Unsupervised Learning with Wholesale Data

July 14, 2023 Jennifer Zhao

Objective

Our project aims to identify the crucial features or variables that contribute significantly to the variance in a grocery sales dataset.

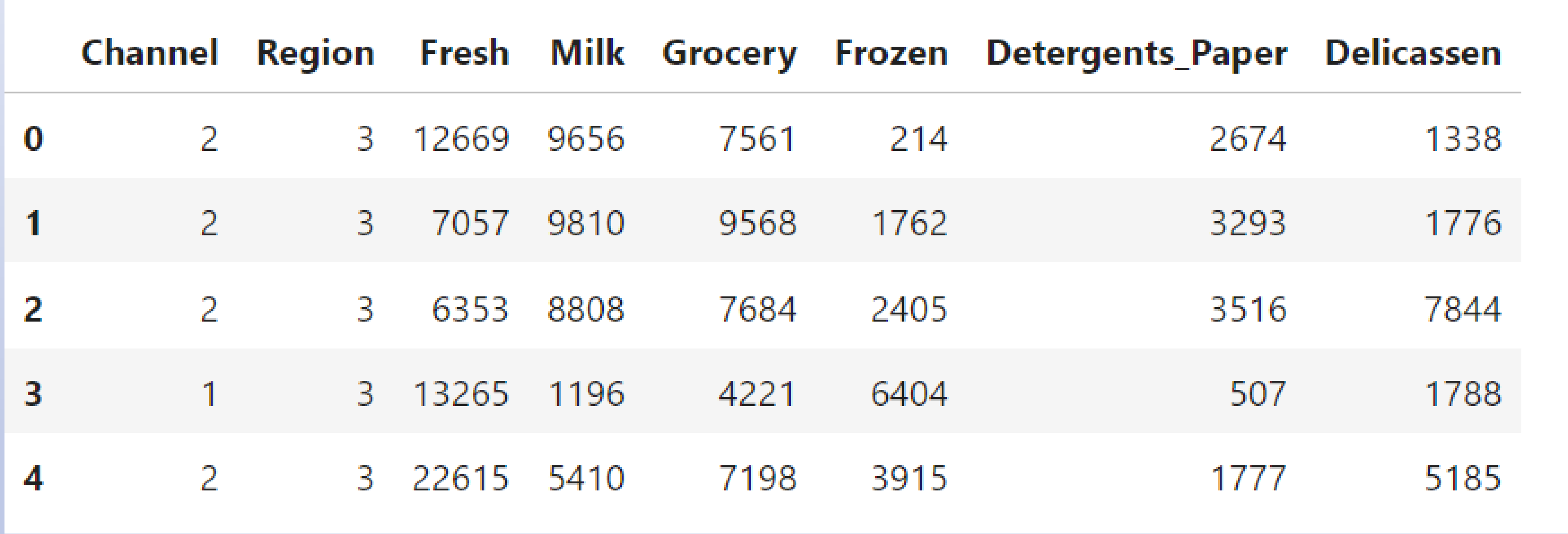
By analyzing this information, we will gain insights into the key factors influencing sales patterns and customer preferences. These findings will enable informed decision-making regarding product selection and inventory management, leading to enhance customer satisfaction.

Main tasks

- Data Import
- Data Cleaning
- Data Visualization
- Outlier Detection
- Correlation Analysis
- Data Transformation
- Feature Selection

Exploring the dataset

The dataset:



	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

Features

Channel: Horeca (Hotel/Restaurant/Cafe or Retail channel (Nominal)

Regions: Lisbon, Oporto, or Other (Nominal)

Fresh: annual spending (m.u.) on fresh products (Continuous)

Milk: annual spending (m.u.) on milk products (Continuous)

Grocery: annual spending (m.u.)on grocery products (Continuous)

Frozen: annual spending (m.u.)on frozen products (Continuous)

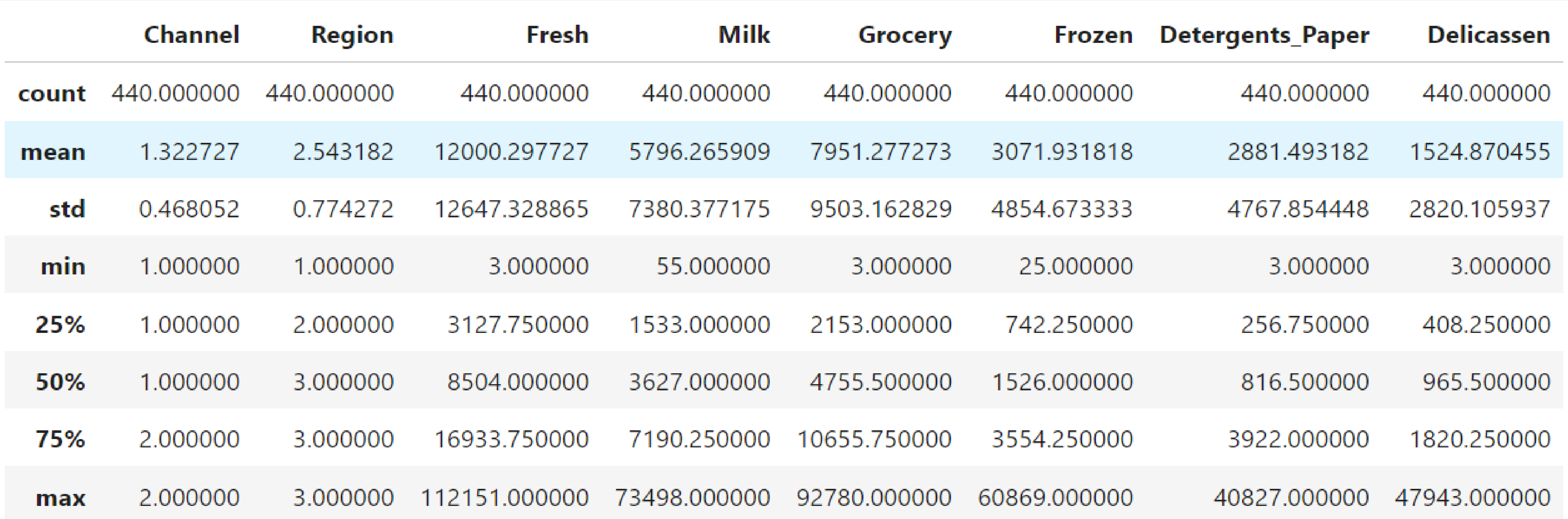
Detergents_Paper: annual spending (m.u.) on detergents and paper products (Continuous)

Delicassen: annual spending (m.u.)on and delicatessen products (Continuous)

Data cleaning and analysis

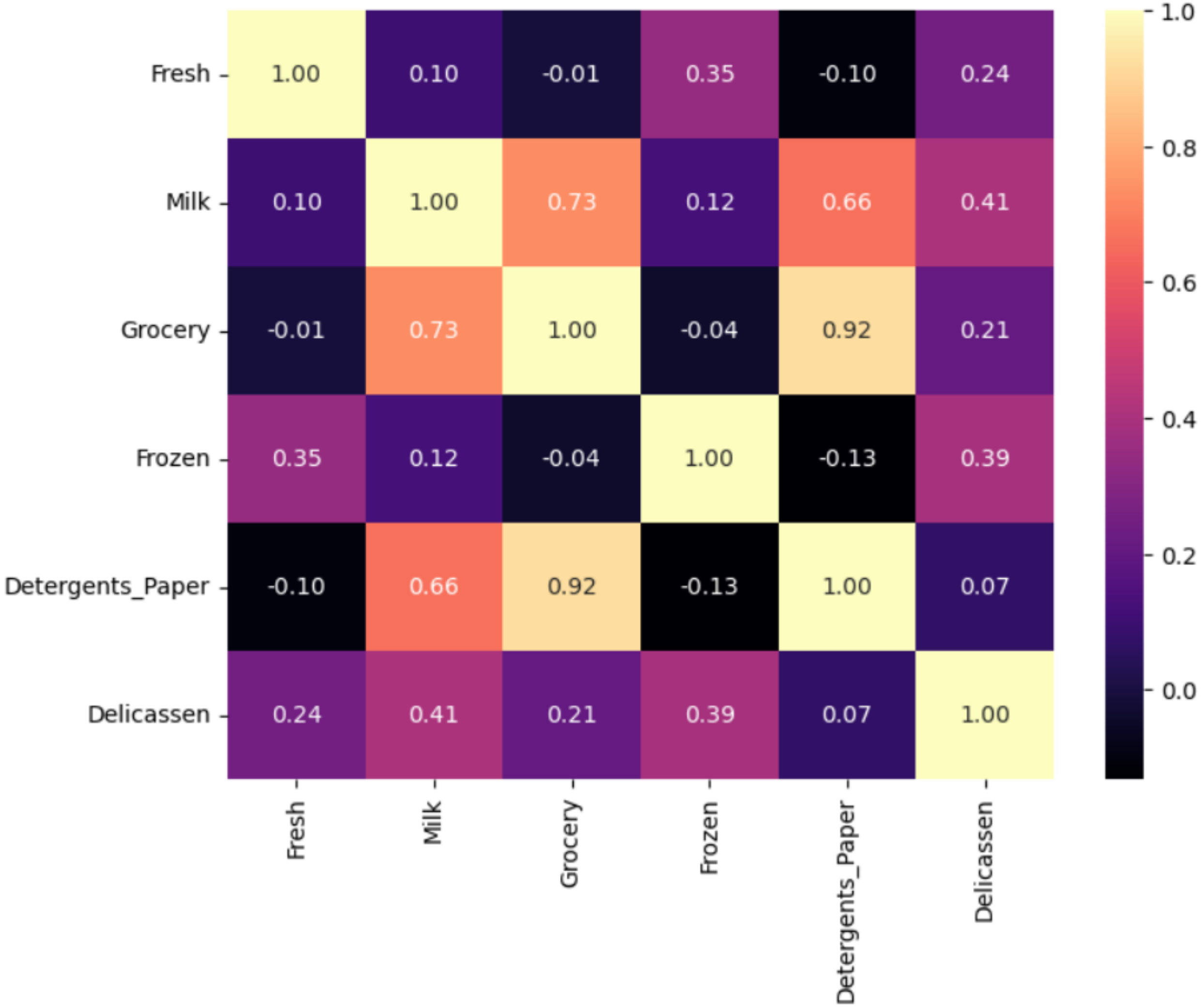
1. Column Analysis:
    - Check column meaning and type.
    - Verify dataset shape, min, and max values.
2. Duplicates Check:
    - Ensure no duplicates exist in the dataset.
3. Checking Missing and Zero Values.
4. Outlier Handling:
    - Use the 1.5 times Interquartile Range (IQR) to identify outliers.
    - Remove outliers that are the same for multiple features.
5. Scaling data using the log function.

Description the dataset



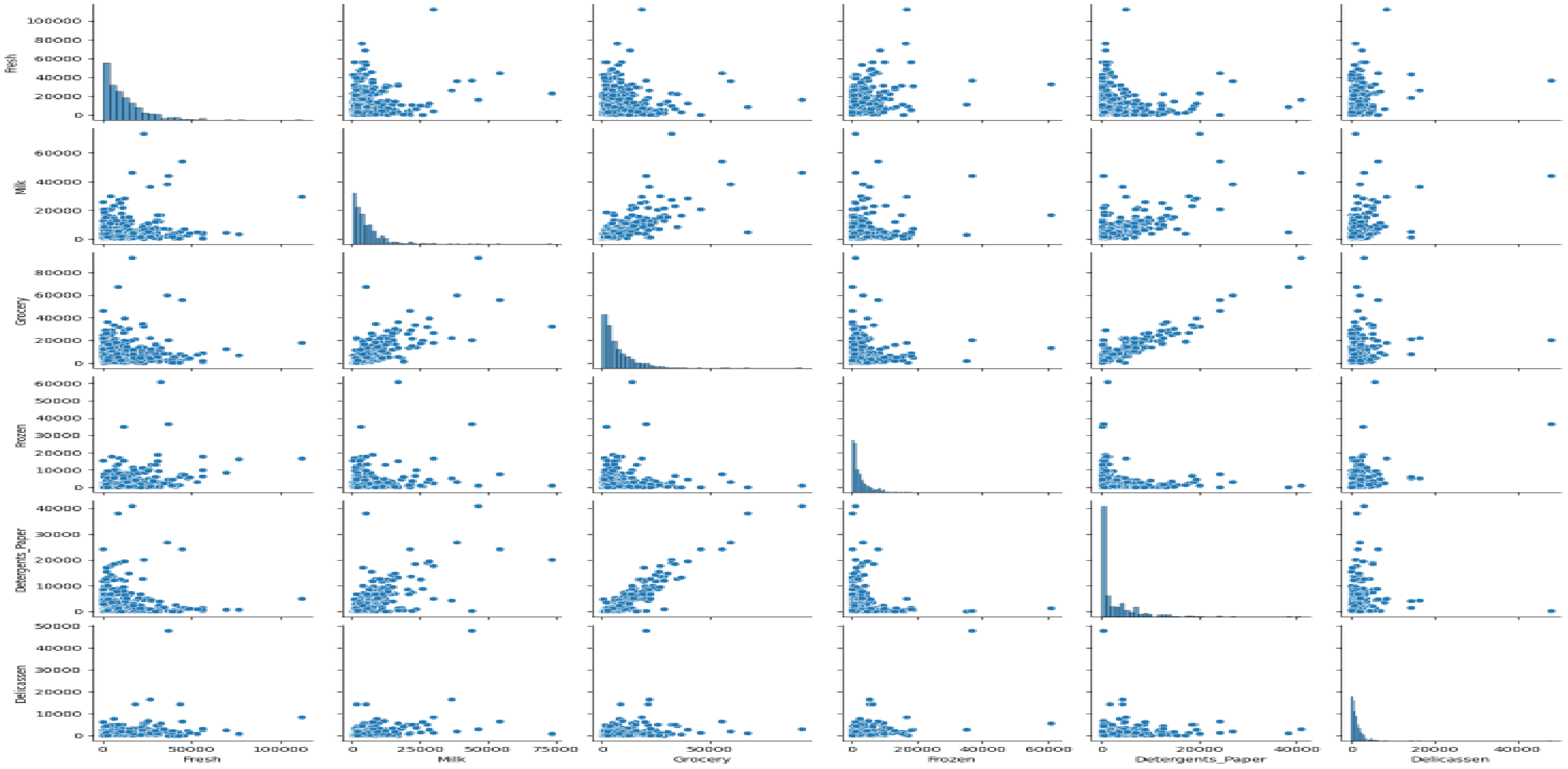
	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	1.322727	2.543182	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	0.468052	0.774272	12647.328865	7380.377175	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

Correlation Matrix



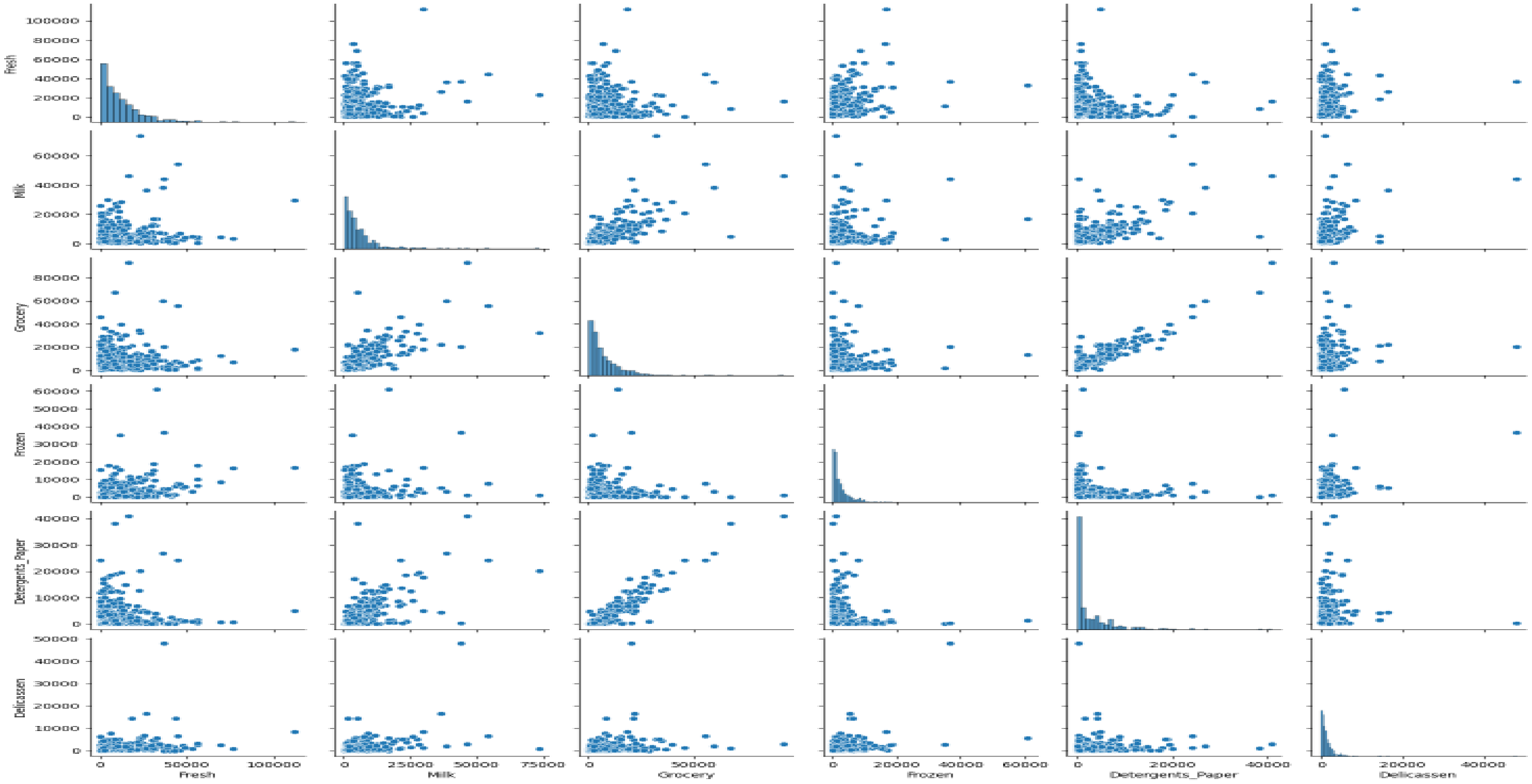
	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Fresh	1.00	0.10	-0.01	0.35	-0.10	0.24
Milk	0.10	1.00	0.73	0.12	0.66	0.41
Grocery	-0.01	0.73	1.00	-0.04	0.92	0.21
Frozen	0.35	0.12	-0.04	1.00	-0.13	0.39
Detergents_Paper	-0.10	0.66	0.92	-0.13	1.00	0.07
Delicassen	0.24	0.41	0.21	0.39	0.07	1.00

Pairplot of feasures

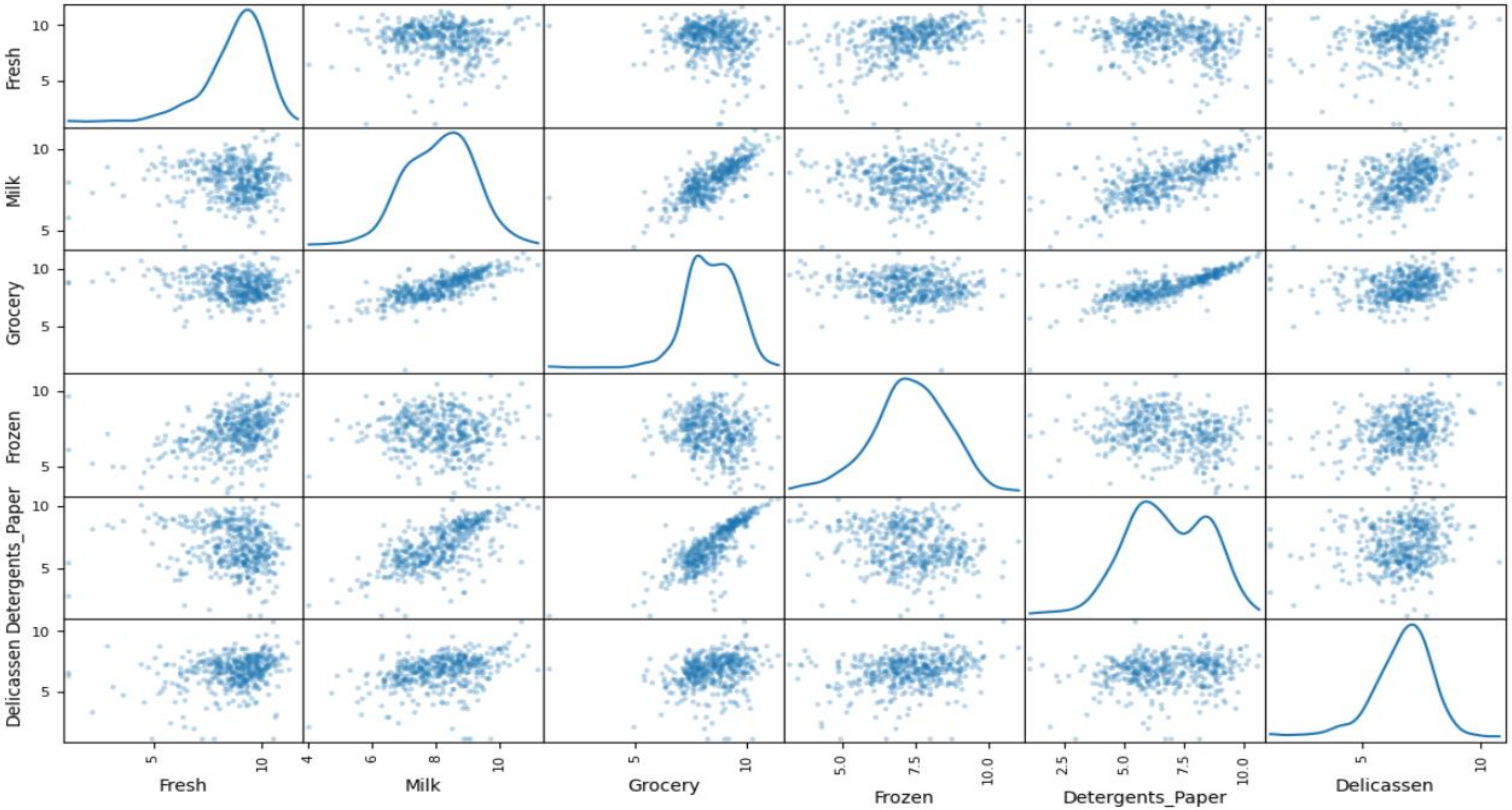


The distribution of all the features appears to be right-skewed.

Pairplot of feasures (after scaling the data)

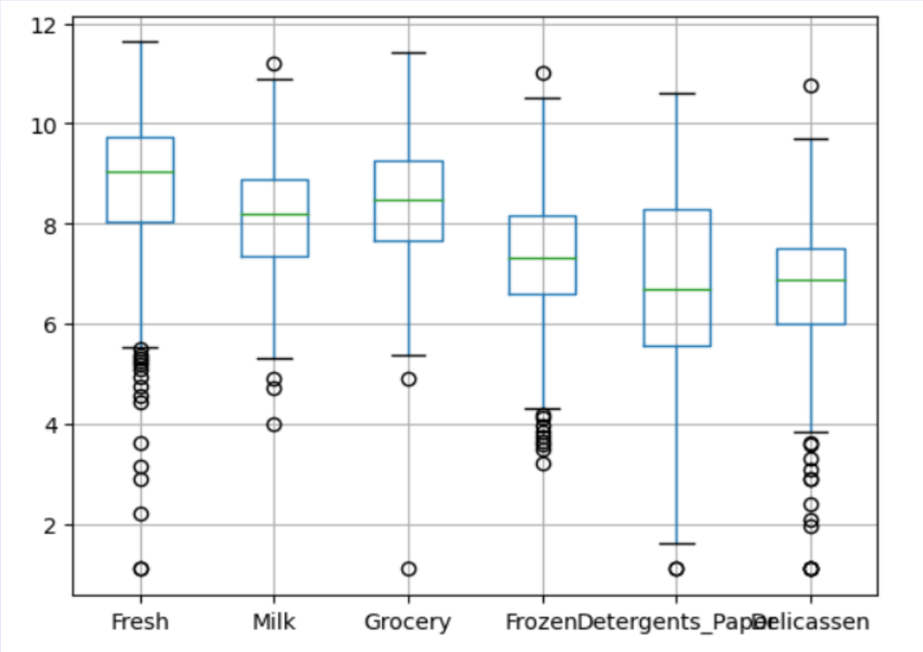


scatter matrix



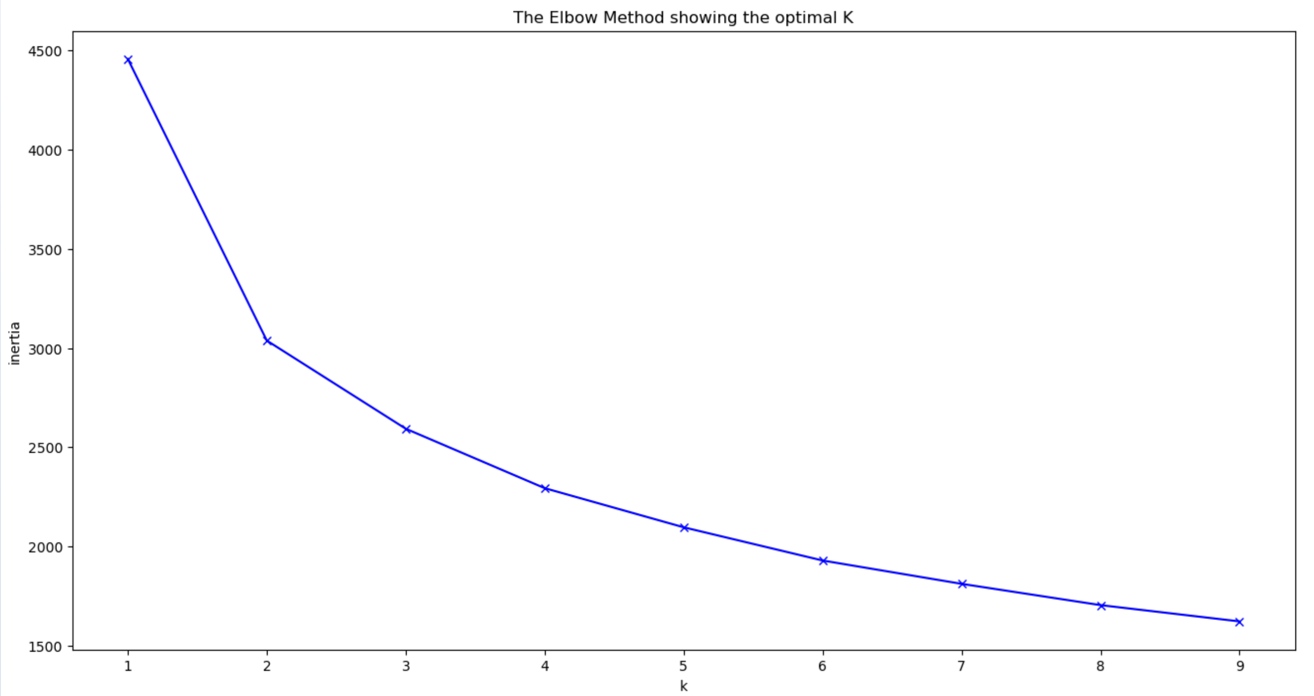
Identify the outliers

Calculate the Interquartile Range (IQR) and apply a threshold of 1.5 times the IQR. Remove outliers that exceed this threshold across multiple features.



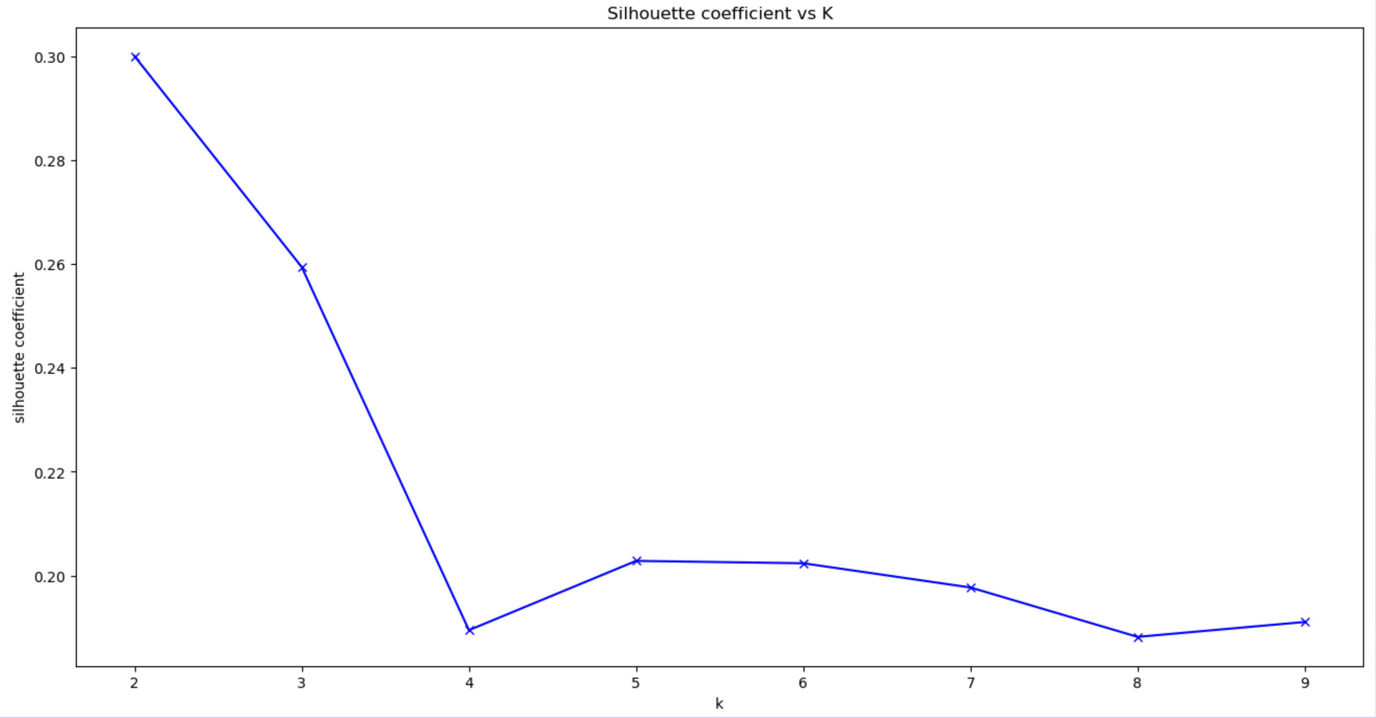
KMeans Clustering

Perform the k_mean clustering analysis, plot the elbow method to find 2 is the optimal k number.

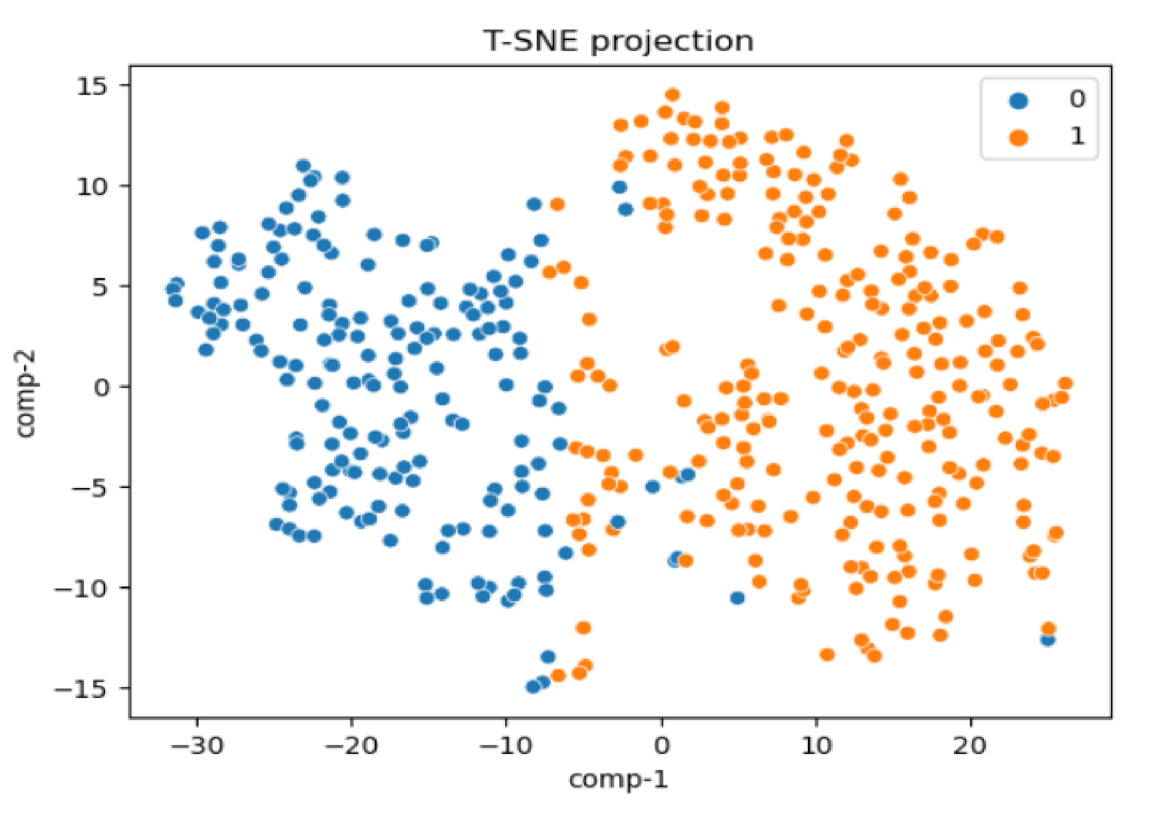


KMeans Clustering

Perform the k_mean clustering analysis, and plot silhouette. From both elbow point and silhouette coefficient, we can identify 2 clusters has the best score.

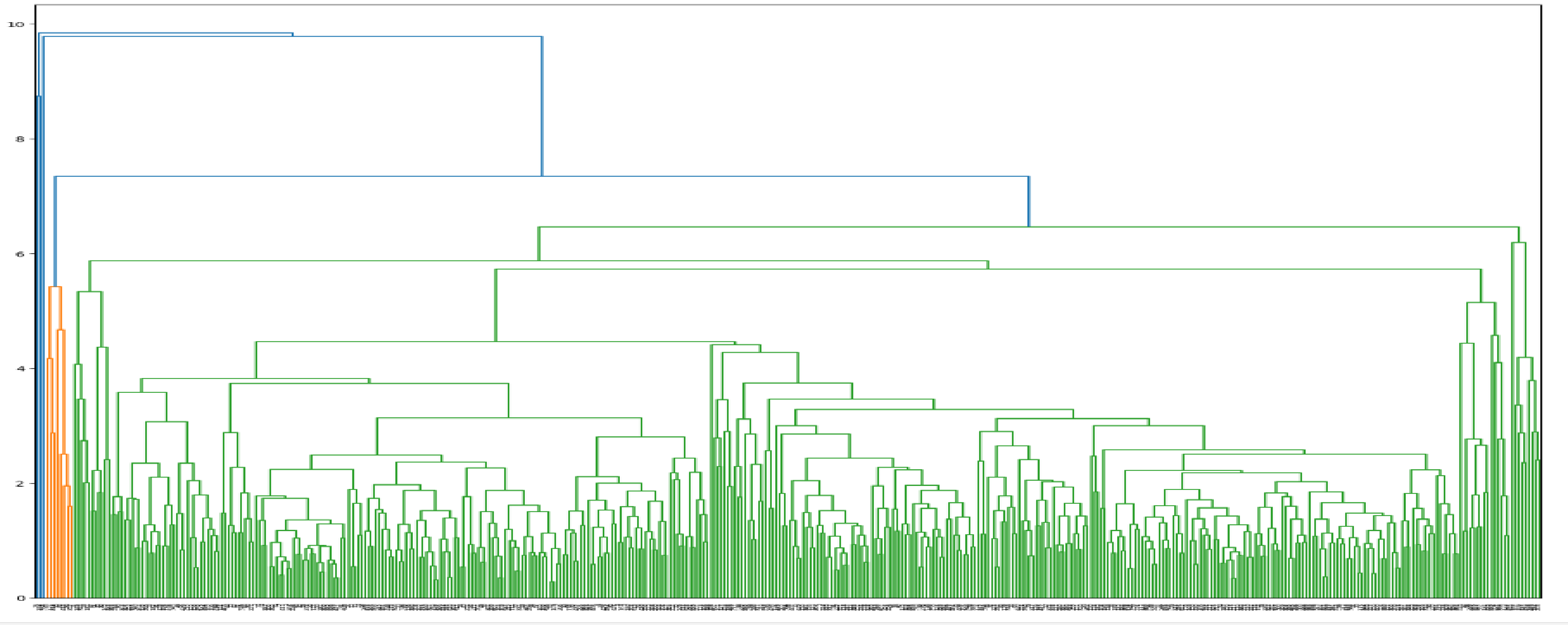


T-SNE projection



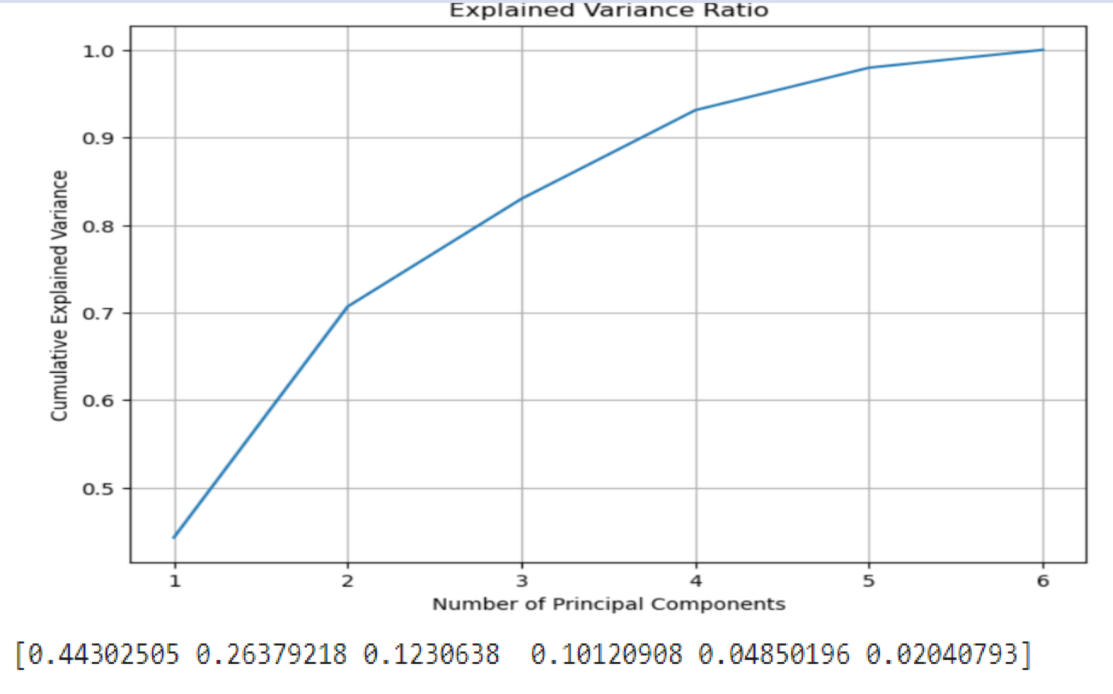
Hierarchical Clustering

Perform Hierarchical Clustering Analysis using Single, Complete, and Average Linkage Methods, and Select the Average Score as the Criterion, and 2 clusters has the best score.



PCA

Perform principal component analysis (PCA) and found 4 combinations of features best describe customers. (explained_variance_ratio > 10%)



```
[0.44302505 0.26379218 0.1230638  0.10120908 0.04850196 0.02040793]
```

Conclusion

After exploring data by using PCA, Hierarchical Clustering, and K-Means for clustering. We can get the findings below:

1. PCA: the data dimensions can be reduced to 4, The first principal component explains approximately 44.3% of the variance, followed by the second component with 26.4%, the third component with 12.3%, and the fourth component with 10.1%. These components capture the most significant patterns or trends in the data.

2. Hierarchical Clustering: the results of hierarchical clustering indicate that using 2 clusters is the best option and the single and average linkage methods achieved higher silhouette scores.

3. K-mean Clustering: the best-performing clustering result is obtained with 2 clusters, supporting the effectiveness of a 2-cluster solution.

Thank you!