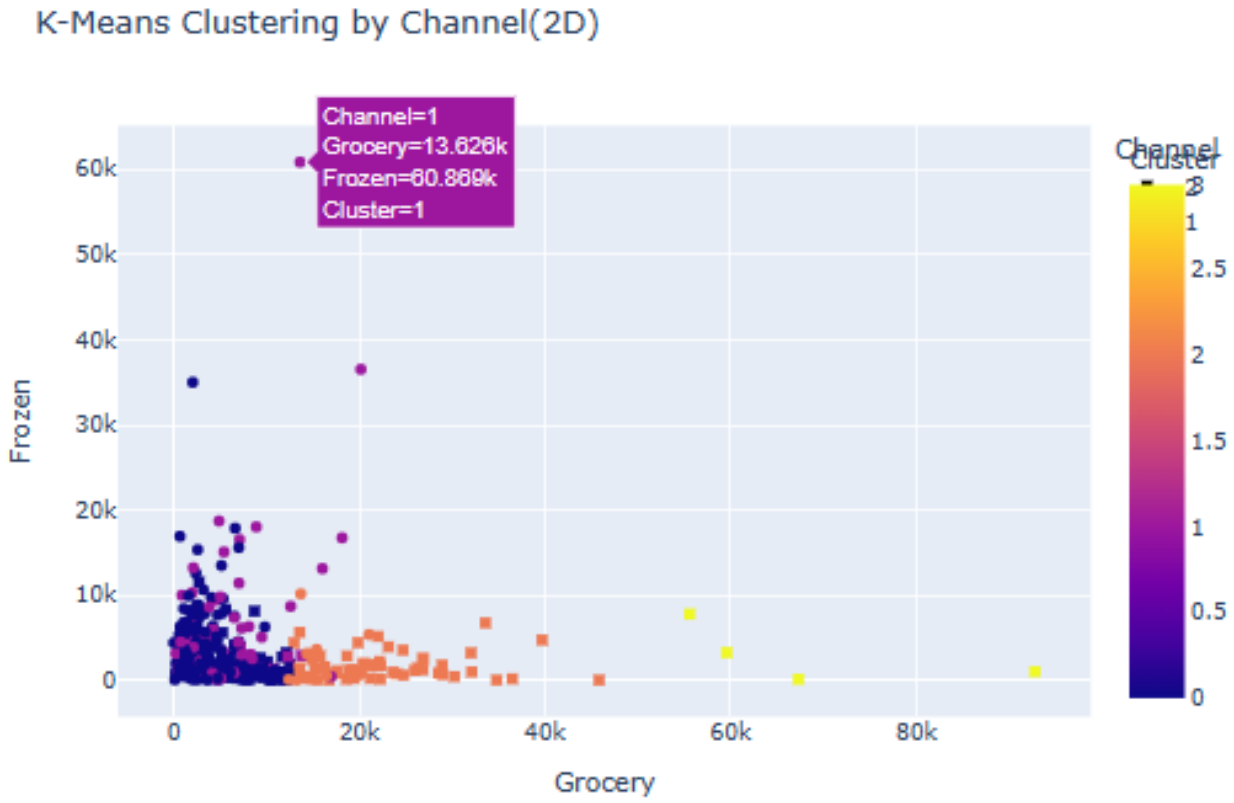


Elbow Plot for K-Means

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.

Visualize the K-means Clusters in 2D

Notice, analyzing the graphs in tandem allows us to compare regions and channels simultaneously



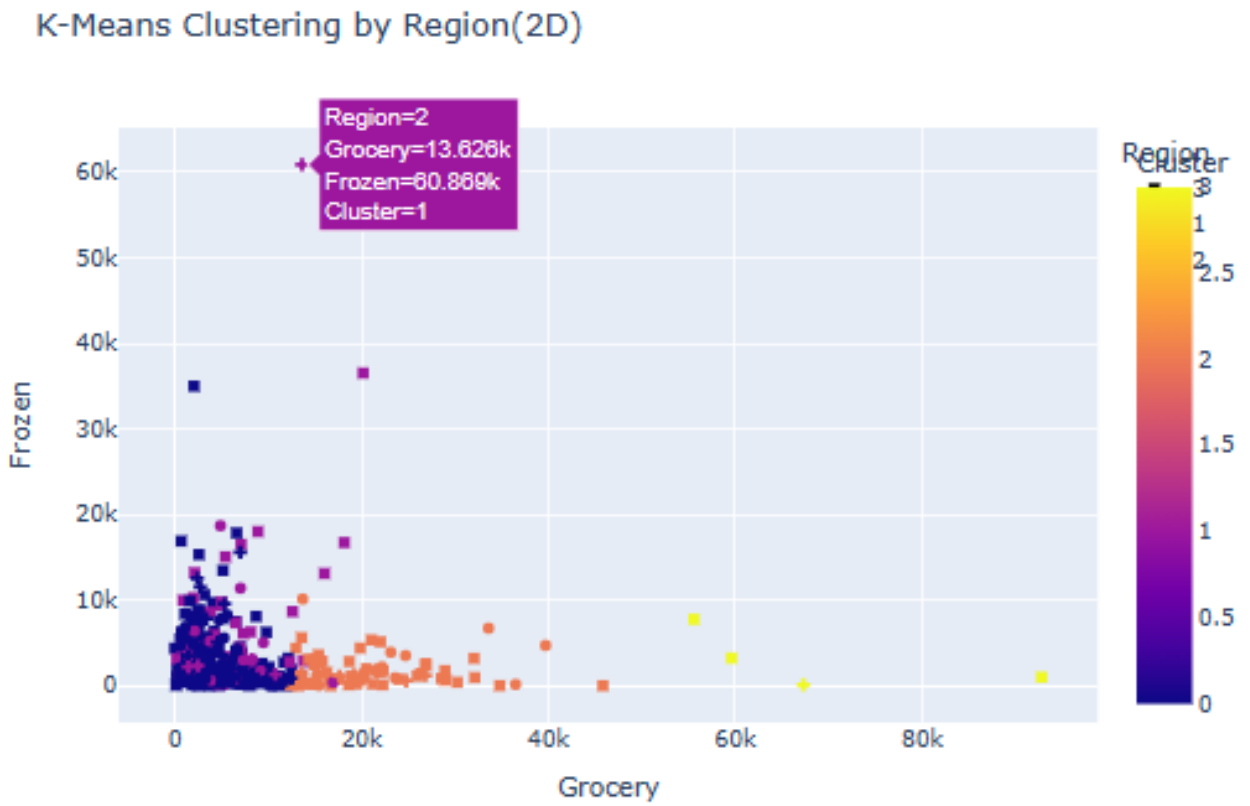


Figure 4: Clusters visualized by Channel and Region, for 2D scatterplot representation, respectively.

Visualize the K-means Clusters in 3D

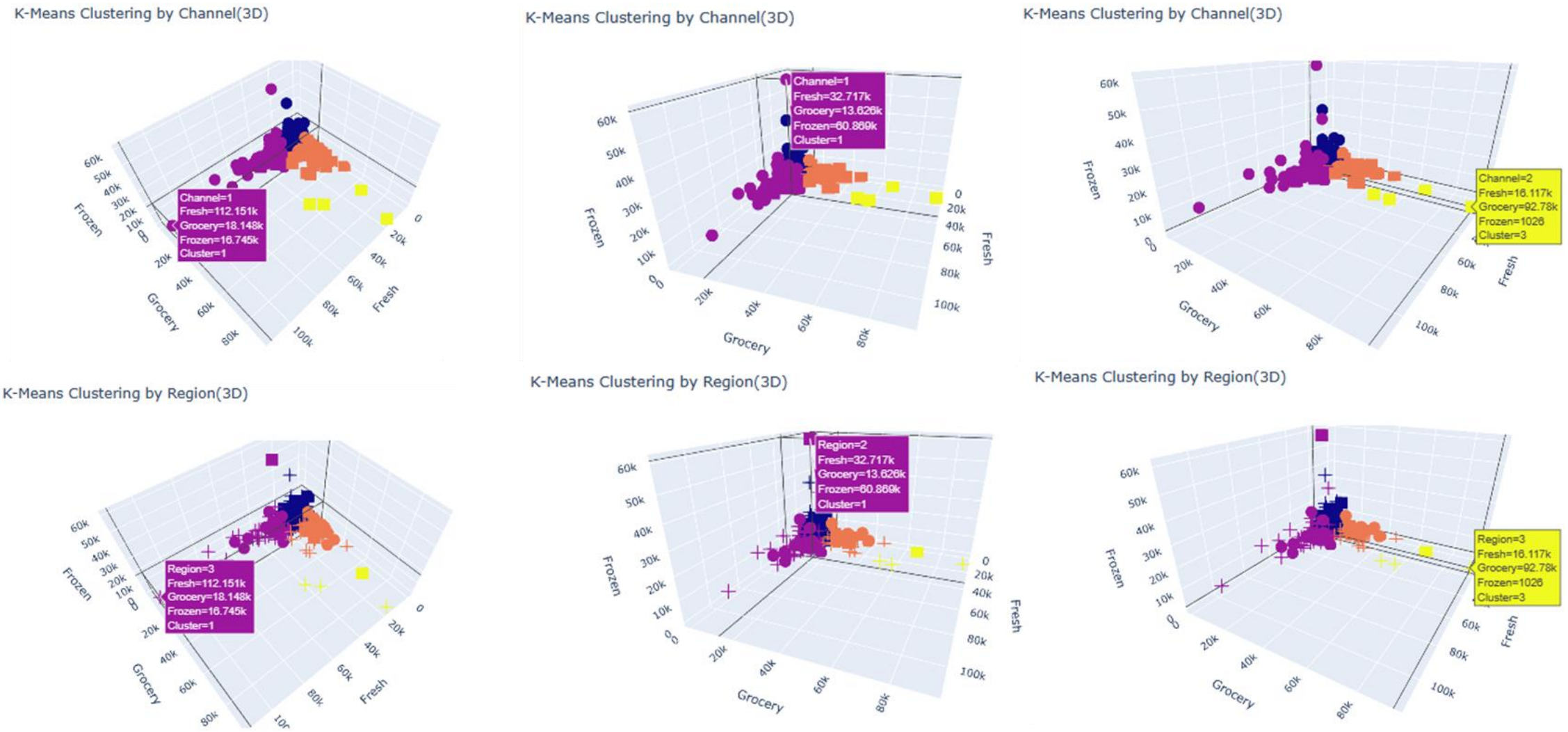


Figure 5: Clusters visualized by Channel and Region, for 3D scatterplot representation, respectively.

Customer Segmentation from K-Means Clustering

- Segment 1

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

- Segment 2

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

- Segment 3

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

Dendrogram from Agglomerative Clustering

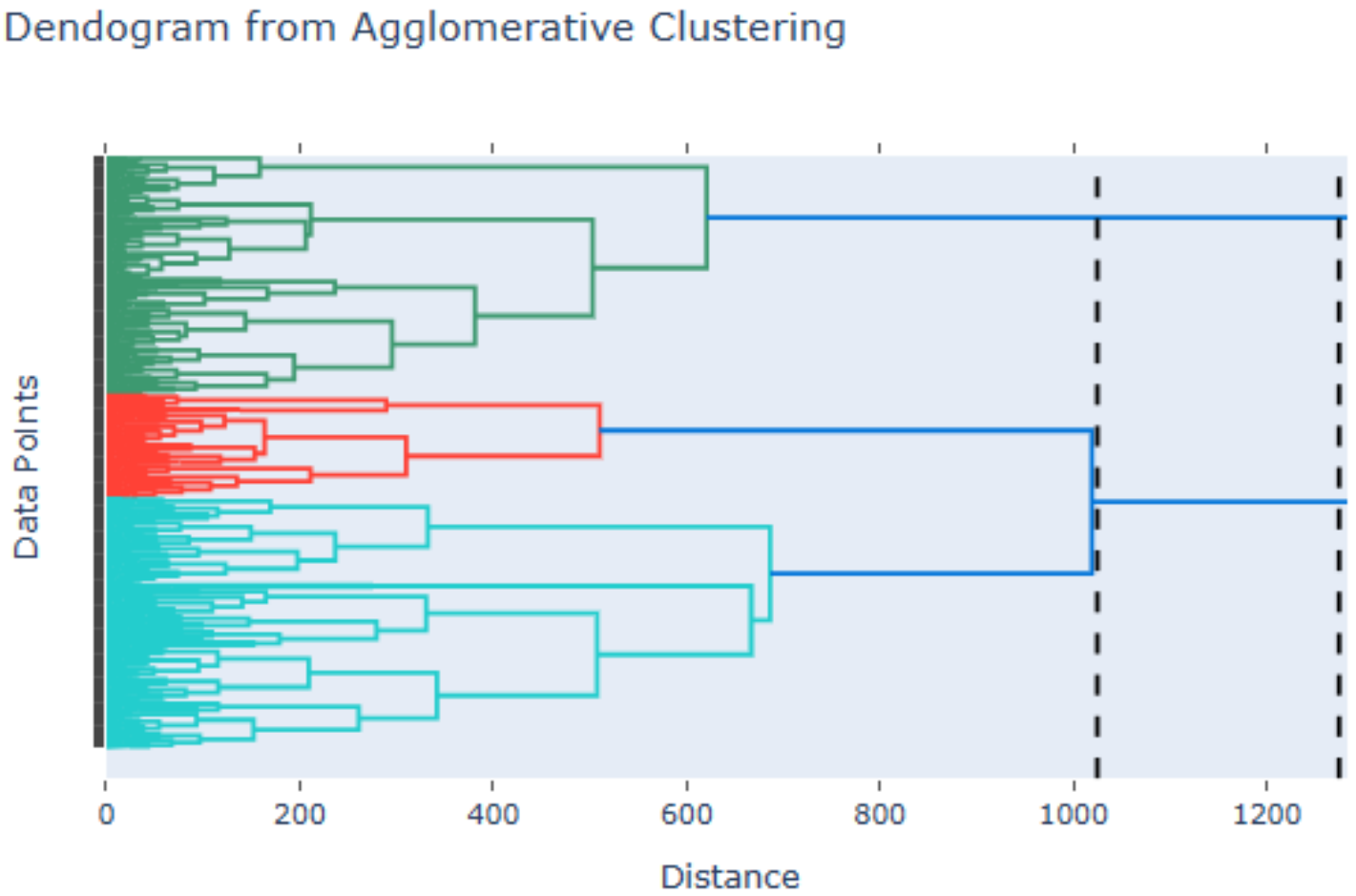


Figure 6: Algorithm automatically clusters the data, no need to explicitly specify the desired number of clusters. In this case, the algorithm determined 2 clusters to be the ideal number. This is supported by the only cluster labels resulting being “0” & “1”

Customer Segmentation from Hierarchal Clustering

- Segment 1

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

- Segment 2

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

- Segment 3

A non-HORECA customer in region “Other”, from cluster 1 that has most of the annual spending in the ‘Grocery’ category

Model Comparisons

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 overlapping between the 2 clusters identified.

Additionally, there are data points so different from one another that intuitively one would assume they'd be in different clusters.

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.