# Customer Segmentation Analytics Report

**Module Name:Analytics Specialization and Applications**

**Academic Year: 2018/19**

**Name: Shangrui Zhao**

**Student ID: lixsz44**

**Word count: 2439**

## Executive Summary

By analysing four files 3000 customers over six months, this report aims to segment customers into six segments with respective profiles and a statistical summary for the national convenience store chain. Since given data is transactional, behavioural analysis is stressed with 'recency', 'frequency', 'average_spend', 'average_quantity' as inputs companying with topic modelling features: 'we live here', 'tobacco love', 'drinks love', 'we need cash', 'lottery love' and 'we don't cook' done by non-negative matrix factorization (NMF). Although those features are selected, the topic modelling features are also run separately to have a comparison. Three techniques: K-means clustering, Hierarchical clustering, and Spectral clustering are utilised, and with ten features and topic modelling features alone to run multiple clusters. By comparing the silhouette score, segmentation distribution and characteristic different, the final cluster is K-means with 6 clusters. The detailed pen portraits of each segment can be found in the result section. As the high average revenue per person from segment 2 and 3 with high frequency and low recency, they are recommended as the target group. Positioning on 'quality', 'convenience', 'joy' and 'experience' can be more efficient. As a further improvement, more detailed data are suggested to collect for further demographic, geographic psychographic analysis. Besides, a recommender system will be helpful and network analysis to identify influential products.

## Feature description

### Data preparation

There are four tables which are 'baskets', 'category_spends', 'customers' and 'lineitems'. For better data quality, null and duplicated records are checked. Only 114 duplicated records are found in 'lineitems'. However, there are some negative in 'baskets' 'quantity' and 'baskets_spend' columns which result from the customer return. Since there are only 19 occurrences in 195547 records, negative occurrences are deleted. Also, 'lottery' in 'category_spends' displays negative as a result of lottery redemption. As only 29 records and smaller value, the negative value is replaced by 0. For further exploration and modelling, '£' is deleted and data type is converted to float. 'bakery' in the category_spends table are all zero value. The column is deleted.

### Feature Selection and Engineering

Based on the given four files, point-of-sale data is given for behavioural analysis but lack enough data to do the demographic, geographic, psychographic analysis. Hence segments based on behaviours are examined. RMF segmentation (recency, frequency and monetary) is utilised to understand purchase occasion, purchase behaviour (loyalty or not), and how profitable the customer is. As required by the Chief Data Officier, 'monetary' represents the 'total_spend' to indicate the value of a customer and 'frequency' represents the 'Number_of_visit' to show the different shopping patterns over time. 'average_spend' and 'average_quantity' in 'customer' table represent the 'average_basket_spend' and 'average_item_count' indicating the involvement of products. 'Average_spend_per_item' is calculated through 'total_spend' divided by 'total_quantity' from 'customer' table. Figure 1 shows the scatter plot between RFM variables. As a result of the linear relationship between 'frequency' with 'monetary' and 'average_spend' with 'average_spend_per_item', 'frequency' and 'average_spend_per_item' are dropped to avoid double counting high correlated features influencing distance measure.

Another critical factor to consider when analysing behaviour is the motivation for consumption which means to understand categories which customer engage with. After cleaning, there are 19 categories, to discovering the trend of purchasing behaviours and have a comprehensive picture of each segment topic modelling are utilised which also avoid the curse of dimensionality and deal with the high correlation between categories. Principle component analytics (PCA), Non-negative matrix factorisation (NMF) and Latent Dirichlet Allocation (LDA) are used. 'Category_spends' are logged for normal distribution before fitting in the clustering model. As some value in category spends are between 0 and one resulting in a negative value after logarithm which is unacceptable for NMF and LDA. Since the small occurrences and less in value, they are replaced by 0.

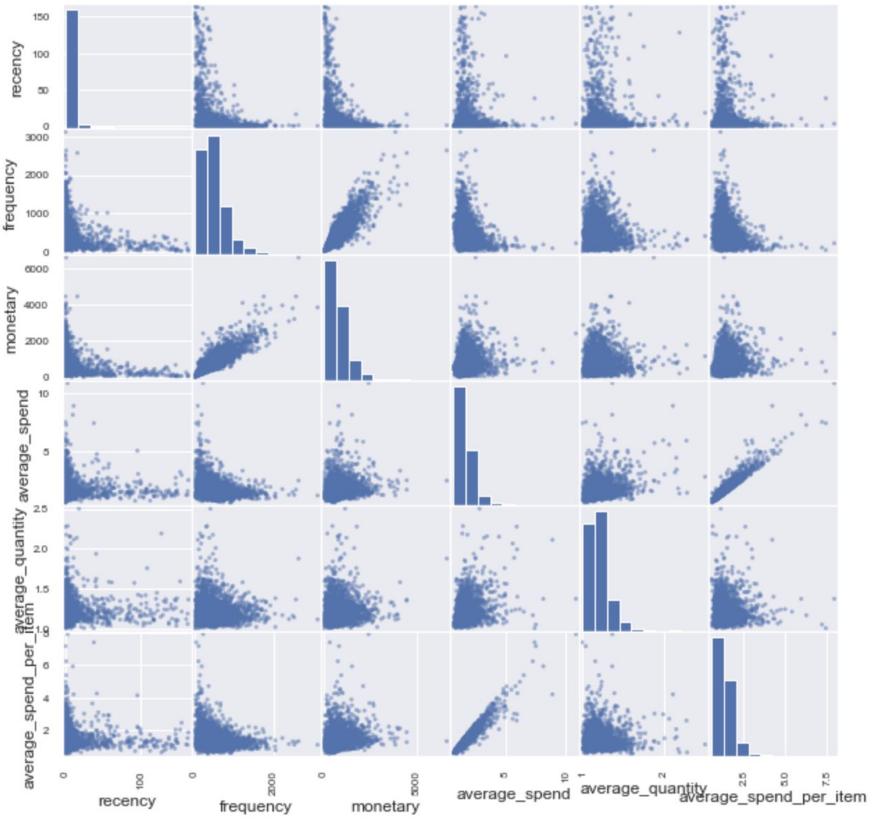


Figure 1 scatter plot

In all three topic modelling methodology, the bar chart represents a new feature, and the heights mean its contributions to the new feature. In PCA 72.95% of the variance in the data is explained by the first six principal, n_components equals to six are selected as well for NMF and LDA. As the negative value in PCA bringing complication in interpretations of new features and the insignificant differences for original feature weights to each new feature in LDA, NMF is employed for topic modelling. Figure 2 visualises the result of NMF. For better understanding the meaning of new features, Figure 3 list the new features with top 5 important original features in contribution to them respectively.

Based on the weighted importance from original feature, topic 0 to 5 are defined as ‘we live here', ‘tobacco love', ‘drinks love', ‘we need cash', ‘lottery love',‘we don't cook'. Since in topic 0 ‘dairy', ‘fruit_veg','grocery_food','confectionary','grocery_health_pets', 'meat','frozen', ‘prepared_meals' all highly contributed to the new feature which means people influenced by this feature shop all they need in the convenience store. Other are titled for their highest weighted original feature.

To summarise, to segments based on behavioural features, RFM is illustrated with topic modelling by NMF to understanding the trend and motivation for shopping. The final features to put in clustering model are ‘recency','frequency','average_spend','average_quantity', ‘we live here', ‘tobacco love', ‘drinks love', ‘we need cash', ‘lottery love',‘we don't cook'. For comparison, features generated by NMF is also put in the model independently. Although generating those features for clustering, there are many features handed not in use but still important such as purchasing time implying the shopping time and also product information specifying the most popular products. Those factors will be examined after clustering to reveal the characteristics of segments.

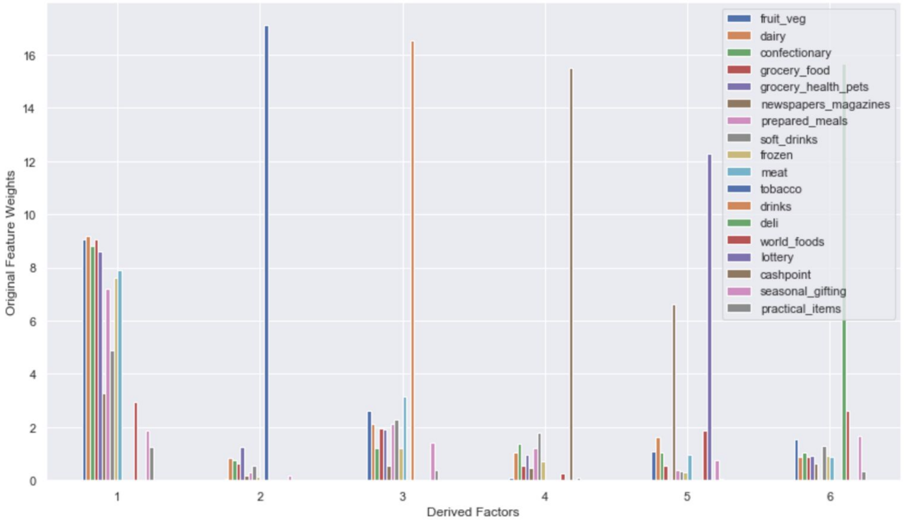
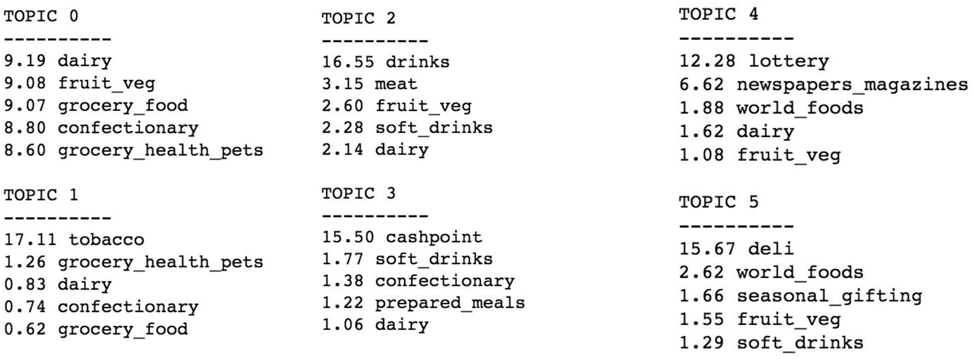


Figure 2 Non-negative matrix factorization



```
TOPIC 0
----------
9.19 dairy
9.08 fruit_veg
9.07 grocery_food
8.80 confectionary
8.60 grocery_health_pets

TOPIC 1
----------
17.11 tobacco
1.26 grocery_health_pets
0.83 dairy
0.74 confectionary
0.62 grocery_food

TOPIC 2
----------
16.55 drinks
3.15 meat
2.60 fruit_veg
2.28 soft_drinks
2.14 dairy

TOPIC 3
----------
15.50 cashpoint
1.77 soft_drinks
1.38 confectionary
1.22 prepared_meals
1.06 dairy

TOPIC 4
----------
12.28 lottery
6.62 newspapers_magazines
1.88 world_foods
1.62 dairy
1.08 fruit_veg

TOPIC 5
----------
15.67 deli
2.62 world_foods
1.66 seasonal_gifting
1.55 fruit_veg
1.29 soft_drinks
```

Figure 3 New features importance

## Customer Base Summary

As figure 4 suggests, the distribution of these four features is all left skewed with a long tail in the right. Since the day uses to calculate recency is first September 2007, many zero values are existing in recency. Also, many customers are no more active again. The kurtosis is high with a long tail. The mean for ‘frequency', ‘averge_spend' and ‘average_quantity' are 487, 1.68 GBP and 1.20 respectively suggesting 20.29 times visit a week. Figure 5 ranks the revenues per category over the six months from the category spends table indicating tobacco is most revenue driving one followed by dairy, fruit vegetables, drinks, Grocery health pets, and grocery food while less revenue from practical items, discount bakery and bakery.

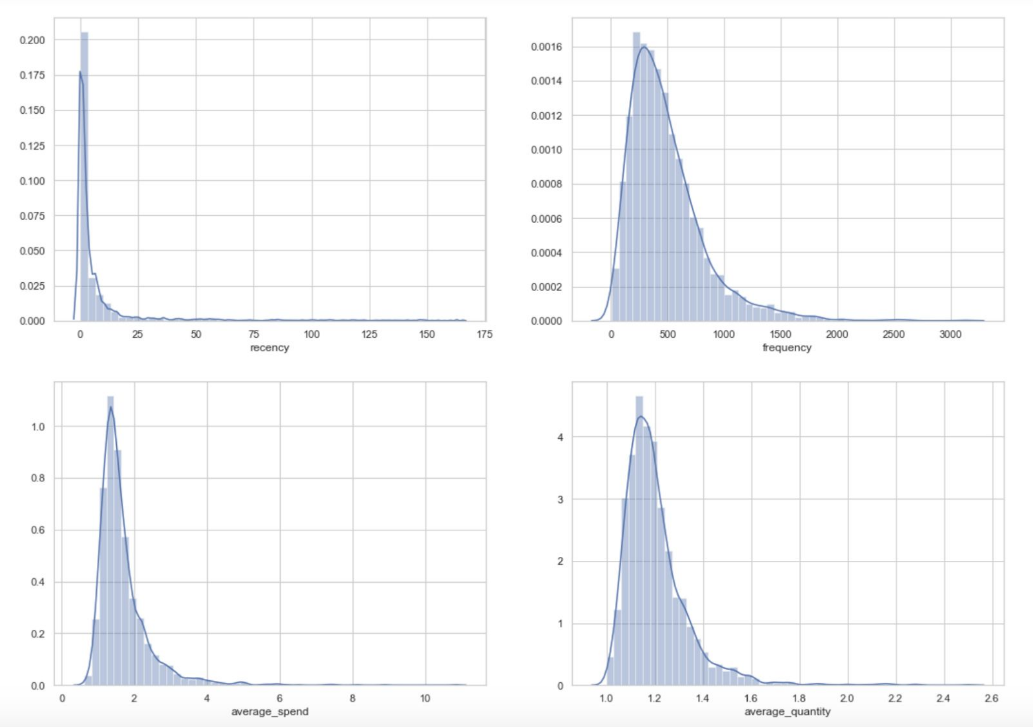


Figure 4 Histogram about ‘recency’ ‘frequency’ ‘average_spend’ ‘average_quantity’

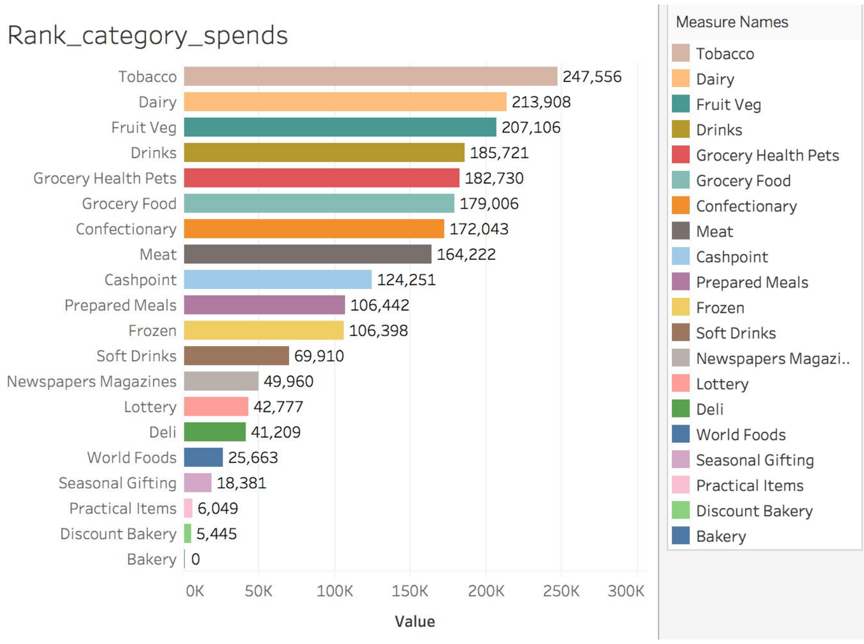


Figure 5 Tableau, Rank_category_spends

## Segmentation Methodology

K-means, hierarchical, Spectral Clustering is selected as clustering methodology to have an initial examination about the clustering performance based on silhouette score which indicated how well the in-class similarities and between classes discrepancies. Since the units for RFM are diverse hence logarithm is used on RFM features for pre-processing. ‘we live here', ‘tobacco love', ‘drinks love', ‘we need cash', ‘lottery love',‘we don't cook' runs as model II, and they are run with ‘recency','frequency','average_spend','average_quantity' as model I.

K-means is selected as its high speed and low in cost in generating clusters. The parameter for n tested from 5 to 7 as the number of segmentation requirement. Cosine similarity is selected as the distance measure since the topic modelling features. Hierarchical clustering is achieved through sklearn AgglomerativeClustering(). Since the data is already logarithmic hence ‘ward' linkage is suitable for the current distribution. Since the linkage is "ward', only Euclidean distance is allowed. N is set from 5 to 7. It is selected as it does not assume spherical data distribution. N and distance measurement are set the same as K-means.

Adopting silhouette score as clustering performance comparison, top 3 scores are from Spectral Clustering (model I, n=6), K-means (model II, n=6), K-means (model I, n=5) which are 0.6237, 0.3517, 0.2729 respectively. High in silhouette score not necessarily mean a significant clustering in business insight. Hence, clusters from these three methods are listed to have a detailed examination. As a result, the distribution of clusters generated by spectral clustering is exceptionally uneven with one dominant cluster having 2636 customers which means little targeting. Hence, the two K-means are selected as the final models for the reason that high in silhouette score and workable cluster size. Besides, K-means has O(n) in time complexity which is computationally faster and cheaper than the other two algorithms with the increase in customer data for clustering. Although it is not possible to visualise clusters with all features, Figure 6 shows the 3D plot of clusters with for three features from the two K-means models.

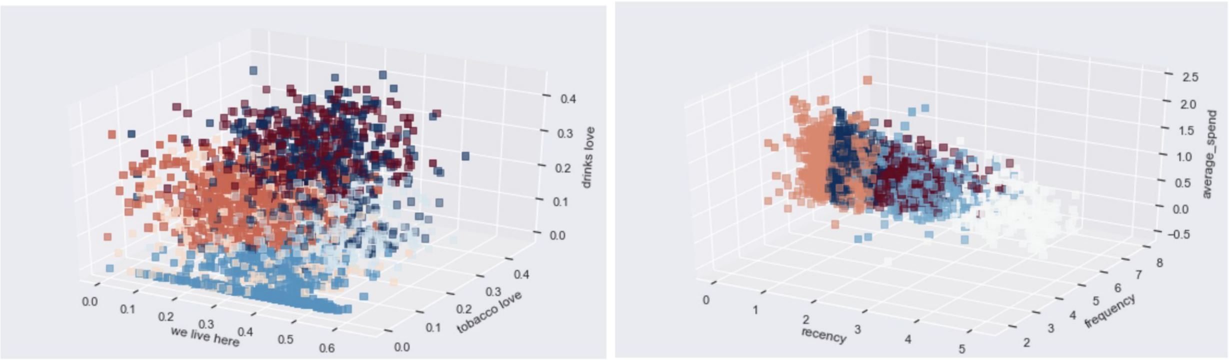
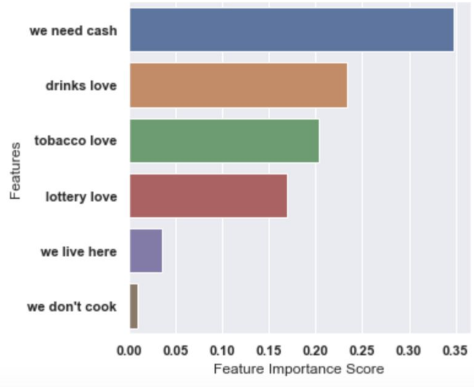


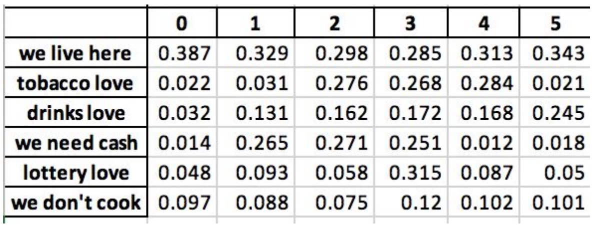
Figure 6 3D plot of clusterer I and II

**Result**

As K-means is a central clustering technique, central points of each cluster can be seen as a representative of each segmentation. Figure (tableau) visualise the centres of the model I. There is a significant difference in values, but their contributions of categories are similar which is harmful in understanding customer consumption topic and motivation (see Tableau, Cluster1_category_spends). Hence the final segments are finalised using the cluster II.

Assisting by decision tree classifier, figure 7 exhibits the feature importance which indicating ‘we live here’ and ‘we don’t cook’ contribute less in classify customer segmentation. Consisting with NMF value (figure 8), ‘we live here’ and ‘we don’t cook’ is the common theme for all clusters while the other four features can distinguish clustering significantly. The violin plot (figure 9) shows the characteristics of each cluster in ‘recency’,‘frequency’,’average_spend’,’average_quantity’. Segments 3 possesses a higher ‘frequency’ and a lower ‘ recency’ than others indicating they are more frequently and actively engage with transactions. Referring to ‘average_spend', segment 0, 1 and five distributed more concentrated and lower than other segments which means they consume less each time compared to segments 2,3 and 4. There is no considerable variation in ‘average_quantity' where segment 3 distributes slightly lower suggesting that they may engage more with higher value products. Overall, segments 2 and 3 high in ‘monetary'. ‘Average_spend_per_item' hints segments 2,3 and 4 possess actively engage with valuable products.





| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| we live here | 0.387 | 0.329 | 0.298 | 0.285 | 0.313 | 0.343 |
| tobacco love | 0.022 | 0.031 | 0.276 | 0.268 | 0.284 | 0.021 |
| drinks love | 0.032 | 0.131 | 0.162 | 0.172 | 0.168 | 0.245 |
| we need cash | 0.014 | 0.265 | 0.271 | 0.251 | 0.012 | 0.018 |
| lottery love | 0.048 | 0.093 | 0.058 | 0.315 | 0.087 | 0.05 |
| we don't cook | 0.097 | 0.088 | 0.075 | 0.12 | 0.102 | 0.101 |

Figure 7 Feature importance score Figure 8 NMF value

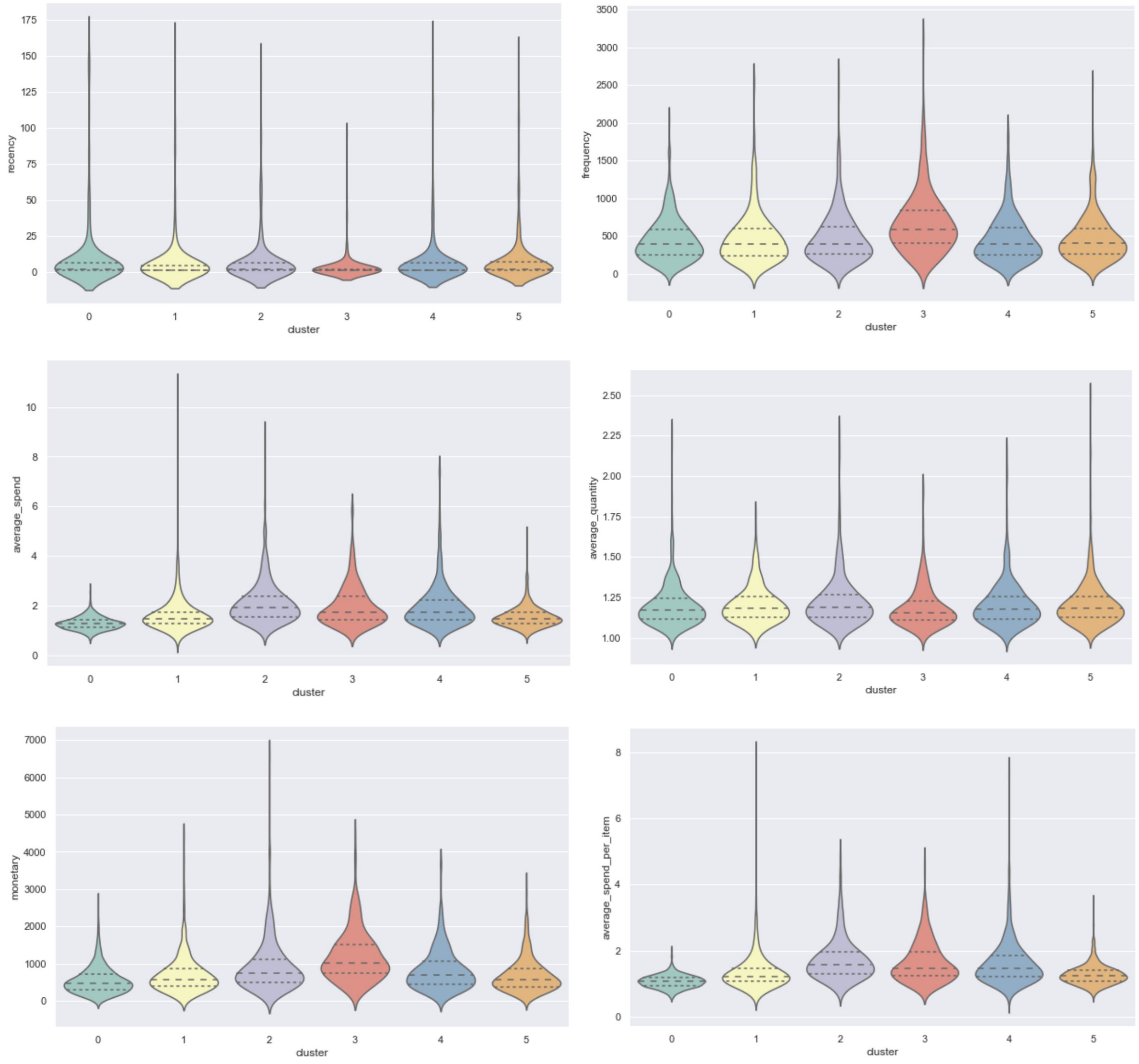


Figure 9 Violin plot

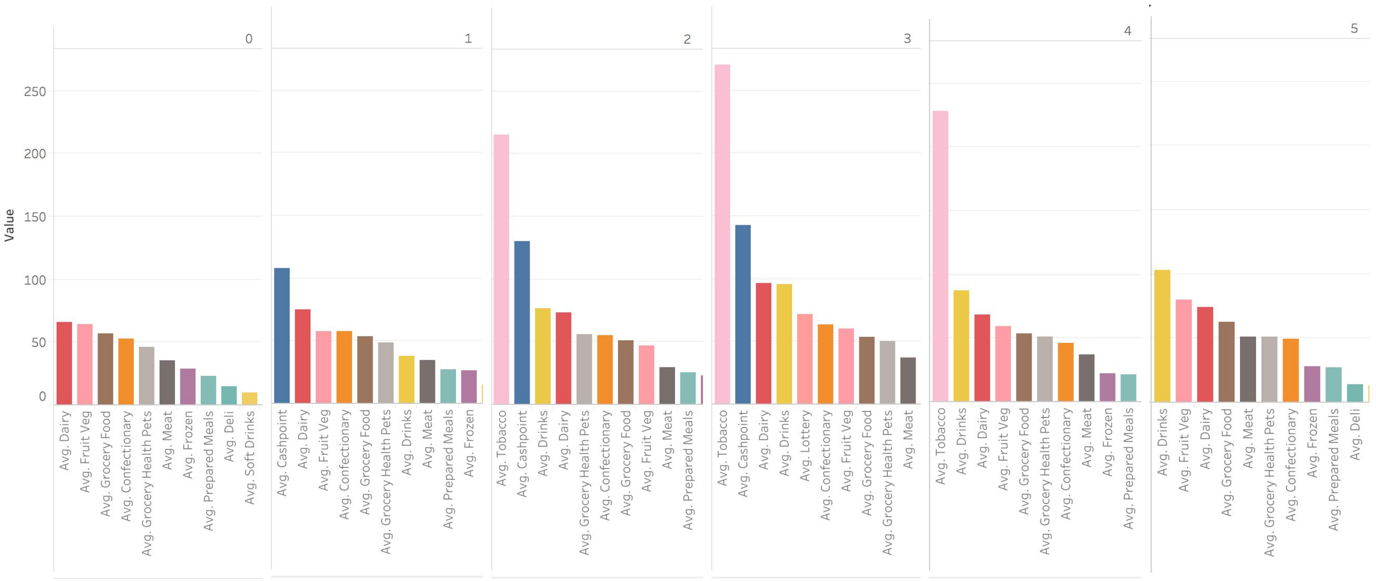


Figure 10 Tableau , Clusterer2_category_spends(final clusterer)

Segment 0 (736 customers)

As figure 10 indicates the ‘we live here' is the main topic for the consumers. The consumers actively buy products in ‘Dairy', ‘Fruit Veg', 'Grocery Food', ‘Confectionary', ‘Grocery Health Pets' and ‘Meat'. Sometimes they also buy some prepared meal or deli. They gain basic meal materials in this store and shop 18 times a week, and high spend in confectionary and pets products may indicate they have children and pets. From the comparison on recency, average_spend_per_item, weekly_frequency_per segment (figure 11), the average spend per item is 1.08 pounds which means they may be more price sensitive. Moreover, the higher recency may indicate some are not loyal anymore.

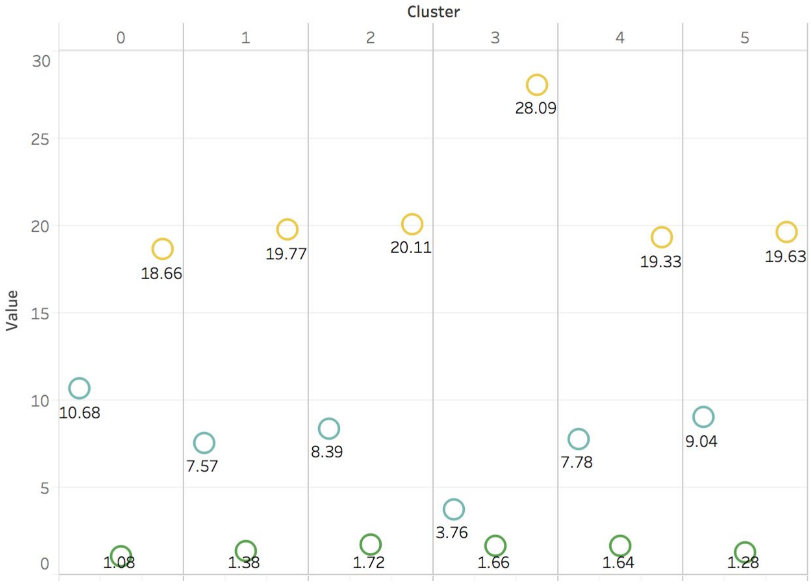


Figure 11 Tableau, Comparison on recency, average_spend_per_item, weekly_frequency_per seg(final clusterer)

Segment 1 (366 customers)

This segment focus on ‘we live here' and ‘we need cash' mostly. They mostly come to the store for getting cash followed basic meal materials. They visit the store 19 times a week and with a moderate average spend per item. The high concentration on cashpoint may suggest they less willing to use the bank card in daily life, and they may also have kids and pets.

Segment 2 (393 customers)

‘We live here', ‘tobacco love' and ‘we need cash' stimulate consumption for this division. Tobacco is their purchasing mission accompanied by acquiring cash, drink, and dairy. They may follow a casual lifestyle as they are less interested in obtaining fresh meal material rather than tobacco, cashpoint and drinks. They appear 20 times weekly in store and less price sensitive as having a 1.72 pounds average spend per items.

Segment 3 (309 customers)
‘lottery love' is the common shopping theme here. Generally, consumption motivation of this segment is likewise to segment 2 beside segment 3 engage more with 'lottery'. They come most frequently in all segments which is 28 times per week and more active with the least recency.

Segment 4 (468 customers)
‘We live here’, ‘tobacco love’ motivate consumption. Again, similar to segments 2 except obtaining cash, their priority is tobacco, drinks and some fresh foods with 19 visits per week. Moreover, they are less price sensitive than segments 2 since the average spend per item is 1.64 pounds.

Segment 5 (728 customers)
‘We live here', and ‘drinks' love are the primary motivation in this segment. ‘Drinks' is most concentrated followed by some meal related categories. They reach 19 times a week, and the high recency shows their less active than other segmentation. Same as segment 1, a low average spend, they may also price sensitive.

**Summary and recommendation**

To summarise, there are six segmentations clustered. Although ‘we live here' appears to be the common theme for all segments and there are similarities in their purchasing themes, they still have sole motivation with different recency, frequency and product involvement. As shown in the figure, the revenue from segment 5 and 0 rank are the top 2 which is 508,801 GBP and 411,821 GBP respectively. Nonetheless, the uneven population distribution among different segments leads segments 3 (1,206.8 GBP) and segments 2 (928.8 GBP) ranks as top 2 in average revenue per person from each segment. This is an essential index when select which segment to target as its positively related to customer lifetime value and ROI. Accompanied by the high average spend per item and frequency with lower recency, segments 3 and 2 are selected as the final two segments to targets.

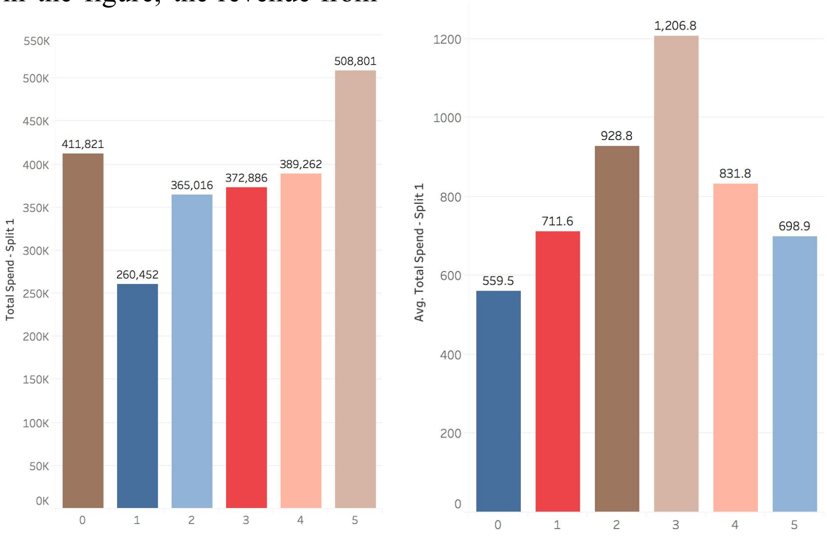


Figure 12 Tableau, Average_revenue_pp_seg (final clusterer, Total_revenue_per_seg (final clusterer)

This task is mainly focused on the behavioural and transactional data without too much information to do the demographic, geographic and psychographic analysis. With more demographic and geographic information, the characteristics of each segment can be much more precise. Surveys to investigate product preference can be done, and communication channel data can be analysed. Assisting by these, the company can better target segment 2 and 3 with a more effective channel to promote products under their shopping topic. Since they are less sensitive with the price positioning on ‘quality', ‘convenience', ‘joy' and ‘experience' may be more effective than the discount. Segment 3 possesses the highest frequency with the lowest recency may demonstrate they may enjoy the experience being present in the convenience store and segment 2 still high in recency. Mapping a customer journey may help in identifying the stage required improvement in service. The final suggestion is a two-mode network analysis can be done to identify the most influential product, and a recommender system is helpful to recommend products to similar customers.