PRACTICAL DATA SCIENCE(COSC2670) Assignment 2

Outline

- Data Set
- Research goal
- Analysis & Results
- Conclusion

Data Set

-Wholesale customer data set
-From UC Irvine Machine Learning Repository
-440 customer records from a wholesale distributor
-Contains annual spending (in monetary units m.u) on 6 product categories
-Contains two additional attributes of customer information.

Numerical Attributes (Product Categories)
FRESH
MILK
GROCERY
FROZEN
DETERGENTS
DELICATESSEN

Categorical Attributes	Labels	Frequencies
CHANNEL	Horeca (Hotel/Restaurant/Cafe)	298
	Retail	142
REGION	Lisbon	77
	Oporto	47
	Other	316

-Summary Statistics of the Dataset

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
count	440.0	440.0	440.0	440.0	440.0	440.0
mean	12000.0	5796.0	7951.0	3072.0	2881.0	1525.0
std	12647.0	7380.0	9503.0	4855.0	4768.0	2820.0
min	3.0	55.0	3.0	25.0	3.0	3.0
25%	3128.0	1533.0	2153.0	742.0	257.0	408.0
50%	8504.0	3627.0	4756.0	1526.0	816.0	966.0
75%	16934.0	7190.0	10656.0	3554.0	3922.0	1820.0
max	112151.0	73498.0	92780.0	60869.0	40827.0	47943.0

Research goal

is to analyze and model the wholesale customer data set which will eventually enable the wholesale distributor to best structure their services in order to optimally cater the needs and requirements of different customers.

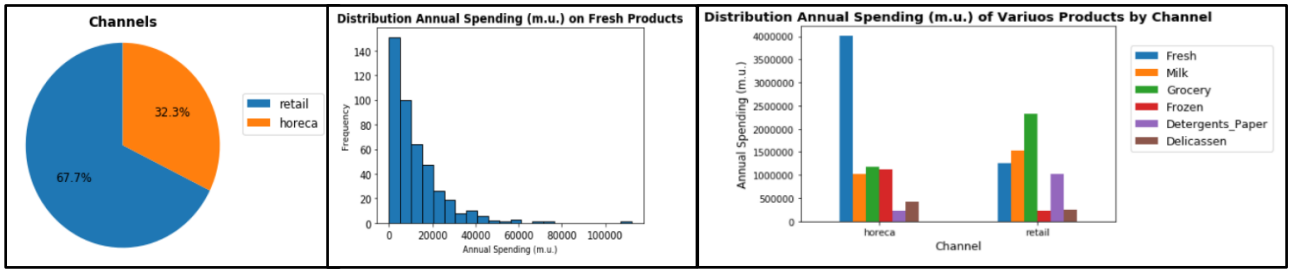
PRACTICAL DATA SCIENCE(COSC2670) Assignment 2

Analysis & Results

-Data cleansing

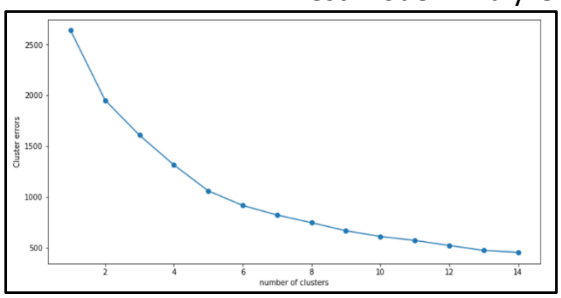
-Data exploration- to get deep understanding of the data and its characteristics.

- Explore each Attributes using graphs
- Explore relationship between each pair of attributes of the Data set
  - Scatter matrix (numerical attributes), Bar chart (Categorical attributes)
  - Strong relationships- Milk-Grocery, Milk-Detergent, Grocery- Detergent



-Data Modeling- Clustering (K-Means, DBSCAN)

- K-Means Clustering
  - Optimal number of k: Using the elbow method (i.e. 2, 3,5)
  - Build models: For different k-values (i.e. for 2,3 and 5)
  - Best model: Analyze cluster mean to determine best k for the data set (i.e. k=5)



Elbow analysis

cluster	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	48777.0	6607.0	6198.0	9463.0	932.0	4435.0
1	4654.0	11296.0	17856.0	1433.0	7794.0	1574.0
2	21200.0	3886.0	5139.0	4120.0	1132.0	1690.0
3	18192.0	35362.0	48052.0	3308.0	23535.0	4461.0
4	6089.0	3253.0	4020.0	2475.0	1182.0	980.0

Cluster mean for K-Means model

- Validate model: Using **Confusion matrix**

(Comparing K-Means clustering results with the true observation label-Channel)

Cluster	Channel	Count
0	Horeca	22
	Retail	2
1	Horeca	7
	Retail	74
2	Horeca	83
	Retail	21
3	Horeca	0
	Retail	10
4	Horeca	186
	Retail	35

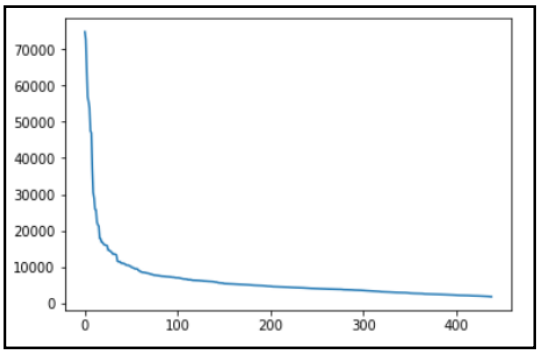
Confusion Matrix for K-Means clustering

Cluster	Observations	
0	High Fresh Spenders	Mostly Horeca Customers
1	High Milk and Grocery Spenders	Mostly Retail Customers
2	Medium Fresh Spenders	Mostly Horeca (but 1/5th of Retail) Customers
3	Overall High Spenders	Mostly Retail Customers
4	Overall Low Spenders	Mostly Horeca Customers

Results of K_Means

PRACTICAL DATA SCIENCE(COSC2670) Assignment 2

- DBSCAN Clustering
  - Optimal eps: Using k Distance Graph (i.e. 11000)
  - Best model-2 clusters



k Distance Graph

cluster1	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
-1	33933.0	22582.0	28345.0	12472.0	10381.0	7316.0
0	10956.0	4997.0	6980.0	2624.0	2524.0	1249.0

Cluster mean for DBSCAN model

  - Validate model: Using **Confusion matrix** (comparing DBSCAN clustering results with the true observation label-Channel)

Cluster	Channel	Count
-1	Horeca	12
	Retail	8
0	Horeca	286
	Retail	134

Confusion Matrix for DBSCAN clustering

Cluster	Observations
-1	High Spender
0	Low spender

Results of DBSCAN

-Compare two clustering Models (K-Means, DBSCAN)

K-Means model	DBSCAN model
5 Clusters	2 Clusters
reasonably be validated against the true observations	Difficult to be validated against the true observations
Sufficient to draw detail insight of the dataset	not sufficient to draw detail insight of the dataset

Conclusion

-DBSAN model is underperforming.

-K-Means clustering model clearly identified five cluster groups of the costumers based on their annual spending on six products types

- the identified customer market segments would be quite critical for the wholesale distributor to structure its business to best cater the customer specific requirements and needs.

-The results from this study can be used for deploying any direct marketing and distribution strategies to promote the sales.