Module: BUSI4370 UNUK

Analytics Specialization and Applications

WENMIN NIU(alywn1)

20157952

2019/20

Executive Summary

The main purpose of this report is to provide a clustering analysis for a national convenience store chain to help understand its customers better. RFM-related data, average values on basket dimension and category spending data were considered as features. After engineering features manually and adopting PCA, K-means algorithm produced 5 clusters.

"**Champions_Shopping-dispersed_ALL/Tobacco&Alcohol**" is the most valuable groups, from which customers account for only 21% but generate 37% of total spend and each customer spends about £1300 over 6 months. As its name indicates, people in this group like to shop dispersed and they like all categories especially tobacco and alcohol comparing with others.

There are three middle-valuable groups. "**Loyal_Shopping-together_Food**", "**Potential-loyalist_Shopping- dispersed_ Tobacco&Alcohol**" and "**Moderate-to-low_Shopping-dispersed_All**" have middle-to-high, middle, middle-to-low monetary respectively. Their shopping habits can be seen from their name. 9% of customers are assigned to "**Low_About-to-sleep_No-preference**", with low monetary, low recency, low frequency.

Some recommendations are given:

1. The group "**Moderate-to-low_Shopping-dispersed_All**" needs attention. Bundling sales and lottery campaign when monetary of a basket surpasses a certain value may be useful to improve the spend per basket and then total spend in this group.
2. The company can offer some special discount to recall and retain customers in "**Low_About-to-sleep_No-preference"** group.
3. Further analysis of demographic information may be helpful to devise more creative and effective marketing strategies.

Section A: Feature selection

The national convenience store chain provides 4 tables of its transaction data, which includes 3000 customers' purchase behavior from 1st March to 31st August in 2007. The line-items table shows the most detailed information, such as time and quantity that customers buy a specific product, which category the product belongs to and how much they pay for. Basket table, category table and customers table, as their names suggest, summarize data to a certain degree respectively.

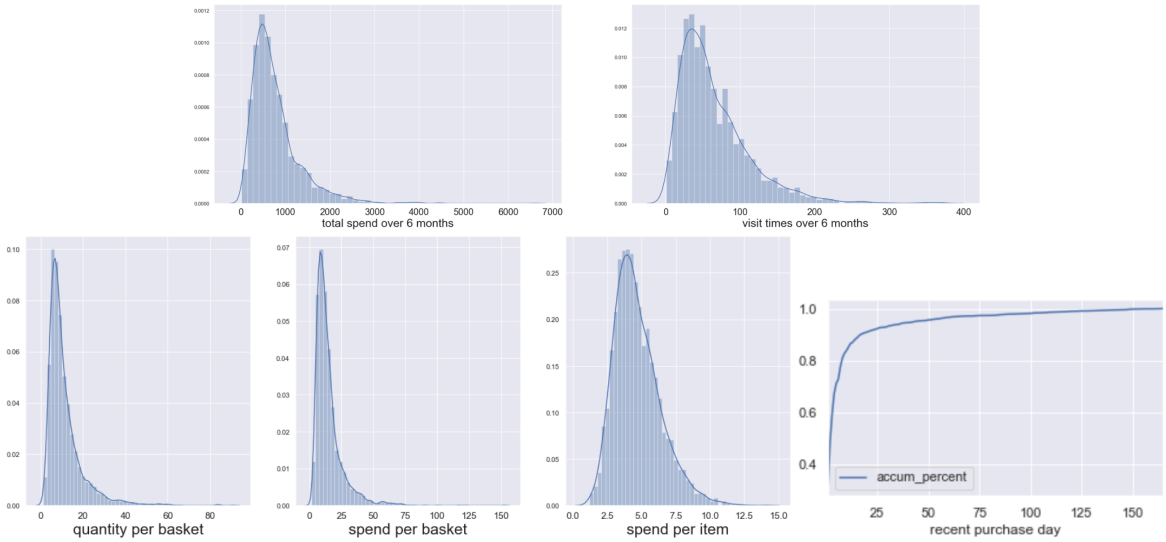
Based on the information from dataset and requirements from CDO , following features were picked:

1. RFM, being widely used in retail industries, uses sales data to segment a pool of customers based on their purchasing behavior(Blattberg, Kim and Neslin, 2008). It stands for three dimensions: recency, frequency and monetary. In this case, three dimensions are:
    a. Recency: the number of days that have passed since customers' last purchase.(Assuming today is 1st September in 2007)
    b. Frequency: the number of purchase(baskets) over 6 months.
    c. Monetary: total spend over 6 months
2. Customer behavior varies in terms of each visit. Two possible completely different scenarios are: (1)people just buy a few items in each baskets but visit store frequently, (2)the frequency of shopping is low but the item quantity of each basket is high. Therefore, just as CDO suggests, average spend per basket, average quantity per basket and average spend per item, as well as category number per basket, were picked as features.
3. Different customers may prefer different products. For example, people who have pets usually buy pets-related products and those who try to eat healthy favor vegetables and fruits. Analyzing product-related features is necessary but picking a fine level is crucial, because higher level may obscure some characters while lower level cause data sparse. In this case, category level was chosen because product_id level is too detailed which would cause too many 0 values.

Section B: Customer Base Summary

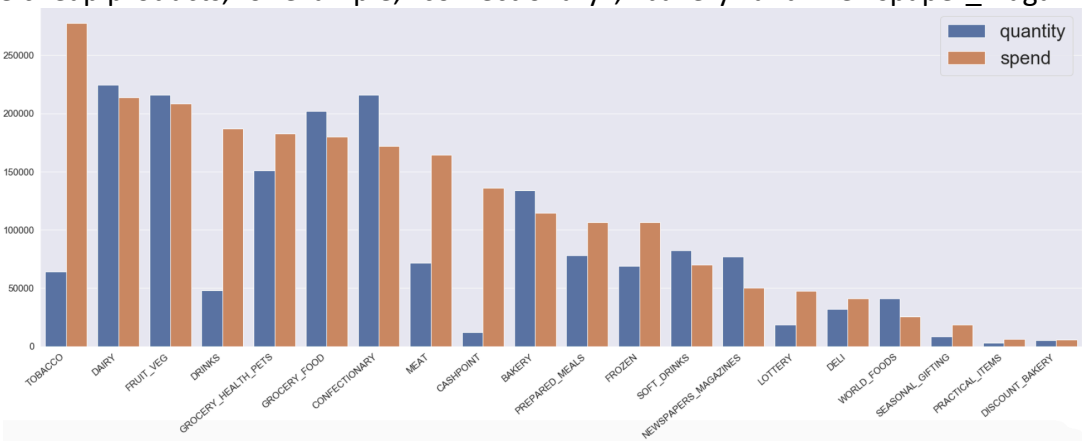
Among 3000 customers, most people(2985) have purchase behavior in March and earliest purchase of others(15) are all in April, so new customer scenario would not be considered in this report. Some important summaries are listed below:

1. On average(mean), each customer spends £771 and visits store 65 times over 6 months. They buy 11 items and pay £15 for each basket approximately. Besides, 60% and 80% people buy in recent 3 days and 8 days respectively.

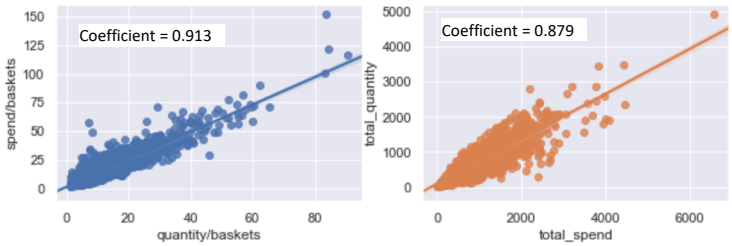


2. In terms of category, spend on tobacco accounts for most. Expenditure on those for maintaining daily life(food) are also high, such as "dairy", "fruit_veg" and so on. The five lowest value categories are "deli", "world_food", "practical_items", "seasonal_gifting" and

"discount_bakery". In addition, low quantity and high spend of some categories – such as "tobacco", "drinks", "meat" and "cashpoint" – imply high unit price. High quantity and low spend indicate cheap products, for example, "confectionary", "bakery" and "newspaper_magazines".



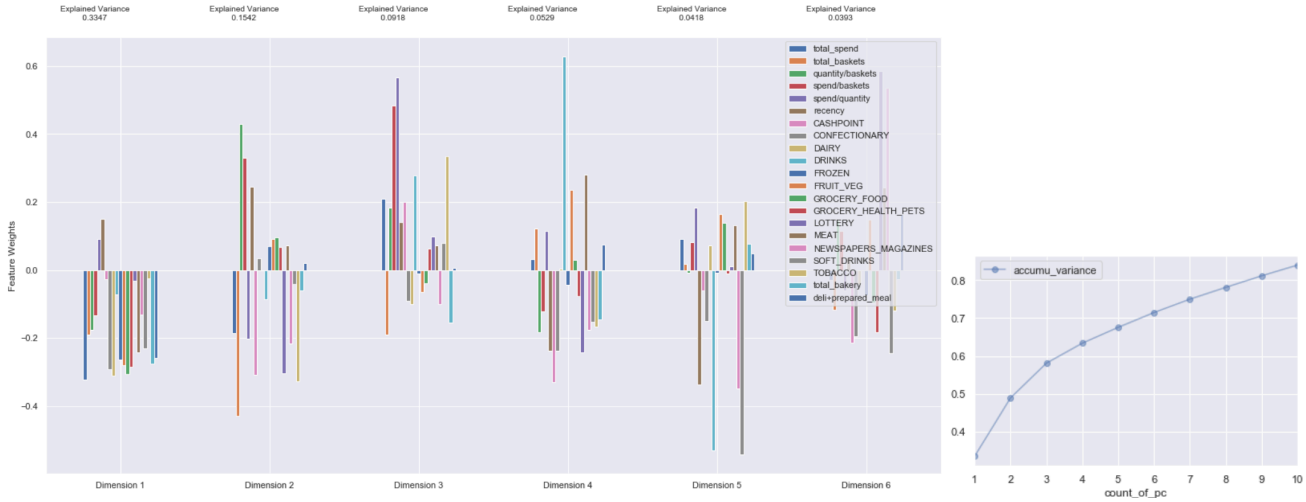
3. Strong correlations exist. More quantity means more spend which is consistent with common sense.



Section C: feature Engineering

In this case, dropping some low spending categories is acceptable, because higher value categories deserve more analysis. Consequently, the five lowest value category features were adjusted based on common sense and statistics from Section B. Firstly, using "total bakery" to instead ("bakery" + "discount bakery"), as they both mean bakery and there are too many 0 values in "discount_bakery". Secondly, combining "deli" and "prepared meal" is reasonable because they are both pre-cooked, possible having same target customers. Thirdly, it is plausible to cut out "world food", "practical items" and "seasonal gifting" rather than integrating them to "others". The reason is that if any of them was important, it would be better to consider it independently. If all were trivial, "others" would be not informative too.

After the above step, 21 features were remained, but correlations still exist among them. PCA was used to not only reduce dimensions but also eliminate the correlations, as they would skew results towards partitioning customers based on those correlated features. 71% of the variance was explained by the first 6 principal components. Here gives the explanation of first four.



- The first principal component(PC1): an increase in PC1 is mainly associated with the decrease of "total_spend" and food_related categories (such as "confectionary", "dairy", and so on), because absolute values of their weights are higher. They have similar height which means that they together distinguish customers and no one is more significant. Thus, PC1 can be summarized as "**spend and food**".
- The second principal component(PC2): here an increase in PC2 is associated with an increase in "quantity/baskets", "spend/baskets" and "recency" and a decrease of "total_baskets", and non-food-related categories (such as "cashpoint", "lottery", and so on) that affect just a little in PC1. Food categories in PC1 have small weights here. This makes sense as different PC should represent different features. PC2 can be described as "**value of baskets and non-food**"(high total basket means low corresponding value on per basket).
- The third principal component(PC3): in this case an increase of "spend/baskets", "spend/quantity", "cashpoint", "drinks" and "tobacco" causes an increase in PC3. Coincidentally, these three categories have relatively higher unit price than other categories. I refer PC3 as "**high unit_price**" dimension.
- The forth principal component(PC4): spend on drink influences most, so I refer this as "**drinks**".

Section D: Segmentation Methodology

This report chose K-means, which is a popular and robust clustering algorithm. The reason for using this technique is its pros are significant:

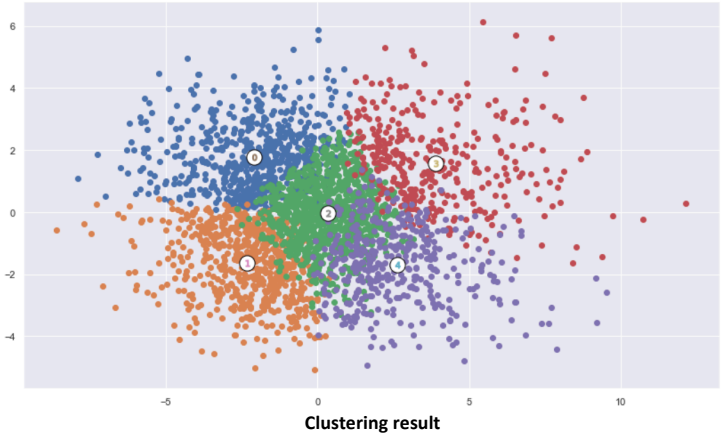
- It is simple and intuitive, so it is easy to understand.
- Easy to implement and adapt to new examples.
- It gives elliptical clusters and centers that can represent a group of customer, which is suiltable for this scenario.

To do K-means, the first step is to select k centers randomly—where k is specified in advance(Albon, 2018:288). Then compute the distance of each customer to each centroid separately and choose the nearest centroid as its cluster. Next, using the mean of each cluster as new centers to instead previous ones. After computing centroids and getting there cluster members iteratively until they do not change or complete m iterations, K-means divides customers into K clusters in which intra-class similarity is high and inter-class similarity is low.

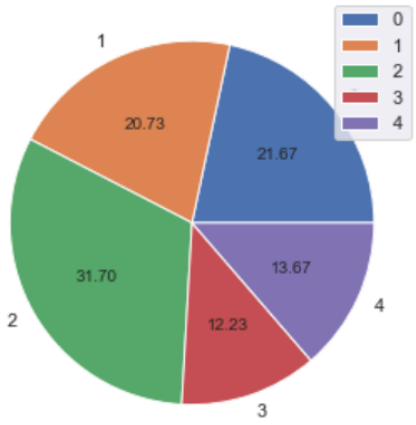
Silhouette score is used to find the best K. It is a measure how similar an object is to its own cluster (cohesion) compared to other clusters (separation), ranging from -1 to 1, where a high value indicates better clustering(Wikipedia. n.d.).

In the light of the company requirements(generating 5~7 segments) and silhouette score, 5-segments is the best choice because it has higher score than the other two and its plot is clearest among three choices.

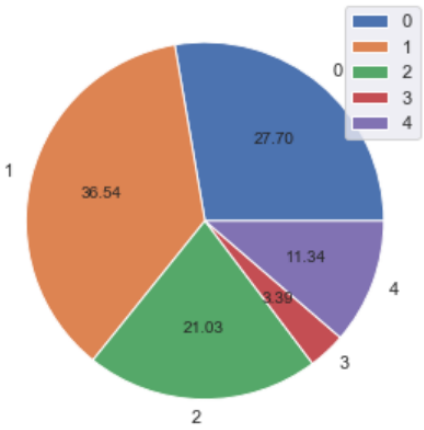
Section E: Segmentation Result



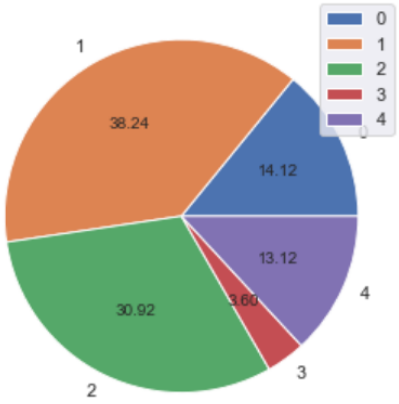
5 groups were produced in Section D. The scatter graph shows that segment 0, 1, 3, 4 are clearly separated, and segment 2 has some overlaps with 1, 4 respectively, which would be explained later. In this part, each segment was analyzed both statistically and descriptively. Segment name would also be given and its format is “value + shopping habits + favoring categories”. To do better comparison, the analyzing would not follow the order of 0 to 4.



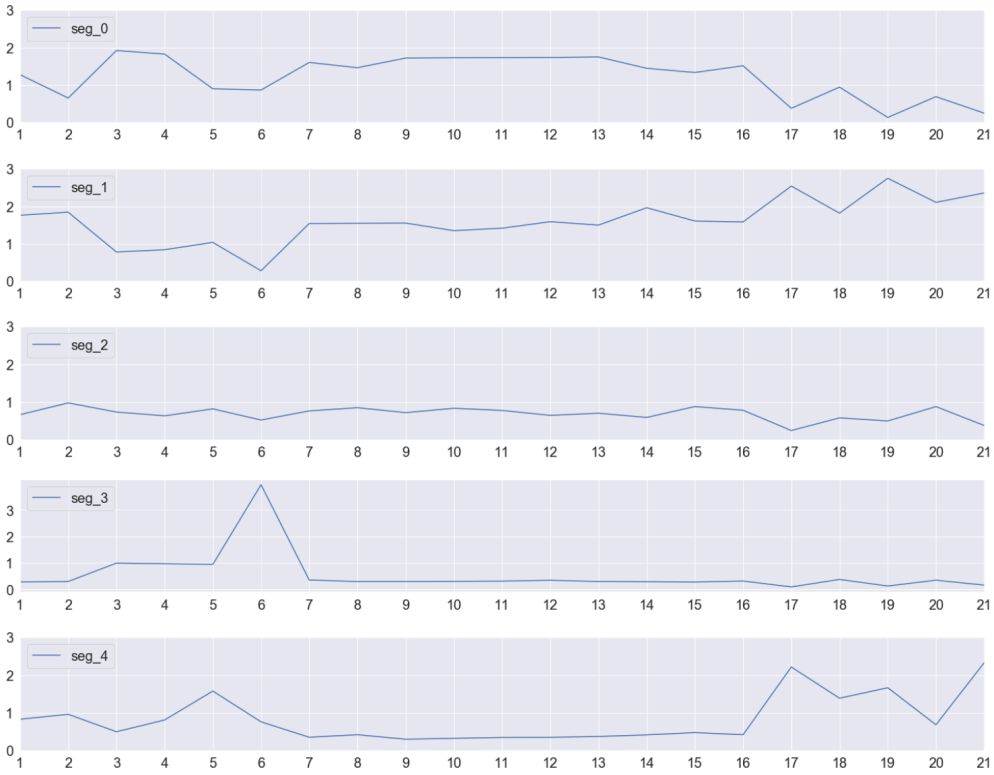
number of customers



total spend



total baskets(ferquency)



features	
1	total_spend
2	total_baskets
3	quantity/baskets
4	spend/baskets
5	spend/quantity
6	recency
7	CONFECTIONARY
8	DAIRY
9	FROZEN
10	FRUIT_VEG
11	GROCERY_FOOD
12	GROCERY_HEALTH_PETS
13	MEAT
14	SOFT_DRINKS
15	total_bakery
16	deli+prepared_meal
17	TOBACCO
18	DRINKS
19	LOTTERY
20	NEWSPAPERS_MAGAZINES
21	CASHPOINT

Note:

1. Value of vertical axis(y) equals to mean of each segment over mean of all customers, so y=1 means average, y=2 means data is twice the average.

2. Recency: low recency value means high recency level and vice

value comparison of each segment on each feature

	total_spend	total_baskets	quantity/baskets	spend/baskets	spend/quantity	recency
0	985.728600	42.470769	21.669782	27.077503	1.256521	7.901538
1	1358.965289	120.213826	8.794029	12.469553	1.449311	2.556270
2	511.627361	63.580442	8.279480	9.403377	1.142956	4.770768
3	213.982616	19.160763	11.163948	14.306167	1.315766	36.108992
4	639.653195	62.582927	5.631898	12.018017	2.201714	6.948780

value of each non-category feature

- **Segment_0:**

650 samples(22%) belong to this segment. Spending £986 each person over 6 months which is higher than average of all customers(£771), people here have middle-to-high monetary value and their total spend contributes 28% of all customers. Meanwhile, compared to average, relatively lower frequency (42 vs 65) and higher spend per basket(27 vs 14) are also their traits. In short, they do not go shopping frequently very much but purchase a lot each time, with a recency of 8 days which is similar to average(9 days).

In terms of categories, they have high value on 7^{th}~16^{th} features that are daily-life-maintaining food as the almost all parts of the line graph between 7^{th}~16^{th} are higher than 1.5(y>1.5), which means they are 1.5 times as much as the average.

Therefore, I refer this cluster as "**Loyal_Shopping-together_Food**", with approximate pen profile and corresponding level comparing with all customers:

	Monetary	Frequency	recency	Spend / basket	Category
Value	£1000	40 times	One week	£27	Food
Level	Middle-to-high	Middle-to-low	middle	high	

- **Segment_4:**

This segment includes 410 customers(14%). Although this segment has lower spend(640) than Segment_0, it has higher frequency(63) and low-level quantity per basket(5.6). In a nutshell, people here purchase more times but only buy a little each time.

From the trend of "seg_0" and "seg_4" line, we can clearly see that they have totally different category preference. The line of seg_4 is near zero on 7^{th}~16^{th} features, but far away from x-axis on 17^{th} ~21^{st} features. More detailed, people here like tobacco, drinks, lottery, newspaper_magazines and cashpoint. "Tobacco&Alcohol" is used to describe their category preference as spending on these two categories are much more than others(see Section B).

This segment performs moderately on RFM-related features, but has higher value on the 5^{th} feature – "spend/quantity". It is reasonable because "tobacco", "drinks", "lottery" and "cashpoint" have high unit price(see Section B).

Therefore, I refer this cluster as "**Potential-loyalist_Shopping- dispersed_ Tobacco&Alcohol**", with approximate pen profile and corresponding level comparing with all customers:

	Monetary	Frequency	recency	Spend / basket	Category
Value	£650	60 times	One week	£12	Tobacco&Alcohol
Level	Middle	middle	middle	middle	

- **Segment_1:**

There are 662 clients(21%) here. Contributing 37% of total spend, this group fully deserves "champions", having total spend(£1359) twice the average and recent purchase day (2.6 days) one third of the average.

This group combines the advantages of Segment_0 and Segment_4. Like Segment_0, this group buys food 1.5 times as much as average. The similar part with Segment_4 is that it buys non-food category("Tobacco&Alcohol") much more. Besides, people here have same purchase habit – buy more times but only a little each time. More specifically, their frequency is 120 which is twice the average but spend per basket(£12) is a little bit lower than average level(£14). What is necessary to mention is that most "newspaper_magazines" are bought by this group, because value on this feature of all other

groups is less than 1. It makes sense as people who have interest in reading news need to buy them regularly and frequently.

Hence, I describe this segment as "**Champions_Shopping-dispersed_ALL-more-Tobacco&Alcohol**", with approximate pen profile and corresponding level comparing with all customers:

	Monetary	Frequency	recency	Spend / basket	Category
Value	£1300	120 times	2.5 days	£12	All-more-Tobacco&Alcohol
Level	High	High	High	middle	

- **Segment_2:**

This segment includes 951 customers(31%). The line is so flat that no special trait is exactly the trait of this group. People in this group also like to shopping dispersed because the spend per basket (£9.4)is the lowest among all groups and total basket is at middle level(64). This group has middle-to-low monetary, because it total spend(£512) is two thirds of average value(£771).

This group do not have obvious favor on categories. They buy both food and Tobacco&Alcohol, which explains the overlap mentioned above.

So this segment is "**Moderate-to-low_Shopping-dispersed_All**": with approximate pen profile and corresponding level comparing with all customers:

	Monetary	Frequency	recency	Spend / basket	Category
Value	£500	65	5 days	£9.5	All
Level	Middle-to-low	Middle	Middle	Low	

- **Segment_3:**

12% of customers(367) here only generate 2.4% of total spend and 3.6% of total baskets. This is a low value segment, with both total spend(£213) and total basket(19) being less than 1/3 of average of all customers. The recency is 36 days, much longer than average and any other group. In addition, their expenditure on every category is low too.

Therefore, I refer this segment as "**Low_About-to-sleep_No-preference**": with approximate pen profile and corresponding level comparing with all customers:

	Monetary	Frequency	recency	Spend / basket	Category
Value	£200	20	More than 1 month	£12	No
Level	Low	Low	Low	Middle	

Section F: Summary

	group name	customer percent	total spend percent	Monetary	Frequency	recency	Spend per basket	Category
Segment_1	Champions_Shopping-dispersed_ALL/Tobacco&Alcohol	21%	37%	High	High	High	middle	All, especially Tobacco&Alcohol
Segment_0	Loyal_Shopping-together_Food	22%	28%	Middle-to-high	Middle-to-low	middle	high	Food
Segment_4	Potential-loyalist_Shopping- dispersed_Tobacco&Alcohc	14%	11%	Middle	middle	middle	middle	Tobacco&Alcohol
Segment_2	Moderate-to-low_Shopping-dispersed_All	32%	21%	Middle-to-low	Middle	Middle	Low	All
Segment_3	Low_About-to-sleep_No-preference	12%	3%	Low	Low	Low	Middle	No

Important traits of all 5 segments were summarized in the above table. 21percent customers are champions, contributing 37 percent of total spend. The company can focus attention to "**Moderate-to-**

low_Shopping-dispersed_All" group, which is a large group. But here customers do not have special preference and perform middle-to-low on RFM value which means that the company's products do not attract them very much and their loyalty is insufficient. So they are prone to churn when they find better options from competitors. In addition, spend per basket of this group is lowest. Therefore, to retain these customer, the company can do campaigns to increase their spend per basket. For example, bundling sales and lottery campaigns when spend per basket surpasses a certain number.

"**Low_About-to-sleep_No-preference**" group is also necessary to be pay more attention, because they are about to sleep, with low recency, low frequency and low monetary(PUTLER n.d.). Recalling non-active customers is cheaper than acquiring those who are totally lost. The company can offer some special discounts to this group of customers.

However, to analyze customers better, the company can do further. Exploring demographics distribution, such as gender, age and occupation, in each group resulted from this report is a good choice. If there are significant character in one group, more particular marketing strategy can be used. For instance, if office workers are majority in "Potential-loyalist_Shopping-dispersed_Tobacco&Alcohol", then advertising in office building may be useful.

Reference

1. Albon.C (2018*) Python Machine Learning Cookbook*. United States of America: O'Reilly Media
2. Blattberg R.C., Kim B.D., Neslin S.A. (2017) *Database Marketing Analyzing and Managing Customers.* United States of America: Springer
3. PUTLER n.d. *RFM Analysis for Successful Customer Segmentation.* [online] available from https://www.putler.com/rfm-analysis/#visuals
4. Wikipedia n.d. *Silhouette(Clustering)* [online] available from https://en.wikipedia.org/wiki/Silhouette_(clustering)