**ANALYTICS SPECIALIZATIONS & APPLICATIONS COURSEWORK**

**SECTION-A: EXECUTIVE SUMMARY**

The report below, provides an analysis of the market segmentation performed on the four provided transactional datasets. The purpose of this analysis is to identify distinctive traits of the customer segments, the knowledge of which will streamline targeted marketing in subsequent campaigns. This will help in personalizing customer experience, which in turn will translate to higher customer satisfaction and consequently, higher customer profitability.

Behavioural segmentation was deemed most appropriate based on data exploration. After data cleaning, relevant datasets were selected based on the need to pick features that would best describe the customer segments. The features were then selected and engineered using PCA. Then using a compressed set of features, segmentation was carried out for different number of clusters. Based on the evaluation metric of silhouette coefficient for different numbers of clusters for the algorithm, 7 was picked as the final number of customer segments.

The segments can be broadly classified into high, moderate and thrifty spenders with distinct purchase pattern in each category. Based on the engagement and inclination of segments towards a certain set of categories, the cluster analysis was conducted to develop archetypes of segments and the recommendations were chalked out accordingly. The purpose of this study is to improve targeted marketing and boost sales. The recommendations are to pose tailored offers and deals to the moderate spenders so as to improve engagement and revenue. Further, loyalty discounts and benefits can be provided to ensure retention of the existing customer base.

**SECTION-B: FEATURE DESCRIPTION**

The four transactional datasets provided were studied to identify trends and patterns in customer behavior. The datasets namely customers\_sample, category\_spends, baskets\_sample and lineitems\_sample record customer behavior across store visits, item categories, customers’ individual visit and all individual purchases respectively. The features that would most distinctively define customer segments based on their transactional behavior were picked from these datasets after data exploration. Exploring the datasets highlighted the need to clean the data first. The process of data cleaning included:

1. Eliminating pound signs and commas from the rows of the tables for subsequent mathematical operations using python.
2. It could be observed that the category\_spends table had not recorded customer behavior across the bakery category. This was resolved by using the lineitems\_sample dataset to aggregate spend across categories for each customer and creating an updated version of category\_spends as a new table df\_spend.

On data exploration, behavioural segmentation seemed to be the most appropriate based on the provided datasets. The statistical dataset summary provides a fair idea about the overall behavior of the customer base as discussed at a greater depth in the following section. The behavioural pattern of the customer can be best captured from the combination of the RFM (Recency, Frequency, Monetary) features from customers\_sample table and the product type features from df\_spend, the cleaner updated version of category\_spends. So, these two tables were joined using python and the resultant table named *Data* was considered for subsequent feature selection and engineering. The steps undertaken during this process are outlined below:

1. The new resultant table Data had all the features from customers\_sample and df\_spend. To investigate the customer behaviour with respect to average spend on item, an additional feature had to be engineered using the total\_spend and total\_quantity features in customers\_sample.
2. Using a heatmap, the Pearson correlation between the pairs of features could be visualized. Based on this knowledge, the *total\_quantity* feature in customers\_sample which had an extremely high correlation (r=0.9) with baskets and total\_spend was eliminated. Multiple category feature pairs (16 pairs) have a moderate positive correlation. Dairy, grocery\_foods and confectionary are the categories which recur most frequently in the list of moderately correlated feature pairs. However, no category feature pairs displayed strong correlation. Therefore, no product category was eliminated.

So, the set of features used for segmentation are:

* Total\_spend: As an indicator of the customer’s contribution to the profit of the company, the boosting off this is our final target. This feature highlights the segments that are most profitable to the store chain.
* Baskets: This feature stands for the number of store visits of a customer which in turn will help us gauge the loyalty of the customer.
* Average\_spend: Knowledge about this will help the company tailor offers and deals for the customers. This is helpful to understand the degree of consumer engagement.
* Average\_item\_spend: This is crucial to understand the customer’s purchase pattern better as to whether the customer tends to purchase fewer expensive products or a large number of relatively cheaper products.
* Product categories: The spends of the customer across the item categories will provide accurate insights about the interests of the customer and can be employed later in sculpted advertising.

1. Using this feature set, on visualizing the data, it could be observed that very few features had a normal distribution. As per the scatterplot, most distributions were hugely skewed to the left. Hence, the non-positive values were changed to 1 as a prerequisite for smooth log transformation. This meant interpreting the 29 customers who had won the lottery the same way as a customer having zero investment in lottery. Following this, log transformation was performed on the dataset to normalize the distribution of the data across features. This is an important pre-clustering step since a fair number of clustering techniques seek to find globular clusters.
2. The total number of features in the resultant dataset alarmingly increase the risk of overfitting the data. To tackle this curse of dimensionality, principal component analysis (PCA) was employed to draw conclusions about the underlying structure of the customer base by reducing the entire set of features to fewer components, each component being a combination of the features based on the feature weights. The PCA was done in a manner to capture at least 70% variability of the data in hand. Keeping the cumulative explained variance ratio in consideration, six components were picked to explain 70.28% variability of data. These dimensions were studied and have been used to explain the overall behaviour of the customer base in the following section. These six components were used to transform the logged data and the results were used for employing the clustering technique which has been explained in the methodology section.

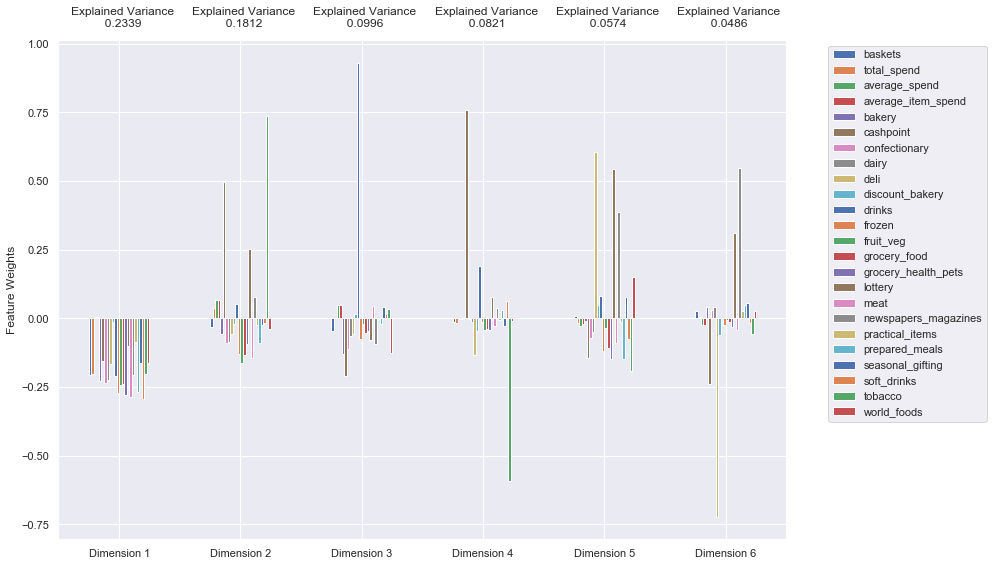


Fig: Dimensions from Principal Component Analysis

**The first principal component (PC1):** In this case a decrease in PC1 is associated with large increases in "Frozen", "Confectionary", “Groceries for health and pets”, "Prepared Meals" and “Soft drinks” spending. These features best represent PC1 and are in line with our initial findings where the feature pairs are moderately correlated. This dimension can be referred to as: **“NON-STANDARD GROCERIES”**.

**The second principal component (PC2):** Here an increase in PC2 is associated with large increases in “Cashpoint”, “Lottery” and “Tobacco” spending. This dimension represents **“LOTTO-TOBACCO”**.

**The third principal component (PC3)**: An increase in PC3 is associated with a large increase in "Drinks" and a relatively smaller decrease in "Cashpoint" spending. These features best represent PC3, and indicate an outlet where drinks are dispensed. This dimension can be referred to as **"LIQUOR".**

**The fourth principal component (PC4)**: An increase in PC3 is associated with a large increase in "Cashpoint" and “Drinks” and a large decrease in "Tobacco" and “Deli” spending. This might indicate alternate stocking of the drinks and tobacco categories. The dimension represents **"OFF-LICENSE GOODS"** vastly.

**The fifth principal component (PC5):** An increase in PC5 is associated with large increase in "Deli", "Lottery", “Newspapers” and "World foods" spending with relatively smaller decrease in “Tobacco” and “Groceries”. This dimension can be named **"NON-ESSENTIAL PURCHASES".**

**The sixth principal component (PC6)**: An increase in PC6 is associated with a large increase in "Lottery" and “Newspapers” and a large decrease in "Cashpoint" and “Deli” spending. This dimension can be referred to as **"STREET-VENDOR GOODS".**

**SECTION-C: CUSTOMER BASE SUMMARY**

The four datasets provided record the transactional behavior of 3000 customers across a national convenience store chain over a period of six months. Prior to performing market segmentation, the customer base was studied as to gauge overall behavior across the features selected above. Understanding the data distribution across the features determined the segmentation methodology to be undertaken.

The statistical summary of the entire dataset shows that the customer sample (n=3000) typically makes 487 store visits and spends £769 on an average during six months. Of the 20 product categories, it was observed that a customer on an average spends significantly more on certain categories. For instance, a typical customer has a higher mean spend for categories like dairy (£71), fruit and veg (£69), groceries (£60) and confectionary (£57) as compared to non-standard products like delicatessen (£14), newspaper (£17), seasonal gifting (£6), practical items (£2), world foods (£8.5), lottery (£14) and soft drinks (£23). Also, the average amount spent on drinks (£62) and tobacco (£92.5) is noticeably high with an unusually high standard deviation of £121 and £202. A customer usually spends moderately on categories like bakery (£38), meat (£55), prepared meals (£35) and frozen food (£35). This erratic spending pattern across the categories is the reason as to why a scatterplot visualization of the whole dataset indicates a high level of skewness in the data which is resolved through normalization as discussed in the previous section.

As per the engineered features from PCA, it can be concluded that customers often opted to buy non standard groceries like frozen foods, confectionary and groceries related to health and pets. The second dimension is associated with a simultaneous high spending in lottery and tobacco. Similarly, the third component indicates a high spend on drinks in cases where the customer did not opt for cashpoint. The fourth principal component shows that the customer either spends on drinks and cashpoint or delicatessen and tobacco. The fifth dimension is an indicator of high spends on non-essential products like lottery and deli when the spend on tobacco and groceries is low. The final principal component indicates spend on newspaper and lottery when there is low spend on deli and cashpoint and vice versa.

**SECTTION-D: SEGMENTATION METHODOLOGY**

The task is to perform a market segmentation based on the behavior of the customers. Segmentation is classified under unsupervised learning since there is no target class to learn from. For customer segmentation, K-means was chosen as the clustering technique employed on the reduced data obtained after PCA on logged data. This technique was opted for owing to its robust, light weight and flexible structure in addition to the normalized distribution of the logged data. The latter would greatly boost the efficiency of k-means in labelling the datapoints to globular clusters.

The only drawback of the K-means technique is that the ideal value of K is not known beforehand. K stands for the number of clusters the data is to be divided into. However, as per the message of the CEO, the number of customer segments could range between 5 and 7 both inclusive. To determine the best value of K for segmentation, each data point's silhouette coefficient was calculated. The silhouette coefficient for a data point measures how similar it is to its assigned cluster from -1 (dissimilar) to 1 (similar). This score helped us pick the best value of K, i.e. the number of customer segments for targeted marketing. The then labelled data points from the clusters would be then transformed back into their original dimensions and scaled to understand their significance.

So, the K-means algorithm was applied to the reduced data to create K clusters. The cluster for each data point was predicted and after assigning the cluster centers to their respective clusters, the cluster for each sample data point was predicted and the silhouette score was calculated for different number of segments varying between 5 and 7.

The approach yielded 7 as the number of clusters with the highest silhouette coefficient i.e., the clusters produced by the algorithm exhibited maximum intra-similarity and inter-dissimilarity. The distinct segments were then analyzed for further understanding and targeted marketing.

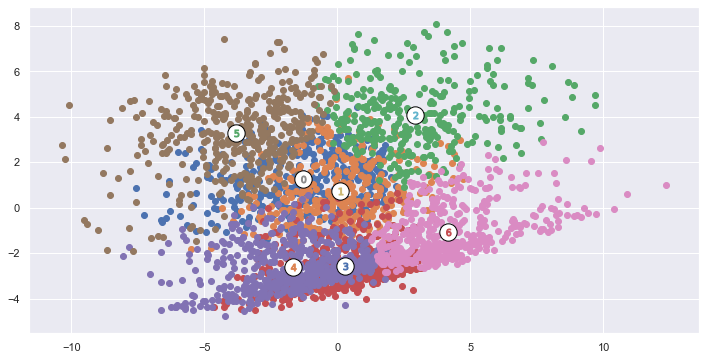


Fig: Cluster representation

**SECTION-E: RESULTS AND CLUSTER ANALYSIS**

As discussed in the section above, the number of clusters K was picked as 7 since the algorithm yielded the high average silhouette coefficient for that value of K. The clusters along with their segment centres can be visualized in the figure attached above. Each segment portrays a highly typical behavioural pattern the knowledge of which can be channeled to improve the targeted marketing approach. The statistical summary for cluster is present in the python code for perusal. Based on the set of features used for the segmentation, the customers can be classified as either high, moderate or thrifty spenders. So, each segment displays one of these spending patterns with distinctive behavior across the rest of the features. The clusters are described according to their spending pattern below because the total spend is the sole relevant monetary factor that accurately gauges a customer’s contribution to the revenue of the company.

*HIGH SPENDERS:* This spending behavior is exhibited by three clusters (0,1 and 4). With extremely high total spends, this set of loyal and repeating customers contribute most significantly to the company’s revenue. The distinguishing factors across the segments are outlined below:

*Segment 0:* With moderate spending trends across essential categories like groceries, fruit and veg, dairy, this cluster surprisingly spends extensively on categories like deli, discount bakery, seasonal gifting and world foods. With a high average spend, they majorly have a low engagement with cashpoint and can be categorized as spendthrift connoisseurs.

*Segment 1:* With the highest average spend across all clusters, they are the most frequent class of shoppers. They have a high spending pattern across all the product categories. Hence, this set of customers can be classified as prodigal shoppers.

*Segment 4:* These are mainly loyal and habitual customers who spend on almost every product category. With a low average spend, these non-smokers spend mostly on essential products excepting for deli and drinks. The most noticeable observation is the high spending trends across the seasonal gifting category and its extremely low lottery investments. They can be called loyal and responsible shoppers.

*MODERATE SPENDERS:* This spending behavior is endorsed by two clusters (2 and 5). They have a moderate spending pattern with a low average spend per item. The overall monetary aspect is the only similarity between the two cluster. The distinguishing features are highlighted below:

*Segment 2:* Despite having the lowest average spend per item, as contrasted to the other cluster in this category, this set of customers does spend extensively on few product categories like dairy, fruit-veg, groceries, delicatessen and frozen which can be broadly categorized under the umbrella of essential food items. They exhibit a low engagement with cashpoint and do not spend much on lottery and other non-standard products. They are largely an intersection of non-smokers and non-drinkers. They can be described as essentials buyers.

*Segment 5:* With a low average spend per item, this cluster exhibits moderate spending pattern across all essential product categories and low spending pattern across the remaining categories. With moderate spending on drinks, they are also non-smokers. The most noticeable feature is this cluster’s engagement with cashpoint which is unusually high and moderate lottery investments. With a cautious spending pattern, this set of customers can be broadly categorized as economical customers.

*THRIFTY SPENDERS:* The remaining clusters (3 and 6) are segments that spend sparingly across product categories and have a low total spend. These two clusters are similar in terms of low engagement of customers across most of the product categories. However, they are distinct with respect to few factors as outlined below:

*Segment 3:* This set of customers are shown to be rare visitors at the store but they have a high average spend which can be attributed mostly to their heavy spends on tobacco and moderate spends on drinks and lottery. Their behavior is akin to consumers who frequent only off-license liquor stores. As can be observed from the statistical analysis, with a moderate engagement with cashpoint, they have a low spending when it comes to each of the other essential and non-standard product categories. They can be called thrifty smokers.

*Segment 6:* The final customer segment is easily the most frugal set of consumers. They are extremely rare store visitors with a moderate average spend which explains their low total spend i.e., their contribution to the profits of the company is least. They do not spend on any product category in the store. They are majorly non-smokers who spend only on drinks, in moderation. No other assumption can be made about this set of customers because of their low engagement with the product categories. This segment can be referred to as cautious and disengaged customers.

**SECTION-F: SUMMARY AND RECOMMENDATIONS**

For market segmentation of the customer sample, behavioural segmentation was opted for and K-means was used as the clustering technique to divide the customer base into 7 segments. This will help targeting the customer base based on their purchasing behavior. The targeted advertising and sculpted marketing based on the customer’s inclination towards certain categories more than the rest can help boost the spends of the customers, in turn the total revenue of the company. Based on the cluster analysis, it is recommended to focus most strongly on the two moderately spending segments (clusters 2 and 5) because of sure shot engagement boosting. With tailored deals and offers, the average spends and frequency of store visits is most likely to improve for these segments which are currently moderately engaged. Apart from the focus on the two customer segments mentioned, for segments tending to invest in lottery can be targeted with lucrative offers. The third recommendation would be to provide more loyalty benefits to high spenders to ensure their retention and unwavering engagement. Finally, the company can aim to improve the reach of their services among thrifty segments and provide them with discounts. This approach will help in converting occasional shoppers to habitual customers. In conclusion, with efficient targeted marketing of the two recommended segments, the store chain is sure to maximize its revenue collection over time.