Unsupervised Learning Wholesale data

Bernard Adeboye

Project/Goals

- Apply unsupervised learning techniques to a real-world data set and use data visualization tools to communicate the insights gained from the analysis

- Perform exploratory data analysis,

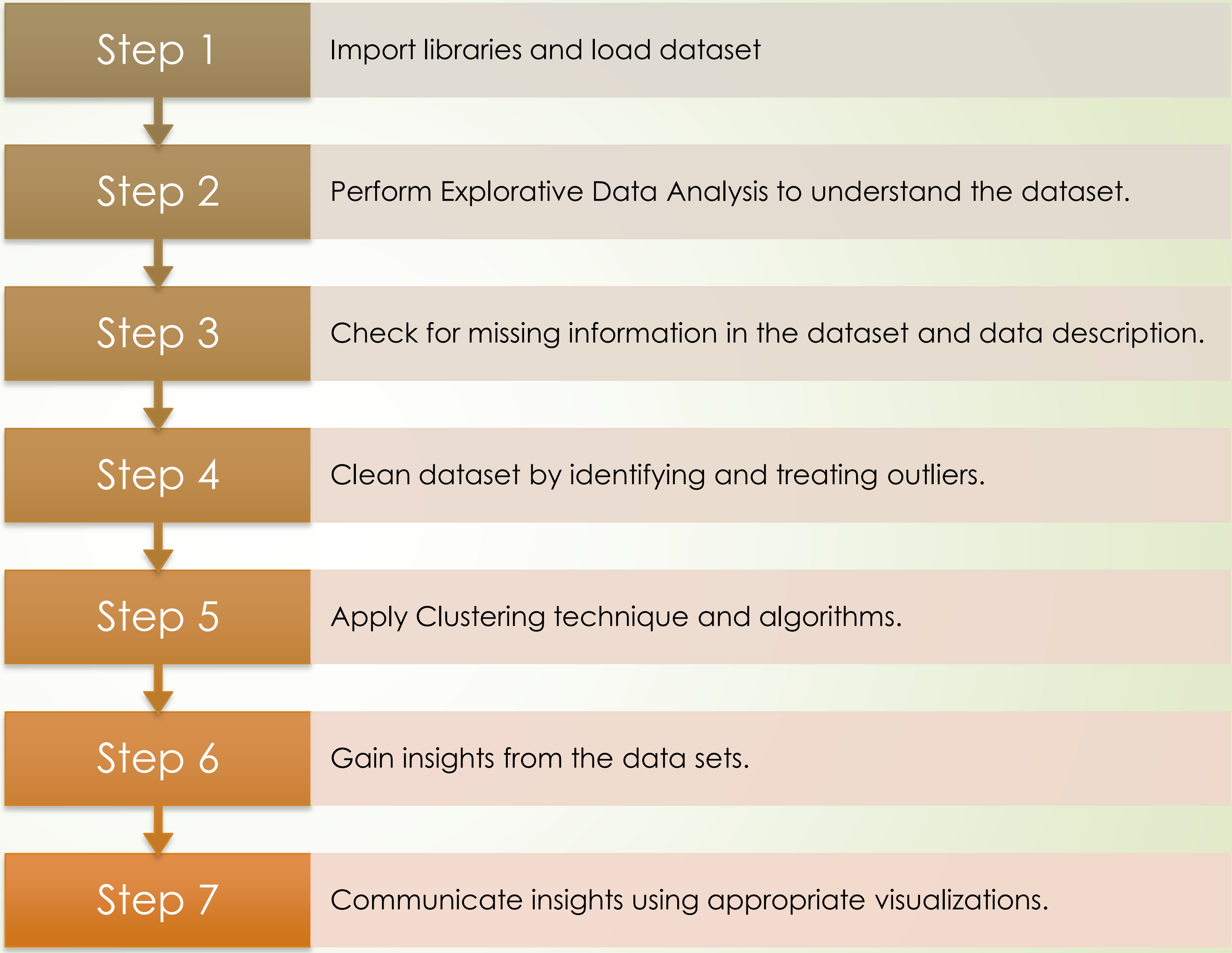
- Preprocessing and feature engineering,

- Perform KMeans clustering,

- Perform hierarchical clustering,

✓ - Perform PCA

Process



Step 1 Import libraries and load dataset

Step 2 Perform Explorative Data Analysis to understand the dataset.

Step 3 Check for missing information in the dataset and data description.

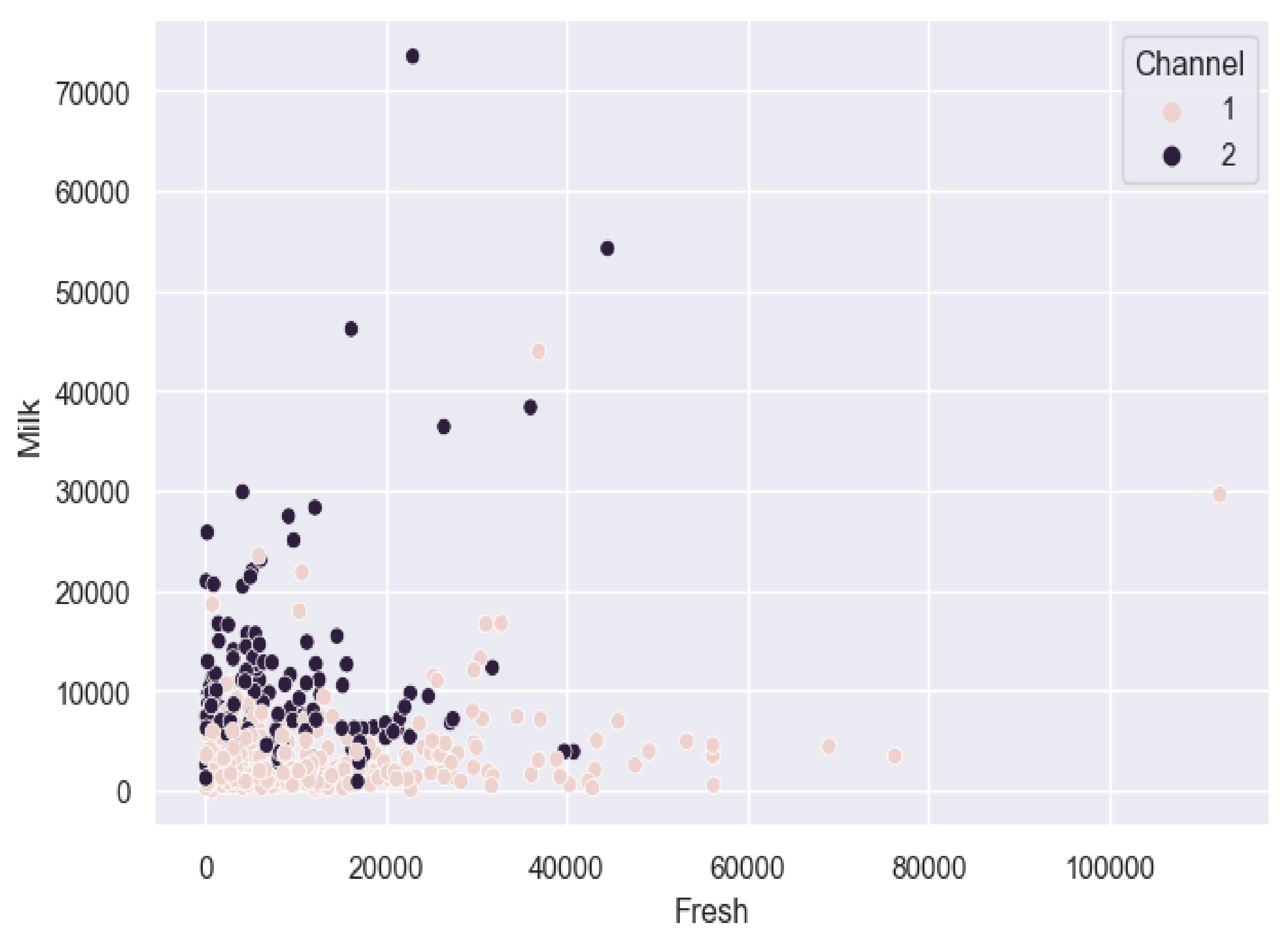
Step 4 Clean dataset by identifying and treating outliers.

Step 5 Apply Clustering technique and algorithms.

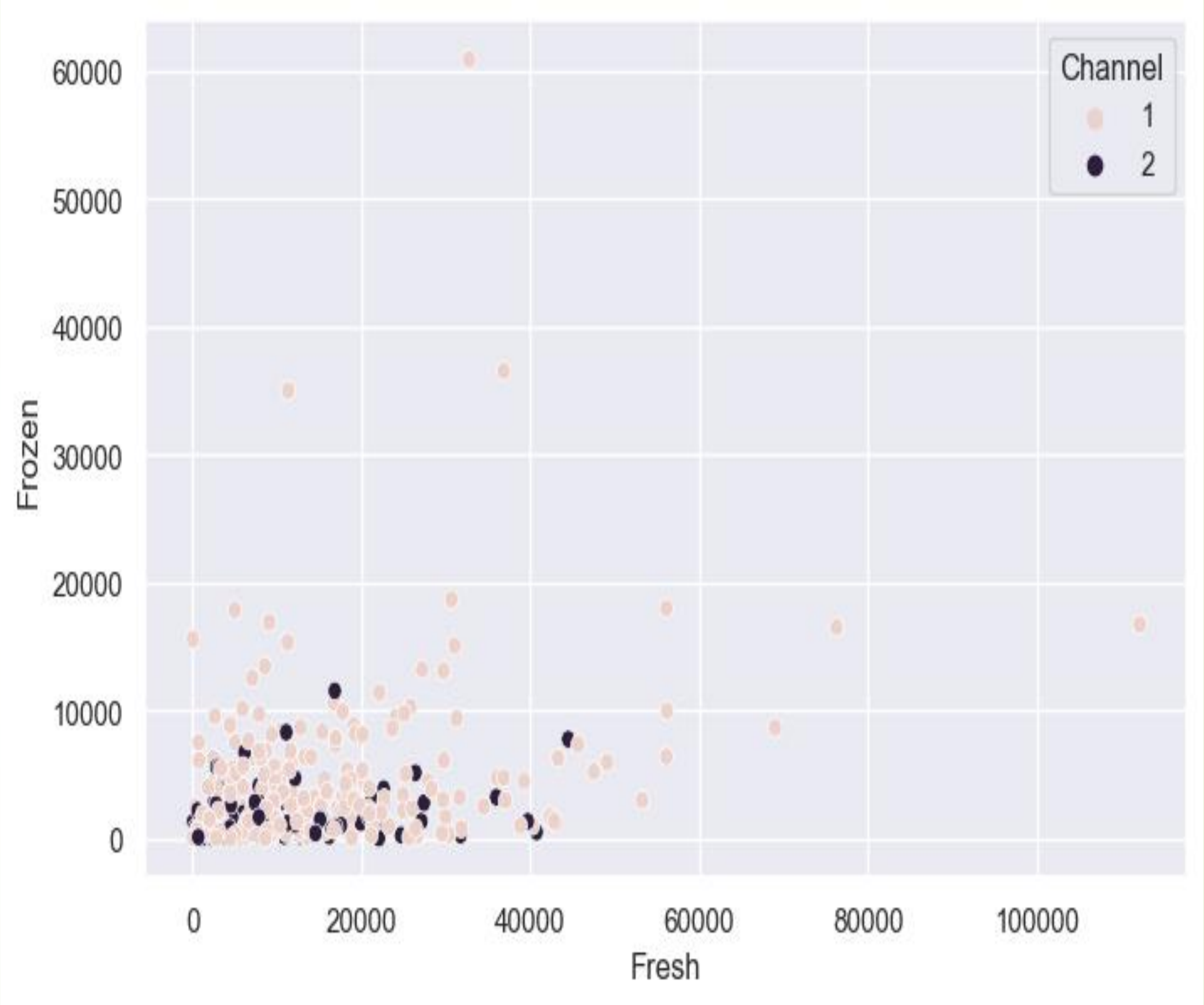
Step 6 Gain insights from the data sets.

Step 7 Communicate insights using appropriate visualizations.

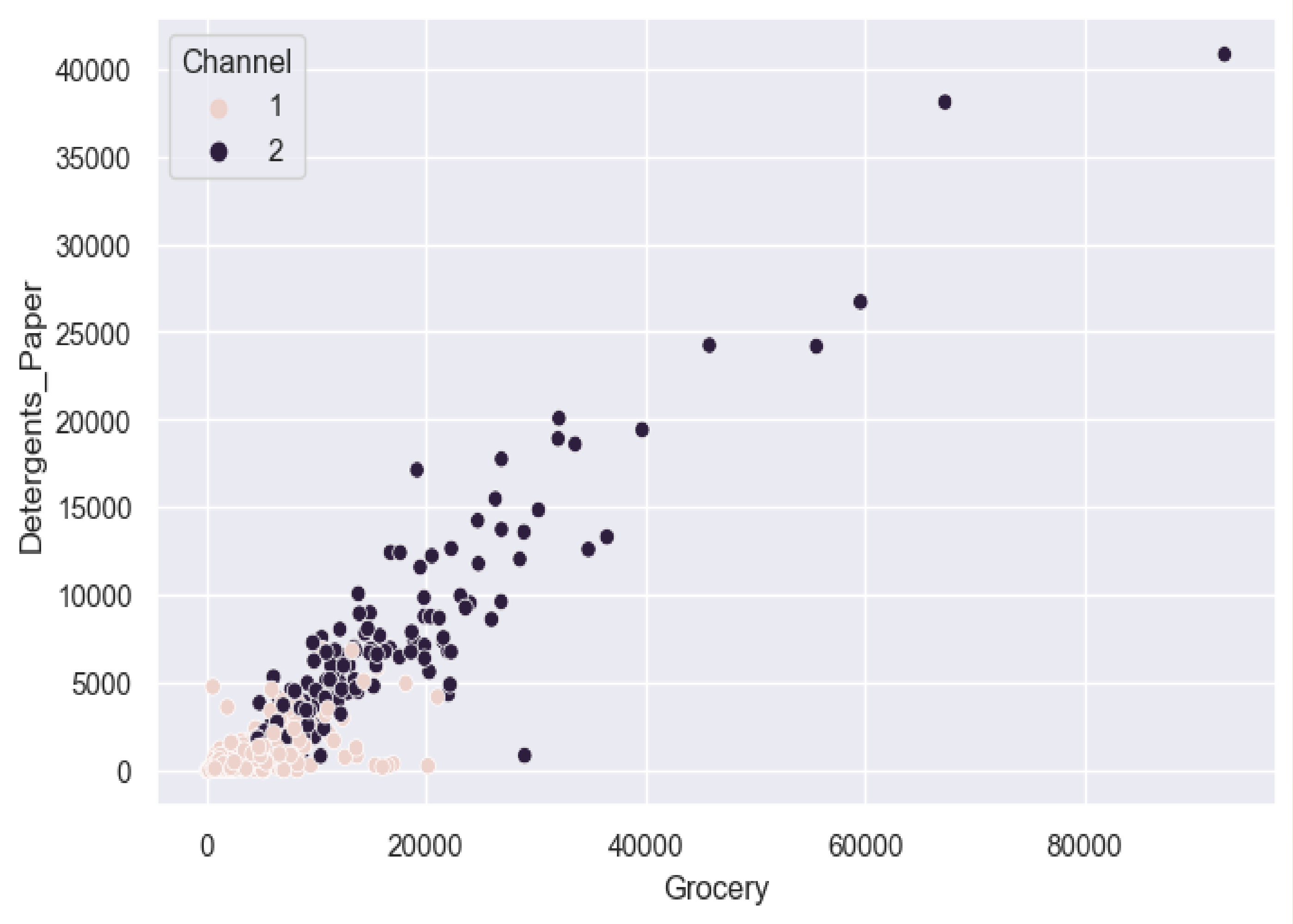
Explorative Data Analysis Relationship between the Fresh, Milk and Channel variables



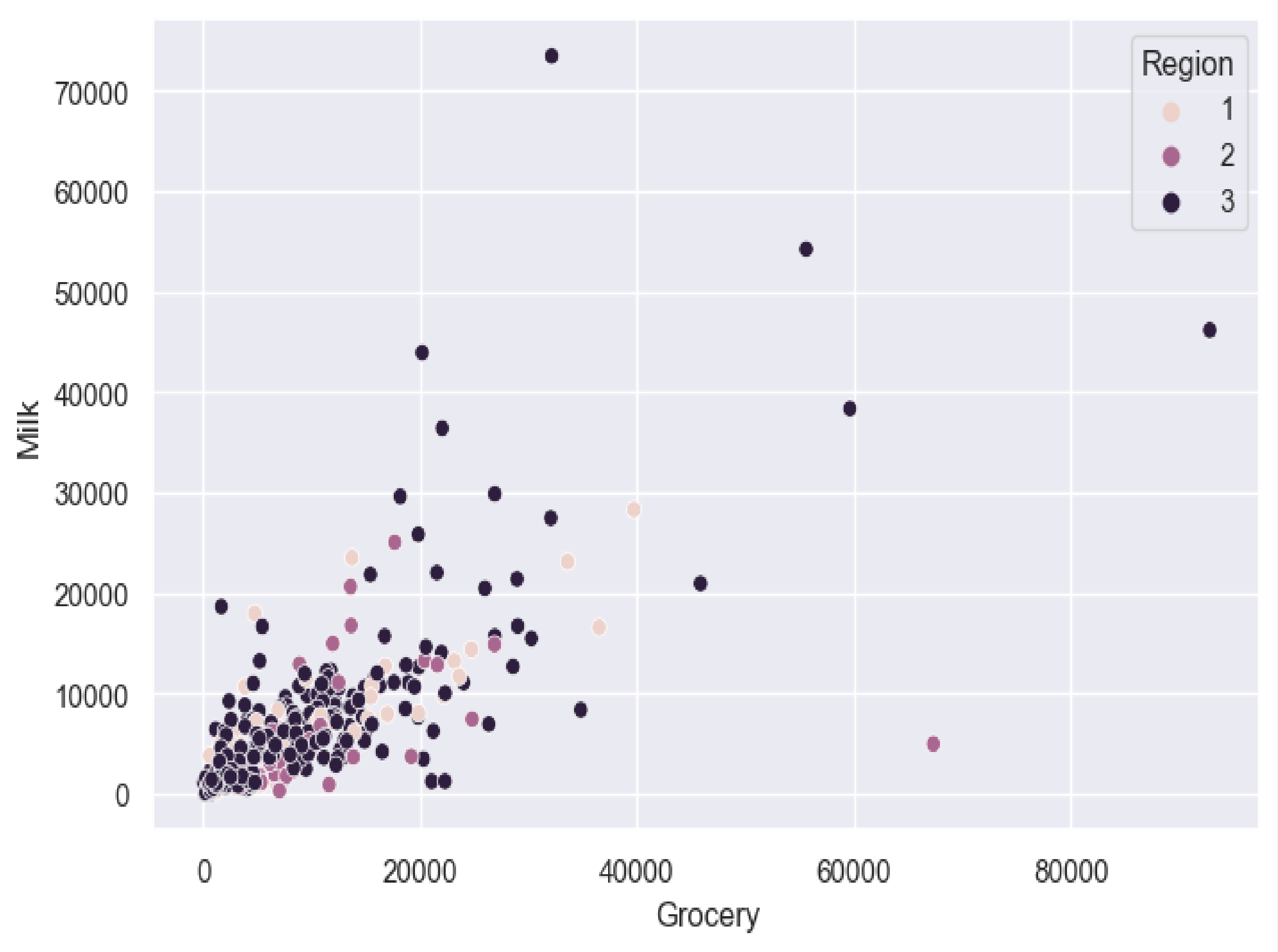
Relationship between the Fresh, Frozen and Channel variables



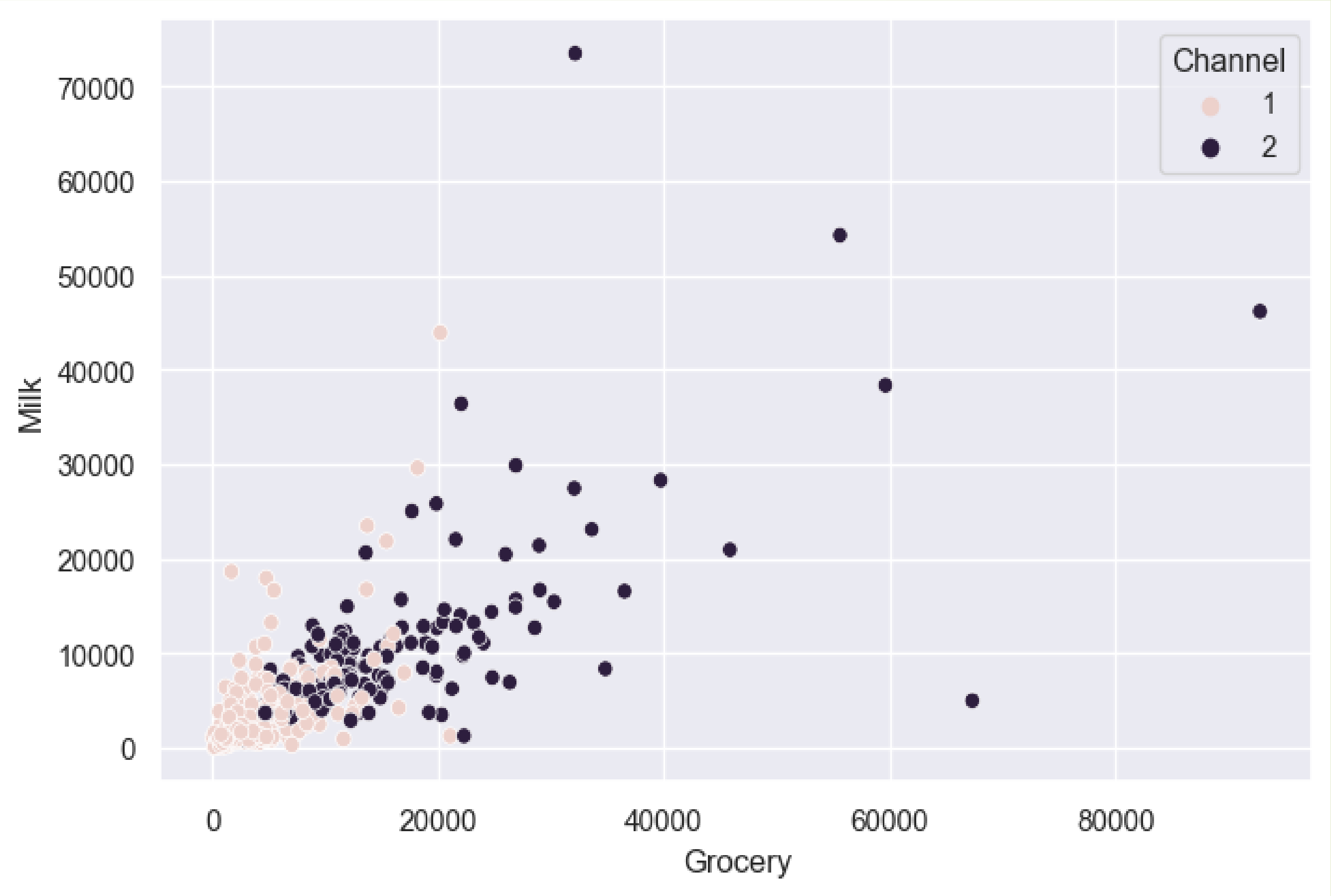
Relationship between the Grocery, Detergents_paper and Channel variables



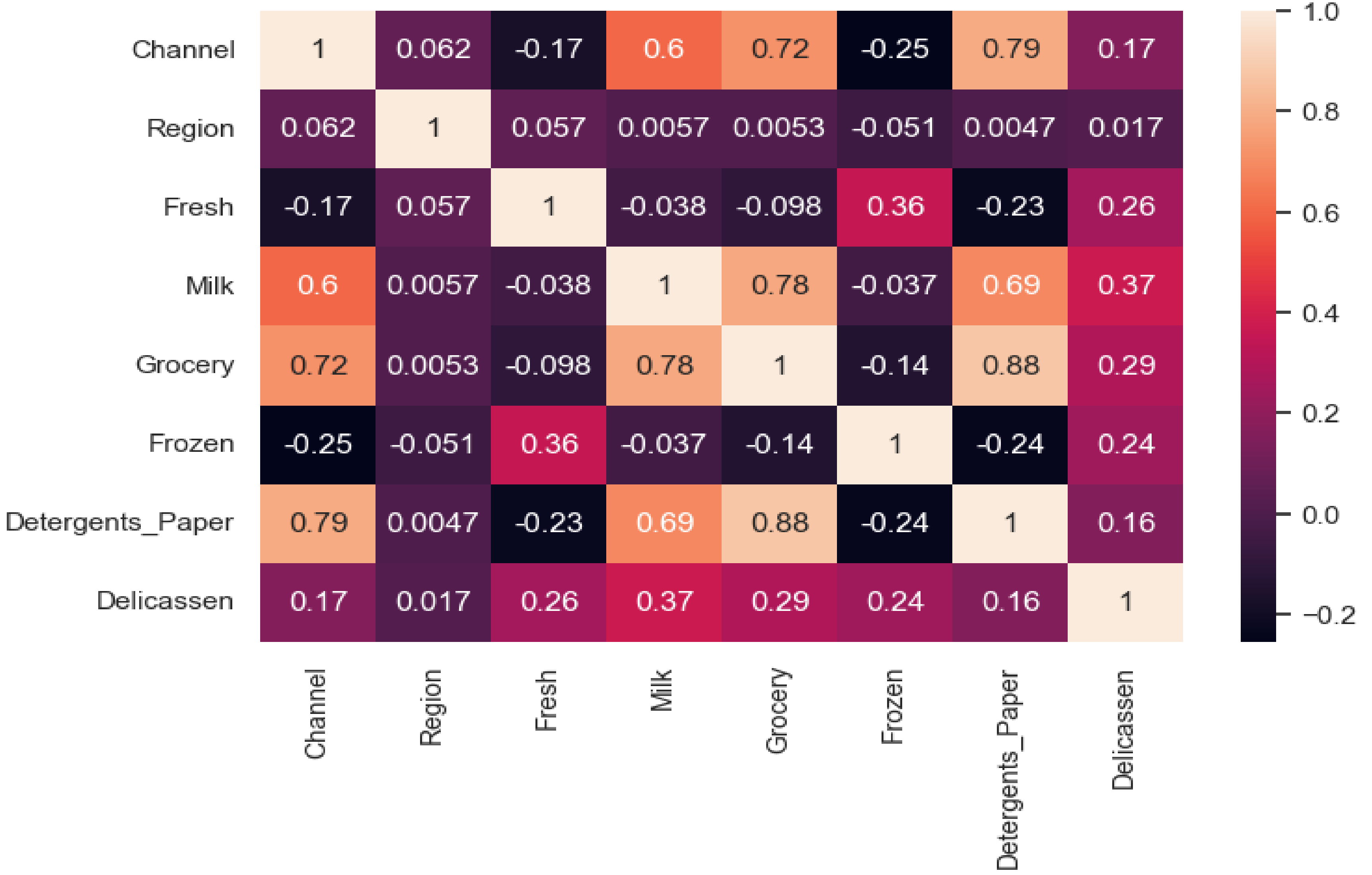
Relationship between the Grocery, Milk and region variables



Relationship between the Grocery, Milk and Channel variables

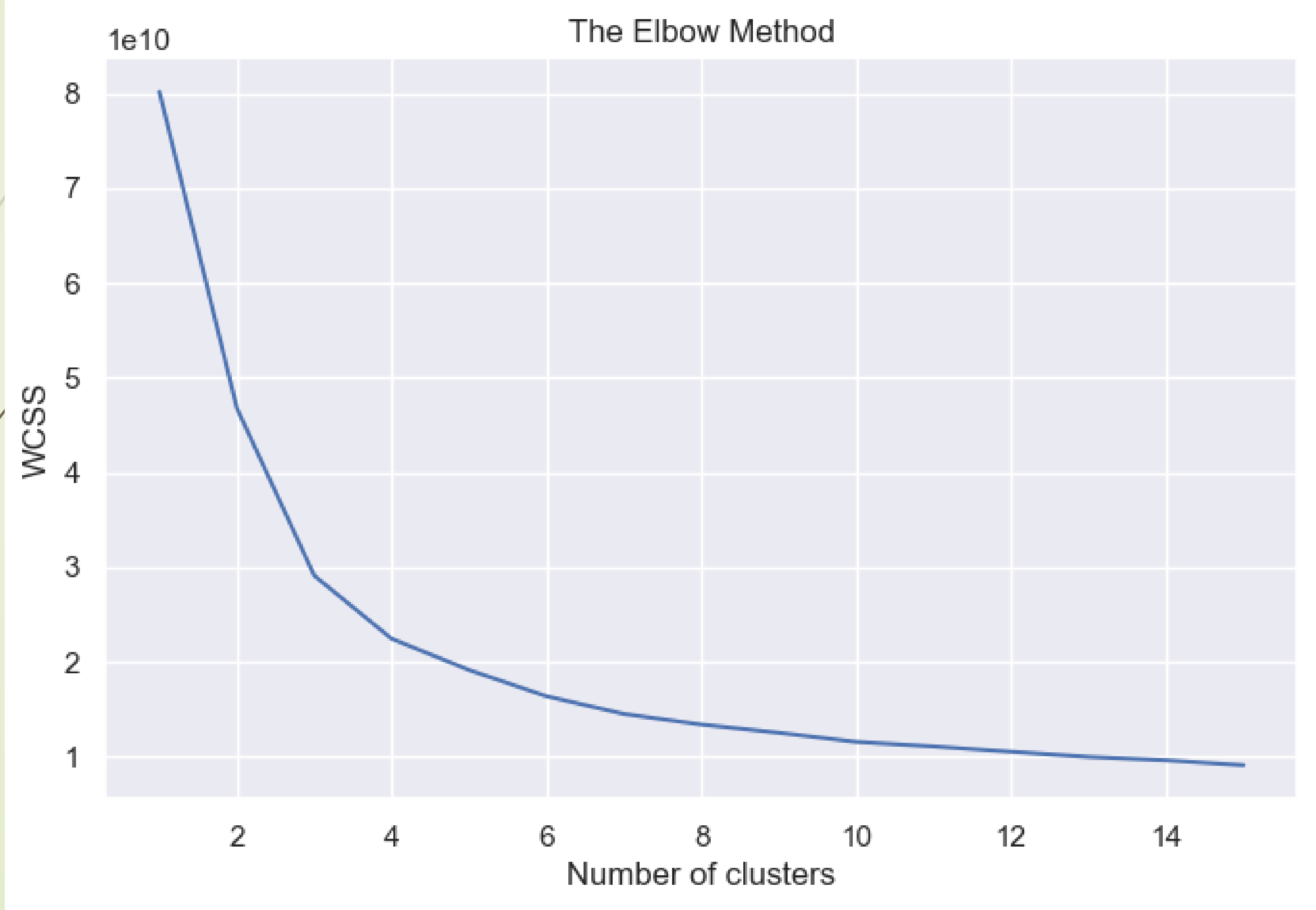


Correlation heatmap

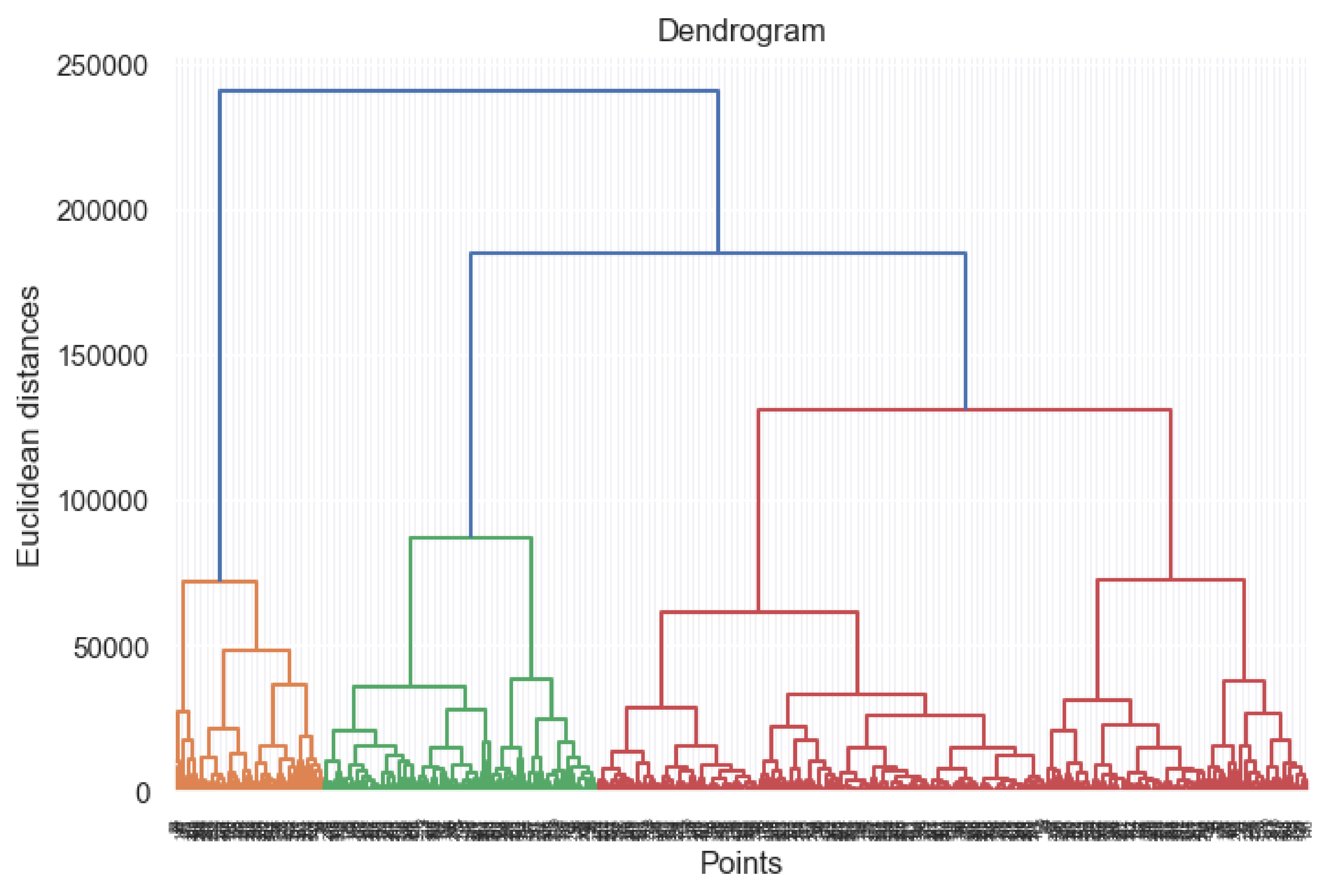


	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Channel	1	0.062	-0.17	0.6	0.72	-0.25	0.79	0.17
Region	0.062	1	0.057	0.0057	0.0053	-0.051	0.0047	0.017
Fresh	-0.17	0.057	1	-0.038	-0.098	0.36	-0.23	0.26
Milk	0.6	0.0057	-0.038	1	0.78	-0.037	0.69	0.37
Grocery	0.72	0.0053	-0.098	0.78	1	-0.14	0.88	0.29
Frozen	-0.25	-0.051	0.36	-0.037	-0.14	1	-0.24	0.24
Detergents_Paper	0.79	0.0047	-0.23	0.69	0.88	-0.24	1	0.16
Delicassen	0.17	0.017	0.26	0.37	0.29	0.24	0.16	1

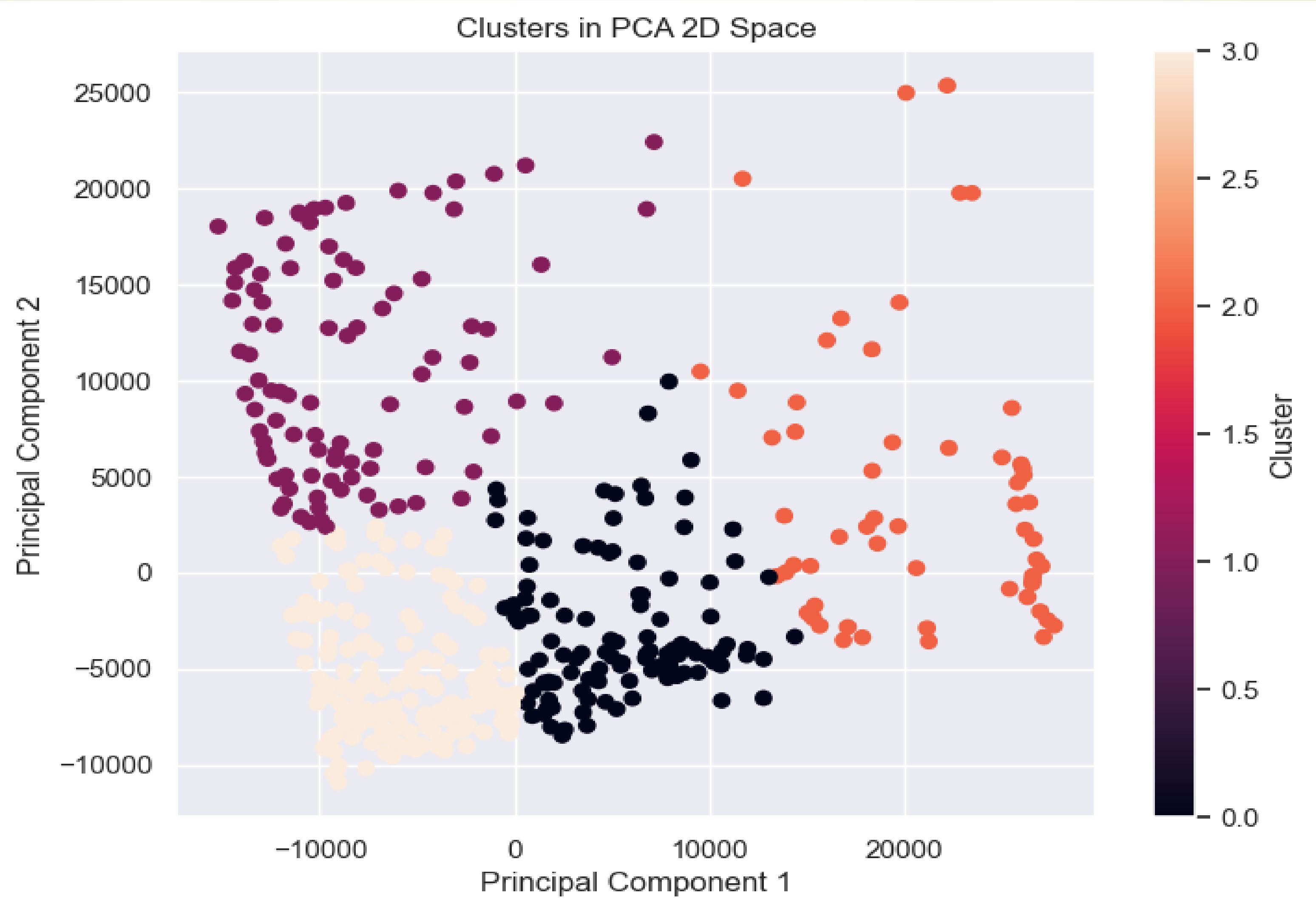
The Elbow Method



Hierarchical Clustering - Dendrogram



Principal Component Analysis



Conclusion

- Performing PCA on dataset for feature reduction yielded four clusters;

    1. Cluster 3 is the combination of products that falls around zero and negative. Stakeholders should consider promo or discount to encourage customers to engage in the purchase of those items.

    2. Cluster 2 is the combination of products that are of high interest to customers, therefore, stakeholder should provide measure to retain this group and provide ways to increase their purchasing activities.

- Performing K-means on the dataset grouped the products into 4 clusters.
- Right Skewness: Features such as 'Fresh', 'Milk', 'Grocery', 'Frozen', ‘Detergents Paper', and 'Delicassen' show a right skew (mean > median).This could impact the performance of some machine learning algorithms.
- Combination of Milk and detergents paper/Grocery shows a positive linear relationship with channel 2 and region 3 playing a significant part.
- There is a very strong Relationship between the Grocery, Detergents_paper and Channel variables
- The Hierarchical Cluster shows two clusters based on agglomerative clustering.