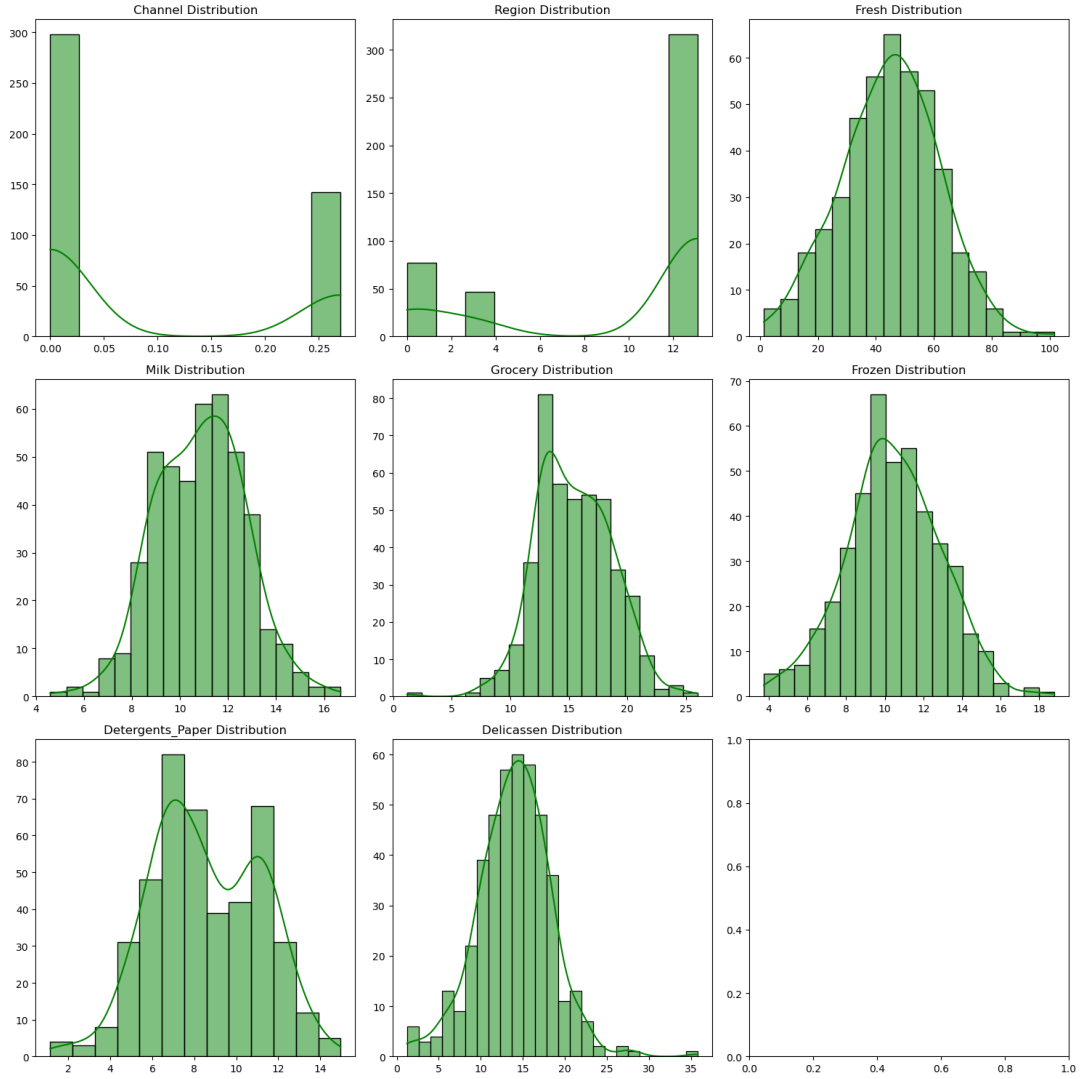
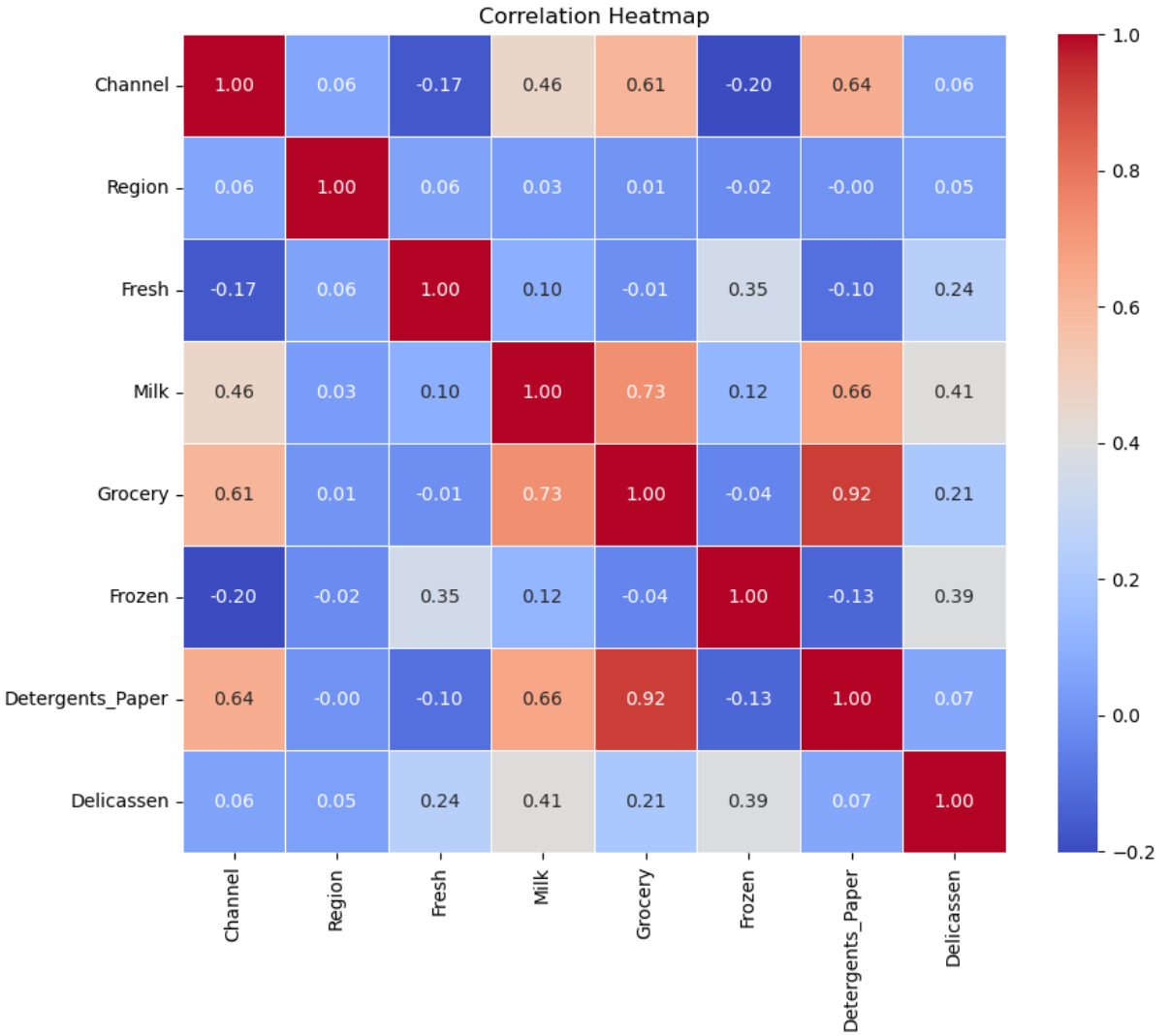
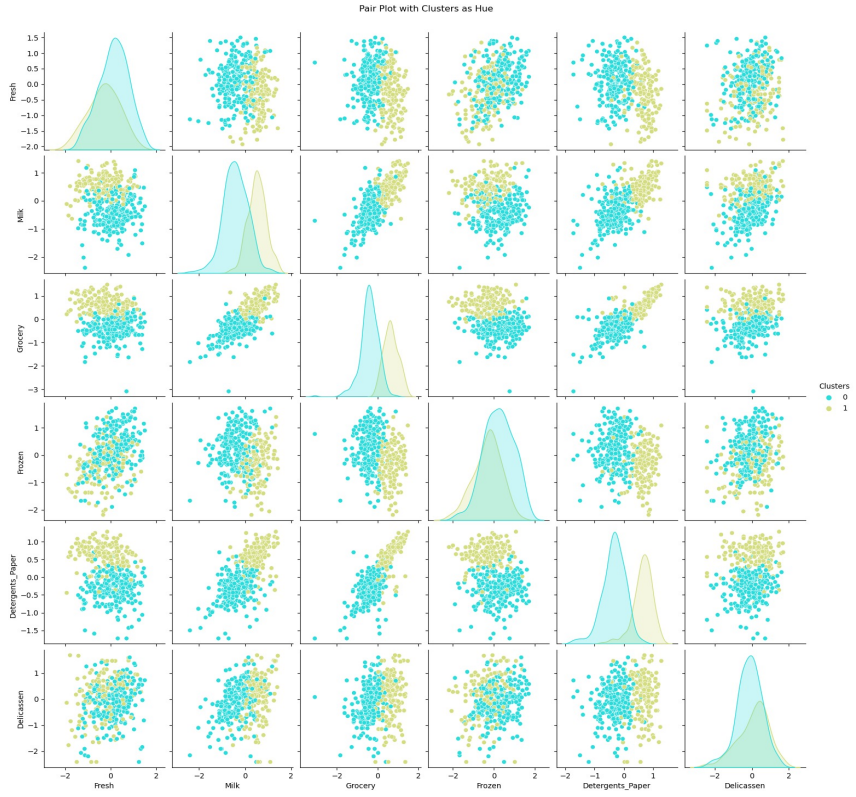
UNSUPERVISED LEARNING PROJECT

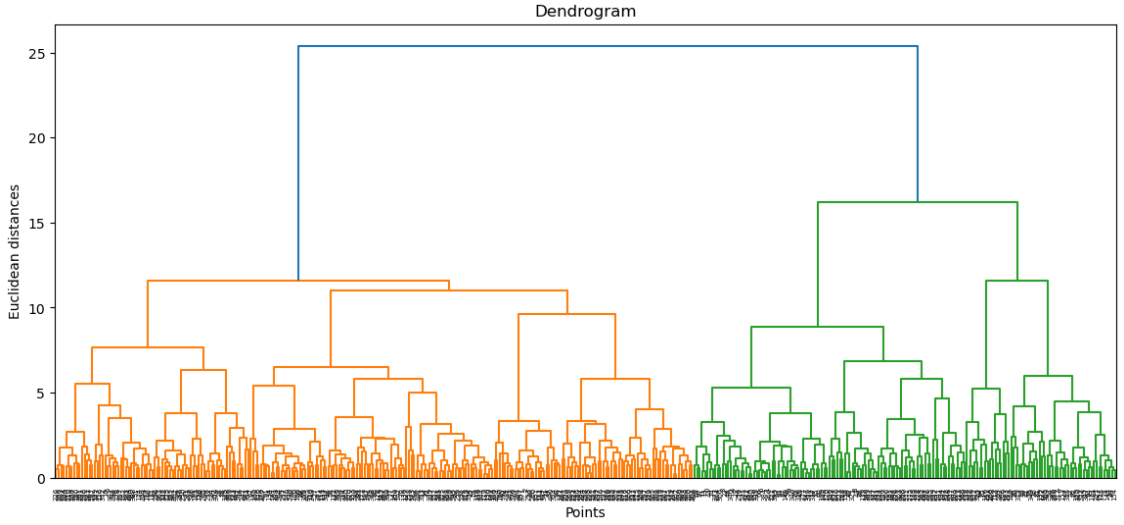
Skyler Wilson

EDA - Visualizations

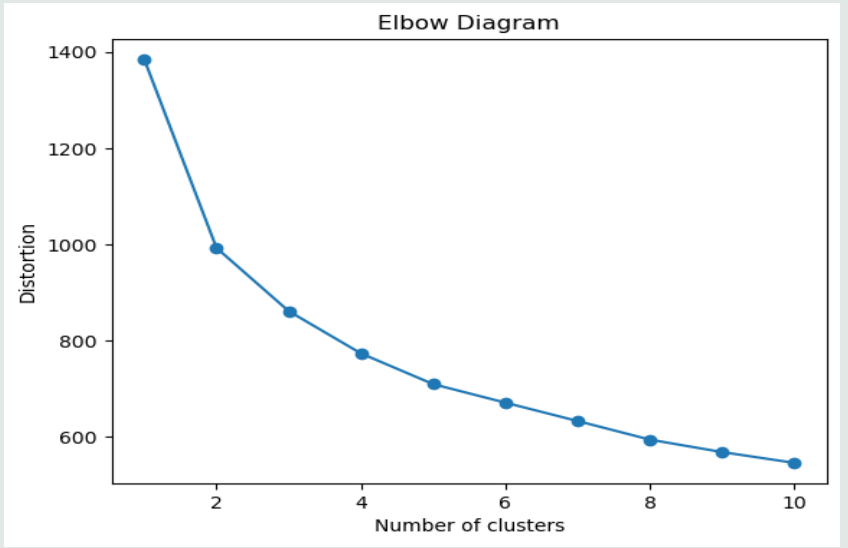


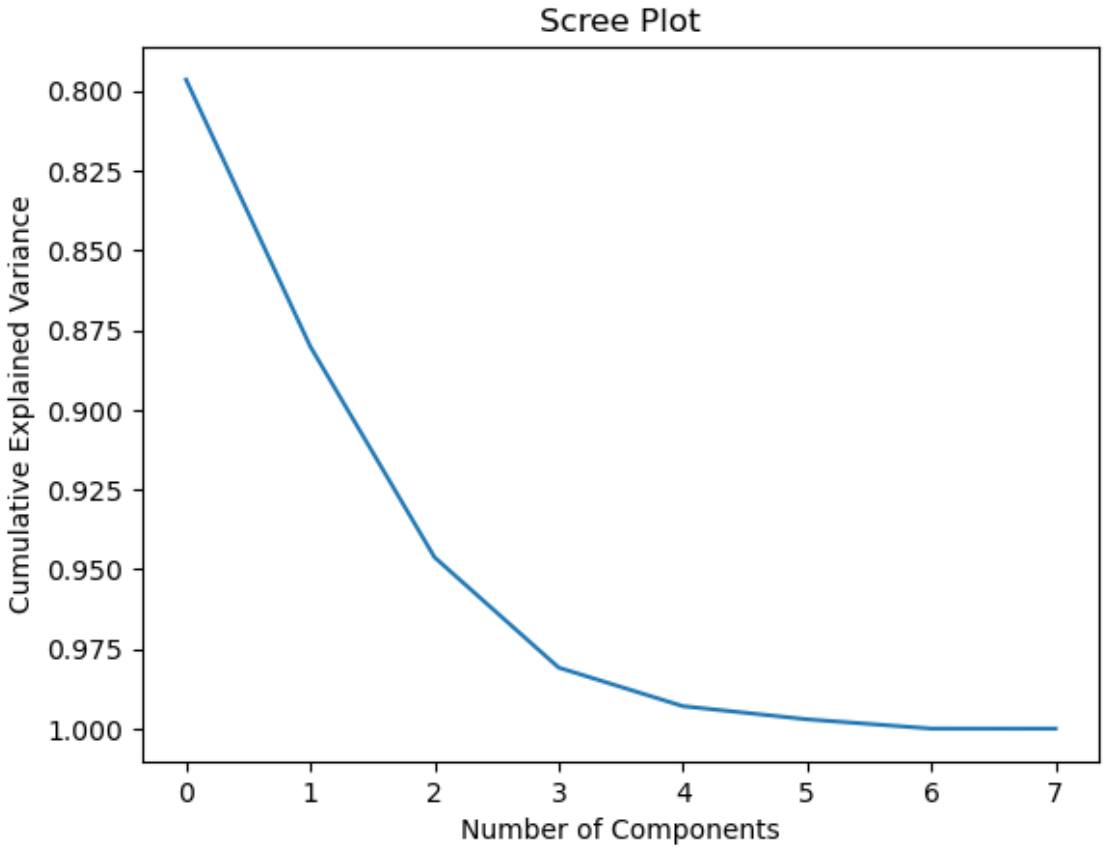






Cluster Visualizations





	Percentage Explained Variance (%)	Cumulative Explained Variance (%)
Component 1	79.646222	79.646222
Component 2	8.367752	88.013974
Component 3	6.611339	94.625313
Component 4	3.461448	98.086761
Component 5	1.208817	99.295577
Component 6	0.412290	99.707867
Component 7	0.290274	99.998142
Component 8	0.001858	100.000000

PCA Visualizations

Analysis of Results

- Based on the results of the K-means and hierarchical clustering, the optimal number of clusters for the model was 2. The K value was calculated using a silhouette calculation and gap statistics. Additionally, the optimal number of components that explained over 90% of the variation in data was 3 as seen in the scree plot.

- 94.6% of the variation in the data is explained by the first 3 components, thus a model with 3 components distilled by PCA can explain nearly all of the variation of the 8-dimensional model in only 3 dimensions. While it is less accurate, just over 88% of the variation in the data can be explained with just two components which means that 88% of the data can be explained by PCA decomposed data in just two dimensions rather than 8.

- The most influential features in the PCA for component 1: was Grocery, with a Loading of 0.536 and for component 2: Delicassen with a Loading of 0.623 and for component 3: Delicassen with a loading of 0.670. Interestingly the 3 most influential features were the same for components 2 and 3 which suggests that these 3 features (Delicassen, Fronzen, and Region) contribute most to the variation in those components

- Detergents_Paper and gerocery have a strong positive correlation of 0.92 which suggests that clients of the wholesale distributor annual spending on both grocery and detergents_paper products is highly related.