Wholesale Data

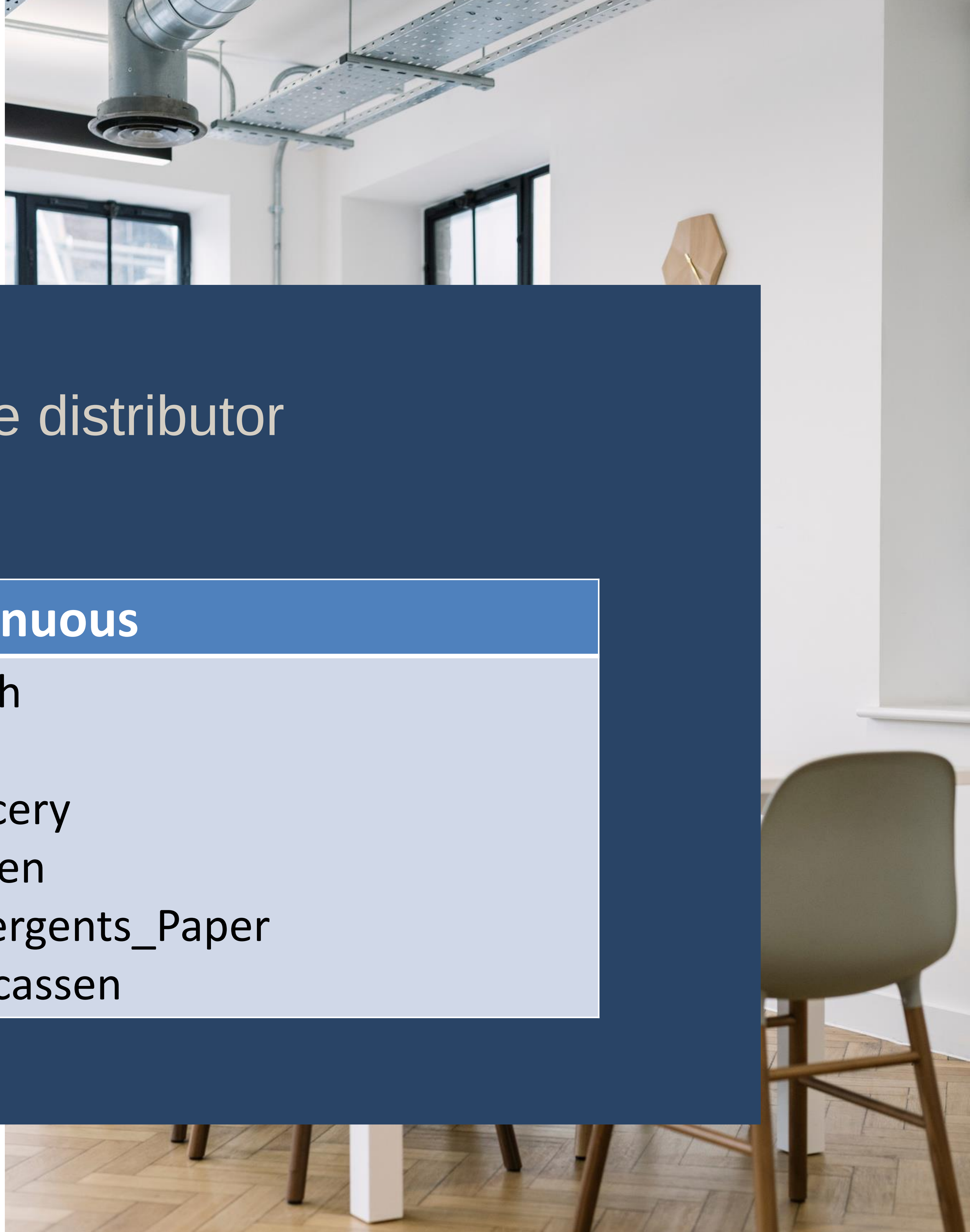
Unsupervised Learning

By Vidusha Wilpita

Contents

- Introduction
- Flow Structure
- Results
- Conclusion
- Challenges
- Future Goals

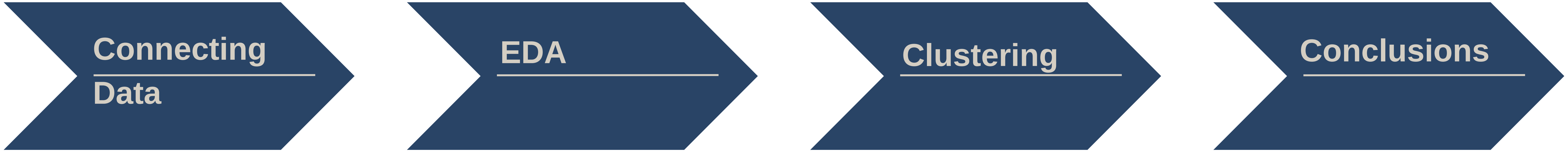
Introduction



- Data about 440 clients of a wholesale distributor (Numerical Data only)

Discrete	Continuous
• Channel • Region	• Fresh • Milk • Grocery • Frozen • Detergents_Paper • Delicassen

Flow Structure



Connecting Data

EDA

Clustering

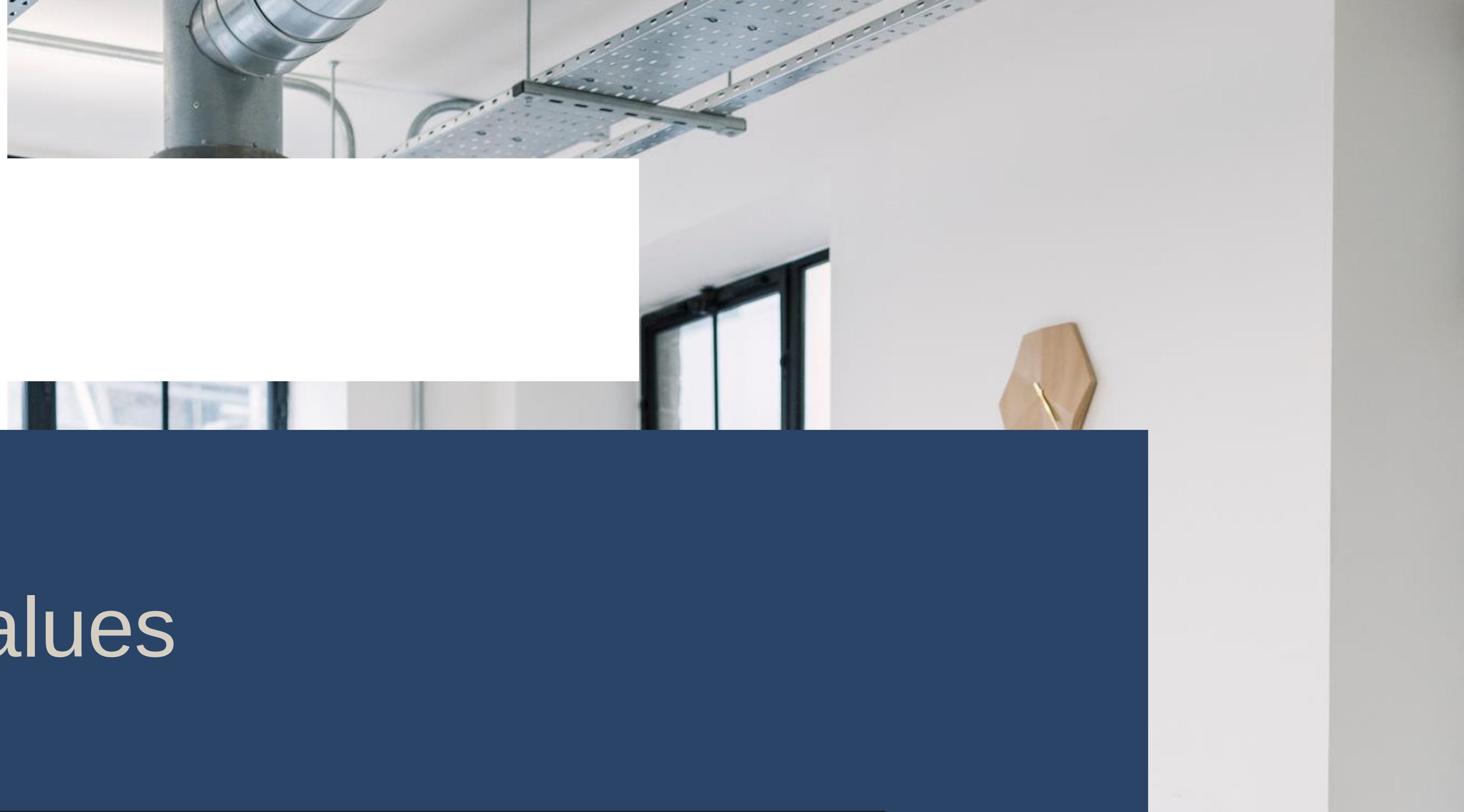
Conclusions

Connected Wholesale customers data.csv to analyze the data

Created different visualizations to learn more about the data, Done data cleaning and preprocessing

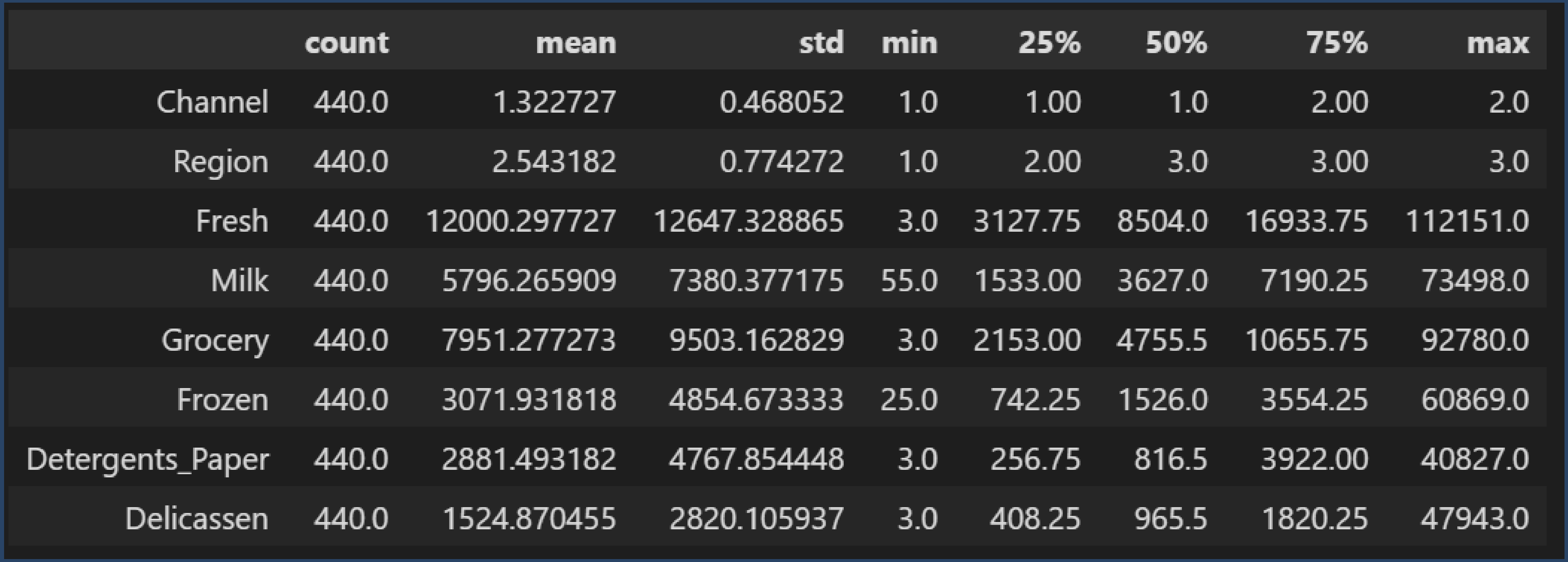
K-means clustering, Hierarchical Clustering and PCA

Coming up with conclusions using EDA, and unsupervised learning techniques



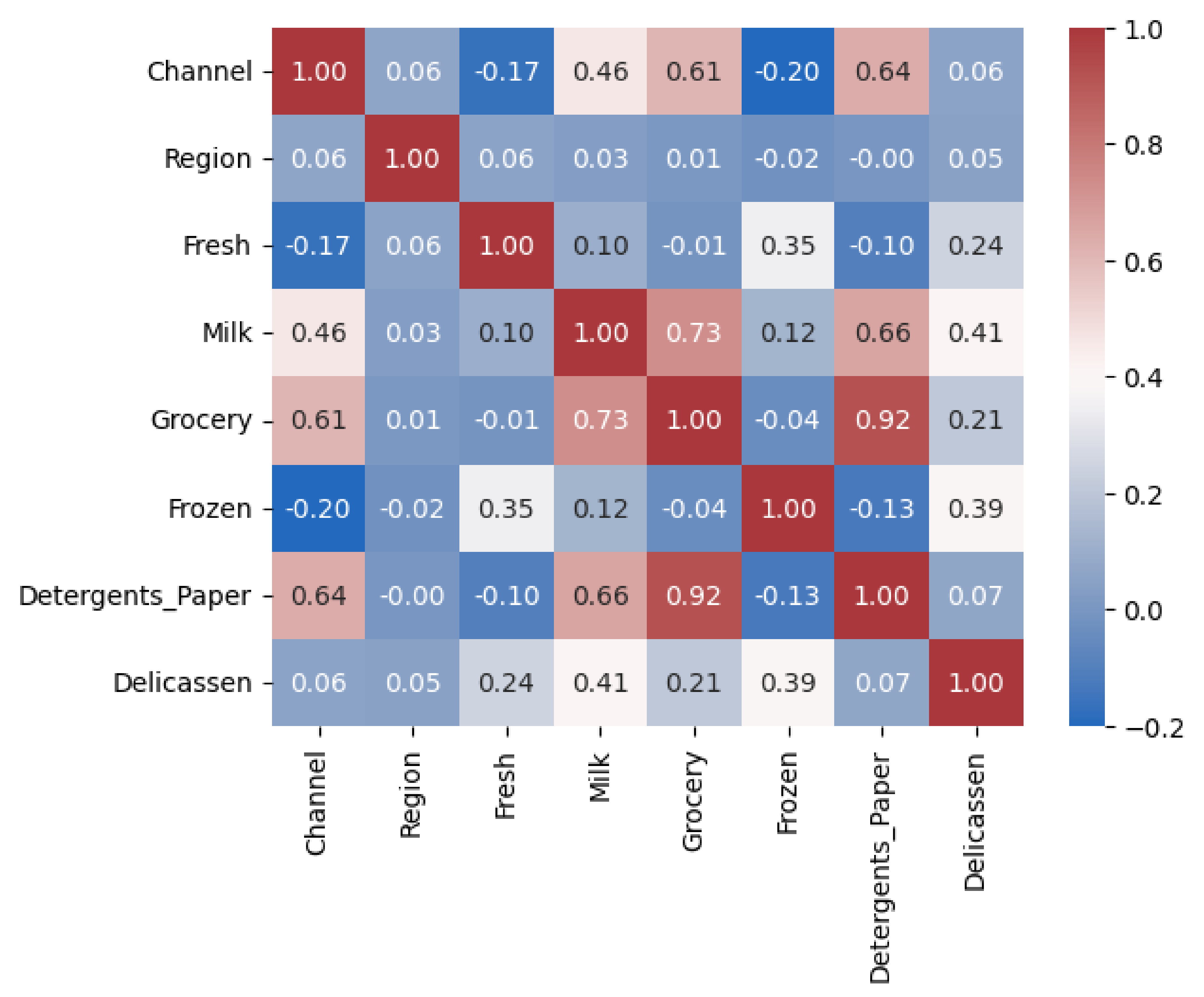
EDA

- Dataset does not contain any null values
- Dataset Description:



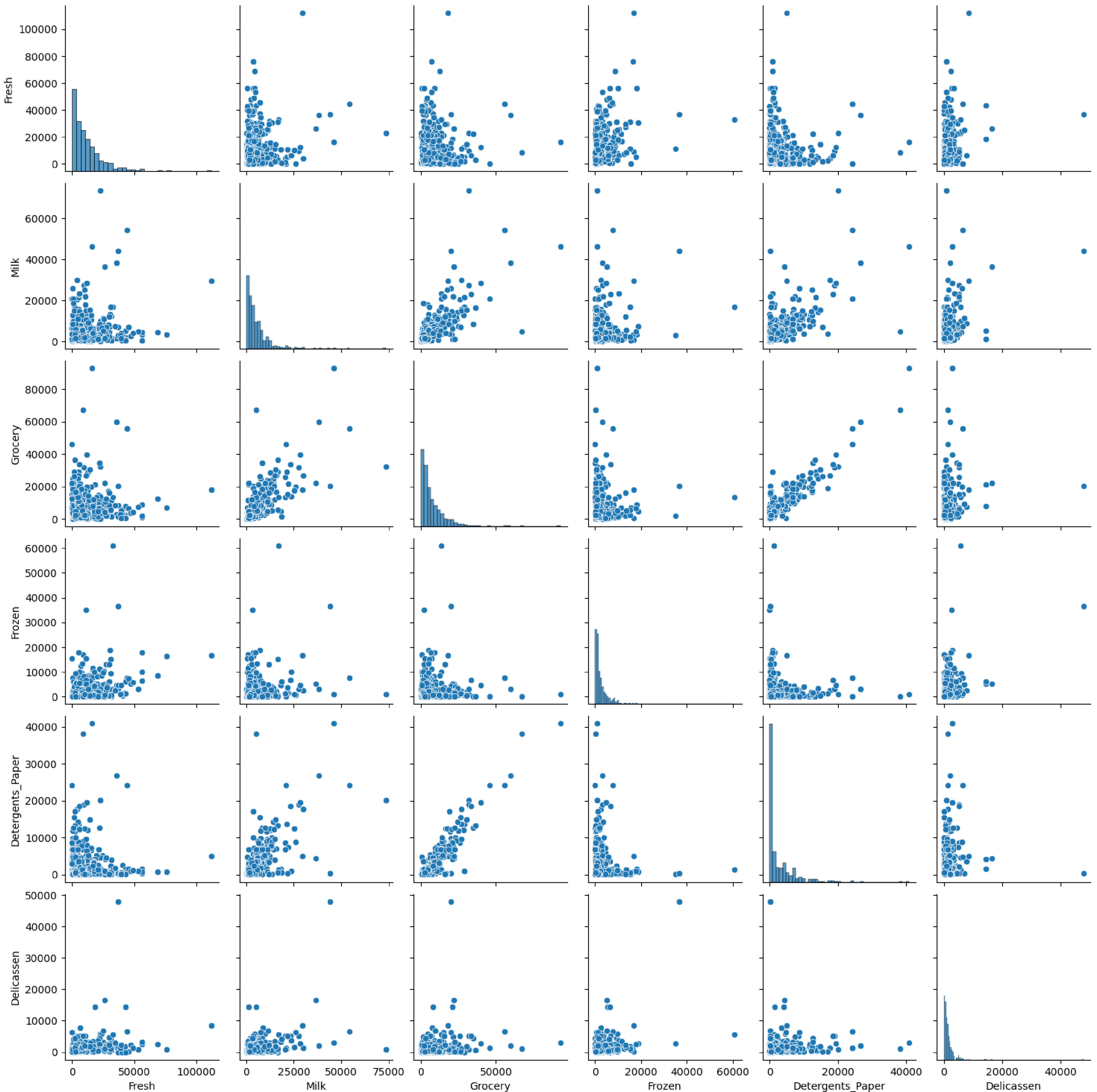
	count	mean	std	min	25%	50%	75%	max
Channel	440.0	1.322727	0.468052	1.0	1.00	1.0	2.00	2.0
Region	440.0	2.543182	0.774272	1.0	2.00	3.0	3.00	3.0
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	7951.277273	9503.162829	3.0	2153.00	4755.5	10655.75	92780.0
Frozen	440.0	3071.931818	4854.673333	25.0	742.25	1526.0	3554.25	60869.0
Detergents_Paper	440.0	2881.493182	4767.854448	3.0	256.75	816.5	3922.00	40827.0
Delicassen	440.0	1524.870455	2820.105937	3.0	408.25	965.5	1820.25	47943.0

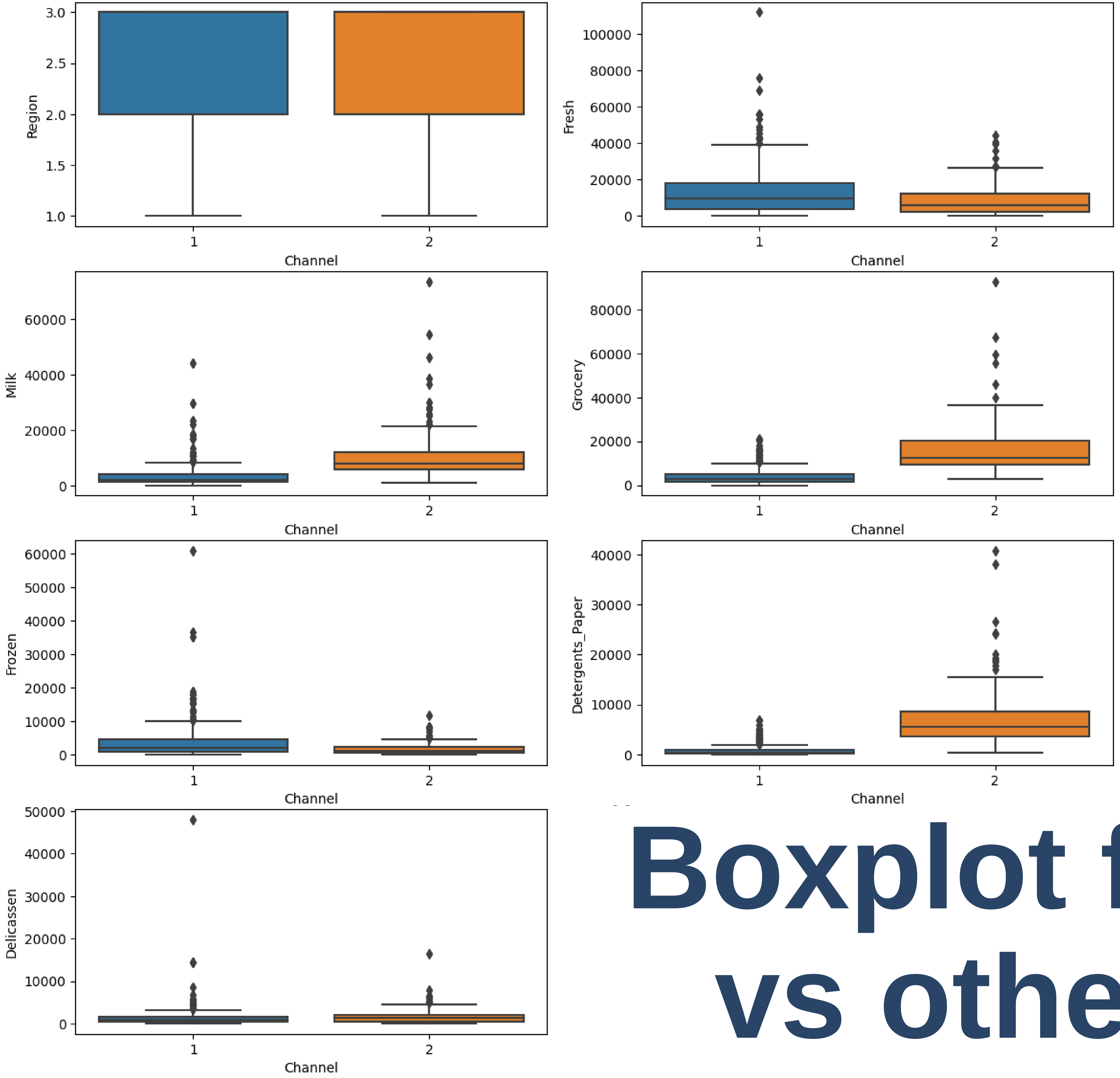
Heatmap of features



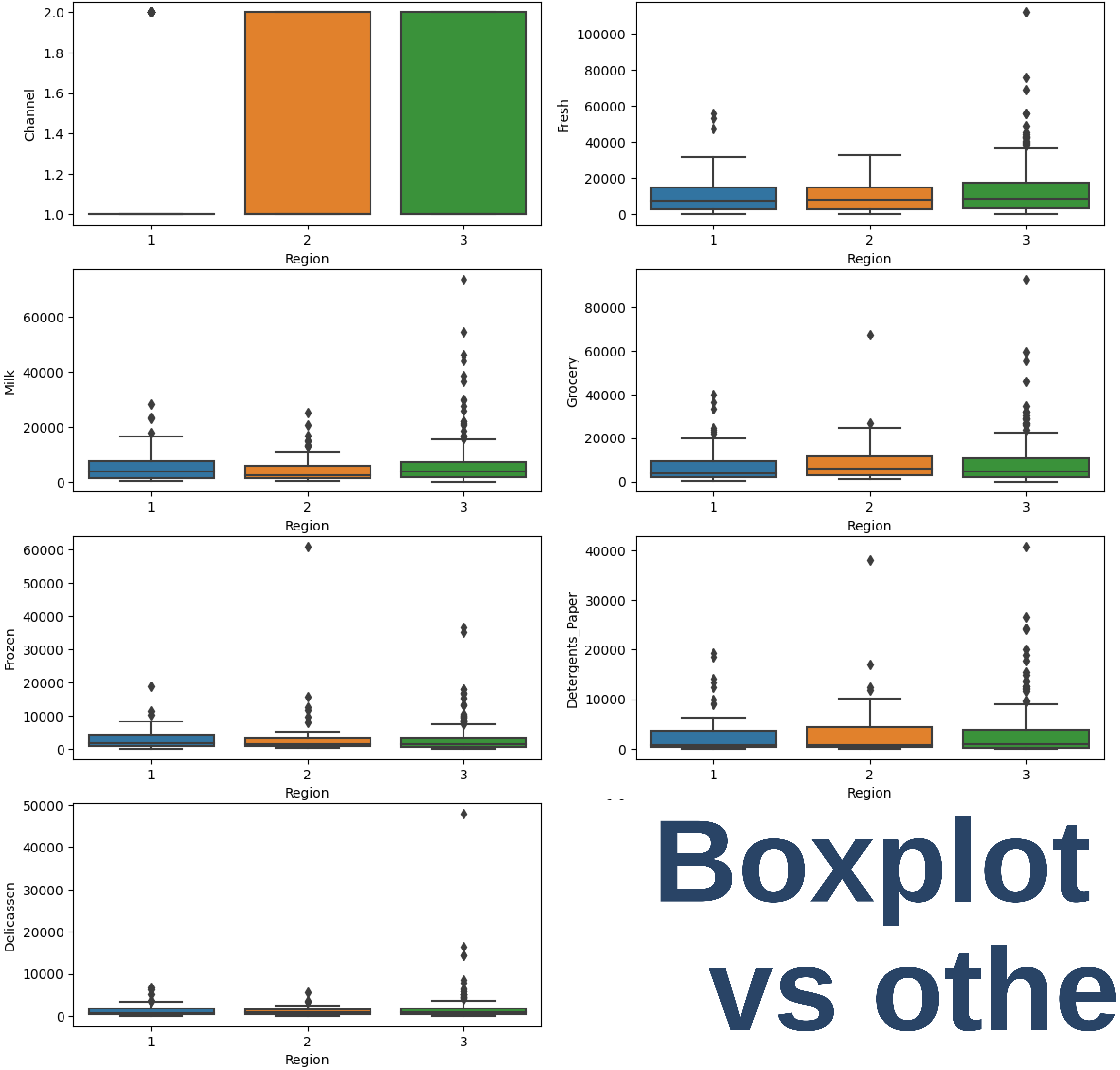
	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Channel	1.00	0.06	-0.17	0.46	0.61	-0.20	0.64	0.06
Region	0.06	1.00	0.06	0.03	0.01	-0.02	-0.00	0.05
Fresh	-0.17	0.06	1.00	0.10	-0.01	0.35	-0.10	0.24
Milk	0.46	0.03	0.10	1.00	0.73	0.12	0.66	0.41
Grocery	0.61	0.01	-0.01	0.73	1.00	-0.04	0.92	0.21
Frozen	-0.20	-0.02	0.35	0.12	-0.04	1.00	-0.13	0.39
Detergents_Paper	0.64	-0.00	-0.10	0.66	0.92	-0.13	1.00	0.07
Delicassen	0.06	0.05	0.24	0.41	0.21	0.39	0.07	1.00

Pairplot for continuous features





Boxplot for Channel vs other features

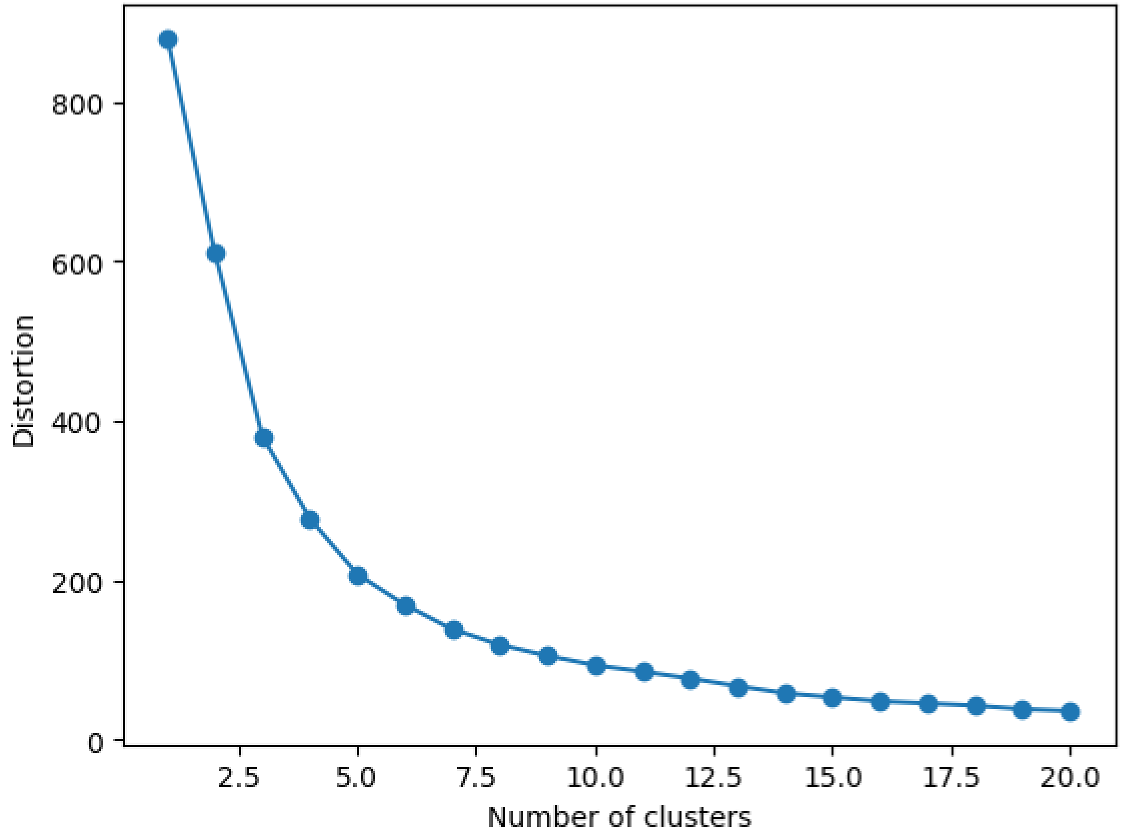


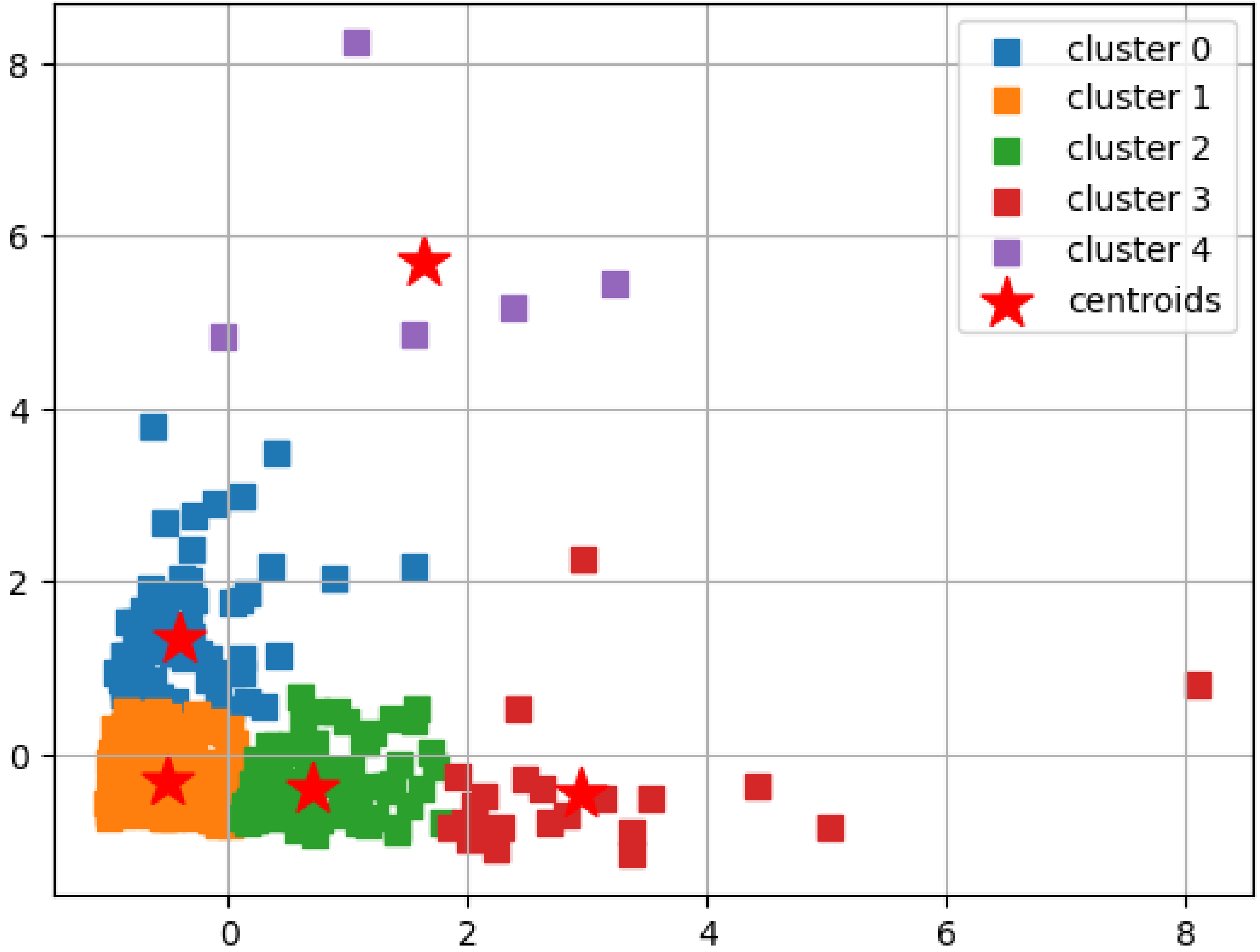
Boxplot for Region vs other features

Results/ Conclusions

Kmeans Clustering

Elbow Rule

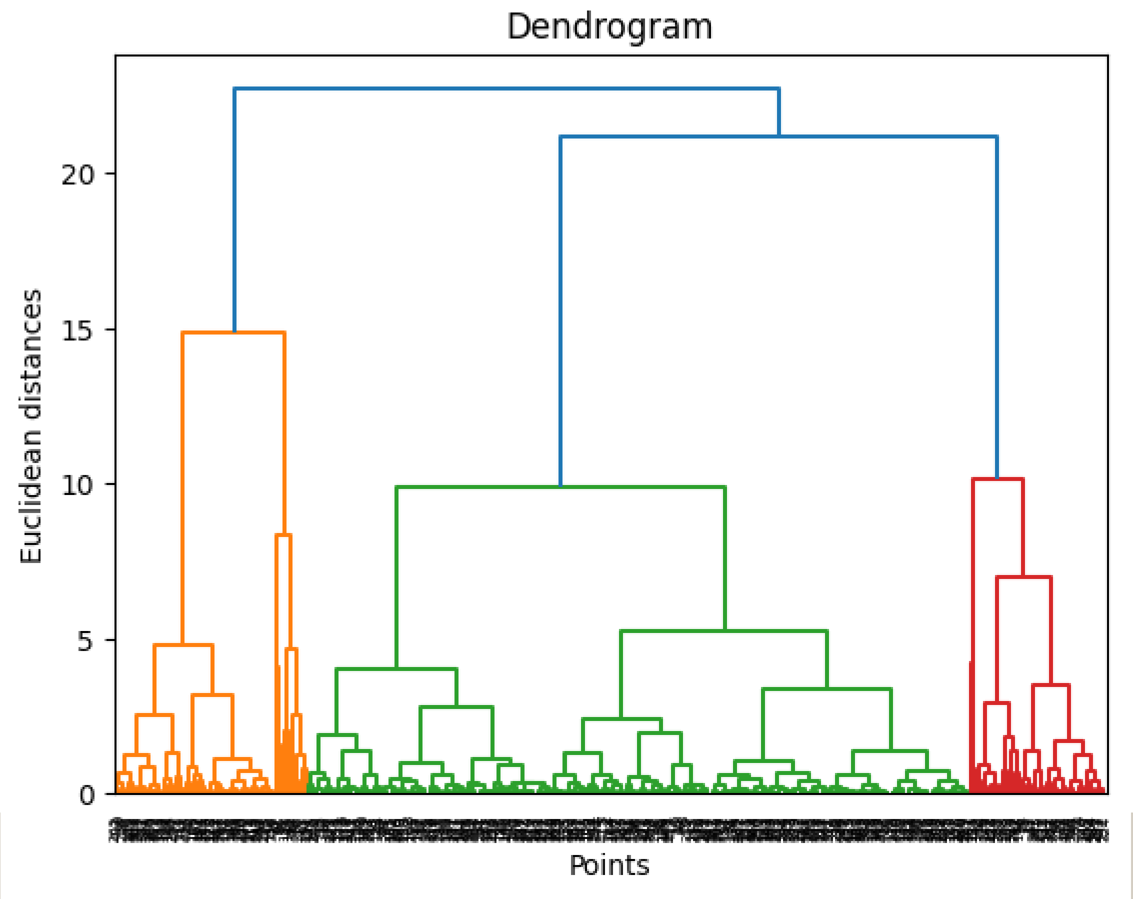


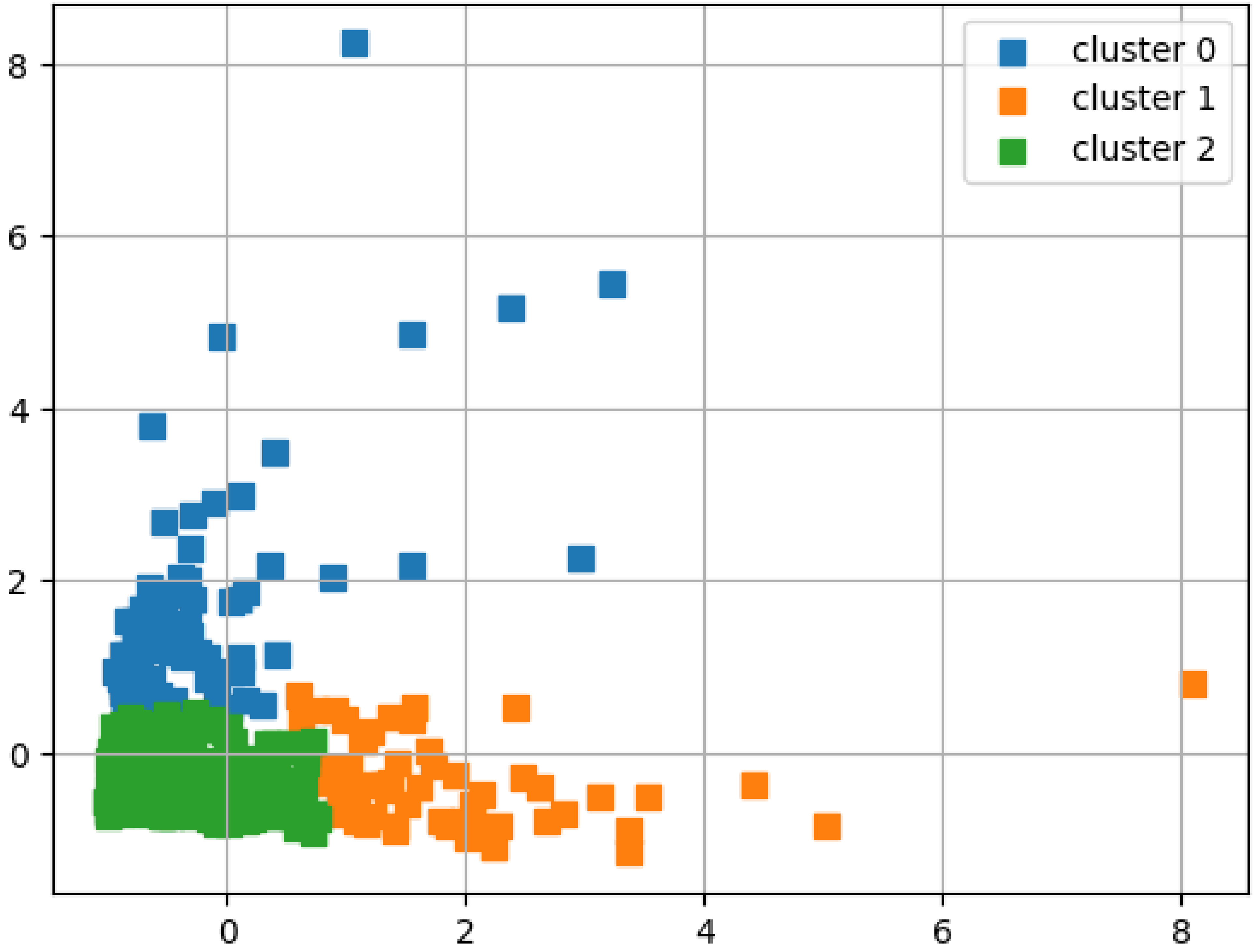


- Selected number of clusters is 5

Hierarchical Clustering

Dendrogram

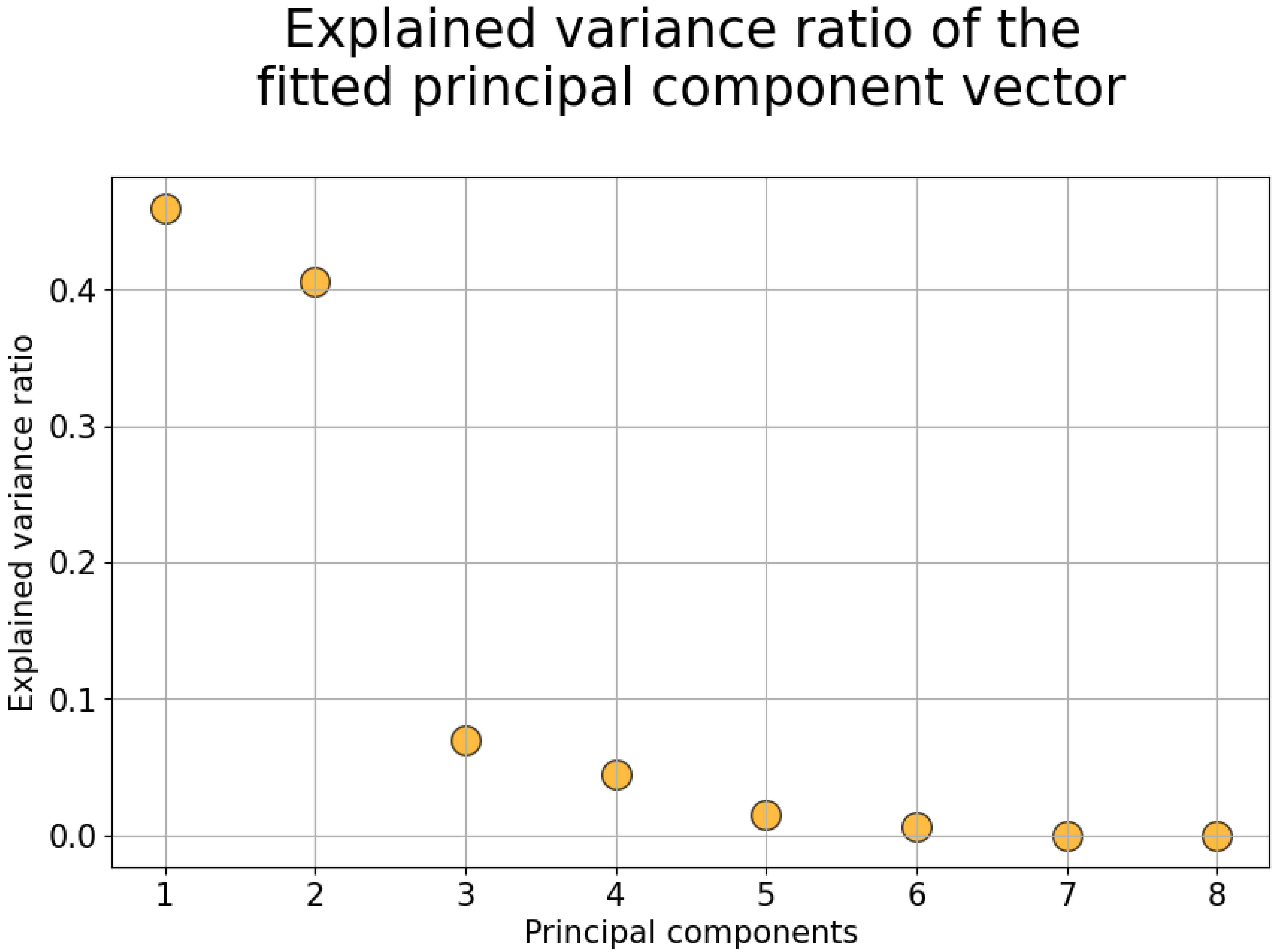




- Selected number of clusters is 3

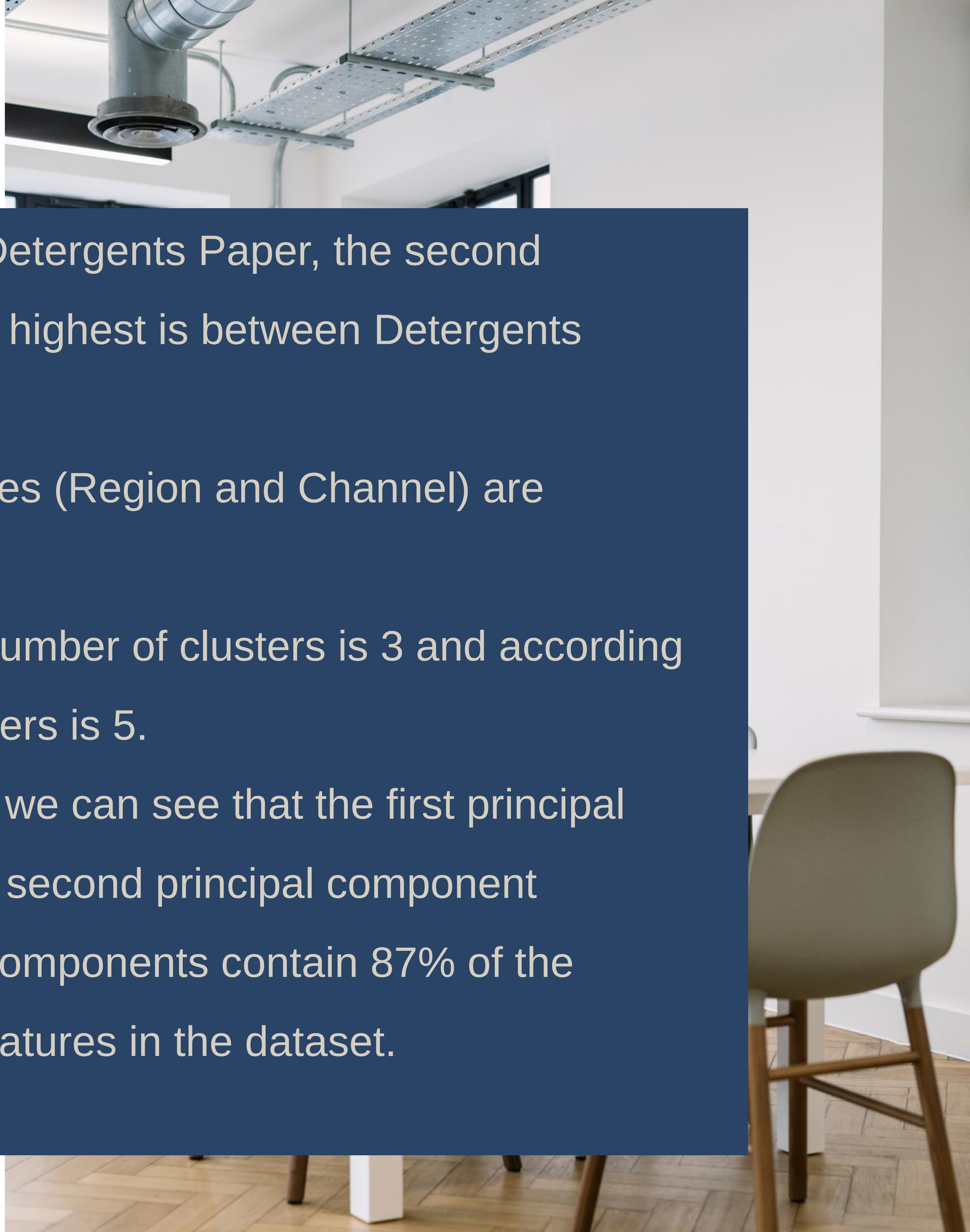
PCA

- By using the attribute explained_variance_ratio_, we can see that the first principal component contains 46% of the variance and the second principal component contains 41% of the variance. Together, the two components contain 87% of the information.



Conclusions

1. The highest correlation is between Grocery and Detergents Paper, the second highest is between grocery and milk and the third highest is between Detergents Paper and Milk.
2. All the features are numerical however two features (Region and Channel) are discrete while others are continuous.
3. According to hierarchical clustering, the optimal number of clusters is 3 and according to k-means clustering the optimal number of clusters is 5.
4. By using the attribute explained_variance_ratio_, we can see that the first principal component contains 46% of the variance and the second principal component contains 41% of the variance. Together, the two components contain 87% of the information. Therefore, there are two important features in the dataset.



Challenges

- Limited time

Future Goals

- More EDA to learn more about the dataset
- More data cleaning

Thank you!