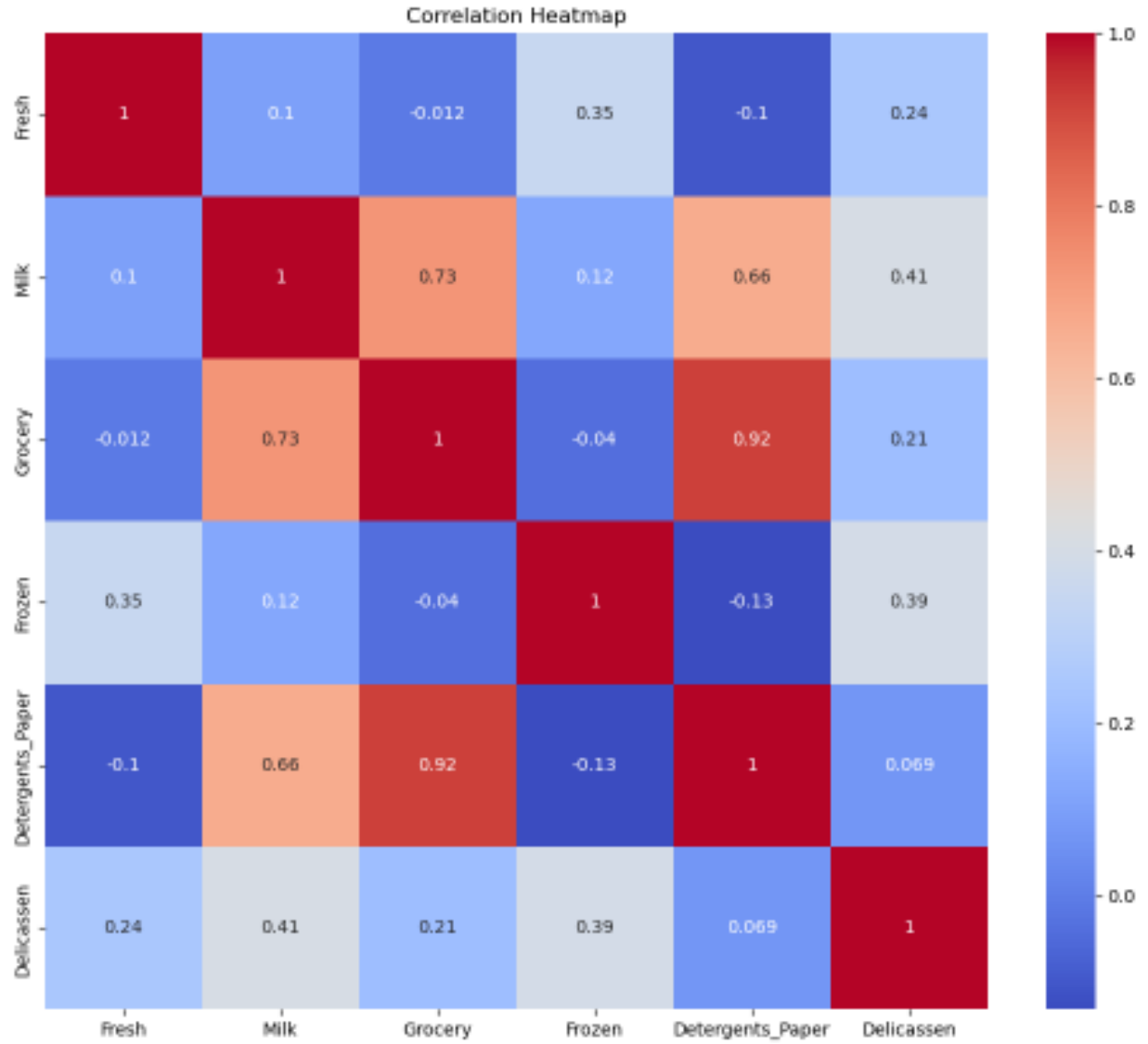
Project-Unsupervised Learning

TES BEYENE

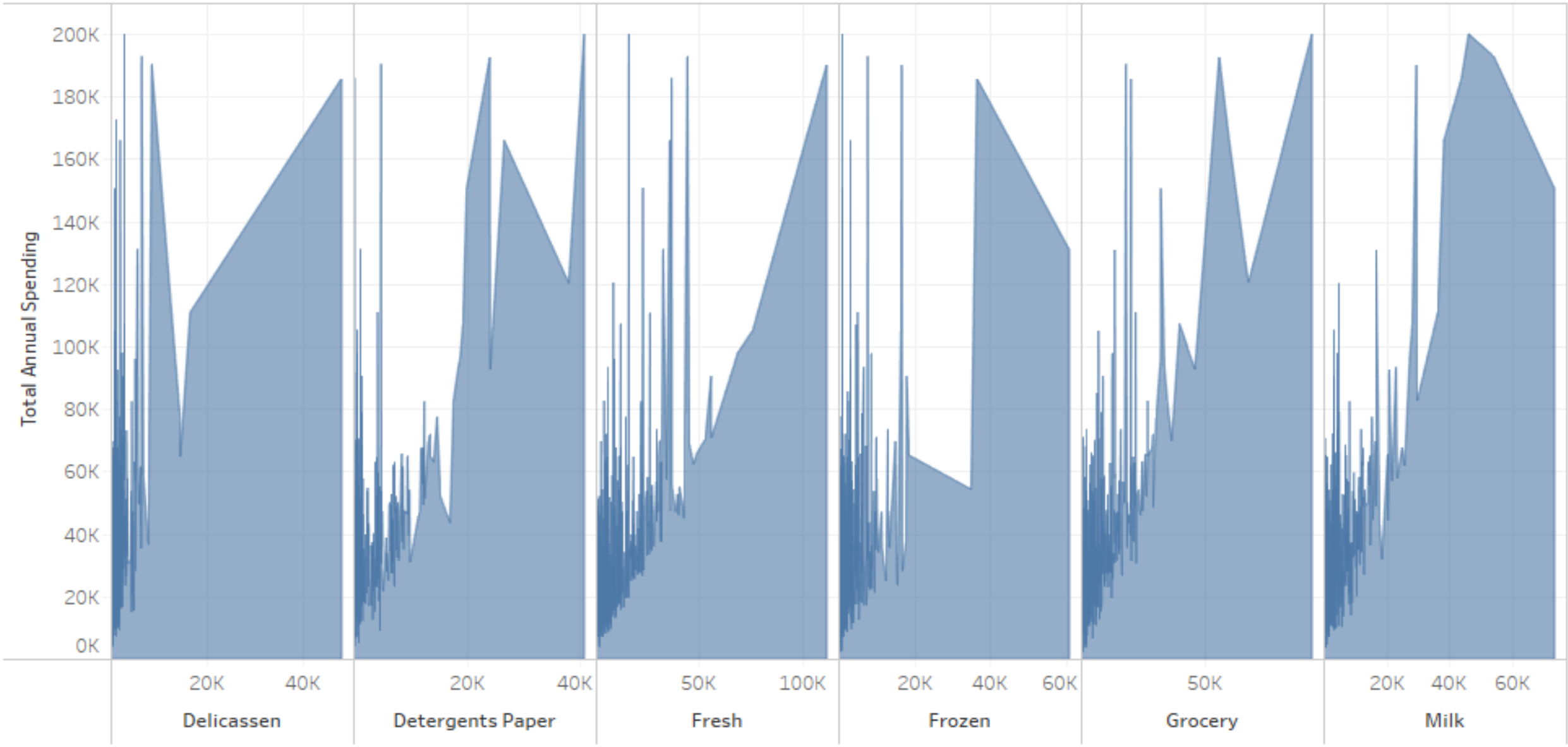
Project Goal

Perform unsupervised learning techniques specifically annual spending of product categories on a wholesale data dataset using KMeans clustering, hierarchical clustering, and PCA and communicate insights on the clustering result and what it means for our business.

EDA & Visualization

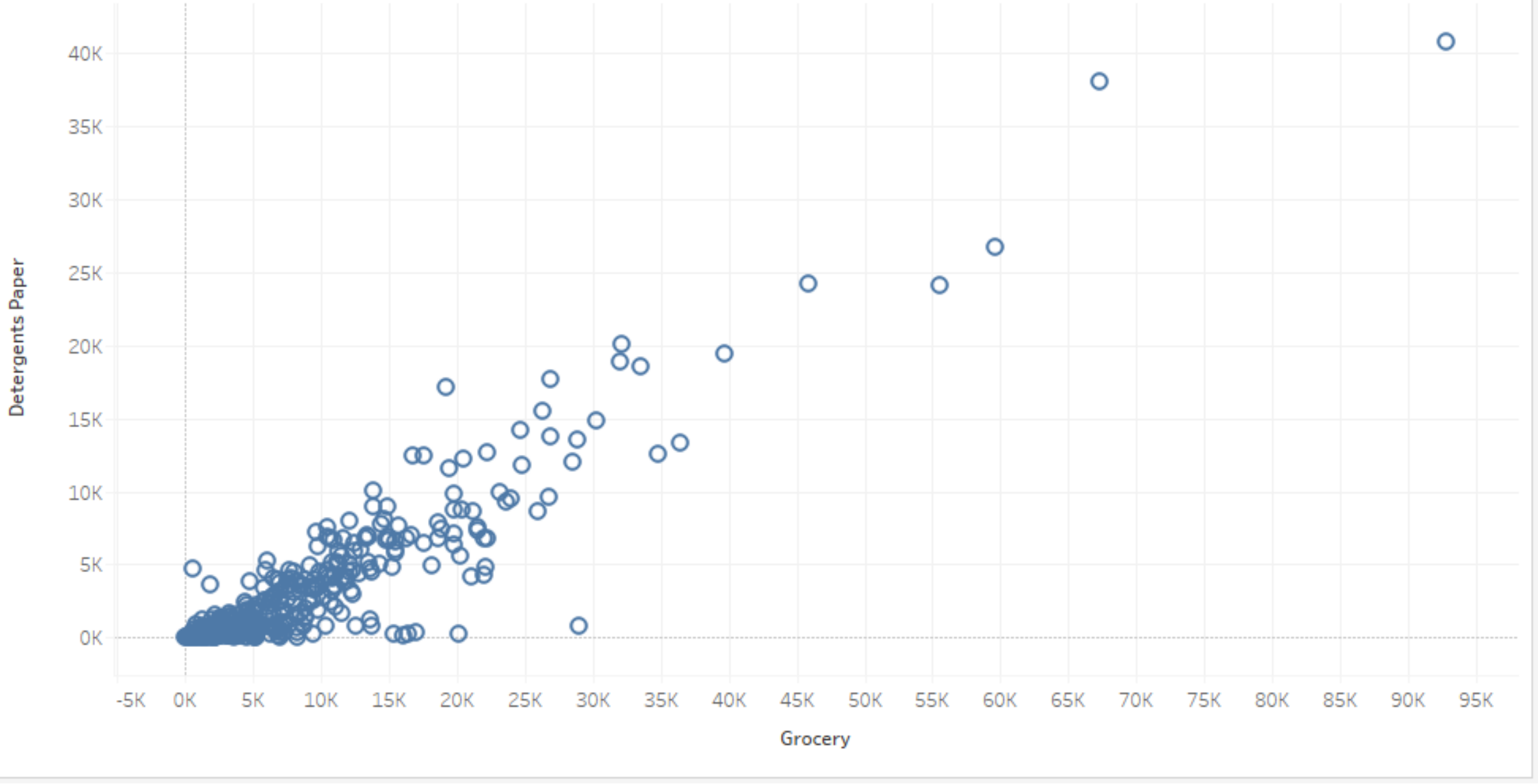


Total Spending by Categories

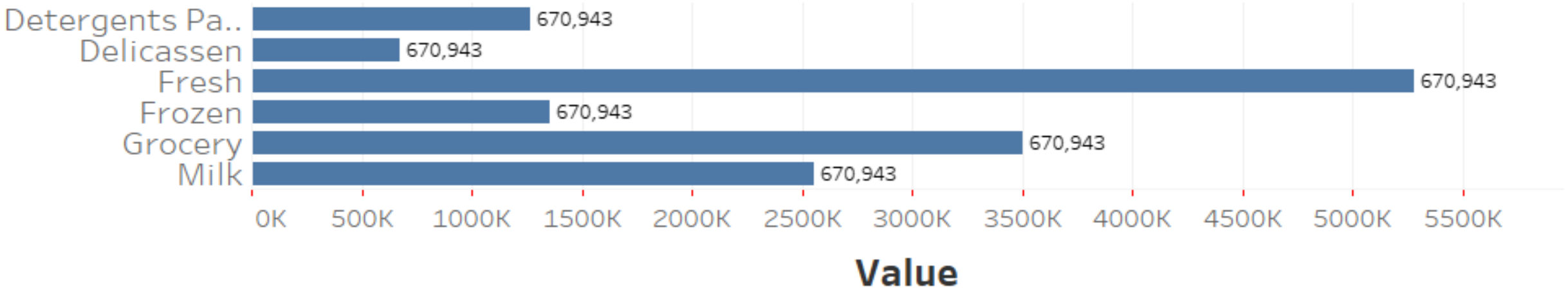


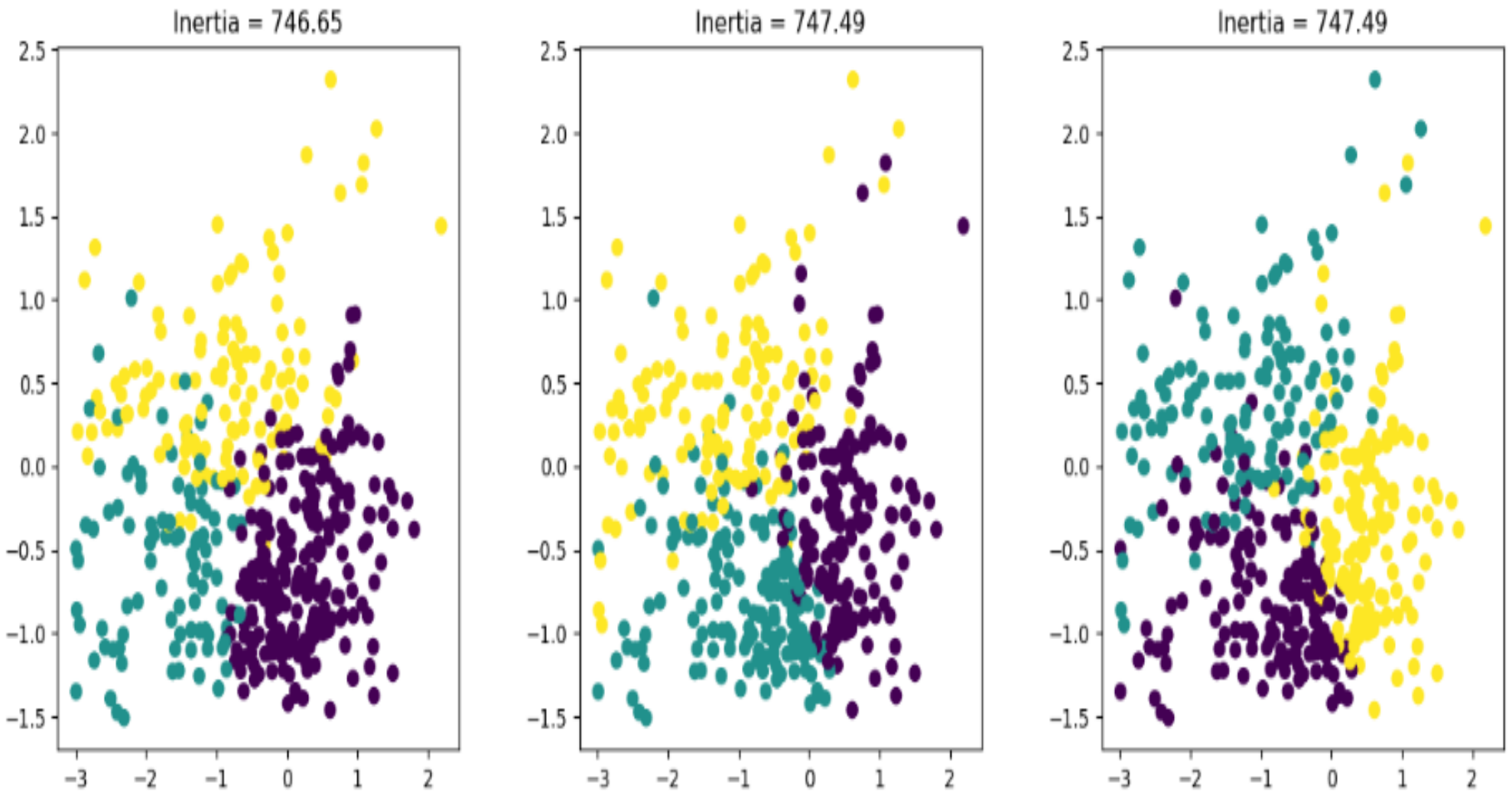
double-click to start a new calculation.

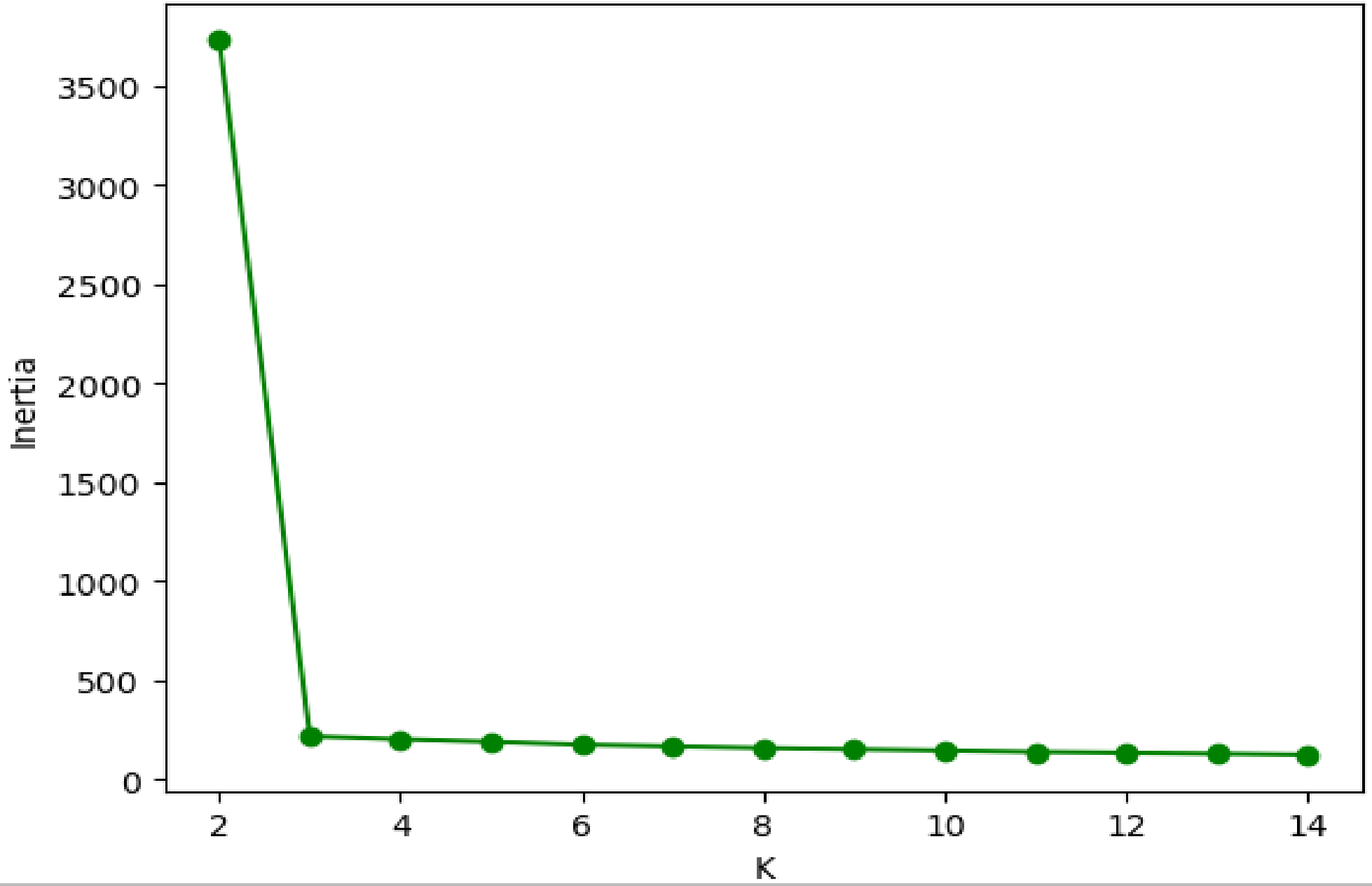
Grocery Vs Detergent Paper

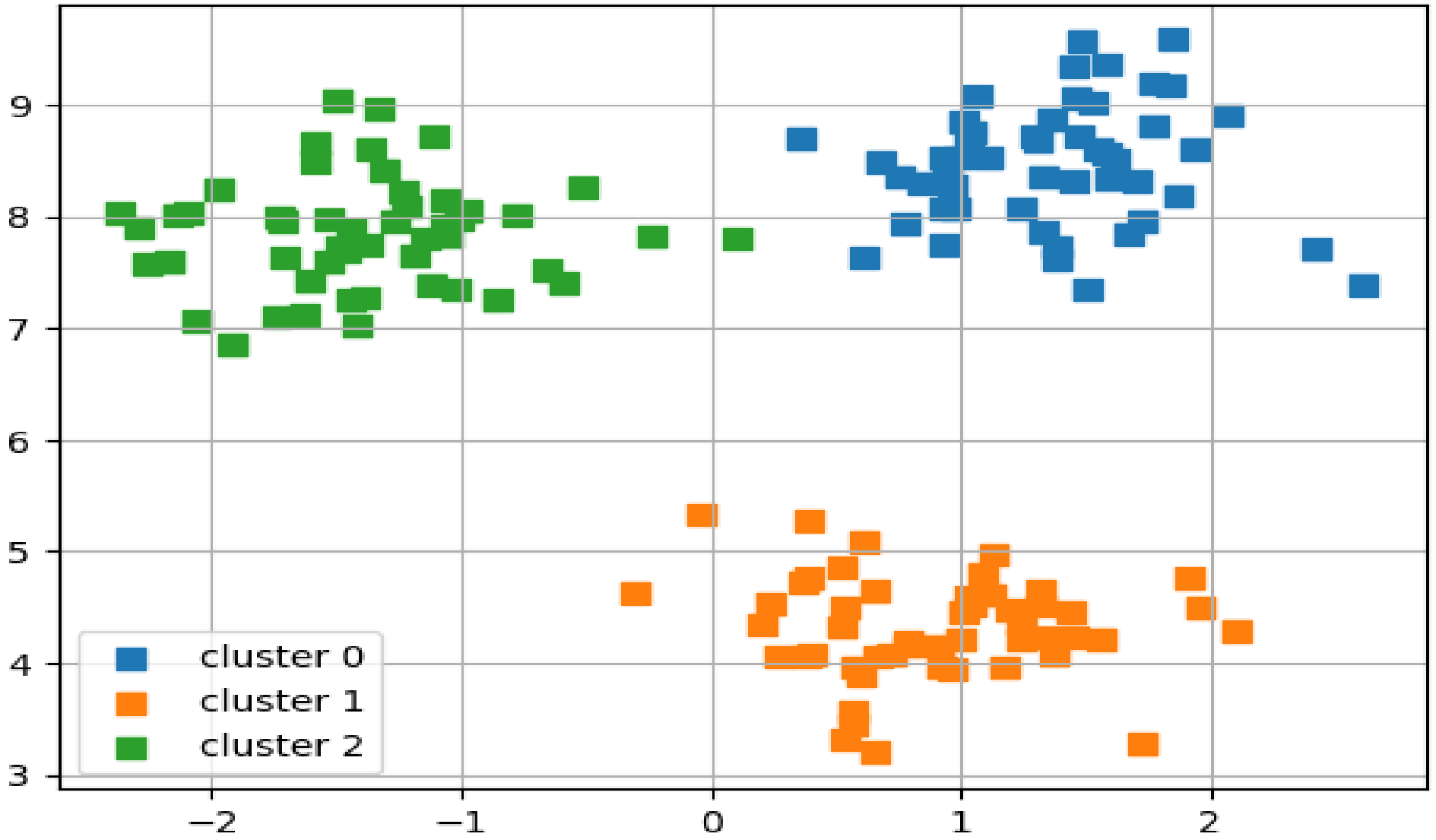


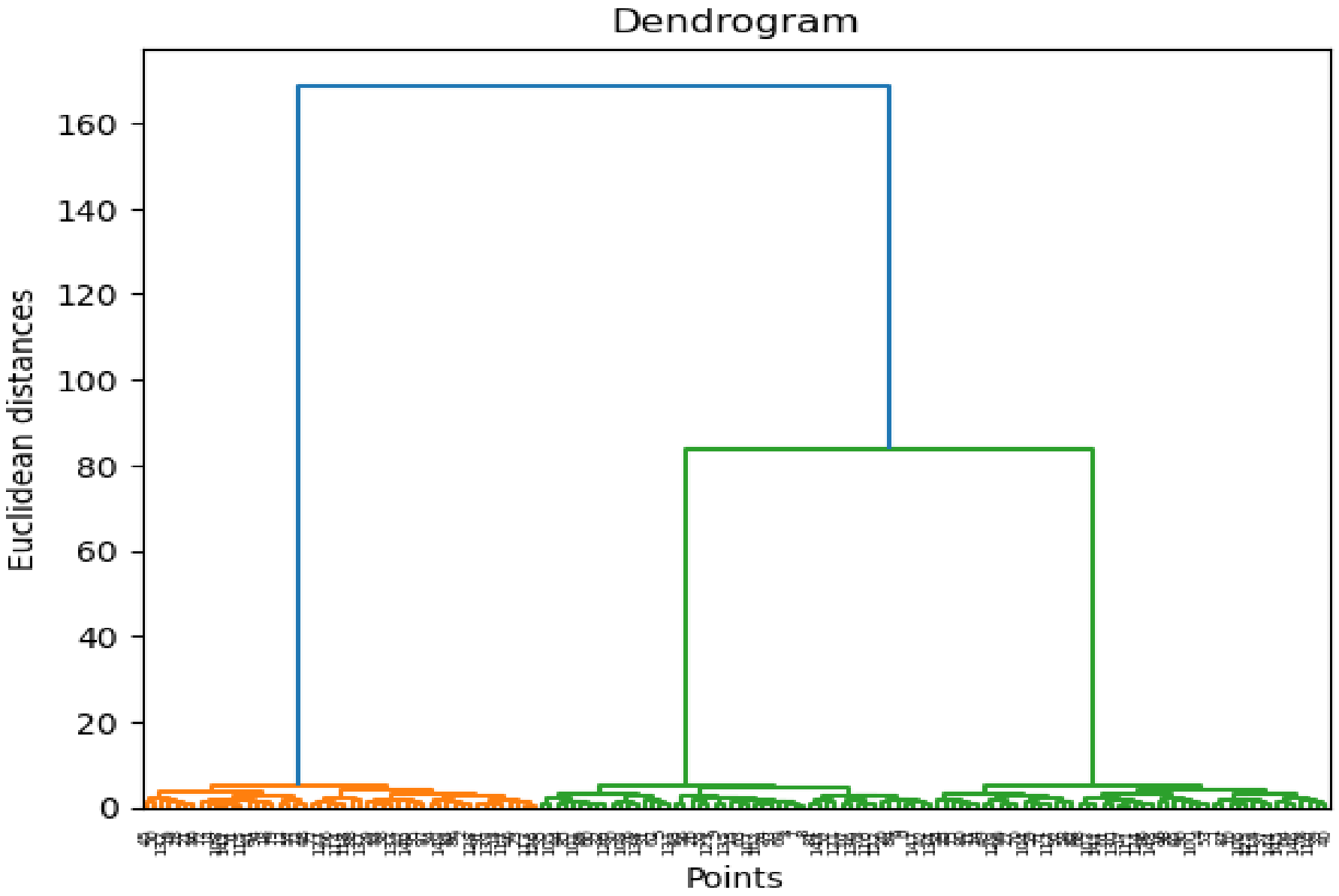
Annual Spending By Product Category

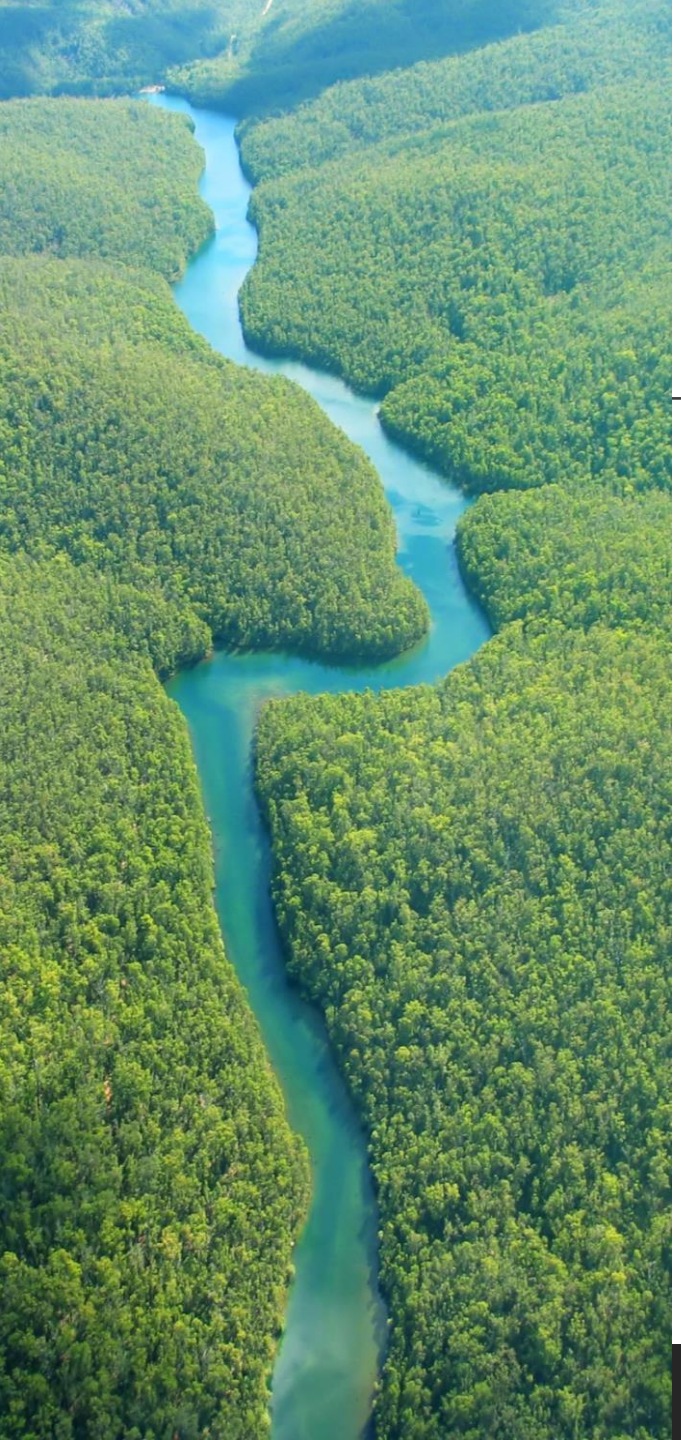



Inertia = 746.65
Inertia = 747.49
Inertia = 747.49
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−3
−2
−1
0
1
2


Inertia
3500
3000
2500
2000
1500
1000
500
0
2
4
6
8
10
12
14
K


9
8
7
6
5
4
3
−2
−1
0
1
2
cluster 0
cluster 1
cluster 2


Dendrogram
Euclidean distances
160
140
120
100
80
60
40
20
0
Points

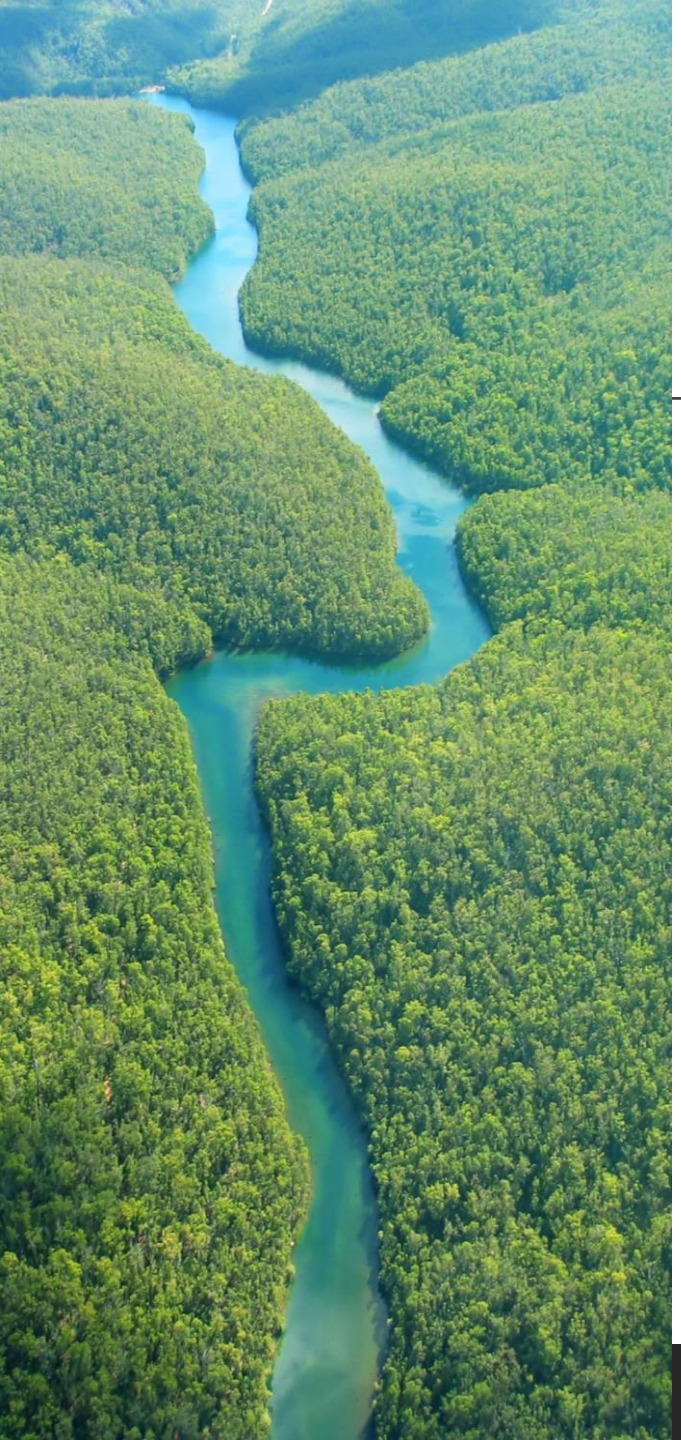


Findings

1. Noticed from the distribution plot of the 6 product that they are not normally distributed.

2. From the above heatmap we noticed that there was significant relationship between Milk & Grocery, Milk & Detergent Paper and at finally between Grocery & Detergent Paper.

3. The result of the Kmeans clustering show that the optimal number of clusters is three. That means the 6 products features can be reduced in to three categories based on their underlying patterns and relationships.



Findings

4 . The hierarchical as well grouped the products in to three clusters based on the dendrogram plot.

5. The result of the PCA show that dimensions can be reduced to 3(from the 6-dimensional dataset) as the three components explain around 85% of the variability in the dataset.

6. Since all the results are the similar, we can conclude the annual spending can be categorized in to three categories based on the underlying relationship and correlation on the product category features.

Challenge & Future Plans

- Focusing was difficult given that I knew I was a bit behind for some reason on my materials and preparation for the final project.

- If I had more time, I could have done more commits, improved the analysis and tried more clustering algorithms.