Coursework 1

Customer Analytics

Module title: Analytics Specializations and Applications
Student ID number: 20171924
Current academic year: 2019 / 20

1. An Executive Summary

Description of the task

In this Customer Analytics Study Coursework1, I performed a market segmentation on four files of 3,000 customers' transactional dataset, and as a result, six segments represent all customers' behaviours. From four dataset files, I made three different datasets, '*RFM*', '*spend_habit*' and '*item_spend*'. After selecting features in each dataset, it merged into one final dataset, '*df*', and continued pre-processing to finally applying K-Means clustering methodology to make segments of customers and target specific marketing strategy.

Summary of the data

Mainly I used three datasets to perform a market segmentation. Recency frequency monetary (RFM) model, average item count, average basket spend, average spend per item from customers sample and baskets sample datasets, as well as nine categories from spends sample and line items sample datasets.

Summary of technical approach

After feature selection, the main pre-processing for various types of features would be log transformation and standardisation. The final number of features is 15, and most of them show the right skewness with 0 values. Thus I implied log1p transformation instead of logarithmic transformation to remove skewness and avoid the error of infinity value. After log1p transformation, I applied standard scaler to standardises the data. It arranges the data in a standard normal distribution. Next step is feature engineering by using principal component analysis (PCA) to reduce dimensions and eliminate the correlations in our data. The number of components is four in PCA, which it explains variance ratio over 70 per cent. Finally, K-Means clustering algorithms used to create our segments and the optimal number of a segment is six with 0.228 silhouette score.

Summary of the results

Customers from a national convenience store can be grouped into six segments. Each segment has its uniqueness as following:

Segment 1: Occasionally purchasing a small amount of expensive product, especially in dairy.
Segment 2: Occasionally purchasing a small amount of low-priced product, especially in dairy.
Segment 3: Frequently purchasing a large amount of low-priced product in all categories including health food for pets and meat.
Segment 4: Frequently purchasing a small amount of low-priced product, especially in meat and prepared food
Segment 5: Seldom purchasing a small amount of low-priced product, especially in dairy.
Segment 6: Seldom purchasing a large amount of low-priced, especially in dairy, meat and prepared food.

Summary of the insights

From the analysis, I highly recommend to target Segment 3 and Segment 4 and provide personalised marketing on two separate groups. Suggesting upselling marketing strategy on segment 3 to purchase more expensive products focused on healthy food for pets and meat. Moreover, providing discount pricing strategy by limited time particular order, especially in fruit and vegetable, and dairy products, to segment 4 to induce them to purchase a large amount of product.

2. A Feature Description section

In this section, data preparation and feature selection are mainly discussed. Before feature engineering, the final data set is merged by three data sets, '*RFM*', '*spend_habit*', and '*item_spend*'. The first dataset is '*RFM*' which includes '*Recency*, *Frequency*, and *Monetary*'. Secondly, '*spend_habit*' dataset contains '*average_item_count*, *average_basket_spend*, and *average_spend_per_item'*. Lastly, *'item_spend'* dataset holds eight categories of product.

RFM	spend_habit	item_spend		
***Recency*	*average_item_count*	*fruit_veg*	*grocery_food*	*frozen*
***Frequency*	*average_basket_spend*	*dairy*	*grocery_health_pets*	*meat*
**Monetary*	**average_spend_per_item*	*confectionary*	*prepared_meal*	****Bakery*

- All features in *'item_spend'* dataset are from *'category_spends_sample.csv'* file
- * is from *'customers_sample.csv'* file
- ** is from *'baskets_sample.csv'* file
- *** is from *'lineitems_sample.csv'* file

'RFM' dataset preparation

'*RFM'* dataset shows the last order date by '*Recency'* feature, and it indicates the freshness of the customer activity. '*Frequency'* shows how many times the customer visited in-store by days and '*Monetary'* means the purchasing power of customers by the total spend. Customer value can be measure by these three individual criteria.

More details, *'RFM'* dataset used '*customers_sample.csv'* and '*baskets_sample.csv files'*. From those two files, '*total_spend'* and '*purchase_time'* was collected and then merged into '*RFM'* dataset. '*total_spend'* feature requires to change data type into 'float64', and this allows to replace '£' and ','. '*purchase_time'* feature need to change data type into 'datetime64' to extract the year, month and day, and this made a new feature of '*purchase_day'* which is last day of each customer's purchase. Based on these data preparation, '*Recency'* feature calculated by misusing '*purchase_day'* from today, which is the day after of the very last day of purchase, 1 September 2007. '*Frequency'* feature counting the number of visits by days and '*Monetary'* feature is '*total_spend'*. To see potential correlations, corr() function is used, and '*Monetary'* and '*Frequency'* shows 0.56 correlations. This correlation is going to be eliminated in feature engineering step by using principal component analysis (PCA).

Finally, logarithm transformation applied into '*RFM'* dataset to change right skew distribution into a normal distribution. This step is essential in pre-processing because K-Means clustering technique is the plan to use in segmentation step, which seeks to find globular clusters. In this step and further logarithm transformation step in each dataset, it used log1p instead of log in python code. This is because the 0 value contains in the features, and log 0 is undefined, which is not a real number, and it gives error in python.

'spend_habit' dataset preparation

On top of that, *'spend_habit'* dataset consist of '*average_item_count'*, '*average_basket_spend'*, and '*average_spend_per_item'*. This dataset explains the average spend of customers and this information is used after clustering by determining which groups are purchasing many small items or single higher-cost products. Furthermore, it helps to indicate the type of customer.

To create these three features, *'total_quantity'* and '*total_spend'* features in '*customers_sample.csv'* file and '*baskets_sample.csv'* file is required. '*total_spend'* feature requires to change data type into 'float64', and this allows to replace '£' and ','. In *'baskets_sample.csv'* file, I created a new column name of '*basket'* which counted the number of customer visit in-store by days. This is because current '*baskets'* feature counted the number of customer visit in stores several times in days, I created a new feature as '*basket'* which scored the visits by days.

After these two files merge into the name as '*spend_habit'* dataset, feature engineering applied to make '*average_item_count'* feature to show the average number of item purchased for each visit by days, '*average_basket_spend'* feature to show the average spend in each visits by days, and ‘*average_spend_per_item’* features to show the average spend per items. The formula of each feature is the following:

average_item_count = total_quantity / basket

average_basket_spend = total_spend / basket

average_spend_per_item = total_spend / total_quantity

To see potential correlations, corr() function is used and ‘*average_item_count’* and ‘*average_basket_spend’* shows 0.91 correlations. This correlation is going to be eliminated in the next step by using principal component analysis (PCA). Furthermore, log1p transformation applied to change right skewness to the bell shape.

***‘item_spend’* dataset preparation**

The last dataset is *‘item_spend’* that holds eight categories of the product, *‘fruit_veg’*, ‘*dairy’*, ‘*confectionary’*, ‘*grocery_food’*, ‘*grocery_health_pets’*, ‘*prepared_meal’*, *‘frozen’*, ‘*meat’* and *‘bakery’*. From this dataset, it gives the insight to distinguish customers in each segment by the different categories of product. These features can analyse customer’s interest and preference.

Step 1. Removing the *‘bakery’* feature because all the value is zero.

Step 2. Adding a new *‘bakery’* feature from *‘lineitems_sample.csv’* file

Step 3. Removing features that have zero values until 25 per cent. This includes *'lottery’*, *‘cashpoint’*, ‘*discount_bakery’*, ‘*practical_items’*, ‘*tobacco’*, ‘*drinks’*, ‘*deli’* and *‘seasonal_gifting'* features.

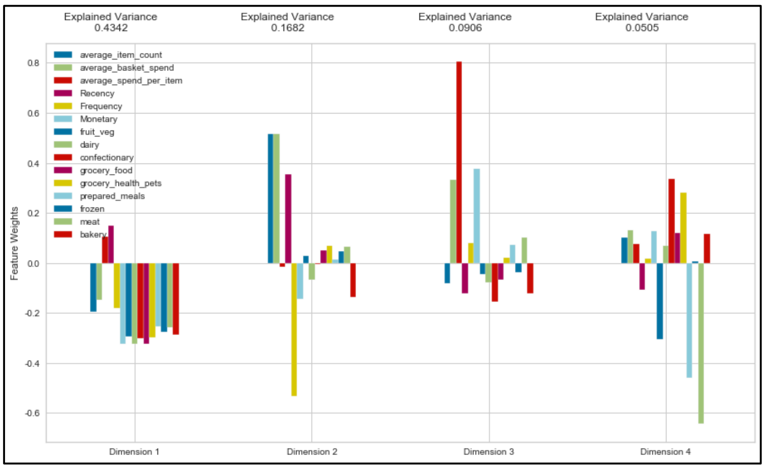
Step 4. Removing features that total amount of spend is less than £ 100,000 and this includes' *newspapers_magazines’*, ‘*soft_drinks’* and ‘*world_foods'* features.

‘*item_spend’* dataset is from ‘*category_spends_sample.csv’* file and ‘*lineitems_sample.csv’* file. After importing these two files, it requires to replace ‘£’ and ‘,’, and change data type into ‘float64’. Several features are needed to delete in these files. Firstly, the *‘bakery’* feature removed because all the value is zero. Secondly, however, from ‘*lineitems_sample.csv’* file, the spend of each customer on ‘*bakery’* can be found. Thus adding a new ‘*bakery’* feature is a must. Thirdly, after statistical analysis of each features by using describe() python code, features of zero value until 25 per cent is deleted. *'lottery’*, ‘*cashpoint’*, ‘*discount_bakery’*, ‘*practical_items’*, ‘*tobacco’*, ‘*drinks’*, ‘*deli’*, ‘*seasonal_gifting'* is included in this step. These features show highly left-skew distribution containing extreme outliers, and this might interrupt general analyses. After the third steps, a total of 12 features remain. The last step is removing features that total amount of spend is less than £ 100,000 and this includes *'newspapers_magazines’*, ‘*soft_drinks’*, ‘*world_foods'* features. From 12 features, only three features’ total amount of spend is lower than £ 100,000, and the other nine features’ total amount of spend is between £ 106,398 to £ 213,908. By removing three features that the total amount of spending lower than £ 100,000, it diminishes total variance. Through the third and last step, it dealt with the concentration spend issue that mentioned in the message from the company’s Chef data officer.

To see potential correlations, corr() function is used, and *‘fruit_veg’*, ‘*diary’*, ‘*grocery_food’* and other features show high correlation with other feature that over 0.6, however, these correlations are going to be eliminated with PCA. Moreover, log1p transformation applied to change right skewness to the bell shape. After preparation of appropriate data, three datasets ‘*RFM’*, ‘*spend_habit’* and ‘*item_spend’* merged into ‘*df_log’* dataset. And ready for the feature engineering to feature selection.

Feature Engineering

Last part of data preparation is feature selection by feature engineering. ‘*df_log’* dataset is composed of 15 features from 3 features in ‘*RFM’* dataset, three features in ‘*spend_habit’* dataset and nine features in ‘*item_spend*’ dataset, and all the features applied log1p transformation. Before feature engineering, standardisation is used in *‘df_log’,* and the dataset saved as ‘*df_scaled*’. The main reason for standardisation is transforming the dataset into a distribution that has a mean value 0 and standard deviation of 1. Moreover, units of each feature are different from the count, currency (£) and RFM measurement; thus, standardisation is essential to remove the groups and allow compare data that correspond to different units.



After standardisation, principal component analysis (PCA) is used to eliminate the correlations in ‘*df_scaled’* data. Furthermore, PCA provides a report of explained variance ratio of each dimension. In this part, I decided to use the number of a component from PCA that explains the variance ratio over 70 per cent. From the result of PCA, 74.35 per cent of the variance in the data is defined by the first four principal components.

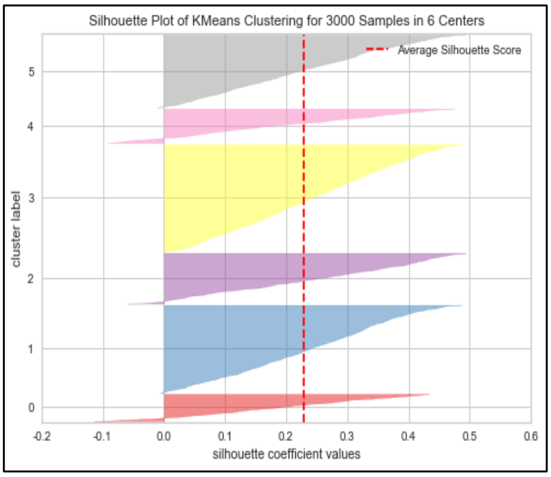
The first principal component (PC1): An decrease in PC1 is associated with increases in ‘*Monetary*’, ‘*dairy*’ and ‘*grocery food*’ spending. Even though a lot of features has a similar degree of decreases, these three features are highly correlated and represent of PC1. I might call this: ‘GROCERY_FOOD_INCLUDING_DARIY’. The second principal component (PC2): In this case, an increase in PC2 is associated with large increases in ‘*average item count*’, ‘*average basket spend*’, ‘*Recency*’ and ‘*grocery health pets’* spending. These features best present PC2. This makes sense as PC1 presents different features. And in PC2, the features in PC1 have relatively small weights. So there I might refer to the dimensions as: ‘FREQUENTLY PURCHASING_A_LOT_OF_GROCERY_HEALTH_FOOD_FOR_PETS’. The third principal component (PC3): An increases in PC3 is only associated with a large increase in ‘*average spend per item*’, and this feature best represent PC4. So here I might refer to the dimension as: ‘PURCHASING EXPENSIVE PRODUCT’. The fourth principal component (PC4): An increase in PC4 is associated with a large decrease in ‘*meat’* and ‘*prepared food*’. So here I might refer to the dimension as: ‘MEAT_AND_PREPARED_FOOD’. Thus in this model, four components of PCA result from ‘*df_scaled*’ dataset is used to make a new dataset, ‘*reduced_data’* to applying to K-Means Clustering algorithm.

3. A Customer Base Summary section

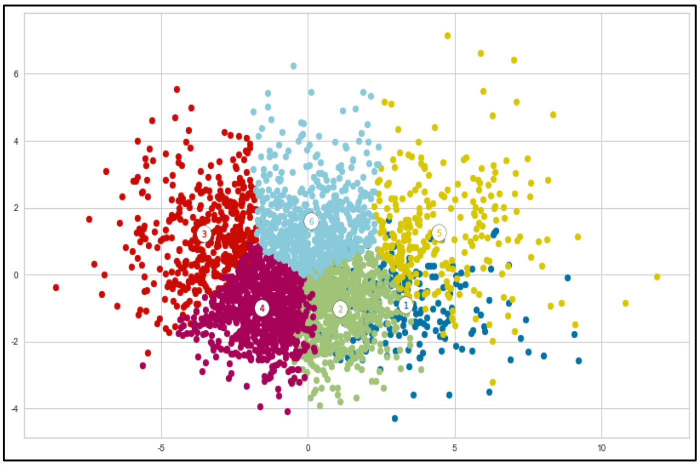
	average_item _count	***average_bakset _spend***	***average_spend _per_item***	***grocery_food***	***grocery_health _pets***	***prepared _meals***
Mean	11.27	14.80	1.39	60.01	60.91	35.48
Std	8.54	11.16	0.57	57.68	69.81	41.24
50%	8.73	11.77	1.25	44.03	39.08	23.10
	fruit_veg	***dairy***	***confectionary***	***frozen***	***meat***	***bakery***
Mean	69.46	71.30	57.35	35.47	54.74	38.21
Std	70.50	57.97	55.96	41.48	67.80	36.50
50%	50.94	56.88	42.29	22.28	32.93	29.27
	Recency	***Frequency***	***Monetary***			
Mean	8.12	65.19	769.41			
Std	20.94	47.47	552.77			
50%	2	53	627.17			

In this section, a total of 15 features’ general the descriptive statistic will be discussed. Since all the features’ standard deviation and units of each features are different, it is impossible to compare each other; thus, each cell’s background is coloured by based on values, small values with light colour and large amounts with dark colour. Generally, customers buy around £ 1 price of 11 products in each visit, and the last visit of 2 days ago. Customer visits 53 times in total and spends £627. Mostly, customer purchases a lot of fruit and vegetable, and dairy and least in frozen food.

4. A Segmentation Methodology section



In this section, description and justification of K-Means clustering methodology are covered. ‘*reduced_data’* that diminished into four components from ‘*df_scaled*’ by using PCA is applied K-Means clustering Algorithm. To find the optimal number of K in K-Means, silhouette score is used to measure the performance of each K-Means from 2 clusters to 10 clusters. Silhouette score represents consistency within groups of data, and 6 clusters in this model show the highest silhouette score, 0.228. Using the number of 2 in K in K-Means shows the highest silhouette score; however, it is a too small number of clusters to represent the total characteristics of customers. From 3 clusters to 10 clusters shows similar silhouette score nearby 0.2, but the number of 6 groups are the highest score.



Since six is the optimal number of segments for the clustering algorithm using a scoring metric, this visualisation is the results of clustering. Segments 3, 4 and 6 are relatively well divided without much noise. On the other hand, Segment 1 and 2 shows, sort of mixed in the middle of two groups, while segment 1 and 5 also incorporated in the middle of two groups. To target a specific group and suggest personalised marketing strategy would be more reasonable with well separated and high cohesiveness group, which has their uniqueness.

5. A Results section

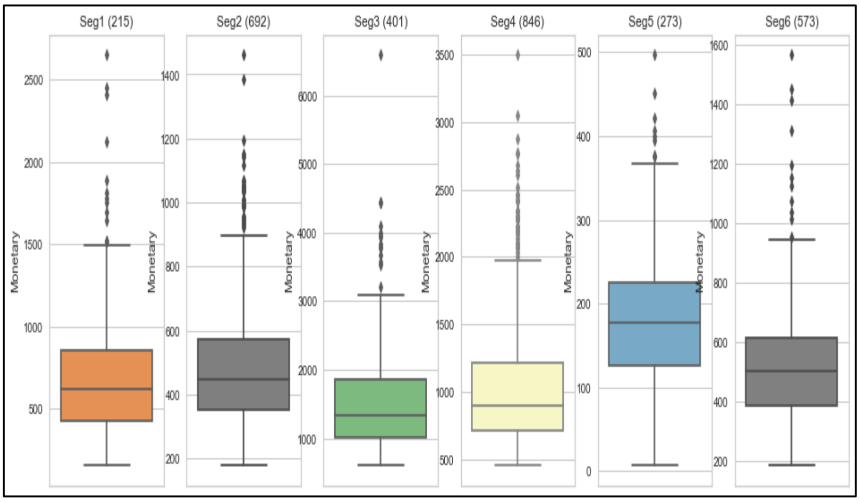
In a results section, analysis and descriptive of results will be dealt with, including overall customer base summary, Individual statistical reviews of clusters and Pen Portraits of Clusters.

Overall customer base summary

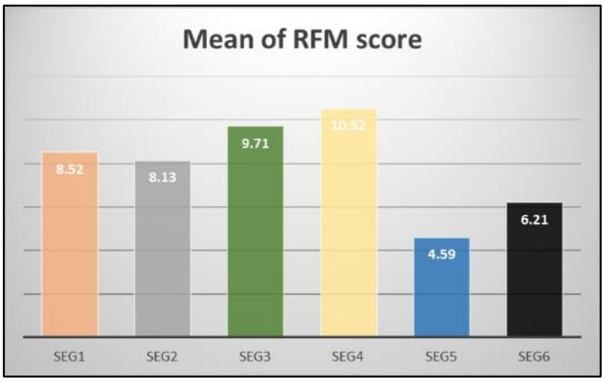
Segment 1: 215 customers, occasionally purchasing a small amount of expensive product, especially in dairy.
Segment 2: 692 customers, occasionally purchasing a small amount of low-priced product, especially in fruit and veg, and dairy.
Segment 3: 401 customers, frequently purchasing a large amount of low-priced product in all categories including health food for pets and meat.
Segment 4: 846 customers, frequently purchasing a small amount of low-priced product, especially in meat and prepared food
Segment 5: 273 customers, seldom purchasing a small amount of low-priced product, especially in dairy.
Segment 6: 573 customers, seldom purchasing a large amount of low-priced, especially in dairy, meat and prepared food.

Individual statistical summaries of clusters

In the individual clusters analysis, 3,000 customers are divided into six segments, and the following boxplot used one of the feature in the RFM model, monetary. Comparing each customer’s total spend by displaying the distribution helps to understand each segment’s values and outliers.



More details, Seg3 shows the highest variance among the six segments, while Seg5 has the smallest variance. In terms of average point, seg3 and seg4 show highest in mean and median. All of the segments have outliers, however, I decided to include all of the outliers in K-Means clustering technique. The impact of outliers in monetary values can be diminished by using an RFM score with recency and frequency value.

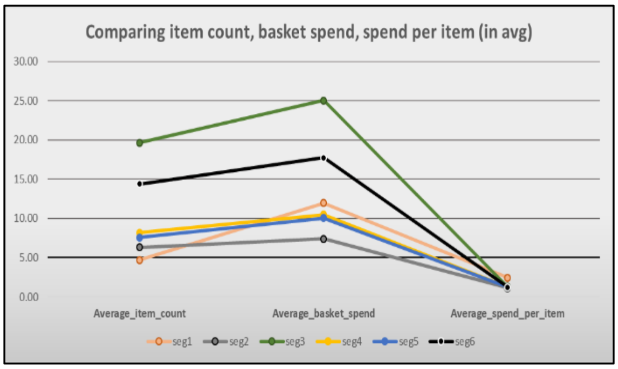


In the analysis, '*RFM*' dataset contains three features, '*Recency*, *Frequency*, and *Monetary*', however, in the report 'mean of RFM score' graph is used, which represent three features. From this 'mean of RFM score graph', Segment 3 and Segment 4 shows the highest purchasing power group.

Values in each '*Recency*', '*Frequency*' and '*Monetary*' feature is assigned to a range of 1 to 4 and saved in new features '*R*', '*F*' and '*M*'. For instance, value 1 in '*Recency*' feature, which means the customer went to store on the last day, is assigned to value 4 in a new feature. Thus these new three features contain the numbers between 1 to 4, and this allows to create '*RFM_score*' feature by adding every three numbers. Now, all the customers have their RFM score between 3 to 12.

Comparing six segments with RFM score could find out which segment are more likely to purchase soon. The higher RFM score group has more purchasing power than other groups. From this graph, segment three and segment 4 shows over 9.5 RFM score, which is higher than the other four groups. Seg1 and 2 indicated similar RFM score around 8.3, and Segment 5 and 6 shows the lowest RFM score as below 7.

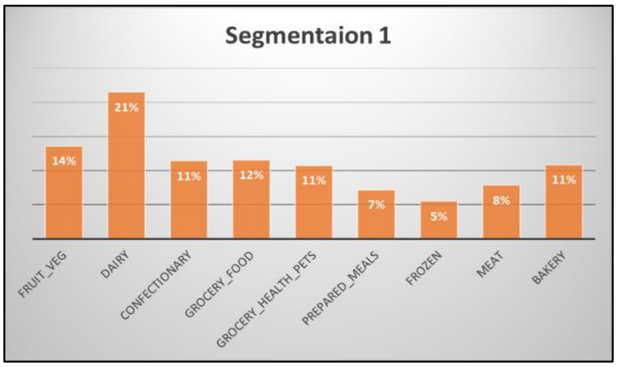
'Comparing item count, basket spend, spend per item (in avg)' graph is based on '*spend_habit*' dataset, which explains overall customer spending habits by the average spend of customers.



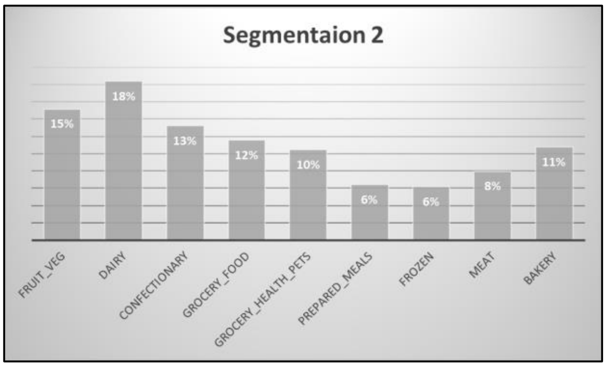
From the graph, six segments can be divided into three groups based on '*average item count, average basket spends,* and *average spends per item'* features. Firstly, Segment3 and Segment6 show high numbers in '*average item count*' and '*average basket spend'*, however, a low amount in '*average spend per item*', this could interpret as purchasing a lot of low-priced products. Secondly, Segment1 shows a small number in '*average count item*' with relatively high numbers in '*average basket spend'* and '*average spend per item'*. This means that Segment1 is purchasing only a small amount of expensive products. Lastly, Segment2, 4 and 5 can be grouped based on a low number in all three features. These segments are purchasing a small number of low-priced products.

Pen Portraits of Clusters

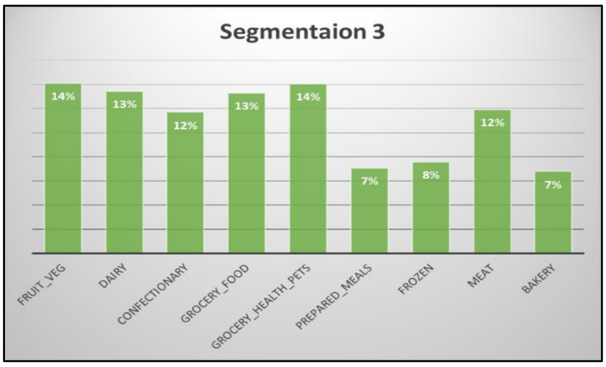
From 'Comparing item count, basket spends, spend per item (in avg)' graph, overall spending behaviours can be found out, and 'Mean of RFM score' graph explains purchasing power of each segment. These two insights need to be connected with more specified purchasing habit, such as answering the question of ‘what kind of product categories do they buy usually?’. This question can be answered with ‘*item_spend’* dataset analysis that shows the ratio of each nine categories, fruit and veg, dairy, confectionery, grocery food, grocery health pets, prepared meals, frozen, meat and bakery. All segments show a small amount of purchase on prepared meals and frozen category, and a large portion of spending in fruit and veg and dairy category. Main insights with the above two graphs with ‘*item_spend’* dataset analysis are the following:



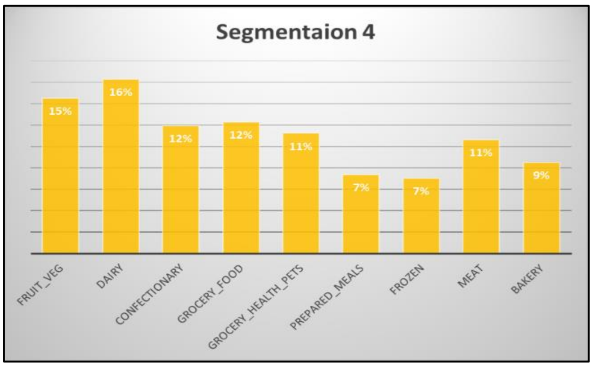
Segment 1: Total of 215 customers is clustered into segment 1. The primary behaviour is purchasing a small number of expensive products occasionally. As it can see through the bar chart above, diary, fruit and vegetable show large portion over 15 per cent. Segment 1’s unique characteristic is it shows a large amount of purchase on dairy products, and it is connected to PC3, ‘PURCHASING EXPENSIVE PRODUCT’.



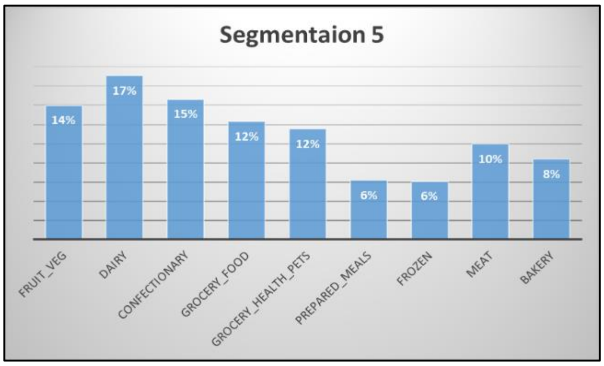
Segment 2: Total of 692 customers is clustered into segment 2. The primary behaviour is purchasing a small number of low-priced products occasionally. Fruit, vegetable and dairy show large portion above 15 per cent. Segment 2’s main feature is they also purchase a large amount of fruit and vegetable and dairy. Segment2 is represent of PC1, ‘GROCERY_FOOD_INCLUDING_DARIY’.



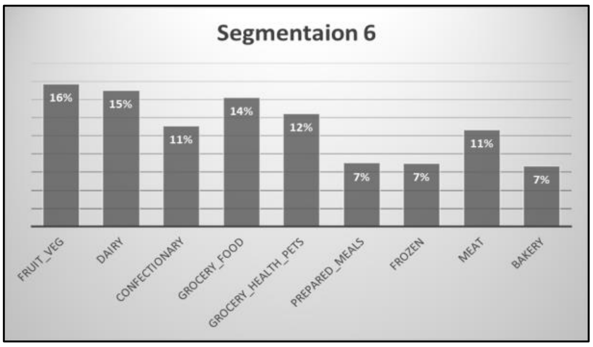
Segment 3: Total of 401 customers is clustered into segment 3. Purchasing a large number of low-priced products frequently and recently is their significant behaviour. More details, there is no category that over 15 per cent purchase, however, Segment 3’s important insight is all types purchase similarly, including health food for pets and meat. Through PC2 and PC4 can be explains segment 3.



Segment 4: Total of 846 customers is clustered into segment4. The dominant behaviour is purchasing a small number of low-priced products recently and frequently. Fruit and vegetable show large portion above 15 per cent. Segment 4’s main characteristic is they also purchase a large amount of grocery health for pets. This characteristic is linked to PC 4.



Segment 5: The 273 customers are clustered into segment 5. Seldom purchasing a small number of low-priced products is their primary behaviour. Fruit, vegetable and dairy show large portion above 15 per cent. Segment 5’s uniqueness is confectionary take a lot of portion in total spend. Moreover, PC1 is connected to these features.



Segment 6: Total of 573 customers is clustered into segment 6. The primary behaviour is seldom purchasing a large number of low-priced products. Fruit, vegetable and dairy show large portion above 15 per cent. Segment 6's main feature is they also buy a large number of confectionery products. In terms of PCA result, PC1, 'GROCERY_FOOD_INCLUDING_DARIY' and PC4, 'MEAT_AND_PREPARED_FOOD' well explains this uniqueness.

6. A Summary Section

Summary of result

In this coursework, 3,000 transactional data clustered into six segments using 15 features. Dividing into six groups depending on their spending habit, type of product and RFM model. By RFM and spending habit, it helps to grasp the overall concept of customer's behaviour, and by analysing product's categories, it could figure out what items they most interested in.

Recommendation for two segments

From the analysis, Segment 3 and Segment 4 need to be targeted into the specific personalised marketing strategy. Segment 3 has 401 customers that are frequently purchasing a large amount of low-priced product in all categories, including health food for pets and meat. Segment 4 contains 846 customers who are regularly purchasing a small amount of low-cost product, especially in fruit and vegetable and dairy. From the result of these two segments, upselling marketing strategy should be on segment 3 to purchase more cost healthy food for pets and meat products. In both type of products has premium products such as pet snacks that have various vitamins or high-quality meats. Furthermore, a limited period of discount pricing strategy is suggested to Segment 4 to induce purchasing a large amount of product in fruit and vegetable and dairy products. Since customers in Segment 4 visit store frequently, the next approach to them would be purchase a large amount of product instead of upselling to buy expensive products. The marketing strategy should be focused on meat and prepared food product.

Draw together any key that has dropped out

The main dropped out features would be half of the category spend features. Since those categories' total spend is low and has 0 values more than 25 per cent, I did not use in the analysis. However, if I could use them, I might have specified segment that used minor features, for instance, a segment that has a lot of spend in tobacco or drinks.

Suggestions for further analysis

Companies need a deeper understanding of their customers amid intensifying competition. The development of information technology allows data on each activity; the strategic use of customer information has become very important. Thus, the company must use the customer segment to do advanced marketing strategy in terms of customer management based on deeply understanding their customers' behaviour. Further customer segment analysis, customer needs to be narrow into groups based on their loyalty level, and make each segment. The level of loyalty could be defined into customers who are using loyalty card such as premium membership card. The other benefits of using loyalty card are that it allows collecting more diverse data such as age, sex, occupation and location.

From loyalty card to dividing customers based on the degree of loyalty, diversifying clustering algorithms could improve clustering performance. In this coursework, only K-Means clustering is used, however, hierarchical and density-based clustering also implied and should be compared with each other. From a different point of view, it could search for various aspects of customers.