Executive Summary:

The main aim of the research was to analyze the transactional data for 3000 customers across different product categories. The research was focused on identifying distinct buying patterns which could be used as a basis of formulating customer segments. The initial statistical analysis suggested that there was extensive variation in the amount of money spent by the customers, how often they purchased the products and the time frame in which products were purchased. This led to the selection of RFM technique for customer segmentation which used recency, frequency and monetary values as features. In order to apply RFM, the data was cleaned, and relevant variables were selected. It was analyzed that RFM features themselves were not sufficient for effective clustering, so average spend along with average quantity was also used. 5 main clusters were identified which were labelled as Golden Goose, Low Spending Active Loyal Customers, Inactive Potential Loyalists, Churn Customers and Infrequent High Spending Customers. In order to analyze these clusters further, the category data was used which reflected the different product preferences across clusters. The analysis suggested that the company could focus on Inactive Potential Loyalists and Infrequent High Spending Customers as they offer high value to the company. In order to keep these customers engaged, the business can incentivize them by developing loyalty schemes and having more personalized offers.

Data Cleaning:

All the 4 files were used in the analysis. The customer and basket files were used as an input for clustering process whereas the category and line item files were used to further do an in-depth analysis of the clusters obtained. As a part of making the data ready for use, all the columns with monetary values were stripped off the "£" and "," signs. All of these columns were converted into numeric data type. In the category file, it was also found that the "bakery" column had no values. The line item file was used to filter the observations with bakery items. The unique customer transactions were then grouped together to calculate the total spend of each customer on bakery items. This new column of values was then integrated into the category file and the missing values were replaced with 0 as it was assumed that the customers with no record of bakery items did not spend on this category. The negative values representing refunds to the customers were removed from the category as well as baskets file.

Methodology:

RFM model was chosen for the purpose of segmentation. It is based on 3 measures; recency, frequency and monetary values. These factors influence the future purchase possibilities of the customers. Recency refers to the time period since last purchase date whereas frequency is the number of transactions made by a customer within a certain period. Monetary on the other hand refers to the cumulative money spent by a particular customer. RFM model is useful because in case of convenience stores, the past purchases of customers can effectively predict their future purchase behavior. By this technique, the company can identify which customer is worthy to be contacted based on their past purchase behavior.

	customer_number	baskets	total_quantity	average_quantity	total_spend	average_spend
count	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000
mean	8095.724333	487.105000	583.722000	1.204499	789.412937	1.682477
std	4686.259488	332.824524	405.006359	0.136323	552.769022	0.733105
min	14.000000	6.000000	6.000000	1.000000	7.280000	0.620000
25%	4044.750000	257.000000	307.750000	1.119625	406.120000	1.260000
50%	8218.500000	417.000000	495.000000	1.175889	627.170000	1.490000
75%	12115.500000	628.250000	744.250000	1.250430	957.675000	1.860000
max	16316.000000	3119.000000	4949.000000	2.503688	6588.650000	10.840000

Fig 1.1

The basic statistical analysis suggests that there is a huge variation among how often the customers visit the store and how much they spend on each visit. The highlighted figures show that the standard deviation of baskets is high. This variable indicates that there is a lot of variation among customers regarding how often they visit the store. Apart from this, money spent among customers

varies extensively which indicates that customers might be inherently different in their purchase patterns which will provide an excellent opportunity to segment them.

RFM will help the company devise a marketing strategy to approach different customers in the most impactful way. Frequency measure will be based on the assumption that customers who buy more products are more likely to make greater number of purchases than customers with fewer purchases. Similarly, recency feature will suggest that most-recent customers are more likely to repeat purchases than less-recent purchasers.

Feature Engineering:

1) **RFM Score** - In order to implement the RFM technique, the customers were split into quartiles. On each of the 3 features, customers were given the scores from 1-4 based on the quartiles they belonged to. Customers were given scores for frequency, recency and monetary. For example, from the most to least frequent, the top 25% of the customers were given a score of 4, and the less frequent quartiles as 3, 2, and 1. Finally, an RFM score was calculated by concatenating R, F, and M values.

2) **Total Score** - A total score was also calculated based on summing the R, F and M values. For example, a customer that fell in the top 25% for all 3 features had an RFM score of 444 and his total score was 16. Similarly, the lowest total score was 3 i.e. an RFM score of 111 across all 3 features.

3) **RFM Level** - The customers were then labelled as 'Diamond, Gold, Silver, Bronze and Brass' on the basis of their Total Score as shown in **Fig 2.1**. These labels were just used to have a basic idea about how customers will be classified on the basis of RFM factors only. However, in the final clustering technique, these labels were not used.

4) **Selection of Additional Variables -** In order for the clusters to reveal more information, 2 more features were selected – average spend and average quantity per basket. These 2 features were expected to reveal information regarding the buying habits of customers i.e. the spread of customer interest in either high or low involvement products.

5) **Checking for Data Skewness** -It was analyzed that the 5 features to be used for clustering were highly skewed. This implied that the data had to be transformed to make for it to appear normally distributed.

6) **Log Transformation –** To account for data skewness, the data was log transformed.

7) **Trying PCA**- In order to account for correlation between the monetary and frequency measures as shown in **Fig 2.2**, PCA was used. However, later this approach was discarded which is discussed in detail in later section.

8) **Data Scaling –** As the units of variables were different, the data was then scaled.

Scores	Label
3-5	Brass
6-7	Bronze
8-9	Silver
10	Gold
11-12	Diamond

Fig 2.1

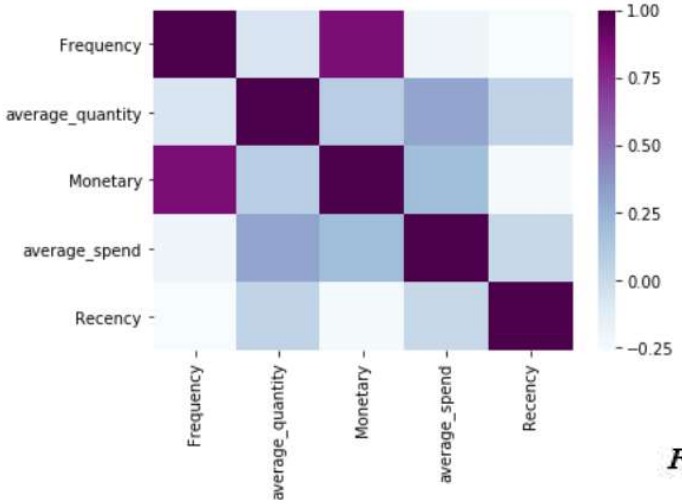
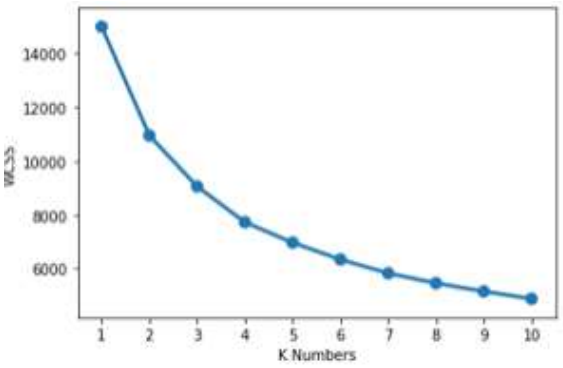


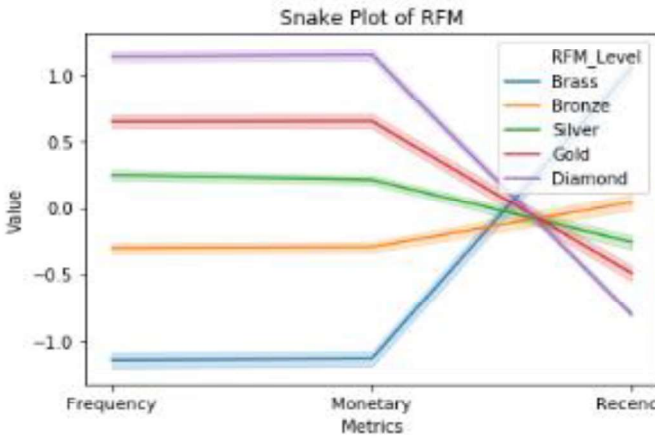
Fig 2.2

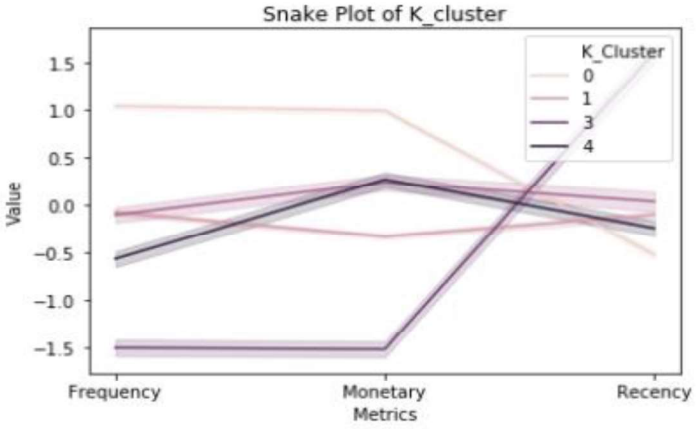
Clustering:

K-means++ algorithm was used instead of simple k-means for clustering as it ensures better initialization of centroids and better-quality clusters. The optimal number of clusters were chosen to be 5 using the silhouette score and elbow method. The features chosen for clustering were frequency, monetary, recency, average spend and average quantity.



In order to compare the clustering output of k-means algorithm and the initial 5 categories obtained through RFM features solely, snake plots were used. The scaled values of the 3 features were plotted on the axis to facilitate comparison. It can be analyzed that diamond customers align well with cluster 0 and depict the most valued customers. Similarly, brass customers were depicted by cluster 3 and show the least valued customers. It can however be seen that bronze, silver and gold customers cannot be precisely allocated to the remaining clusters 1,2 and 4. These differences are primarily due to average spend and quantity. This suggests that it might be a good idea to incorporate these 2 features as they might reveal significant information related to the customer consumption patterns.





Clustering Comparison Using PCA:

To counter the issue of correlation, PCA was used. After implementing PCA on the same 5 variables, 3 dimensions were chosen which explained almost 98% of the variance as shown in **Fig 3.2**. However, when the data was reduced to 3 dimensions and clustering was applied, the clusters generated were not very distinct as shown in **Fig 3.3**. The silhouette score suggested that 6 clusters should be made. The clusters summary in **Fig 3.1** shows that it is very difficult to differentiate between clusters as there is no stark difference between the values. Thus, clustering was done without PCA as explained earlier above which led to the formation of 5 distinct clusters explained in the later section.

Cluster	Frequency	average_quantity	Monetary	average_spend	Recency
0.0	480.801980	1.192586	661.034851	1.438614	8.277228
1.0	511.210884	1.178751	729.318299	1.469456	6.170068
2.0	516.916667	1.198708	819.253056	1.586944	13.888889
3.0	479.561404	1.211495	677.099123	1.458246	9.333333
4.0	459.043011	1.211007	646.176344	1.483441	7.043011
5.0	457.000000	1.181210	629.335065	1.463117	6.428571

Fig 3.1

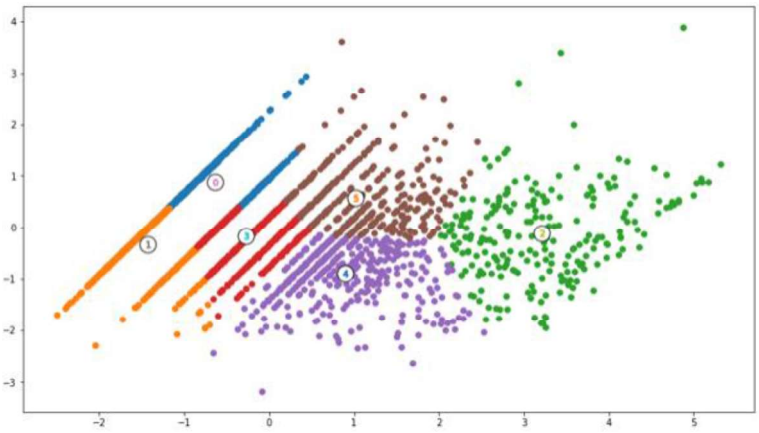


Fig 3.2

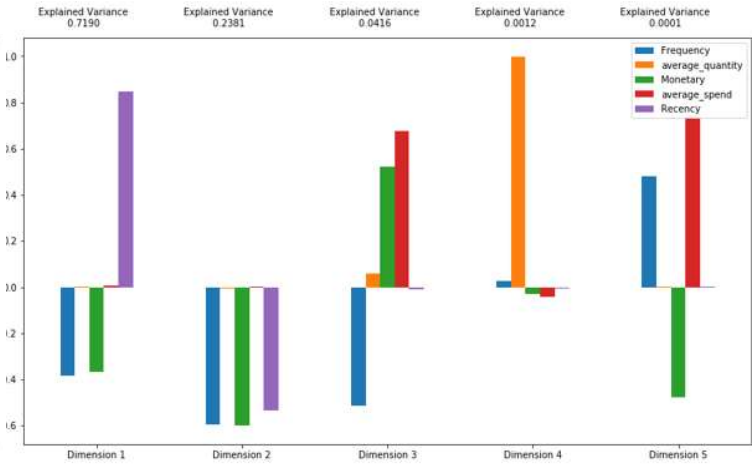


Fig 3.3

Overall Customer Based Summary:

	customer_number	Frequency	average_quantity	Monetary	average_spend	Recency
count	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000
mean	8095.724333	487.105000	1.204499	769.412937	1.682477	8.121333
std	4686.259488	332.824524	0.136323	552.769022	0.733105	20.938531
min	14.000000	6.000000	1.000000	7.280000	0.620000	0.000000
25%	4044.750000	257.000000	1.119625	406.120000	1.260000	0.000000
50%	8218.500000	417.000000	1.175889	627.170000	1.490000	2.000000
75%	12115.500000	628.250000	1.250430	957.675000	1.860000	6.000000
max	16316.000000	3119.000000	2.503686	6588.650000	10.840000	164.000000

The overall customer analysis shows that over the period of 6 months, an average customer visits the store 487 times. On average, the customers spend £1.68 per visit. The average quantity they purchase is around 1.2 units per visit. Furthermore, the average customer visits the store once after every 8 days. The high standard deviations of frequency and recency figures show that there might be some customers who visit the stores more frequently, but the average transaction size is lower. On the contrary, there might be customers who visit less often but purchase greater volume and value of products. This could be an excellent basis of segmentation.

fruit_veg	dairy	confectionary	grocery_food	tobacco	drinks	deli	world_foods	lottery	cashpoint
2971.000000	2971.000000	2971.000000	2971.000000	2971.000000	2971.000000	2971.000000	2971.000000	2971.000000	2971.000000
69.418775	71.134174	57.347883	60.102605	92.134551	61.937967	13.807105	8.569017	14.398320	45.428859
70.220421	57.715027	56.013047	57.781169	200.898314	120.757133	25.583499	14.810689	48.628111	123.952605
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
22.720000	31.435000	21.070000	21.085000	0.000000	0.000000	0.000000	0.890000	0.000000	0.000000
51.080000	56.880000	42.260000	44.170000	2.030000	12.800000	3.500000	3.750000	0.000000	0.000000
93.460000	95.050000	75.100000	81.175000	80.670000	65.085000	15.940000	10.695000	6.000000	30.000000
1262.970000	708.040000	614.370000	1017.070000	2488.94	...00	316.190000	321.490000	946.000000	2137.010000

Fig 4.1

If the data is further analyzed, it is observed that the average customer spends a greater share of the money on food items such as grocery, fruit_veg, confectionary etc. Interestingly, the standard deviation of food items is relatively lower than certain other items such as tobacco, drinks, cashpoint and lottery where the variation was more than double the value of the mean as highlighted by **Figure 4.1**. This might imply that the spending on food items might not vary to a great extent among different customers however, spending across non-food items might vary significantly. This provides the opportunity to further analyze the clusters on the basis of types of products purchased.

Statistical Summary of Clusters:

SEGMENT 0

	Cluster	customer_number	Frequency	average_quantity	Monetary	average_spend	Recency	TotalScore
count	865.0	865.000000	865.000000	865.000000	865.000000	865.000000	865.000000	865.000000
mean	0.0	7476.409249	865.855491	1.182865	1321.572694	1.547145	1.673988	10.756069
std	0.0	4002.703416	335.738040	0.082613	572.847633	0.366590	4.257899	1.102150
min	0.0	67.000000	415.000000	1.003717	660.160000	0.820000	0.000000	7.000000
25%	0.0	4163.000000	636.000000	1.122427	915.320000	1.290000	0.000000	10.000000
50%	0.0	7587.000000	772.000000	1.174843	1145.100000	1.480000	0.000000	11.000000
75%	0.0	10728.000000	997.000000	1.236295	1541.140000	1.740000	2.000000	12.000000
max	0.0	16292.000000	3119.000000	1.463687	4448.190000	3.050000	56.000000	12.000000

Cluster 0 has a total of 865 customers which is around 29% of the total customer base. This cluster is ranked the best among all the RFM variables. During the 6-month period, the average customer in this cluster visited the store approximately 865 times and spent an average amount of £1321 which is the highest among all the clusters. Furthermore, the recency figure suggests that the average customers are quite active as the time period that has lapsed since their last purchase is only 1.67 days. The average total score also falls under the "Diamond Category" which means that these customers could be the ideal group for the business. This cluster contributes around 49.5% to the total revenue of the company.

SEGMENT 1

	Cluster	customer_number	Frequency	average_quantity	Monetary	average_spend	Recency	TotalScore
count	1167.0	1167.000000	1167.000000	1167.000000	1167.000000	1167.000000	1167.000000	1167.000000
mean	1.0	7811.487575	389.651243	1.157902	510.265159	1.348920	3.706084	6.941731
std	0.0	4949.229996	130.989463	0.074533	154.905306	0.272571	6.609948	1.635802
min	1.0	45.000000	106.000000	1.006849	126.420000	0.620000	0.000000	3.000000
25%	1.0	3122.000000	289.000000	1.103001	390.875000	1.150000	0.000000	6.000000
50%	1.0	7789.000000	378.000000	1.148472	505.190000	1.330000	2.000000	7.000000
75%	1.0	12232.000000	473.500000	1.203445	618.265000	1.520000	5.000000	8.000000
max	1.0	16316.000000	943.000000	1.427184	1020.410000	2.060000	74.000000	11.000000

Cluster 1 has the greatest chunk of customers i.e. 1167 customers which form 39% of the total customers of the company. This cluster ranks second lowest on the monetary value. This implies that although there are a lot of consumers in this segment, on average they do not spend a lot of money in the store. It is however important to understand the distinction between the average spending of the consumers and the total money that the segment generates for the business. When the totals are analyzed, it can be seen that segment 1 generates **26% of the total revenue** for the business which is 2nd highest after cluster 0. It is interesting to note that this segment ranks better than 3 other segments in terms of recency and frequency. The recency figure is second highest among all other clusters which means that an average customer in this segment made his last purchase only 3.7 days ago. Frequency score of 389 also suggests that the customers visit the store often. The total score classifies this segment into the "Bronze Category". The average spend and quantity purchased is the lowest among all the clusters which might indicate that these customers shop for low involvement and low-cost products more often.

SEGMENT 2

	ıstomer_number	Frequency	average_quantity	Monetary	average_spend	Recency	TotalScore
count	260.0	260.000000	260.000000	260.000000	260.000000	260.000000	260.000000
mean	2.0	8261.488462	405.315385	1.503989	826.237692	2.138500	5.730769	7.580769
std	0.0	4756.753131	227.600511	0.193092	560.676206	0.927154	12.689818	2.101024
min	2.0	14.000000	90.000000	1.283117	205.880000	0.890000	0.000000	3.000000
25%	2.0	4229.000000	259.250000	1.374414	522.970000	1.610000	1.000000	6.000000
50%	2.0	7808.000000	383.000000	1.463268	704.395000	1.880000	2.000000	7.000000
75%	2.0	12618.500000	477.500000	1.546123	1000.975000	2.390000	6.000000	9.000000
max	2.0	16294.000000	2638.000000	2.503688	6588.650000	8.890000	128.000000	12.000000

Cluster 2 has around 260 customers which form 8.7% of the total customer base. This segment has a high average monetary value of £826. It can be analyzed that the average quantity purchased is the highest with the customers on average purchasing 1.5 items

per visit. On the other hand, the average spend is second highest with a value of £2.14. However, it is important to note that despite having a high average monetary value, this segment only **contributes 9% to the total revenue** of the company. This could be attributed to the small size of the segment. If the frequency is considered, the average customer visits the store as frequently as 405 times in a span of 6 months which is the second highest across clusters. However, the recency value ranks 2nd lowest among all the clusters. The last time an average customer visited the store was around 6 days ago. The high variation in this variable could also indicate that some customers might not be actively buying from the store and need attention from the business. This cluster falls closely under the "Silver Category" according to the total score which means that this might be an important segment for the business.

SEGMENT 3

	Cluster	customer_number	Frequency	average_quantity	Monetary	average_spend	Recency	TotalScore
count	402.0	402.000000	402.000000	402.000000	402.000000	402.000000	402.000000	402.000000
mean	3.0	9151.422886	154.885572	1.198440	246.129577	1.657438	40.315920	3.482587
std	0.0	5074.054745	80.956790	0.109117	135.158498	0.537386	42.142641	0.815037
min	3.0	110.000000	6.000000	1.000000	7.280000	0.740000	0.000000	3.000000
25%	3.0	4583.750000	94.750000	1.121611	152.497500	1.312500	9.250000	3.000000
50%	3.0	10462.000000	147.000000	1.177257	221.070000	1.530000	23.000000	3.000000
75%	3.0	13444.750000	203.000000	1.246717	314.970000	1.860000	57.000000	4.000000
max	3.0	16231.000000	519.000000	1.753623	784.760000	5.050000	184.000000	7.000000

Cluster 3 comprises of around 402 customers which form 13.4% of the total customers. This segment performs the worst on all 3 RFM variables. The average customers spend only £246 on the store. The average quantity and spend is also the lowest among all the segments. This cluster only contributes **4% to the total revenue** of the business. If recency is considered, the last time an average customer made a purchase was around 40 days ago. This might imply that these customers initially made some purchases from the business however they have stopped buying. Similarly, the frequency of visits is also only 154 times which is the lowest among all segments. These customers fall into the "Brass Category" according to the total score which means that these could be the least valued customers. It is however important to note that this is the 3rd largest segment of the business in terms of number of customers so this segment might need some attention if the company wants to reduce its churn rate.

SEGMENT 4

	Cluster	customer_number	Frequency	average_quantity	Monetary	average_spend	Recency	TotalScore
count	306.0	306.000000	306.000000	306.000000	306.000000	306.000000	306.000000	306.000000
mean	4.0	9402.660131	294.055556	1.196852	836.055882	2.982549	2.921569	7.454248
std	0.0	4388.652329	141.225132	0.092966	449.903282	1.076434	5.492205	2.105309
min	4.0	85.000000	37.000000	1.000000	156.610000	1.940000	0.000000	3.000000
25%	4.0	5657.000000	186.000000	1.129087	538.672500	2.270000	0.000000	6.000000
50%	4.0	9892.500000	277.500000	1.187570	735.270000	2.685000	1.000000	7.000000
75%	4.0	13196.000000	371.500000	1.250759	985.130000	3.237500	3.000000	9.000000
max	4.0	16295.000000	835.000000	1.556818	2683.550000	10.840000	44.000000	12.000000

Cluster 4 comprises of 306 customers which form around 10% of the customer base. This segment performs well on the monetary and recency variables; however, the frequency variable is second lowest among all the segments. The highest average spend value reveals that these customers might be involved in the purchase of high involvement and relatively expensive products. Recency figure suggests that on average the customers visited the store 3 days ago. It is a good sign as the customers might be actively involved in making purchases in the future as well. However, this segment doesn't shop as frequently as other customers. Despite being less in number, these customers contribute **11% to the total revenue**. These customers fall under the "Bronze Category" according to the total score.

Pen Profiles:

In order to get a deeper understanding about the customer segments and their distinct features, all the segments were also analyzed on the basis of the types of products that they purchase.

K_Cluster	fruit_veg	dairy	confectionary	grocery_food	cashpoint	tobacco	drinks	deli	world_foods	lottery
0	121.404481	121.013477	101.508495	106.248005	63.340980	152.220910	95.589510	23.593862	13.900035	25.078891
1	58.507692	58.497442	44.031452	47.725385	18.079153	23.060199	32.971366	12.009101	7.199948	6.751772
2	63.617868	71.693953	62.407481	61.119302	41.104496	151.713411	62.470388	10.248992	9.686667	10.315891
3	21.684598	24.141281	20.543995	19.832889	12.855905	19.334121	23.452286	4.679045	3.432437	4.173367
4	31.436080	39.349734	27.128738	28.670299	146.334651	231.763256	127.901163	7.973223	4.487076	30.400332

1) **Golden Goose –** These customers fall under Cluster 0. These are the most valuable customers for the company as they purchase regularly and contribute towards half of the revenue of the company. If their spending patterns are analyzed, it can be seen that these customers spend across all product categories, however majority of their spend is concentrated around food items which include fresh food (fruit_veg, dairy) as well as frozen and prepared meals. They also spend heavy amount on other items such as tobacco and drinks. These are likely to be well off customers who might adopt new products quickly. They will also play an important role in promoting the business. These can be long term loyal customers so the business must retain them.

2) **Low Spending Active Loyal Customers –** The customers that fall under Cluster 1 are those who buy more frequently however they spend on low value products. If the categories are analyzed, it can be seen that these customers spend primarily on basic food items such as fruit_veg, grocery and dairy. These are all low value items. These are the customers who spend on a limited number of basic necessities on regular basis. They look for products that offer value for money. Although the average transaction size and value is small, they form the largest segment and are the second largest contributors to the total revenue.

3) **Inactive Potential Loyalists**- These are the customers that fall under Cluster 2. These customers buy in high quantities and spend heavily which is why they have the potential to become loyal customers in the future. However, they haven't visited the store recently. This implies that these customers are currently not active and need to be incentivized to encourage purchases. If their category spend is analyzed, they spend across a variety of categories. Their main spending is focused around food items (fruit_veg, meat, dairy, frozen) as well as tobacco and drinks.

4) **Churned Customers-** These customers fall under Cluster 3 and are those who haven't purchased from the company in over one and a half month. Their spending amount has been insignificant which makes them the least value generating segment. These customers might have left the company for a competitor. If their category spend is analyzed, they spent mostly on dairy, meat and other low value food items.

5) **Infrequent High Spending Customers –** These customers fall under Cluster 4. They have the highest average spend. These customers are currently active, but they do not make purchases as often. If their category spend is analyzed, they spend majorly on tobacco, drinks, cash point and lottery instead of food items. As they are high value customers, they should be kept engaged as they could be a source of sustainability for the business. It might also be easy to attract them towards other high cost products as this segment seems relatively well off. This segment could be retailers/bartenders as well who operate bars/café and make bulk purchases few times a month. However, more information is needed to arrive at this conclusion.

Recommendations:

It is important to understand that all the 5 segments have unique characteristics and different strategies need to be adopted for each of them. The company should prioritize two segments which have the potential to generate most value for the business in the long run. According to the analysis, the first segment that the company could focus on is "**Inactive Potential Loyalists**". As these customers haven't been active recently, it is important to attract them again through relevant promotions. They should be targeted with their wish list items and should be given limited time promotions so that their interest in the store could be restored and they continue making repeat purchases. Surveys could also be run to find out why the customers haven't been active to avoid losing them to the competitors. The second segment that the company could target is "**Infrequent High Spending Customers**". As these customers are heavy spenders, it is important to make these customers feel valued. Incentives such as loyalty points and coupons should be given to them so that they continue to interact with the brand on more frequent basis. Furthermore, their baskets should be analyzed, and product recommendations should be given to them via personalized emails.

Apart from targeting two main segments, the company shouldn't completely withdraw its attention from other segments. **Golden Goose** is another very important segment. As these customers generate a disproportionately high percentage of overall revenue, the company should develop good long-term relationships with them. They could be made part of loyalty schemes in which their individual preferences could be analyzed. They could be given personalized offers and messages to retain their interest. Moreover, **low spending active loyal customers** could also be incentivized so that they increase their average spend. As these customers look for value for money and shop frequently, they could be updated about weekly/daily offers by in store display. This would keep them engaged. These high number of customers could be a great source of spreading positive word of mouth. On the other hand, the segment which the company can potentially pay less attention to are the **churn customers.** As these customers might have left the business and are of low value, it might not be useful to spend resources to re-engage them.

Limitations of Model:

The clustering model uses 5 variables out of which frequency and monetary are correlated. As these variables reflect different information, both of them had to be included in the model. To account for the correlation, initially PCA was used before clustering. However, the clusters obtained as a result were not very distinct. It was then decided to formulate the model without PCA which led to more interpretable clusters. It is understood that correlation would lead to higher weightage being allotted to frequency and monetary in the model which can affect the clustering performance. However, it must be taken into account that these two variables are the most relevant ones in case of convenience store. If weighted RFM were to be used these two variables might have still been given more weightage than the rest of the variables.

Areas of Further Analysis:

For further in-depth analysis of buying patterns, association rule mining can also be applied. It is likely to reveal frequent purchase patterns of each segment. The purchase patterns of people with similar RFM values and demographic characteristics can be analyzed and product recommendations could be formulated. To apply this technique, more information about the description of the products is needed along with relevant customer information. Apart from this, a weighted RFM model could also be used in which weights can be attached to each of the features depending on their relative importance. As the store primarily sells food related items, more weight might be given to the frequency of visits of the customers. This analysis can further be extended to predict the Customer Lifetime Value as well.