Unsupervised Learning Project

Samuel Croft

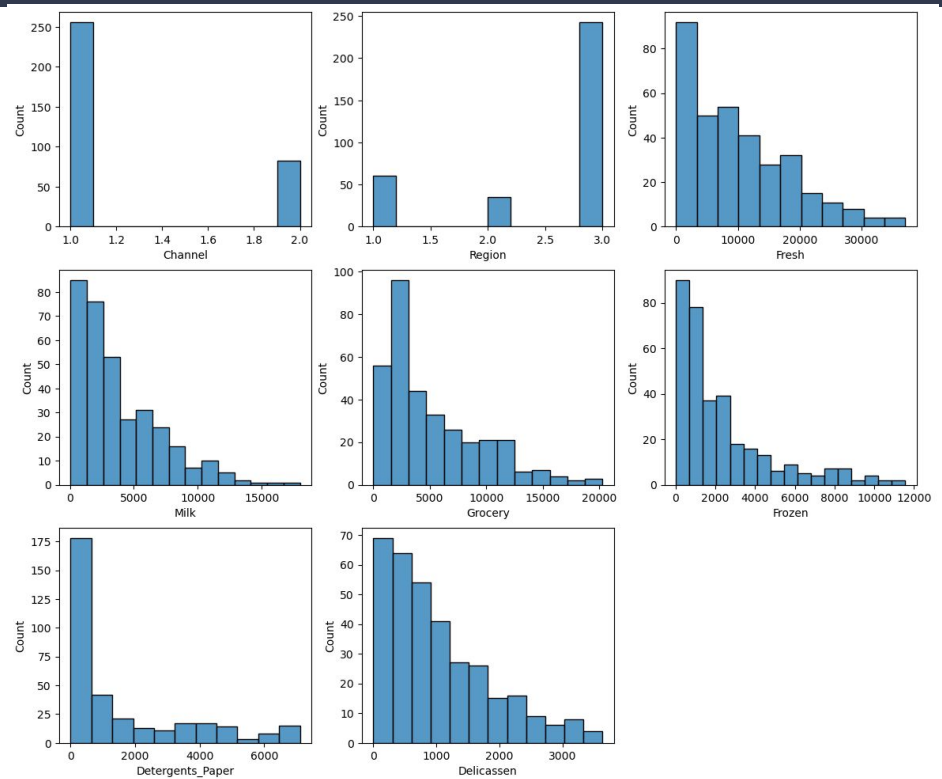
LHL Data Science (May 29th Cohort)

Project Goals

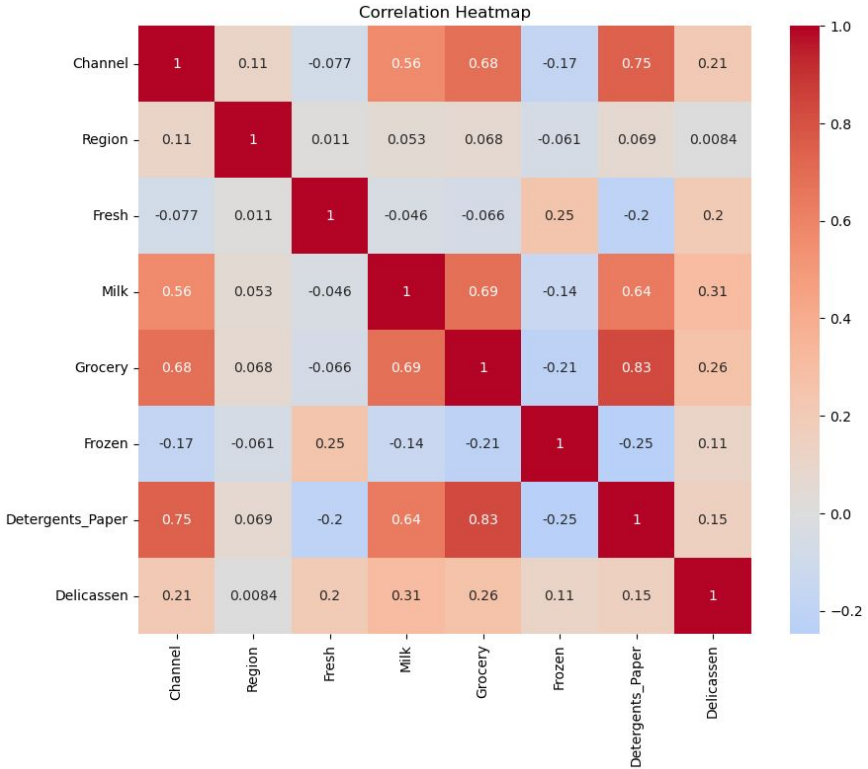
1. Perform exploratory data analysis and pre-processing on the Wholesale Customers Data Set
2. Perform K Means clustering, determine optimal value for K, and converge on cluster centroids
3. Perform hierarchical clustering, and confirm optimal value for K
4. Perform Principal Component Analysis and determine how to best reduce the number of features

EDA and Preprocessing

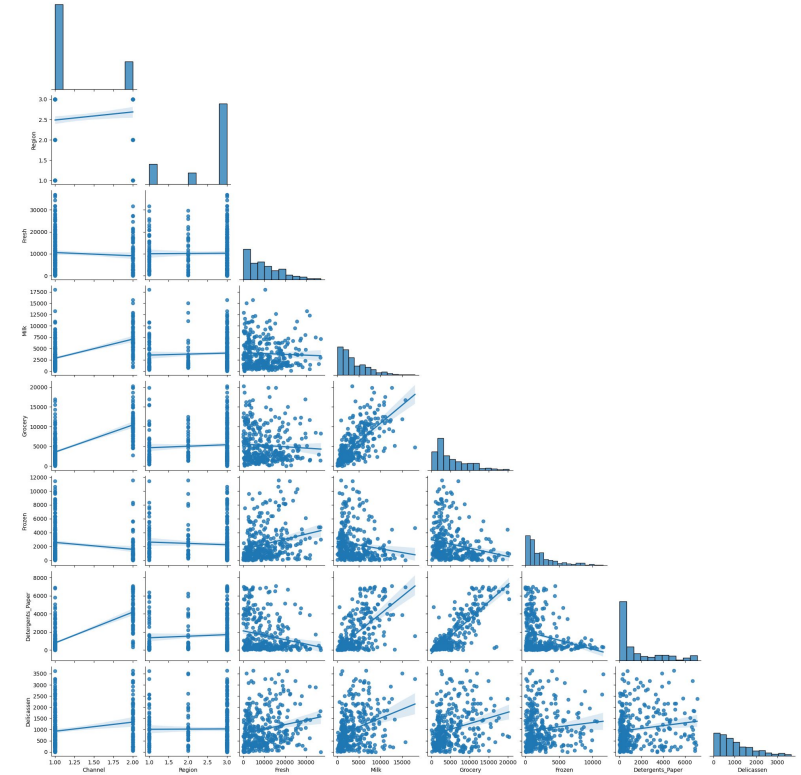
EDA – Distributions of Features

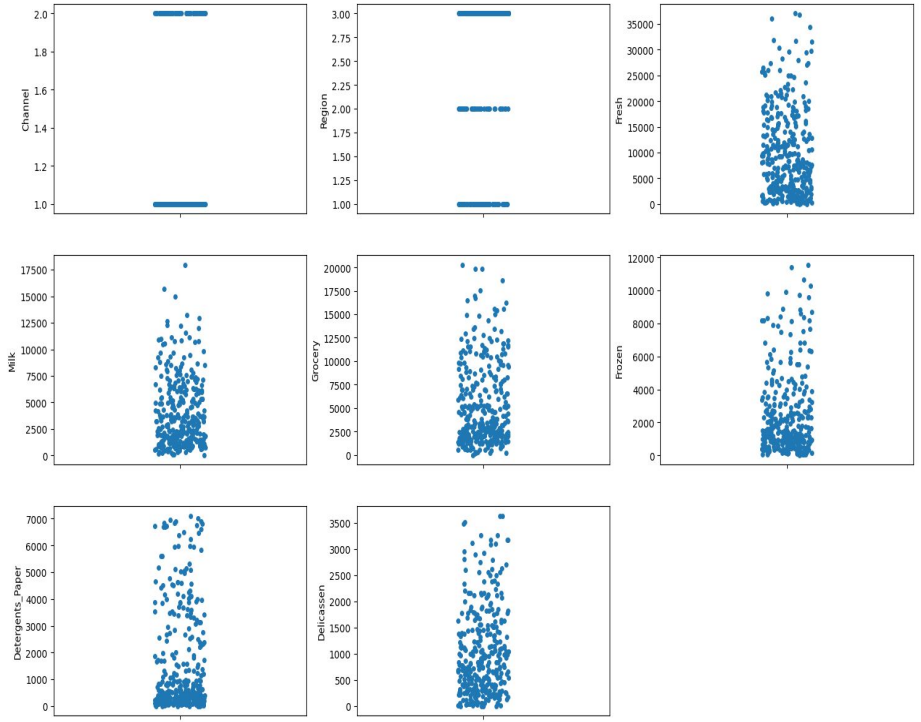


EDA – Correlation Heatmap



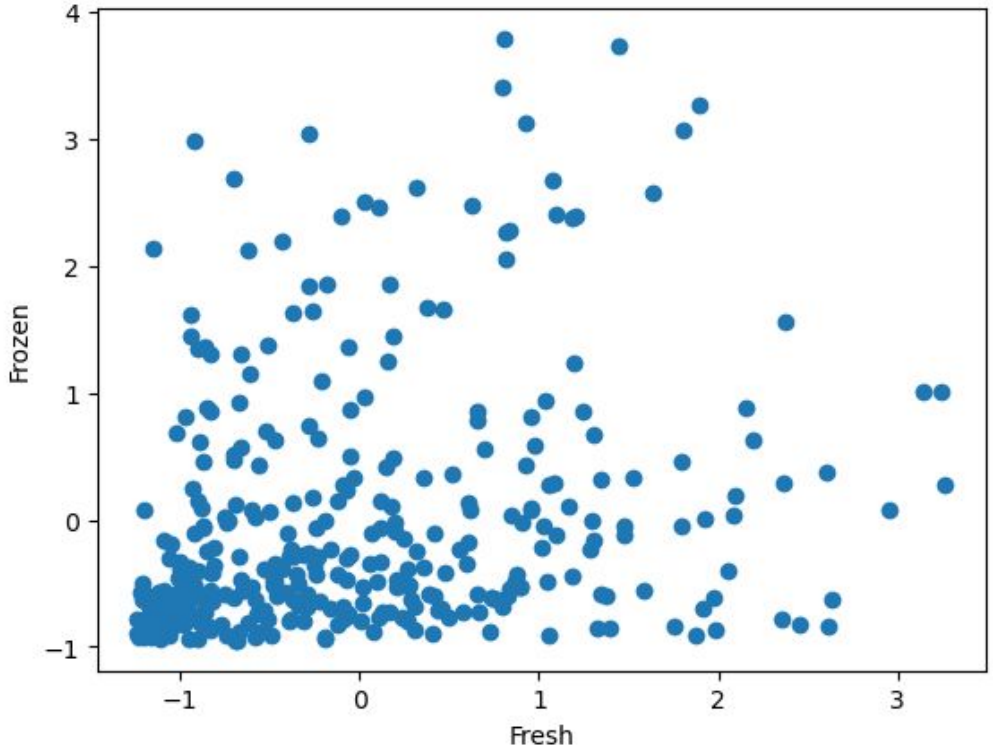
EDA – Correlations and Strip Plots



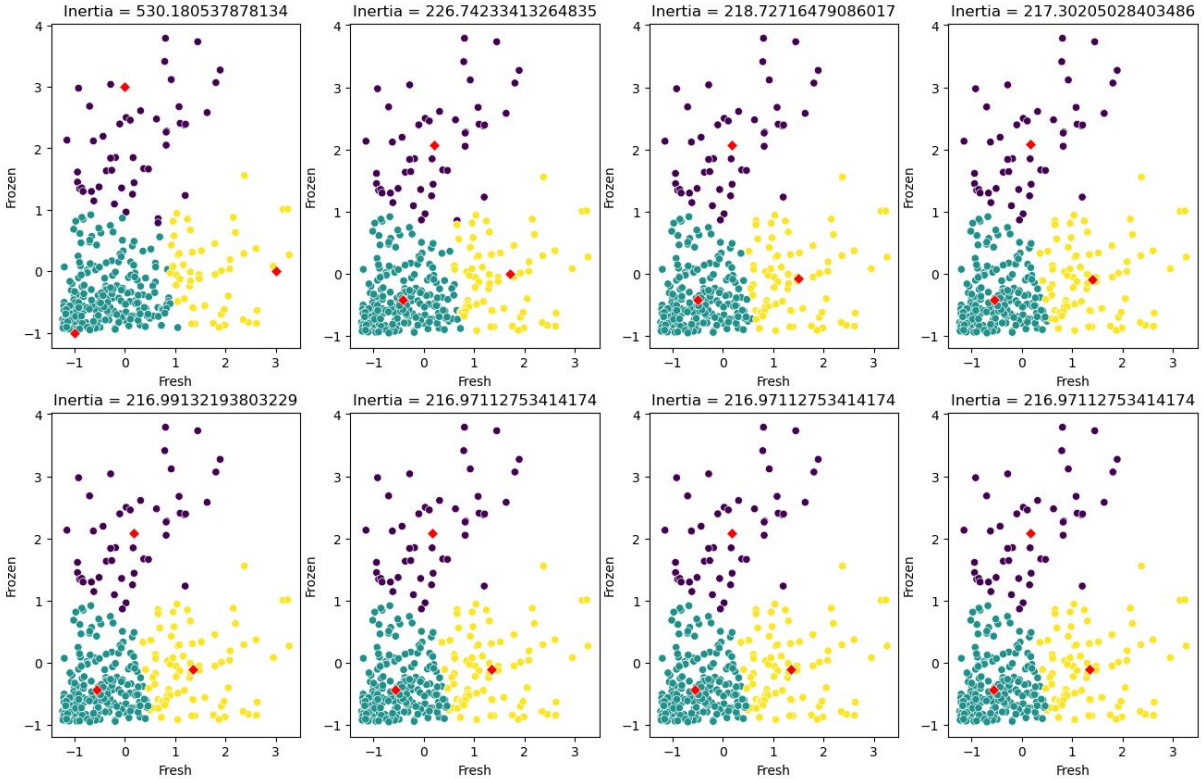


K Means Clustering

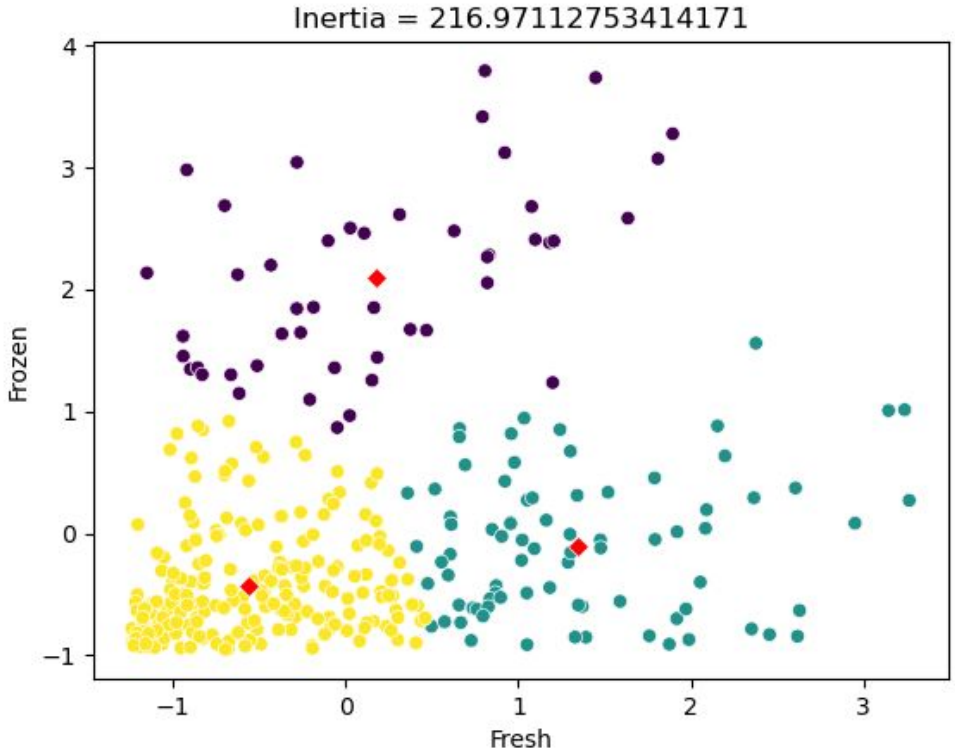
K Means - Fresh vs Frozen



K Means - Manually Converging on Centroids

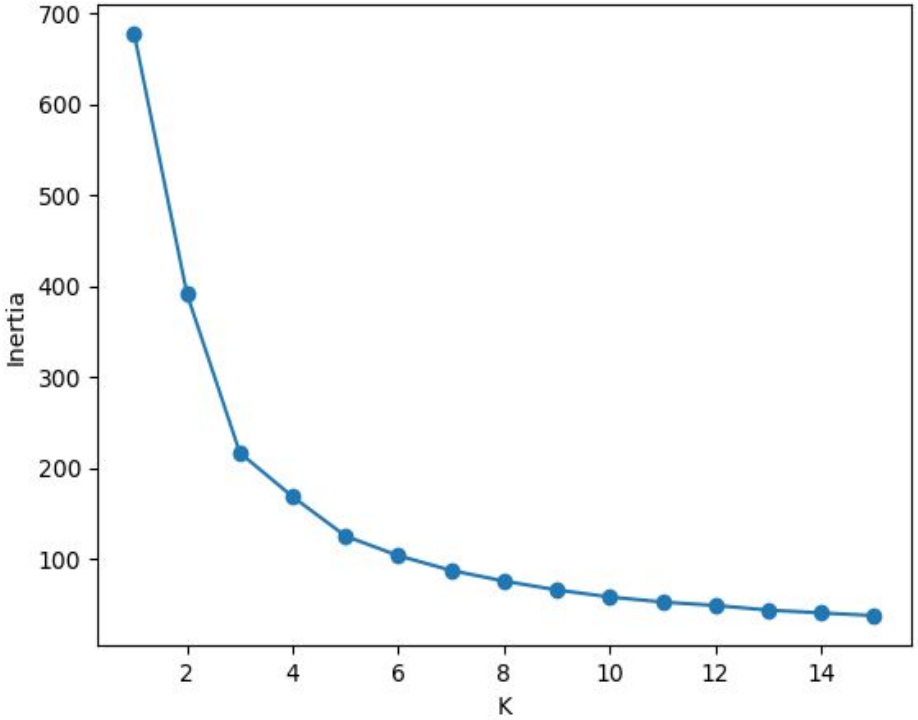


K Means - Using KMeans Function



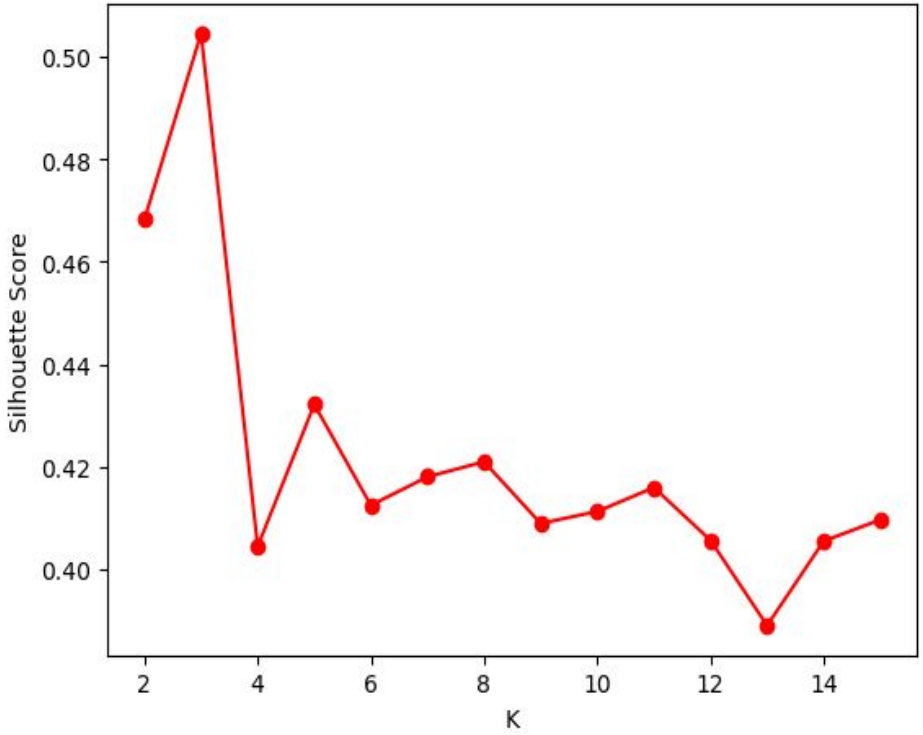
K Means - Optimizing K

The “elbow” occurs at K = 3



K Means - Optimizing K

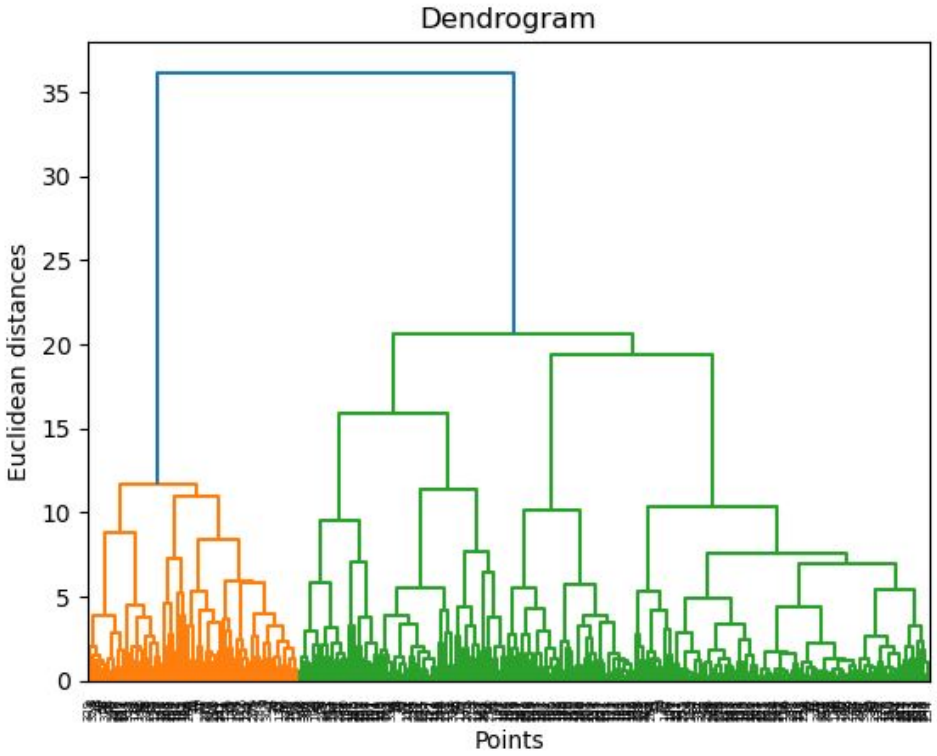
The silhouette score is maximized at K = 3



Hierarchical Clustering

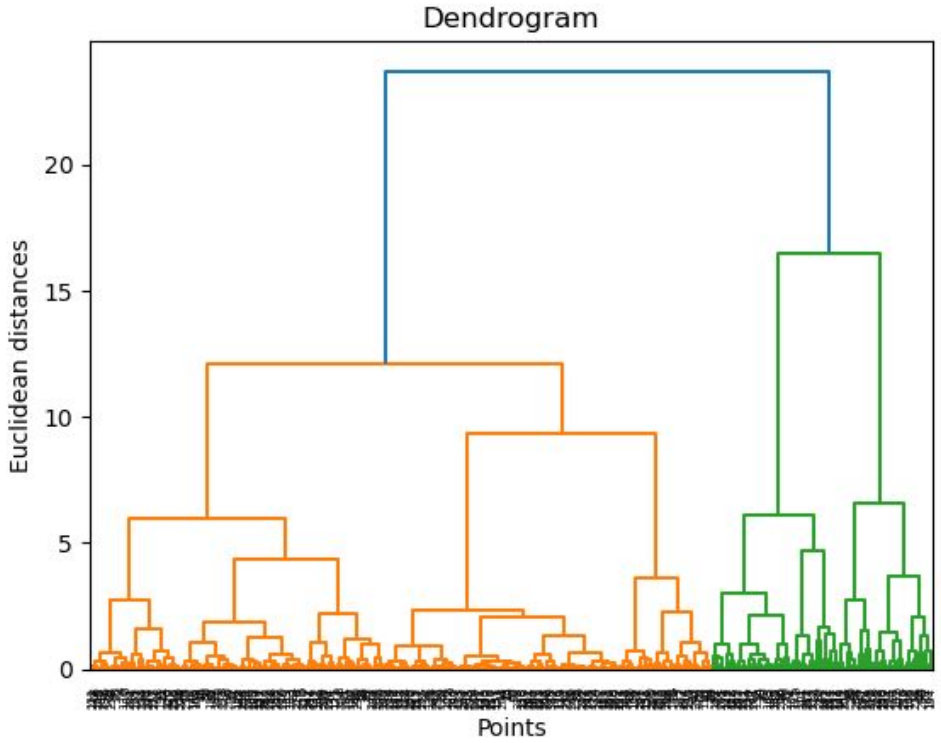
Hierarchical Clustering - Dendrogram

When using all discrete features, 2 clusters will give us the best results

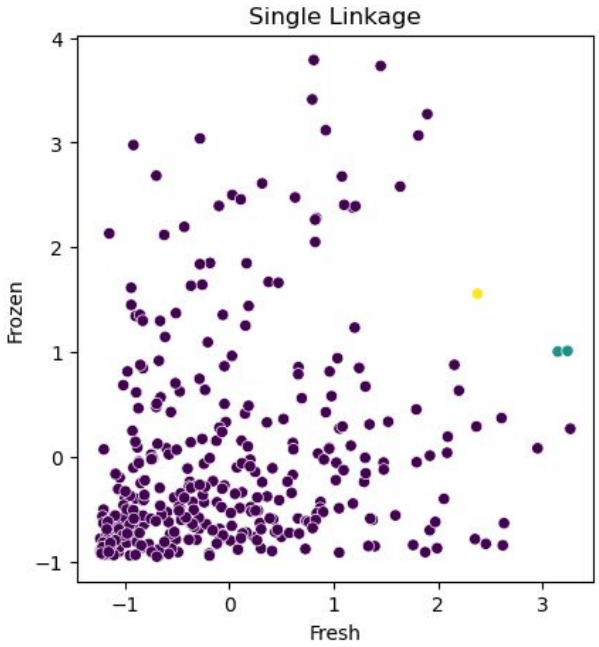


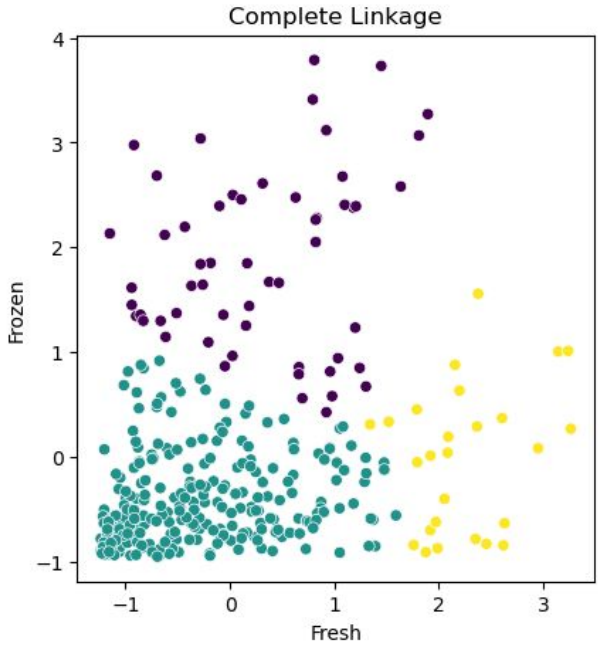
Hierarchical Clustering - Dendrogram

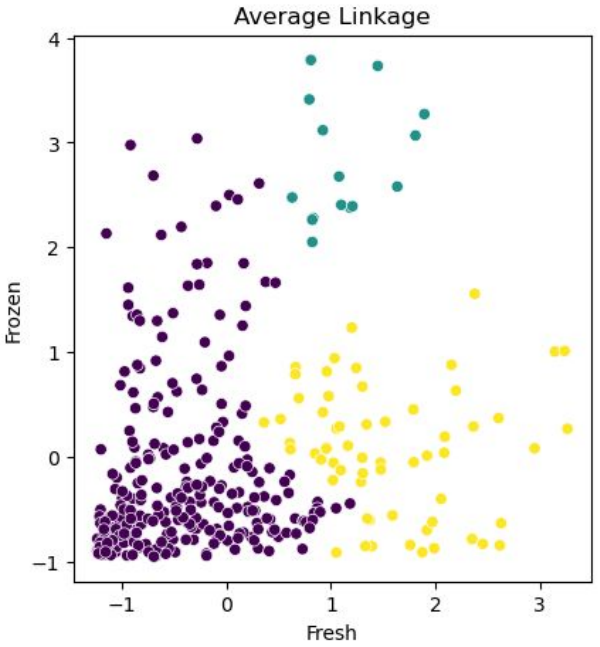
When using only the Fresh and Frozen features, 2 or 3 clusters could be used



Hierarchical Clustering - Linkage Methods



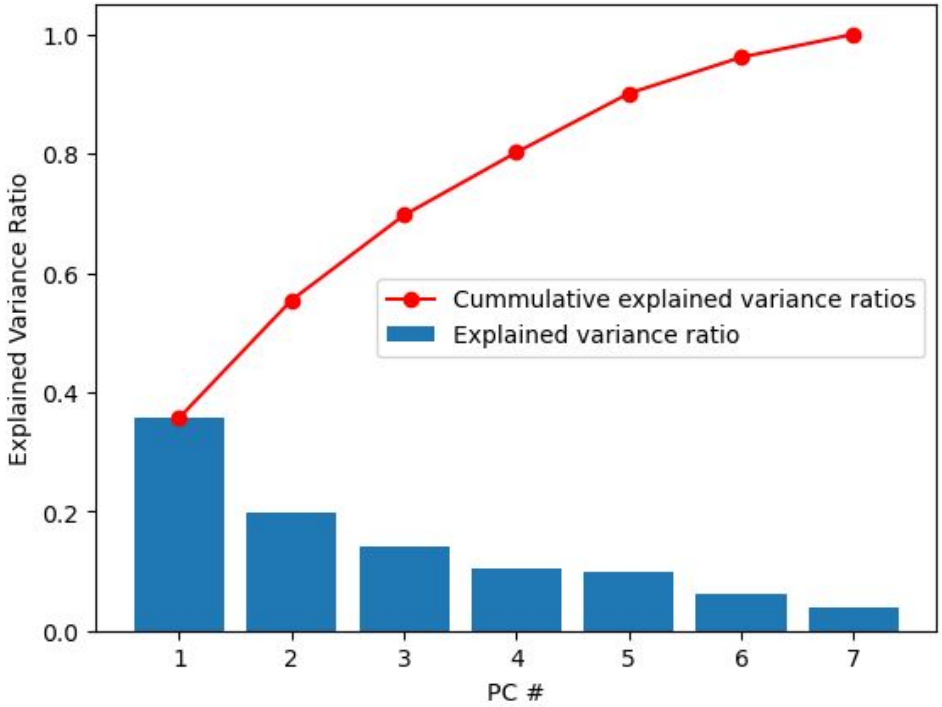




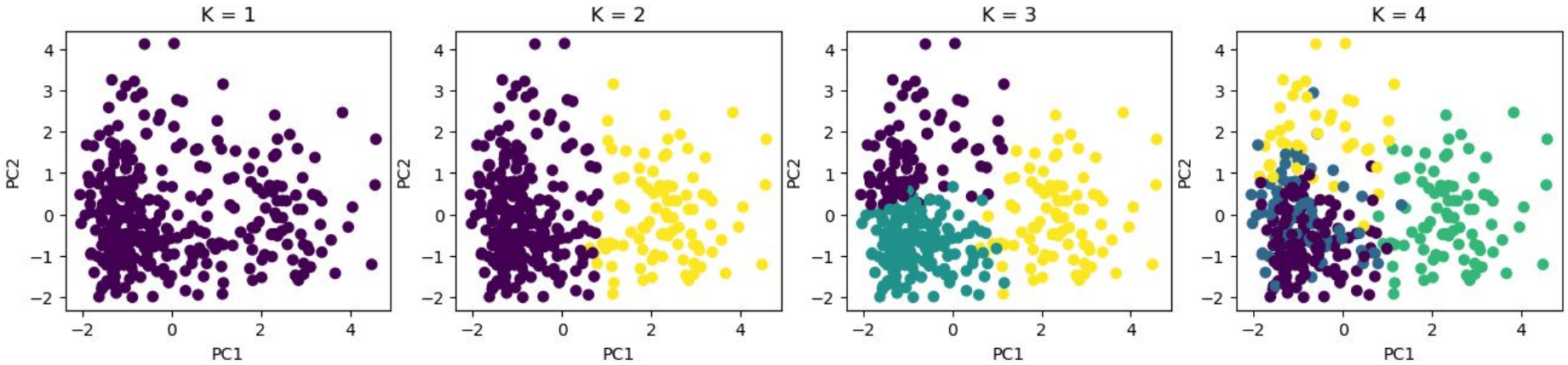
Principal Component Analysis (PCA)

PCA – Scree Plot

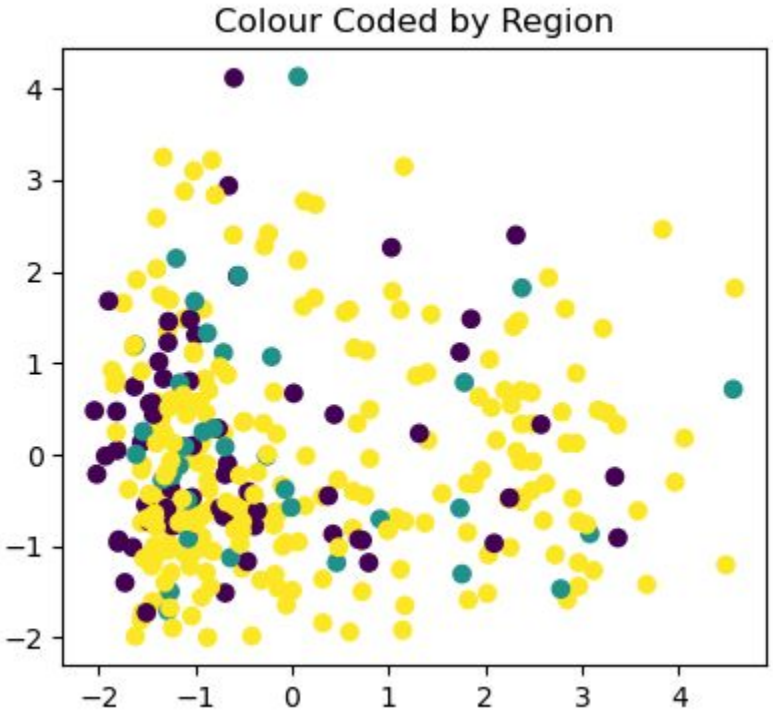
PC's 1 and 2 together account for ~55% of explained variance

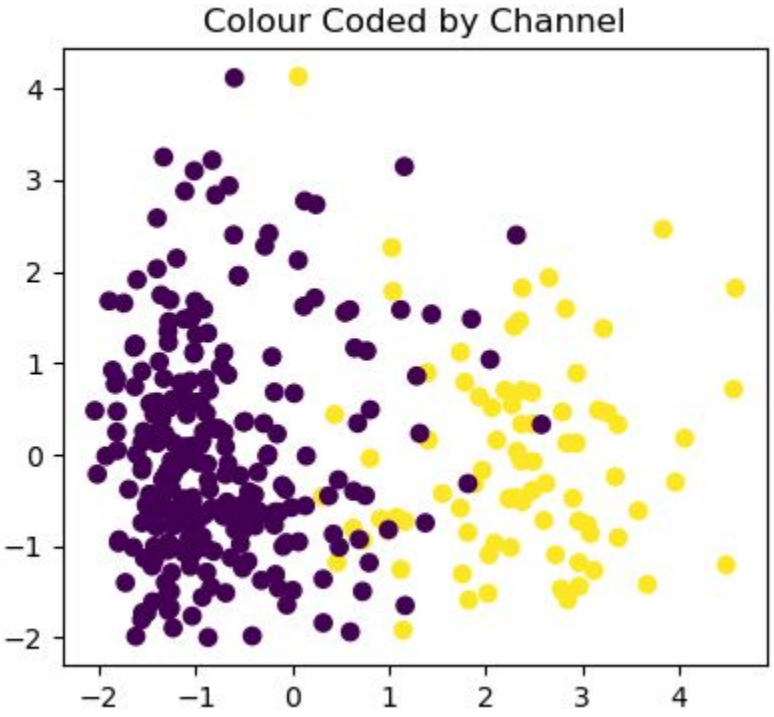


PCA – Colour coding by K Means clusters



PCA - Colour coding by Region and Channel





Conclusions

1. The optimal value for K when only looking at features Fresh and Frozen is 3. If we assume that each row is a grocery store (customers of a wholesaler), these three clusters could be:
    a. "healthy" grocery stores that specialize in fresh food,
    b. big chains like Walmart and No Frills that make large selections of frozen food, and
    c. grocery stores that sell a variety of both fresh and frozen.

2. When using all of the discrete features, the optimal value of K is either 2 or 3. Based on the Dendrogram, K = 2 is better.

3. The discrete values can be used to fairly accurately estimate the Channel.

4. Reducing the discrete features to 2 principle components can still produce an accurate clustering model.