ML Project- Unsupervised Learning

Presented by

Mehzabeen Gheesah

DS May 29th

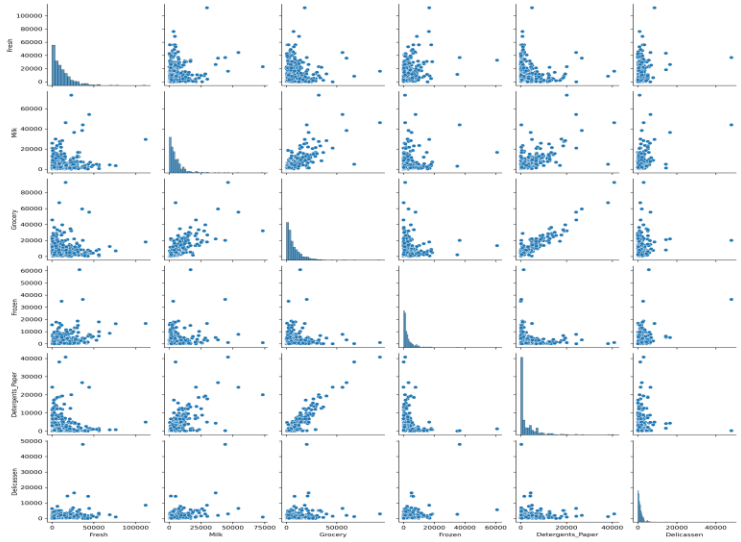
Problem Statement

To identify and understand distinct categories of customers from a wholesale distributor based on their annual spending on diverse product categories. The main goal is to segment the customers in such a way that the distributor can devise targeted marketing and service strategies for each segment, thus improving business performance and customer satisfaction. This is achieved by applying unsupervised learning techniques (like K-means clustering, hierarchical clustering) and dimensionality reduction (Principal Component Analysis) on the spending data

STEPS

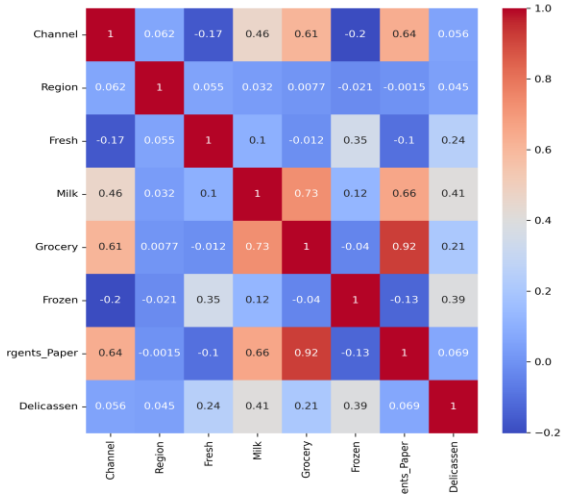
- Part I : EDA - Exploratory Data Analysis & Pre-processing
- Part II - KMeans Clustering
- Part III - Hierarchical Clustering
- Part IV - PCA
- Part V - Conclusion

Dataset Facts

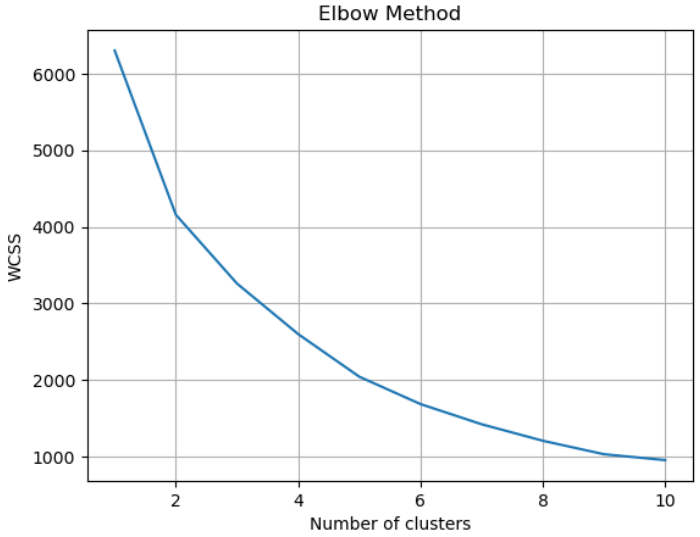


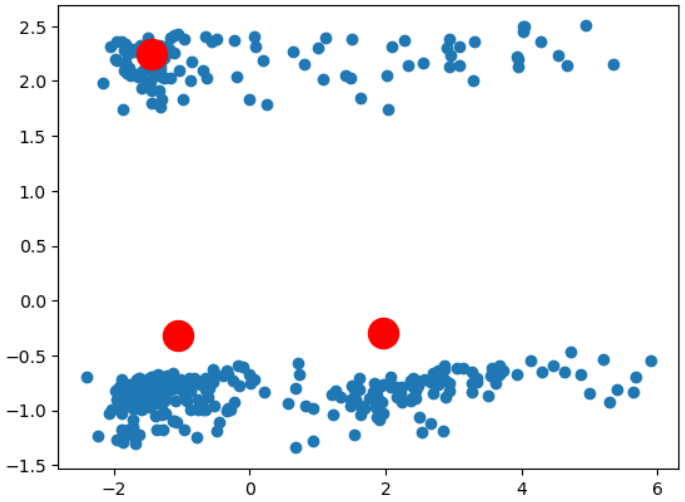
	count	mean	std	min	25%	50%	75%	max
Channel	440.0	1.322727	0.468052	1.0	1.00	1.0	2.00	2.0
Region	440.0	2.543182	0.774272	1.0	2.00	3.0	3.00	3.0
Fresh	440.0	12000.297727	12647.328865	3.0	3127.75	8504.0	16933.75	112151.0
Milk	440.0	5796.265909	7380.377175	55.0	1533.00	3627.0	7190.25	73498.0
Grocery	440.0	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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Channel           440 non-null    int64
 1   Region            440 non-null    int64
 2   Fresh             440 non-null    int64
 3   Milk              440 non-null    int64
 4   Grocery           440 non-null    int64
 5   Frozen            440 non-null    int64
 6   Detergents_Paper  440 non-null    int64
 7   Delicassen        440 non-null    int64
dtypes: int64(8)
memory usage: 27.6 KB
```

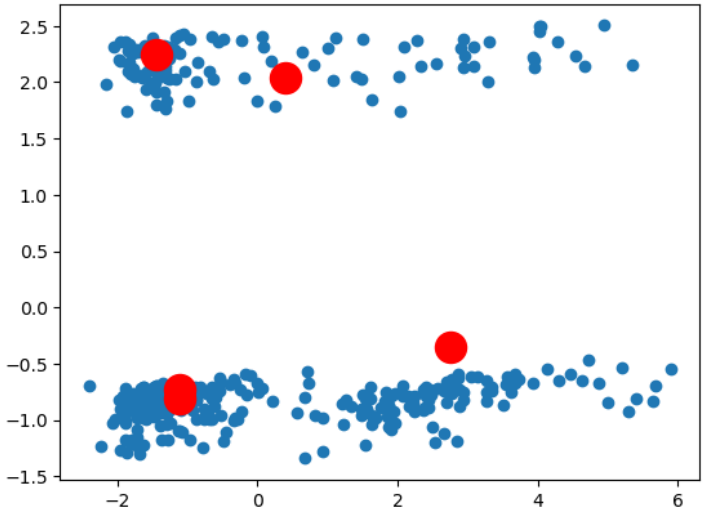


KMeans Clustering

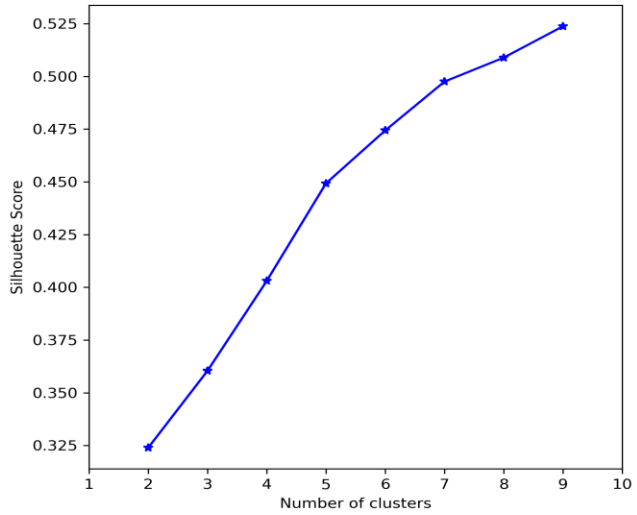




K=3



K=5



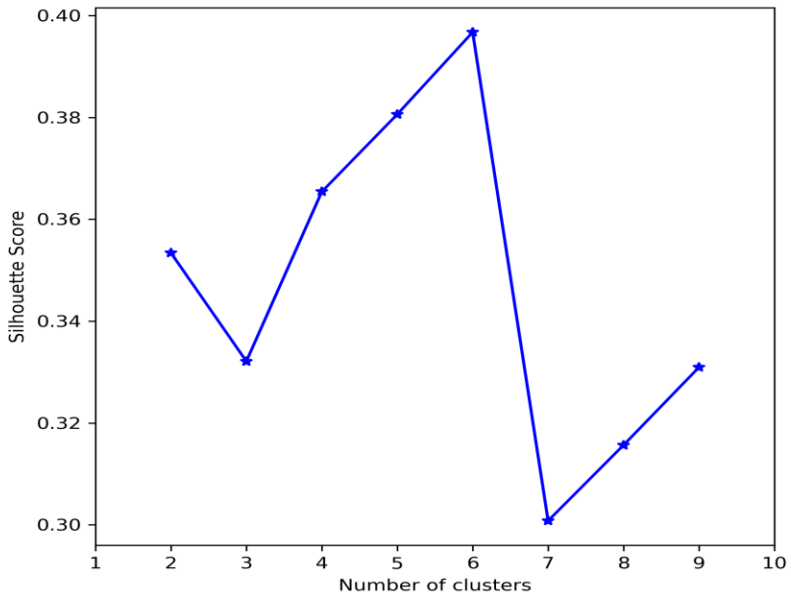
Silhouette Score

Hierarchical Clustering

```
For n_clusters = 2, silhouette score is 0.3423722051476725
For n_clusters = 3, silhouette score is 0.36550220472671796
For n_clusters = 4, silhouette score is 0.4075824421996507
For n_clusters = 5, silhouette score is 0.4641194577827786
For n_clusters = 6, silhouette score is 0.49654176057262844
For n_clusters = 7, silhouette score is 0.5082824554256299
For n_clusters = 8, silhouette score is 0.525798202032903
For n_clusters = 9, silhouette score is 0.5420309004233974
Best number of clusters is 9 with silhouette score of 0.5420309004233974
```

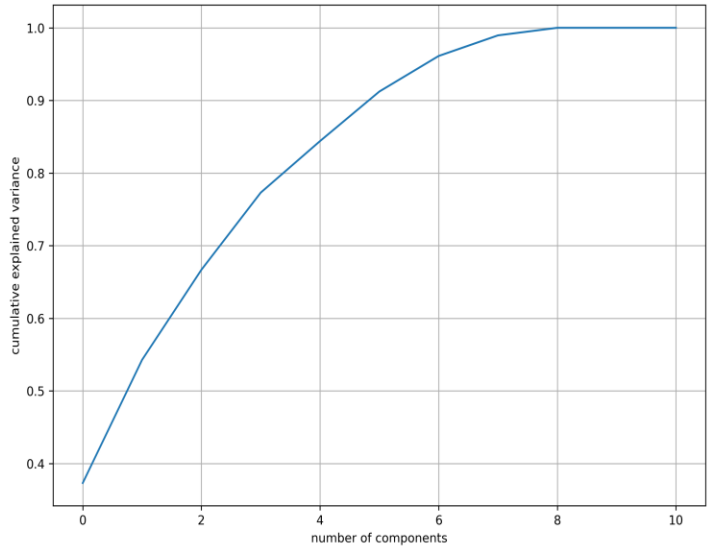
]:

	principal component 1	principal component 2	principal component 3	principal component 4	principal component 5	principal component 6	cluster_labels
0	1.937437	-0.923862	-0.148889	-0.096314	0.727773	-0.297254	1
1	2.220110	-0.846524	0.068570	-0.105262	-0.013138	-0.055382	1
2	2.672210	-1.123797	3.079351	-0.347032	-1.630211	-2.996143	8
3	-1.497146	-0.885827	0.871633	0.252696	-0.536869	0.478758	2
4	1.536685	-1.221745	2.724719	0.212680	0.366659	-1.414877	8



Silhouette Score

PCA



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Region_1	Region_2	Region_3	Channel_1	Channel_2
0	-0.130690	0.377809	0.445794	-0.144222	0.442946	0.131687	-0.045413	0.033196	0.015110	-0.449284	0.449284
1	-0.122059	0.005656	0.029315	0.019399	0.043145	-0.056811	0.550384	0.382615	-0.727143	0.018595	-0.018595
2	0.514069	0.250323	0.065419	0.522079	-0.062001	0.603355	0.132693	-0.077561	-0.057619	0.026462	-0.026462
3	0.154503	-0.072200	-0.018043	0.210891	-0.015963	-0.059572	-0.571919	0.765770	-0.050702	-0.055885	0.055885
4	0.781926	-0.246492	-0.030125	-0.311772	0.011550	-0.311800	0.128540	-0.062046	-0.064887	-0.232151	0.232151
5	-0.035046	-0.004406	0.096530	0.723901	0.155605	-0.618515	0.086351	-0.180345	0.052515	-0.089136	0.089136

Conclusion

1. Key Differentiating Features: The heatmap and Principal Component Analysis (PCA) revealed that 'Channel', 'Grocery', 'Detergents_Paper', and 'Milk' are the features that contribute most significantly to the data's variance. These findings suggest that these features are crucial in distinguishing between different customer groups.

2. Distinct Customer Groups: The application of the KMeans clustering algorithm led to the identification of three distinct customer groups, each exhibiting unique purchasing behaviors. These groups could potentially represent various types of establishments, including restaurants, delis, or grocery stores. Although experimentation with different numbers of clusters was carried out, a three-cluster solution offered the most distinct segmentation. However, a five-cluster solution also presented some insightful groupings.

3. Regional Influence: In addition to the above, The PCA analysis also highlighted the significant impact of the 'Region' feature in differentiating customers. This indicates that customer purchasing behavior varies considerably across different regions.(I did a previous analysis without removing outliers and without hot-encoding Region and Channel and the result was different)

4. Tailored Marketing Strategies: In summary, given an understanding of the different distinct customer segments and their purchasing patterns, businesses can develop a tailored marketting strategies to better target these customers.