RMIT UNIVERSITY MELBOURNE
COSC2670 - PRACTICAL DATA SCIENCE

Assignment 2

(Data Modelling for Wholesale Customers)

Udeshika Dissanayake (S3400652)

Udeshika.dissanayake@student.rmit.edu.au

21-05-2018

Table of Contents

Executive Summary

The objective of this assignment is to analyze and model a customer data set that contains annual expenditure data on various product categories from a wholesale distributor. The dataset has been obtained from UC Irvine Machine Learning Repository and it contains annual spending data for various product types recorded in monetary units (m.u.) of 440 customers. In detail, it contains the expenditure data on six different product categories: Fresh, Milk, Grocery, Frozen, Detergents Paper, and Delicatessen. Also, it has two auxiliary labels (i.e Channel and Region) that can be used to validate the model by treating them as true observations. A better insight of data through this analysis would enable the wholesale distributor to best custom their services in order to optimally cater the needs and requirements of different customers.

In this exercise, two different clustering methods (K-Means and DBSCAN) have been used to model the data and selected the better model by comparing the model results against the true observations. Confusing Matrices have been constructed for each clustering method and for different parameters to select the best clustering method and it's optimal parameters.

The best performed clustering model in this study identified five groups of customers based on the similarities of their annual expenditure on aforementioned product types. Identifying such similarities of customers in terms of their annual spending would enable the wholesale distribute obtain additional knowledge on customers which was not available to prior to this study. The interpretation of the insight has been done using the results of this cluster model and they are discussed in this report. It is recommended that the wholesale distributor to use the knowledge and the insight presented in this study to be used in structuring its marketing and service strategies to enhance the overall sales.

Introduction

Market Segmentation is vital concept of business marker to identify any similarities of large and broad set of customers and to cater for their needs in much optimal and efficient way. More importantly, the insight of customer and market segmentation could promisingly be used in for deploying efficient and direct marketing strategies.

In this exercise, the dataset chosen is from a wholesale distributor on his customers' annual expenditure figures on six different product categories. The annual expenditures are recorded in monetary units (m.u.) and the data types are numerical. The corresponding attributes in the dataset are,

- Fresh: Annual spending on fresh products
- Milk: Annual spending on milk products
- Grocery: Annual spending on grocery products
- Frozen: Annual spending on frozen products
- Detergent Paper: Annual spending on detergents and paper products
- Delicatessen: Annual spending on delicatessen products

The objective of the exercise is to identify the similarities of different customers in regards to their annual expenditures on above product groups. Two attributes of the dataset identify the customer groups in very high level. Those are,

- Channel – Horeca (Hotel/Restaurant/Café) or Retail (Normal Grocery)
- Region – Lisbon, Oporto, or Other

However, the high-level groupings identified by above two attributes do not necessarily provides sufficient insight on the customers, therefore the wholesale distributer has limited or no knowledge to structure his sales or service strategies.

This report focuses on the modeling of aforementioned dataset using two clustering techniques (K-Means and DBSCAN). Each of the implementation steps are briefly outlined in the report before presenting the detail analysis of the modeled data and recommendations.

Methodology

Methodology for Data Retrieving

This modeling task used a dataset from [UC Irvine Machine Learning Repository]. The dataset contains annual expenditure data of 440 customers on six different product types recorded in monetary units (m.u.). Firstly, the dataset was loaded and then the labels of the two attributes, 'Channel' and 'Region' were replaced by descriptive label instead of original numerical labels. For an example, the original Channel data 1 and 2, representing Horeca and Retail, respectively were replaced by descriptive labels of horeca and retail. Similarly, the numerical labels of the Region have been replaced by descriptive labels of Lisbon, Oporto, or Others. The data types of the changed attributes were changed from numerical to categorical.

Secondly, the data cleansing has been undertaken by checking and fixing the data types, typos, null values, extra white spaces, and capital letter mismatches. The outlier of each numerical attributes was analyzed using box diagrams as follows:

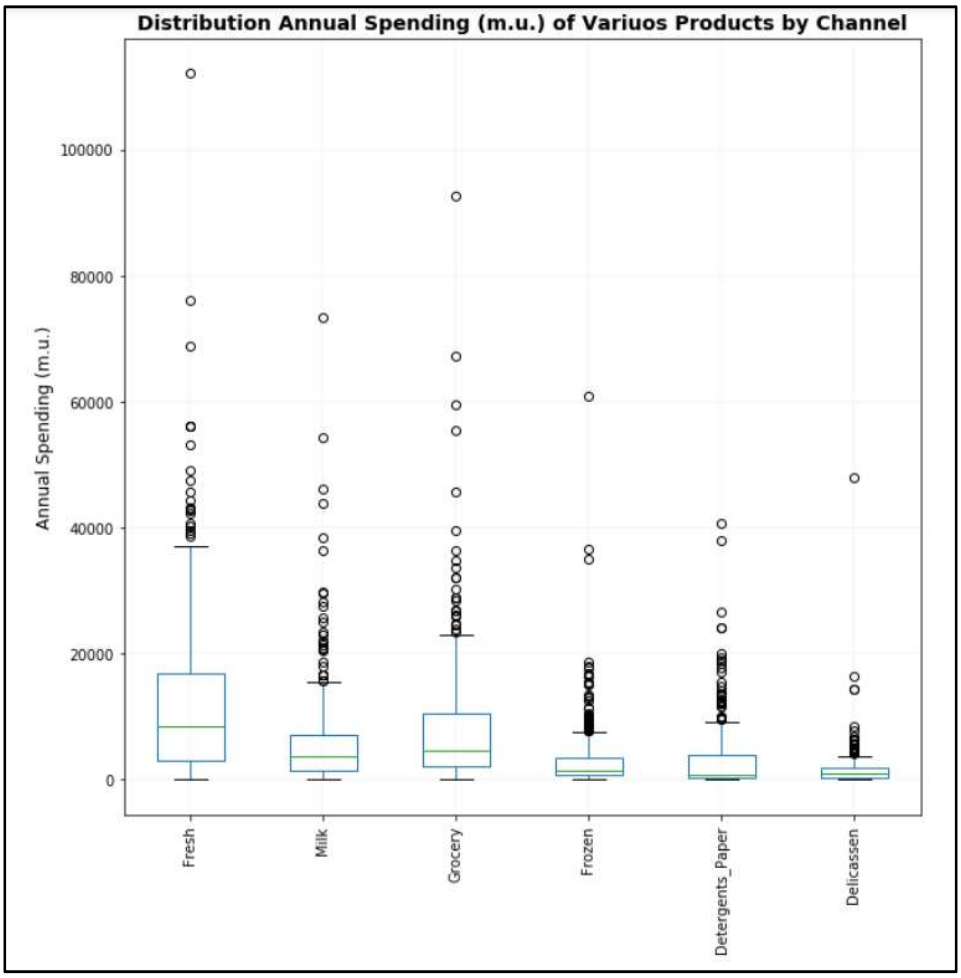


Figure 1: Distribution of the Dataset-Box plot

Despite a significantly high variances between the minimum and maximum value were observed for each data attribute, the outlier removal was not performed as the annual spending figures are real world vales, therefore it could be any real positive value. For an example, annual spending on Fresh products vary from 3 to 112151. Since these are real values on spending, the far end spending data has not been considered as outlier and not removed from the dataset.

The set of attributes after cleansing and transforming to more suitable data types looks as follows:

Attribute Information of Wholesale Customers Data Set

- FRESH: annual spending (m.u.) on fresh products - Numerical
- MILK: annual spending (m.u.) on milk products - Numerical
- GROCERY: annual spending (m.u.) on grocery products - Numerical
- FROZEN: annual spending (m.u.) on frozen products - Numerical
- DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products - Numerical
- DELICATESSEN: annual spending (m.u.) on and delicatessen products - Numerical
- CHANNEL: customer sales Channel - Horeca (Hotel/Restaurant/Cafe) or Retail channel - Categorical
- REGION: customer sales Region - Lisbon, Oporto or Other - Categorical

Methodology for Data Exploration

Each column of the dataset has been explored carefully to get deep understanding of the data and its characteristics. Histograms have been drawn to identify the descriptive statistics and distribution information of the numerical attributes as can be seen in the Results section. Similarly, proportional plots have been drawn for each categorical attribute. And also, for further analysis the metadata for the dataset has been defined.

In order to identify possible relationship between each pair of attributes, scatter matrix has been plotted for numerical attributes and bar charts have been plotted for categorical data. All these plots are presented in the result section.

Methodology for Data Modelling

The modeling of the dataset was carried out using clustering technique in this exercise. Two famous clustering models were selected to carry out the analysis: K-Means Clustering and DBSCAN Clustering.

K-Means Clustering

As the first step of K-Means clustering columns Channel and Region have been dropped as those fields are not useful in clustering. Determining the optimal number of clusters (k) in the data is most important step in K-Means Clustering. In order to find optimal k, the elbow method has been used in this study.

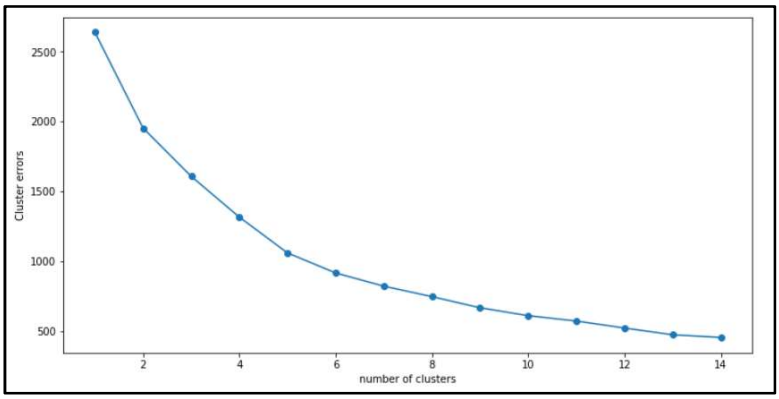


Figure 2: Elbow Analysis

As can be seen in the elbow diagram, after values 2, 3, and 5, the curve remains less changing. Therefore, it could be concluded stating 2, 3, or 5 as the optimal number of clusters for this dataset. All of these three K values have been considered and each scenario has been analyzed to obtain the optimal model.

Finally, the results from the clustering analysis have been compared with the true observation labels by constructing the Confusion matrix. The Channel information has been chosen as the true observation labels for this dataset.

DBSCAN Clustering

Determine the optimal eps in the data is most important step in DBSCAN Clustering. In order to find optimal eps, k Distance Graph has been generated.

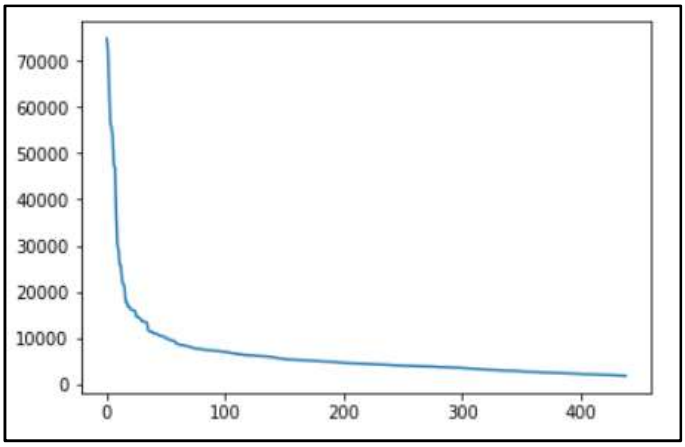


Figure 3: k Distance Graph for finding eps

Concentrating the drastic change in the k Distance Graph, 11000 can be selected as the optimal eps. Data set has been modelled by considering above obtained eps and analyzed cluster mean to get the optimal model.

Finally, the results from the clustering analysis have been compared with the true observation labels by constructing the Confusion matrix.

Results

Data Exploration

Before modeling the dataset using the clustering techniques, the dataset has been carefully explored to get the abstract statistical details of each attribute and to identify any possible relationship between any attribute pairs.

Below table shows the 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

Table 1: Summary Statistics of the Data in monetary units

As can be noted from the above summary table, in total there are 440 customers in the dataset. It is evident that the Fresh food has the highest mean out of all six product types. This means on average; the Fresh products are the type that customers are spending heavily on. On the other hand, the Fresh products do have the largest standard deviation, which means the variation of the customer spending on Fresh products is are high as well. This leads to a requirement to have a more complex modeling of the dataset to obtain statistically significant insight.

The skewness of each numerical attributes was calculated and presented in below table:

Fresh	2.56
Milk	4.05
Grocery	3.59
Frozen	5.91
Detergents_Paper	3.63
Delicassen	11.15

Table 2: Skewness of attributes

According the skewness figures, the Delicassen has the highest skewness while Fresh products has the lowest skewness out of all six attributes.

The frequencies for Categorical attributes have been tabulated below:

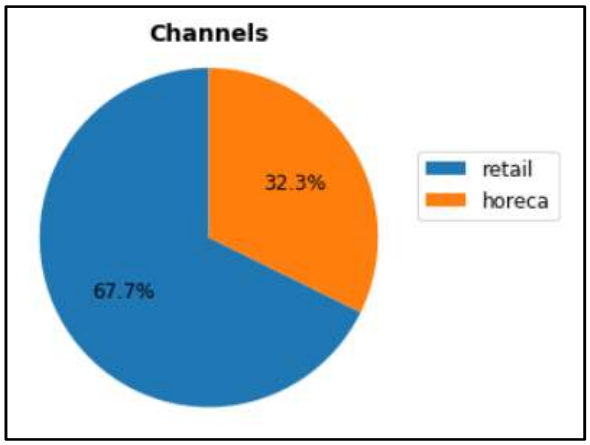
Region	Lisbon	77
	Oporto	47
	Other	316
	Total	**440**
Channel	Horeca	298
	Retail	142
	Total	**440**

It is evident that the Horeca customer count are significant high (as twice as much) compared to the Retail counterparts.

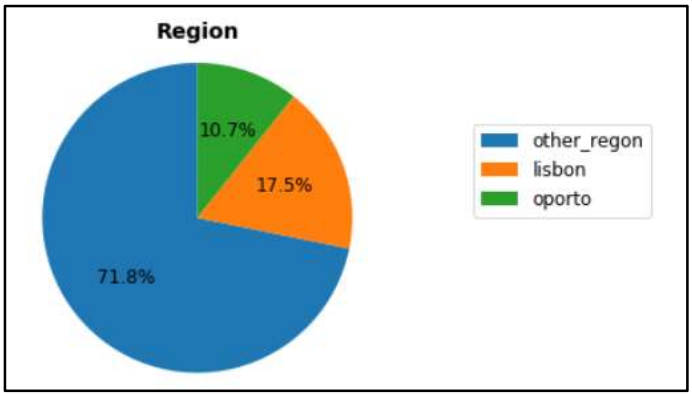
Proportional charts and Histograms have been plotted for each categorical and numerical attribute, respectively.

Visualization for each column by producing graphs

The proportions of customer types (horeca vs retail) are represented in below pie chart. As explained previously, the horeca customers in the dataset are twice as much as the retail customers.



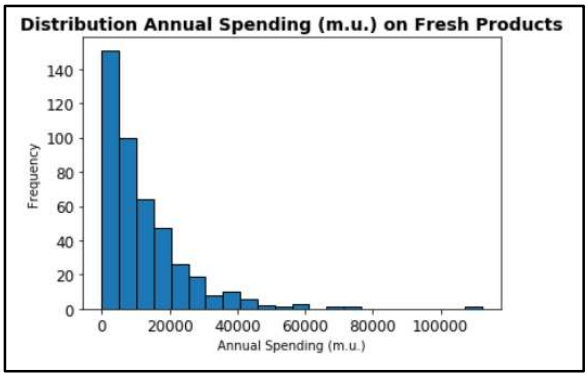
To visualize the proportions of the customer by regions, below pie chart has been drawn. It is evident that the majority of the customers are neither from Lisbon nor Oporto. Therefore, Region data is not sufficient to use as true observation to validate the clustering model.

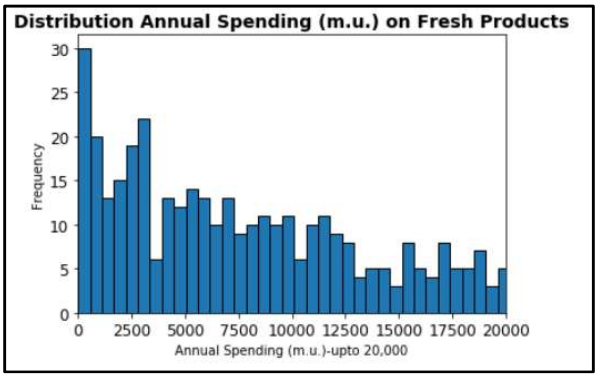


Each of the numerical attributes has been explored by plotting histograms as shown in the below section.

Exploration of Annual spending on Fresh Products

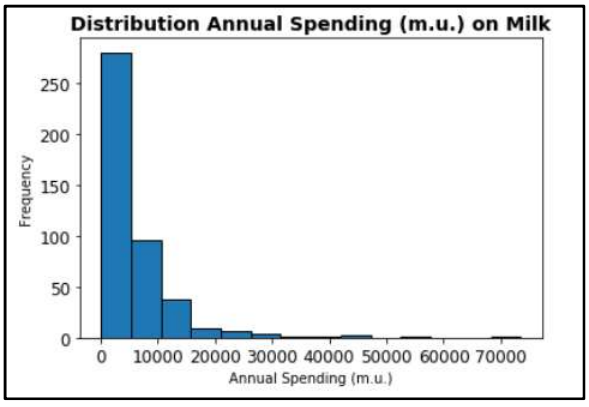
The left graph shows the histogram of the customers' annual spending on Fresh product for the full range. The bin sizes chosen here are comparatively broad to capture the full range (i.e. from 3 to 112151). As can be seen from the left plot the majority of the distribution is spanning over fairly narrower range, a histogram for a smaller range (only up to 20000) with narrower bin size has been plotted in the right-hand side. The histogram in the right for reduced range does show more details in regards to the customer expenditure distribution for the fresh products.

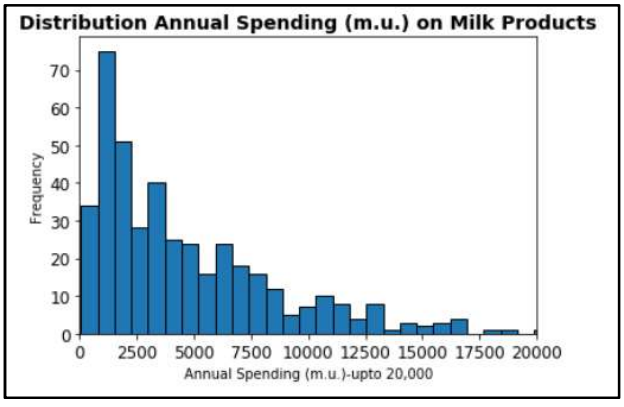




Exploration of Annual spending on Milk

The left graph shows the histogram of the customers' annual spending on Milk product for the full range. The bin sizes chosen here are comparatively broad to capture the full range (i.e. from 55 to 73498). As can be seen from the left plot the majority of the distribution is spanning over fairly narrower range; therefore, a histogram for a smaller range (only up to 20000) with narrower bin size has been plotted in the right-hand side. The histogram in the right for reduced range does show more details in regards to the customer expenditure distribution for the milk products.

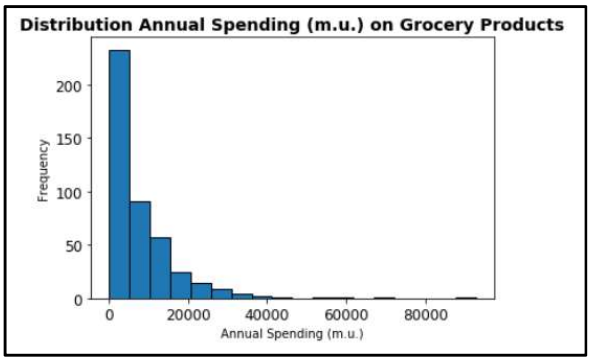


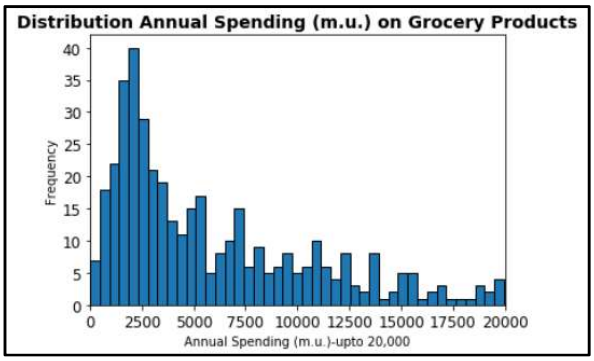


Exploration of Annual spending on Grocery

The left graph shows the histogram of the customers' annual spending on Grocery product for the full range. The bin sizes chosen here are comparatively broad to capture the full range (i.e. from 3 to 92780). As can be seen from the left plot the majority of the distribution is spanning over fairly narrower range; therefore, a histogram for a smaller range (only up to 20000) with narrower bin size has been plotted in the right-hand

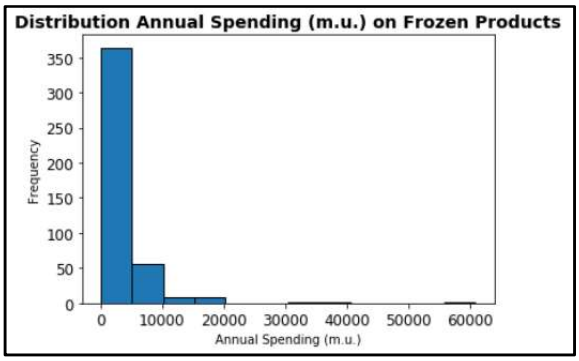
side. The histogram in the right for reduced range clearly shows peak around 2500 in opposed to the histogram in the right.

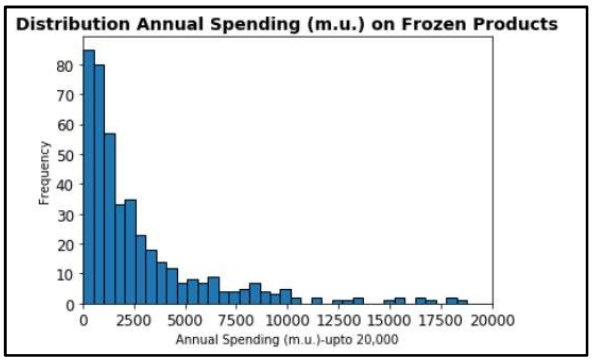




Exploration of Annual spending on Frozen Products

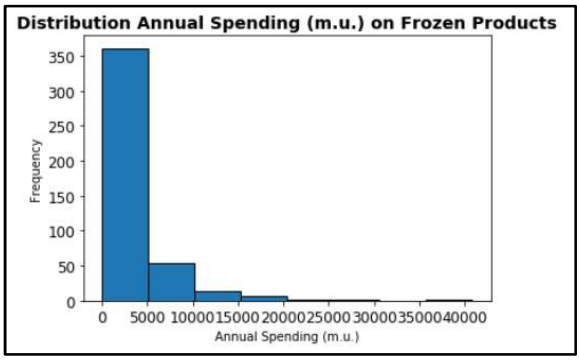
The left graph shows the histogram of the customers' annual spending on Frozen product for the full range. The bin sizes chosen here are comparatively broad to capture the full range (i.e. from 25 to 60869). As can be seen from the left plot the majority of the distribution is spanning over fairly narrower range; therefore, a histogram for a smaller range (only up to 20000) with narrower bin size has been plotted in the right-hand side. The histogram in the right for reduced range does show more details in regards to the customer expenditure distribution for the frozen products.

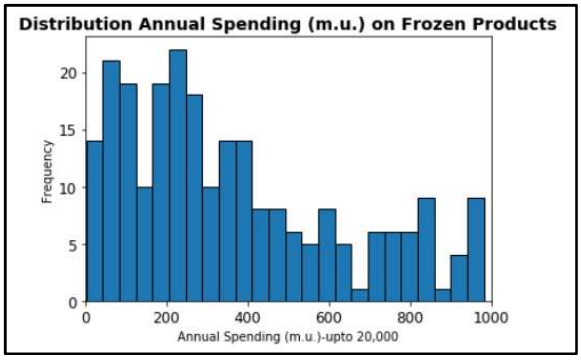




Exploration of Annual spending on Detergents_Paper

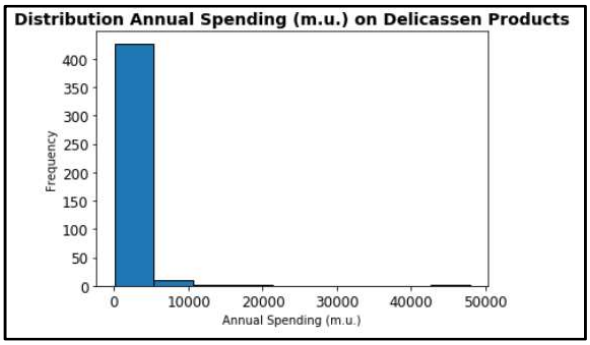
The left graph shows the histogram of the customers' annual spending on Detergent paper for the full range. The bin sizes chosen here are comparatively broad to capture the full range (i.e. from 3 to 40827). As can be seen from the left plot the majority of the distribution is spanning over fairly narrower range; therefore, a histogram for a smaller range (only up to 1000) with narrower bin size has been plotted in the right-hand side. The histogram in the right for reduced range does show more details in regards to the customer expenditure distribution for the detergent paper.

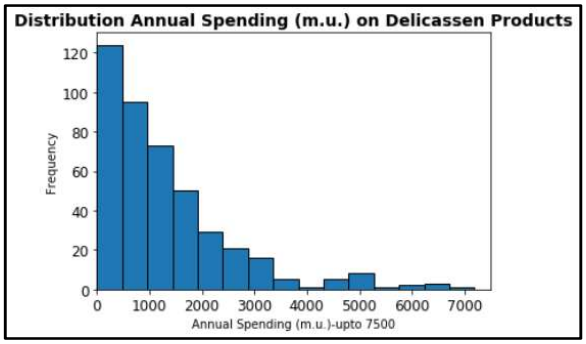




Exploration of Annual spending on Delicassen

The left graph shows the histogram of the customers' annual spending on Delicassen for the full range. The bin sizes chosen here are comparatively broad to capture the full range (i.e. from 3 to 47943). As can be seen from the left plot the majority of the distribution is spanning over fairly narrower range; therefore, a histogram for a smaller range (only up to 7000) with narrower bin size has been plotted in the right-hand side. The histogram in the right for reduced range does show more details in regards to the customer expenditure distribution for the delicassen products.



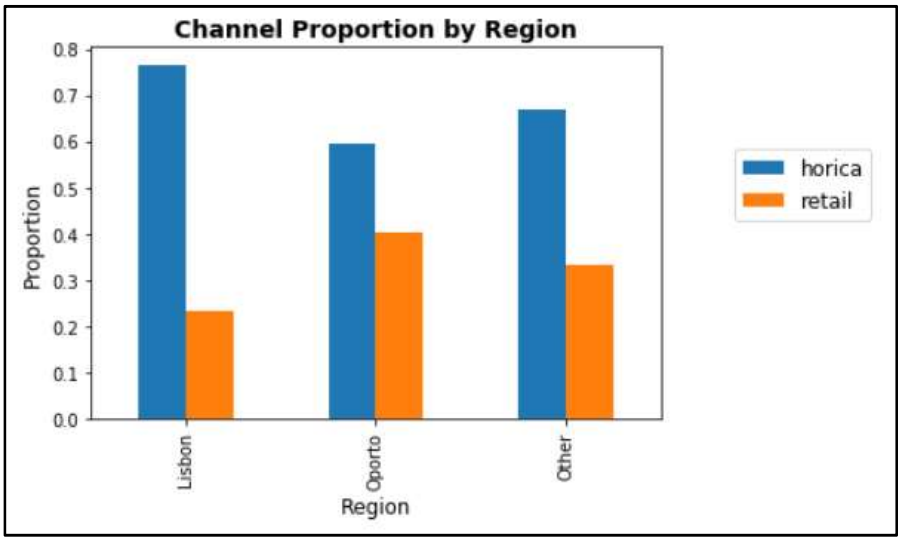


Explore relationship between attributes

Explore relationship between Region and Channel

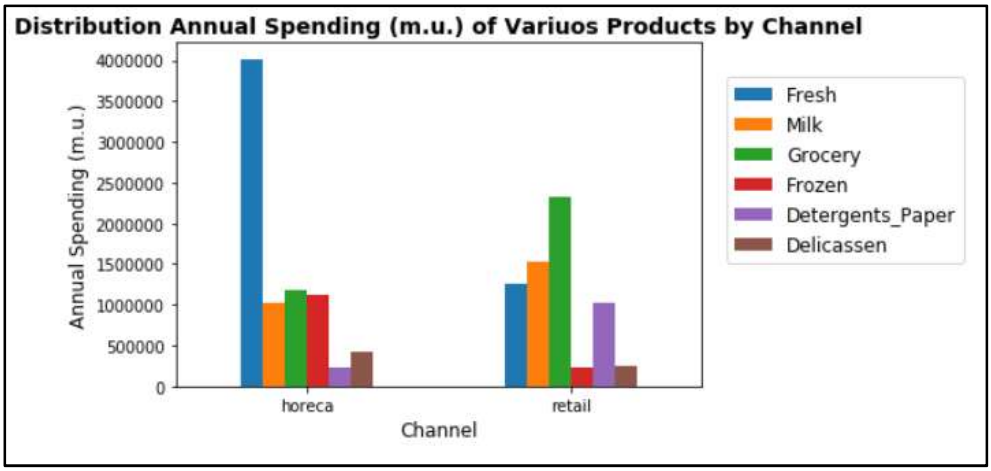
In order to explore the relationship between categorical attributes, following proportional bar chart has been plotted.

The x-axis represents the label Region while y-axis represents the proportions of customers under each category. The sub grouping of customers based on the Channel label (i.e. horeca vs retail) is distinguished by two colors in the figure.
According to the figure, Lisbon got the highest proportion of horeca customers and the lowers proportion of retail customers compared to other regions.



Explore relationship between Channel and products

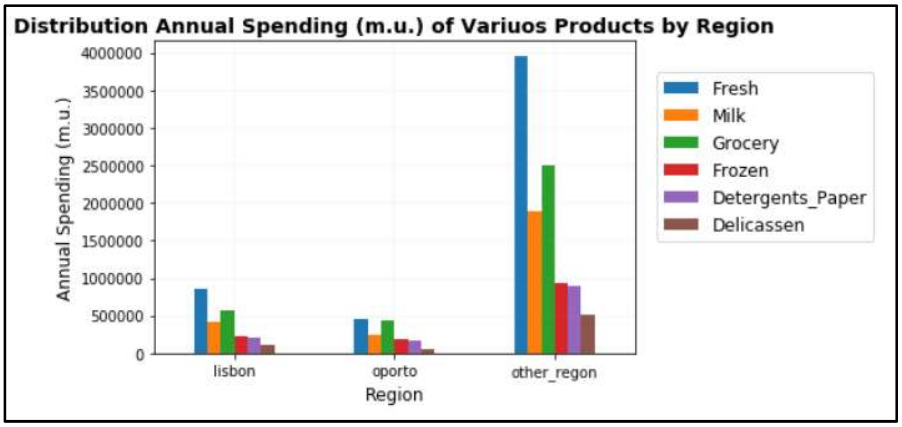
In order to explore the relationship between channel vs products, following proportional bar chart has been plotted. The x-axis represents the label Channel while y-axis represents the proportions of customers under each category. The sub grouping of customers based on the products is represented by different colors as shown in the legend.



The plot prominently shows a relatively higher amount of expenditure on Fresh products by horeca customers compared to retail customers. Also, it is clearly evident within horeca customers that significantly largest amount of expenditure is on Fresh products. The least expenditure by horeca customers is for Detergents papers. By a similar observation for retail customers, it is not difficult to notice that the largest amount of spending is for Grocery while the least amount of spending is for Frozen products.

Explore relationship between Region and products

In order to explore the relationship between regions vs products, following proportional bar chart has been plotted. The x-axis represents the label Regions while y-axis represents the proportions of customers under each category. The sub grouping of customers based on the products is represented by different colors as shown in the legend.



Scatter Matrix for numerical columns

Below scatter matrix shows the relationships between each pair of numerical attributes (i.e. spending types – Fresh, Milk, Grocery, etc.). Any healthy relationships can easily be found by this scatter plot by identifying any linear dependencies of two variables.

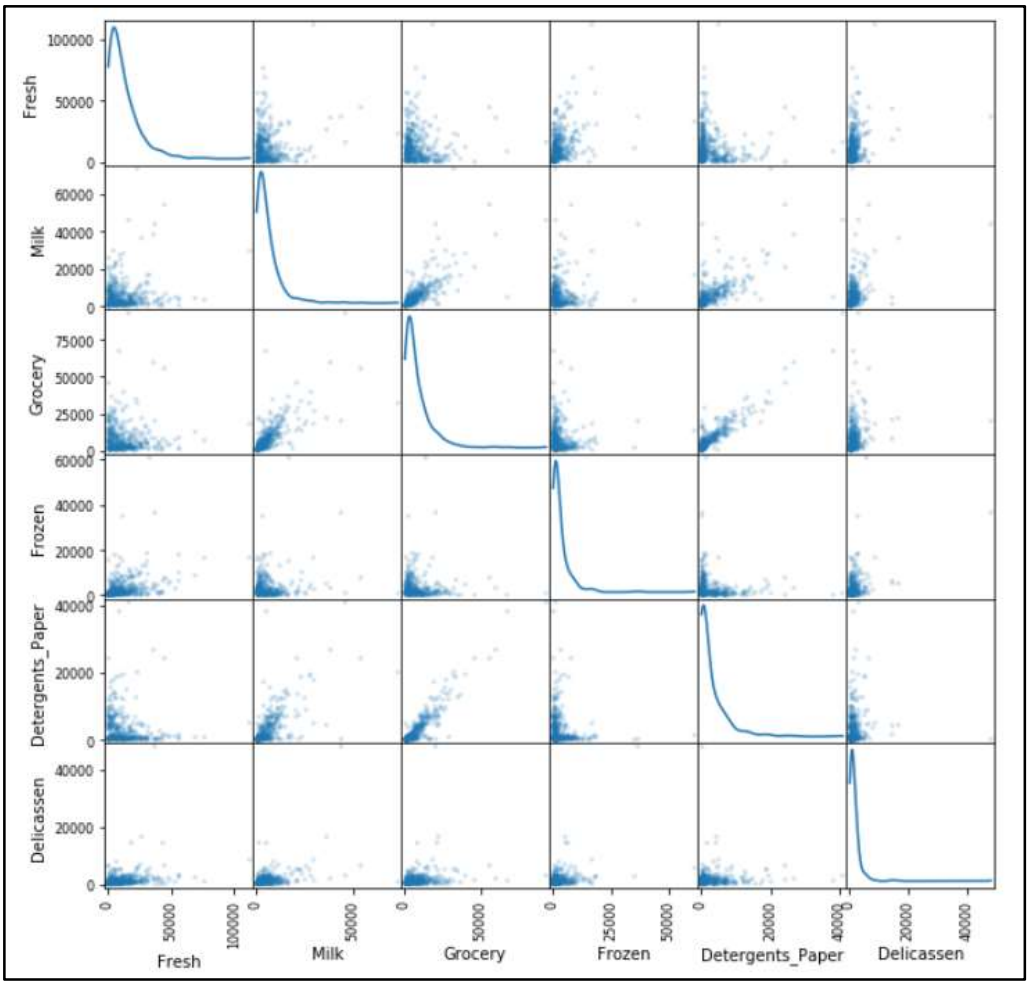
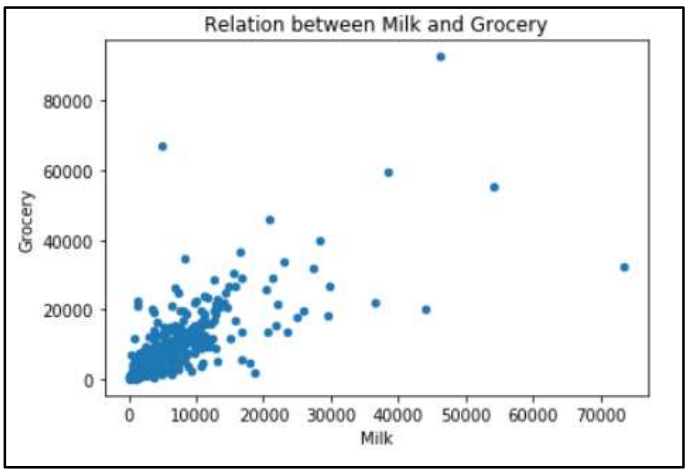
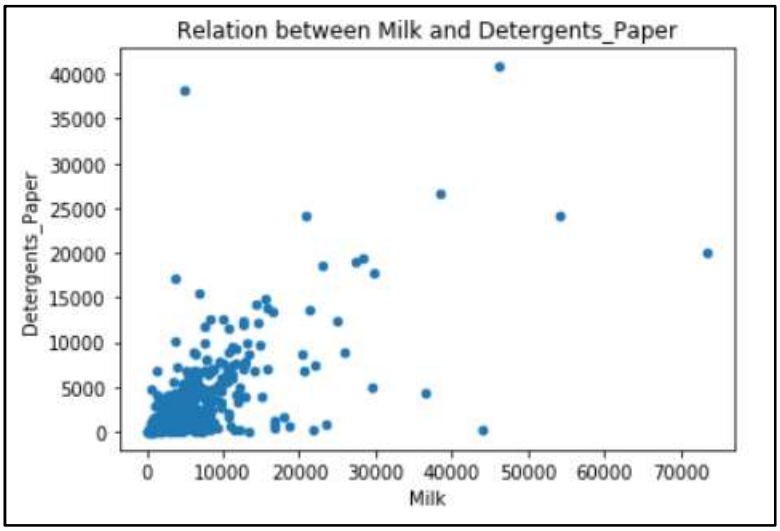


Figure 4: Scatter Matrix of the Dataset

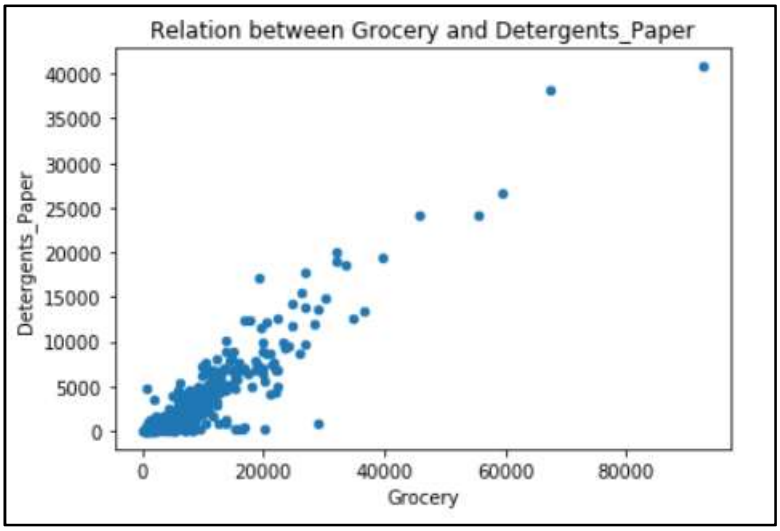
For an example, it is clearly evident a linear type relationship between Milk and Grocery as both attributes tends to have "y = mx" type of connection. However, it should also be noted that there are a number of outlier data points in this relationship plot that do not adhere to the linear relationship.



Similar linear relationship is being hold by Milk against Detergent_papers as can be seen from below enlarged relationship plot. It is also worth noting that there are number of outlier data points in the plot that do not fall in to the linear relationship.



Lastly, it was observed that Detergent_paper and Grocery hold quite healthy linear relationship as can be seen in the below plot. Compared the number of outliers noted in above two relationships, this plot has significantly less number of outlier data points. That further suggest the respective relationship is comparatively more linear.



Data Modelling

As mentioned in the methodology section, the modeling of the dataset was carried out using clustering technique. Two popular clustering models were selected to carry out the analysis: K-Means Clustering and DBSCAN Clustering.

Clustering K-Means

Using the elbow method (refer to methodology section) the possible optimal numbers of clusters (k) in the dataset were observed visually as 2, 3, or 5. Therefore, the training of the dataset was carried out to all of these possible optimal K numbers independently.

Below cluster mean table corresponds to K = 5 training dataset.

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

Table 3: Cluster mean for k=5

Similarly, cluster mean tables were derived by training the dataset for K = 4, 3, and 2, respectively. The respective tables are shown below:

cluster	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	5473.0	4108.0	5593.0	2255.0	1984.0	1051.0
1	49331.0	6823.0	6339.0	9666.0	951.0	4558.0
2	20801.0	3794.0	5013.0	4023.0	1120.0	1648.0
3	8150.0	18716.0	27757.0	2035.0	12523.0	2282.0

Table 4: Cluster mean for k=4

cluster	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	36156.0	6124.0	6367.0	6811.0	1050.0	3090.0
1	7752.0	17911.0	27038.0	1971.0	12105.0	2186.0
2	8342.0	3780.0	5152.0	2577.0	1721.0	1137.0

Table 5: Cluster mean for k=3

cluster	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	35401.0	9514.0	10346.0	6463.0	2933.0	3317.0
1	7944.0	5152.0	7536.0	2484.0	2873.0	1214.0

Table 6: Cluster mean for k=2

By carefully observing the cluster mean values, the clustering for K = 2, 3, 4 could be safely ignored for further analysis. For an example, at K = 4, both cluster group 1 and 2 seems to similar characteristics: both represent high Fresh spenders. Therefore, it is assumed K = 4 is not the optimal K number for this analysis. Considering similar misclassification arguments K = 2 and 3 were also ignored.

For further investigation K = 5 would be selected. By deriving a confusion matrix, the clustering results have been compared against the true observation (Channel data).

Confusion Matrix

The confusion matrix for K = 5 is shown below:

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

Table 7: Confusion Matrix for K=5 K-Means clustering

The label values under the categorical attribute of 'Channel' are being considered as the true observation to validate the clustering model. Since, the attribute 'Channel' has only two labels (Horeca or Retail), each cluster group has been counted for Horeca and Retail as can be seen in the above Confusion Matrix. It is evident from the count values that the results obtained from K = 5 clustering, do have real world interpretation in regards to their channel grouping.

Only the Cluster group 2 has around 21/83 (i.e. ~20%) mix grouping of Horeca and Retail customers. The clustering group 4 has about ~15% of mix grouping of Horeca and Retail customers. All the other clustering group has better that ~8% of mix grouping. This is obviously due to the availability of limited true observation to validate the clustering results.

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

By considering the cluster mean values again for each product types for K = 5 clustering, and the target counts in the confusion matrix, below definition to each of the clusters could be provided.

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

It is quite evident that the Horeca customers are spending heavily on Fresh products (Cluster 0 and 2) or in overall, they are low spenders (Cluster 4) compared to Retail customers. On the other hand, Retail customers are overall high spenders (Cluster 3) or spending heavily on Milk and Grocery (Cluster 1).

Clustering DBSCAN

As mentioned in the Methodology section, determining the 'esp - Neighborhood' value in the dataset is the most important step in DBSCAN clustering technique. Considering the drastic change in the k Distance Graph, 11000 was selected as the optimal eps. Data set has been modelled by considering above obtained eps and analyzed cluster mean to get the optimal model.

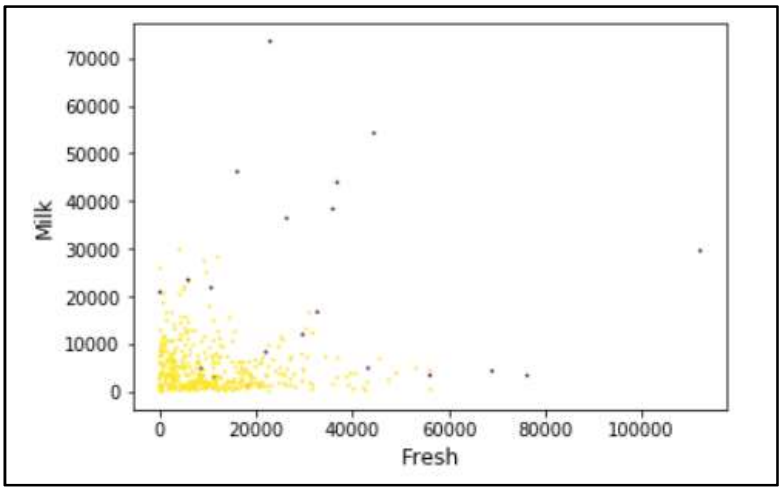


Figure 5: DBSCAN cluster results

As can be seen from below cluster mean table, DBSAN identified only two clusters for eps = 11000. Primally, the two clusters represent high spenders (Cluster -1) and low spenders (Cluster 0)

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

Table 8: DBSCAN Cluster mean

The clustering results obtained from this method was compared against the true observation label (í.e. 'Channel') to validate the results. Below is the derived confusion matrix. Its is evident that DBSAN clustering

results is difficult to be validated with available true observations of 'Channel' as there are quite a lot of mix-classifications.

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

Table 9: Confusion Matrix for DBSCAN clustering

However, considering the cluster mean value table, below descriptions could be provided for two cluster groups:

- Cluster -1 – High Spenders
- Cluster 0 – Low Spenders

Compare the Clustering Models

The two clustering models carried out in this study are K-Means and DBSCAN and they presented 5 and 2 cluster groups, respectively. It is clearly evident that 2 cluster group presented by DBSAN is not sufficient to draw detail insight of the dataset in comparison to 5 cluster group presented by K- Mean group. Also, as described in the above section, the results from K-Means method at K = 5 could reasonably be validated against the true observations. It should also be noted that the K-Mean classification model for K = 2, 3, and 4 were not relatively weak (compared to K = 5 model) in terms of validated against the true observation using the respective confusion matrices.

Therefore, it is recommended to choose the model presented by K-Means method as the most suitable clustering data model for this study in oppose to the DBSCAN.

The reason for limited cluster groups presented by DBSCAN method could be due to the high-dimensional nature of the data in dataset and its large differences in the densities.

In contrast, the reason for K-Means method to be more suitable for this dataset is due to its robust and efficient nature of the algorithm for distinct and well understood datasets.

Discussion

The K-Means method outperformed the DBSAN method in clustering the dataset of customers' annual expenditure on different product types from a wholesale distributor. The clustering results from the K-means model has been validated against the true observation and eventually below customer segments were identified:

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

Having such insight on the customer segments would enable the wholesale distributor to structure its organization to cater for specific customer needs in opposed to treat all the customers equally. Prior to this study, only a high-level classification of customers was available to the wholesale distributor. Those are 'Channel' and 'Region' of the customers. Such a limited knowledge on the customers did not allow the wholesale distributor to address the specific needs and requirements of the customers.

By identifying the similarities of the customers (i.e. knowledge on market segments) in regards to their annual spending through this study, the above-mentioned knowledge gap is narrowing down. For an example, though the results of this study, the wholesale distributor should now be aware that the Horeca (Hotels/Restaurants/Café) customers are much focused on Fresh products, hence the direct marketing & distribution campaigns on Fresh products could directly be carried out for Horeca customers. Similarly, the Retails customers are spending heavily on Milk and Grocery, therefore it would be much beneficiary to the wholesale distributor to focus on direct marketing strategies on Milk and Grocery products to the retail customers.

Also, this study enables the wholesale distributor to identify where specific customer groups are spending less on. For an example, the costumers on cluster 0 who spends heavily on Fresh products spends the least on the Detergent papers. With such knowledge, the wholesale distributor can investigate reason for this pattern and if possible deploy a promotional campaign on Detergents papers for customers in cluster 0. A few such interesting spending patters are listed below so that the wholesale distributer can structure its future marketing strategies to address these:

- Cluster 0 – Spends heavily on Fresh products but very less on Detergent Papers
- Cluster 1 – Spends heavily on Milk and Grocery but very less on Frozen and Delicassen
- Cluster 1 – Spends heavily on Milk and Grocery but very less on Fresh products
- Cluster 3 – High spenders, however spending on Frozen and Delicassen are relatively low

Conclusion

This study used two clustering techniques to model the data from a wholesale distributor on its customer annual spending. By comparing the results of the two clustering models, it was quite evident that the K-Means model outperformed the DBSAN model. The most appropriate K-Means clustering model identified five cluster groups (i.e K = 5 model) of the costumers based on their annual spending on six products types: Fresh, Milk, Grocery, Frozen, Detergent papers, and Delicassen. The comparison of the model was carried out by comparing the model results against the confusion matrices that have been derived by true observation labels. The clustering results from this study can be used by the wholesale distributor in deploying any direct marketing and distribution strategies to target customer segments in order to promote the overall sales. Also, 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.

References

1. Lecture notes: COSC 2670 Practical Data Science, RMIT
2. UCI data repository: http://archive.ics.uci.edu/ml/index.php
3. https://www.dataquest.io/blog/pandas-python-tutorial/
4. https://www.learnpython.org/en/Pandas_Basics
5. http://scikit-learn.org/stable/modules/clustering.html
6. https://bl.ocks.org/rpgove/0060ff3b656618e9136b
7. https://www.datascience.com/blog/k-means-clustering