# Analytics Specializations and Applications

**BUSI4370 UNUK - SPR1 23-24**

**Student ID: 20535493**

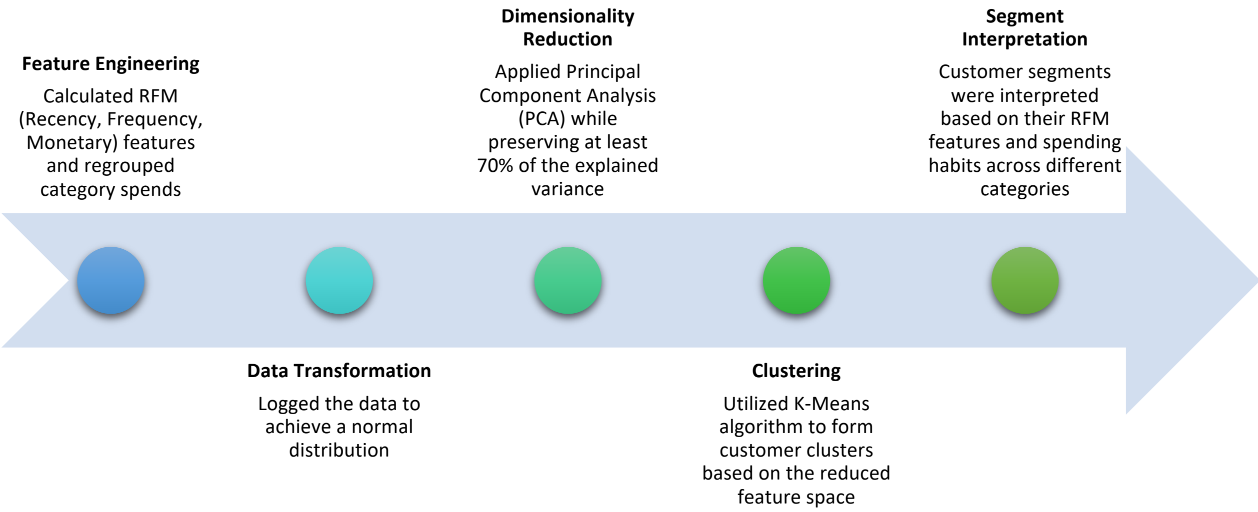
## Customer Segmentation Report

# Executive Summary

This report details the customer segmentation process conducted to understand purchasing behavior. Considering customer behavior metrics like purchase frequency, item variety, and spending habits, the analysis revealed **six** distinct customer clusters. Each segment exhibits unique characteristics, providing valuable information for targeted marketing strategies, personalized promotions, and enhancing customer experiences.

The dataset consisted of historic transaction data, down to the level of each product purchased. This data was utilized to calculate RFM features, customer lifetime value, lottery spending to winning ratio, and product category-wise spending.

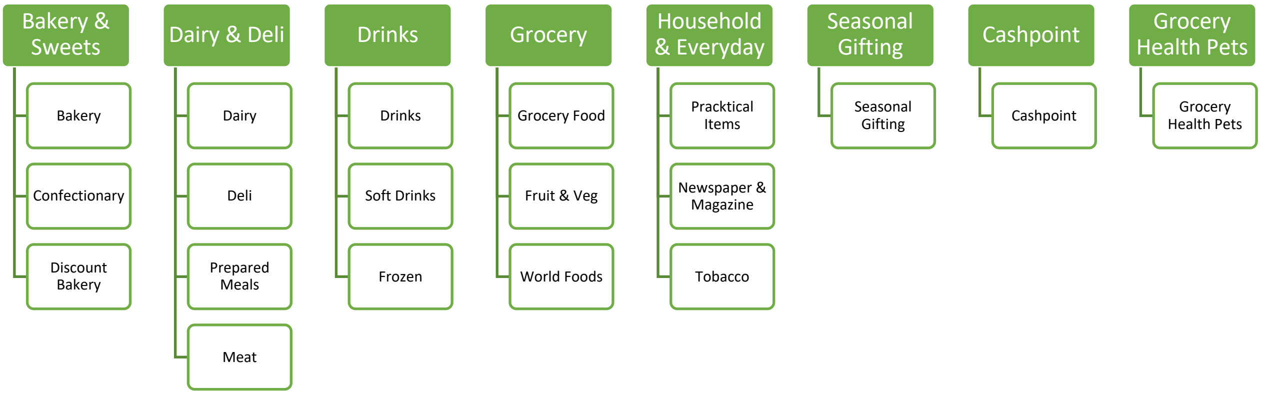
Technical Approach taken to achieve the segmentation is as follows:




Six distinct customer segments were identified. Each segment exhibited unique characteristics in terms of average days between orders, item variety per bucket, and spending habits across different categories. The segments were named based on their distinguishing features and behaviours. Segments formed:

# Feature Description

The dataset provided comprises of transaction-level information, allowing for a granular analysis of customer spending across various product categories. The first step involved aggregating customer expenditures into broader categories in order to extract valuable insights from this data. The dataset was further reduced from its original 20 distinct categories to **8 higher-level categories** by combining related product categories. By understanding consumer preferences and spending habits, this consolidation enables businesses to base their decisions on broad category trends. **The categories were regrouped** as shown below –




RFM (Recency, Frequency, Monetary) features, which are well-known for their ability to accurately describe consumer behaviour and inform segmentation tactics are calculated for feature space. This decision reflects a comprehensive understanding of customer segments by capturing important aspects of customer engagement and value.

**Customer Lifetime Value (CLV)**

- Overall value a customer provides to the company during the course of their association. It serves as a key metric for customer profitability and loyalty

**Average Days Between Visits**

- The average time interval between two consecutive purchases made by a customer. It offers insights into customers' purchasing frequency and helps identify patterns in their buying behavior

**Item Variety per Bucket**

- Diversity of products purchased within each product category by a customer. It indicates the breadth of a customer's shopping preferences and their willingness to explore different product offerings

**Lottery Spend to Win Ratio**

- Reflects the ratio of money spent on lottery tickets to the amount won. It sheds light on customers' gambling behavior and their propensity to engage in risky spending activities, providing insights into their disposable income and recreational preferences.

Therefore, the feature set comprises eight high-level categorical spends and **four calculated RFM features** (as shown to the left of this text), which are considered fundamental in customer segmentation and analysis.

This approach ensures that the feature set captures essential aspects of customer spending patterns while **maximizing the utilization** of available information in the dataset.

## Customer Base Summary

The dataset consists of four CSV files, each containing detailed behavioral data from **3000 customers**, recorded through loyalty card transactions. These records span a period of **six months**, providing valuable insights into customer behavior. Upon review, it was noticed that the features in some files were derived from **transaction line items**, leading to inconsistencies. To ensure data accuracy, all required features were reconstructed using the transaction lines file. This approach maintains data integrity throughout the analysis.
Customers spend was aggregated and grouped across 20 categories. Below chart shows the distribution of category spend across all customers.

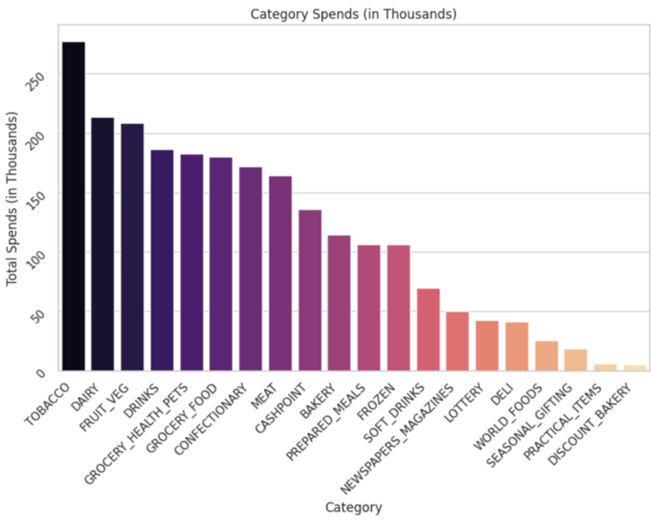



Figure 1: Category Spends (in Thousands) across Categories

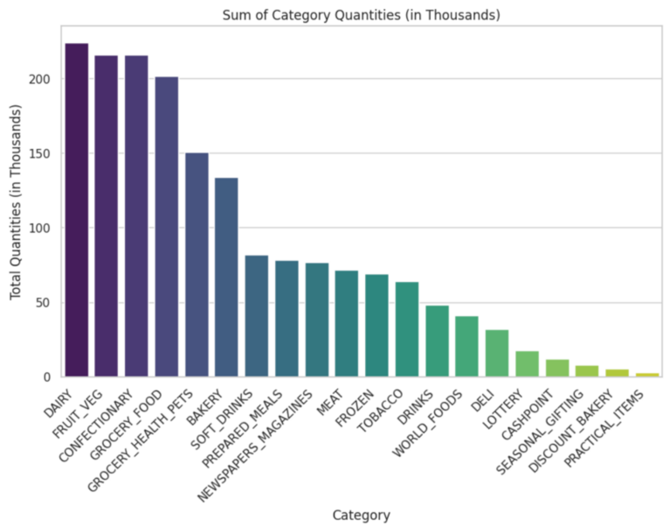



Figure 2: Category Quantities (in Thousands) across Categories

The visualizations depicting category spend and quantities purchased reveal a strong correlation across most categories, indicating consistent spending patterns relative to the quantities acquired. Still, there's a noteworthy outlier in the tobacco category. It is ranked highest in terms of customer spend, but when it comes to quantity purchased, it contrasts greatly and is in the lower half of the categories. Figure on the right shows correlation matrix of additional features created. All these fields have low mutual correlation as desirable for applying clustering algorithms like K-Means clustering.

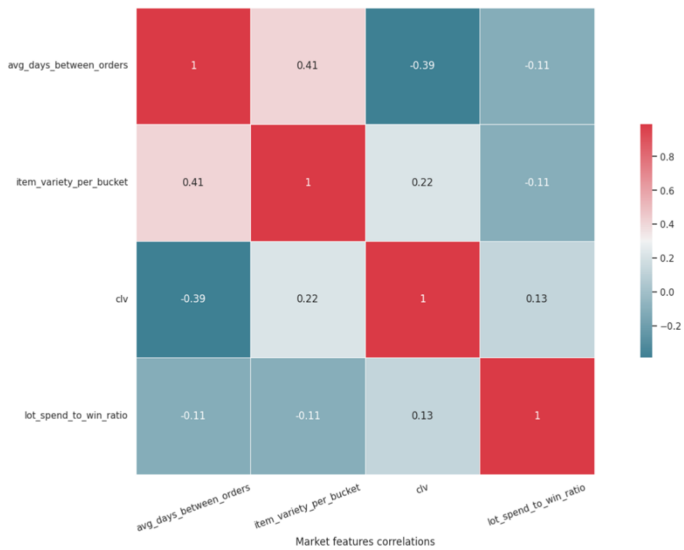



Figure 3: Correlation matrix of RFM features

## Segmentation Methodology

Several steps were involved in the segmentation methodology used in order to effectively cluster and comprehend the underlying patterns in the dataset:

**Dimensionality Reduction with PCA**: The modified dataset had twelve features at first. Principal Component Analysis (PCA) was selected as a technique to decrease dimensionality while maintaining data variance because of the high dimensionality and possible redundancy in features. It was crucial to deal with the data's skewed distribution before using PCA. This was accomplished by transforming the dataset using a logarithmic function to bring it closer to a more normal distribution. PCA was then applied to the preprocessed dataset to reduce the number of dimensions. The aim was to capture the maximum variance in the data with a smaller number of principal components. By retaining the principal components that explained at least 70% of the cumulative variance, the dataset was effectively reduced to three dimensions, whose distributions can be seen on the left.

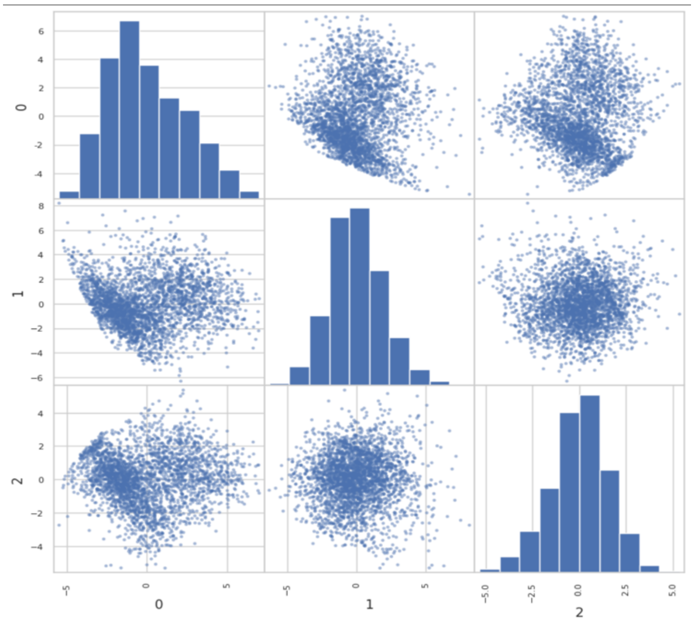


*Figure 4: Scatter Matrix of the reduced PCA components*

**Evaluation of Clustering Algorithms**: After the dimensionality was decreased, the data was divided into distinct segments using clustering algorithms. For this task, the unsupervised learning algorithm K-means clustering was selected. It was necessary to determine the ideal number of clusters (k) before moving further. A measure of cluster cohesion and separation called the silhouette score was calculated for a range of values of k, from 2 to 11. To further illustrate the connection between k and the total squared distances from every point to the designated cluster centroid, an elbow curve was also plotted as shown to the right. Based on the evaluation metrics and visual inspection of the elbow curve, the optimal number of clusters was determined to be six. The K-means clustering algorithm was then applied with k=6 (highest Silhouette score between 5 and 7) to segment the dataset into distinct clusters.

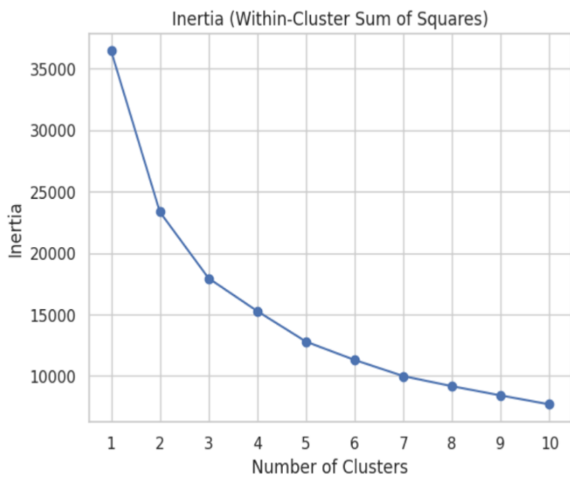



*Figure 5: Elbow curve to find optimum k value*

# Results

The cluster visualization presented in the figure provides a graphical representation of the six segments derived from the K-means clustering algorithm. Each point in the plot represents a customer, with their position determined by the values of the three principal components obtained through PCA. By plotting these points in a three-dimensional space, the clusters formed by the algorithm are visualized, allowing for an intuitive understanding of their distribution and separation.

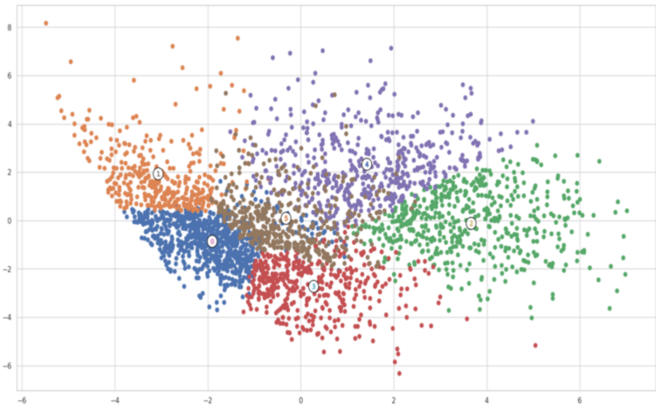


*Figure 6: Visualization of clusters*

Furthermore, detailed insights into the customer profiles and archetypes are provided in the next paragraph. These insights delve deeper into the characteristics of each cluster, highlighting key attributes and behaviours that distinguish one segment from another. By analysing the features that contribute most significantly to the formation of each cluster, such as average days between orders, item variety per bucket, customer lifetime value (CLV), and others, a comprehensive understanding of the customer segments is developed. The cluster centres are as shown below:

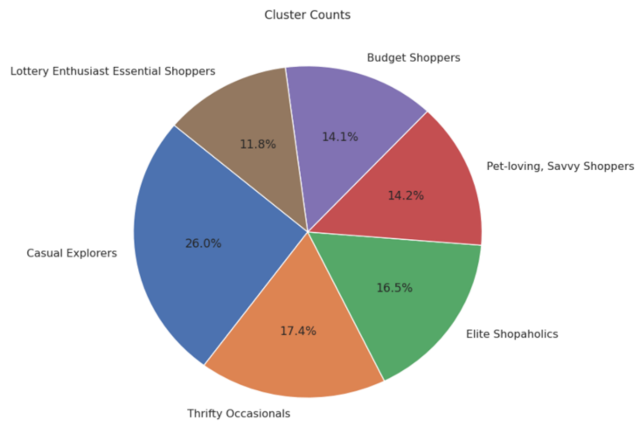



*Figure 7: Proportion of customers in clusters*

| | Segment 1 | Segment 2 | Segment 3 | Segment 4 | Segment 5 | Segment 6 |
|---|---|---|---|---|---|---|
| Avg days between visits | 4 | 6 | 1 | 2 | 3 | 3 |
| Item variety per bucket | 11 | 7 | 7 | 11 | 6 | 6 |
| CLV | 537 | 218 | 1265 | 1192 | 476 | 525 |
| Lottery Spend to Win Ratio | 0 | 0 | 4 | 1 | 2 | 3 |
| Cashpoint | 0 | 0 | 85 | 0 | 59 | 0 |
| Grocery Health Pets | 45 | 9 | 67 | 104 | 18 | 20 |
| Seasonal Gifting | 2 | 0 | 4 | 6 | 1 | 2 |
| Bakery & Sweets | 80 | 26 | 120 | 150 | 46 | 48 |
| Dairy & Deli | 159 | 49 | 201 | 296 | 76 | 86 |
| Drinks & Frozen | 77 | 21 | 128 | 170 | 41 | 45 |
| Grocery | 132 | 40 | 139 | 235 | 54 | 65 |
| Everyday & Household | 6 | 4 | 134 | 70 | 23 | 120 |

**<u>Segment 1 - Casual Explorers</u>:**

*Casual Explorers* represents a customer profile characterized by high item variety per bucket and moderate customer lifetime value (CLV), suggesting a consistent but not overly substantial level of spending. The archetype of this segment indicates that the customers enjoy exploring a variety of products and categories during their shopping experiences. They are likely to be open to trying new items and may value novelty and variety in their purchases.

In terms of specific category spends, *Casual Explorers* allocate a significant portion of their budget to Grocery, Dairy and Deli indicating a focus on essential household items and convenience food items. The absence of spending on cashpoint and low lottery spend-to-win ratio suggests that these customers may not heavily engage in lottery or gaming activities.
Overall, the customer profile suggests a segment of consumers who value variety, convenience, and practicality in their shopping habits, while also demonstrating a degree of openness to exploring new products and experiences.

**Segment 2 – Thrifty Occasionals:**
*Thrifty Occasionals* represent a customer profile characterized by infrequent purchasing behaviour and a relatively limited variety of items per shopping occasion. These customers exhibit lowest average customer lifetime value (CLV), indicating more restrained spending habits. In terms of specific category spends, these customers tend to spend most on Grocery, Dairy and Deli.
Overall, the customer profile of *Thrifty Occasionals* suggests these customers are likely to make purchases only when necessary or on specific occasions, rather than engaging in regular shopping routines. They may prioritize saving money and being frugal in their purchasing decisions.

**Segment 3 – Elite Shopaholics:**
*Elite Shopaholics* represents a group of customers characterized by their extremely high spending habits and frequent visits to the store. With an average of just 1 day between visits, these customers exhibit a strong preference for regular shopping trips. They also demonstrate a remarkable item variety per bucket, suggesting a broad range of interests and preferences when it comes to products. The Customer Lifetime Value (CLV) for this segment is notably high, indicating that these individuals are high-value customers, contributing significantly to the store's revenue. Moreover, the lottery spend to win ratio of 4 suggests that these customers are willing to take part in promotional activities and invest in lottery tickets, potentially driven by a desire for excitement or the chance to win rewards.
Therefore, the individuals here are highly engaging and affluent customers.

**Segment 4 – Pet-loving, Savvy Shoppers:**
*Pet-loving, Savvy Shoppers* comprises customers who exhibit a combination of high spending and a particular affinity for the Grocery Health Pets category. With an average of 2 days between visits, these individuals maintain a relatively frequent shopping tendency. With a notably high CLV of 1192 and diverse purchasing behaviour in each visit, the individuals here can be considered high value customers.
A distinguishing characteristic of *Pet-loving, Savvy Shoppers"* is their substantial spending in the Grocery Health Pets category. This suggests that these customers prioritize the well-being of their pets and are willing to invest in high-quality pet products.

**Segment 5 – Budget Shoppers:**
*Budget Shoppers* is characterized by customers who exhibit moderate frequency in their shopping behaviour, with an average of 2 visits per week. While their item variety per bucket is relatively lower at 6, suggesting less diversity in their purchasing patterns compared to other segments, their CLV is also moderately low at 476.

Budget Shoppers display a notable preference for cash transactions, as indicated by their higher-than-average spending through cashpoints. They allocate a significant portion of their budget to essential categories such as Bakery & Sweets, Dairy & Deli, and Drinks & Frozen, albeit to a lesser extent compared to other segments. This indicates a practical approach to their shopping habits, focusing on essential items while maintaining budget-conscious behaviour.

**Segment 6 – Lottery Enthusiasts Essentials Shoppers:**
Customers in this segment visit the store approximately every 3 days on average and have a moderate variety of items in their buckets. They exhibit a relatively lower CLV compared to other segments, indicating moderate spending habits. They are somewhat interested in lottery spending, with a spend-to-win ratio of 3. They do not use cashpoints frequently. In terms of product categories, they spend relatively less on Grocery Health Pets and Bakery & Sweets compared to other segments. *Lottery Enthusiasts Essentials Shoppers* tend to buy everyday items like newspapers, magazines and tobacco the most. Their moderate spending habits and frequent visits suggest a balance between practicality and enjoyment in their shopping behaviour.

## Summary

Six unique consumer segments were identified by the clustering analysis based on their spending and shopping habits. Every segment denotes a distinct customer archetype with particular traits and inclinations. Based on the analysis, two segments that are of particular importance for the company to focus attention on are:

| | |
|---|---|
| Segment 3: Elite Shopaholics | These customers exhibit high spending habits, frequent visits, and a diverse range of product preferences |
| | **Recommendation**: Implement loyalty programs, VIP perks, and exclusive offers to incentivize continued spending and enhance customer loyalty |
| Segment 4: Pet-loving, Savvy Shoppers | This segment shows a strong affinity for pet-related products and demonstrates savvy shopping behavior. |
| | **Recommendation**: Introduce pet-focused promotions, partnerships with pet brands, and personalized recommendations to cater to their specific needs and preferences |

From marketing recommendations point of view, product recommendations, website content, and email marketing campaigns can all be made more personalized by using customer segmentation data. Use social media platforms to advertise in a way that is specifically targeted to each segment's preferences. Get input from clients in each market segment to improve marketing tactics and product lines over time.
However, there is some scope for further analysis. By conducting A/B testing, the effectiveness of different marketing strategies and promotions for each segment can be evaluated. Furthermore, demographic or psychographic variables can help in refining customer segmentation.

All things considered, the business can increase sales, forge closer bonds with customers, and maintain its lead in a crowded market by utilizing customer segmentation insights.