Data Cleaning & analyzing

EDA

Building a Machine LearningModel

Use CSV For collect data

Unsupervised Learning Project

Rinkal raj

Project/Goals

- Use unsupervised learning techniques on a wholesale data dataset. The project involves four main parts: exploratory data analysis and pre-processing, KMeans clustering, hierarchical clustering, and PCA.
- The project involves three main parts: exploratory data analysis, preprocessing, feature engineering, and training a machine learning model.

Tool Installation & Set Up

Use Jupyter lab for python programming.

Use Visual Studio for editing files which clone from GitHub.

Use GitHub for storing files of project data and share on local server it is easy way to access and store data.

Find Missing values:

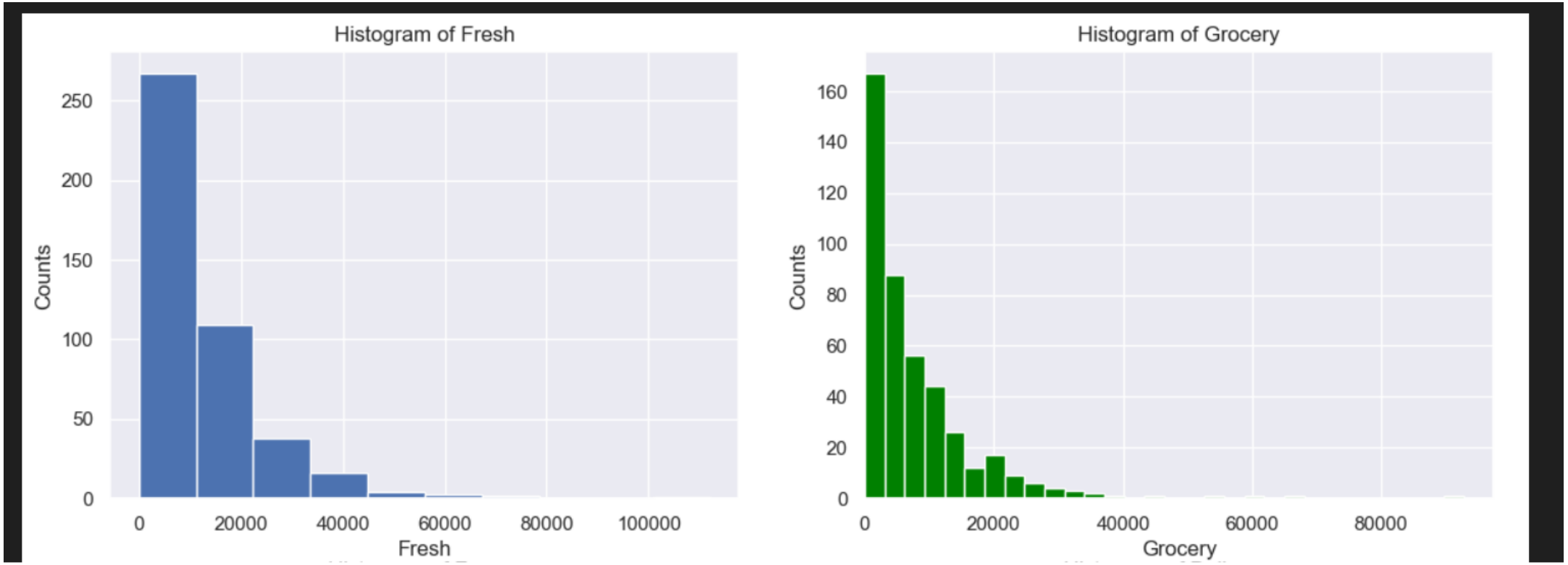
```
#missing values
def missing(x):
    n_missing = x.isnull().sum().sort_values(ascending=False)
    p_missing = (x.isnull().sum()/x.isnull().count()).sort_values(ascending=False)
    missing_ = pd.concat([n_missing, p_missing],axis=1, keys = ['number','percent'])
    return missing_
missing(df)
```

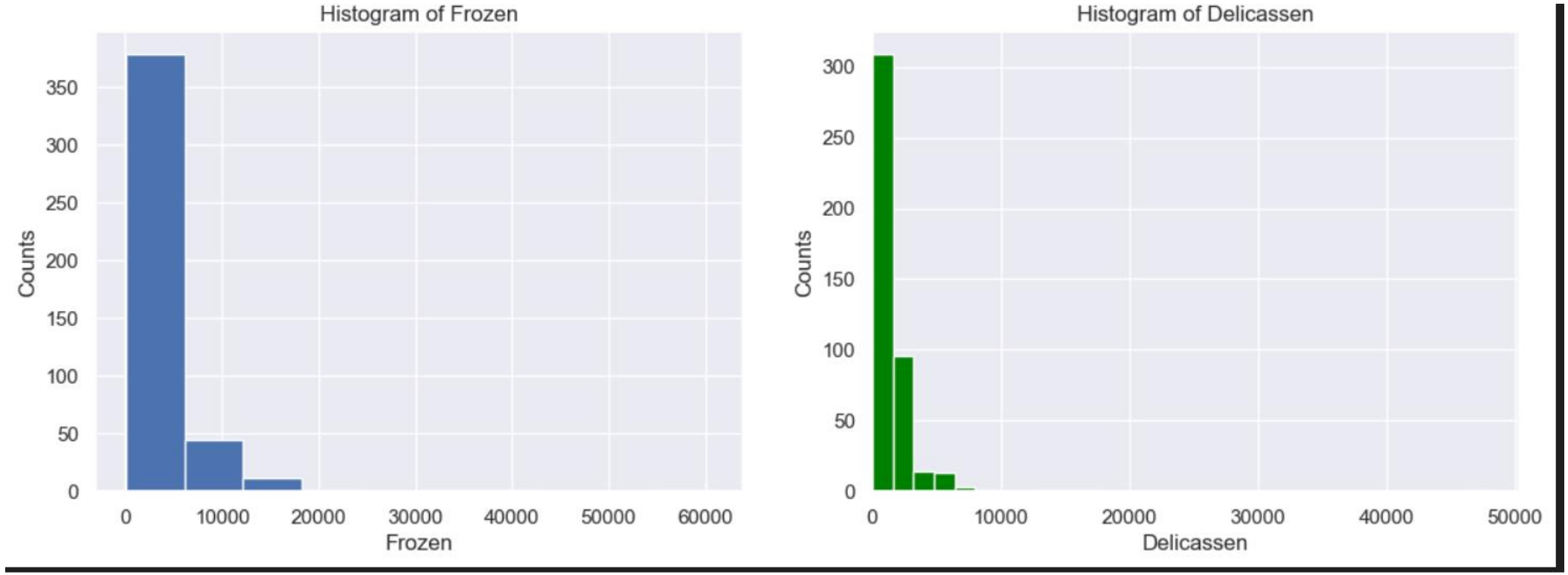
✓ 0.0s

	number	percent
Channel	0	0.0
Region	0	0.0
Fresh	0	0.0
Milk	0	0.0
Grocery	0	0.0
Frozen	0	0.0
Detergents_Paper	0	0.0
Delicassen	0	0.0

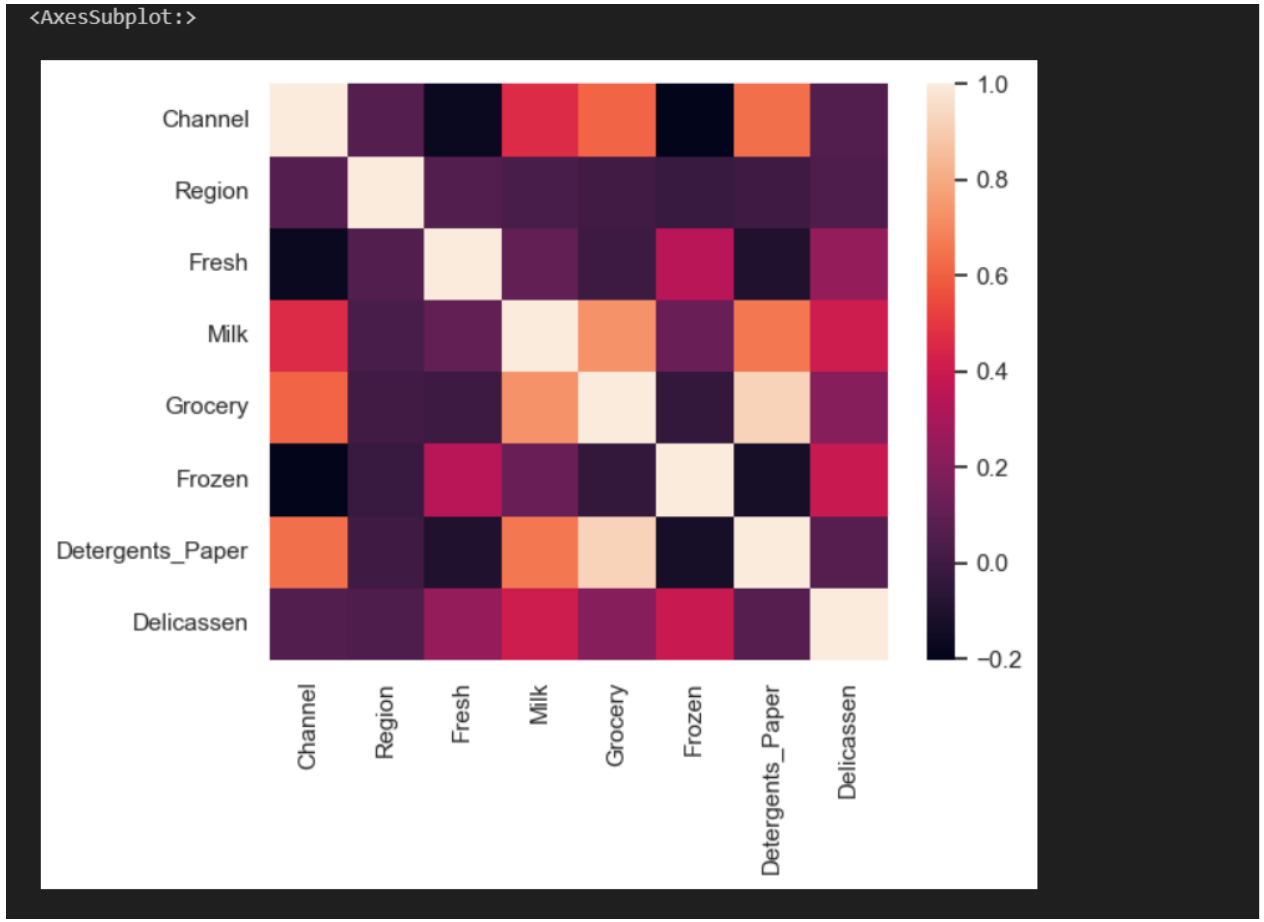
Find missing and null values in dataset and we can see that there is no any null values.

Distribution of data



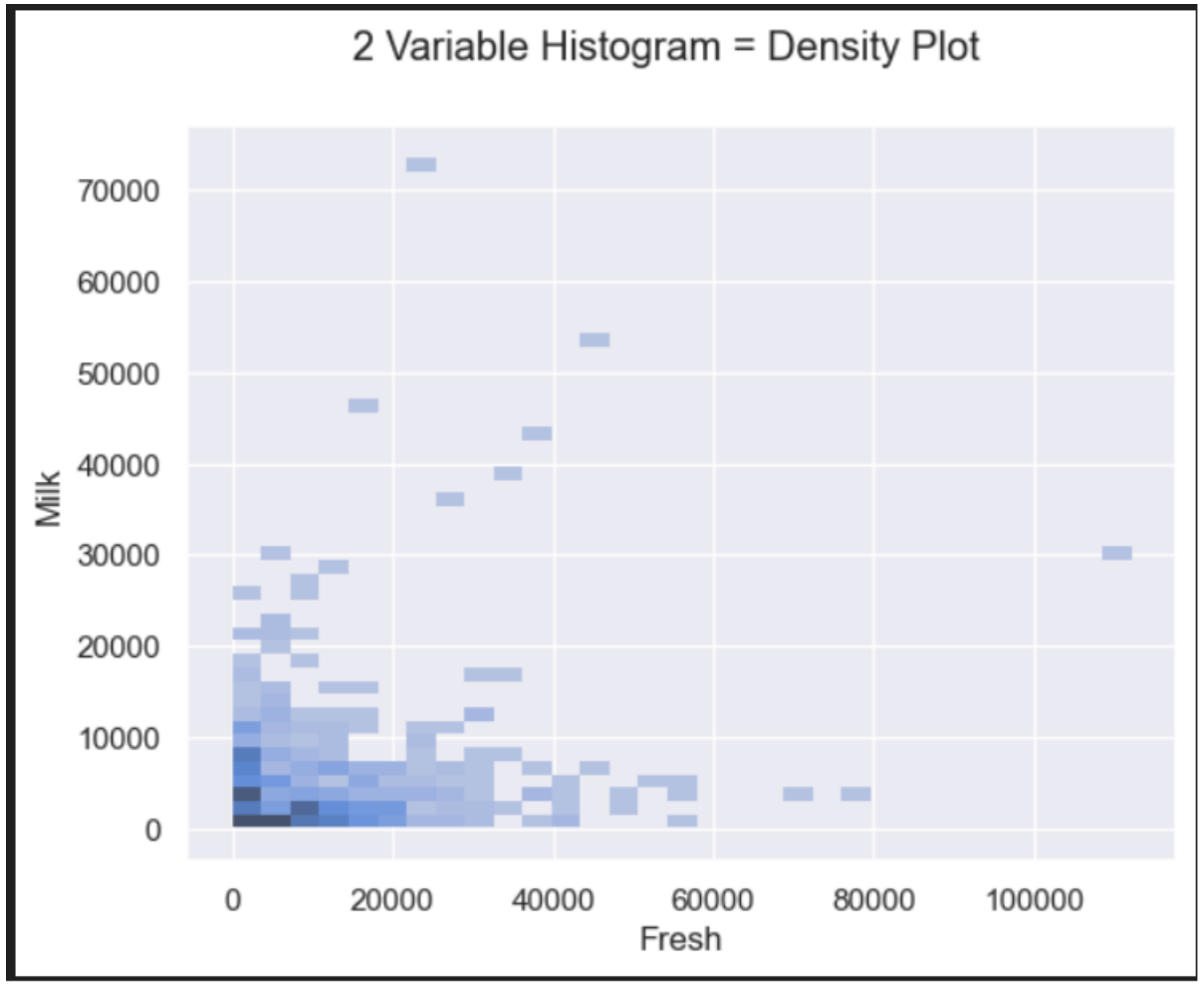

Histogram of Frozen
Counts
0
50
100
150
200
250
300
350
0
10000
20000
30000
40000
50000
60000
Frozen
Histogram of Delicassen
Counts
0
50
100
150
200
250
300
0
10000
20000
30000
40000
50000
Delicassen

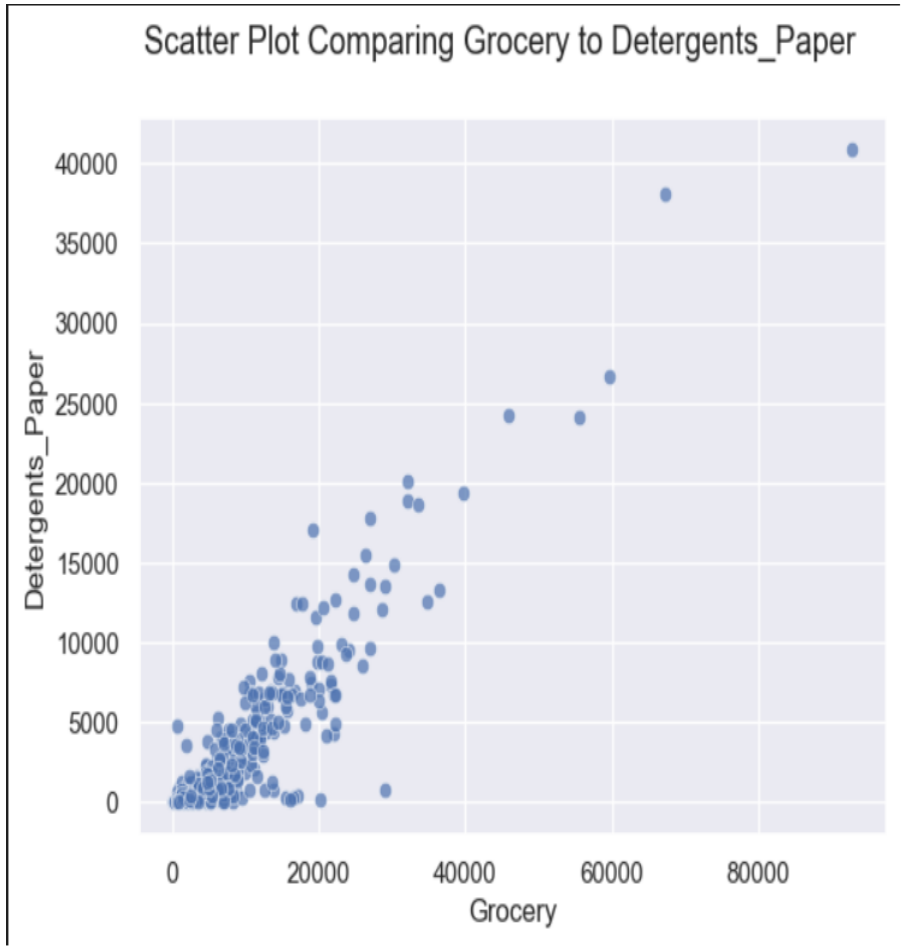
Heatmap

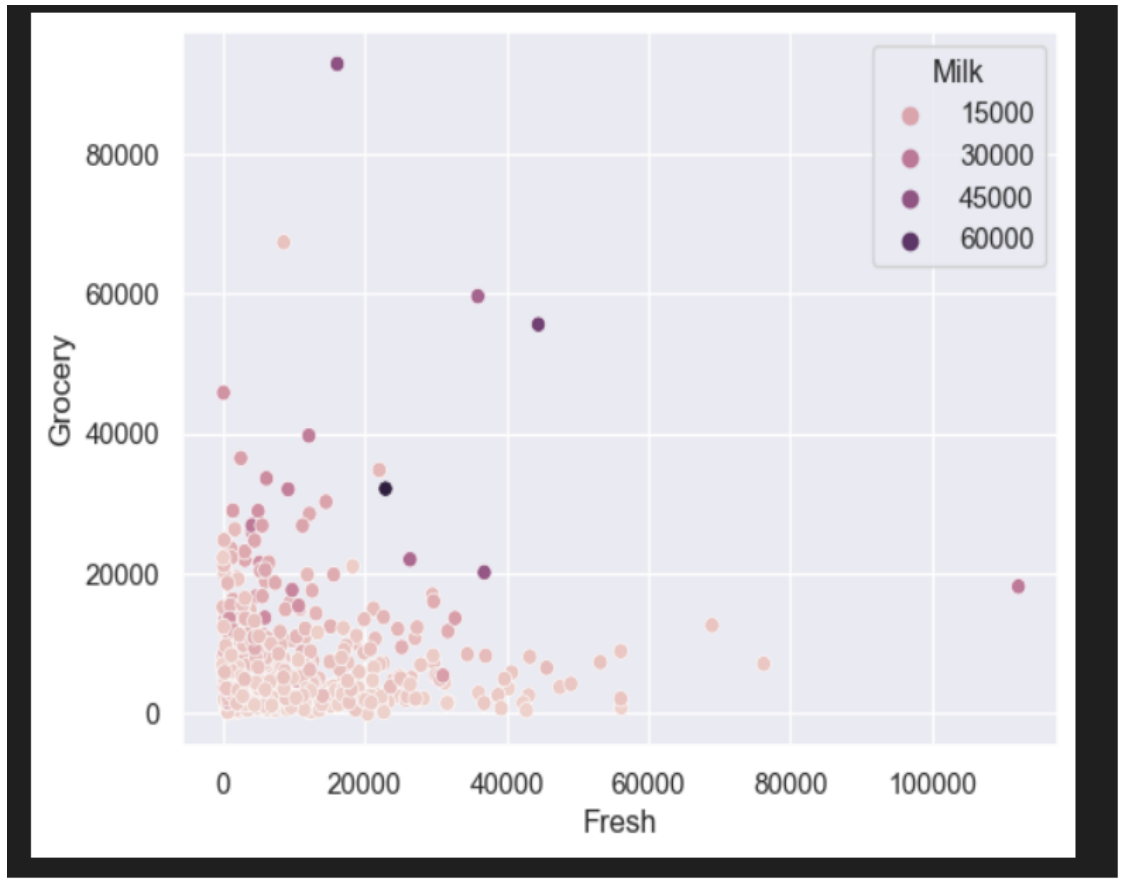


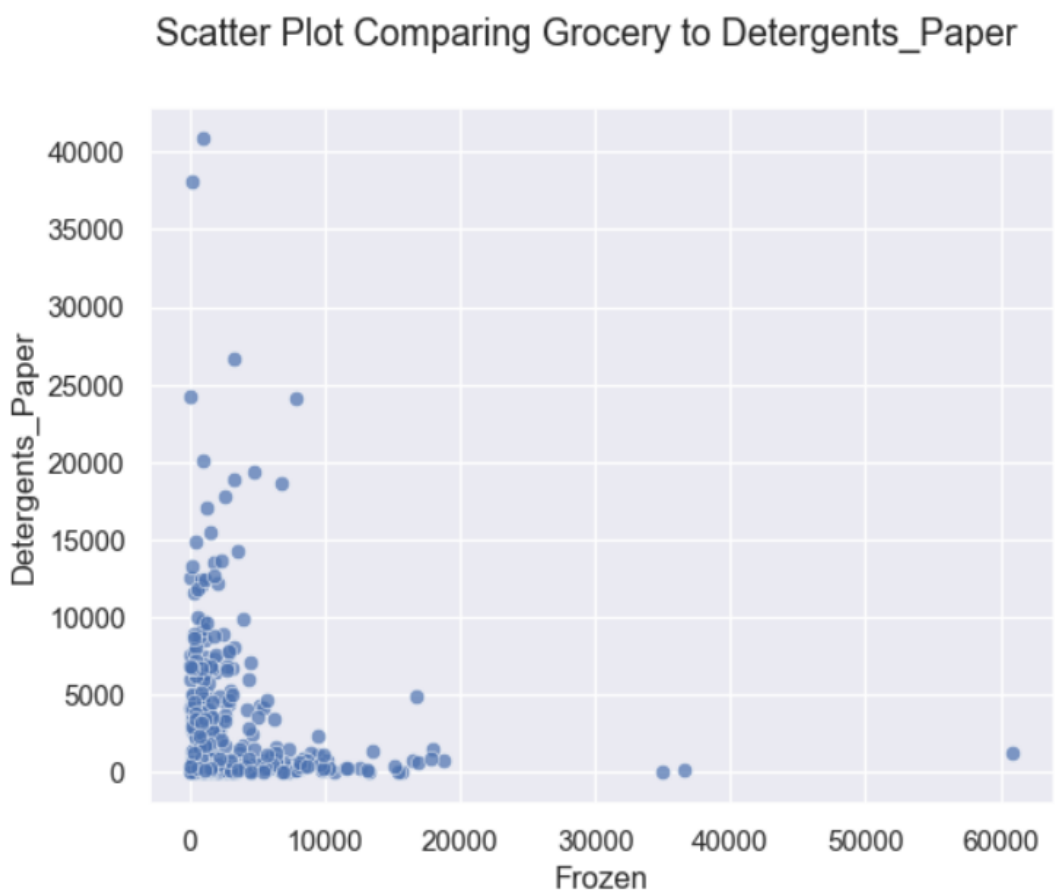
Define the relation between variables.

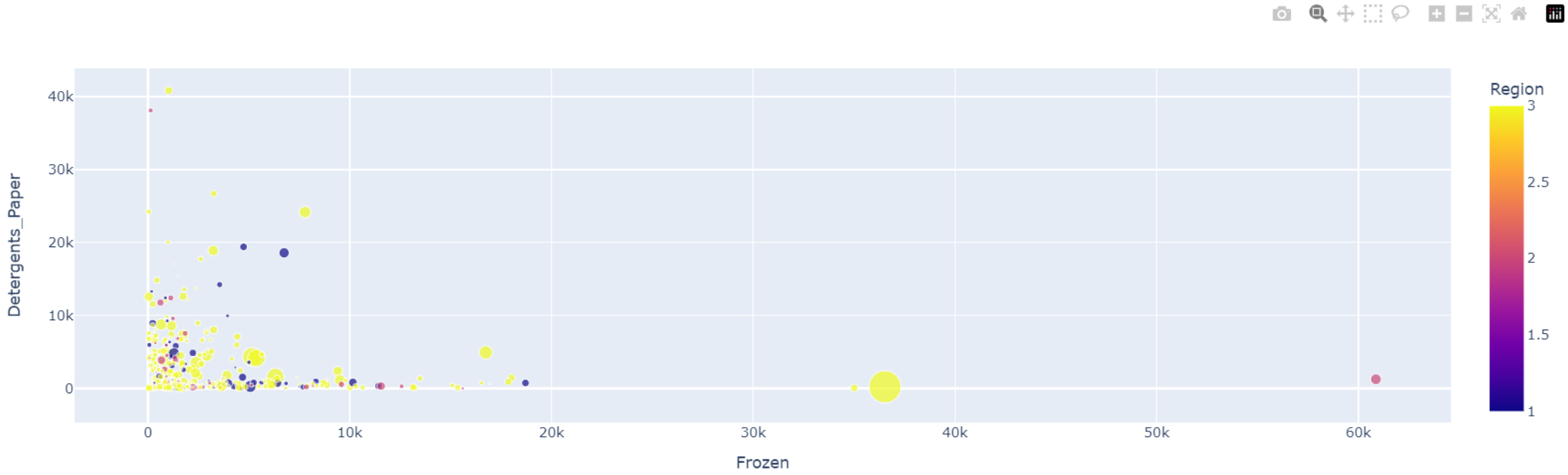
Co-relation between variables





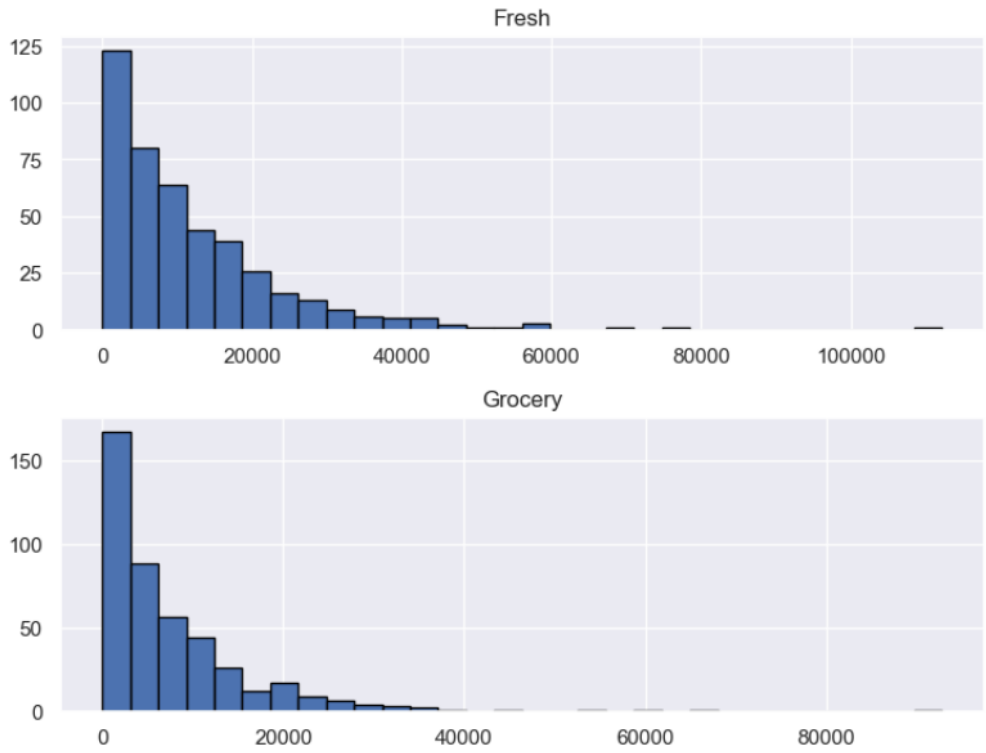

Milk
15000
30000
45000
60000
Grocery
80000
60000
40000
20000
0
0
20000
40000
60000
80000
100000
Fresh

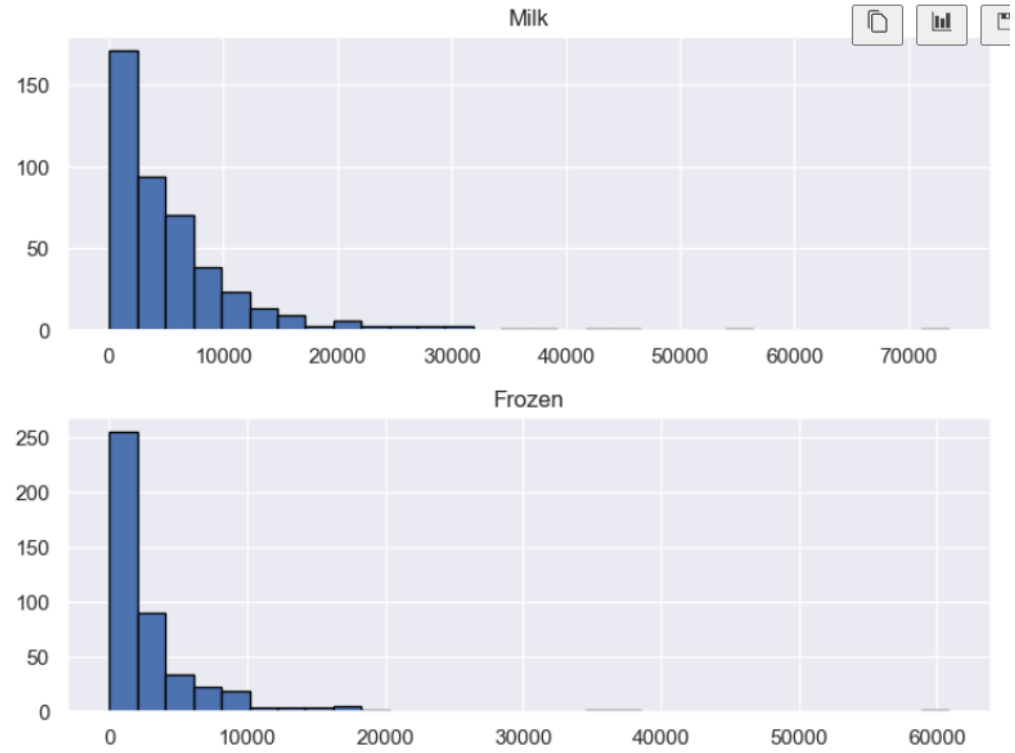

Scatter Plot Comparing Grocery to Detergents_Paper
Detergents_Paper
40000
35000
30000
25000
20000
15000
10000
5000
0
0
10000
20000
30000
40000
50000
60000
Frozen

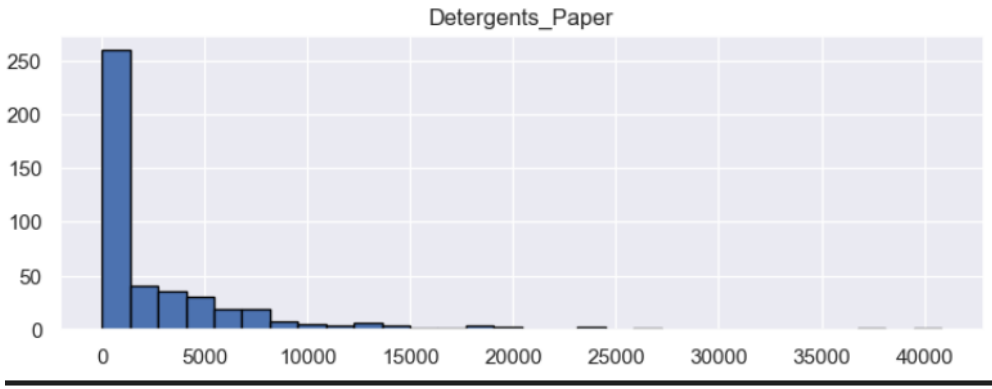


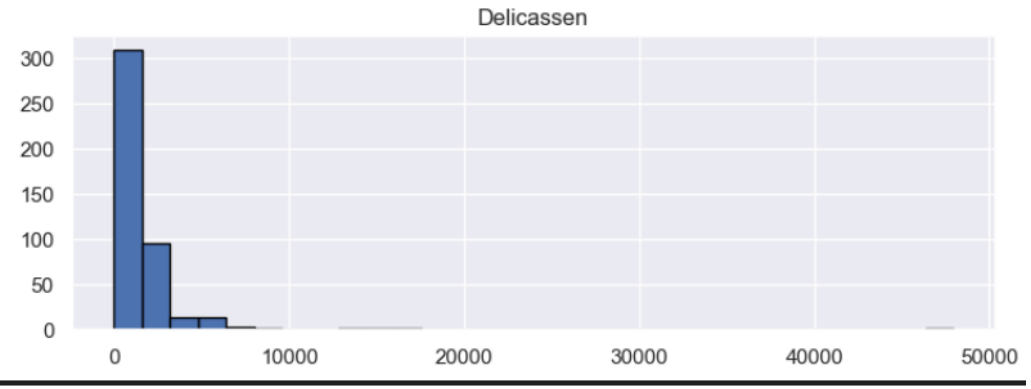
Relation between products and region as well as the channel.
Each EDA shows that the distributor’s annual spending is not more than 20k on every product.

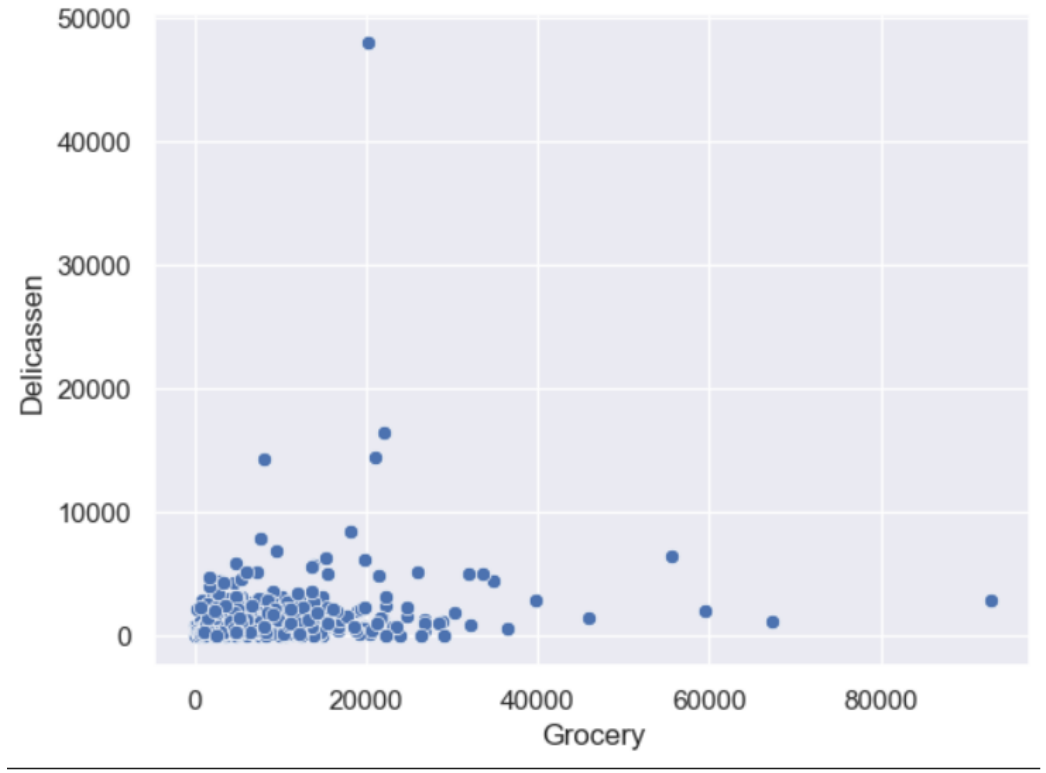
Outliners


Fresh
125
100
75
50
25
0
0
20000
40000
60000
80000
100000
Grocery
150
100
50
0
0
20000
40000
60000
80000


Milk
150
100
50
0
0
10000
20000
30000
40000
50000
60000
70000
Frozen
250
200
150
100
50
0
0
10000
20000
30000
40000
50000
60000

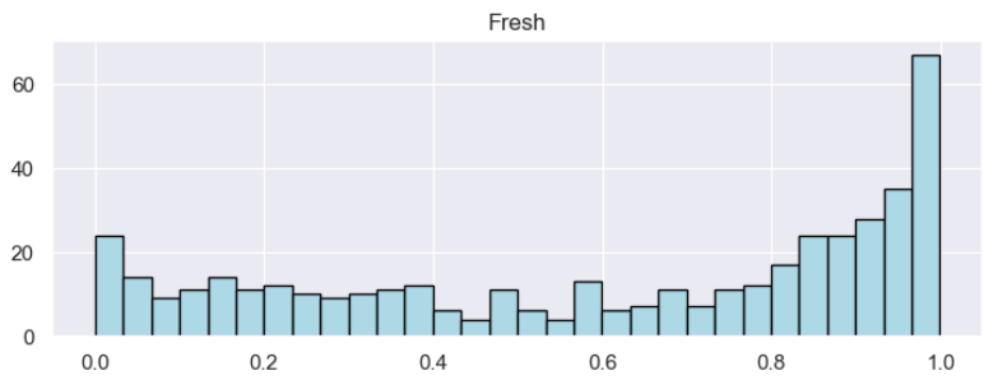

Detergents_Paper
250
200
150
100
50
0
0
5000
10000
15000
20000
25000
30000
35000
40000

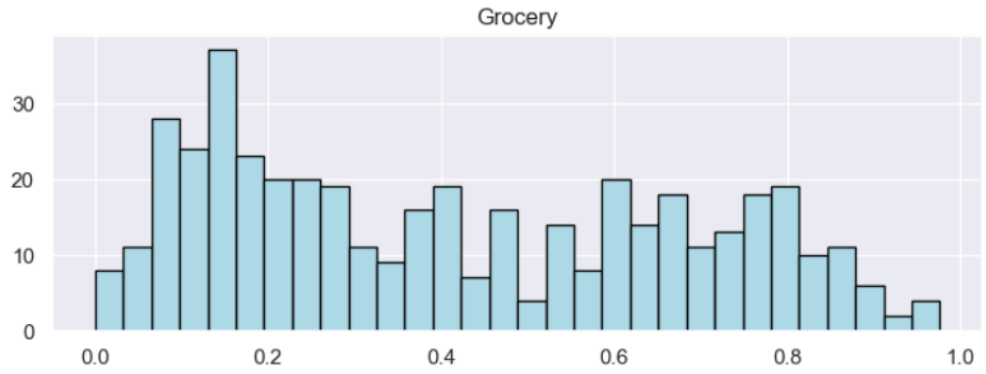

Delicassen
300
250
200
150
100
50
0
0
10000
20000
30000
40000
50000

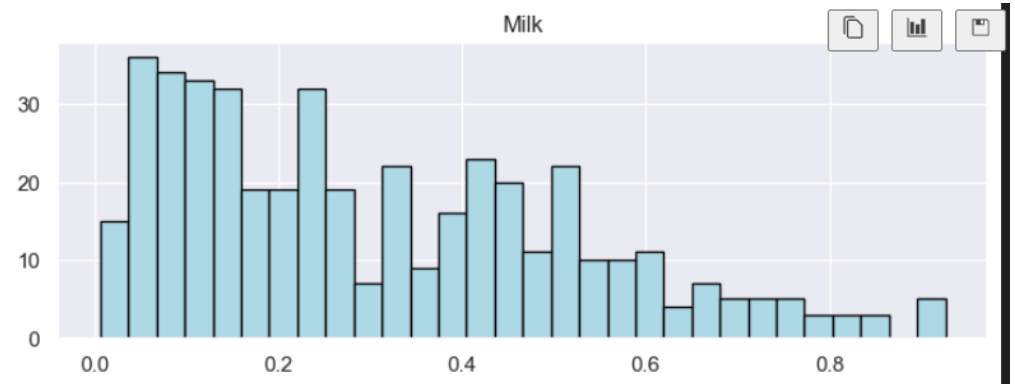


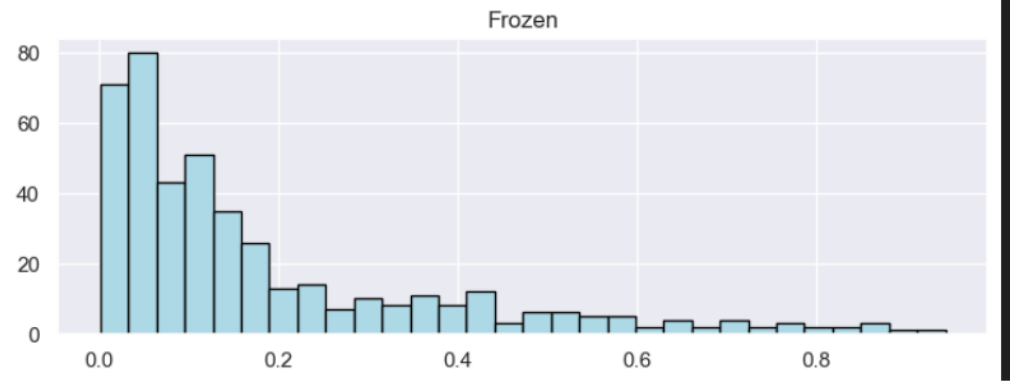
The data appears to be very skewed and "rich" in outliers which can negatively affect our analysis.

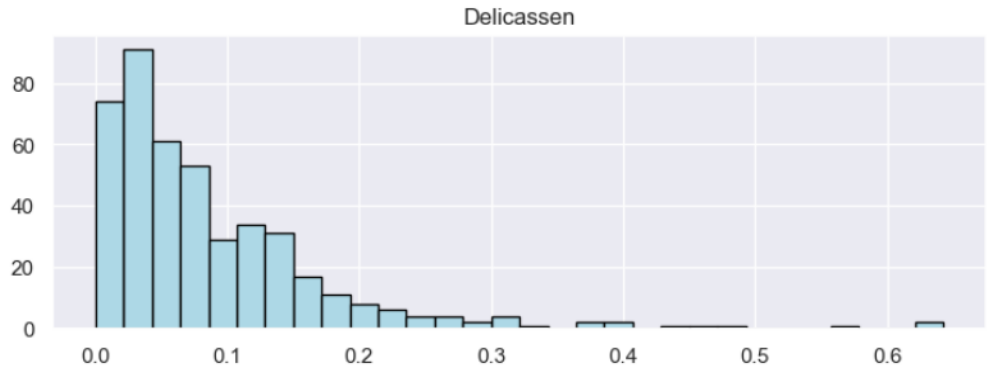
Data Normalization

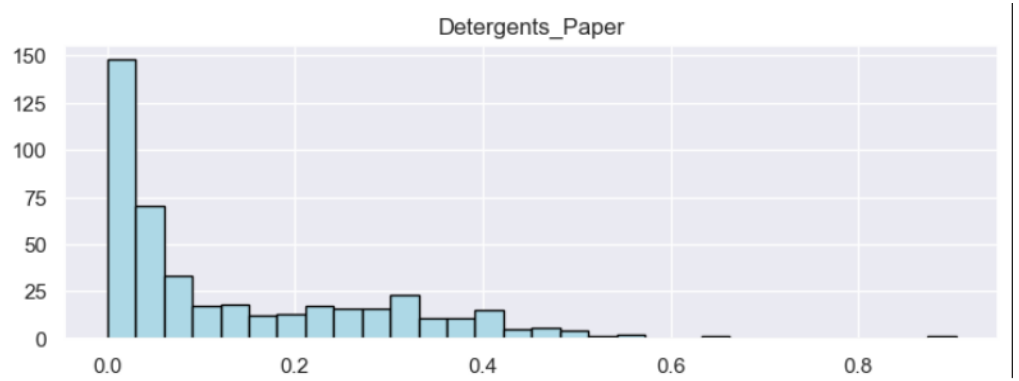






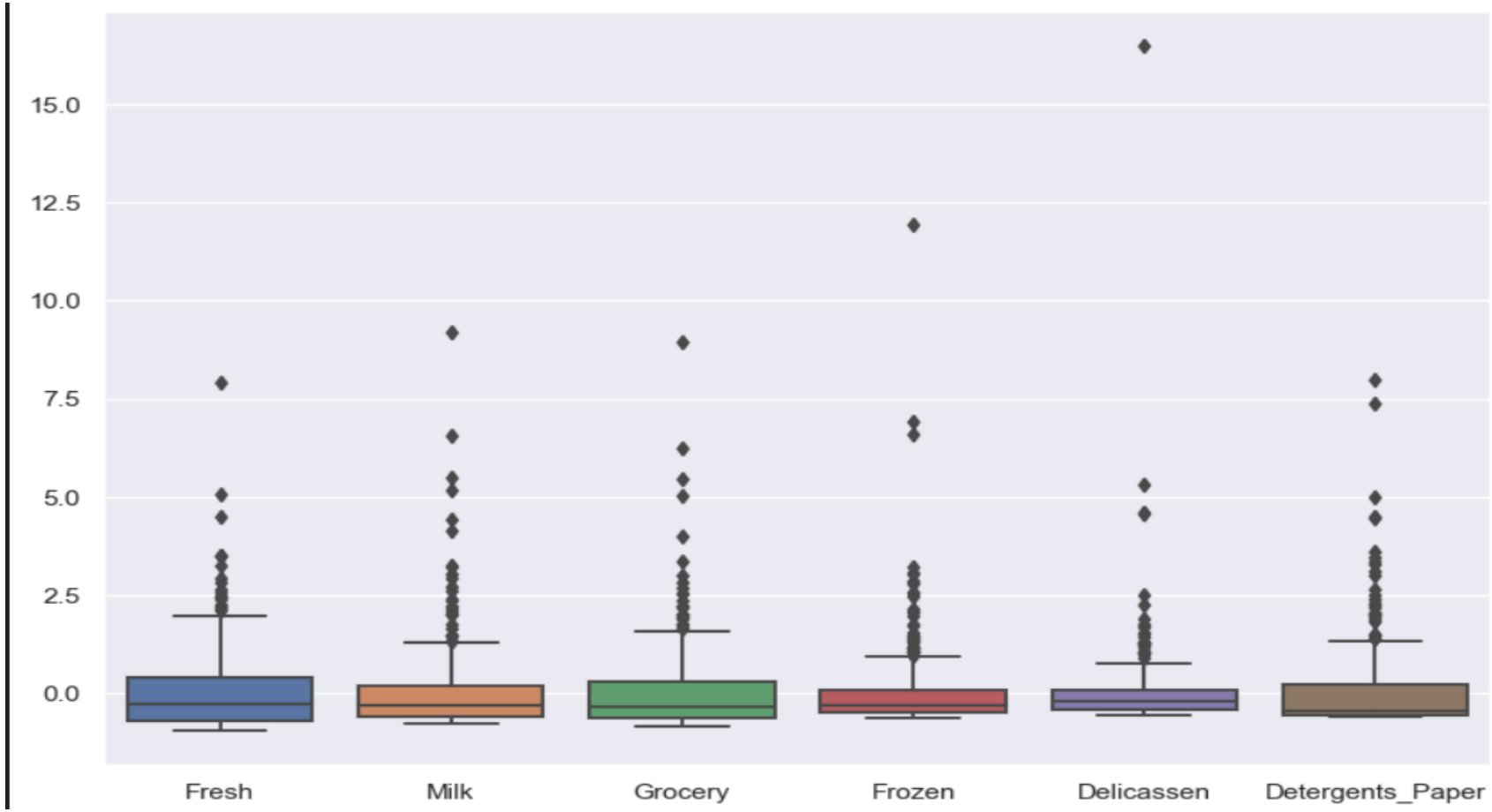




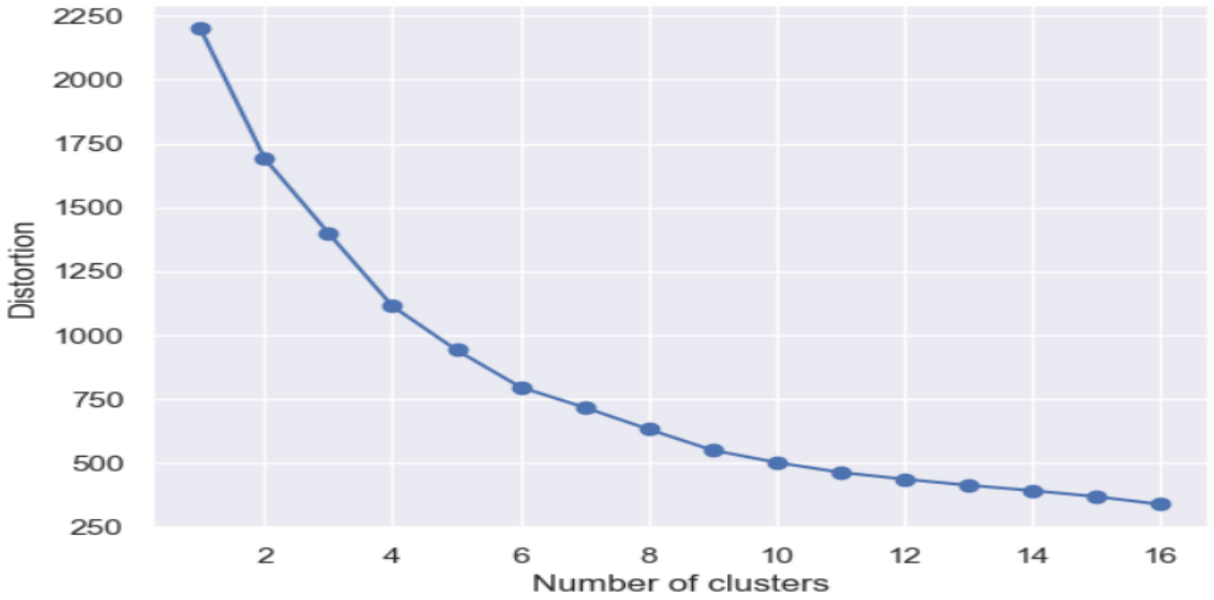


We can see that the high level of outliners thus, data is not normalized and scaled. For example, if one variable is expressed in kilograms and another is in grams, Kmeans will consider 1000 grams as 1000 times higher than 1 kg as it considers only values and not units. We have to scale the data to the same unit. So, we need to normalize and scale data after that, we can get the proper output from the algorithm of K-means clustering, Hierarchical clustering, and PCA.

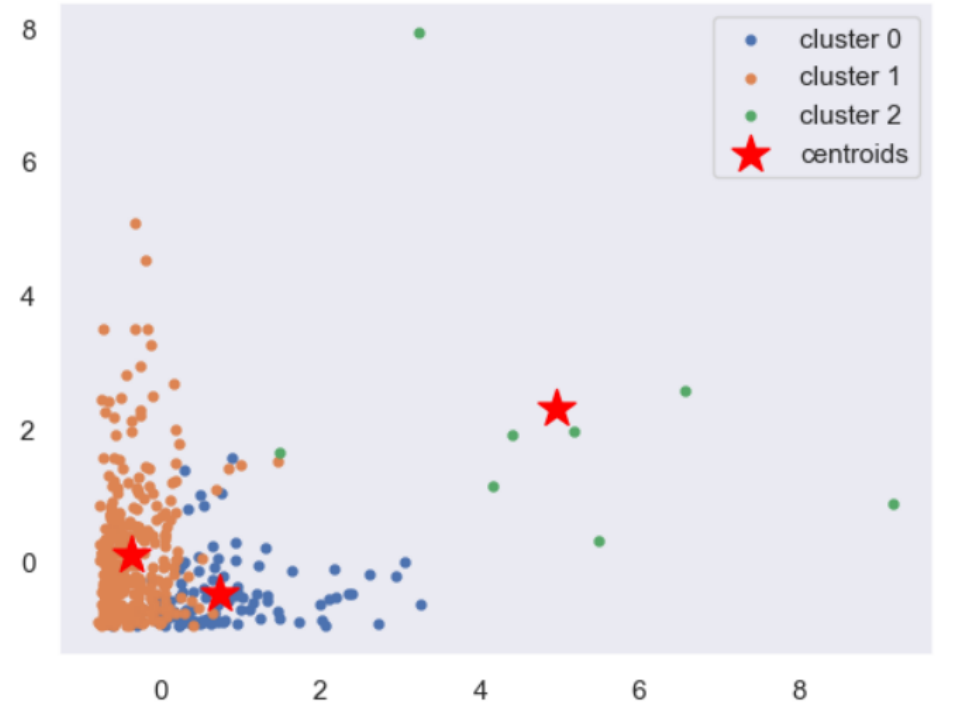
Data Scaling

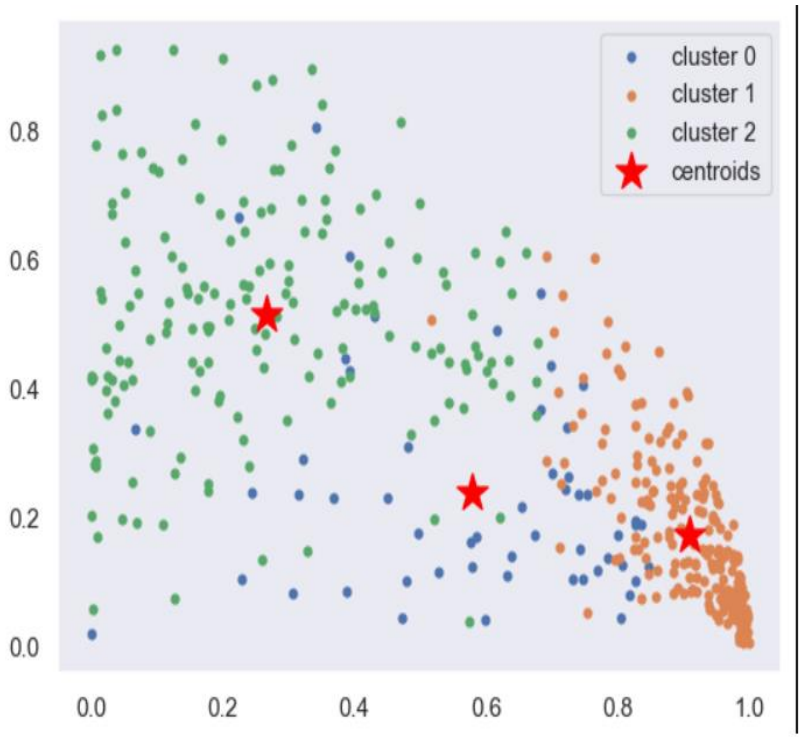


K-means Clustering



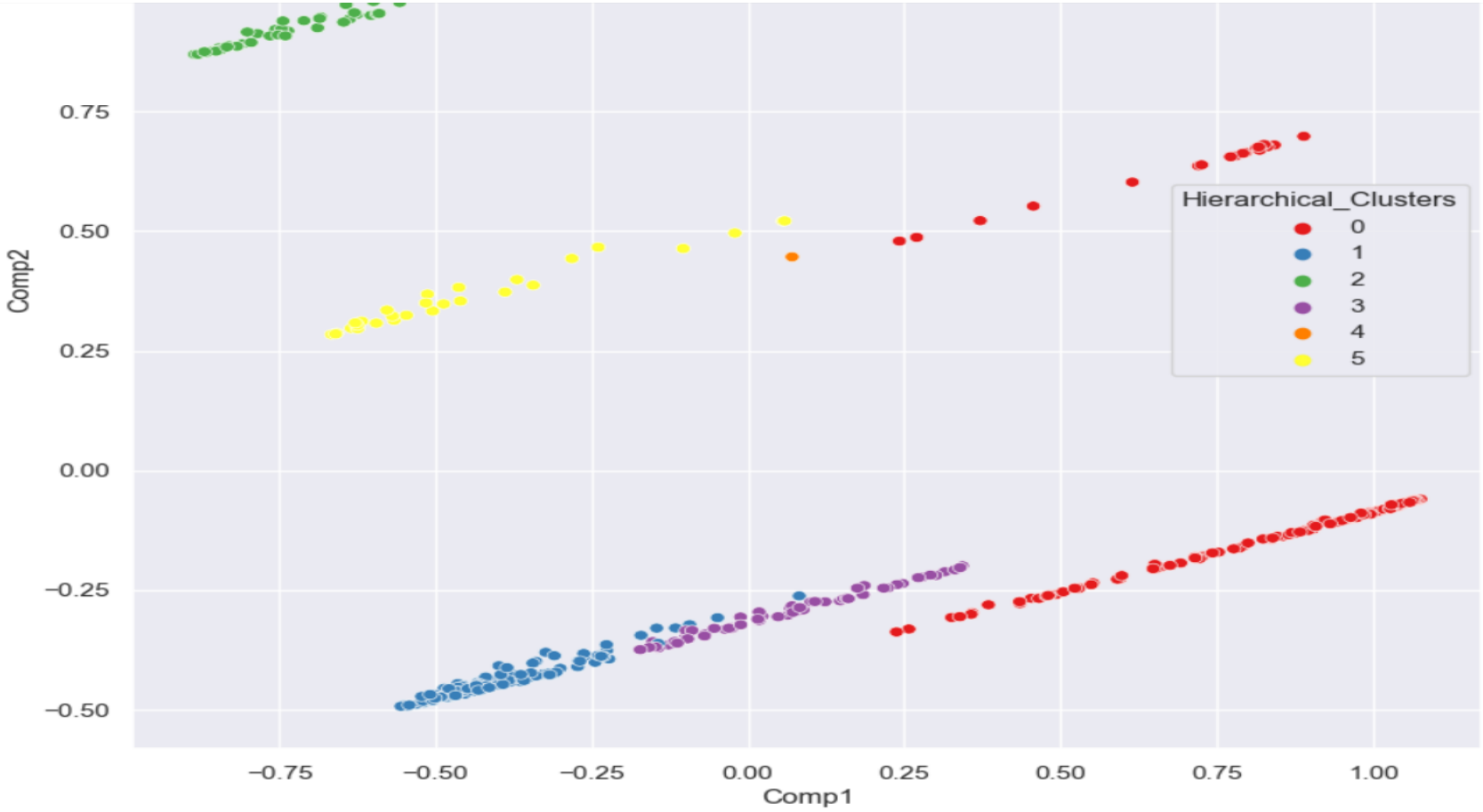
This is the Elbow rule which helps us to decide the number of clusters in Kmean clustering.
In this graph, our inflection point and number of clusters are 10.



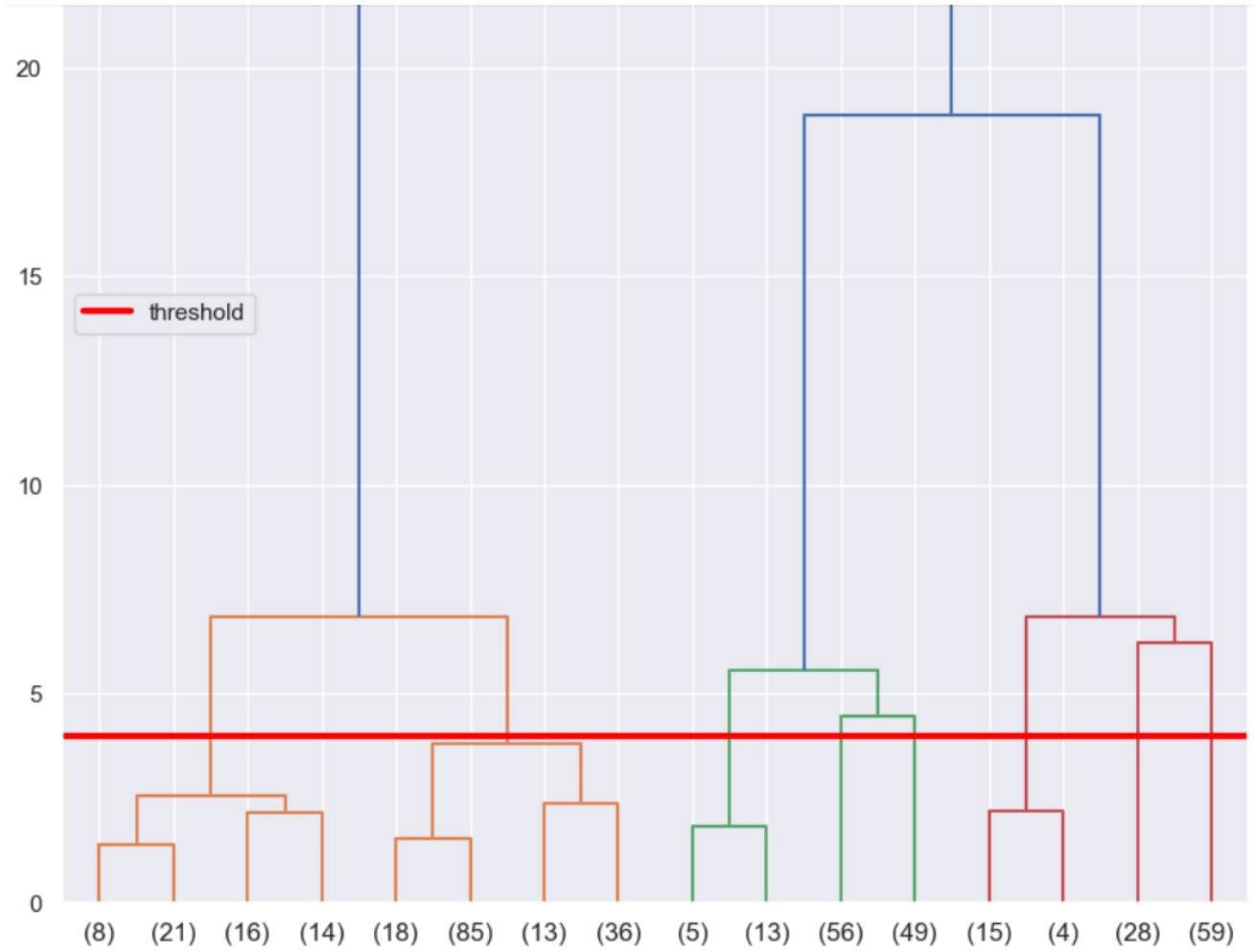


K mean clustering with centroid.

Hierarchical Clustering

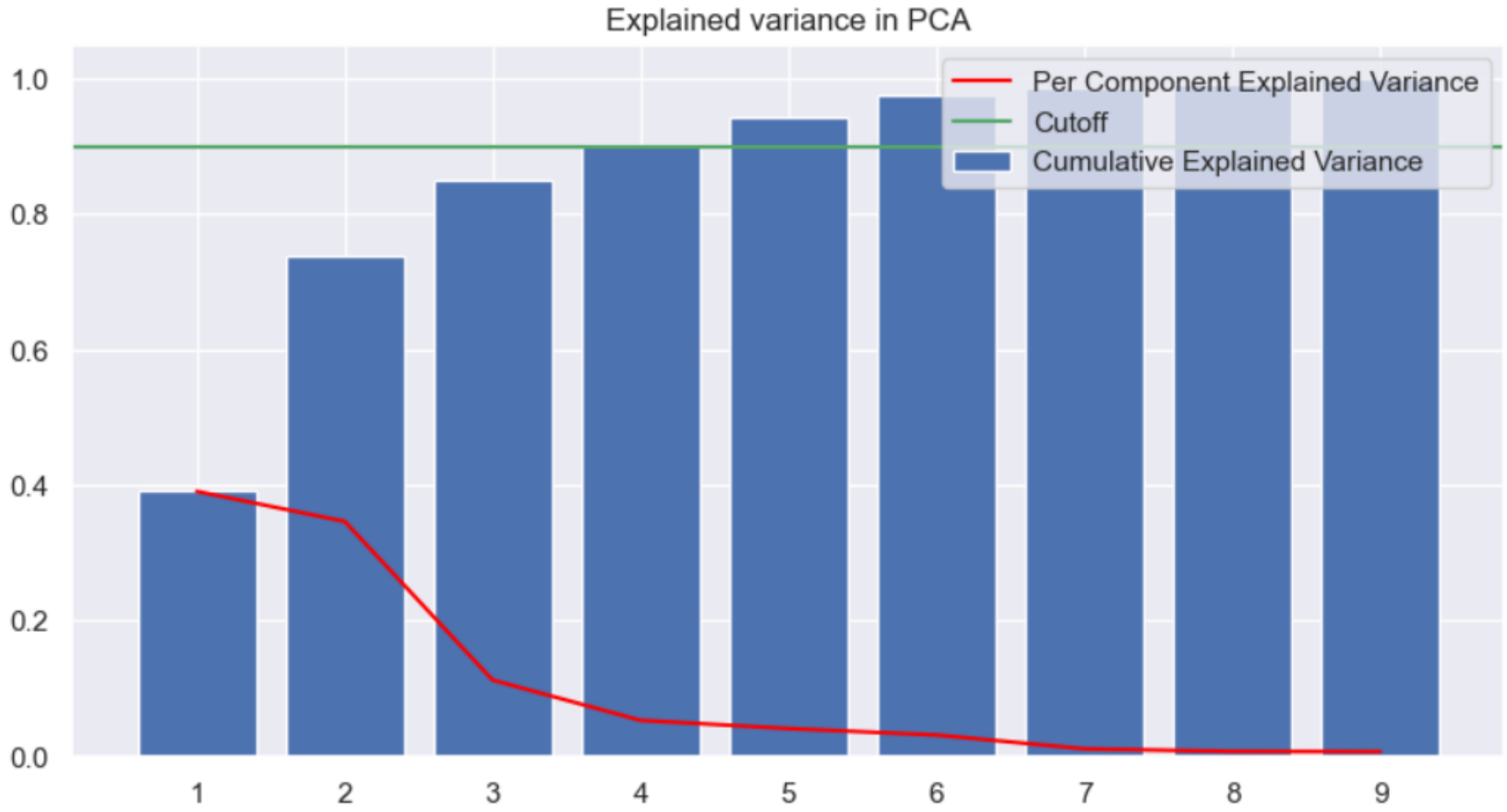


We can see that the clusters are now formed adequately. But what happens if we want to see what are options in forming clusters, or how the

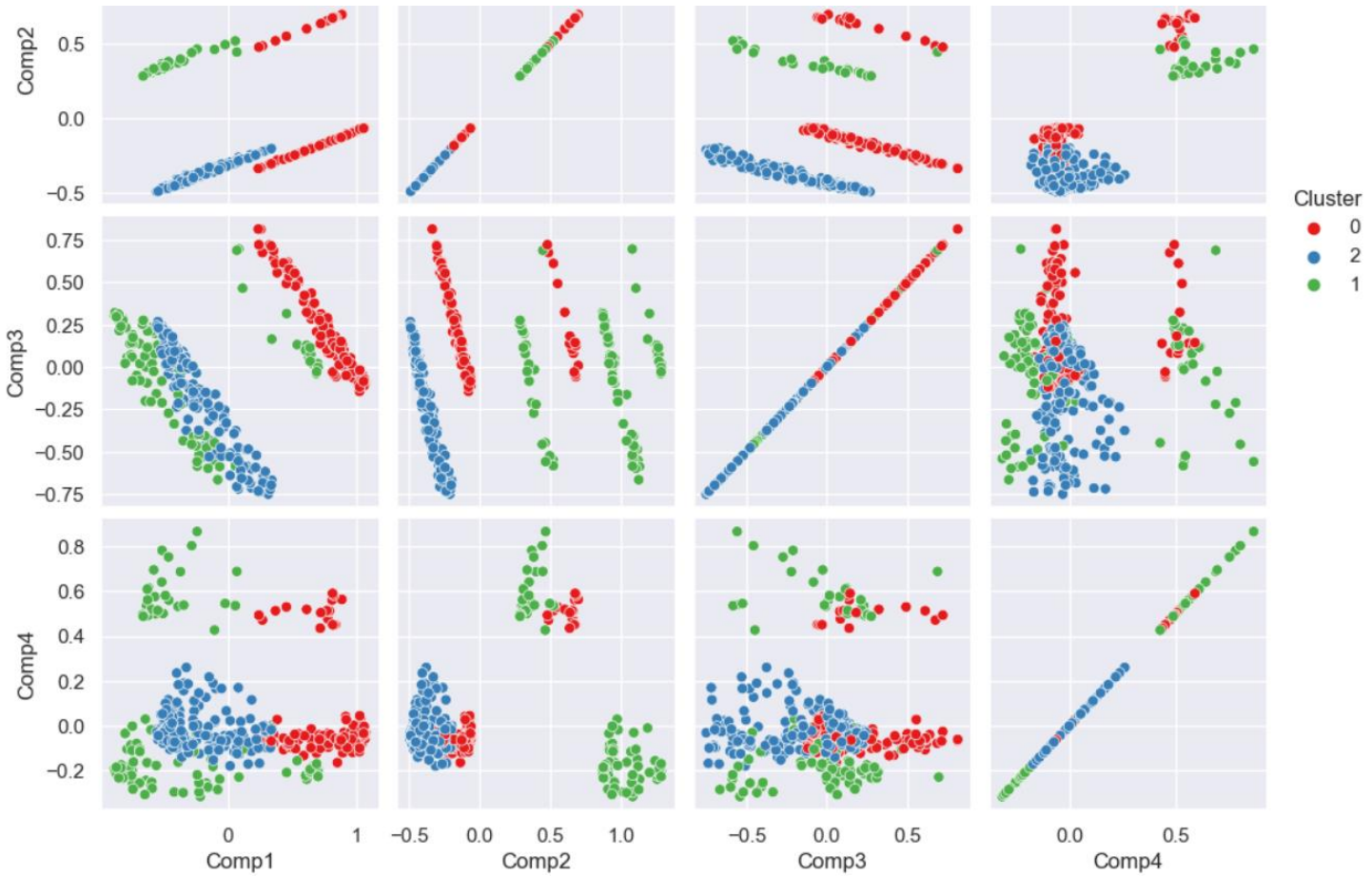


Pay attention that there are vertical lines crossing the horizontal red line. These vertical lines denote the 5 clusters formed. The numbers in parenthesis are numbers of observations.

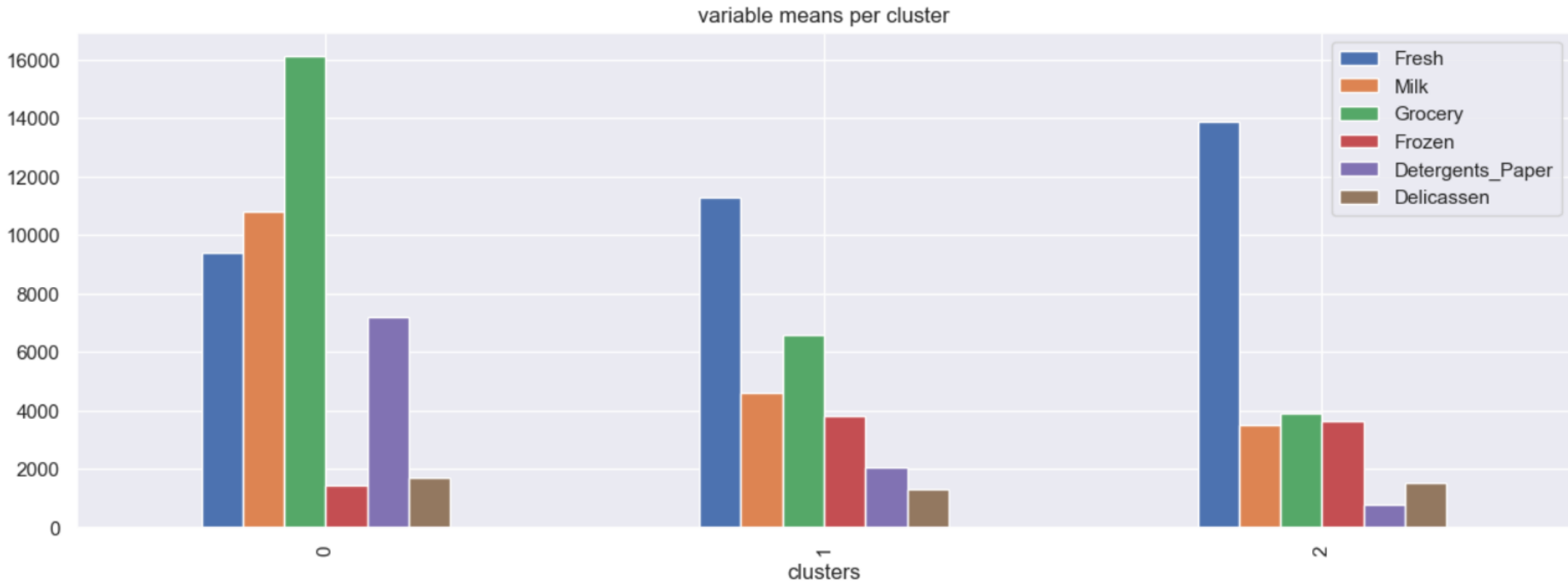
PCA



We can see that 90% of explained variability is reached when we use 4 components. Therefore we will implement dimensionality reduction to 4 components.



It´s difficult to grasp the clusters like that. Let´s perform a component reduction to two variables.


variable means per cluster
16000
14000
12000
10000
8000
6000
4000
2000
0
Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicassen
0
1
2
clusters

Conclusion

- In this project, I built various EDA for data exploration and understanding of datasets. Thus, I found the distributor’s annual spending between 10k to 20k on every product not more than that.
- Dataset is not normally distributed and scaled thus, I did normalize and scale data, so our machine learning model work properly and we get proper output.
- In k mean clustering use the elbow rule and I found a number of clusters 8 also I created a data point and centroid with various clusters. And also, to minimize the sum of distances between the points and their respective cluster centroid.
- In Hierarchical clustering, I get the numbers in parenthesis are numbers of observations.
- In PCA, I implemented dimensionality reduction to 4 components because on 4 components I get 90% of explained variability.

THANK YOU