Unsupervised Machine Learning Project by *Ikenna Odinye*

Project goal

The goal is to analyze product categories of a Wholesale dataset, visualize patterns, develop unsupervised machine learning algorithm and communicate insights to stakeholders based on findings

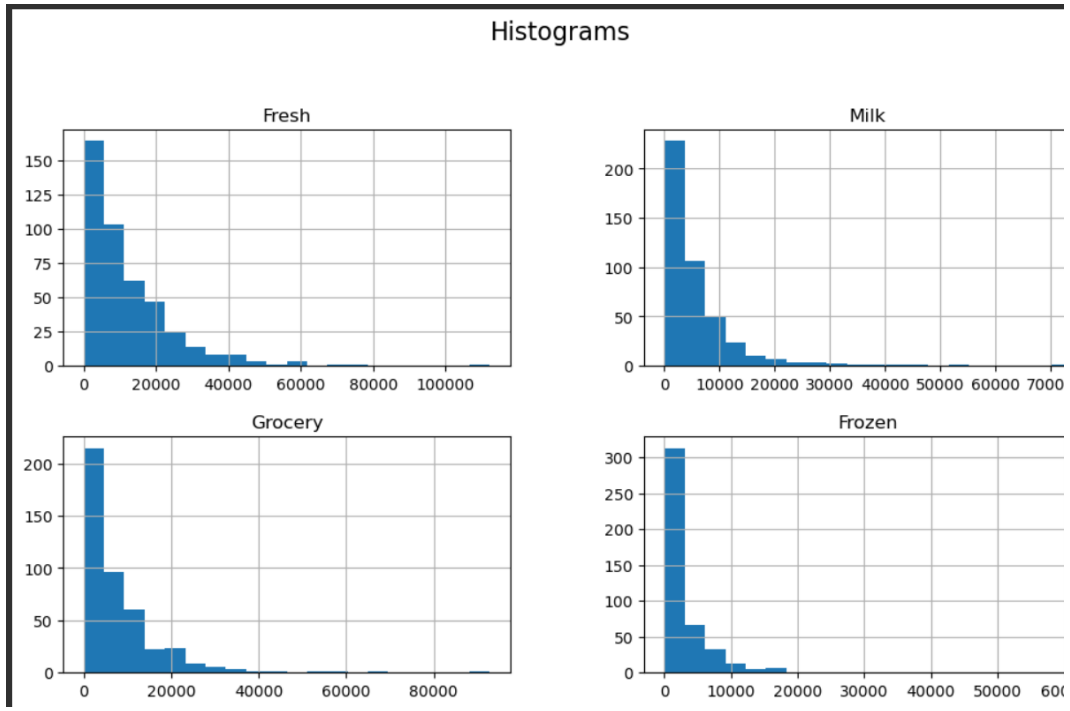
Part I : EDA - Exploratory Data Analysis & Pre-processing

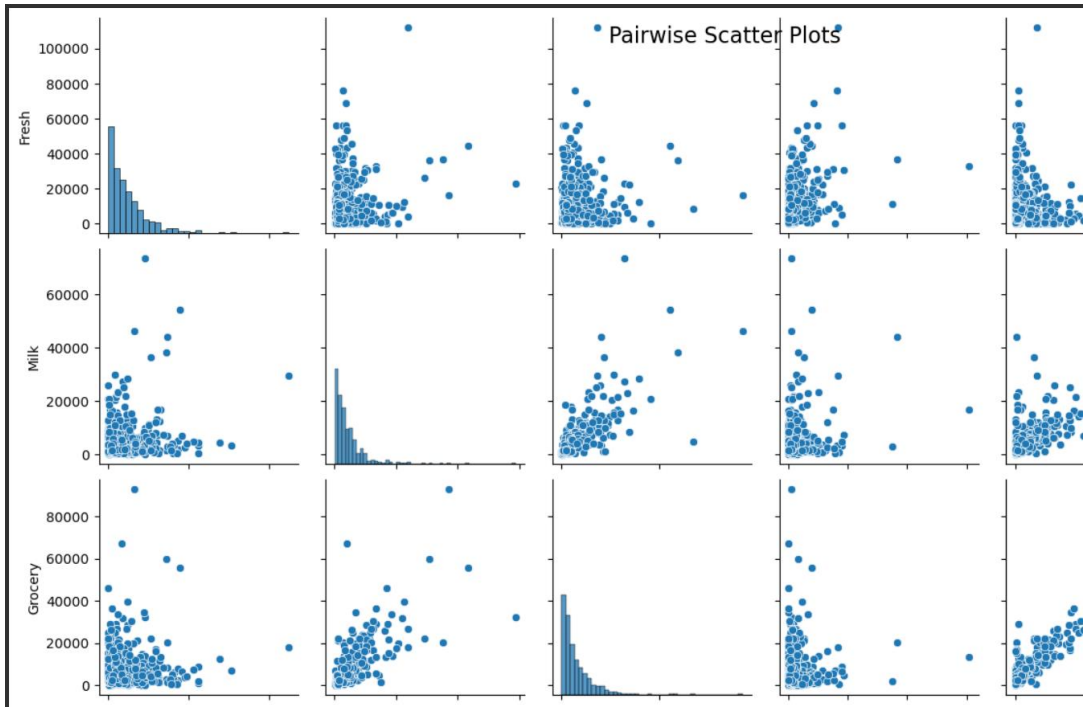
```
df.head()
```

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	2	3	12669	9656	7561	214	2674	1338
1	2	3	7057	9810	9568	1762	3293	1776
2	2	3	6353	8808	7684	2405	3516	7844
3	1	3	13265	1196	4221	6404	507	1788
4	2	3	22615	5410	7198	3915	1777	5185

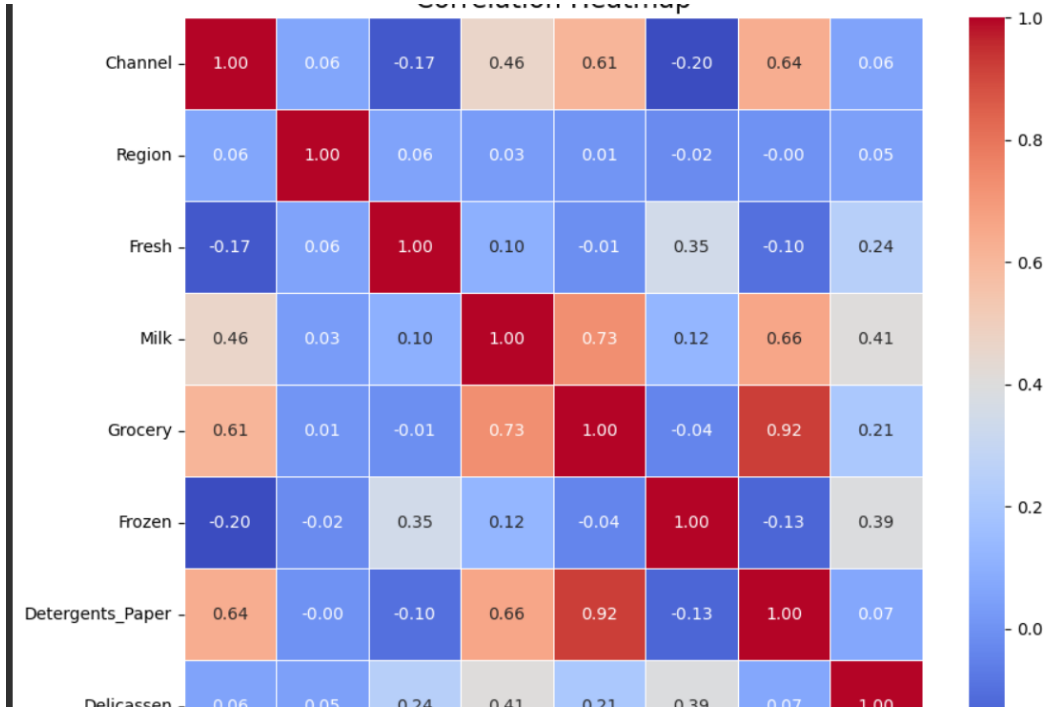
	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	1.322727	2.543182	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	0.468052	0.774272	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	1.000000	1.000000	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	1.000000	2.000000	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	1.000000	3.000000	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	2.000000	3.000000	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	2.000000	3.000000	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

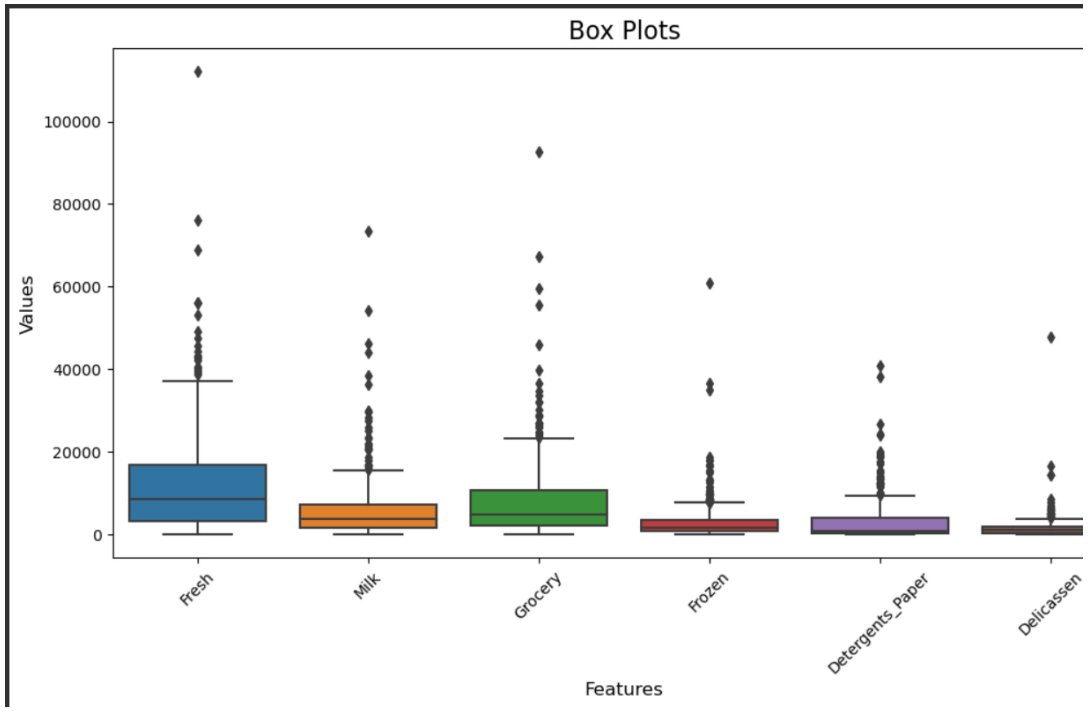
Data Visualization:Finding relationship and distribution between features with pairplots and histograms

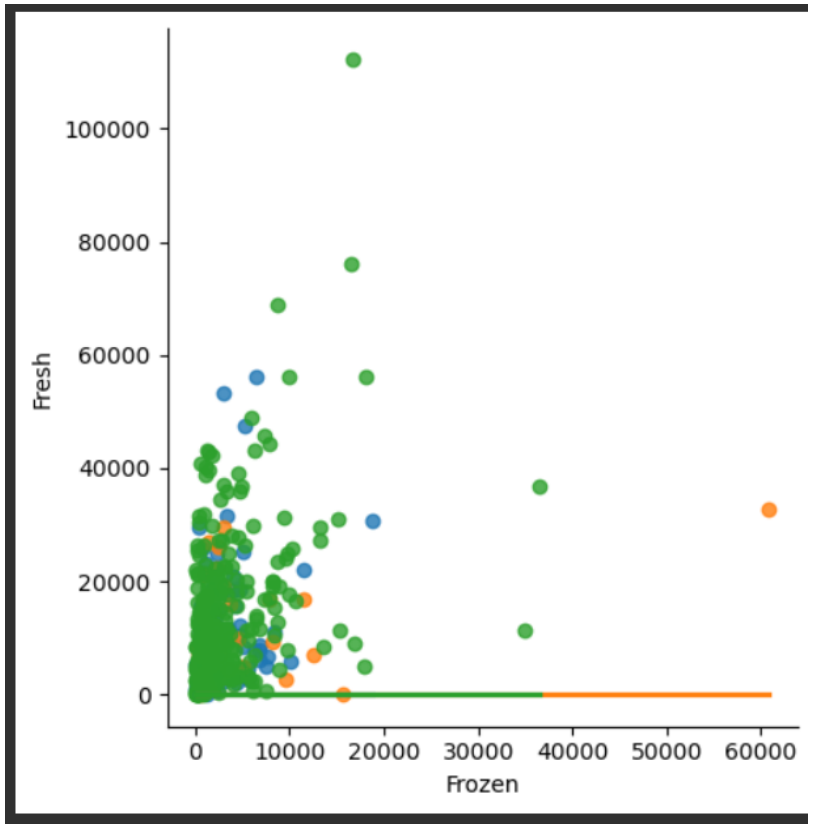


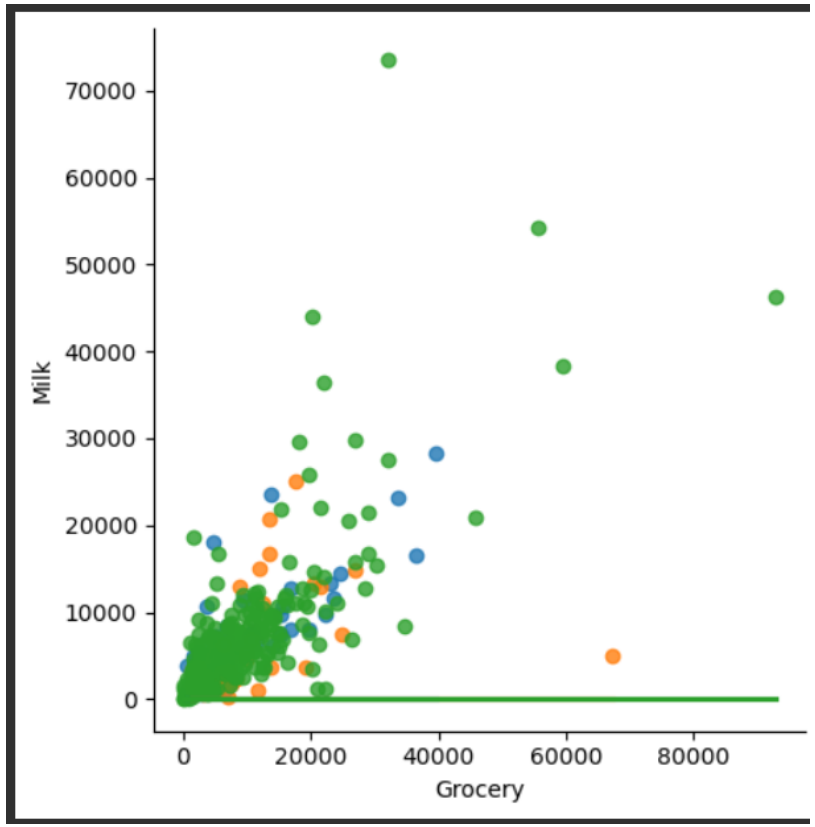


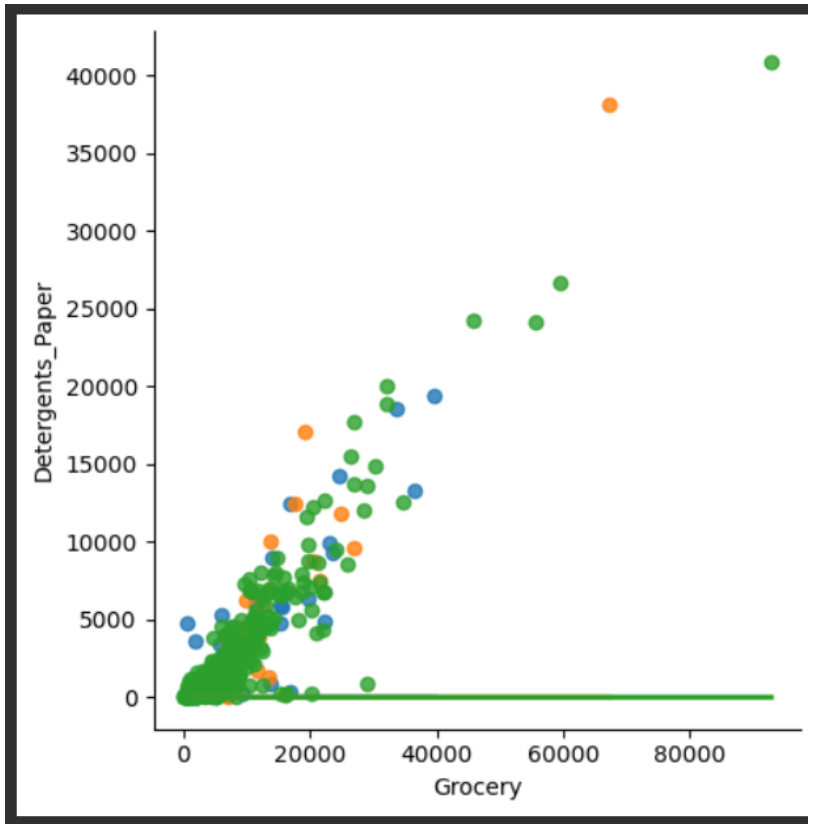
Data Visualization: Outlier detection and correlation matrix (heatmap)





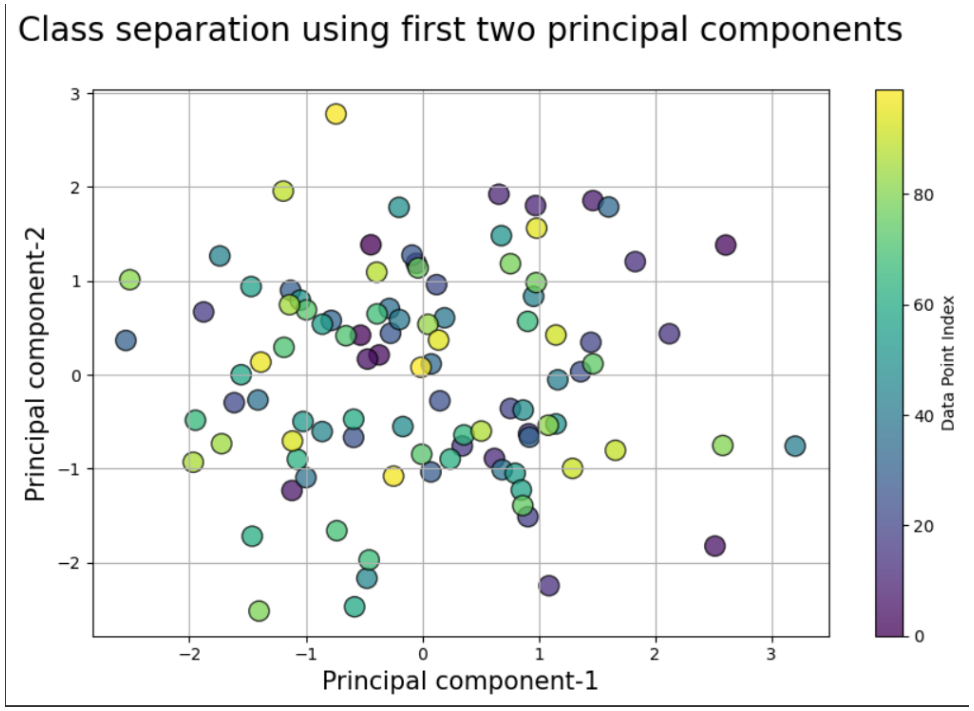


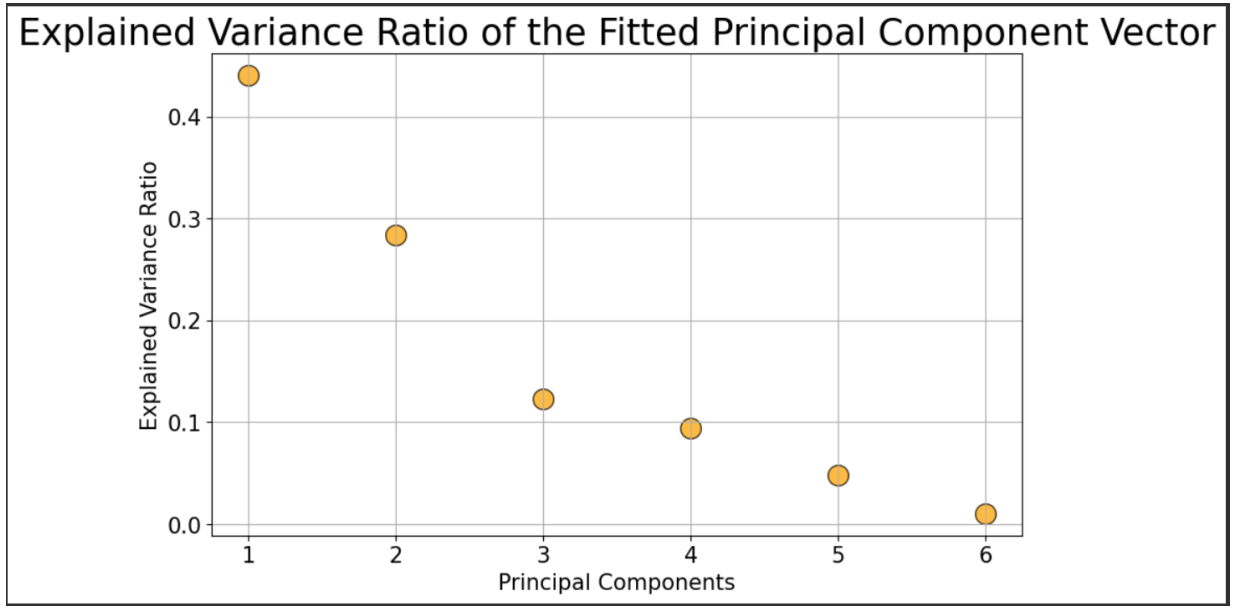




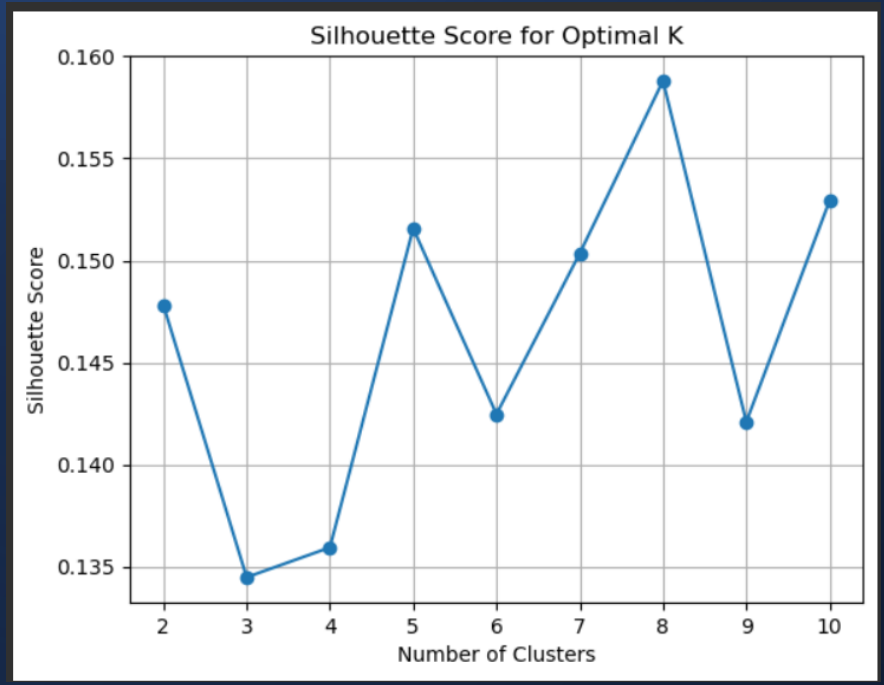
Data Visualization: Correlation Analysis

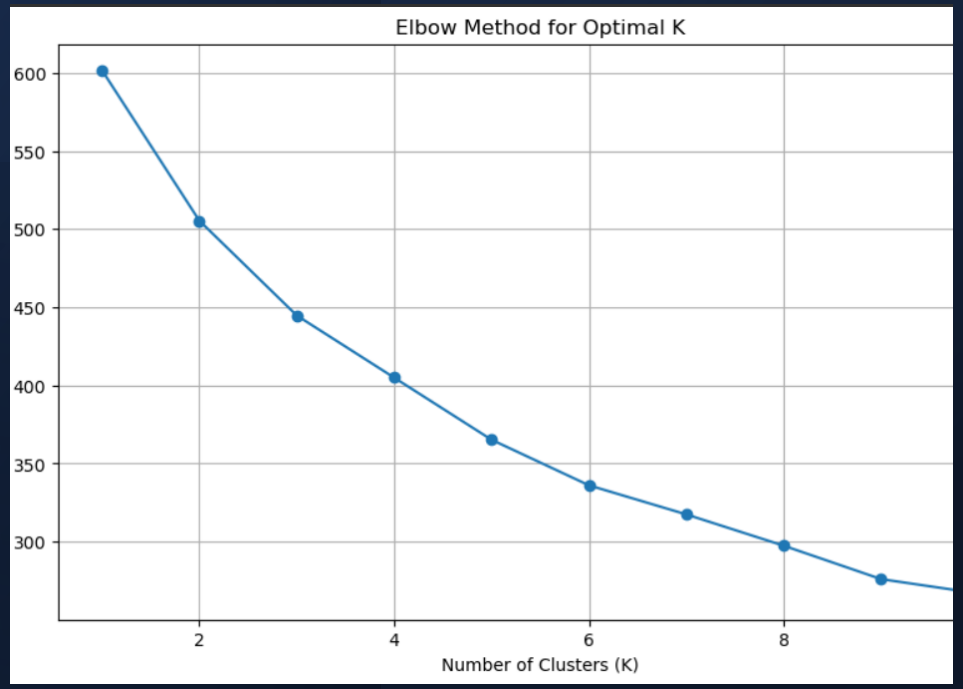
Data Transformation: Explained Variance ratio and class separation plot

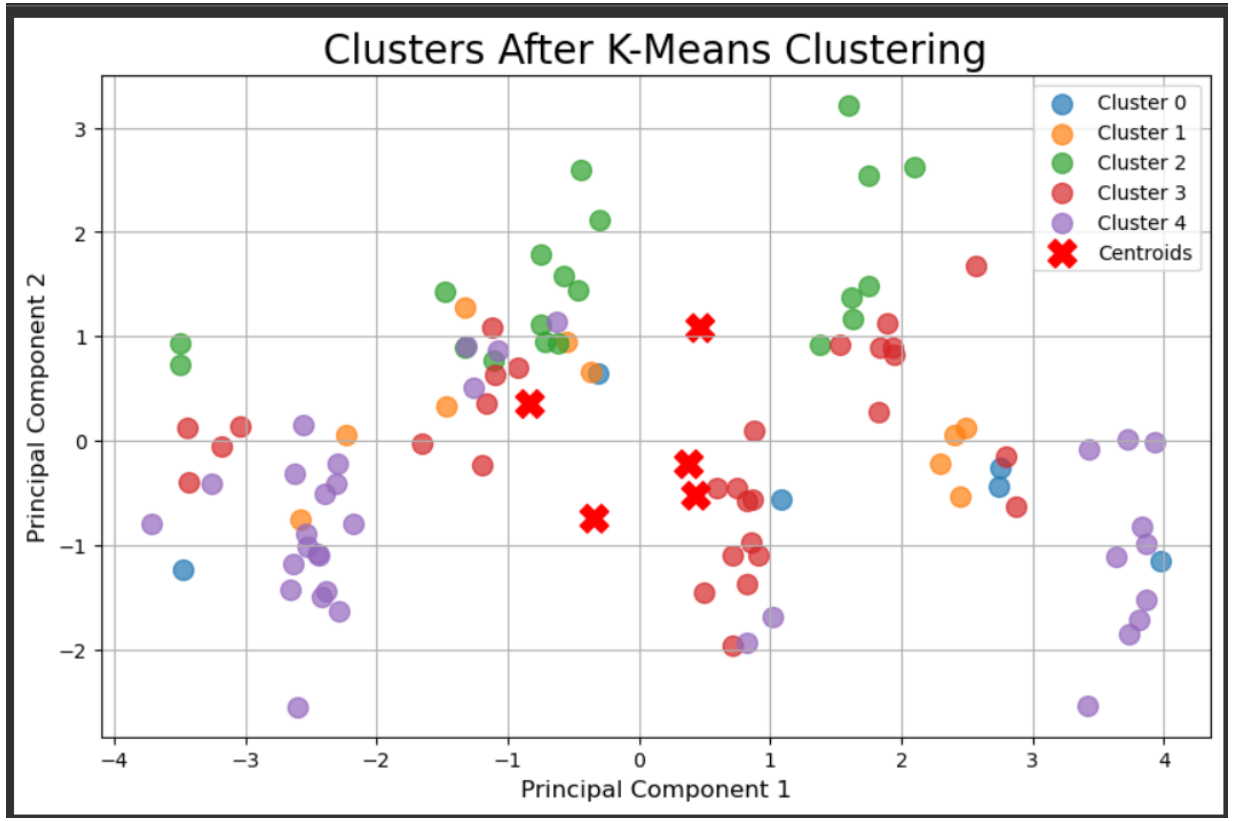




Part II - KMeans Clustering

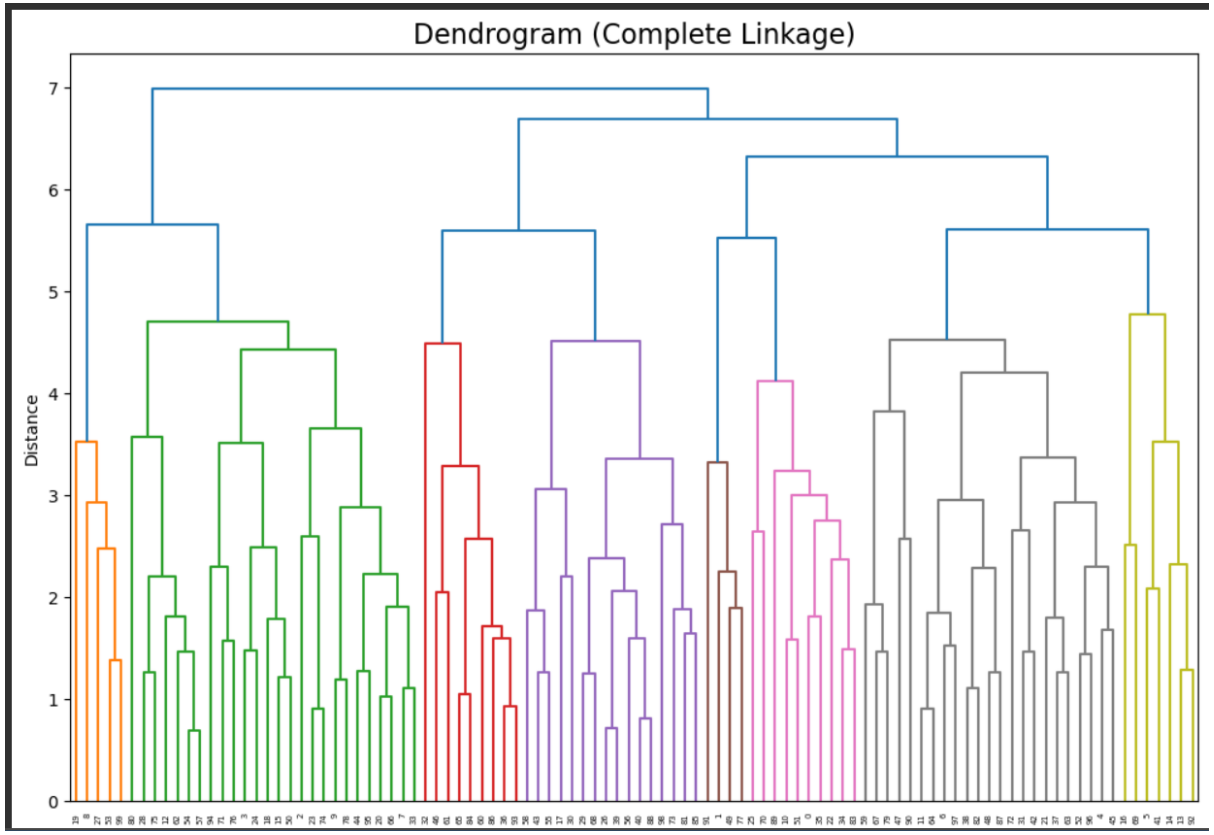


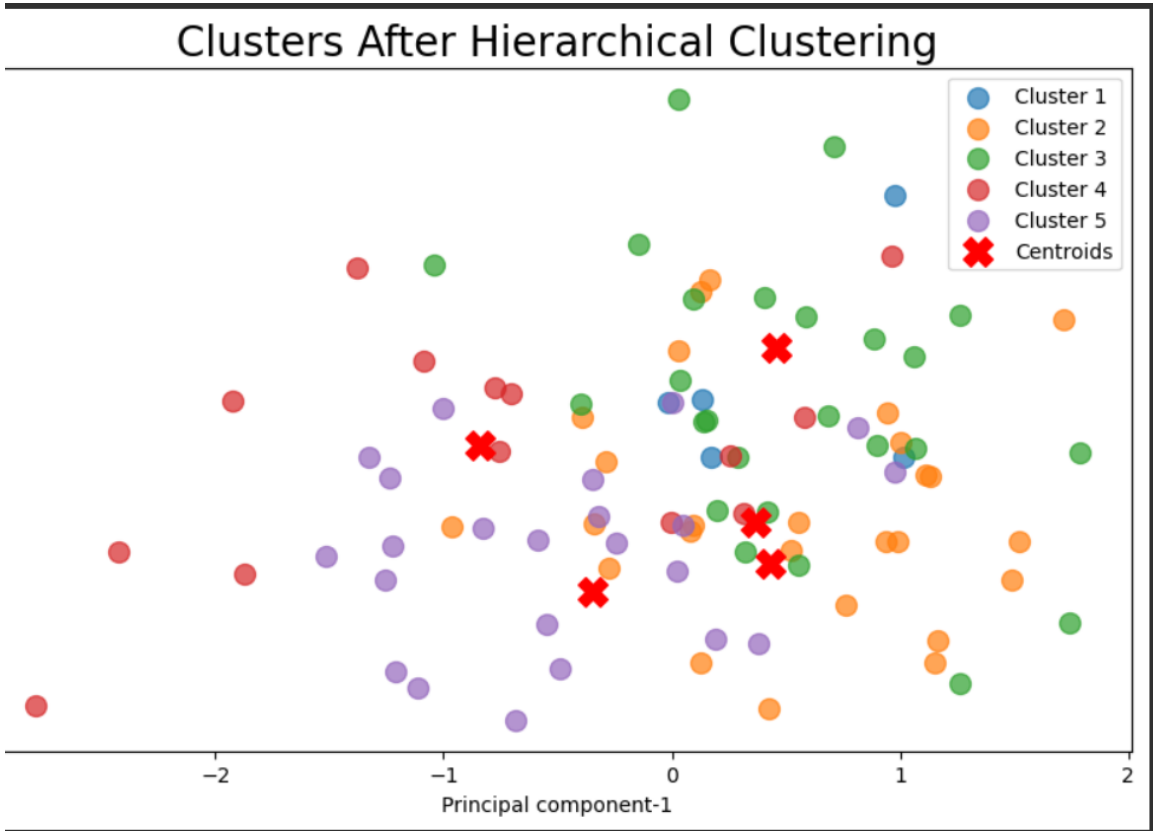




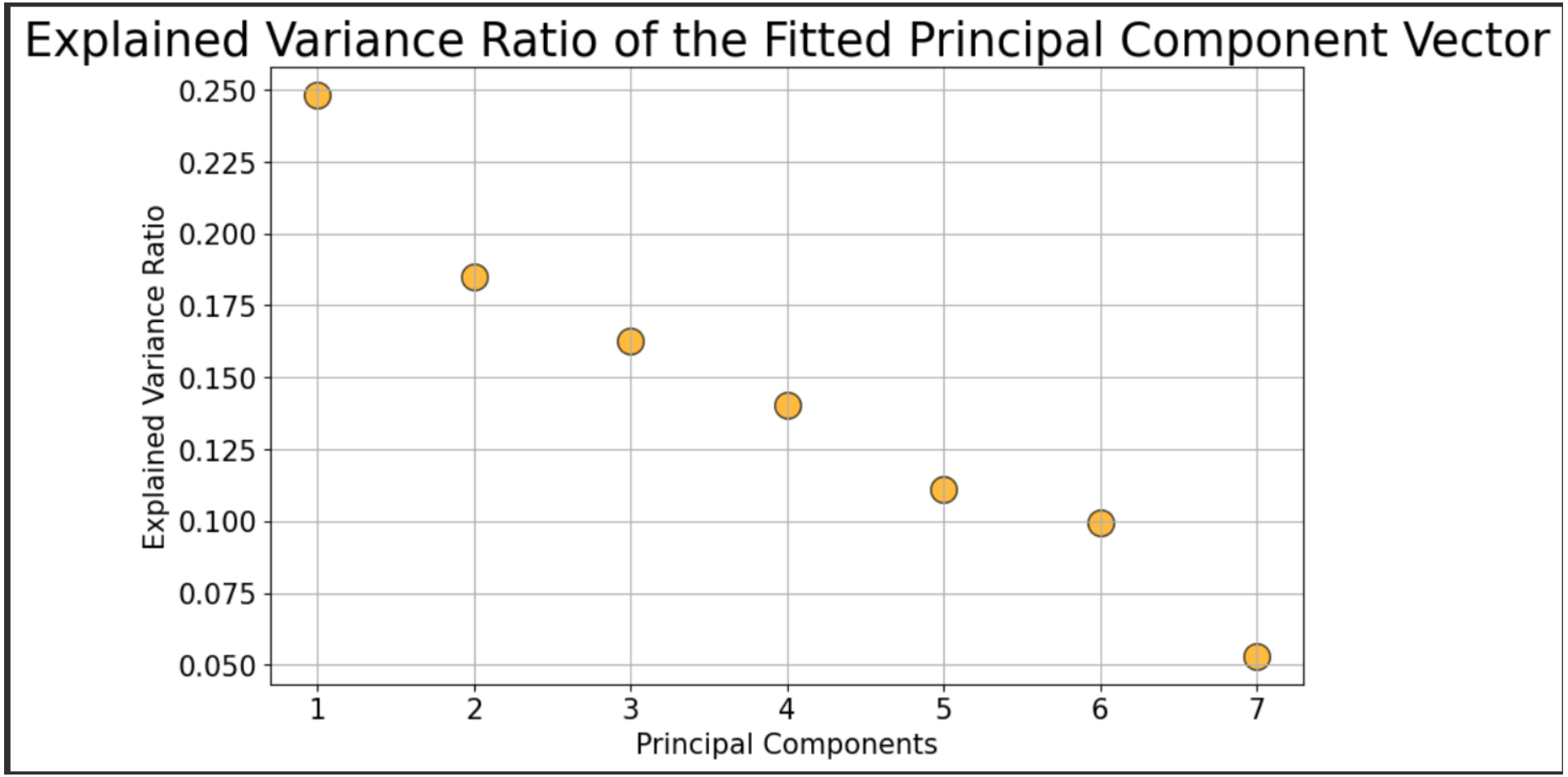
K-mean clustering

Part III - Hierarchical Clustering and Dendrogram





Part IV - PCA


Explained Variance Ratio of the Fitted Principal Component Vector
Explained Variance Ratio
0.250
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
1
2
3
4
5
6
7
Principal Components

Part V - Conclusion

- The correlation analysis between "Grocery" and "Milk“ and "Grocery" and "Detergents Paper" at 0.728 and 0.925 respectively, indicating that there is a tendency for customers to spend more on groceries when they spend more on milk, detergents paper and vice versa. However, it's essential to remember that correlation does not imply causation. The observed correlations do not necessarily mean that one variable causes the other; there may be other underlying factors or external influences influencing the relationship between the variables.
- The elbow method did not yield a clear optimal number of clusters.
- The silhouette score from K-means and the hierarchical clustering both suggested different cluster numbers (8 and 6, respectively). This discrepancy may indicate that the data is not perfectly clustered, and the choice of the number of clusters could depend on the business's specific needs or domain knowledge.
- The PCA analysis provides essential information about the features' contribution to the data variance, aiding in better understanding customer behavior and needs.