Analytics and specialization- BUSI4370

COURSEWORK 1- CUSTOMER ANALYTICS STUDY 2022-23

STUDENT ID: 20494451

An Executive Summary:

We have analyzed a transactional data set which was recorded by a loyalty card transactional log and performed market segmentation on it. A national convenience store chain provided the data set, which contained 4 files and data to describe the behavior of 3000 consumers over the course of six months. Based on the point-of-sale data, the analysis will create 5-7 consumer groups, and each group will get statistical and a pen profile. With the aid of these profiles, the retailer will be able to comprehend their client base better and plan out the next marketing initiatives.

To complete this task, a technical approach involving a Python script in Jupyter Notebook is implemented. Several libraries were utilized, including Pandas, Sklearn and NumPy. To further aid in the visualization of a dataset, seaborn and matplotlib were employed. The report and python script are accompanied by a csv file that lists the customers' assignments to the clusters generated in the analysis.

The first step involves data preprocessing and feature engineering, where we have cleaned the "**basket_spend**" column in the "**baskets**" data frame since the column is non-numeric and it contains commas and currency symbols (in this dataset, "£"). We also convert the resulting strings to floating-point numbers, making the column numeric so that we can perform calculations on the column, then we transform to create new features. These features are then used as input for PCA, which reduces the dimensionality of the data by identifying the most important features that explain the variability in the dataset. As clustering algorithms use distance metrics between attribute values, normalization is crucial.

In total, 3000 customers were segmented based on purchases over the past 6 months. Consumers from segment 5 make purchases from us frequently and spend a lot of money each time they come in, but segment 4 customers make purchases from us less frequently and spend a lot less money overall. segment 1 customers have high average spending and high average quantities bought, whereas segment 2 customers have low average spends and low average quantities bought. Based on these segments, we observe that segments 4 and 2 need to be converted using marketing campaigns to better meet their needs, ultimately leading to increased sales and customer loyalty. On the other hand, segments 5 and 1 are the most loyal ones.

Feature Description section:

Feature selection and engineering are critical steps in preparing data for machine learning models. We have taken three features:

- **purchase_time** is a valuable component in market segmentation since it sheds light on the timing of when customers are most likely to purchase a good or service and we can modify our marketing tactics and promotions to reach the appropriate audience at the right time by knowing when customers are most likely to make a purchase.
- By analyzing **basket_spend** and **basket_quantity** i.e., the amount of money customers spends on every visit and the no. of items purchased, businesses can identify high-spending customers who are willing to pay a premium price for products or services, and they can target their marketing efforts towards this

group. We may also recognise clients who are price-sensitive and target them with special offers and discounts to persuade them to spend more.

	frequency	total_spend	average_spend	average_quantity
0	56	675.72	12.066429	9.482143
1	33	585.73	17.749394	19.848485
2	59	222.18	3.765763	4.983051
3	37	547.87	14.807297	13.486486
4	48	293.34	6.111250	5.854167

Figure 1: The image shows the selected features.

From the above taken features, we have generated additional features such as **frequency**, **total_spend**, **average_spend** and **average_quantity** and clubbed them into a new data frame called "**newbaskets**". We observed that these variables are the most informative in differentiating between customer segments.

- '**Frequency**' represents the number of times a customer has made a purchase within the time frame of 6 months. This feature can be useful for predicting consumer loyalty and locating valuable customers who buy frequently.
- **'total_spend'**: This feature shows the overall sum that each client has spent on all their purchases. It is determined by summing each set of "customer number" items with the feature "basket_spend" from the baskets data frame. As consumers that spend more money are probably more important to the company, this attribute may be helpful in determining the overall worth of each client to us.
- **'average_spend'**: This feature displays the average amount each consumer spends on a single purchase. The mean of the "basket spend" entries in each group of "customer number" from the baskets data frame is used to compute it. While consumers who frequently spend more each transaction may have different purchasing patterns than those who consistently spend less, this feature might be helpful in determining the normal purchasing behavior of each client.
- **'average_quantity'**: This feature displays the typical number of products each consumer buys for each transaction. Although customers who frequently buy more things in each transaction may have different wants or preferences than those who consistently buy fewer items, this feature might be helpful in identifying the normal purchasing behavior of each customer.

To initialize the feature engineering, Principal Component Analysis (PCA) is the adopted methodology for market segmentation utilizing unsupervised machine learning techniques. Creating segments of clients with similar behavior or traits can enable marketers to create more specialized and successful marketing campaigns.

Since we are working with a high number of features, which is a problem because the data will be high dimensional which means there will be a curse of dimensionality. There are 3 techniques to reduce dimensionality such as LDA, NMF and PCA. When we normalize our data by using log(), we get negative values, but we cannot work with it. This is the reason we do not use LDA and NMF instead here, PCA is an effective method that we employed to decrease the curse of dimensionality.

Each principal component's explained variance reflects the proportion of total variation in the original dataset captured by that component. In this situation, the first main component accounts for 54.86% of total variance, the second component accounts for 41.65%, the third component accounts for 3.36%, and the fourth component accounts for 2.81e-32%, which is insignificant.

We have described these four dimensions; the respective image is attached in the appendix.

Dimension 1: The first principal component explains the highest amount of variance in the data. The features with the highest loadings on this dimension are frequency and average quantity, with negative values. This suggests that customers who make more frequent purchases tend to buy fewer items per transaction. The feature with a positive loading is average spend, which suggests that customers who make larger purchases tend to do so less frequently.

Dimension 2: The second principal component explains the second highest amount of variance in the data. The features with the highest loadings on this dimension are total spend an average spend, with negative values. This suggests that customers who make larger purchases tend to spend less on average per transaction.

Dimension 3: The third principal component explains the third highest amount of variance in the data. The feature with the highest loading on this dimension is average spend, with a positive value, suggesting that customers who spend more on average per transaction also tend to spend more overall. The feature with the highest negative loading is average quantity, suggesting that customers who buy fewer items per transaction also tend to spend more overall.

Dimension 4: The fourth principal component explains the lowest amount of variance in the data. The feature with the highest loading on this dimension is average spend, with a positive value, suggesting that customers who spend more on average per transaction also tend to buy more items per transaction. The loading of average quantity on this dimension is close to zero, indicating that it does not contribute much to this dimension.

A Customer Base Summary section:

	frequency	total_spend	average_spend	average_quantity
count	3000.000000	3000.000000	3000.000000	3000.000000
mean	65.182333	769.412937	14.801139	11.273373
std	47.464717	552.769022	11.161440	8.538046
min	1.000000	7.280000	1.456000	1.200000
25%	32.000000	406.120000	8.036819	6.114316
50%	53.000000	627.170000	11.770923	8.732520
75%	86.000000	957.675000	17.436190	13.388537
max	374.000000	6588.650000	152.621667	90.750000

Figure 2: The image shows the statistical summary of the selected features.

Based on the given data, we can summarize the company's market as follows:

Frequency: The average number of purchases made by customers is 65, with a minimum of 1 and a maximum of 374. 25% of customers make 32 or fewer purchases, 75% of customers make 86 or fewer purchases.

total_spend: The average amount spent by customers is £769.41, with a minimum of £7.28 and a maximum of £6,588.65. The 25th percentile is $406.12, meaning that 25% of customers spend £406.12 or less, while the 75th percentile is £957.68, meaning that 75% of customers spend £957.68 or less.

average_spend: The average amount spent per transaction by customers is £14.80, with a minimum of £1.46 and a maximum of £152.62. The 25th percentile is £8.04, meaning that 25% of customers spend £8.04 or less per transaction, while the 75th percentile is £17.44, meaning that 75% of customers spend £17.44 or less per transaction.

average_quantity: The average quantity purchased per transaction by customers is 11.27, with a minimum of 1.20 and a maximum of 90.75. The 25th percentile is 6.11, meaning that 25% of customers purchase 6.11 or fewer items per transaction, while the 75th percentile is 13.39, meaning that 75% of customers purchase 13.39 or fewer items per transaction.

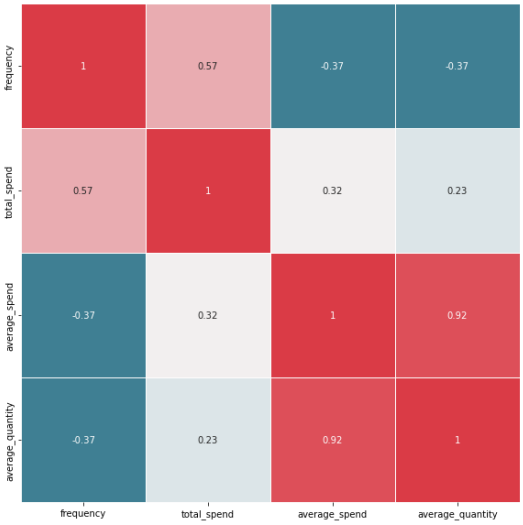

Figure 3: This image depicts a correlation heatmap between the selected features.

We have shown the relationship of the selected features by making a scatter plot as given in the python script file and a correlation heatmap which tells us that **average_spend** and **average_quantity** is highly positively correlated with a correlation of 0.92 which means that amount of money which per customer spends is directly dependent on their number of items bought by each customer. **Total_spend** and **frequency** are moderately correlated to each other which tells us that the total amount of money spent by customer depends upon number of times visited. On the flip side of the coin, **average_spend** and **frequency** are negatively correlated which tells us that per customer spend and number of times visited are not independent of each other.

Segmentation Methodology section:

K-Means clustering is used as a clustering approach to group customers into segments based on their similarity in terms of the selected features. The silhouette score, which assesses the similarity of items inside a cluster to objects in other clusters, is used to calculate the number of clusters for K-Means clustering based on the PCA results. Based on the given silhouette scores, the best value of k would be 6. A score closer to 1 indicates better-defined clusters, whereas a score closer to -1 indicates that data points may have been assigned to the wrong cluster.

In this case, the silhouette score is highest for k=6, which is 0.3289, indicating that this configuration of clusters has better-defined clusters than for k=5 and k=7. A higher silhouette score also suggests that there is good separation between clusters and that each cluster is distinct from the others.

The resulting clusters are visualized using a scatter plot given in the python script file, and the centers of each cluster are calculated and interpreted in terms of the original features. The final step involves assigning each customer to their respective cluster, and the resulting segment assignments are saved to a CSV file for further analysis.

We have chosen 6 segments. **Segment 1** has a high average quantity but the lowest frequency, indicating that customers in this segment buy a lot of items but do not step in much for a purchase. **Segment 2** has a high frequency. However, they have a moderate total spend and they have the lowest average spending and the lowest purchase an average. **Segment 3** has a low frequency. They have less total spending, and they have a low average spending of and low purchase of an average quantity. **Segment 4** has the lowest frequency. However, they have a very low total spend and they have a relatively high average spending and high average quantity. **Segment 5** with the highest frequency. They have a relatively high total spend and purchase very little. **Segment 6** has a moderate frequency. They have a moderate total spend, decent average spends.

Results section:

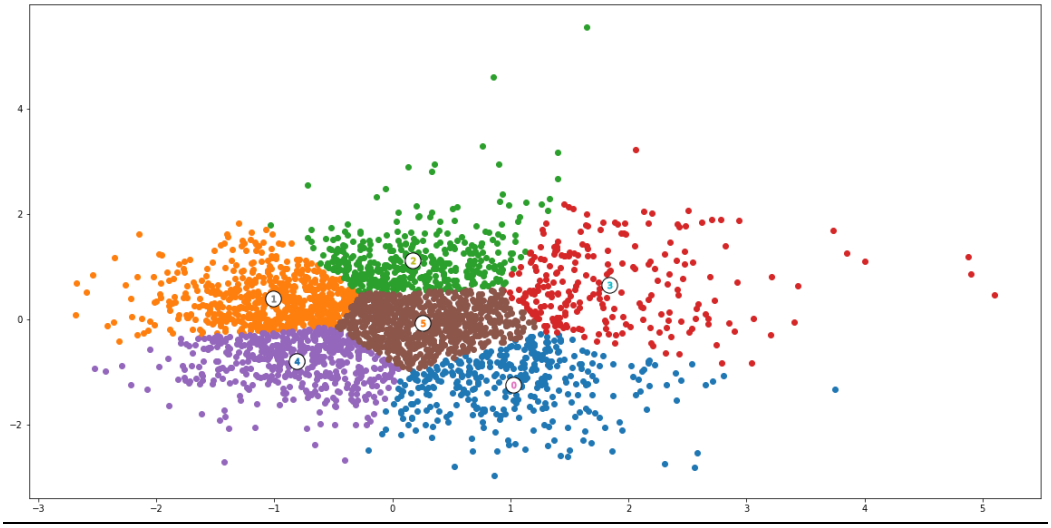

Figure 4: This image shown depicts a scatter plot of all clusters.

Segment 1/Cluster 0, named "**Premium spenders**" has a low frequency of 12, with 358 customers in this segment making purchases just a few times. However, they have a high total spend of £1110.66 and they have a relatively high average spending of £35.84 per purchase and high purchase an average of 27 items per transaction. These are the spenders who spend on a premium level whenever they step in, but visit only a few number of times, on a needy basis.

Segment 2/Cluster 1, named "**Medium tickets**" has a high frequency of 107, with 668 customers in this segment making purchases frequently. However, they have a moderate total spend of £700.12 and they have the lowest average spending of £6.92 per purchase and the lowest purchase an average of 5 items per transaction. This segment likely consists of customers who are price-sensitive and tend to shop for lower-priced items. Although they make frequent purchases, their lower average spending per purchase and fewer items per transaction suggest that they are not making large purchases or buying in bulk. The moderate total spend may be due to the high frequency of purchases, but the relatively low spending per purchase and number of items purchased suggest that these customers are not high-value customers.

Segment 3/Cluster 2, named "**Lapse customers**" has a low frequency of 38, with 425 customers in this segment making purchases rarely. They are inactive buyers with just a total spending of £284.92 and they have a low average spending of £7.75 per purchase and low purchase of an average of 6 items per transaction. It describes a group of consumers that were formerly engaged but have now turned inactive due to a variety of factors. These clients may have shifted to a competitor or lost interest in the company's products and services.

Segment 4/ Cluster 3, named "**Risky customers**" has the lowest frequency of 12, with 245 customers in this segment making relatively less frequent purchases. However, they have a very low total spend of £231.15 and they have a relatively high average spending of £20.26 per purchase and high purchase an average of 15 items per transaction. They are rare purchasers who make smaller purchases but tend to buy a decent number of items when they do make a purchase.

Segment 5/Cluster 4, named "**Big tickets**" has 510 loyal customers with the highest frequency of 113, where customers in this segment make purchases the most often. They have a relatively high total spend of £1536.78,

averaging around £14.32 per purchase and purchasing an average of 9 items per transaction. The customers in this segment are highly loyal and very economical for us, they buy less items but spend the most.

Segment 6/Cluster 5, named " **Moderate spenders**" has a moderate frequency, with 794 customers in this segment making purchases 43 times. They have a moderate total spend of £606.35, decent average spend of £14.33 per purchase and purchase an average of 11 items per transaction. These spenders are the ones who are not massive spenders, but they buy multiple items in one single purchase.

Overall, we can see that the different segments have different average spending levels and purchase quantities, with some segments having customers who are willing to spend more than others. Additionally, there is significant variation within each segment, suggesting that there are likely subgroups of customers with different spending patterns within each segment.

Summary Section:

The company should focus on 2 segments who are bringing high value such as "**Big Tickets**" and "**Premium Spenders**". The "Premium spenders" in Segment 1/Cluster 0 are significant because, despite their low frequency of purchase, they have a high overall spend and a high average expenditure per transaction. This suggests that they are prepared to spend a large amount of money on high-end items, and they might be a great target for premium-focused marketing efforts. But, because they only visit a few times, it is critical to devise tactics to get them to return more regularly. For premium spenders, a loyalty rewards program can be suggested offering them this program with exclusive rewards and discounts can encourage them to increase their loyalty with the company and spend more. In addition, providing personalized product recommendations based on their past purchases and preferences can help improve their shopping experience and increase their spending. Segment 5/Cluster 4, the "Big tickets," are important because they have the highest frequency of purchasing, indicating that they are loyal customers who are regularly engaging with the company. While their average spending per purchase and the number of items they purchase may not be as high as some other segments, their consistent patronage is valuable for generating a steady stream of revenue. Retaining these customers is essential to maintain a stable customer base and constant revenue flow. The company should consider retaining them by offering loyalty programs and discounts in this segment to encourage repeat purchases and build brand loyalty. A marketing strategy used for big tickets would be Upselling since customers in this demographic tend to purchase fewer things but spend more, therefore supplying complimentary or higher-priced items that complement their purchases might raise their spending per transaction.

Appendix:

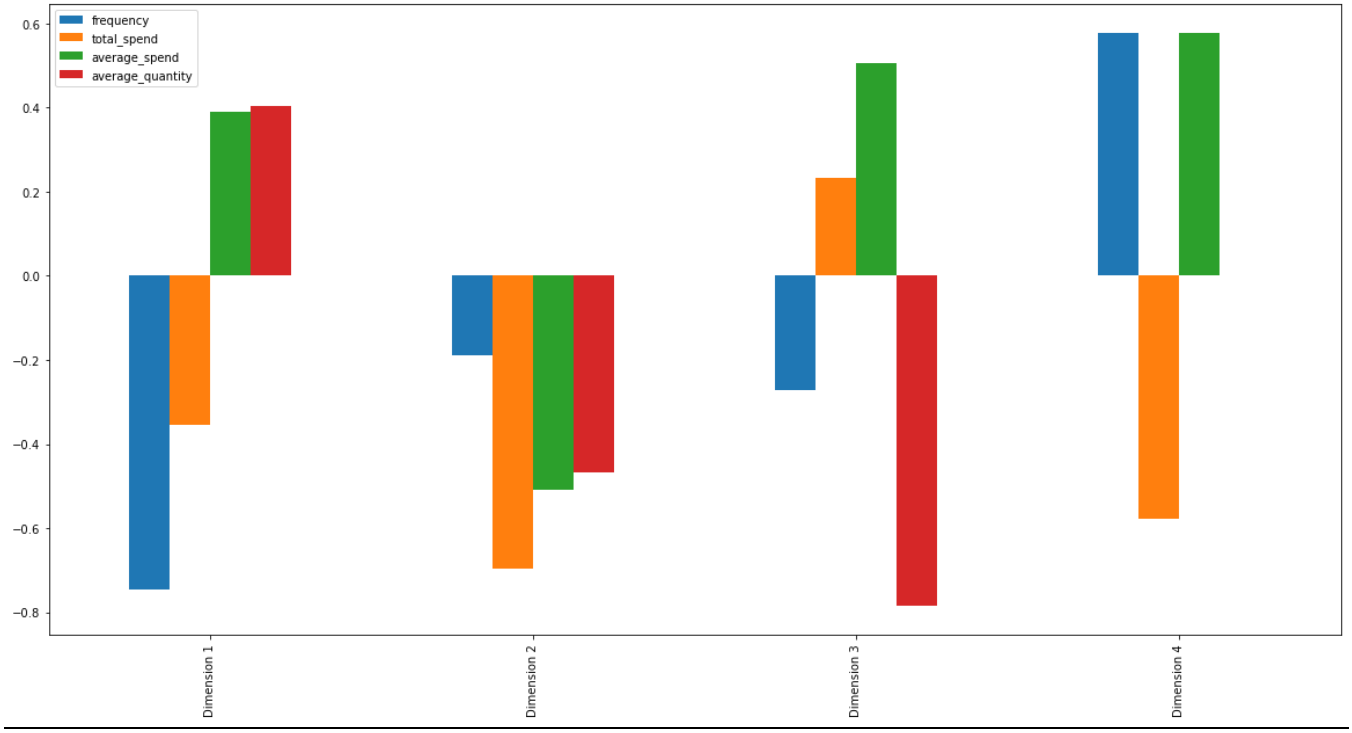

Figure 5: This image depicts 4 dimensions with respect to the selected features.