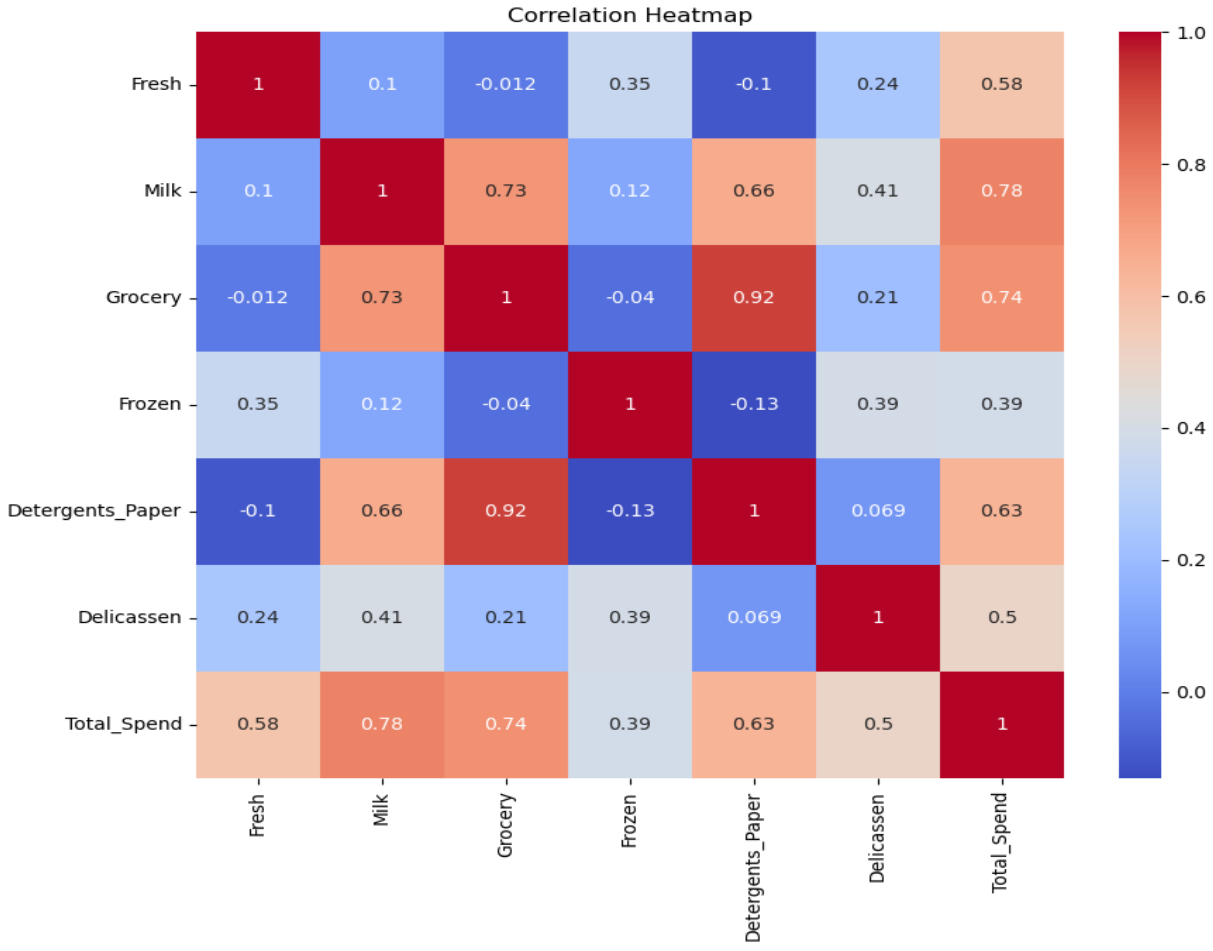
"Unsupervised Learning – Project

Dataset - Wholesale Data"

There are 2 Channels through which 6 products are sold across 3 regions. Below image shows the details:

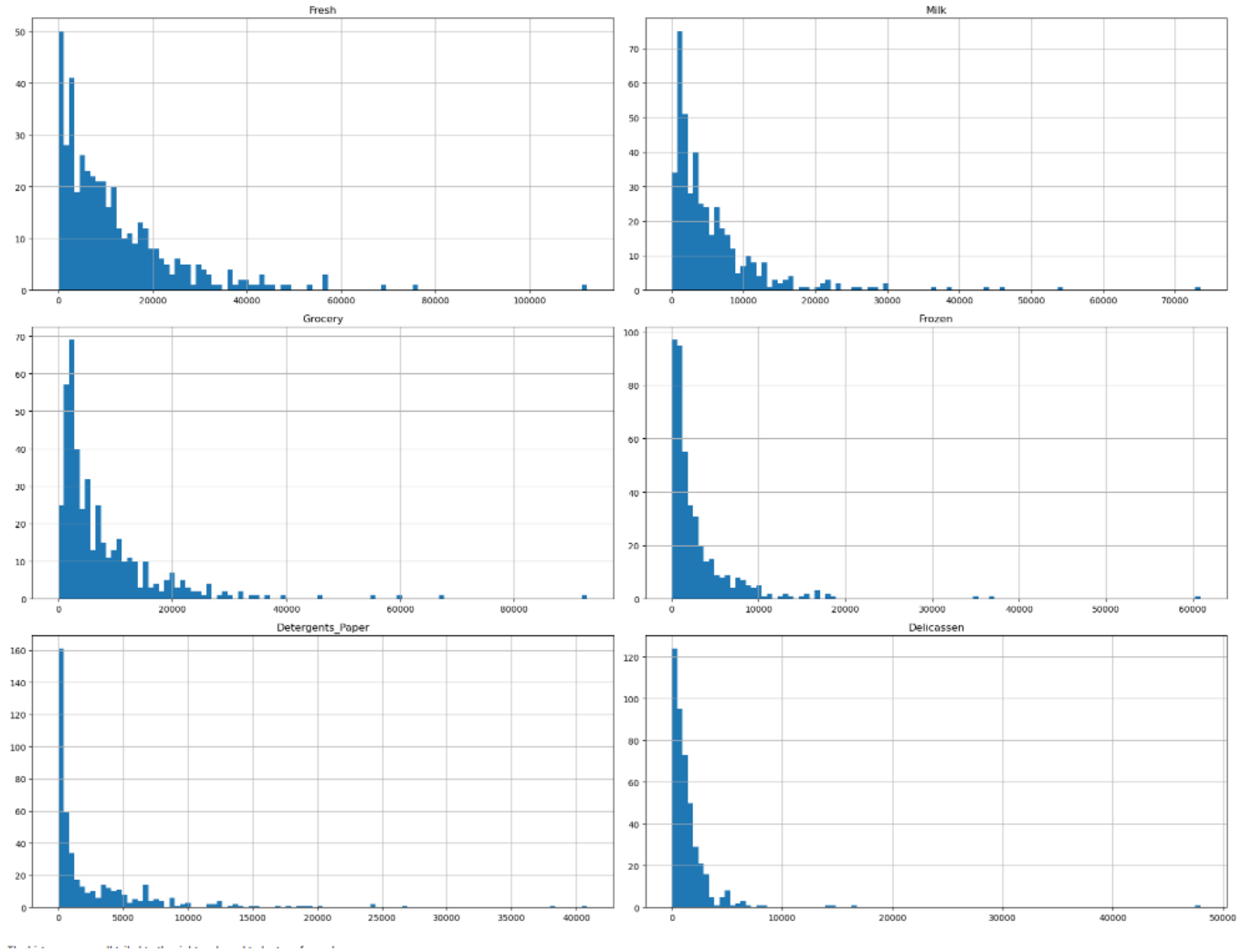
Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
1	1	761233	228342	237542	184512	56081	70632
	2	326215	64519	123074	160861	13516	30965
	3	2928269	735753	820101	771606	165990	320358
2	1	93600	194112	332495	46514	148055	33695
	2	138506	174625	310200	29271	159795	23541
	3	1032308	1153006	1675150	158886	724420	191752

Check for Relationship



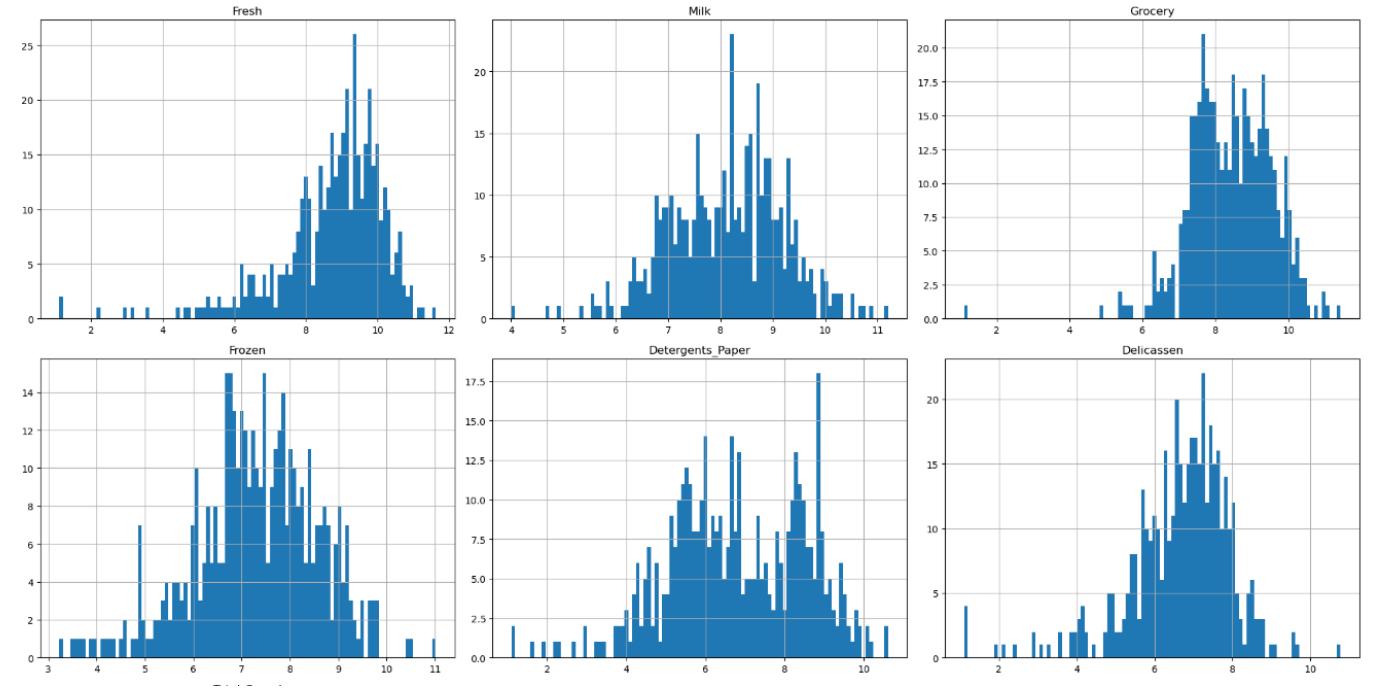
The correlation map shows that the strongest correlation is 0.92 between Detergent Paper and Grocery. This is a positive correlation which tells us that customers spend for both items increase/decrease together.

Check for Distribution



The histograms are all tailed to the right and need to be transformed.

Data Transformation - Logarithmic

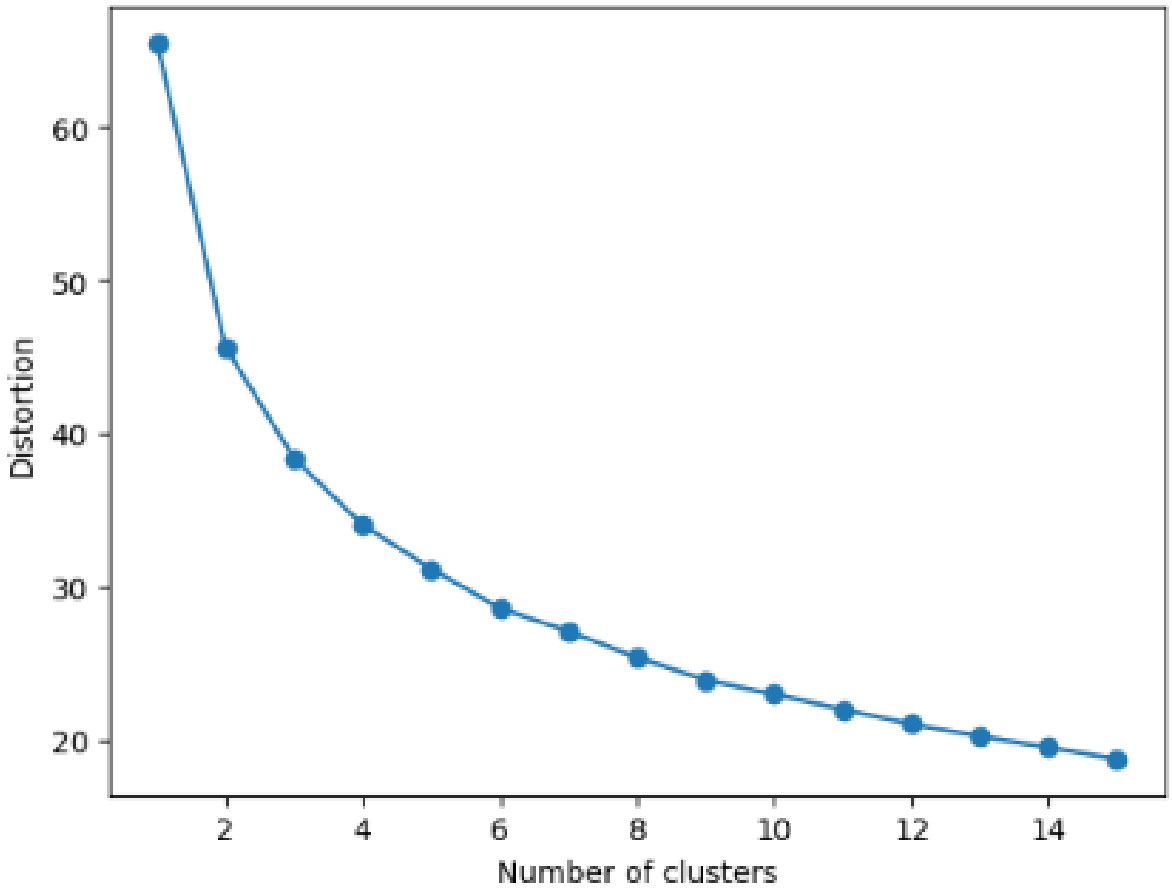


Clustering

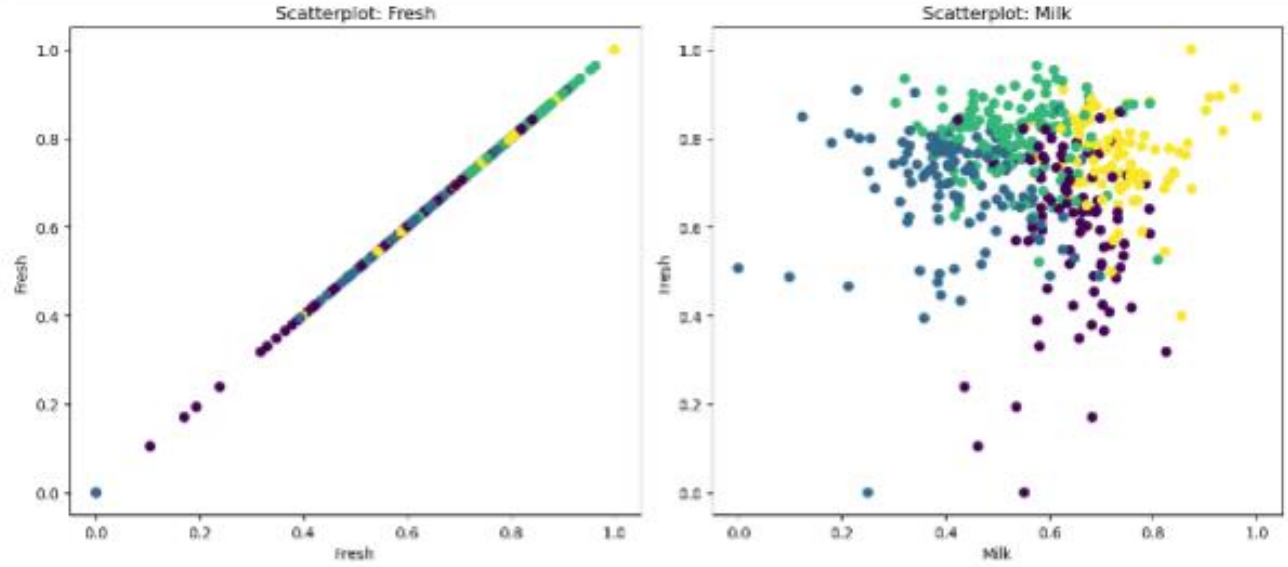
The objective of the analysis is to group similar products together into clusters based on their attributes such as fresh, milk, grocery, frozen, detergents paper, and delicatessen. To perform the k-means clustering analysis, you will need to pre-process the dataset, determine the optimal number of clusters, initialize the centroids, assign data points to clusters, update the centroids, and repeat until convergence.

KMeans Clustering

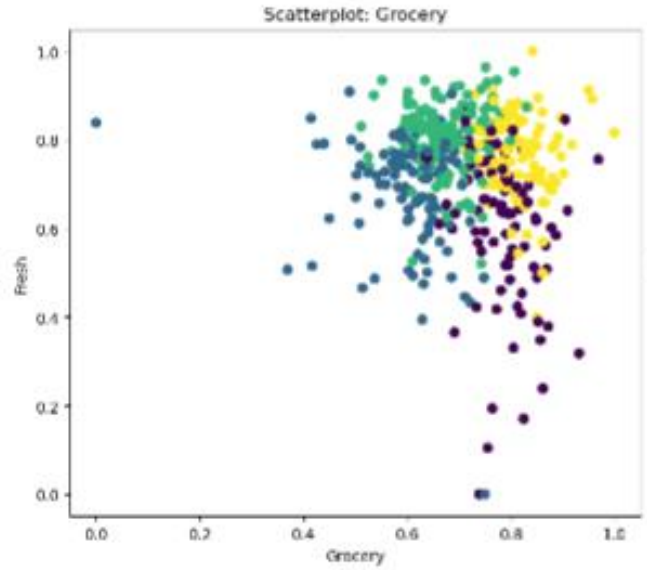
Using the Elbow curve to determine the optimal number of clusters

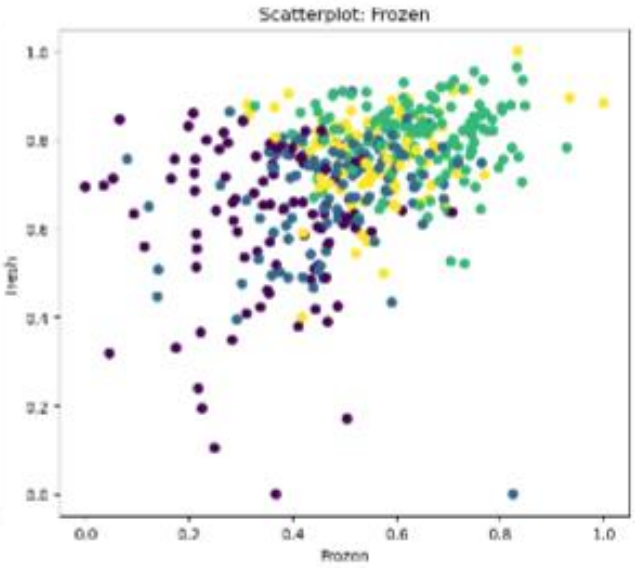


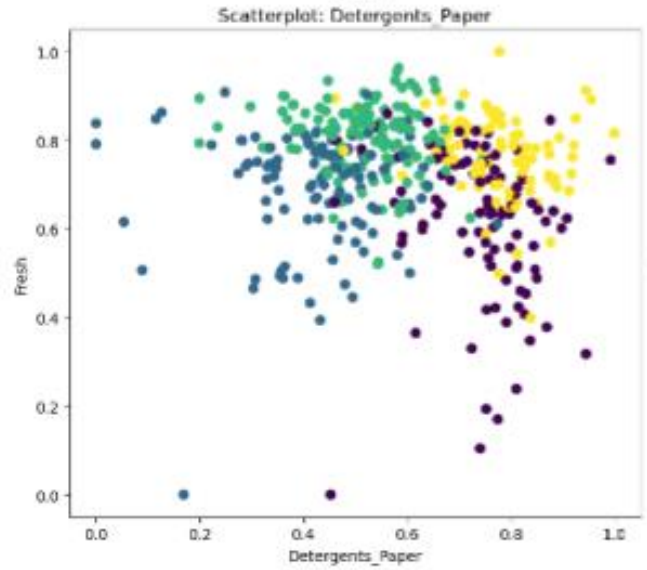

```
KMeans
KMeans(init='random', n_clusters=4, random_state=42)
```

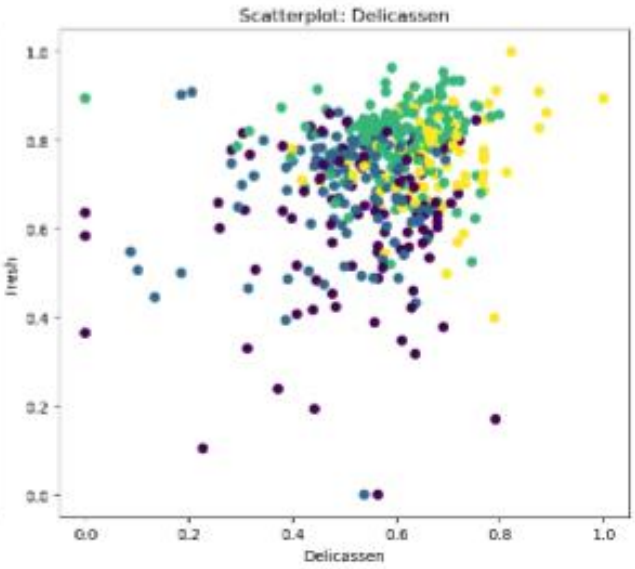

Scatterplot: Fresh
Fresh
Fresh

Scatterplot: Milk
Fresh
Milk


Scatterplot: Grocery
Fresh
Grocery

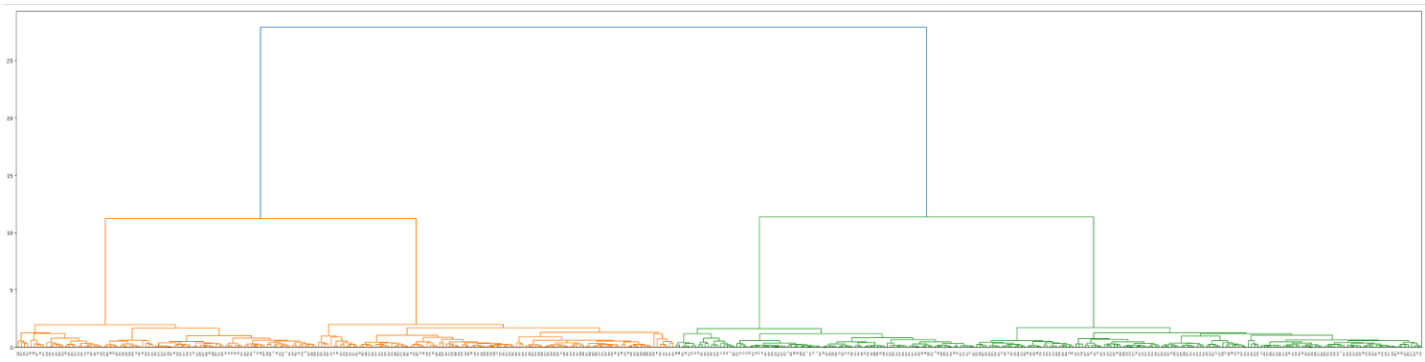

Scatterplot: Frozen
Fresh
Frozen


Scatterplot: Detergents_Paper
Fresh
Detergents_Paper


Scatterplot: Delicassen
Fresh
Delicassen

Hierarchical Clustering

Plotting Dendogram to determine the number of clusters.



PCA

We will find which compound combinations of features best describe customers.

	PC1	PC2	PC3	PC4
0	-0.211857	0.126503	-0.208293	-0.035976
1	-0.223253	-0.032812	0.017461	-0.034480
2	-0.248165	-0.112140	0.026158	-0.160587
3	0.105151	-0.207375	0.053900	-0.020840
4	-0.169566	-0.230296	-0.024893	-0.074294
...	...	...	...	...
435	-0.110982	-0.395349	0.037995	-0.006704
436	0.217006	-0.322126	-0.100859	-0.045259
437	-0.469598	0.044532	-0.145578	0.021803
438	0.177904	-0.046708	-0.100211	-0.129306
439	0.235054	0.464850	-0.167702	0.103059

Random Forest

We will find which compound combinations of features best describe customers.

```
Top 4 Features:
   Feature  Importance
3   Frozen    0.306426
0    Fresh    0.183666
1     Milk    0.170951
2  Grocery    0.152583
Top Feature Combinations:
['Frozen', 'Fresh', 'Milk', 'Grocery']
```

Conclusion

EDA

* The histograms show an exponential decline in the number of orders for the respected products and very high skewness with was normalized with Logarithmic transformation.

* The strongest positive correlation between Detergents paper and Grocery products which indicates that consumers would often spend money on these two types of products and any marketing or sales strategy centres around these 2 products will be profitable for the business.

K-MEANS/DENDOGRAM

* Optimal number of clusters was determined using the "Elbow Curve" and the same number was used in Agglomerative Clustering

PCA

* From the above output, we can observe that the principal component 1 holds about 45% of the information while the principal component 2 holds about 28% , 3 holds only about 10% and 4 holds 9% of the information.

* Top 4 features was determined using Random Forest which are 'Frozen', 'Fresh', 'Milk', 'Grocery' in order of importance.