Unsupervised Learning on Wholesale Dataset

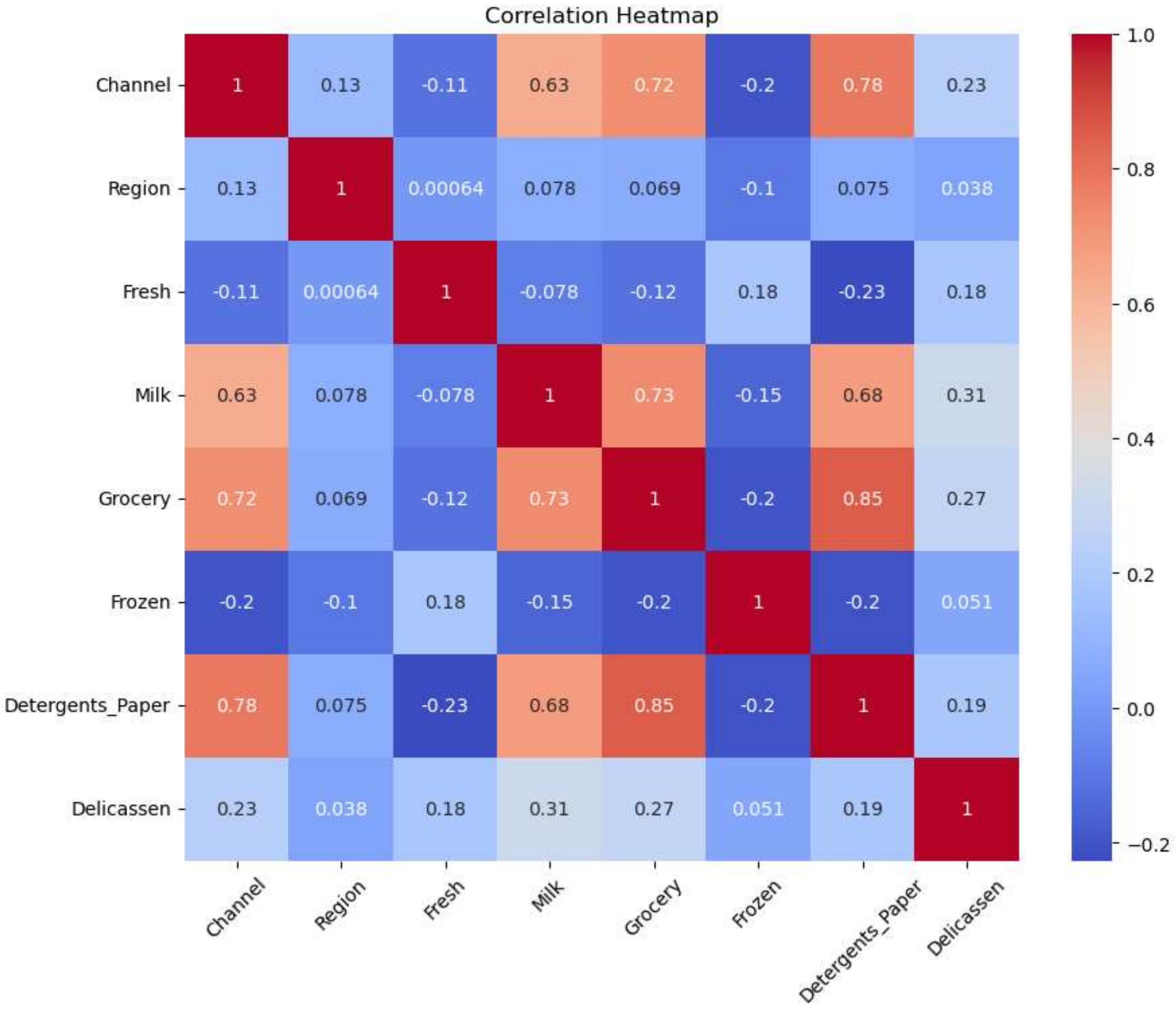
Siyi Mo

Project purpose

The purpose of the project is to use unsupervised learning techniques to build a machine learning model on wholesale dataset to find the pattern within the dataset. The project involves four main parts: exploratory data analysis and pre-processing, KMeans clustering, hierarchical clustering, and PCA.

EDA Results

The correlation matrix showed that Region had low absolute values with the seven other features. It did not hold any great explanatory value. Channel, in comparison, held some high positive values, indicating the dominance of Retail channels in Milk and Detergents_Papers purchases. The negative correlation values for Channel indicated that Hotels/Cafes dominate in Fresh and Frozen purchases.

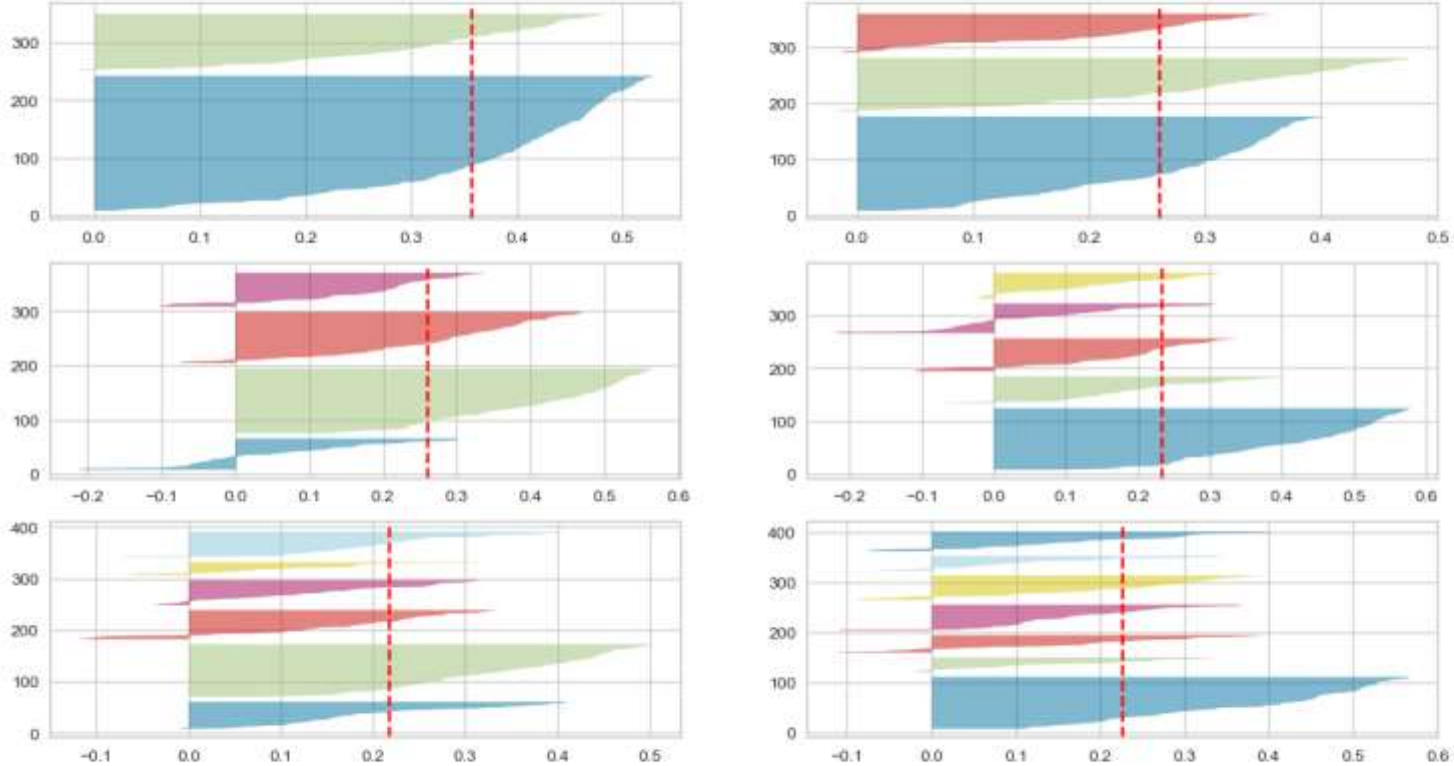


Correlation Heatmap

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Channel	1	0.13	-0.11	0.63	0.72	-0.2	0.78	0.23
Region	0.13	1	0.00064	0.078	0.069	-0.1	0.075	0.038
Fresh	-0.11	0.00064	1	-0.078	-0.12	0.18	-0.23	0.18
Milk	0.63	0.078	-0.078	1	0.73	-0.15	0.68	0.31
Grocery	0.72	0.069	-0.12	0.73	1	-0.2	0.85	0.27
Frozen	-0.2	-0.1	0.18	-0.15	-0.2	1	-0.2	0.051
Detergents_Paper	0.78	0.075	-0.23	0.68	0.85	-0.2	1	0.19
Delicassen	0.23	0.038	0.18	0.31	0.27	0.051	0.19	1

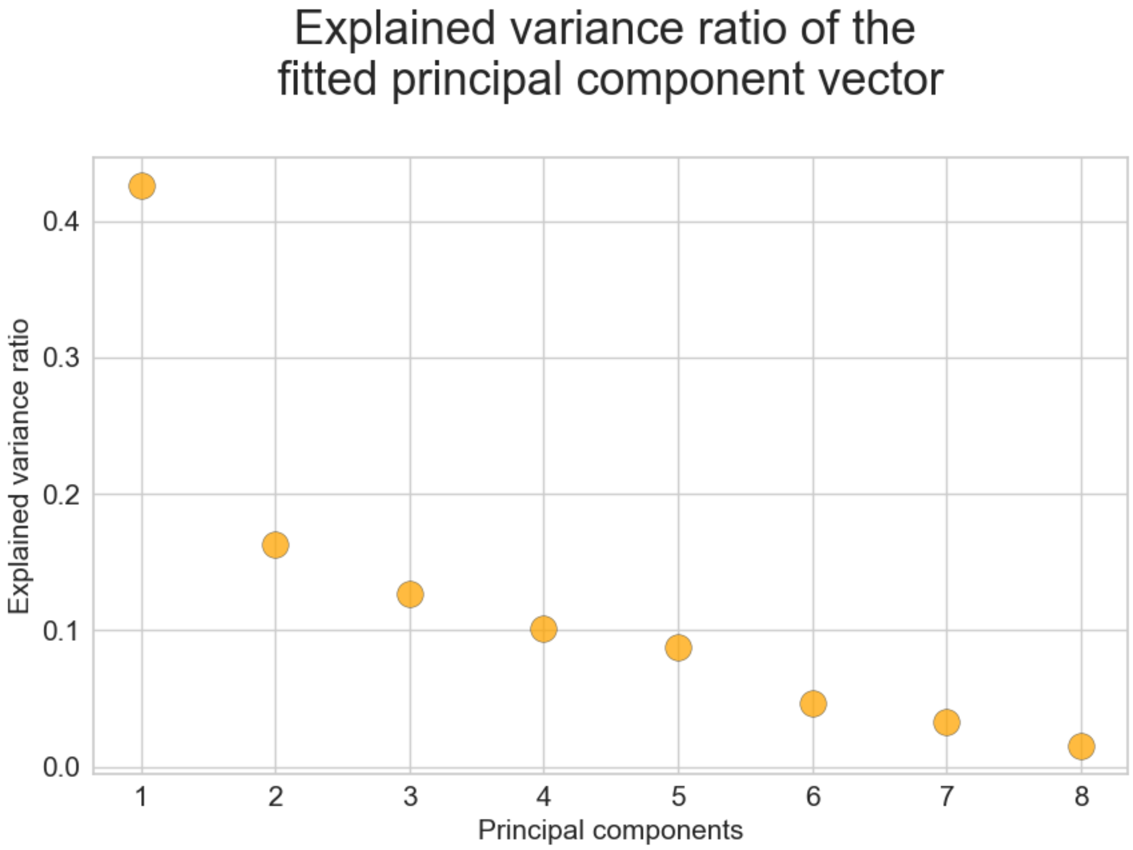
Model Performance

All clusters have scores above the average score for the dataset (represented by the red dashed line), a criterion in Silhouette selection. None of the Silhouette graphs is ideal because of the variance in the width of the clusters, with k=3-5 clusters looking the most even. Combining the results of the Elbow and Silhouette method suggest k=4 is the best.



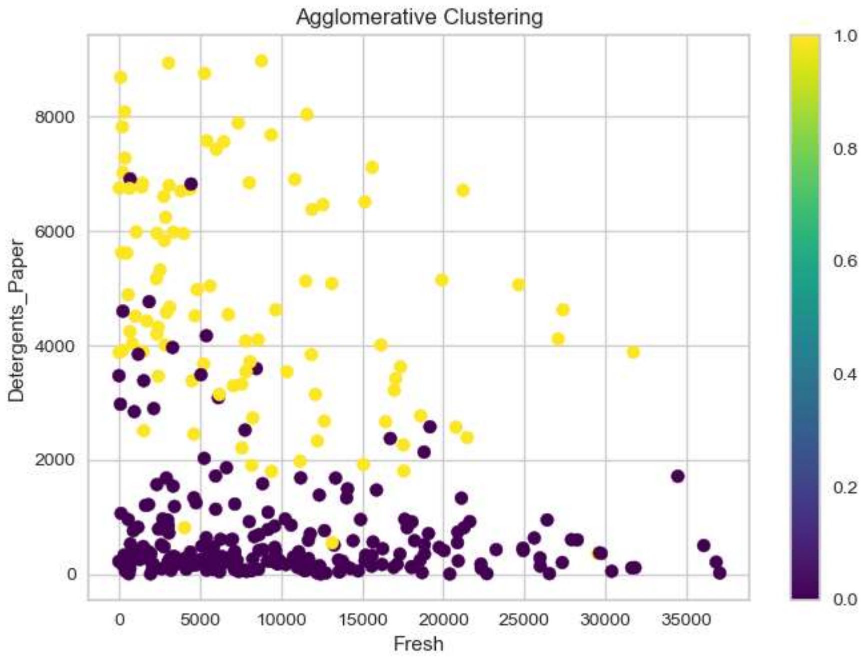
Model Performance

The PCA revealed that PC1 accounted for 42.6% of the variance and PC2 was 16.2% for a combined 58.8% of the variance accounted for with just the first two components.



Model Performance

The similarities in graphs using all eight original features or a reduced model using only four suggest that the reduced model did not lose much explanatory power and could still acount for much of the variance. This is possibly due to certain features like Grocery having high correlations (>=0.73) with two other variables, so eliminating it did not change the model much.


Agglomerative Clustering
Detergents_Paper
Fresh