Unsupervised Learning Project

Christopher Martinez Demmans

Goals of Project

-Perform unsupervised learning techniques on a wholesale data dataset. The project involves four main parts: exploratory data analysis and pre-processing, KMeans clustering, hierarchical clustering, and PCA.

-Use machine learning to segment customers into groups based on spending habits to better serve them in the future.

The Data

Our data set includes information on total sales of a wholesale chain categorized by location and type which is followed by the total annual sales by product category.

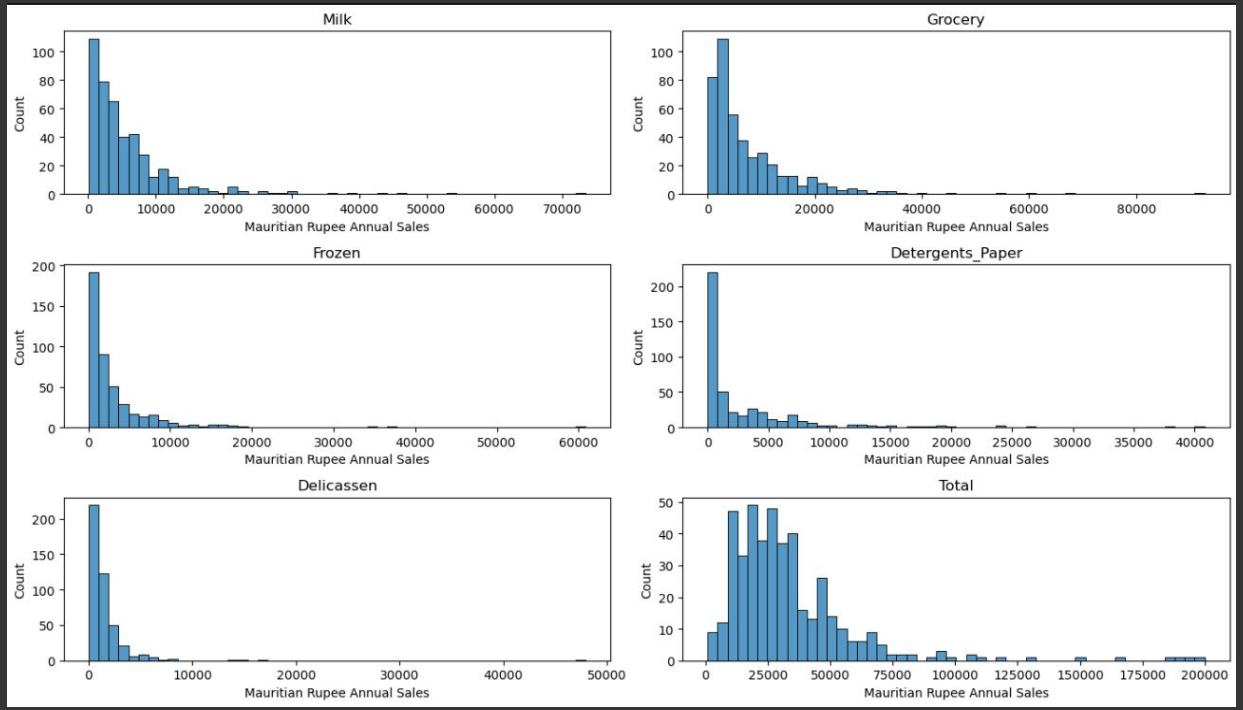
Numerical

- Fresh
- Milk
- Grocery
- Frozen
- Detergents_Paper
- Delucassen

Categorical

- Channel
- Region

EDA:



Observations:

The sales of each item are not random the distribution has a right skew for every product.

Meaning people are more likely to spend less than more.

EDA:

-Meaning of the numbers in the 'region' and 'channel' columns
- There are 2 Channels and 3 regions in this wholesale dataset.
Regions - Lisnon, Oporto or Other (Nominal) (1,2,3)
Channel - Horeca (Hotel/Restaurant/Cafe or Retail channel (Nominal) (1,2)

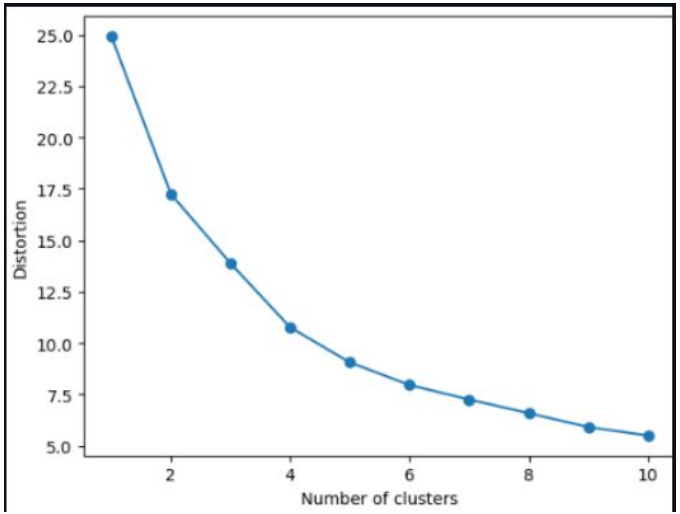
- Meaning of the numbers in the item category columns:
- Annual spending (Mauritian Rupee) on the type of product the category represents.
- Mauritius is where this dataset comes from based on the currency used
- Mauritus is a country in East Africa

- Meaning of the numbers in the item category columns:
- Annual spending (Mauritian Rupee) on the type of product the category represents.
- Mauritius is where this dataset comes from based on the currency used
- Mauritus is a country in East Africa

Part II - KMeans Clustering

K-Means Clustering

The first thing we need to do is find out how many groups we should divide our customers into.

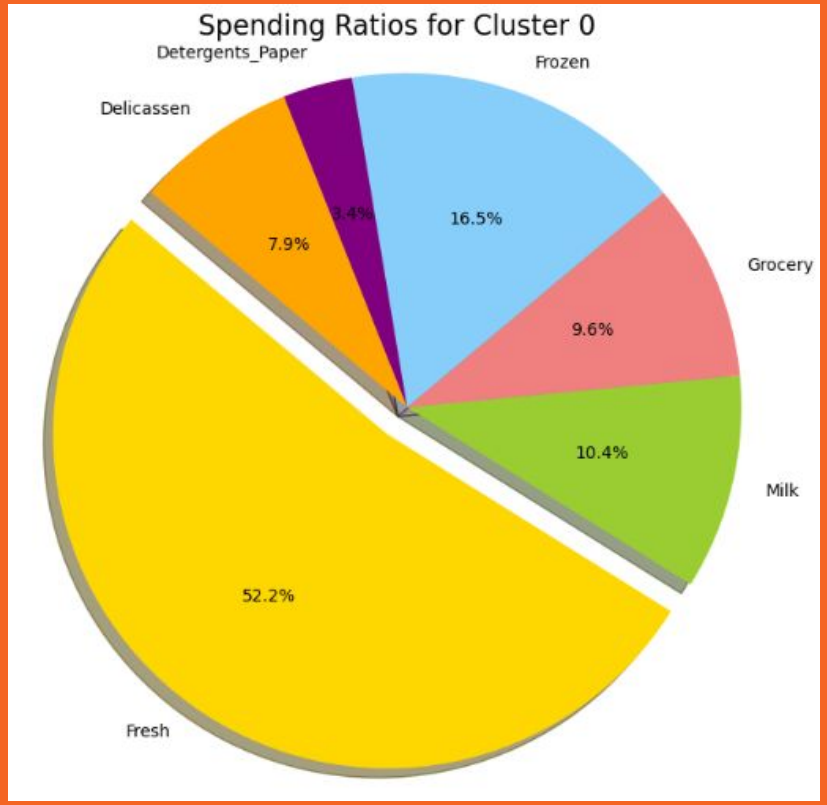


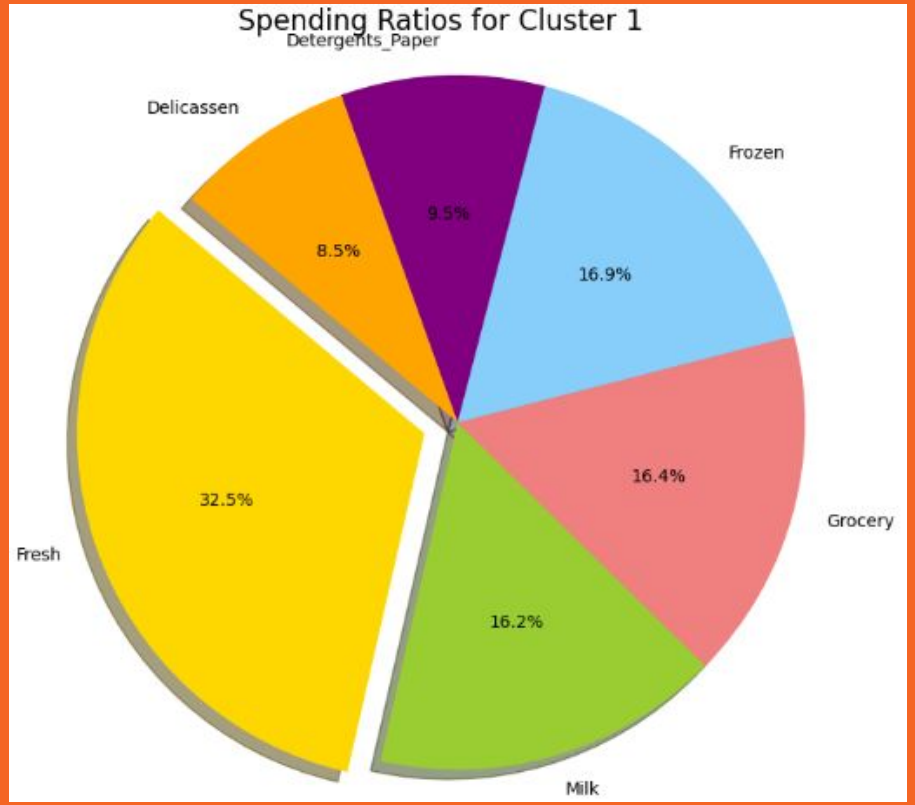
Observations:

Creating more than 4 clusters starts to border on diminishing returns and increases complexity.

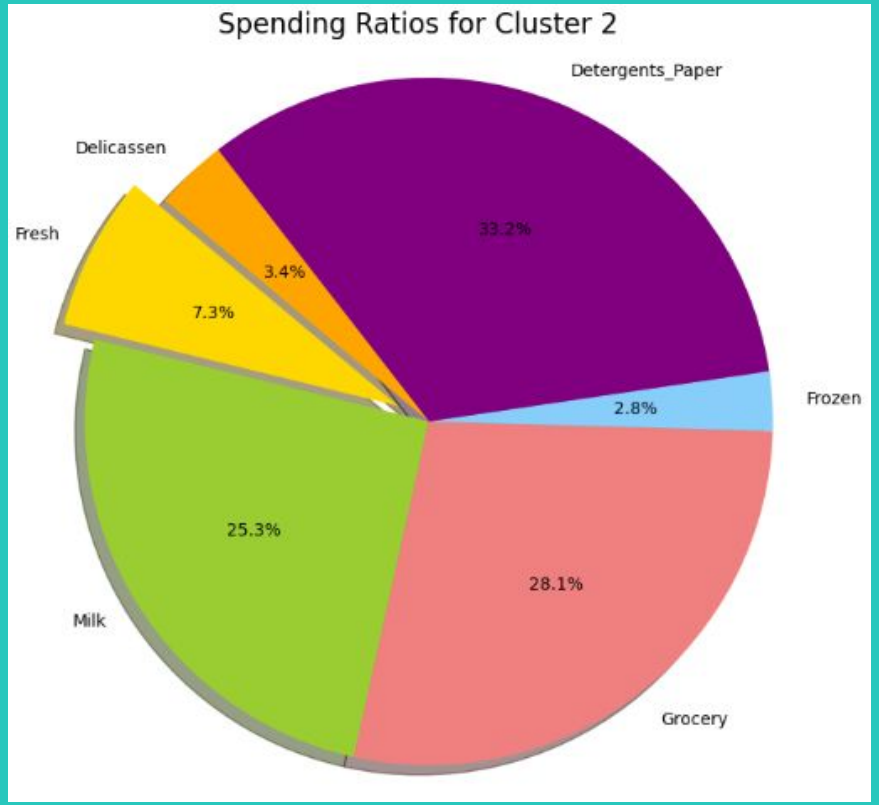
We stuck with four groups.

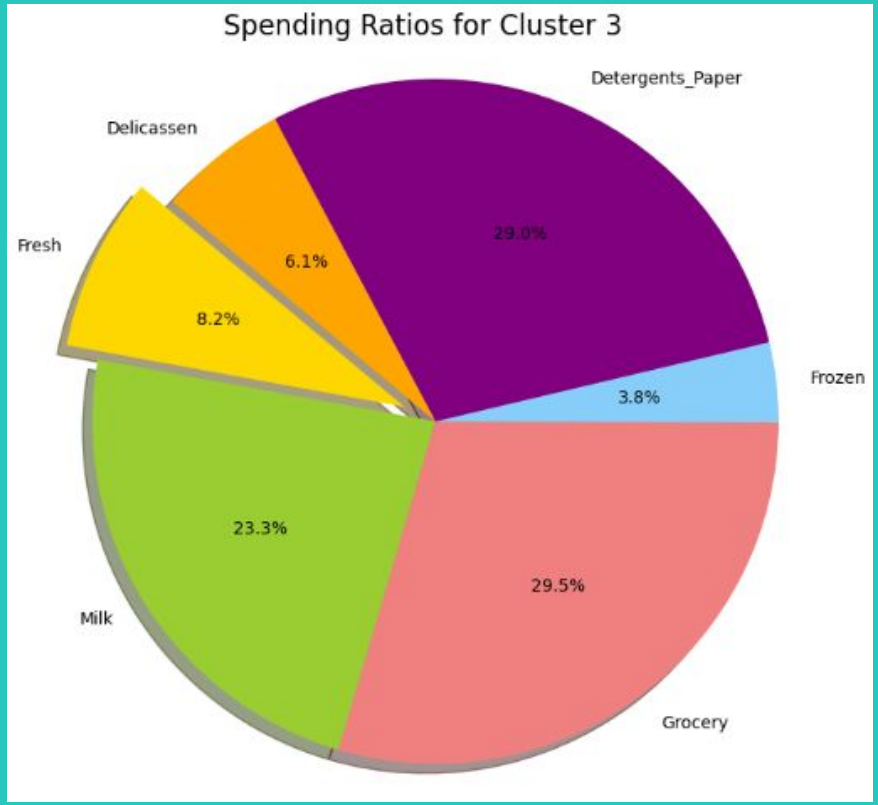
The different spending habits of customers.





The different spending habits of customers.



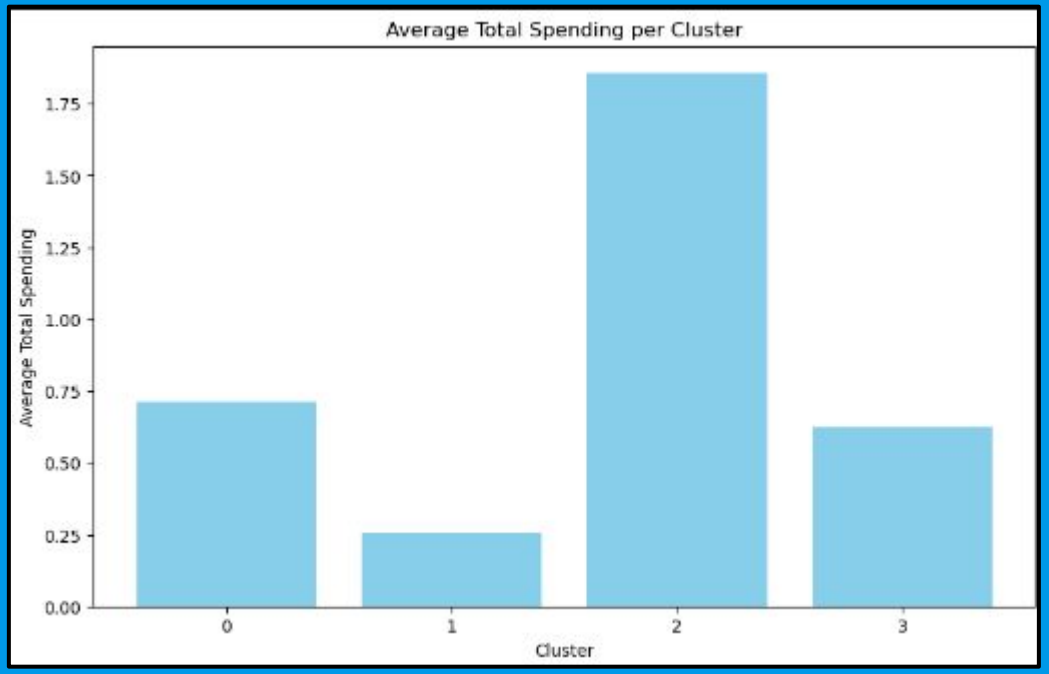


SVM

Each cluster represents a different type of customers spending habits. In this case we have 4 different types of customers:.

Cluster	Spending Level	Shopping Style
0	Medium	Health-conscious or restaurant owners, balanced with a focus on fresh produce and frozen goods.
1	Low	Possibly prefers processed foods, could be eating out often, low on cleaning supplies.
2	High	Convenience or fast-food oriented, good amount of dairy and high on cleaning supplies.
3	Medium	Convenience store types, moderate spending on groceries and cleaning supplies.

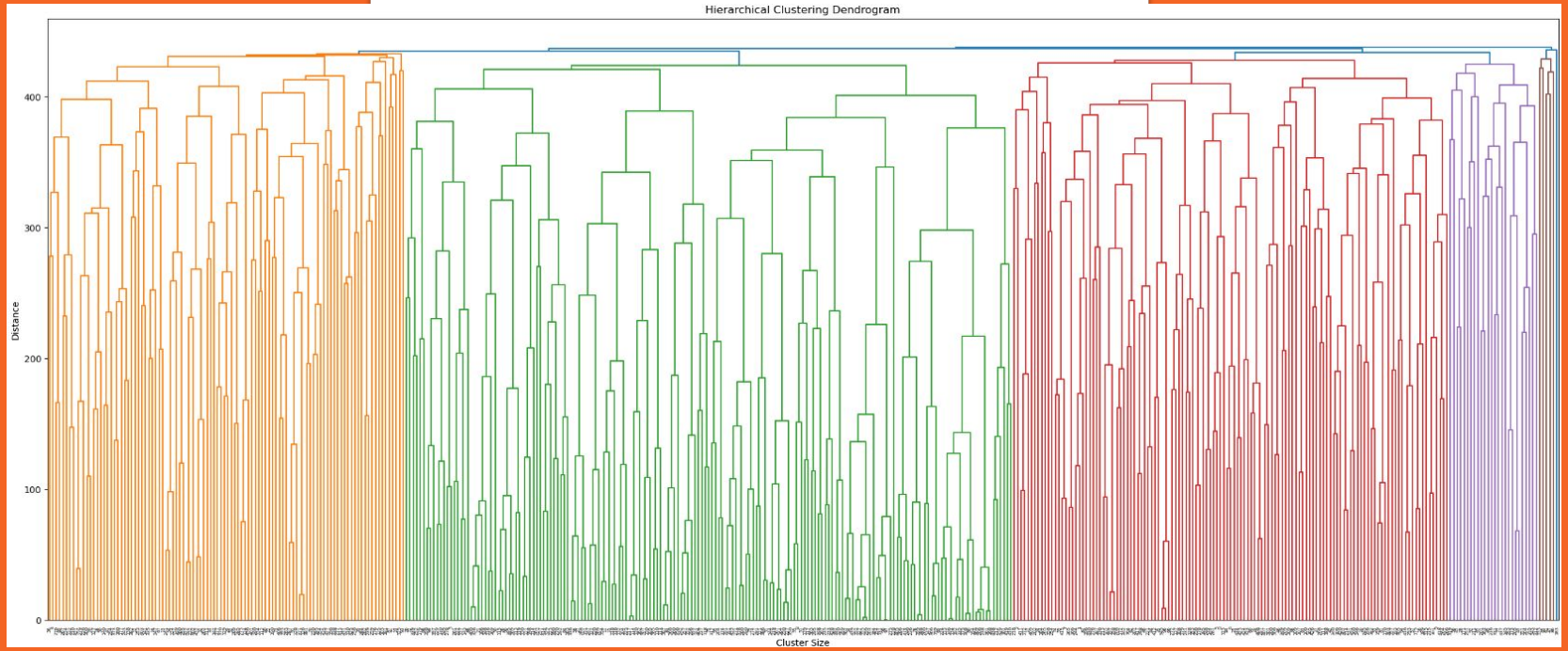
How the spending levels were determined:

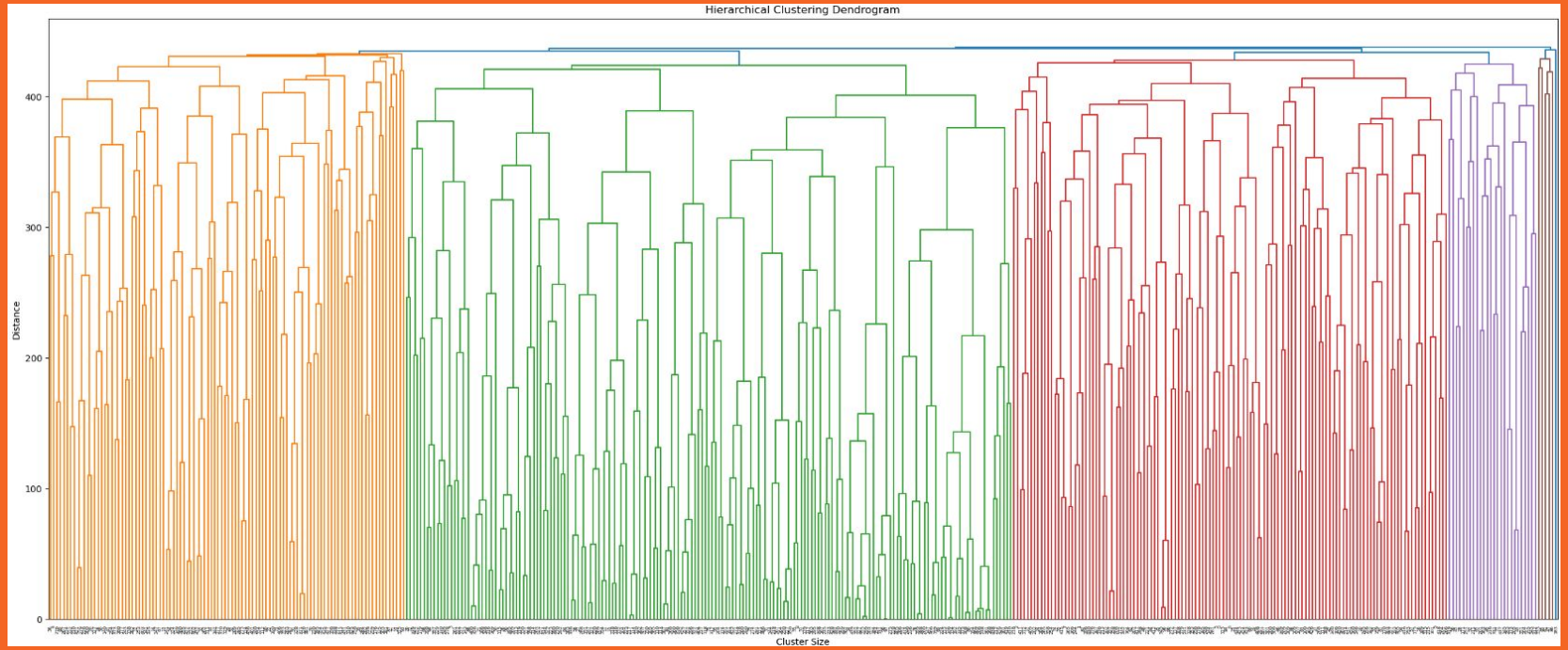


Part III - Hierarchical Clustering

Hierarchical Clustering

The basic idea is that data points that are close to another data point are similar.



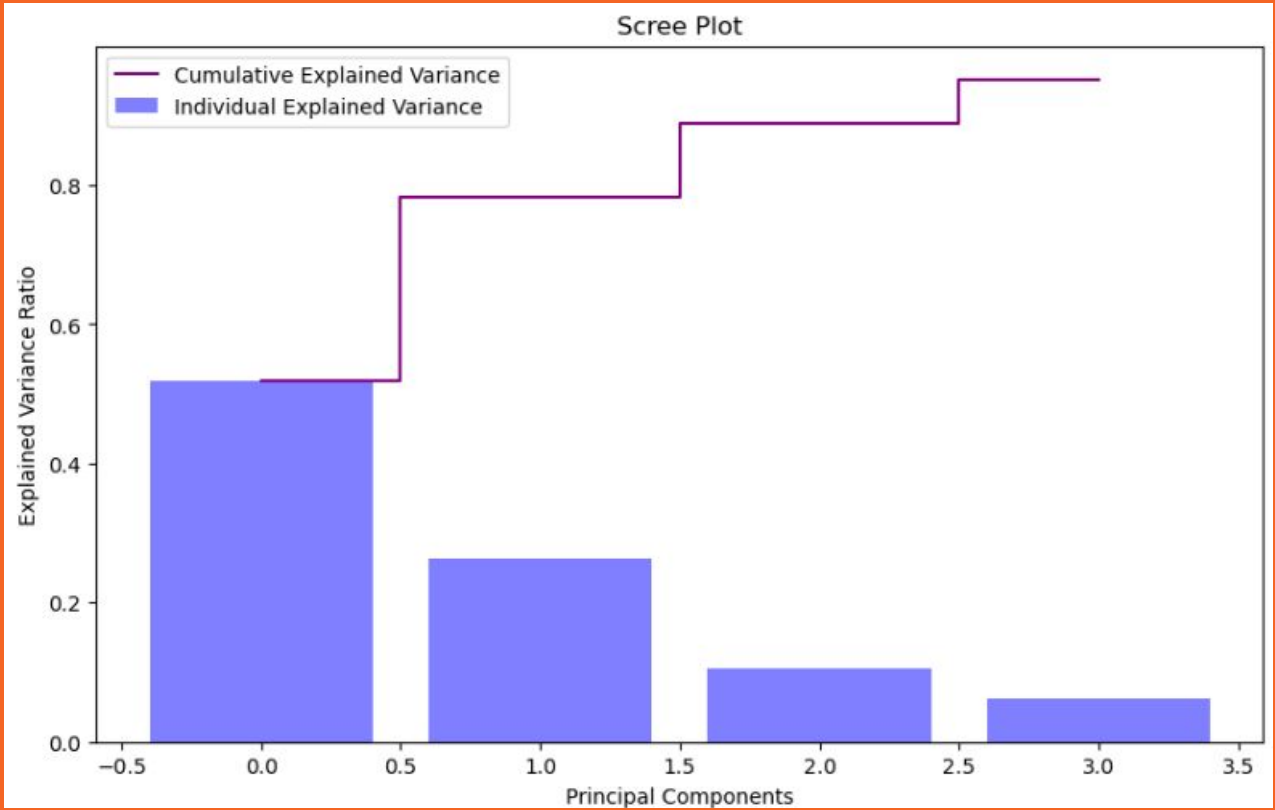


In essence this dendrogram tells us that splitting our customers into four groups is a reasonable idea.

Part IV - PCA

PCA

-The goal here is to reduce the dimensionality of our data to make it easier for the machine learning models to make predictions.



Our first two PCA components can explain about 80% of the variation in our dataset.

PCA Loadings

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
PC1	-0.018545	0.490473	0.576447	-0.017416	0.647882	0.082140
PC2	0.863877	0.181824	-0.003639	0.396910	-0.126527	0.217017
PC3	-0.501784	0.333715	-0.086809	0.658337	-0.220706	0.383657
PC4	-0.024034	-0.627960	0.197409	0.605219	0.350505	-0.277443

PCA components are on a range of 1 to -1

- A value closer to positive 1 represents a strong positive infulence on the PCA component when that variable increases.
- A value closer to negative 1 represents a strong neagative infulence on the PCA component when that variable increases.

PC1: Highlights household essentials like "Grocery" and "Detergents_Paper" with positive values, while fresh items like "Fresh" and "Frozen" have negative influence.

PC2: Focuses on "Fresh" as a major contributor, accompanied by "Frozen" and "Delicassen," suggesting a connection between freshness and deli products.

PC3: Emphasizes "Frozen" and "Detergents_Paper," showing contrasts between cleaning supplies and frozen goods versus fresh produce.

PC4: Illustrates a balance between "Frozen" and "Delicassen," while revealing a nuanced link between categories like "Milk" and "Grocery."

Part V - Conclusion

Four Key Takeaways:

- Classifying customers in to 4 clusters with k-means seems to have worked the best.
- This dataset comes from a wholesale chain in east africa, more specifically Mauritius.
- There are customers who purchase similar ratios of items (eg. 50% grocery and 50% frozen) but spend more or less money (eg. 100 dollars vs 1000 dollars)
- The data we have is a little bit biased to region 3 ('other') since there are many more data points for that region than 1 or 2. However in our use case it didn't have a significant effect. Caution should be taken if using this dataset for future business applications other than this one.

We can use these groups to cater to different customer needs and increase sales.

Thank you for reading!