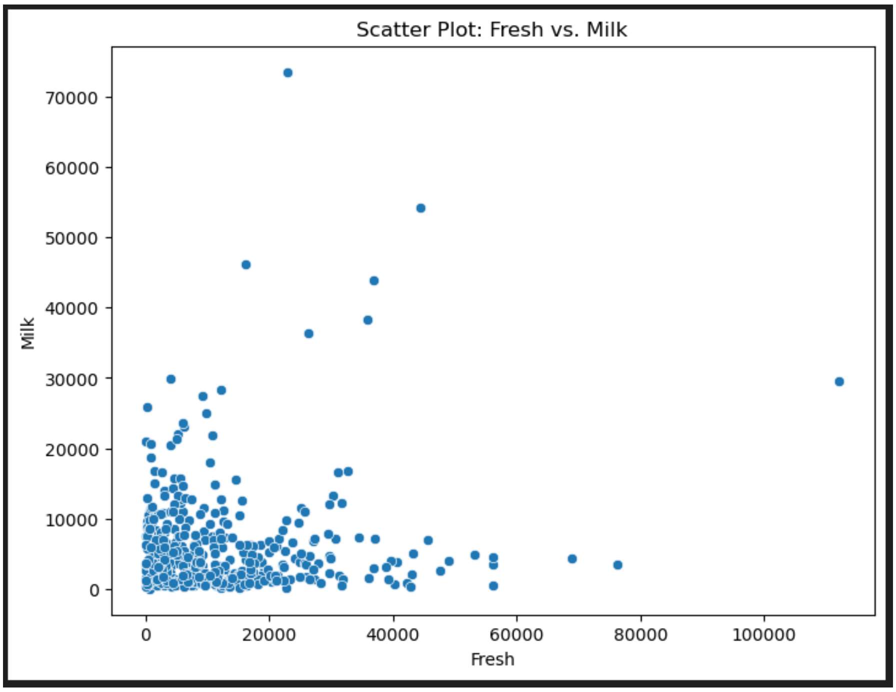
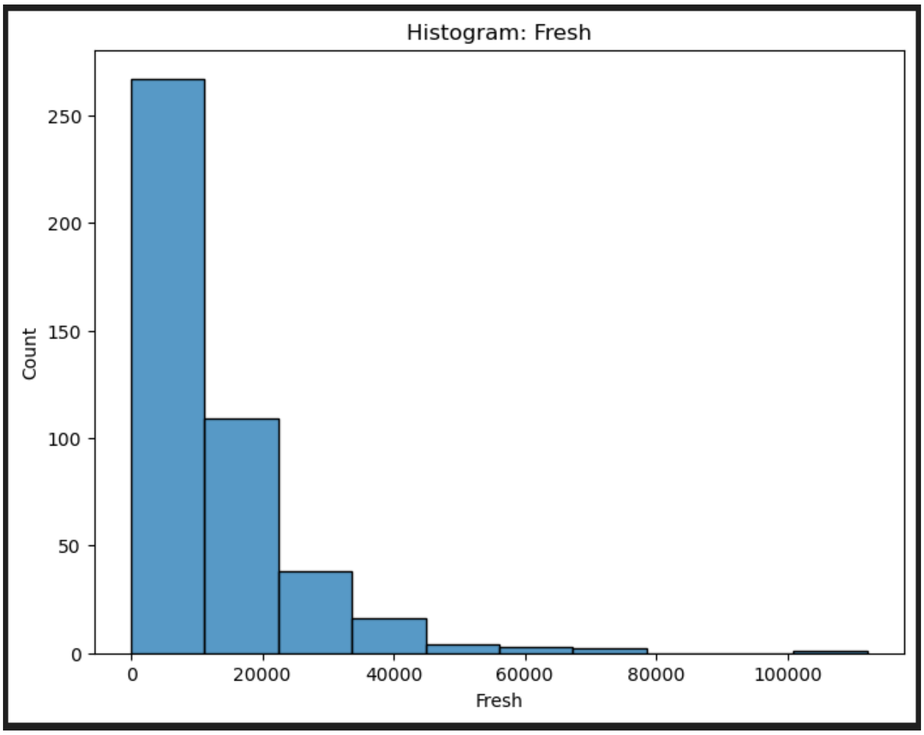
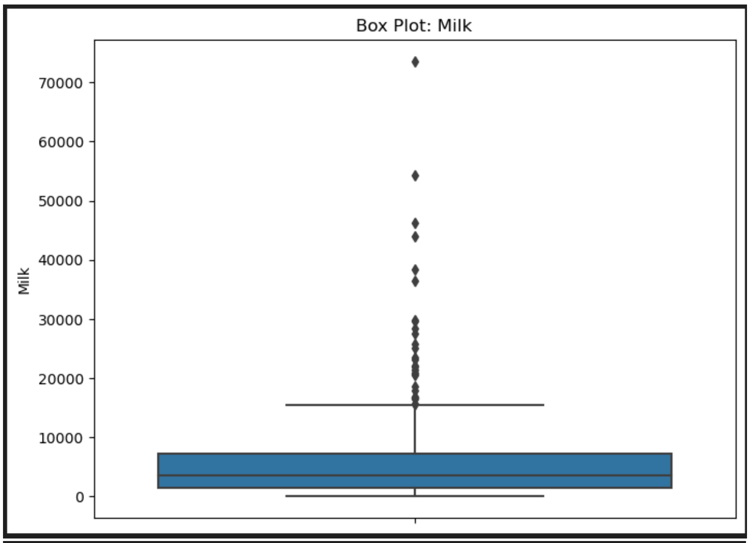
UNSUPERVISED LEARNING PROJECT

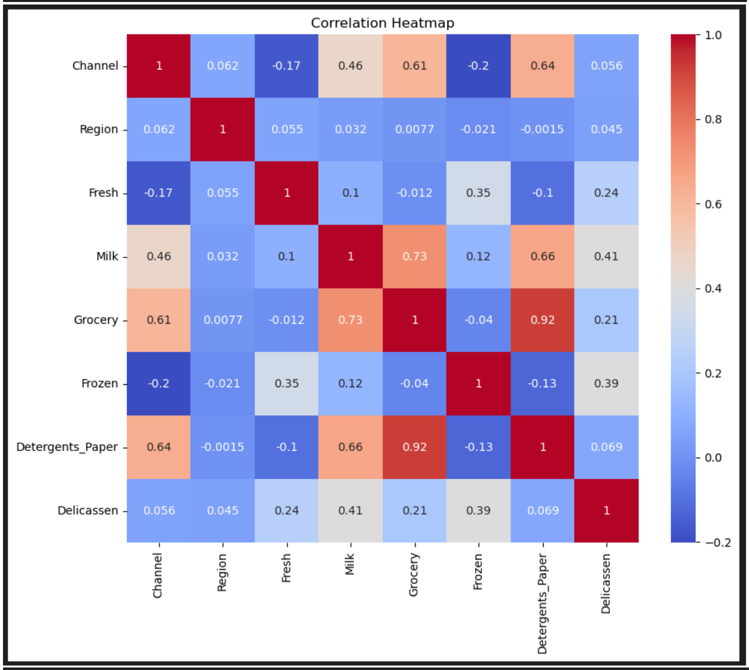
1) Import all the required libraries
2) Load the CSV file
3) Check for missing/null data
4) Used the following visualizations :



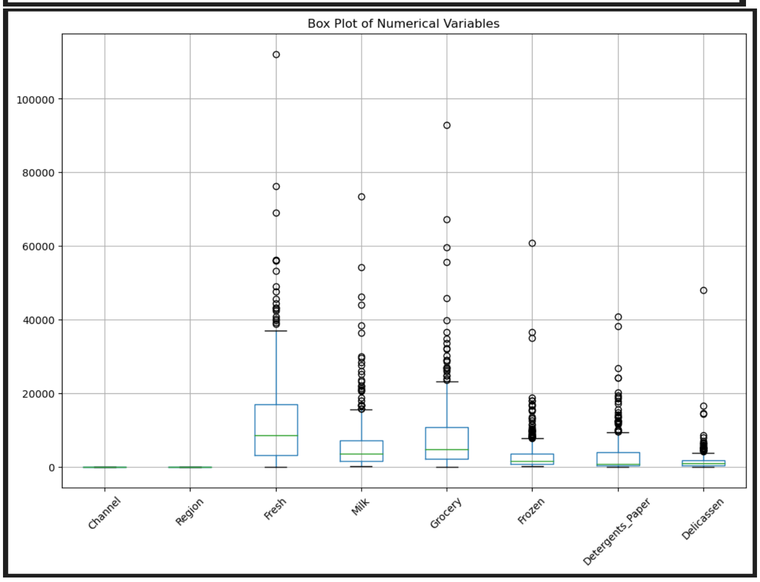




Box Plot: Milk
Milk
70000
60000
50000
40000
30000
20000
10000
0



Correlation Heatmap
Channel
Region
Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicassen
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



Box Plot of Numerical Variables
100000
80000
60000
40000
20000
0
Channel
Region
Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicassen

5) Check for outliers

6) Check for correlation :

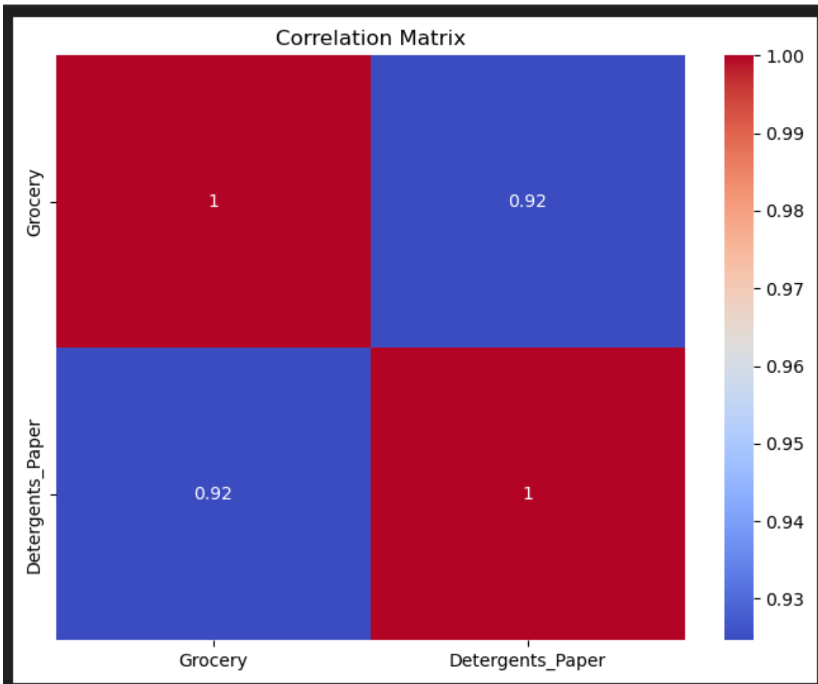
```
# Calculate the correlation matrix
correlation_matrix = df[['Grocery', 'Detergents_Paper']].corr()

# Print the correlation matrix
print("Correlation Matrix:")
print(correlation_matrix)
```

✓ 0.0s

```
Correlation Matrix:
                   Grocery  Detergents_Paper
Grocery           1.000000          0.924641
Detergents_Paper  0.924641          1.000000
```

7) Check for correlation using viz :



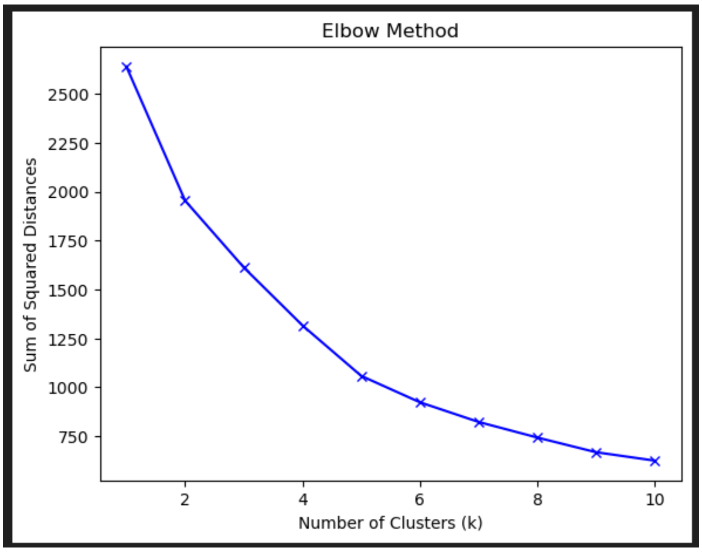

```
grocery_detergents_correlation = correlation_matrix.loc['Grocery', 'Detergents_Paper']
print("Correlation between Grocery and Detergents_Paper:", grocery_detergents_correlation)
```

✓ 0.0s

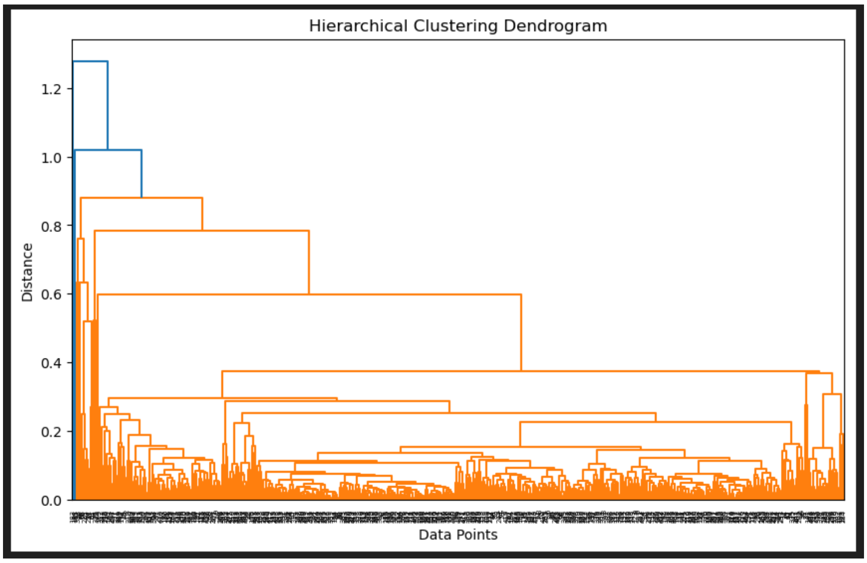
```
Correlation between Grocery and Detergents_Paper: 0.9246406908542676
```

8) Standardize the data and perform PCA

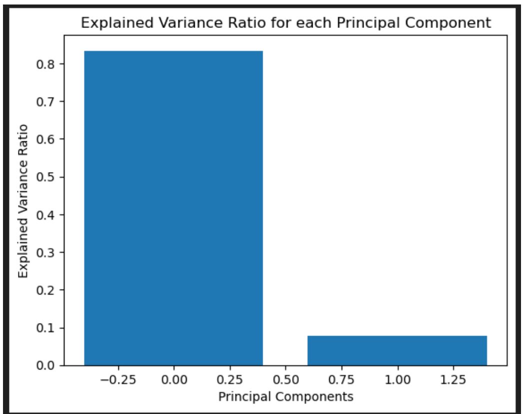
9) Use Scikit Learn to perform KMeans Clustering and find elbow curve :



10) Use Scipy to get Dendogram:



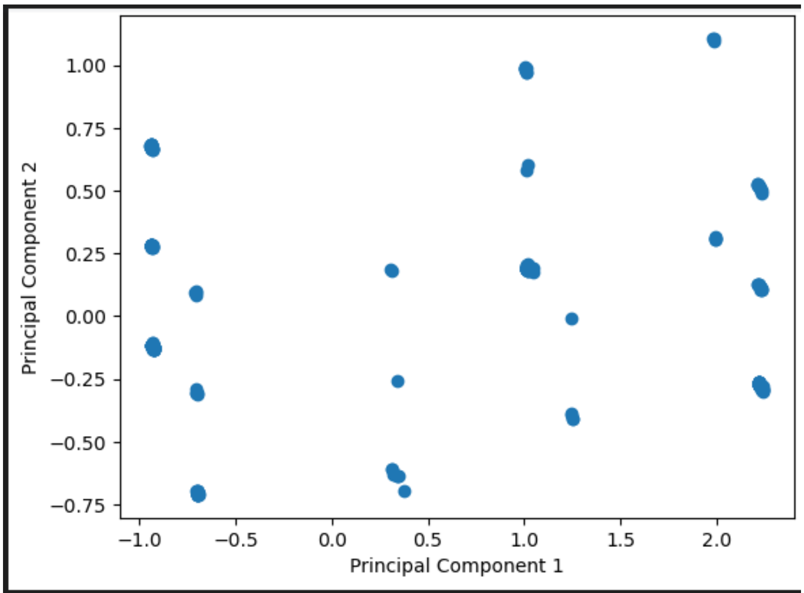
11) Use Scikit Learn to perform PCA :

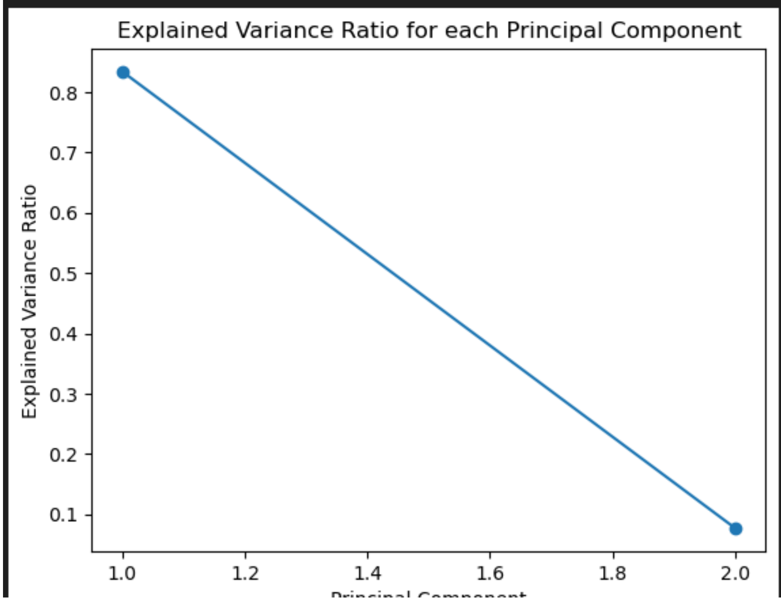



```
Principal Component 1:
Channel: 0.2282064617860256
Detergents_Paper: 0.04658945420484383
Grocery: 0.04396824274389594

Principal Component 2:
Fresh: 0.04210172150120026
Frozen: 0.03694213305561712 4
Delicassen: 0.0015574379684054431
```

12) Plotting the transformed data :





11) Conclusion :

```
**Cluster Analysis**:  By utilizing the K-means clustering algorithm, we were
able to identify and group products with similar attributes into distinct
clusters. These clusters offer valuable insights into the wholesale customers,
enabling us to categorize them into different segments. For instance, we can
identify high-spending customers, low-spending customers, or customers with
specific purchasing patterns. This segmentation empowers businesses to
customize their marketing strategies and offerings, effectively meeting the
unique needs of each customer segment.

**Correlation Analysis**:
The correlation analysis uncovered significant relationships between specific
pairs of variables. Notably, a robust positive correlation was observed
```

```
between the "Grocery" and "Detergents_Paper" variables. This finding indicates
that customers who exhibit higher grocery purchases also tend to acquire more
detergents and paper products. Gaining insights into these correlations
enables us to better understand customer preferences and optimize product
assortments and promotions accordingly. Such knowledge empowers businesses to
make informed decisions that align with customer demands and maximize their
overall effectiveness.

**Principal Components**:
Principal components derived from PCA capture the highest amount of variance
in the data by combining the original features. These components serve as new
variables that are uncorrelated and offer a condensed representation of the
dataset. Analyzing the weights and contributions of the original features to
each principal component allows businesses to understand the fundamental
factors that influence customer purchasing patterns. This insight into the key
drivers of customer behavior can inform strategic decision-making and enable
businesses to tailor their strategies and offerings to meet customer
preferences effectively.

**Variance Explained**:
PCA not only enables dimensionality reduction but also provides insights into
the variance explained by each principal component. This information helps
businesses make informed decisions about the trade-off between reducing
complexity and retaining essential information. By selecting an appropriate
number of principal components that account for a significant portion of the
variance, businesses can effectively balance the reduction in complexity with
the retention of valuable insights. This approach ensures that the resulting
dataset maintains a meaningful representation of the original data while still
reducing the overall dimensionality.
```