Unsupervised Learning Project

By

Zhiyun(Jenny) Liang

Project Objective

Apply unsupervised learning techniques to a real-world data set and use data visualization tools to communicate the insights gained from the analysis.

Project Flow Structure

- Exploratory data analysis
- Preprocessing
- Kmeans clustering, hierarchical clustering
- PCA
- Results and Discussion

Exploratory Data Analysis

- Exploratory data (null value, outlier, etc.)
- Distribution of each variable
- Correlation between the variables

Heat Map

	Correlation Heatmap for Wholesale								
Channel	1	0.062	-0.17	0.46	0.61	-0.2	0.64	0.056	
Region	0.062	1	0.055	0.032	0.0077	-0.021	-0.0015	0.045	
Fresh	-0.17	0.055	1	0.1	-0.012	0.35	-0.1	0.24	
Milk	0.46	0.032	0.1	1	0.73	0.12	0.66	0.41	
Grocery	0.61	0.0077	-0.012	0.73	1	-0.04	0.92	0.21	
Frozen	-0.2	-0.021	0.35	0.12	-0.04	1	-0.13	0.39	
Detergents_Paper	0.64	-0.0015	-0.1	0.66	0.92	-0.13	1	0.069	
Delicassen	0.056	0.045	0.24	0.41	0.21	0.39	0.069	1	
	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	-

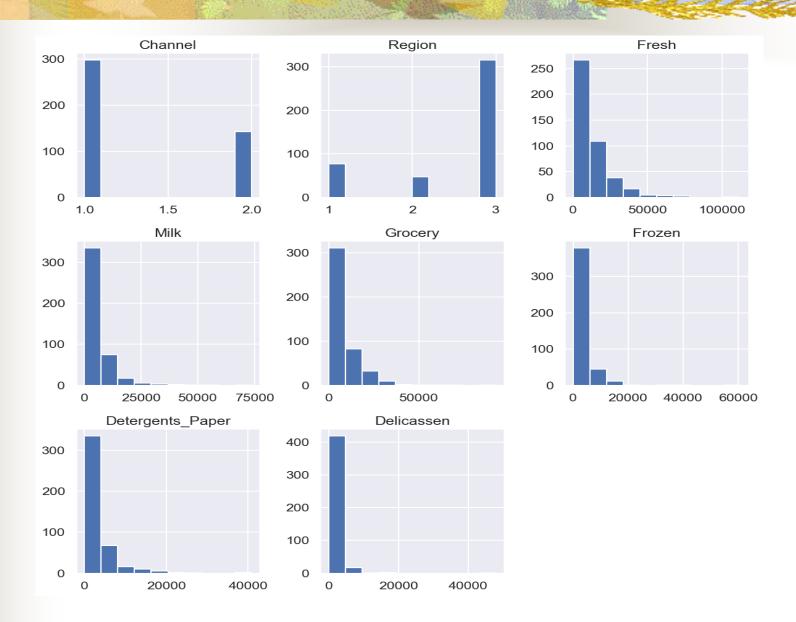
-0.8

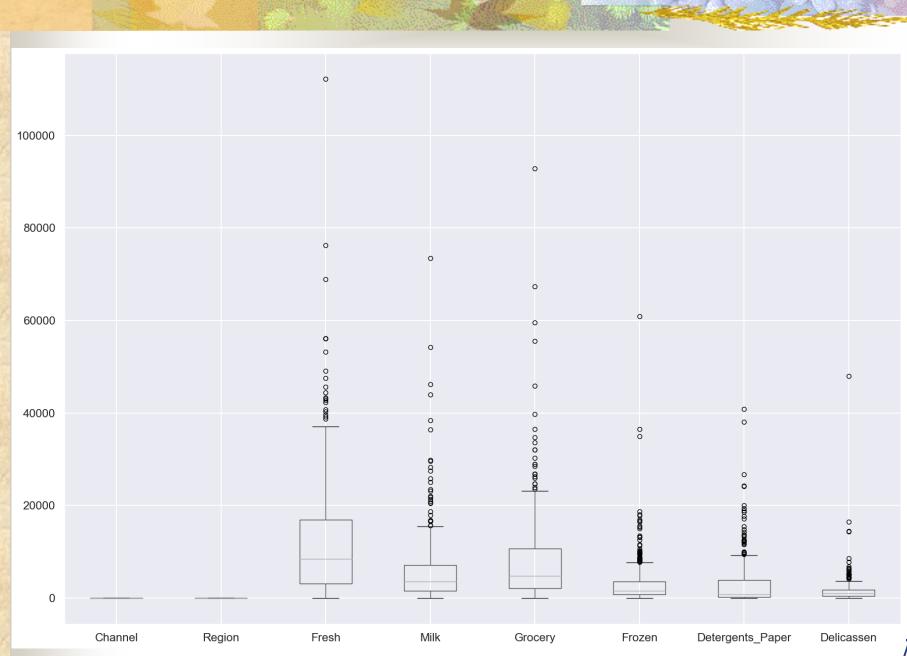
-0.6

-0.4

-0.2

-0.0

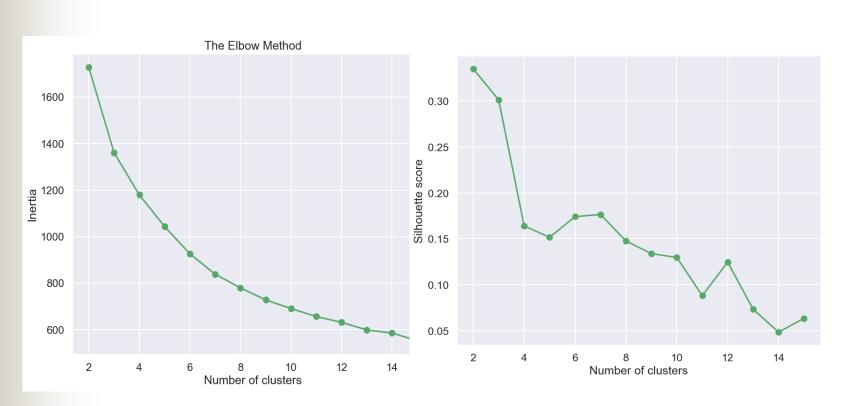




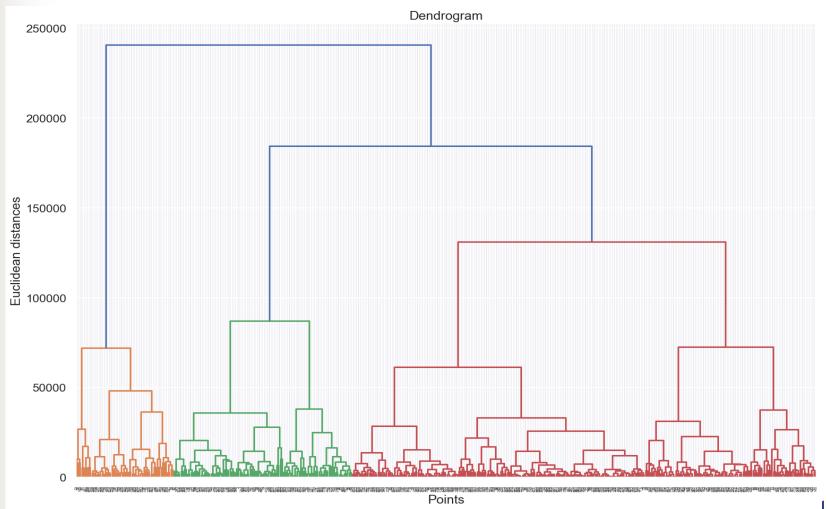
Preprocessing

- Handling outliers
- Scaling and normalization variable

KMeans clustering

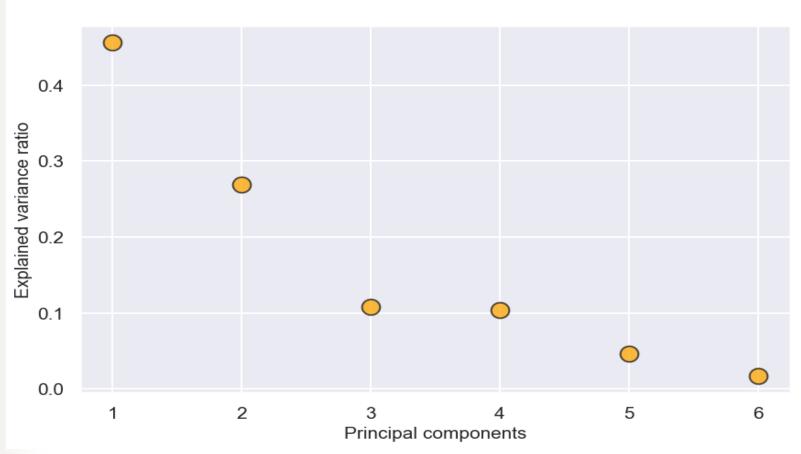


Hierarchical clustering



PCA

Explained variance ratio of the fitted principal component vector



Results and Discussion

- Dataset has 20-40 numbers of outlier for each column
- K-Mean clustering presents 3 or 7 clustering solution
- Hierarchical clustering presents 3 clustering solution
- PCA performs six variables reduction to three variables