CUSTOMER ANALYTICS STUDY ON A NATIONAL CONVENIENCE STORE USING SEGMENTATION

BY: S M SAZID ISLAM

Executive Summary:

This report presents a detail analysis of market segmentation being performed on a transactional dataset provided by a national convenience store chain consisting of 4 files describing the consumer behavior of 3000 customers over a span of 6 months period. The main objective of the report is to analyze the company's point of sale data and create profiles for 5-7 customer segments.

To carry out the segmentation, **Jupyter Notebook (anaconda3)** software has been used to cleanse, process, feature engineer and cluster the data into groups. Different libraries like pandas, seaborn, NumPy and matplotlib are used to analyze and visualize the relationship among the different features of the dataset. Pandas is a data manipulation and analysis tool which has a powerful data structure. NumPy is used to provide a faster array object, it provides a high performance multi-dimensional array object as well as tools for working with these arrays. Seaborn is used to visualize the data which gives high-level interfaces and shows informative statistical graphs for better understanding of the data. Matplotlib is another visualization library which is used to make the boxplots and to present the relationship among different features of the data through visualization. Using these libraries as well as other functions the segmentation has been performed and finally the segmented file is stored in a csv file.

All the files are explored initially and based on the objective of the task, **baskets_sample.csv** file has been chosen to identify the target audiences. The files provided includes customer number as referential key which acts as a unique customer identifier for 3000 individual customers. We identified that there are 195547 datapoints and number of features is 5. The datapoints are then grouped using groupby function to 3000 customers and clustered into 6 clusters using the K-means clustering algorithm, using K-means the hidden segments in the data can be identified. The clusters are then divided into 6 customer segments and pen profiled. Based on the company's requirement, cluster-1.0 and cluster-3.0 are suggested as the best profiles to target and develop marketing plans.

Feature Analysis:

The features of all the files are explored below:

1. baskets_sample.csv- number of datapoints: 195547, features: 5
   Features: customer_number, purchase_time, basket_quantity, basket_spend, basket_categories
2. customers_sample.csv- number of datapoints: 3000, features: 6
   Features: customer_number, baskets, total_quantity, average_quantity, total_spend, average_spend
3. category_spends_sample.csv- number of datapoints: 3000, features: 21
   Features: customer_number and 20 other types of food category
4. lineitems_sample.csv- number of datapoints: 1461315, features: 6
   Features: customer_number, purchase_time, product_id, category, quantity, spend

The file named **baskets_sample.csv** has been used for segmentation. The features of this file are of 2 data types, purchase_time and basket_spend are object and basket_quantity and basket_categories are integers. The data is then checked for missing value and we found no missing value. The basket_spend feature type is changed to float and the pound (£) sign is removed as a part of data cleansing and analysis purpose.

To fulfill the requirements mentioned in the task, we generated a new feature and named it **average_spend**. We grouped the customers based on customer number and **basket_categories** feature is dropped as it is not relevant for the clustering. Purchase time as count is considered as the frequency which is the "**number_of_visits**" of customers, **basket_spend** is the "**total_spend**". **Average_spend** is the average spend of each customer and **basket_quantity** is the average quantity purchased by each customer. The features are then renamed for better understanding of the report. The renamed data dictionary is shown below:

customer_number	Frequency	Total_spend	Average_spend	Average_quantity
14	56	675.72	12.066429	9.482143
45	33	585.73	17.749394	19.848485
52	59	222.18	3.765763	4.983051
61	37	547.87	14.807297	13.486486
63	48	293.34	6.111250	5.854167

As per the message from the company's chief data officer, the main focus of the segmentation is to identify the high spending customers, customers who pays more visits as well as customers who spend more in average. All these requirements are covered with the features present in the dataset above. Hence, baskets_sample.csv file is used for the segmentation.

Customer Base Summary:

This section includes a summary of the exploratory analysis of the dataset we are using to perform the segmentation. After the data is cleansed and pre-processing is done, the relationship among the features are examined to understand how one feature is impacting another feature. To understand the total spending of maximum customers, a graph of total spend against number of customers is plotted.

The figure 1 on the right shows the relationship between number of people vs total spend of each customer, it indicates that most of the customers spend around 400 to 600 pounds in the convenience store.

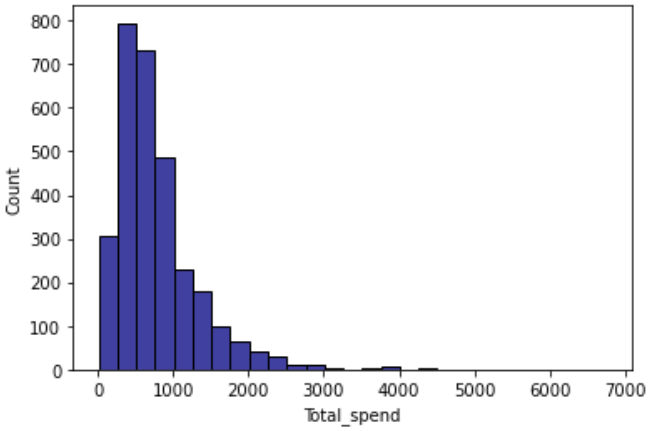

Figure 1: Total Spend Vs Number of People

	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

This is the data after cleansing which shows the customer behavior of **3000 customers**, it indicates that people visited the convenience store an average of **65 times** in the span of 6 months period, while spending a total of around **£770** in average and per customer spent **£15** in average and purchased **11 items** in average form the store. The maximum total amount spend by a customer is **£6588** with maximum average spend of **£153** and maximum average quantity of **91** items per customer in 6 months period.

The scatterplot graphs plotted in the coding file to examine the relationship among the features and for better visualization of the correlation among the features of the file, heatmap of Pearson correlation matrix is plotted, shown in figure.2.

From **figure.2**, this can be stated that **Average_quantity** and **Average_spend** have **strong positive correlation** score of **0.92** and **Frequency** and **Total_spend** have **moderate positive correlation** score of **0.56** which we analyzed. However, features like **Frequency** and **Average_quantity** have **weak negative correlation** score of **-0.37**. This indicates that customers who buys in bulk amount are tend to spend more as well as, customers who visits the store more frequently are somewhat the customers who makes higher purchase decisions.

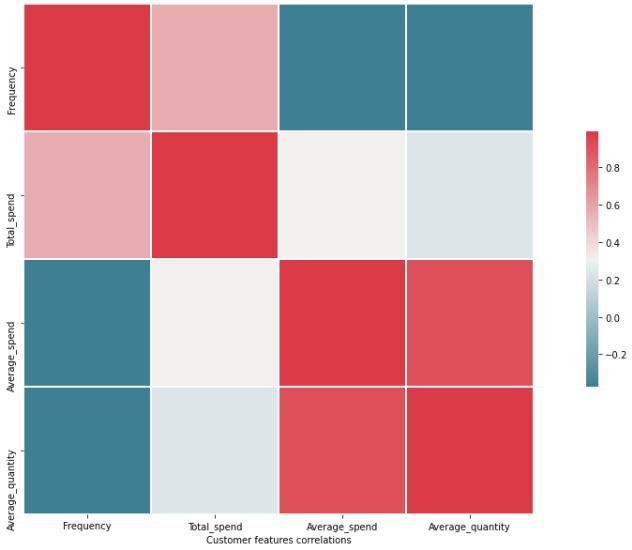

Figure 2: Correlation Heatmap

However, correlation do not have strong impact on clustering but to minimize the effect of correlation on the data and increase the accuracy dimensionality reduction technique is used. There are 3 types of techniques: **Principal Component Analysis (PCA)**, Non-negative matrix factorization (NMF) and Topic modelling (LDA). The importance of dimensionality reduction reduces the time and storage required which minimizes the computational cost to perform learning and removes irrelevant features from the data, these features decreases the accuracy of the segmentation. Before performing dimensionality reduction, it is very important to standardize the data otherwise there might be bias in the outcome as the data will not be in the same scale.

Feature Engineering is initiated where PCA is used to reduce the impact of correlation because PCA helps to identify patterns in data based on the correlation between features. Moreover, PCA is an unsupervised linear transformation technique which has the attribute of finding the directions of maximum variance in high dimensional data and projects onto a new subspace with equal or fewer dimensions than the original dataset. The results of the PCA shows that **96.63%** of the variance in the data is explained by the first and second principal components **(PC1-57.98% and PC2-41.65%)** for which we will consider the PC1 and PC2 for further analysis.

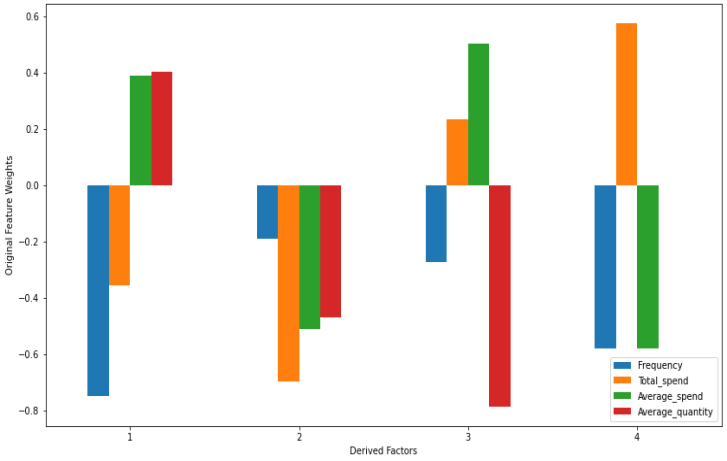

Figure 3: PCA Visualization

Being a convenience store, it is very important to target the right customers and retain a competitive position in the market. The features used in this analysis are very vital to pen profile customers because this will allow the company to increase sales and will also support in minimizing the marketing cost.

Segmentation Methodology:

The next step in the process is the clustering of the data. The main purpose of clustering is to cluster customers with similar behavior among the total number of 3000 customers. In this context, K-means clustering method is used which is an unsupervised learning algorithm and used for clustering tasks which is suitable for complex datasets. Clustering algorithms use the distance in order to separate observations into different groups.

Therefore, the silhouette score of the dataset is visualized to determine the optimum number of clusters. Silhouette score is a measuring score used to calculate the goodness of a clustering technique which ranges from -1 to 1. Whereas, 1 means clusters are well apart from each other. Based on the k values and the requirement by the company, we checked the score of clusters 5 to 7, cluster 5 has a score of 0.32, cluster 6 has a score of 0.33 and cluster 7 has a score of 0.32. Comparing all the cluster scores, we selected 6 clusters as it has the highest score among 5 to 7 clusters, this indicates that the clusters are slightly overlapping but are well separated.

Results:

In this part of the report, the results of the clusters produced are presented and discussed. The clusters are developed using K-means clustering method as shown below:

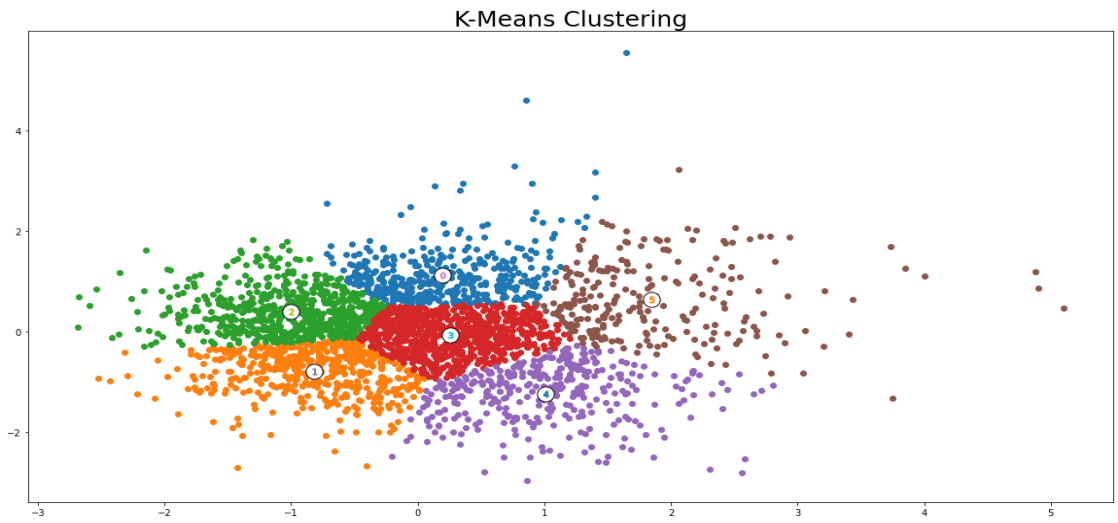

Figure 4: Clustering Visualization

Cluster 0/ Segment 1.0: This cluster is indicated by **blue color** located at the top portion of the figure above consisting a group of **427 customers**. Analyzing the features of the cluster, we named it as the **need-based** customer segment because customers in this segment are the least total spending customers of about **£614** on average among the other clusters. They visited the store on an average of **54 times** in these 6 months, made an average spend of **£15** and bought **12 items** on an average. They visit the store only when a need arises, mostly the students and middle-aged customers falls under this segment.

Cluster 1/Segment 2.0: This cluster is indicated by **orange color** located at the bottom left corner of the clustering consisting a group of **511 customers**. This cluster is named as the **high-spending buyers** as they have the highest number of visits in the time span of about 65 times. As well as, they are the highest spending consumers with a total spend of **£805** on an average while also having an average spending of **£16** per customer and buying **12 items** in average. They are the important customers of the store; marketing campaigns should be designed to address this group of customers.

Cluster 2/Segment 3.0: This cluster is presented by **green color** located at the left side of the figure with a group of **672 buyers**. This cluster is named as the **occasional buyers** because they are neither the maximum or minimum in average spending and average item purchasing, they visit the store occasionally and make purchases. They have an average frequency of **58 times**, total spend of **£628**, average spend of **£14** and buys **12 items** in average. They are the customers who visit the store based on the product preference.

Cluster 3/Segment 4.0: This is the segment marked with **red color** and located at the middle of the clustering with **788 customers** in it. This segment is named as the **active buyers**, they are the buyers who visits the store more often with a frequency of **57**, making an average spend of **£15** and buying an average of **13 items**. They also have a good total spend of **£686**, this cluster of customers in also a very essential target group for the store because it is the largest group of customers.

Cluster 4/Segment 5.0: This cluster can be seen as the cluster with **purple color** in the graph, it consists of **356 individuals**. This is named as the **average spending buyers**, in the span of 6 months, they impacted

sales with a total spending of **£676** in average while they have a low average spend of **£13** and a low average quantity of **11 items** per customer. They are good buyers but their average quantity purchased is low.

Cluster 5/Segment 6.0: Among the 6 clusters formed, this is the last and the smallest cluster with a group of **246 customers** marked with **brown** color. It is named as the **item-focused buyers**, in this cluster where the customers are more item focused, they buy the highest number of items in average of **14 items** among the customers of other clusters and also, they have a high average spend of **£17** and a total spend of **£761** on average. However, this is a very small group to consider as a target group.

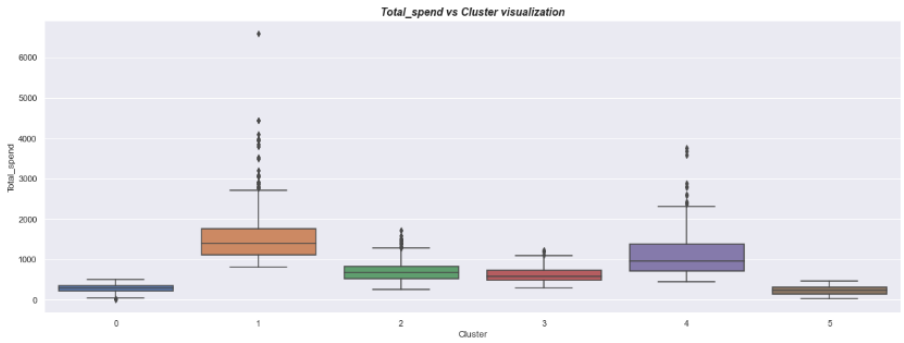

Figure 5: Boxplot of Clusters

In **figure. 5** above, a boxplot of the segmentation performed is plotted which presents that cluster 1 has more scattered dataset covering a customer base with varying behavior and cluster 3 has condensed dataset which means the customer behaviors varies less and is more consistent, these 2 are selected to extract a potential customer base of the convenience store.

Summary:

In a nutshell, the best two profiles to be considered as the idea target groups are the **high-spending** buyers and **active buyers** based on the task provided, focusing on these 2 groups will envelop a total of **1299** customers which is about 43% of the total number of customers of 3000. These two groups are determined as the target group because it include customers with common characteristics. This segmentation process makes it easy to tailor and personalize the marketing, service and sales efforts to the needs of specific groups as well as this assists in boosting customer loyalty and conversions.

The proposed cluster models are given with the lack of details about the changes in customer behaviors, therefore a variety of rules and strategies are necessary to find the hidden patterns and shopping trends of the customers. PCA and K-means helped to find clusters of potential customers in the dataset provided. However, if more time and dataset with more dimensions such as demographic details would have been provided, further study can be performed. A Marketing plan to reach the customers actively can be creation of mobile based application platform to send personalized bundle offers to the customers based on their purchase history and basket quantity.

In conclusion, we tried to cover maximum objectives of the task and portrait a customer segmentation based on the datasets provided which might help the company to identify the ideal target audiences.