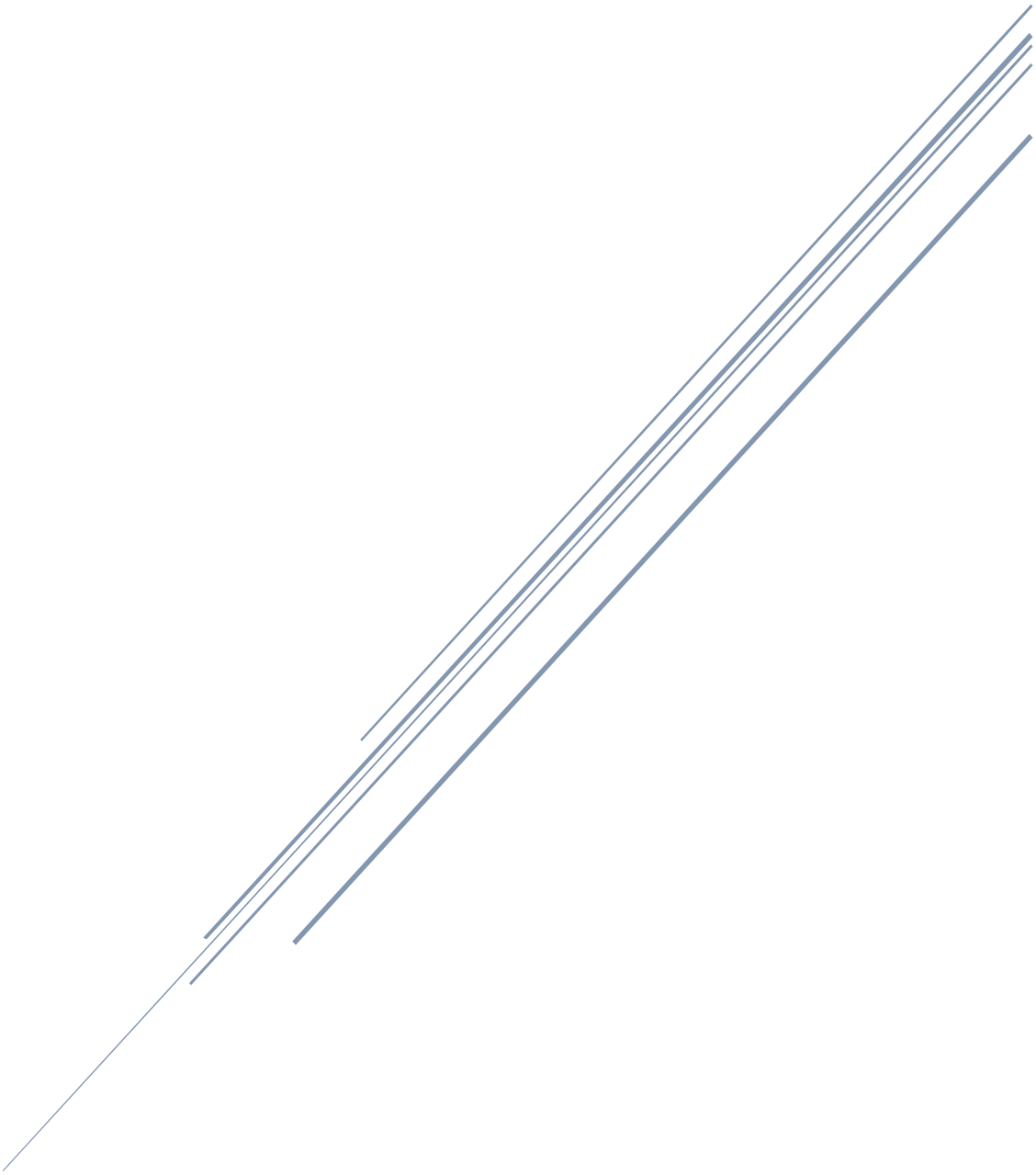
STUDENT_ID: 20549904

BUSI4370 UNUK

Analytics Specializations and Applications

Customer Analytics Coursework

2023/2024



University of Nottingham
MSc Business Analytics

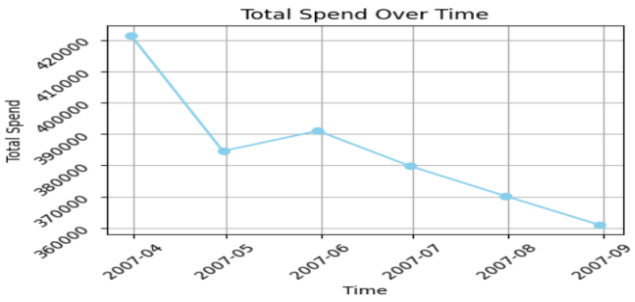
Market Segmentation Analysis of National Convenience Store Transactions

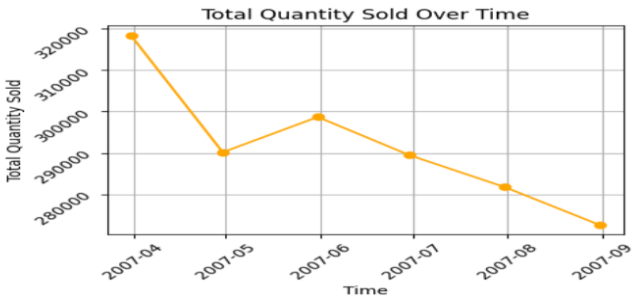
1. Executive Summary:

The National Convenience store aims to segment its customer base for effective marketing. Six distinct segments were identified through transactional data analysis . We conducted data exploration, cleaning, and used feature significance analysis alongside K-means clustering. Despite data complexity, we aggregated features and applied Principal Component Analysis (PCA) for dimensionality reduction. Our analysis uncovered unique consumer behaviors, leveraging Recency, Frequency, and Monetary (RFM) information. We recommend focusing marketing efforts on the top two groups and implementing customized campaigns for improved customer satisfaction and retention.

Data Preparation: The dataset consists of 4 files describing the transactional behavior of 3000 customers over 6 months. There are 20 unique categories and 20,466 unique products in the data. No null values in all four files.

1. **Customers file:**
   - It describes 3000 customers with 5 features- 'baskets', 'total_quantity', 'average_quantity', 'total_spend', 'average_spend'.
   - The mean total_spend is 769.412937 while the average spend per quantity is 1.682477. The average quantity bought is around 583.
   - It is observed here that- customers who spend more also tend to purchase larger quantities of items. And also those who visit more (no. of baskets tell the frequency of visits) tend to spend higher.
2. **Baskets file:**
   - The dataset encompasses 195,547 records, encompassing attributes like 'purchase_time', 'basket_quantity', 'basket_spend', and 'basket_categories'.
   - The mean number of items per purchase stands at approximately 9, with an average basket spend of £11.80. The number of categories purchased in each transaction amounts to around 4.
3. **Category spends file:** The file has 3000 records with 20 features describing how much every customer has spent on an individual category.
4. **Lineitems file:**
   - The file comprises 1,461,315 records with 'purchase_time', 'product_id', 'category', 'quantity', and 'spend' features.
   - Additionally, analysis indicates a declining trend in total spend and quantity over time, underscoring the necessity for fresh marketing strategies to revitalize customer engagement.





Negative values:

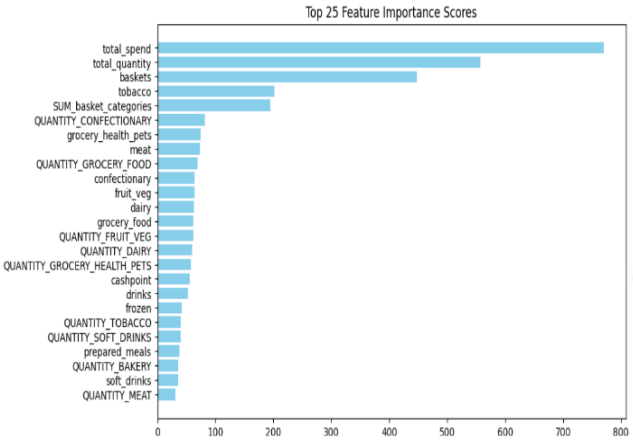
- In lineitems file, 864 entries exhibit negative spend and quantity values, within categories such as 'LOTTERY', 'GROCERY_HEALTH_PETS', 'SOFT_DRINKS', 'DRINKS', 'CONFECTIONARY', 'DELI', and 'MEAT'.
- For the baskets file, 160 occurrences of negative values in 'basket_spend' and 19 instances in 'basket_quantity' are found.
- In the category spends file, only the lottery category has a negative spending value.

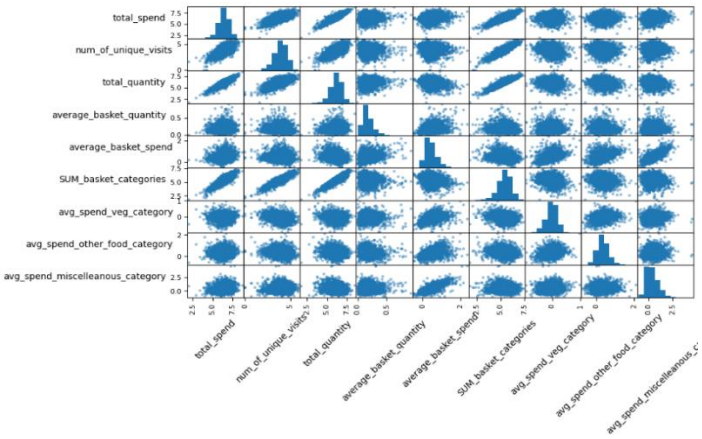
These negative values may signify refunds issued by the store to customers upon product returns, echoing a common retail practice.

Inconsistencies: There is some inconsistency in the **category spends table for the bakery column** where all the values are set to '0', however, when the spend per category is aggregated through the lineitems table, we see the values for bakery spends.

2. Feature Selection/Engineering:

All 4 files were merged by performing the required aggregation in the lineitems and baskets table. Redundant columns were removed. To discern influential features in cluster differentiation, a K-means algorithm conducted a feature importance analysis. Features were sorted based on absolute mean differences in centroids, indicating their significance. Higher absolute mean differences denote greater importance in cluster distinction.





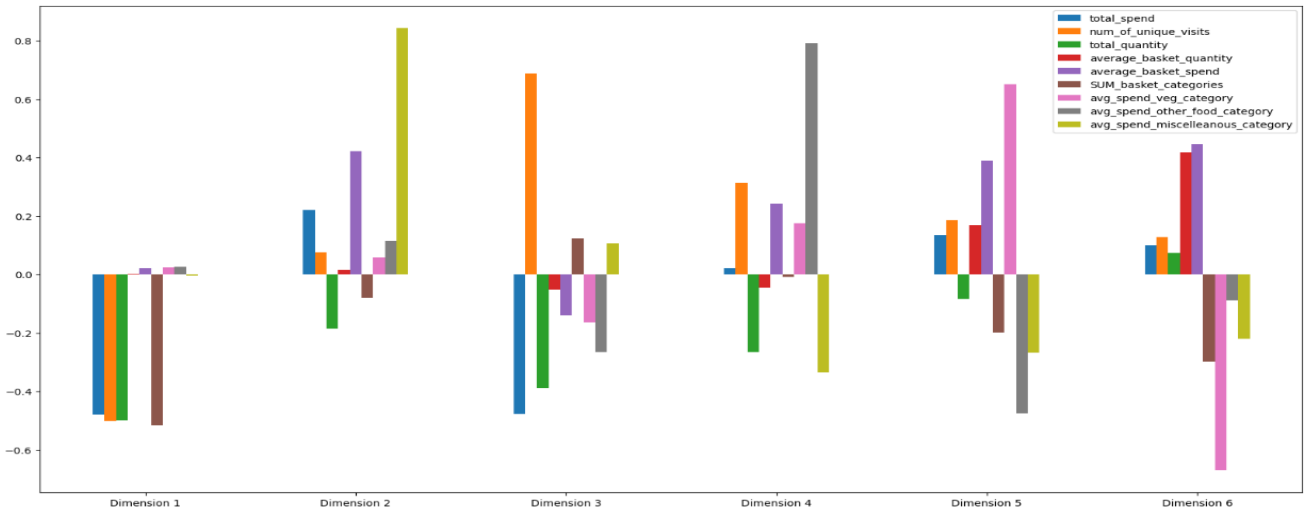
Due to the significant correlation observed among the top 25 categories, a decision was made to aggregate them as a means to address this issue effectively. We combined **'total_spend'** and **'num_of_unique_visits'** from the line items table based on spending and purchase time. Additionally, we aggregated quantities by categories, resulting in **'avg_spend_veg_category'**, **'avg_spend_other_food_category'**, and **'avg_spend_miscellaneous_category'** due to their correlation with total spend. The category_spends table wasn't used due to inconsistencies in the bakery column. From the customers table, we derived **'total_quantity', 'average_basket_quantity', 'average_basket_spend'**, and **'SUM_basket_categories'**, resulting in 9 selected features.

To address non-normal distribution, a log transform was applied since many clustering techniques, notably K-Means, rely on identifying 'globular' clusters, which benefit from appropriately distributed data.

Since there were still some features that were correlated with each other, they may skew the results thus, to solve this we have applied PCA(Principal Component Analysis) on the log-transformed data and converted these 9 features into 6 PCA components.

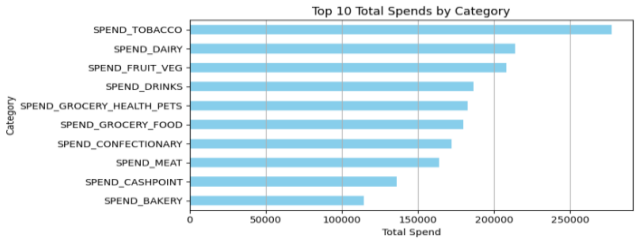
The explained variance of each component ranged from 0.65 to 0.009, totaling approximately 0.993. The **first 2 dimensions explained 83.64% of the variance**, while subsequent dimensions capture additional variance. '**Avg_spend_miscellaneous_category**' holds the highest weight (0.8428), followed by **'avg_spend_other_food_category'** (0.7915). **'Num_of_unique_visits'** and 'avg_spend_veg_category' rank closely with weights of 0.68 and 0.67, respectively. These features aggregate the top categories like 'tobacco',

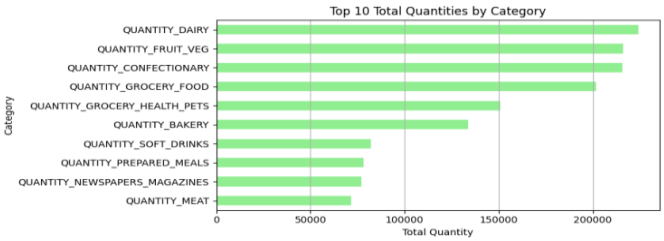
'confectionary', 'meat', 'fruit_veg', 'dairy', 'grocery_food', 'grocery_health_pets', and 'cashpoint' with their quantities.



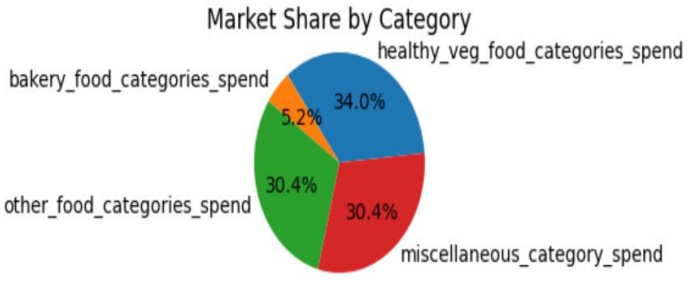
3. Customer Base Summary:

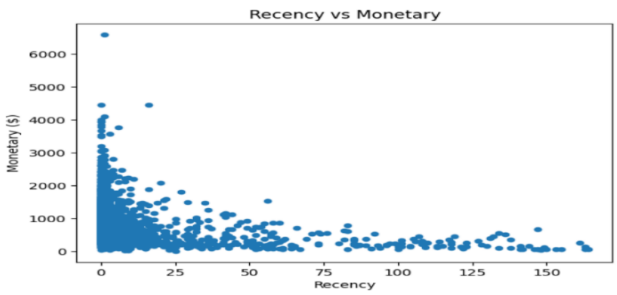
The initial analysis of the dataset reveals that customers with higher spending inclinations also have a higher propensity for acquiring larger quantities. Furthermore, people who visit the store more frequently, as measured by the number of baskets, tend to have more expenditure. Further analysis reveals that the 'Tobacco' category yields the highest revenue, surpassing £250,000, while 'Dairy' accounts for the highest sales volume. Notably, food-related categories dominate both revenue and quantity metrics.





If we explore market share by category, healthy vegetarian foods take the lead with 34%, followed by miscellaneous, which includes tobacco, confectionery, cashpoints, etc. categories. RFM analysis was performed on the data, and it was discovered that consumers with low recency between 0 and 25 days offered higher monetary value than those purchasing after a recency of 25 days.





As per the above findings, it appears that the store has positioned itself to accommodate both health-conscious clients and those looking for tobacco, confectionery, bakery, etc. items. The inclination of customers towards healthy food suggests that customers prioritize health-conscious choices and are willing to invest in products that support their well-being. Thus, there is a strong potential to capitalize on this trend even further by extending the store's variety of healthy and organic food products and aggressively marketing them.
Exploring consumer feedback channels and creating loyalty programs may further improve customer engagement and retention.

4. Segmentation Methodology

The combined file dataset comprised of 46 distinct features, including spending across different categories, quantities purchased, visit frequencies, average spend per basket, and average quantity produced. Initially, the features were carefully aggregated based on their importance and correlations, which brought the dataset finally with 9 unique features. Further, to address the challenge of high dimensionality, we applied Principal Component Analysis (PCA) for dimensionality reduction. Before conducting PCA, the features were log-transformed to ensure normal distribution, a preprocessing step essential for subsequent analysis.

Following PCA, we found that the initial 2 principal components explained approximately **83.64%** of the total variance in the dataset. This discovery led us to focus on these 2 components for further exploration, providing a more concise representation of the dataset while retaining a significant portion of its variance.

To determine the optimal number of clusters (k) for segmentation, we utilized the silhouette score metric, which measures the cohesion and separation of clusters. We systematically assessed silhouette scores for various values of k, ranging from 2 to 11. The **silhouette score peaked at the value of 0.357** at **k = 6** as compared to higher values of k, indicating favorable clustering behavior and distinct separation into 6 groups.

With the optimal k value established, we applied the KMeans clustering algorithm to segment the data into 6 distinct clusters. Further, we utilized PCA inverse to analyze the impact of each feature dimension on the formed clusters, providing valuable insights into the characteristics and behaviors of each consumer segment, and facilitating targeted insights and actionable recommendations.

Overall, our approach involved comprehensive preprocessing, thoughtful dimensionality reduction, and systematic evaluation to uncover meaningful insights from the data. This resulted in a robust segmentation strategy that captures the diverse spending behaviors of consumers.

5. Results

The segmentation technique included both clustering and RFM analysis, resulting in six separate consumer groups. This comprehensive approach provides an in-depth insight into consumer behavior within each group, guiding tailored marketing strategies to engage with diverse customer preferences and habits effectively.

RFM analysis results: Considering the Recency, Frequency, and Monetary value of the customers, they are segregated into 6 segments:

Customer Segment	Recency (days)	Frequency (visits)	Monetary Value (£)	Percentage of Customer Base	Marketing Action
VVIP Customers	0.2	131.5	1399.8	18.7%	Exclusive offers, VIP events, Personalized experiences
Platinum Customers	0.9	87.5	1011.1	10.2%	Loyalty rewards, Upgrades, Cross-selling opportunities
Loyal Customers	2.6	64.8	800.8	24.7%	Loyalty programs, Referral incentives, Surprise gifts
Need Attention Customers	5.9	43.6	567.3	22.2%	Re-engagement campaigns, Personalized recommendations
At Risk Customers	13.3	29.7	403.6	15.9%	Win-back promotions, Targeted discounts, Feedback requests
Lost Customers	46.8	16	209.6	8.4%	Reactivation campaigns, Customer surveys, Last chance offers

Considering the KMeans clustering analysis:

Segment 1: Within this segment, we have identified **321 customers** who exhibit distinctive spending patterns. On average, these customers **spend £570.03** and make approximately **44 visits** to our establishment. Their shopping baskets contain an average of 284.31 items overall, with each basket averaging £2.68 in spending across various categories. Interestingly, while they allocate £1.051 towards vegetarian food, their expenditure on other foods and miscellaneous categories amounts to approximately £1.69 and £4.579, respectively. Notably, the maximum average spend in the miscellaneous category stands at £97.2, **indicating a noteworthy inclination towards non-food items such as tobacco, newspapers, magazines, and confectionery.** Here, the total spend ranges between £238.94 and £2407.45, signaling a wide range of purchasing power. These customers would be considered as **'Need Attention'** customers.

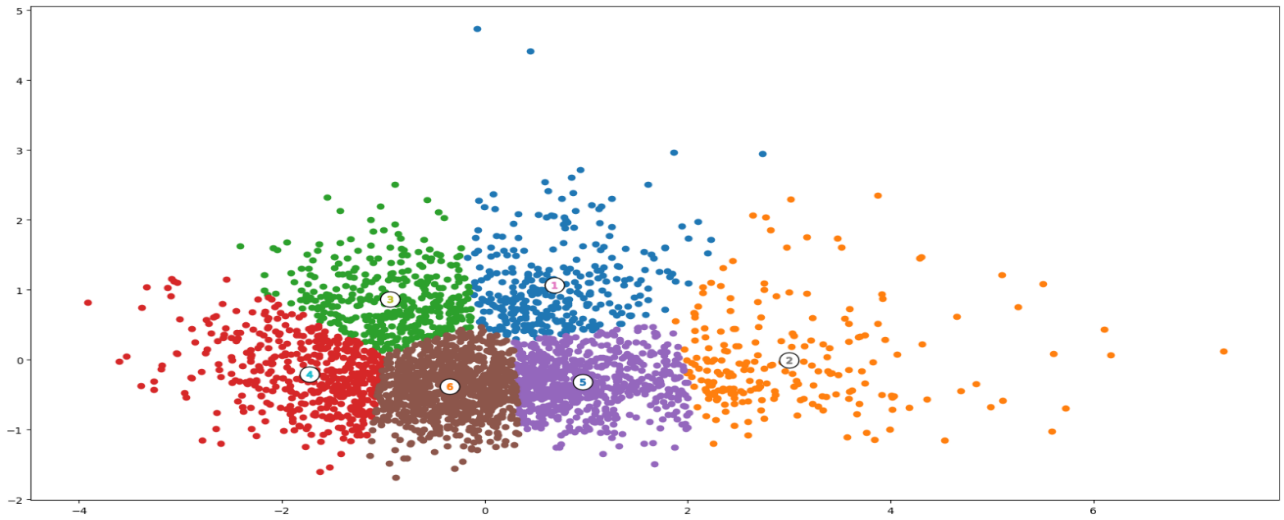
Segment 2: Comprising **214 customers**, this segment demonstrates distinct spending habits. On average, customers in this segment **spend £162.57** and make around **13 visits** to our establishment. They have purchased over 118 items in total, with an average basket spend of £1.77. Notably, they allocate £1.08 towards vegetarian food, £1.63 towards other foods, and £1.77 towards miscellaneous categories. The maximum average spend in the miscellaneous category is £11.5, while for vegetarian and other food is £2.325 and £5.87, **indicating a mixed preference for different items** among this group. Here, the total spend ranges between £7.28 and £422.18, signifying **a very low purchasing power** of customers. Considering purchasing power, these customers can be identified as **'Lost'** customers.

Segment 3: This segment encompasses **376 customers** with an **average spend of £1214.75**. These customers make approximately **93.6 visits** and have purchased a mean total of 657 items overall. Their average basket spend across various categories is £2.29, with allocations of £1.01 towards vegetarian food, £1.586 towards other foods, and £3.29 towards miscellaneous categories. The maximum average spend in the miscellaneous category reaches £13.9, **suggesting a preference for non-food items**. Another important point is the minimum total spend here is £445.75 which is higher among the earlier segments with the highest spend being £3676.46. Based on this, these customers would likely fall under the **'Platinum'** category.

Segment 4: Consisting of **484 customers**, this segment exhibits an **average spend of £1473.92**. Customers in this group make around **130 visits** and purchase over 1212.10 items in total. With an average basket spend of £1.44, they allocate £0.95 towards vegetarian food, £1.382 towards other foods, and £1.343 towards miscellaneous categories. Notably, the maximum spend in other food categories is £4.38, along with £1.5 and £3.9 in vegetarian options and miscellaneous items, respectively indicating **more preference for bakery, meat, prepared meals, and drinks**. The minimum average spend is £543.81 while the maximum spend is £6588.65 which signifies that this group has a **very strong purchasing power**. Based on all this, these customers would likely fall under the **'VVIP'** category.

Segment 5: This segment includes **675 customers** with an **average spend of £384.92**. On average, they make approximately **32.42 visits** and purchase 323.79 items overall. While spending £1.02 on vegetarian food, they allocate £1.47 towards other foods and £1.16 towards miscellaneous categories. Interestingly, the maximum spend in all the categories ranges between £1.91 - £3.85, suggesting an **equal preference for all categories**. Here, the minimum total spend is £115.9 and the maximum spend is £928.01 making it a group with **lower purchasing power**. Considering all this, we can say these customers would be under '**At risk'** customers.

Segment 6: Encompassing **930 customers**, this segment exhibits an **average spend of £710.23**. Customers in this segment make around **62.5 visits** and have purchased over 625.71 items. With an average basket spend of £1.36, they allocate £0.978 towards vegetarian food, £1.409 towards other foods, and £1.10 towards miscellaneous categories **indicating a preference for food categories than miscellaneous** as compared to other segments. Notably, the maximum spend in the other food category is £6.431, followed by miscellaneous with £2.52. The minimum total spend here is £294.38 while the maximum spend is £2211.59 indicating that there is a **wide range in purchasing power** of people. These customers can be considered as the **'Loyal'** customers.



6. Summary:

Upon analyzing the six segments derived from clustering, I recommend focusing marketing efforts on Segment 3 (Platinum Category) and Segment 6 (Loyal Customers) for the following reasons:

Segment 3 (Platinum Category):
Customers here exhibit high average spending and frequent visits, indicating strong purchasing power and loyalty. They have a strong interest in miscellaneous items like confectionery, newspapers and magazines, tobacco, lottery, cashpoint services, seasonal gifting, and practical items presenting opportunities for cross-selling and upselling initiatives. Targeting Segment 3 with personalized marketing campaigns and loyalty programs can further enhance customer satisfaction and retention.

Segment 6 (Loyal Customers):
Customers here demonstrate consistent spending behavior and a preference for food items, mostly in categories of bakery products, discount bakery items, prepared meals, drinks, soft drinks, frozen foods, meat, deli items, and world foods. With high visit frequency and strong purchasing power, they represent a valuable market segment for the company. Implementing targeted promotions, exclusive offers, and personalized communication channels can strengthen relationships with Segment 6 customers and drive repeat purchases.

Marketing Strategy Suggestions:

Segment 3	Segment 6
Personalized Loyalty Programs: Offer exclusive benefits and rewards tailored to high-spenders and frequent visitors.	**Appreciation and Recognition:** Express gratitude through personalized gestures and discounts.
Upselling and Cross-Selling: Recommend complementary products and highlight premium items in miscellaneous categories.	**Exclusive discounts:** Offer exclusive discounts and promotions on preferred food categories like bakery and prepared meals.
Personalized Communication: Utilize targeted emails and promotions based on preferences.	**Referral Programs and Exclusive Loyalty Rewards:** Introduce tiered programs with escalating benefits.

Implementing these strategies fosters strong relationships, drives repeat purchases, and encourages brand advocacy, ensuring sustainable growth and profitability.

Suggestions for Further Analysis: Conduct periodic evaluations of marketing campaigns and customer engagement metrics to measure effectiveness and identify areas for improvement.