ML Unsupervised Learning

Project: Wholesale Data

Conclusions

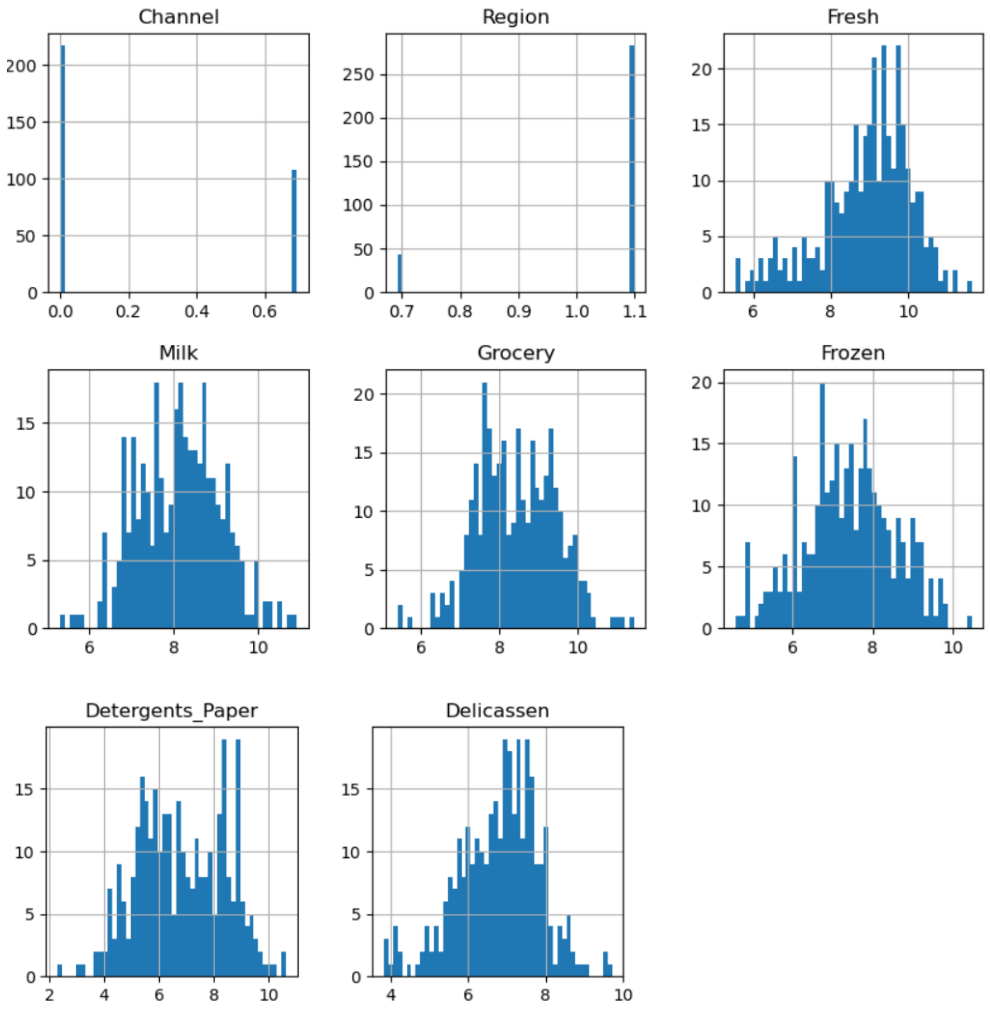
- For this project the goal was to gain insights (find similarities in the data points and group the similar data points together using unsupervised machine learning model) of the Wholesale dataset that contains information about various products sold by a grocery store.
- The variables available in the data set were: Channel, Region, Fresh , Milk , Grocery, Frozen, Detergents_Paper, Delicassen.
- For this project, the unsupervised machine learning models used were Kmeans Clustering and Hierarchical Clustering.
- The PCA method was also applied to draw conclusions about the wholesale customer data and find which compound combinations features best describe the customers.

Conclusions (Continuation)

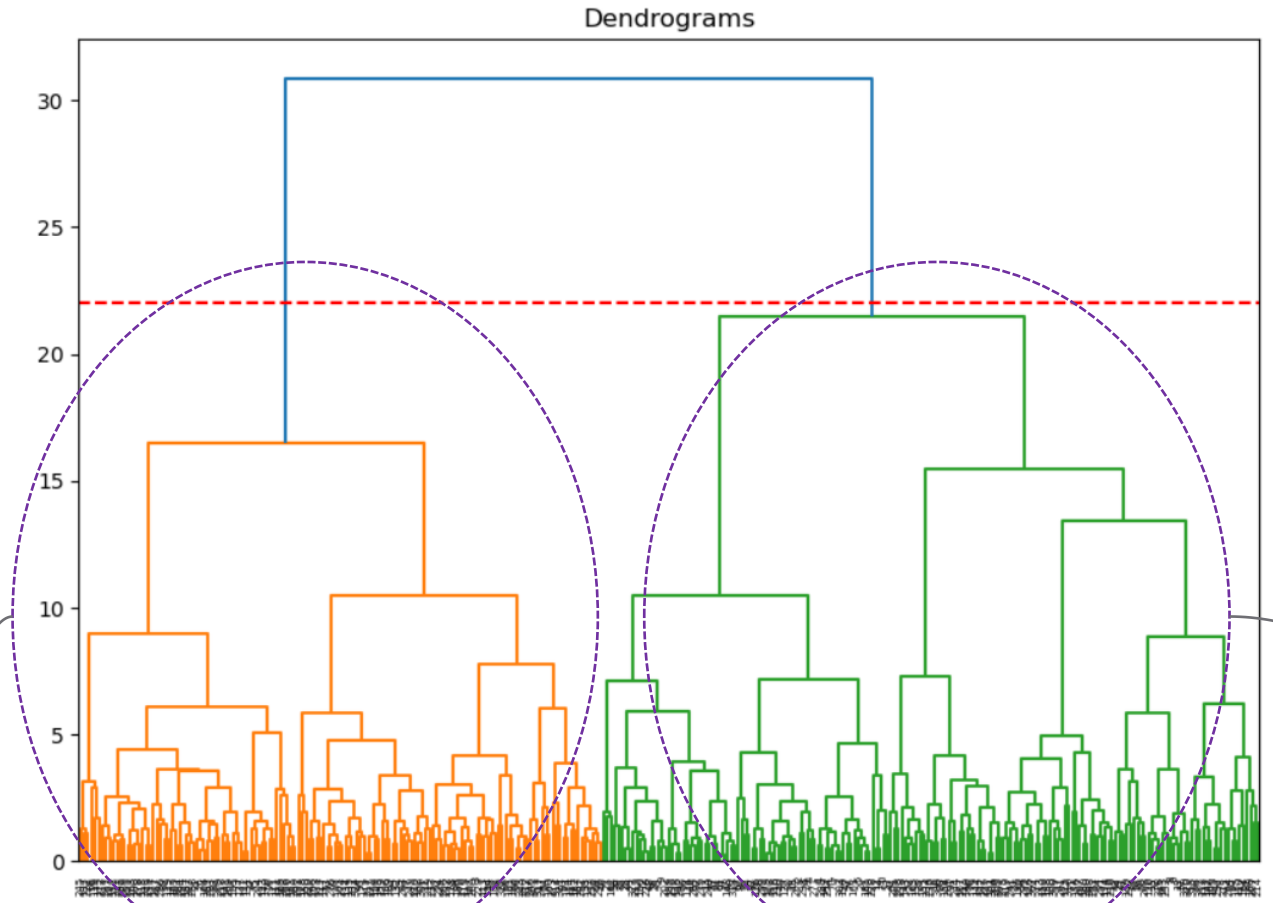
After analysis it was identified:

- Hierarchical clustering did a better job in segregating customers
- Hierarchical clustering presented a cleaner output. The number of clusters were 2 vs 10 in the KM clustering model
- K Means cluster visualization is not very intelligible due the two dimensions
- In the KM clustering model was not clear what was the optimal number of clusters. The elbow method did not show a clear elbow shape
- The PCA method showed that just 3 components where enough to best describe customers. Since 3 components best maximize variance
- For future analysis some option will be to use a different method to deal with outliers, since the IQR resulted in removing a high number of rows.
- And other option for future an analysis could be to try to segregate and prepare data in a different way (remove more columns or add dummy columns).

Data distribution / Variables Histograms



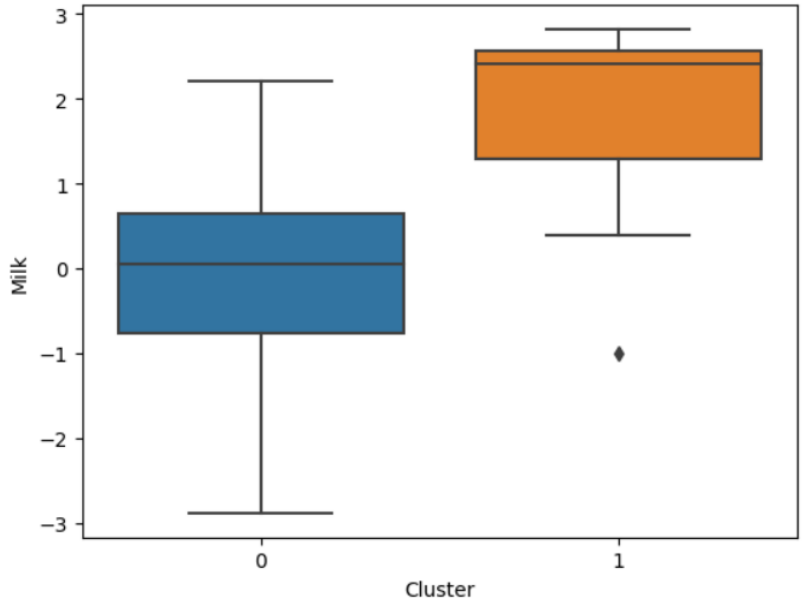
Hierarchical Clustering

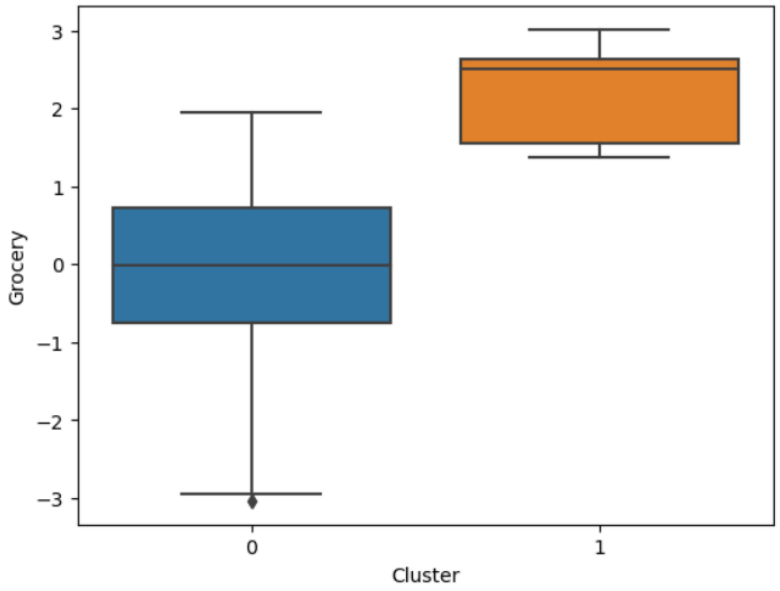


1st cluster

2nd cluster

Some visualizations of the hierarchical clustering output per variable





In these two charts it can be observed that customers from cluster zero spent less than the ones in cluster 1