Title: Data Modelling - Assignment 2

Student ID: s3796107

Student Name: Nushura Islam

Email (contact info): s3796107@student.rmit.edu.au

Affiliations: RMIT University.

Date of Report: 23 May 2022

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honor code by typing "Yes": *Yes*.

CONTENT:

Executive summary

This report aimed to understand consumers' buying behaviour of a wholesale distributor. This will assist the wholesale distributor in understanding the categories of products popular among specific Channels/consumers from certain regions/ cities. In addition to that, by knowing the popularity of those items, the wholesaler can stock those goods to make sure it is always available for its consumers and stock similar or related products that those clients may like. This will further assist the wholesaler in generating revenue. Overall, the result indicated the popularity of wholesale products among specific consumers in certain regions. The report concludes that while specific products are popular within certain Channels and Regions, it can also be seen that the popularity is not the same across all the Channels and Regions. It is recommended that the wholesaler stock products according to individual products' buying behaviour or popularity to provide their service faster, satisfy their clients, and generate revenue.

Introduction

There was no method to analyse how well a particular product or service was doing within the market in the early days. It could only be understood by comparing the product's performance with the individual shops sold. This was a problem because there was over or underproduction happening all the time, which meant that resources were not being used efficiently and were wasted. Hence something had to be done to ensure resources were being utilised at their highest capacity. Nowadays, there are many methods developed to analyse the performance of products. This report will research the methods used to analyse the performance of certain wholesale products with the help of calculating the wholesaler's clients' annual spending on specific goods.

Methodology

The Wholesale customers' data set was used from the 'UCI Machine Learning Repository'website. The research made use of a quantitative method approach. After retrieving the data set, the goal of the dataset was analysed. Then it was followed by data preparation, where the group was cleaned. Afterwards, data exploration was done by analysing each attribute column separately using boxplots and boxplot summary. Data exploration was further done by exploring the relationship between pairs of attributes which was done with the help of scatter plots. Then finally, data modelling was done by clustering, where two different modelling techniques were carried out to determine the model which should be used for the wholesale customers' dataset.

Results

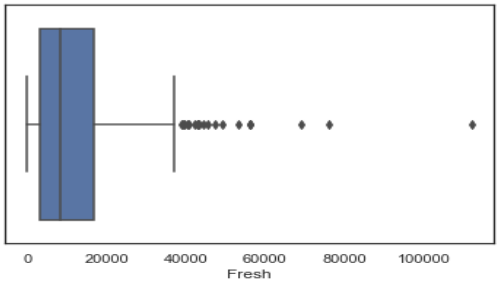
Task 1: Retrieving and preparing the data

For data cleaning purposes, the dropna() method was used to drop all the null or missing values and the drop_duplicates() method was used to drop all the duplicated rows from the dataset. After trying to drop weights using those two methods, it could be observed that the number of rows did not change, which meant that there were no missing values or duplicated rows.

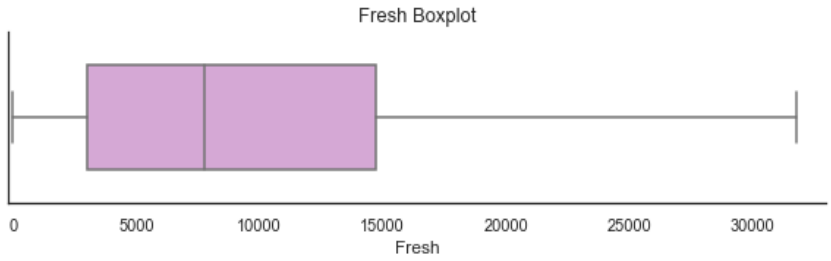
	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	2	3	12669	9656	7561	214	2674	1338
1	2	3	7057	9810	9568	1762	3293	1776
2	2	3	6353	8808	7684	2405	3516	7844
3	1	3	13265	1196	4221	6404	507	1788
4	2	3	22615	5410	7198	3915	1777	5185
...	...	...	...	...	...	...	...	...
435	1	3	29703	12051	16027	13135	182	2204
436	1	3	39228	1431	764	4510	93	2346
437	2	3	14531	15488	30243	437	14841	1867
438	1	3	10290	1981	2232	1038	168	2125
439	1	3	2787	1698	2510	65	477	52

440 rows × 8 columns

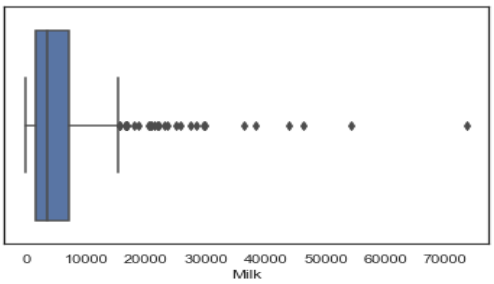
Afterwards, boxplots were used to determine if the wholesale product columns had any outliers. Thus, six boxplots were made to point out the outliers and remove them further as there were six products. The boxplots are demonstrated below where the initial boxplots are the boxplots where outliers are present, and the final boxplots are where the outliers are removed.



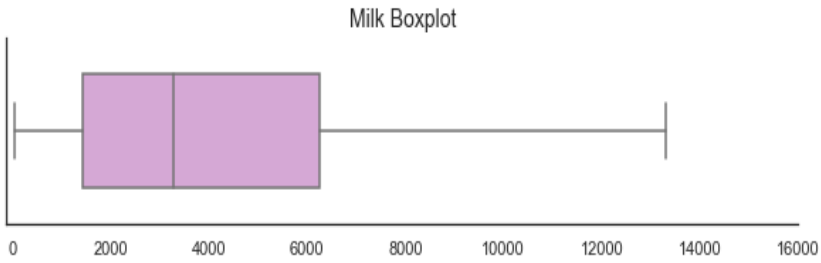
Initial Fresh boxplot



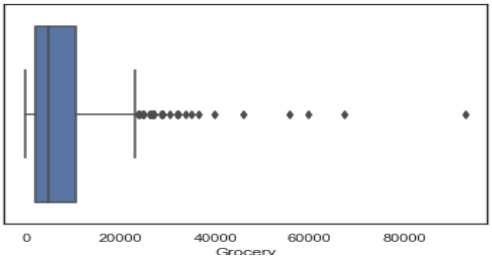
Final boxplot



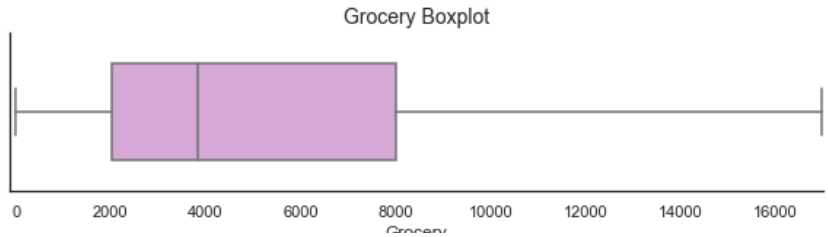
Initial Milk boxplot



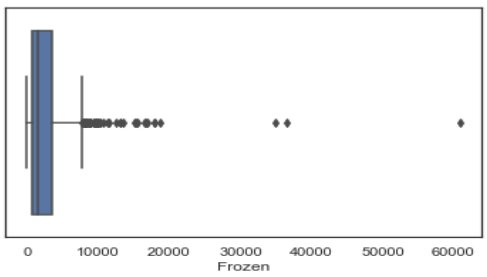
Final boxplot



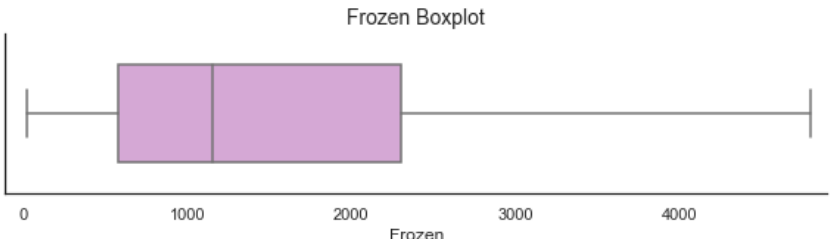
Initial Grocery boxplot



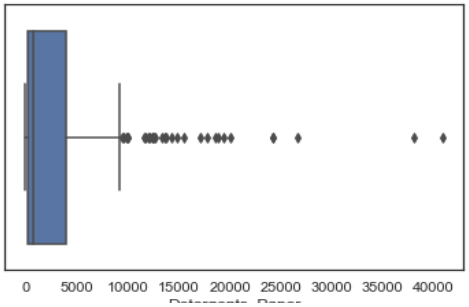
Final boxplot



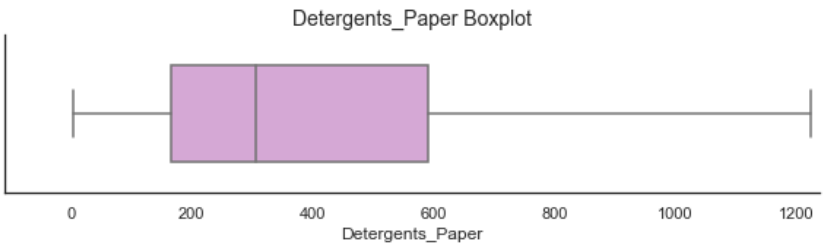
Initial Frozen boxplot



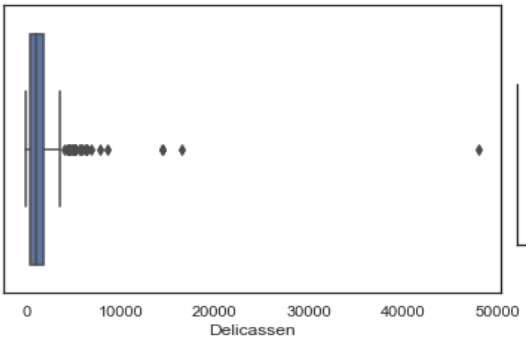
Final boxplot



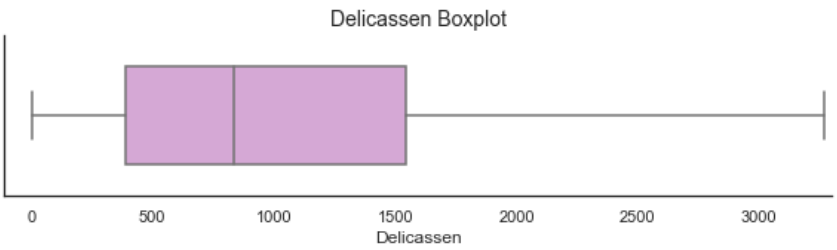
Initial Detergents_Paper boxplot



Final boxplot



Initial Delicassen boxplot



Final boxplot

Task 2.1: Data Exploration

To explore individual product columns, the summary of each cleaned column was derived by using the describe() method.

```
# Summary of the Fresh column
x = new_Delicassen3['Fresh']
x.describe()
```

```
count      407.000000
mean     11348.941032
std      11350.124650
min          3.000000
25%       3077.000000
50%       8257.000000
75%      15999.000000
max      76237.000000
Name: Fresh, dtype: float64
```

```
# Summary of the Milk column
x = new_Delicassen3['Milk']
x.describe()
```

```
count      407.000000
mean      5092.633907
std       6205.684870
min         55.000000
25%       1468.000000
50%       3373.000000
75%       6744.500000
max      73498.000000
Name: Milk, dtype: float64
```

```
# Summary of the Grocery column
x = new_Delicassen3['Grocery']
x.describe()
```

```
count      407.000000
mean      7446.877150
std       9093.010555
min          3.000000
25%       2089.500000
50%       4563.000000
75%       9806.500000
max      92780.000000
Name: Grocery, dtype: float64
```

```
# Summary of the Frozen column
x = new_Delicassen3['Frozen']
x.describe()
```

```
count      407.000000
mean      2737.017199
std       3619.453315
min         25.000000
25%        667.500000
50%       1439.000000
75%       3247.000000
max      35009.000000
Name: Frozen, dtype: float64
```

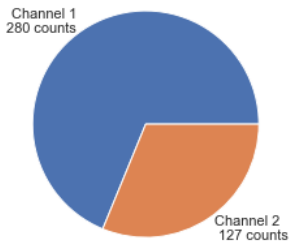
```
# Summary of the Detergents_Paper column
x = new_Delicassen3['Detergents_Paper']
x.describe()
```

```
count      407.000000
mean      2746.036855
std       4638.273986
min          3.000000
25%        245.000000
50%        778.000000
75%       3774.500000
max      40827.000000
Name: Detergents_Paper, dtype: float64
```

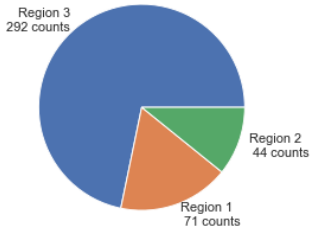
```
# Summary of the Delicassen column
x = new_Delicassen3['Delicassen']
x.describe()
```

```
count      407.000000
mean      1056.034398
std        820.341190
min          3.000000
25%        387.000000
50%        834.000000
75%       1542.500000
max       3271.000000
Name: Delicassen, dtype: float64
```

Afterwards, to explore the Channel and Region columns, pie charts were used.



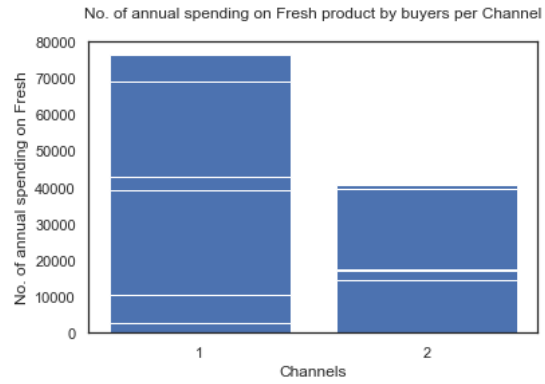
Channel Pie Chart



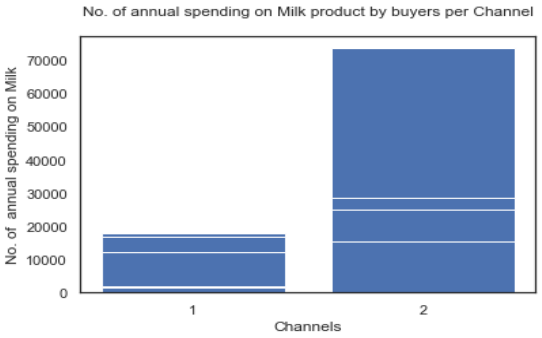
Region Pie Chart

Task 2.2: Exploring relationship between pairs of attributes

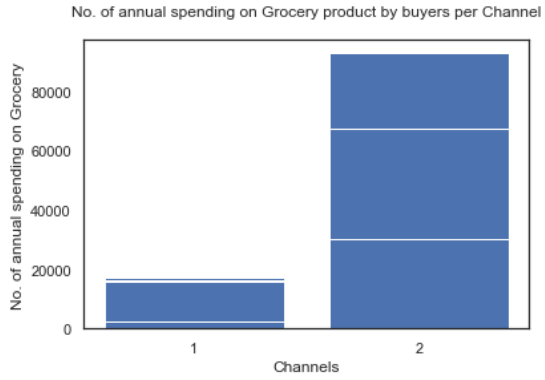
Bar charts were used to explore the relationship between Channel and all the other six wholesale products.



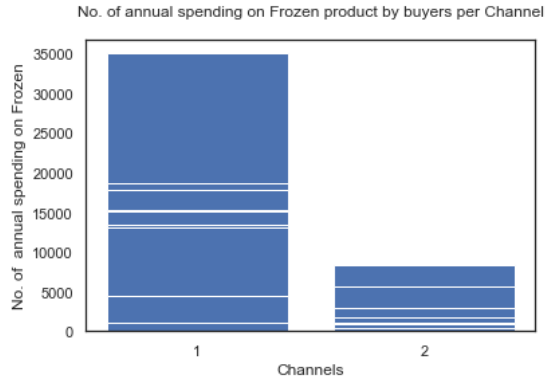
Hypothesis: Number of annual spending on Fresh products by clients per Channel. So, according to the bar chart, the result demonstrates that the highest annual spending is spent on Channel 1 clients, where the annual spending is around 75000.



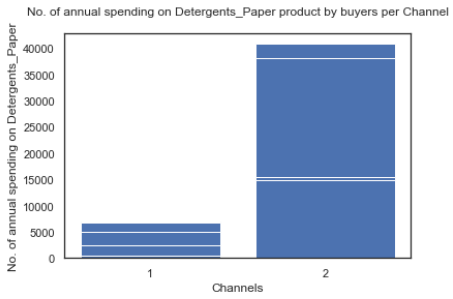
Hypothesis: Number of annual spending on Milk products by clients per Channel. So, according to the bar chart, the result demonstrates that the highest annual spending on Milk is spent from Channel 2 clients, where the annual spending is around 70000.



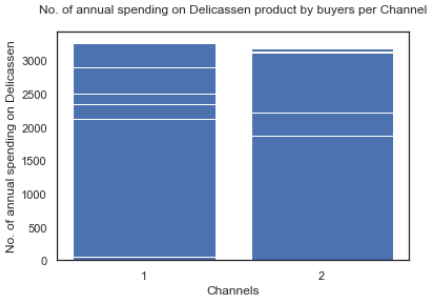
Hypothesis: Number of annual spending on Grocery products by clients per Channel. So, according to the bar chart, the result demonstrates that the highest annual spending on Grocery is spent from Channel 2 clients, where the annual expenditure is around 87000.



Hypothesis: Number of annual spending on Frozen products by clients per Channel. So, according to the bar chart, the result demonstrates that the highest annual spending on Frozen is spent from Channel 1 clients, where the annual spending is around 35000.

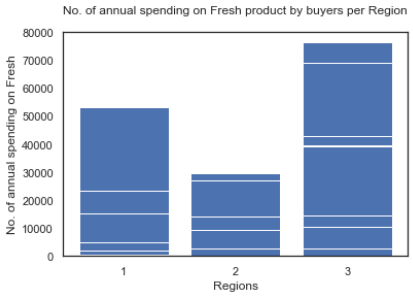


Hypothesis: Number of annual spending on Detergents_Paper products by clients per Channel. So, according to the bar chart, the result demonstrates that the highest annual spending on Detergents_Paper is spent from Channel 2 clients, where the annual spending is around 40000.

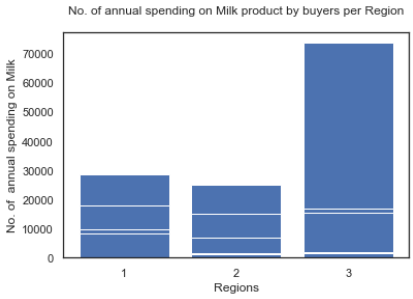


Hypothesis: Number of annual spending on Delicassen products by clients per Channel. The yearly spending on both the Channels is around 3000.

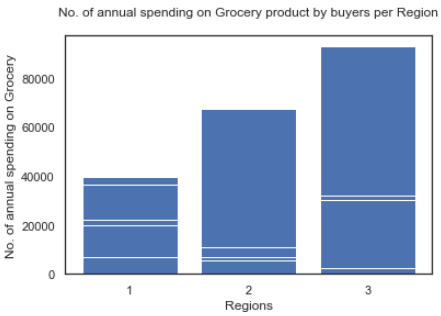
Then, again bar chart were used to explore relationship between Region and all the other six wholesale products.



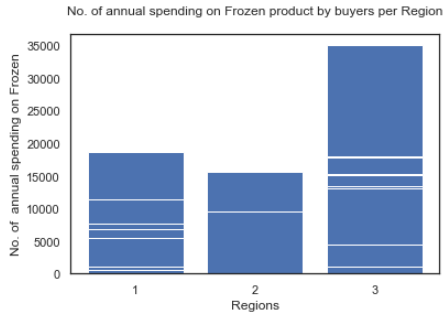
Hypothesis: Number of annual spending on Fresh products by clients per Region. So, according to the bar chart, the result demonstrates that the highest number of annual spending on Fresh is spent from Region 3 clients, where the annual spending is around 75000.



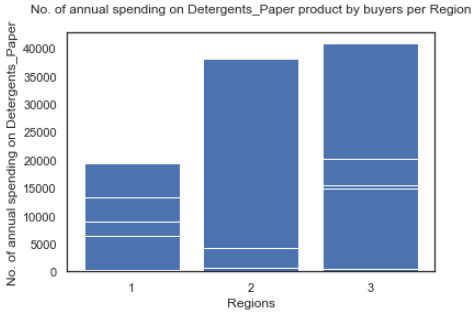
Hypothesis: Number of annual spending on Milk products by clients per Region. So, according to the bar chart, the result demonstrates that the highest number of annual spending on Milk is spent from Region 3 clients, where the annual spending is around 70000.



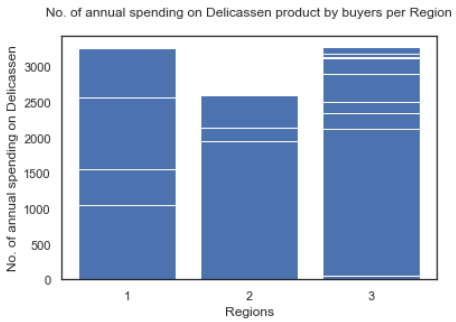
Hypothesis: Number of annual spending on Grocery products by clients per Region. So, according to the bar chart, the result demonstrates that the highest number of annual spending on Grocery is spent from Region 3 clients, where the annual spending is around 90000.



Hypothesis: Number of annual spending on Frozen products by clients per Region. So, according to the bar chart, the result demonstrates that the highest number of annual spending on Frozen is spent from Region 3 clients, where the annual spending is around 35000.



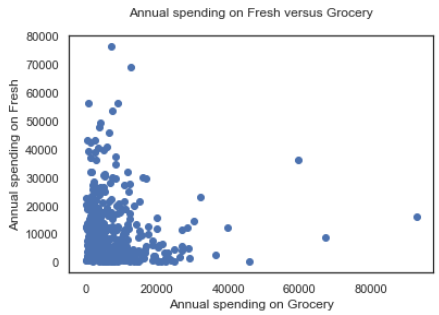
Hypothesis: Number of annual spending on Detergents_paper products by clients per Region. So, according to the bar chart, the result demonstrates that the highest number of annual spending on Detergents_paper is spent from Region 3 clients, where the annual spending is around 40000.



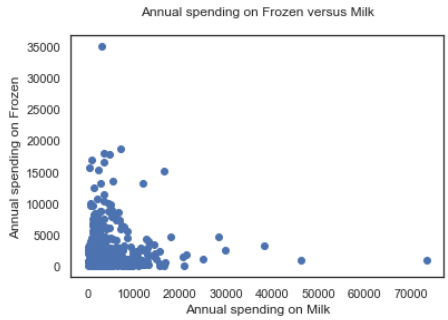
Hypothesis: Number of annual spending on Delicassen products by clients per Region. So, according to the bar chart, the result demonstrates that the highest number of annual spending on Delicassen is spent from Region 1 and 3 clients, where the annual spending is around 3000 each.

Overall, the result of the above bar charts demonstrates that popularity among all the wholesale products is the highest on region 3.

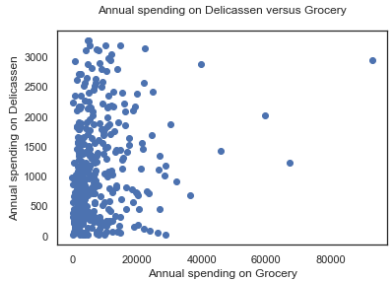
To explore relationship between few numerical columns, scatter-plots are used.



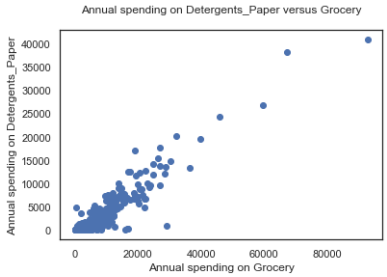
Annual spending on Fresh versus Grocery: It could be observed that the highest number of annual spending on Fresh were between 0 to 40 000 with annual spending on Grocery between 0 to 30 000.



Annual spending on Frozen versus Milk: In this plot, it could be observed that the highest number of annual spending on Frozen were with annual spending on Milk between 0 to 19 000.



Annual spending on Delicassen versus Milk: The result demonstrated that the highest number of annual spending on Delicassen were with annual spending on Grocery between 0 to 15 000.



Annual spending on Detergents_Paper versus Grocery: The highest number of annual spending on Detergents_Paper were between 0 to 10 500 with annual spending on Grocery between 0 to 30 000.

Task 3: Data Modelling

The model used for modelling the wholesale dataset is Clustering which consists of KMeans clustering and DBSCAN. The number of clusters selected was eight when clustering, as there are eight attributes within the wholesale dataset.

```
# Here clustering is done on the whole dataste
# n_cluster is refering to the number of clusters
# so as there are 8 attributes, the number of cluster is going to be 8
model = cluster.KMeans(n_clusters = 8, random_state=14)
```

The fitted model with the dataset using the fit() method. This was followed by creating two new columns called ‘cluster’ and counting as ‘c’. The cluster column values are not yet specified in the following result regarding what those values represent.

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	cluster	c
0	2	3	12669	9656	7561	214	2674	1338	6	count
1	2	3	7057	9810	9568	1762	3293	1776	6	count
2	1	3	13265	1196	4221	6404	507	1788	4	count
3	2	3	9413	8259	5126	666	1795	1451	6	count
4	2	3	12126	3199	6975	480	3140	545	4	count
...	...	...	...	...	...	...	...	...	...	...
402	1	3	29703	12051	16027	13135	182	2204	5	count
403	1	3	39228	1431	764	4510	93	2346	0	count
404	2	3	14531	15488	30243	437	14841	1867	7	count
405	1	3	10290	1981	2232	1038	168	2125	1	count
406	1	3	2787	1698	2510	65	477	52	1	count

407 rows × 10 columns

Then the clustering result was found but carrying out these methods.

```
In [584]: # cluster is the clustering results
          # Region is acting as target here
          # c is the count of Region per cluster
          clustering_result = X[["cluster", "Region", "c"]].groupby(["cluster", "Region"]).agg("count")

In [585]: clustering_result
```

Out[585]:

cluster	Region	c
0	1	2
	3	14
	1	27
1	2	16
	3	92
	1	7
2	2	7
	3	22
3	3	3
	1	15
4	2	11
	3	71
	1	7
5	2	3
	3	26
	1	11
6	2	6
	3	60
	1	2
7	2	1
	3	4

Discussion

Task 1

To begin with, after trying to drop values using the dropna() and drop_duplicates() methods, it could be observed that the number of rows, which was 440, did not change, which meant that there were no missing values or duplicated rows.

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	2	3	12669	9656	7561	214	2674	1338
1	2	3	7057	9810	9568	1762	3293	1776
2	2	3	6353	8808	7684	2405	3516	7844
3	1	3	13265	1196	4221	6404	507	1788
4	2	3	22615	5410	7198	3915	1777	5185
...	...	...	...	...	...	...	...	...
435	1	3	29703	12051	16027	13135	182	2204
436	1	3	39228	1431	764	4510	93	2346
437	2	3	14531	15488	30243	437	14841	1867
438	1	3	10290	1981	2232	1038	168	2125
439	1	3	2787	1698	2510	65	477	52

440 rows × 8 columns

Box-plots for the 6 products to find outliers

All the box plots starting range of the x-axis was high due to the presence of the outliers. Thus, it had to be zoomed in by setting the range of the x-axis. By doing so, the point from which the outliers are present could be figured.

Firstly, the Fresh boxplot's lower quartile value was -180, and the upper quartile value was 38 000. It could be observed through the initial Fresh boxplots that any integer which was above 38 000 was an outlier. Thus, to remove the outliers, this method was used for all the six boxplots repeatedly until there were no outliers present.

```
# This code is enabling all the values before -180 and after 38000 to be removed
# As those values are acting as outliers
# Thus, after removing those values, the new dataset is being stored in 'new_fresh'
new_fresh = wholesaler[wholesaler['Fresh'].between(-180,38000)]
```

- For instance, for Fresh, this method removed all values below -180 and above 38 000. However, it could be observed that even after trying to remove the outliers by running the method above, a new set of outliers was found. Thus, this process had to be repeated three times to achieve outliers free attributes.

- The milk boxplot has a lower quartile value of -120, and the upper quartile value of 155 00. The outlier removal process had to be repeated two times to achieve outliers free attribute.

- The grocery boxplot has a lower quartile value of -110, and the upper quartile value of 23400. The outlier removal process had to be repeated four times to achieve outlier free attribute.

- The Frozen boxplot has a lower quartile value was -110 and the upper quartile value was 7700. The outlier removal process had to be repeated three times to achieve outlier free attribute.

- The detergents paper boxplot has a lower quartile value was -110 and the upper quartile value was 9400. The outlier removal process had to be repeated eleven times to achieve outlier free attribute.

- The delicassen boxplot has a lower quartile value of -110 and the upper quartile value of 4000. The outlier removal process had to be repeated three times to achieve the outlier free attribute.

Hence, the repetition of outlier removal process proved that it is not a normal distribution.

Task 2.1

To explore each column, boxplots could be used, but that was already done in the previous section of the report. So, this time, by using the cleaned dataset, the overall summary of each product column is found. For finding the summary, describe() method was used. The summary of each column included count, mean, standard deviation, minimum, 25 percentile, 50 percentile, 75 percentile and maximum values.

On all the summary values, the count refers to the number of total rows the dataset contains. Generally, the mean represents the average of the column values. The median or 50 percentiles represents the mid-point of the dataset range.

Fresh: Thus, for the Fresh column, the mean value is 11348.94, meaning that the annual spending on Fresh product by consumers are on an average of 11348.94. The mid-point of Fresh is 8257. The minimum and maximum values for Fresh are 3 and 76237, respectively, meaning minimum annual spending of 3 and a maximum of 76237. The lower-quartile or 25 percentile is 3077, and the upper-quartile or 75 percentile is 15999.

Milk: For the Milk column, the mean value is 5092.63, meaning that consumers' annual spending on Milk products is, on average, 5092.63. The mid-point of Milk is 3373. The minimum and maximum values for Milk are 55 and 73498, respectively, meaning minimum annual spending of 55 and a maximum of 73498. The lower-quartile or 25 percentile is 1468, and the upper-quartile or 75 percentile is 6744.50.

Grocery: For the Grocery column, the mean value is 7446.88, meaning that consumers' annual spending on Grocery products is, on average, 7446.88. The mid-point of Grocery is 4563. The minimum and maximum values for Grocery are 3 and 92780, respectively, meaning there is a minimum annual spending of 3 and a maximum of 92780. The lower-quartile or 25 percentile is 2089.50, and the upper-quartile or 75 percentile is 9806.50.

Frozen: For the Frozen column, the mean value is 2737.02, meaning that consumers' annual spending on Frozen products is, on average, 2737.02. The mid-point of Frozen is 1439. The minimum and maximum values for Frozen are 25 and 35009, respectively, meaning there is a minimum annual spending of 25 and a maximum of 35009. The lower-quartile or 25 percentile is 667.50, and the upper-quartile or 75 percentile is 3247.

Detergents_Paper: For the Detergents_Paper column, the mean value is 2746.04, meaning that consumers' annual spending on Detergents_Paper products is an average of 2746.04. The mid-point of Detergents_Paper is 778. The minimum and maximum values for Detergents_Paper are 3 and 40827, respectively, meaning minimum annual spending of 3 and a maximum of 40827. The lower-quartile or 25 percentile is 245, and the upper-quartile or 75 percentile is 3774.50.

Delicassen: For the Delicassen column, the mean value is 1056.03, meaning that consumers' annual spending on the Delicassen products is an average of 1056.03. The mid-point of Delicassen is 834. The minimum and maximum values for Delicassen are 3 and 3271, respectively, meaning that there is a minimum annual spending of 3 and a maximum of 3271. The lower-quartile or 25 percentile is 387, and the upper-quartile or 75 percentile is 1542.50.

Then pie charts were used to explore Channel and Region columns. The Channel pie chart clearly shows that Channel 1, with 280 counts, contains the highest number of counts compared to Channel 2, which comprises 127 counts. Moreover, on the Region pie chart, the highest count is for Region 3 with 292 counts, and the lowest count is for Region 2 with 44 counts.

```
1    280
2    127
Name: Channel, dtype: int64
```

```
3    292
1     71
2     44
Name: Region, dtype: int64
```

Task 2.2

Bar charts were used to explore the relationship between categorical and numerical values. All the bar charts address the research question. They are pointed out below:

- **No. of annual spending on Fresh product by clients per Channel**: Based on the Channel, Fresh product is the most popular on Channel 1.
- **No. of annual spending on Milk product by clients per Channel:** Based on the Channel, Milk product is the most popular on Channel 2.
- **No. of annual spending on Grocery product by clients per Channel:** Based on the Channel, Grocery product is the most popular on Channel 2.
- **No. of annual spending on Frozen product by clients per Channel:** Based on the Channel, Frozen product is the most popular on Channel 1.
- **No. of annual spending on Detergents_Paper product by clients per Channel:** Based on the Channel, Detergents_Paper product is the most popular on Channel 2.
- **No. of annual spending on Delicassen products by clients per Channel:** According to the bar chart, the result demonstrates that both the Channels have similar annual spending on Delicassen, with Channel 1 being a bit more than Channel 2. So, based on the Channel, Delicassen product is the most popular on Channel 1 comparatively.

- **No. of annual spending on Fresh product by clients per Region**: Based on the Region, Fresh product is the most popular in Region 3 and least popular in Region 2.

- **No. of annual spending on Milk product by clients per Region:** Based on the Region, Milk product is the most popular in Region 3 and least popular in Region 2.

- **No. of annual spending on Grocery product by clients per Region:** Based on the Region, Grocery product is the most popular in Region 3 and least popular in Region 1.

- **No. of annual spending on Frozen product by clients per Region:** Based on the Region, Frozen product is the most popular in Region 3 and least popular in Region 2.

- **No. of annual spending on Detergents_Paper product by clients per Region:** Based on the Region, Detergents_paper product is the most popular in Region 3 and least popular in Region 1.

- **No. of annual spending on Delicassen product by clients per Region:** Based on the Region, Delicassen product is the most popular in Region 1 and 3 and least popular in Region 2 comparatively.

Scatter plots were used to explore the relationship between a few numerical columns. The scatter plots illustrated for annual spending on fresh versus grocery, frozen versus milk, and delicassen versus milk demonstrated no trends or patterns associated with the relationship. This means that there is no specific association between the mentioned relationships. However, the scatter plot for annual spending on Detergents_paper versus grocery demonstrated a linear relationship where with the increase in yearly spending, there was an increase of yearly spending on Detergents_Paper.

Task 3: KMeans Clustering

This result illustrates that in cluster 0, there are two regions, Region 1 and 3, which act as targets. Thus, for all the records in cluster 0, there are 2 ones, and 14 threes. For all the records in cluster 1, there are 27 ones, 16 twos and 92 threes. The same goes for all the other clusters.

This is not a suitable data modelling method for this dataset because there could not be a proper target set. This is because for setting a target, proper features are needed. However, for this dataset, all the features are vital to be considered. All the wholesale items available fall under the products of the wholesale distributor. The places the distributor is providing its service to are further determined by the help of the Channel and Region together.

In addition, there cannot be a particular item or set of products popular within a specific Channel and Region as there are a variety of products offered by the wholesaler, with varying levels of popularity within different pairs of Channel and Region.

Conclusion

In general, analysing how well a product is doing within the market or shop is vital in revenue generation. The data analysis process was being done to determine the popularity of wholesale products within specific Channel and Region. Based on the analysis carried out, individual product popularity can be determined. Product popularity with a particular Channel and Region is listed below:

- The fresh product is the most popular within Channel 1 and Region 3
- The milk product is the most popular within Channel 2 and Region 3
- The grocery product is the most popular within Channel 2 and Region 3
- The frozen product is the most popular within Channel 1 and Region 3
- The detergents_paper product is the most popular within Channel 2 and Region 2 and 3
- The delicassen is the most popular within Channel 1 and 2 and Region 1 and 3

Hence, overall, Region 3 is the most popular among the wholesale products.