Unsupervised Learning Project

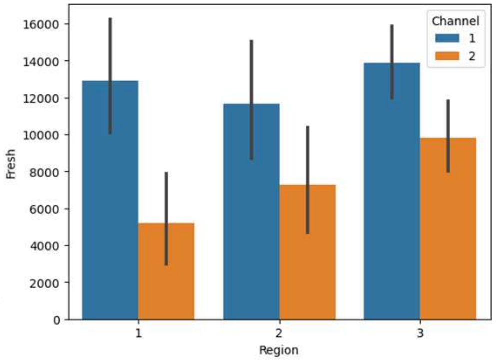
BRETT FORSEN

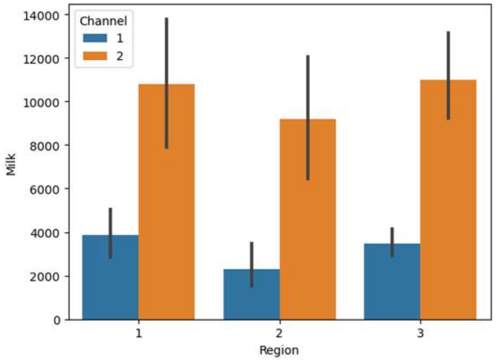
Exploratory Data Analysis

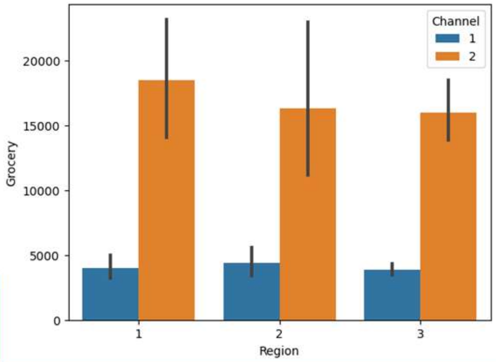
No null or zero values were found in the data.

The outliers were all within the realm of probability, and so no removals needed to be made. More background information regarding the data set could change this, however.

The Region and Channel columns were identified to be categorical. After using them for some visualizations, they were dropped in preparation for the Unsupervised models.



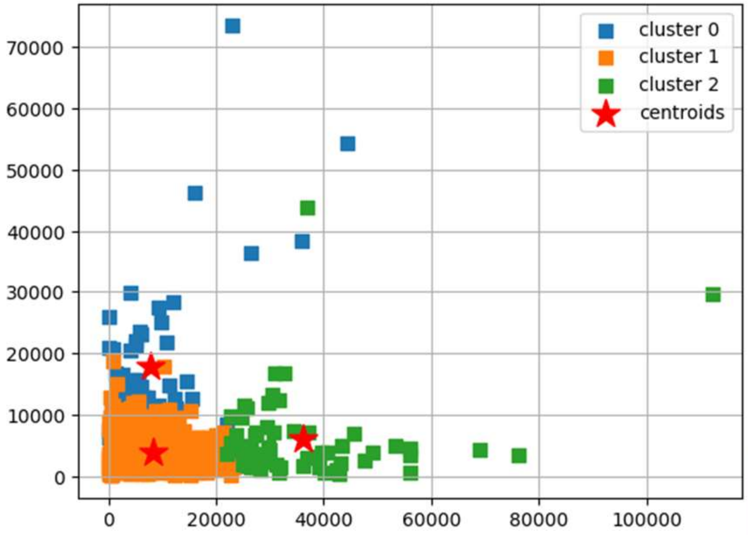


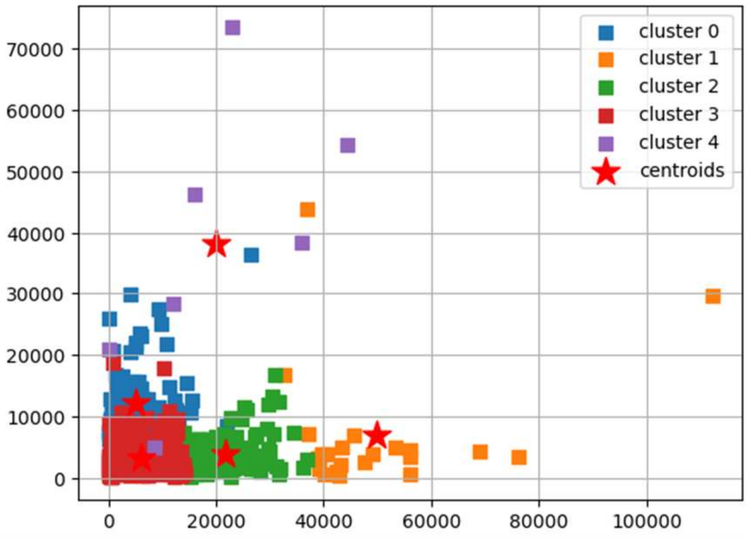


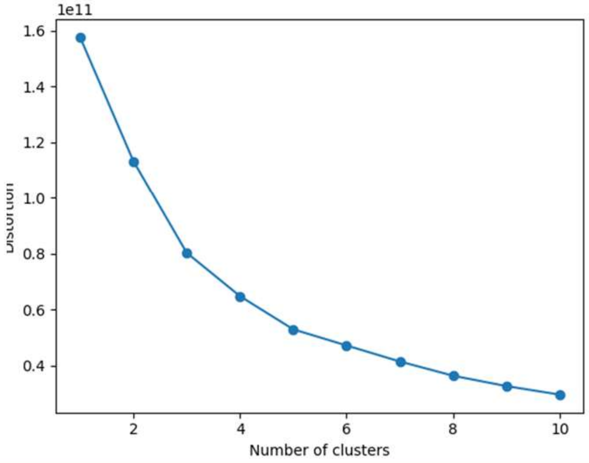
KMeans Clustering

The elbow rule was used to determine the number of clusters to be used for the model. Either 3 or 5 clusters seemed most appropriate.

Both variants of the KMeans model are as follows:



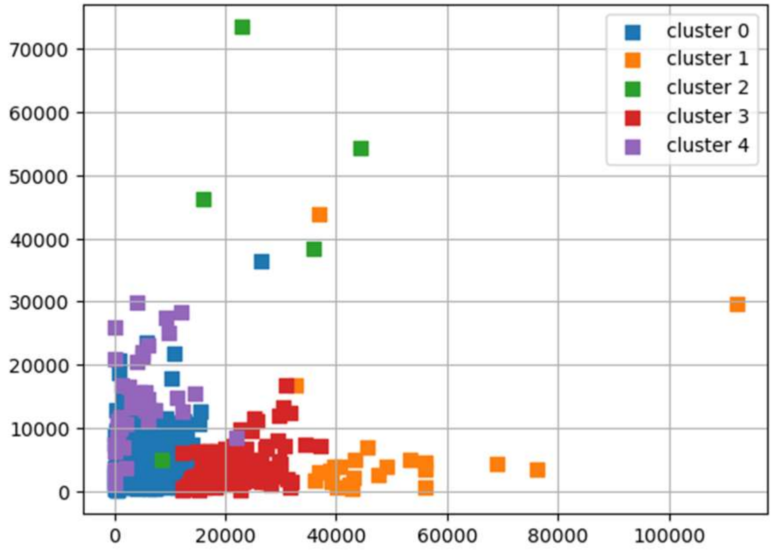


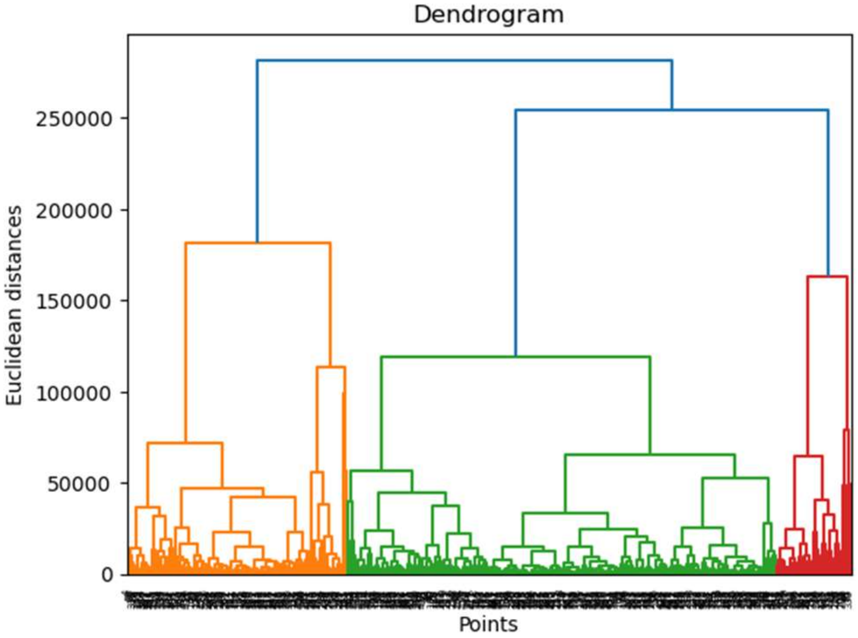


Hierarchical Clustering

A dendrogram was used to determine how many clusters the model should use.

The clusters are not very evenly distributed, and arguments could be made for anywhere from 3 to 5 clusters.





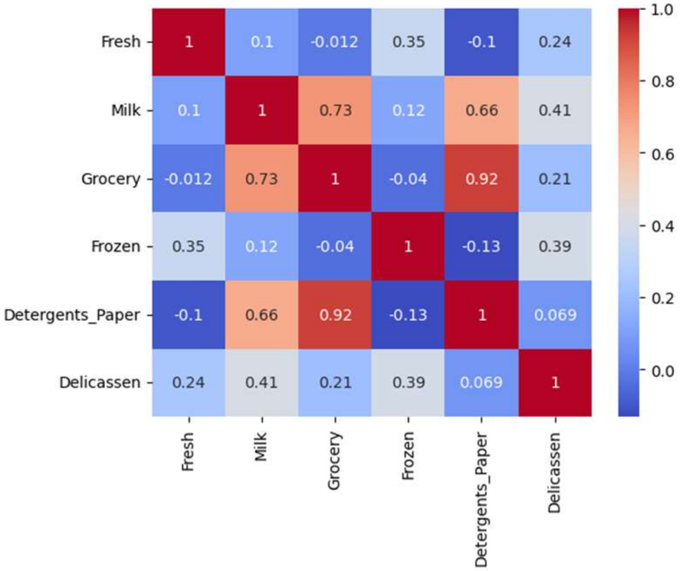
PCA

To get a feel for the relationship between the features, a correlation matrix was created.

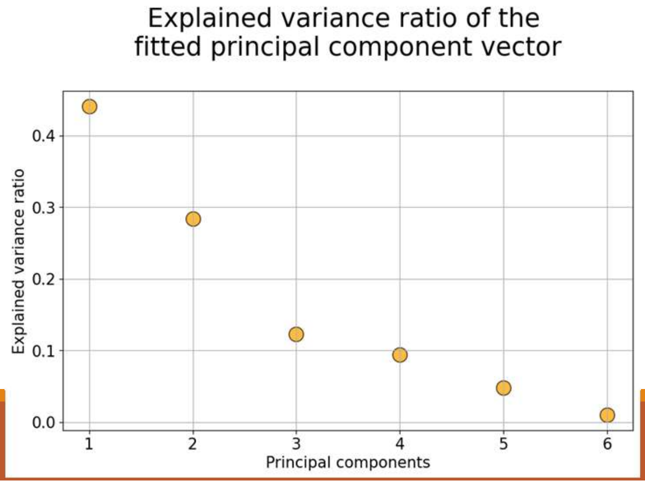
The data was then scaled to increase the effectiveness of the PCA model.

The principal components appear in the following order:

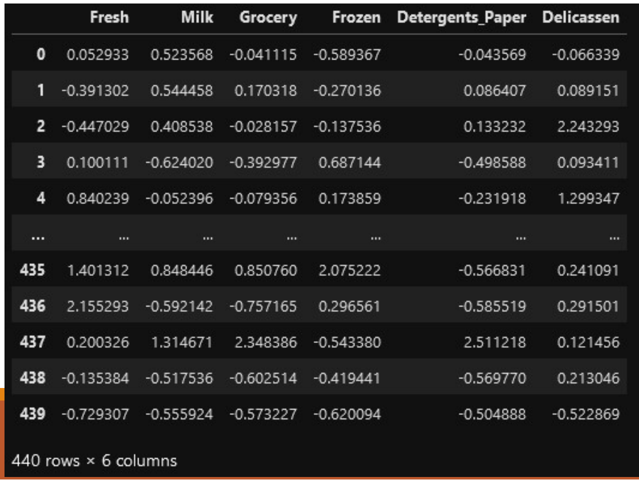
Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicassen.



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Fresh	1	0.1	-0.012	0.35	-0.1	0.24
Milk	0.1	1	0.73	0.12	0.66	0.41
Grocery	-0.012	0.73	1	-0.04	0.92	0.21
Frozen	0.35	0.12	-0.04	1	-0.13	0.39
Detergents_Paper	-0.1	0.66	0.92	-0.13	1	0.069
Delicassen	0.24	0.41	0.21	0.39	0.069	1



Explained variance ratio of the fitted principal component vector



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	0.052933	0.523568	-0.041115	-0.589367	-0.043569	-0.066339
1	-0.391302	0.544458	0.170318	-0.270136	0.086407	0.089151
2	-0.447029	0.408538	-0.028157	-0.137536	0.133232	2.243293
3	0.100111	-0.624020	-0.392977	0.687144	-0.498588	0.093411
4	0.840239	-0.052396	-0.079356	0.173859	-0.231918	1.299347
...	...	...	...	...	...	...
435	1.401312	0.848446	0.850760	2.075222	-0.566831	0.241091
436	2.155293	-0.592142	-0.757165	0.296561	-0.585519	0.291501
437	0.200326	1.314671	2.348386	-0.543380	2.511218	0.121456
438	-0.135384	-0.517536	-0.602514	-0.419441	-0.569770	0.213046
439	-0.729307	-0.555924	-0.573227	-0.620094	-0.504888	-0.522869

440 rows × 6 columns

Conclusion

The optimal number of clusters for the dataset was between 3 and 5.

Fresh and Frozen items were much more likely to be classified under Channel 1 than 2.

More evenly distributed clusters are likely not possible considering the resulting dendrogram in slide 4.

The categories Fresh and Milk accounted for by far the most variance in the model, and it really isn’t close. More knowledge of the data set would be needed in order to explain this relationship.