CUSTOMER SEGMENTATION REPORT

1. EXECUTIVE SUMMARY

This project aims to conduct a market segmentation analysis on a transactional dataset provided by a national convenience store chain. By performing this analysis, the retailer can gain a better understanding of their customer base and use the insights to inform future marketing campaigns.

The analysis involved forming 5 clusters based on various factors such as category spends, RFM, total and average spend, basket quantity, and the time interval between purchases. The main objective was to identify the two most attractive customer segments for the business.

The customer data provided in the form of four files contained information about the behaviour of 3000 customers. The first file summarized the total number of visits, total quantity of items purchased, average quantity per basket, total spend, and average spend per visit. The second file broke down the customers' spending across 20 item categories. The third file provided details about each visit made by the customers, and the fourth file broke down each basket into individual product purchase IDs and their corresponding categories.

Before conducting the analysis, some data cleaning was required, such as correcting the values for the "bakery" column and removing currency symbols from all tables. Feature engineering was then performed, including generating RFM features, extracting visit frequency and spend data, calculating average basket quantity and spend, and determining visit regularity.

After the above steps, all of these features along with the category spends were merged into one dataframe on the customer_number. This dataframe went through log transformation, detecting negative and missing values and replacing them with 0, scaling (standard scaler) and dimensionality reduction (PCA with 3 components). The resulting dataframe was then clustered into 5 clusters using K-means which were then visualised and statistically summarised followed by a T-test to validate the cluster differences.

Cluster pen profiles were then created representing customer archetypes and 2 most attractive segments were chosen. Cluster 0 and Cluster 1 are the most attractive for the company to target. Cluster 0 customers prioritize convenience and practicality, and are willing to pay a premium for high-quality products, while Cluster 1 customers prioritize quality over price and frequently host or attend gatherings.

2. CUSTOMER BASE SUMMARY

- Based on the analysis of the company's customer data, it appears that the customer base is quite diverse and may consist of different demographic segments such as families, bachelors, students, and immigrants. The customers exhibit a wide range of spending patterns, from very high spenders to very low spenders.
- In terms of shopping behaviour, the data suggests that customers tend to prefer shopping on Thursdays and Fridays, and the morning and afternoon time slots are the most popular times for shopping.

- Regarding product preferences, the customer base appears to be varied as well. Customers purchase a range of products from healthy options like fruits and vegetables to cooked food items to addictive products such as tobacco and alcohol. However, the most popular choices seem to be fruits, vegetables, tobacco, drinks and dairy.
- Using the engineered features, we can also infer that the company has a large number of customers who tend to make frequent purchases in moderate quantities. This aligns with the convenience store model, where customers may visit more often for quick purchases rather than stocking up on large quantities. The data also indicates that the usual interval between each purchase is less than 10 days, which further supports this idea.

3. FEATURE DESCRIPTION

Feature Engineering and justification:

Based on domain knowledge we can merge the features in category_spends_sample table:

- daily_groceries: This feature can represent the total amount spent by customers on fruit_veg, dairy, grocery_food, and soft_drinks.
- tobacco_drinks: This feature can represent the total amount spent by customers on tobacco and drinks.
- bakery_sweets: This feature can represent the total amount spent by customers on confectionary, bakery, and discount_bakery items.
- prepared_food: This feature can represent the total amount spent by customers on prepared_meals and deli items.
- newspapers_magazines_practicals: This feature can represent the total amount spent by customers on newspapers_magazines and practical_items.
- Drop the lottery, cashpoint, and seasonal_gifting features since they do not provide significant information for clustering analysis. Other features can stay.

RFM features were generated using lineitems_sample table:

- Recency: The number of days since the customer's last purchase is calculated by subtracting the maximum purchase time for each customer from the overall maximum purchase time in the dataset. It can help identify customers who haven't visited the store recently and may need a reminder or incentive to return.
- Frequency: The total number of purchases made by each customer is calculated by counting the quantity of each purchase for that customer. It can help identify loyal and engaged customers who make frequent purchases and may be more receptive to loyalty programs or special offers.
- Monetary Value: The total amount of money spent by each customer is calculated by summing up the spend of each purchase for that customer. It can help identify high-value customers who may be worth targeting with special promotions or incentives.

Average basket quantity and visit regularity features were generated using baskets_sample table:

- The "basket_quantity" is the total number of items purchased in each transaction, and by taking the average, we get an estimate of the typical number of items purchased by each customer. This information can be useful for understanding customer preferences and behaviour.
- The time difference between visits for each customer is calculated by subtracting the purchase date from the previous purchase date using the shift() method. Finally, the median visit interval

for each customer is calculated using the median() method. The resulting regularity dataframe provides insights into how frequently customers visit the store.

Extracting average spend from customers_sample table:

- 'average_spend': The average amount a customer spends per visit is an important metric as it can give insights into the customer's spending patterns and buying behavior. This can help in understanding which products or categories are most popular among customers and what promotions or discounts might be effective in encouraging customers to spend more.

Features Selected:

'grocery_health_pets', 'frozen', 'meat', 'world_foods', 'daily_groceries', 'tobacco_drinks', 'bakery_sweets', 'prepared_food', 'newspapers_magazines_practicals', 'average_spend', 'Recency', 'Frequency', 'MonetaryValue', 'basket_quantity', 'visit_regularity'

4. SEGMENTATION METHODOLOGY AND JUSTIFICATION

Data Pre-processing:

I. The data had '£' and ',' symbols ***Justification:*** which would be considered as NaN by the program and hence had to be removed.
II. All the columns had to be cast to float and 'purchase_time' column to datetime type.
III. RFM and visit regularity features were newly engineered along with derived features such as average basket quantity.
IV. All of the relevant columns were merged on customer number and stored in a dataframe merged_df which was then tested for missing values and negative values.
Justification: Those values were replaced with 0 as I felt that negative spend and negative quantity might indicate product return; negative spend and positive quantity might indicate had to pay back money but the product was damaged. However, it is important to note that these are my assumptions and have no information on why these values are in negative. Based on my assumptions, I replaced them with 0.
V. Merged_df was highly skewed (Fig1) and thus it was log transformed and 'basket_quantity' and 'visit_regularity' were excluded ***Justification:*** as a log transformation would result in the model predicting changes in the natural logarithm of the quantity and interval between purchases rather than changes in the original quantity itself.
VI. Then the data was scaled using StandardScaler() ***Justification:*** To avoid the dominance of certain features.
VII. The scaled data was then passed through a stage of dimensionality reduction using PCA. On using same number of components as the features and plotting a scree plot(Fig2), ***Justification:*** I found the elbow drops at 3 and thus decided on 3 Principal Components for further analysis.

Clustering:

I. K-means clustering algorithm was chosen to undertake this task.
II. Using elbow method, a WCSS plot (Fig3) revealed that 5 clusters would be a good fit for this data. ***Justification:*** 5 was ideal as it performed decent enough and also the company had requested a range from 5-7 clusters. The elbow dropped at 3 but 5 was the closest number to the minimum cluster range and hence was chosen.

III. Subsequently, K-means algorithm was fit to the data with the hyperparameters: *(n_clusters=5, init='k-means++', max_iter=100, n_init=5, random_state=42)* and the average silhouette score was: 0.27080904120181376.

5. CLUSTER PEN PROFILES

Cluster 0: Bulk Shoppers

- Customers in this cluster have the highest average basket quantity (29.34) *(Tab4)(Fig6)* and spend the most in terms of both frequency (560.80) and monetary value (841.63)*(Tab5)(Fig7).* They tend to purchase a wide range of products, with the highest spending categories being daily groceries, bakery and sweets, and frozen food. They also spend a considerable amount on newspapers and magazines, practicals, and prepared food *(Tab1)(Fig8).* In terms of days between purchases, customers in this cluster tend to make purchases frequently (every 7.4 days)*(Tab2)(Fig5).*
- Customer Archetype: Busy urbanites with high incomes who value convenience and are willing to pay a premium for high-quality products. Customers in this cluster likely have a busy lifestyle and prefer to make frequent trips to the grocery store to ensure that their household is well-stocked. They prioritize convenience and practicality, hence the high spending on prepared food and daily groceries. They may also have a sweet tooth, hence the high spending on bakery and sweets. They value staying informed, hence the high spending on newspapers and magazines. Based on the above information, the customers might be families.

Cluster 1: Extravagant socialisers and Daily Shoppers

- Customers in this cluster have a relatively lower average basket quantity (11.83) but spend more per purchase, with the highest frequency (851.43) and monetary value (1336.86) compared to the other clusters. They tend to spend more on tobacco, drinks, and world foods, as well as newspapers and magazines, practicals, and prepared food. In terms of days between purchases, customers in this cluster make purchases more frequently (every 1.77 days) compared to other clusters.
- Customer Archetype: Affluent and sophisticated customers who prioritize quality over price and are willing to spend more on premium products. Customers in this cluster likely enjoy socializing and frequently host or attend gatherings. They prioritize quality over quantity, hence the higher spending per purchase. They may have a preference for exotic flavours, hence the high spending on world foods. They may also have a habit of reading the news and staying informed while enjoying a drink, hence the high spending on newspapers and magazines. Based on the above information, the customers might be young professionals or middle-aged professionals.

Cluster 2: Budget Shoppers and Smokers

- Customers in this cluster have the lowest average basket quantity (7.01) and spend less frequently (326.51) and with a lower monetary value (748.64) compared to the other clusters. They tend to spend the most on tobacco, drinks, and frozen food, with lower spending in other categories. In terms of days between purchases, customers in this cluster make purchases every 2.64 days on average. They have relatively higher average spend than all other clusters (2.379) *(Tab3)(Fig4)*
- Customer Archetype: Customers in this cluster likely have a smoking habit and prioritize convenience when it comes to their purchases. They may not have a wide variety of products they purchase and are likely to prioritize tobacco and drinks. Frozen foods may also be a convenient

option for them as they require minimal preparation. Based on the above information, the customers might be bachelors.

Cluster 3: Sweet Tooth Snackers

- Customers in this cluster have a relatively low average basket quantity (9.29) and spend less frequently (394.08) and with a lower monetary value (490.78) compared to the other clusters. They tend to spend the most on bakery and sweets, daily groceries, and prepared food, with lower spending in other categories. In terms of days between purchases, customers in this cluster make purchases every 3.06 days on average.
- Customer Archetype: Customers in this cluster likely have a preference for sweet snacks and prioritize indulgence in their purchases. They may not have a wide variety of products they purchase and are likely to prioritize bakery and sweets. Prepared food may also be a convenient option for them as they require minimal preparation. Based on the above information, the customers might be families with kids or young students.

Cluster 4: Budget-conscious shoppers

- Customers in this cluster have a relatively low average basket quantity (8.74) and spend the least frequently (126.19) and with the lowest monetary value (225.46) compared to the other clusters. They tend to spend the most on daily groceries, frozen food, and meat, with lower spending in other categories. In terms of days between purchases, customers in this cluster make purchases every 6.45 days on average. With a median recency of 11, they seem a little off track with reference to days between purchases and an average recency of 31 which is alarming and needs the company's attention.
- Customer Archetype: Customers in this cluster likely prioritize budget when it comes to their purchases. They may not have a wide variety of products they purchase and are likely to prioritize essential items such as daily groceries and meat. Frozen food may also be a cost-effective option for them. They may also prefer to make fewer trips to the grocery store to save on transportation costs. Based on the above information, the customers might be students or immigrants which also explains the high recency.

6. BUSINESS RECOMMENDATION

The two most attractive clusters for the company to target are Cluster 0: Bulk Shoppers and Cluster 1: Extravagant socialisers and Daily Shoppers

a) Cluster 0 is attractive because customers in this cluster have the highest average basket quantity and spend the most in terms of both frequency and monetary value. They prioritize convenience and practicality, and are willing to pay a premium for high-quality products. These customers are likely families and busy urbanites with high incomes who value convenience. Therefore, the company can target this cluster by offering a wide range of high-quality products and emphasizing the convenience factor.

b) Cluster 1 is also attractive because customers in this cluster have a high frequency and monetary value of purchases, and they prioritize quality over price. They are likely young or middle-aged

professionals who enjoy socializing and frequently host or attend gatherings. The company can target this cluster by offering premium and exotic products, as well as convenient prepared food options for hosting events. Additionally, the company can offer loyalty programs or other incentives to encourage repeat purchases from this cluster.

Here are some business case recommendations for each of these clusters:

Cluster 0 (Bulk Shoppers):

- Offer a loyalty program: Given that customers in this cluster tend to make frequent purchases and have a high average basket quantity, a loyalty program that rewards them for their loyalty and incentivizes them to keep shopping at the company's stores could be effective.
- Introduce a subscription model: Since these customers tend to purchase a wide range of products and prioritize convenience, offering a subscription model for regular purchases of their most commonly bought products could be an attractive option.
- Expand the prepared food section: As customers in this cluster spend a considerable amount on prepared food, expanding the company's prepared food section could be a smart move. Offering high-quality, healthy prepared meals that cater to a wide range of dietary requirements could appeal to this cluster's busy urbanites.
- Personalized Offers and Discounts: Based on their high spending on daily groceries, bakery and sweets, and frozen food, the company can offer personalized discounts and deals on these categories to incentivize bulk purchases.
- Convenient Delivery Options: Since these customers prioritize convenience and practicality, the company can introduce more convenient delivery options like same-day delivery, flexible delivery time slots, and subscription services to cater to their busy lifestyle.

Cluster 1 (Extravagant socialisers and Daily Shoppers):

- Offer premium products: As customers in this cluster prioritize quality over price and are willing to spend more on premium products, the company should consider offering a range of high-quality, premium products. This could include exotic flavours and top-shelf alcohol options.
- Host in-store events: Given that these customers enjoy socializing, hosting in-store events such as wine tastings or cooking classes could be a smart way to attract and retain their business.
- Leverage social media influencers: These customers may be influenced by social media influencers who align with their values and interests. Partnering with relevant social media influencers to promote the company's products and events could help attract this cluster's business.
- Customized Gift Sets: Since these customers enjoy hosting or attending gatherings, the company can create customized gift sets and hampers for special occasions like birthdays, anniversaries, and holidays to cater to their gifting needs.
- Personalized Shopping Experience: To enhance their shopping experience, the company can introduce a personalized shopping experience by providing dedicated personal shoppers, exclusive previews, and VIP access to new products and promotions.

APPENDIX

Fig1

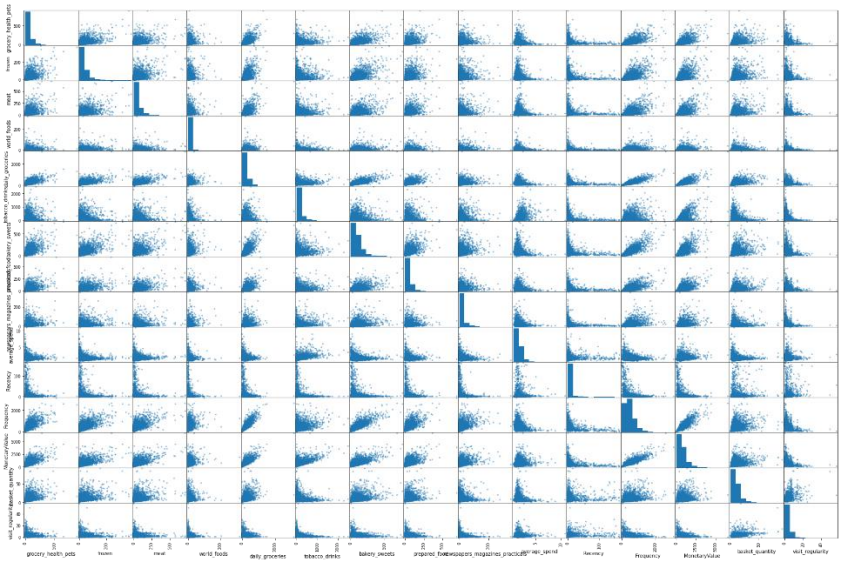


Fig2

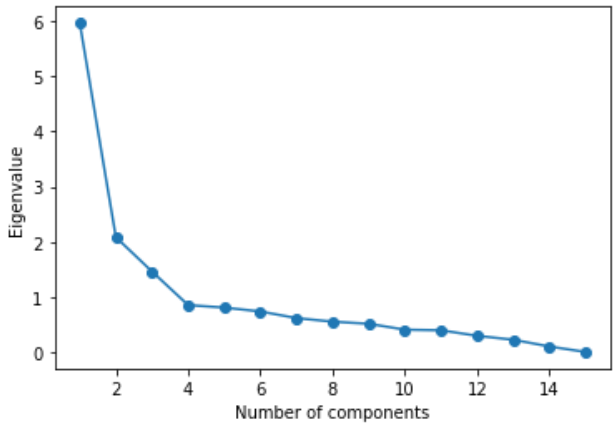


Fig 3

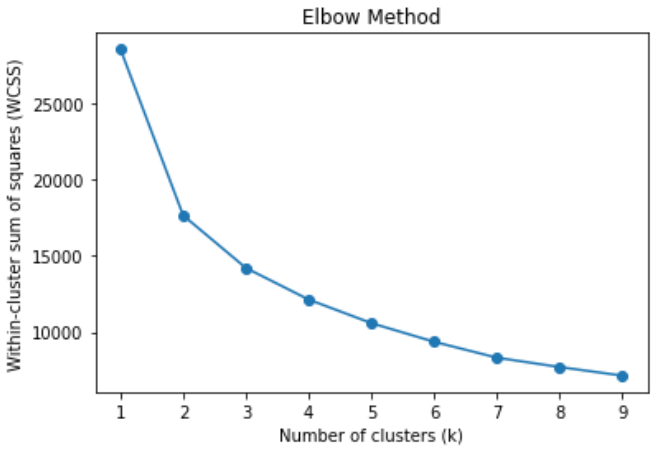


Fig4

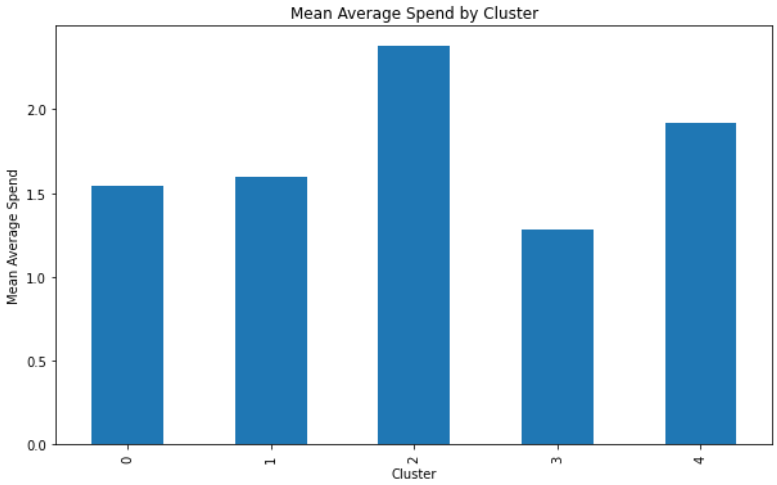


Fig5

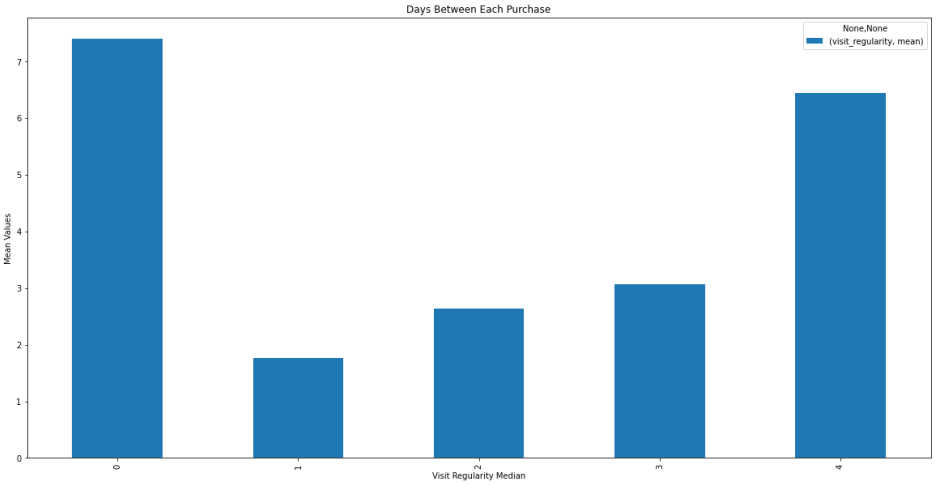


Fig6

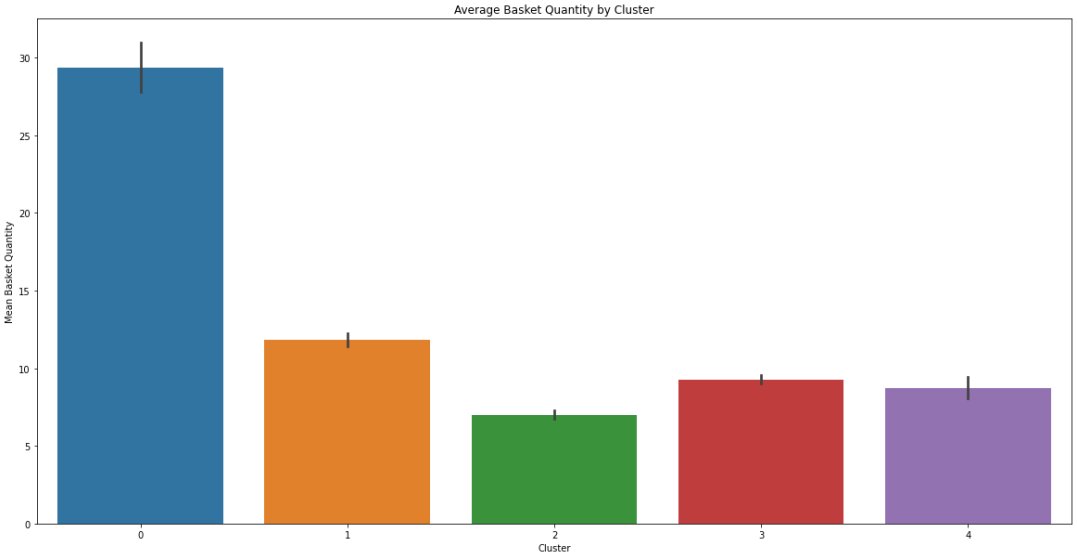


Fig7

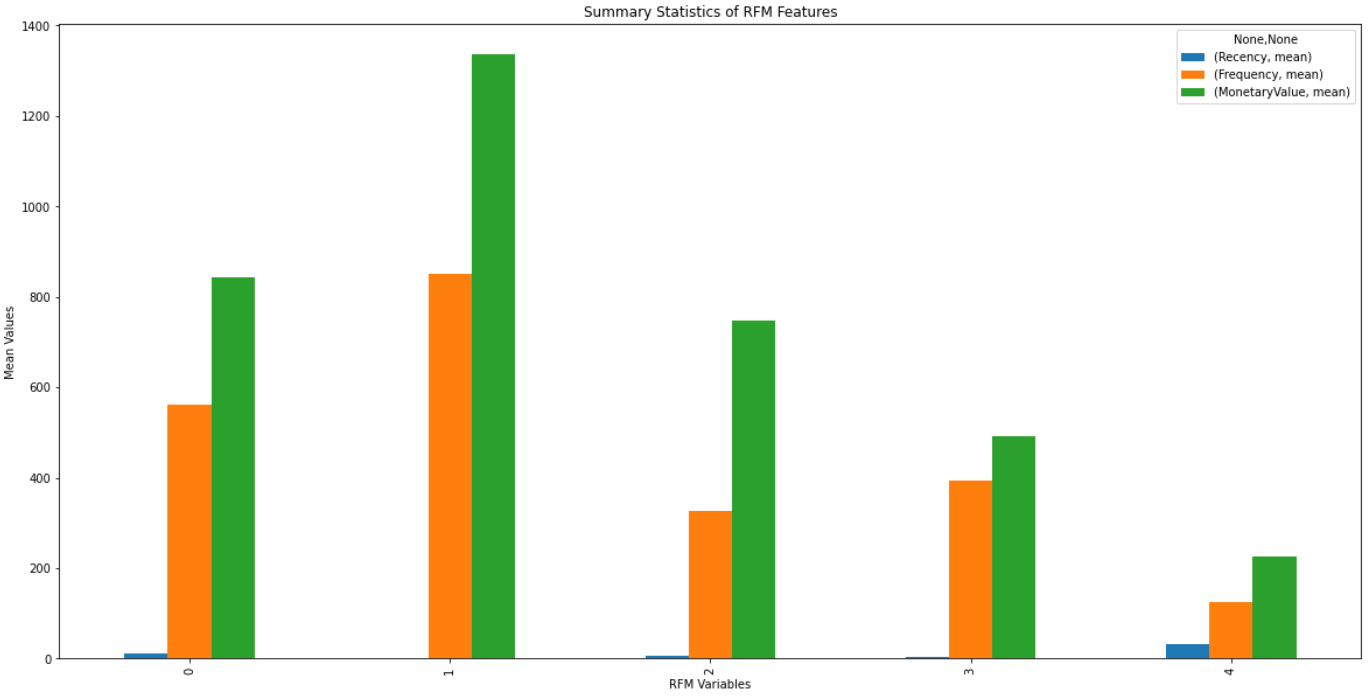
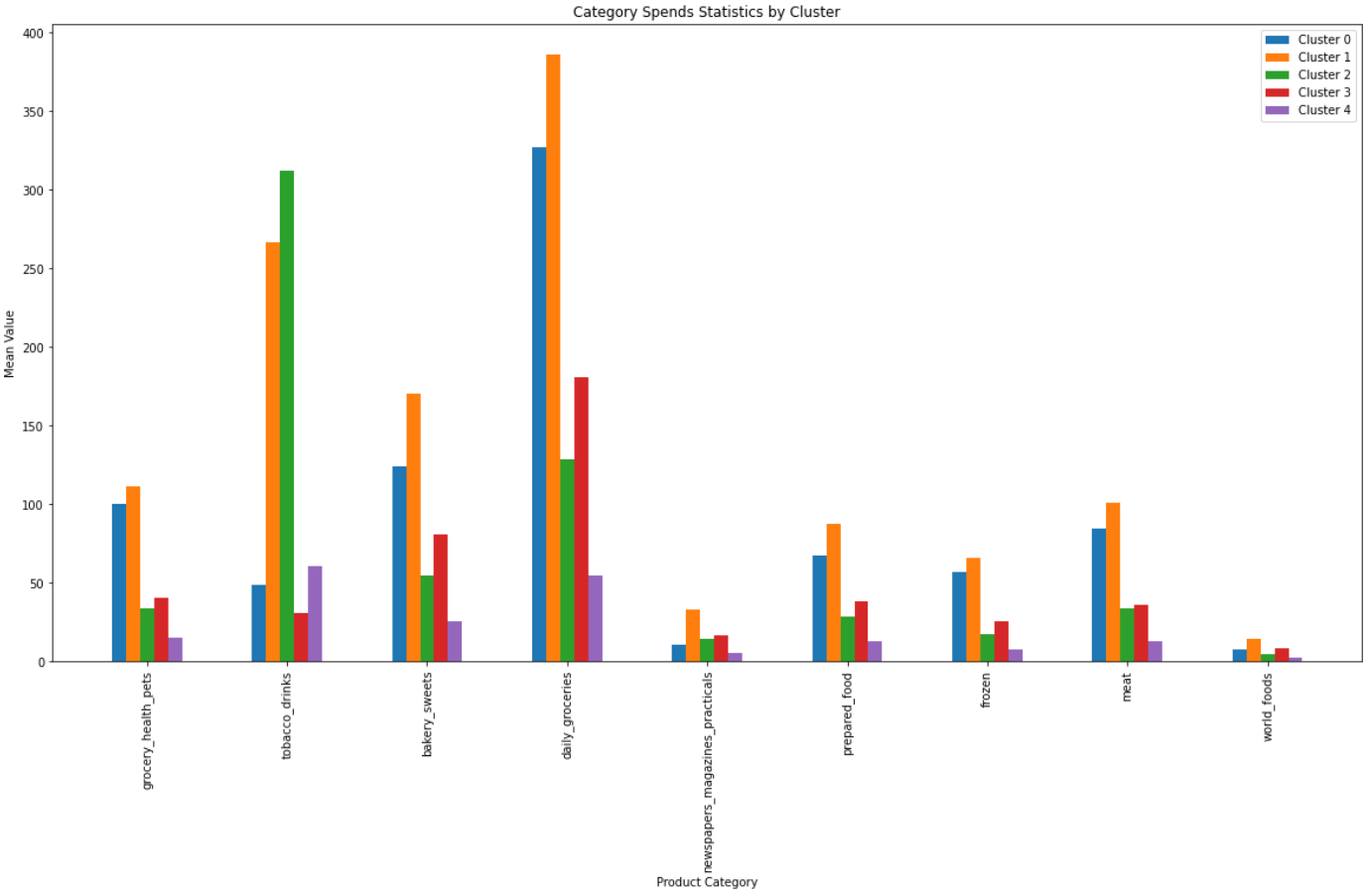

Summary Statistics of RFM Features
None,None
(Recency, mean)
(Frequency, mean)
(MonetaryValue, mean)
Mean Values
RFM Variables
0
1
2
3
4

Fig8


Category Spends Statistics by Cluster
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Mean Value
Product Category
grocery_health_pets
tobacco_drinks
bakery_sweets
daily_groceries
newspapers_magazines_practicals
prepared_food
frozen
meat
world_foods

Tab1 – Mean of category spends

Cluster	Grocery, Health, Pets	Tobacco, Drinks	Bakery, Sweets	Daily Groceries	Newspapers, Magazines, Practicals	Prepared Food	Frozen	Meat	World Foods
0	100.158	48.411	124.439	326.834	11.011	67.391	56.765	84.535	7.613
1	111.508	267.053	170.110	386.469	32.938	87.498	65.704	101.285	14.761
2	33.999	312.396	54.479	128.594	14.599	28.574	17.296	33.664	4.630
3	40.809	30.738	80.784	180.670	16.452	38.262	25.602	35.816	8.265
4	15.230	60.618	25.698	54.637	5.077	12.577	7.466	12.709	2.515

Tab2 – Visit regularity summary statistics

Cluster	Regularity (mean)	Regularity (median)	Regularity (std)
0	7.407407	7.0	4.572906
1	1.770452	1.0	1.059562
2	2.644714	2.0	1.667274
3	3.063874	3.0	1.777186
4	6.448549	5.0	5.556170

Tab3 – Average spend summary statistics

Cluster	Spend (mean)	Spend (median)	Spend (std)
0	1.542296	1.475	0.347817
1	1.597009	1.500	0.439232
2	2.379168	2.150	0.936497
3	1.281749	1.280	0.237445
4	1.916121	1.600	1.024931

Tab4 – Basket quantity summary statistics

Cluster	Qty (mean)	Qty (median)	Qty (std)
0	29.337037	26.0	12.971850
1	11.831502	10.0	6.222322
2	7.008666	6.0	3.297452
3	9.286911	8.0	3.963061
4	8.736148	7.0	6.435602

Tab5 – Mean of RFM

Cluster	Recency (mean)	Frequency (mean)	MonetaryValue (mean)
0	11.962963	560.803704	841.629370
1	1.855922	851.432234	1336.857375
2	5.256499	326.514731	748.640555
3	4.952880	394.082723	490.782963
4	31.261214	126.192612	225.460053