A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is a light green. They are positioned diagonally, with the blue one partially covering the green one.

Unsupervised Learning Project

Prepared by: Chloe Phuong



Project Goals

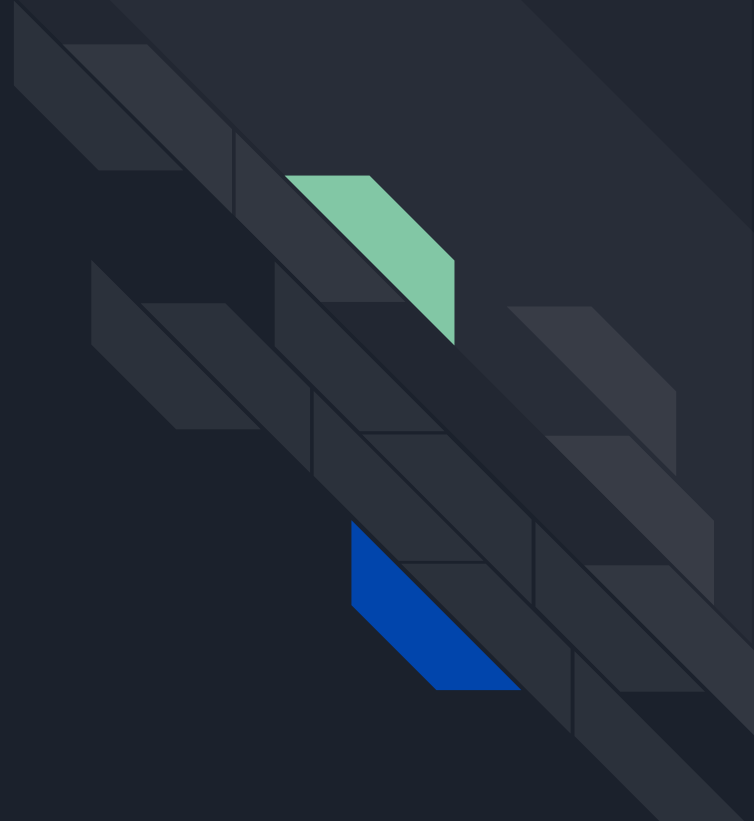
The core objective of this project is to apply unsupervised learning techniques to a wholesale data dataset. The project consists of four main components, including exploratory data analysis and pre-processing, KMeans clustering, hierarchical clustering, and principal component analysis (PCA). Through these steps, we aim to gain insights and discover patterns within the dataset, ultimately providing valuable information for decision-making and business optimization.



Process

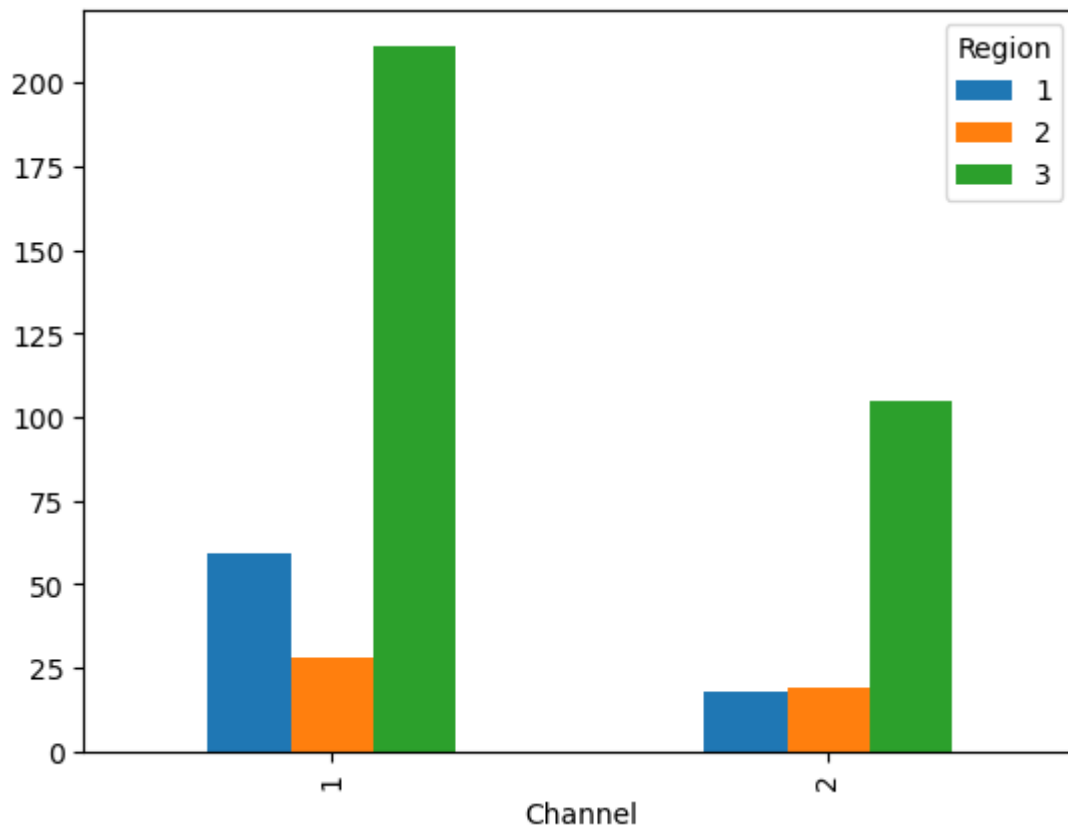
1. EDA - Exploratory Data Analysis & Pre-processing
2. KMeans Clustering
3. Hierarchical Clustering
4. PCA
5. Conclusion

What I have
discovered



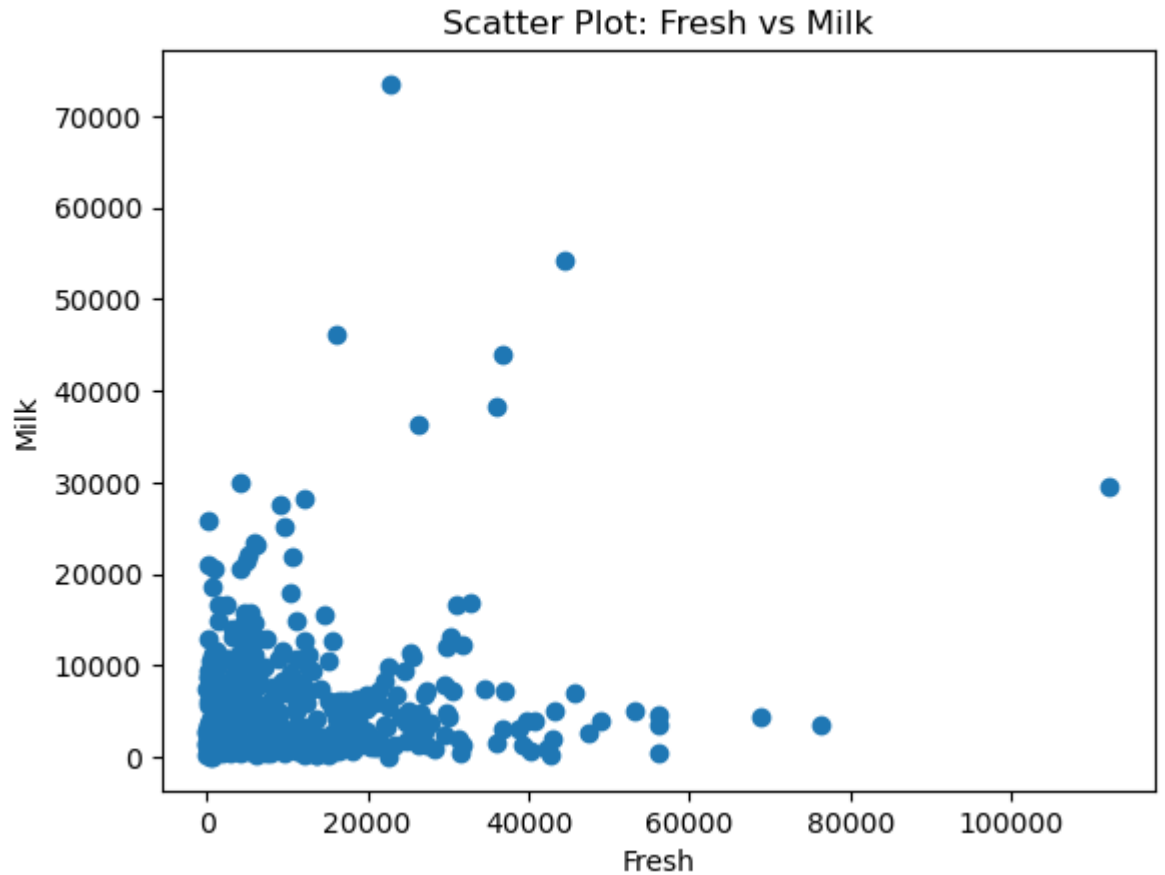
Channel Count


Region	1	2	3
Channel 1	59	28	211
Channel 2	18	19	105





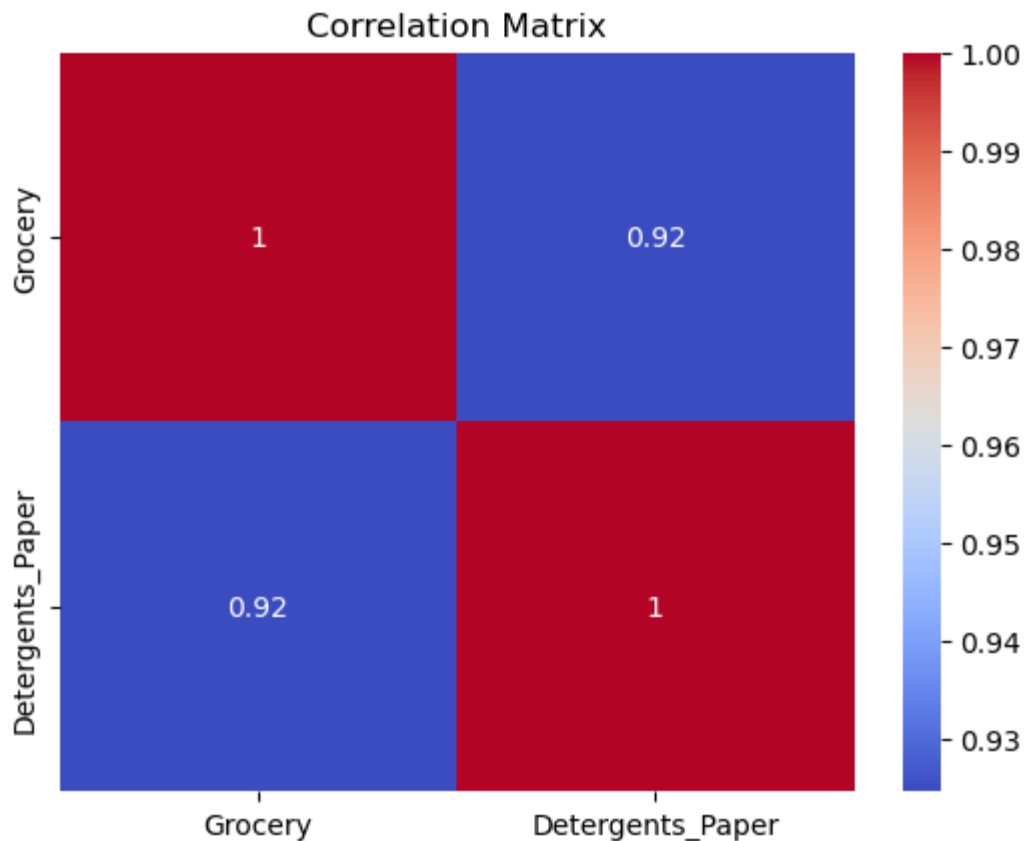
Scatter plot
between "Fresh"
and "Milk" variables





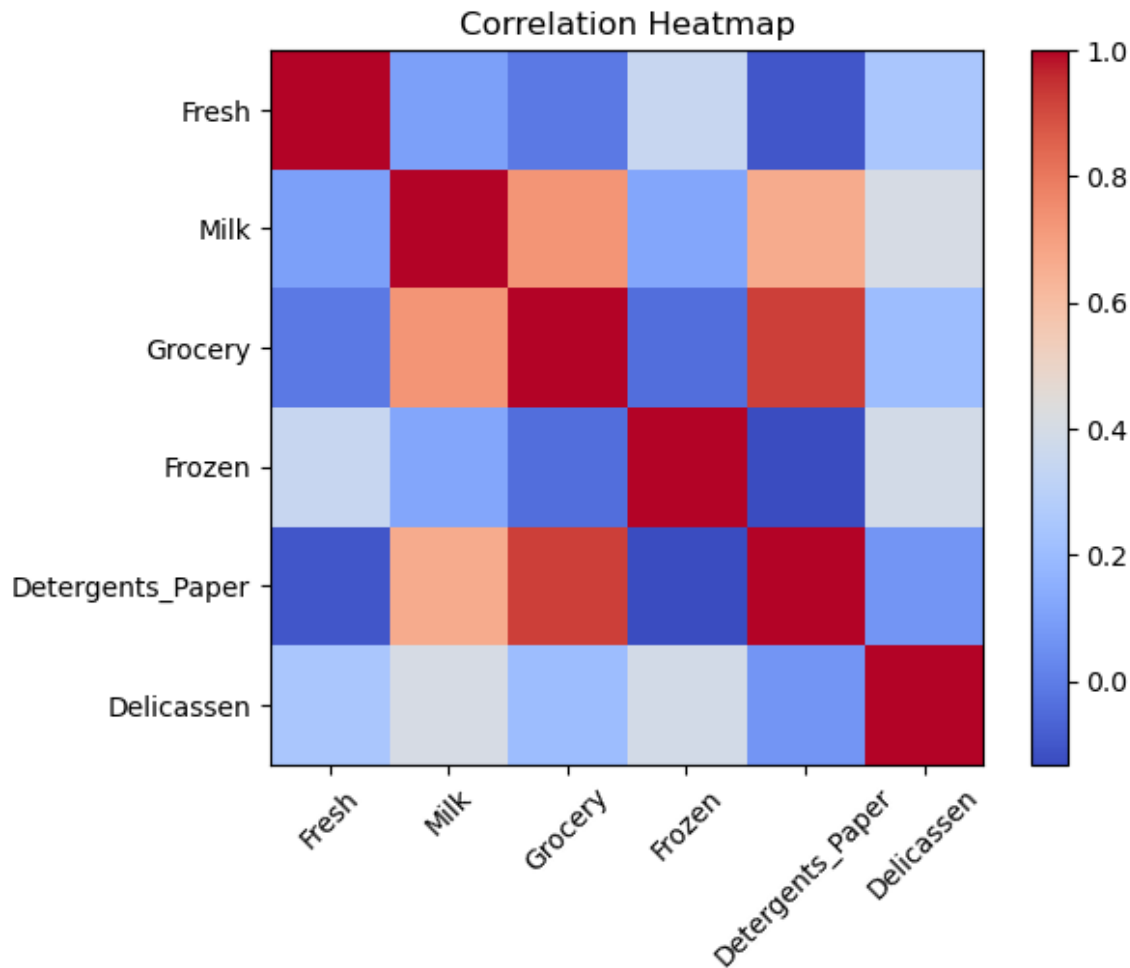
Correlation between Grocery and Detergents paper

There are strong positive correlations observed between the "Grocery" and "Detergents_Paper"



Correlation between multiple variables

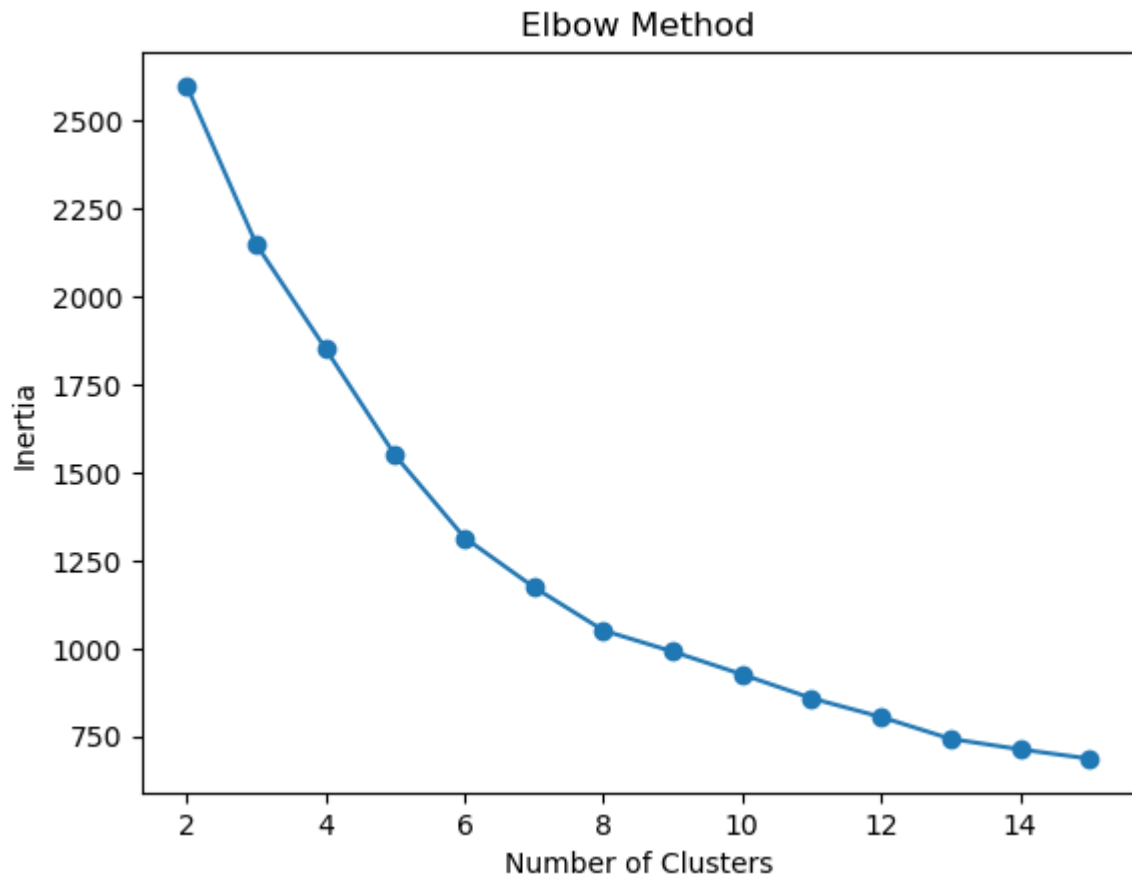
There are strong positive correlations observed between the "Milk" and "Detergents Paper" variables, as well as between the "Milk" and "Grocery" variables.



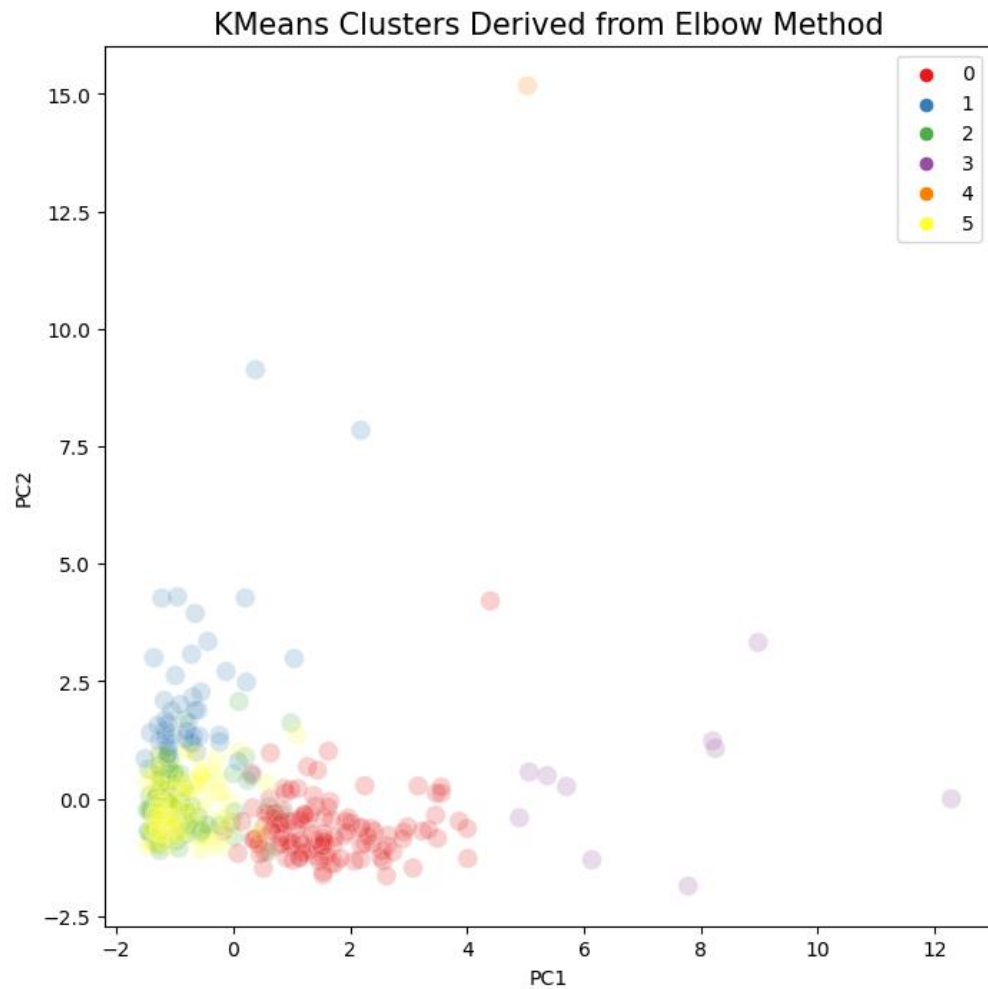


Elbow Method

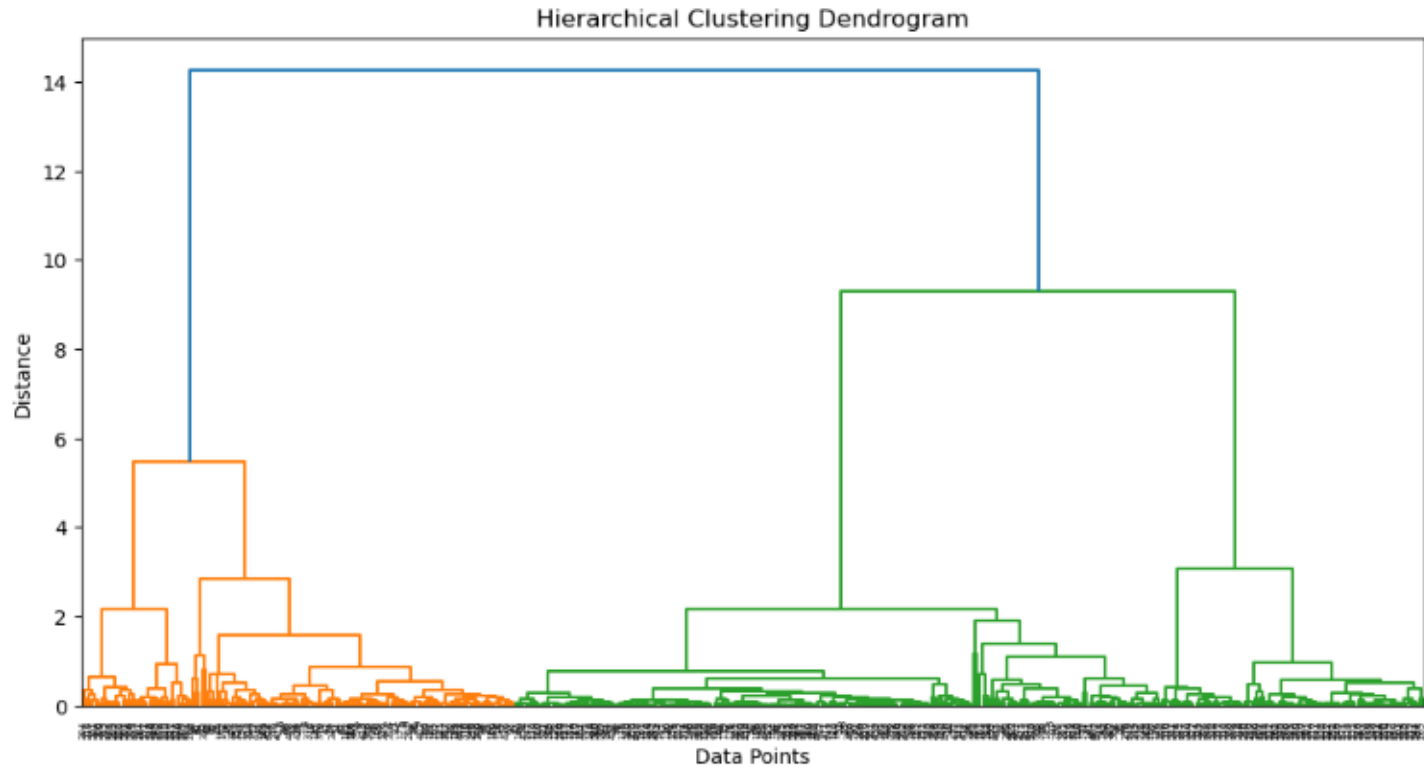
The Elbow Method suggested the number of clusters is 6



Kmeans Clustering

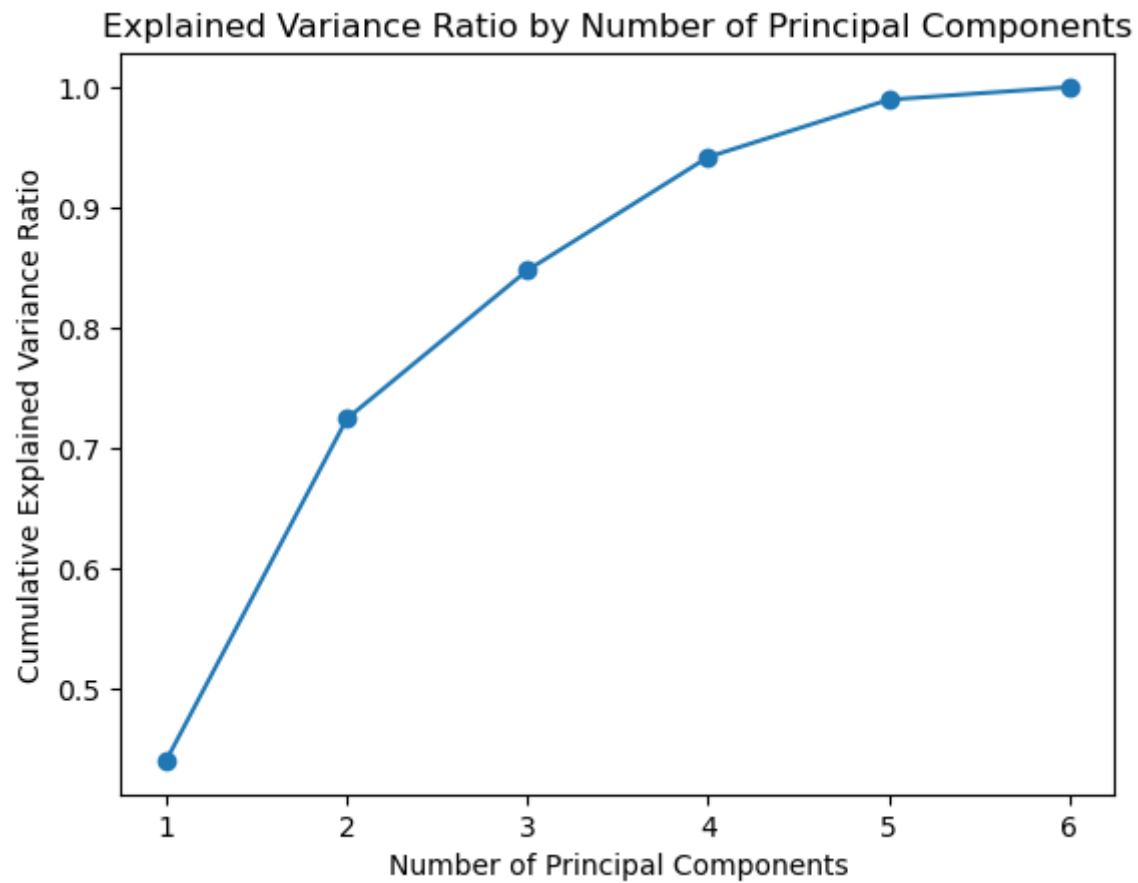


Hierarchical Clustering





PCA





Key insights

Based on the exploratory data analysis (EDA) conducted on the wholesale customer dataset, the following findings can be summarized:

- The distribution of the variables in the dataset is highly skewed, with a majority of the variables having a long tail on the right side. This indicates that there are a few customers who spend significantly more than the majority of customers in each product category.
- There are strong positive correlations observed between the "Grocery" and "Detergents Paper" variables, as well as between the "Milk" and "Grocery" variables. This suggests that customers who spend more on groceries also tend to spend more on detergents and customers who spend more on milk also tend to spend more on groceries.
- The "Fresh" variable shows a relatively low correlation with the other variables, indicating that the spending on fresh products is not strongly associated with the spending on other product categories. This suggests that customers may have different preferences or needs when it comes to fresh products compared to other categories.
- The majority of the customers in the dataset fall into the lower spending range, with a few customers spending significantly higher amounts in each product category. This indicates that there may be different segments of customers with varying purchasing behaviors, such as regular consumers with moderate spending and occasional high-spending customers.

Thank you!

