Executive Summary:

This report consists of a market segmentation on a transactional dataset that has been provided by a national convenience store chain (4 files describing 3000 customers over 6 months). The feature engineering created useful metrics like total spend, average spend, and purchase frequency to understand customer behaviour. Customer base summary reveals average customer spends around £770 with a standard deviation of £553, indicating spending variability. Transactions average 65 with a standard deviation of 47, highlighting customer frequency variation.

Segmentation methodology involved data cleaning, feature selection, and dimensionality reduction using PCA. Silhouette score and Elbow method suggested 5-6 customer segments. The clustering algorithm used was K-means for a fast and efficient visualisation. These segments likely have distinct buying habits based on the features analysed. Each segment is then identified and elaborated upon to gain insights. There are a total of six segments that are generated. Of the six segments two segments show a potential to profit from a targeted marketing campaign.

The recommendation includes a target at the Household Essentials and Regular shopper, targeting them with loyalty programs and incentives to encourage bulk buying. The second recommendation includes focusing on Balanced Basket Shoppers as they are likely to increase their spending with targeted promotions and alerts on discounted household items.

Feature Description:

Feature engineering is an indispensable aspect of the data preprocessing pipeline, when dealing with high dimensions datasets. In such scenarios, where the number of features exceeds the number of observations, the inherent complexity poses challenges for effective analysis and model performance. This is the reason we choose certain features over others.

- **Total basket spend per customer** [1,2] represents the overall sales generated from each customer, providing insights into the revenue potential of individual customers and their spending capacity. Chosen as it reflects direct contribution to revenue generation. It takes into account customers that have less items in their basket but are of high value.
- **Average basket spend per customer** [1,2]indicates the average amount spent by customers in each transaction, this is chosen to reflect the purchasing behaviour and preferences of the customer base. It takes into account customers that have a high number of items but low purchase value.

- **Frequency** [1,2] represents how many times a customer engages with the store. Chosen solely to understand customer loyalty and tailor marketing strategies for high visiting customers.
- **Aggregate features** of Lineitems table is taken into account to understand what kind of preference a group of customers prefer and how they are distributed. The use of all 20 features in the table is taken into consideration.

The first three features are taken to understand customer behaviour and the rest are taken to get an overview of the sales in each category made by each customer.

The following table gives a detailed description of each feature and how it was engineered.

Feature	**Feature Description.**
1) Total Spend	Total basket spend per customer.
2) Average Spend	Average basket spend per customer.
3) Frequency	Count of customer engagement with the store.
4) Bakery	Aggregation of Bakery, Discount Bakery and Confectionary.
5) Meat & Dairy	Aggregation of Frozen, Deli, Meat, Prepared Meals and Dairy.
6) Drinks	Aggregation of Drinks and Soft Drinks.
7) Grocery	Aggregation of Fruit & Veg, Grocery Food, World Foods and Grocery Health Pets.
8) Household	Aggregation of Newspapers & Magazines, Practical Items, Seasonal Gifting, Tobacco and Cashpoint.

Customer Base Summary:

A customer base summary is a concise overview of the people who buy our product or use various services. It helps us understand what kind of customers we are dealing with and what makes them tick. This is used to gain more knowledge about which customer to target while marketing, development of better products and improving

customer retention.Detailed statistical insights into various attributes, including monetary value, transaction frequency, quantities purchased per product category, and basket spend.

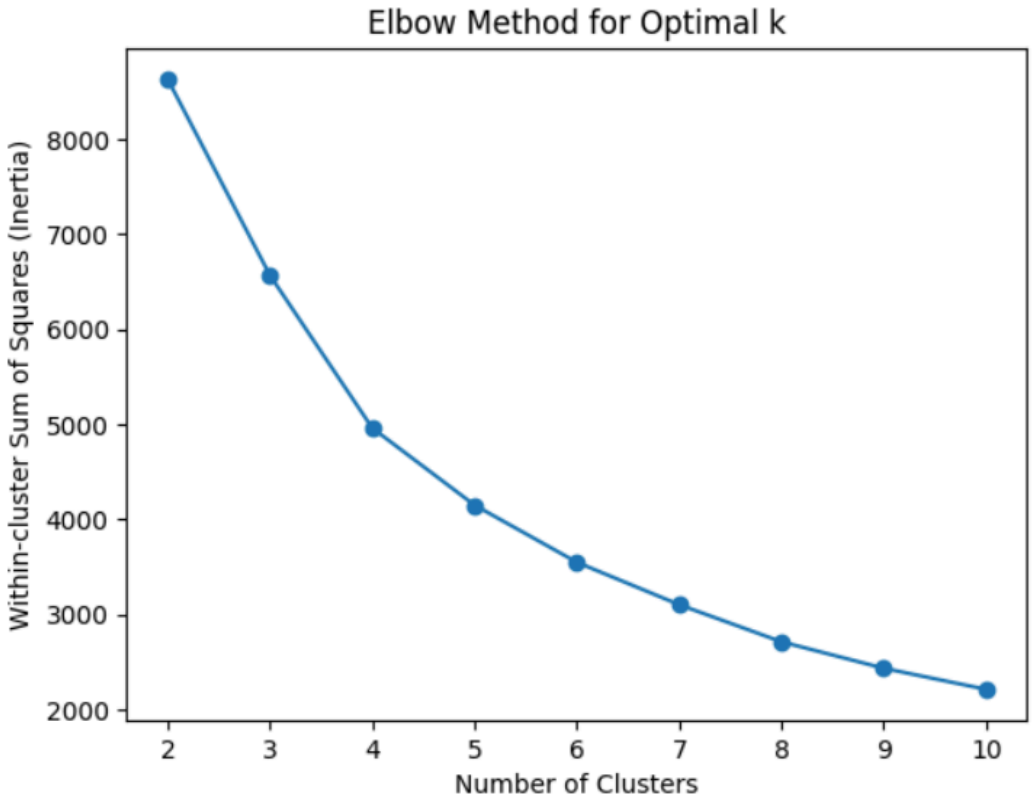
- The mean monetary value indicates an average spending behaviour of approximately £769.78 and a standard deviation of £552.98, showing the variability in spending among customers.
- The frequency of transactions averages at about 65.13, with a standard deviation of approximately 47.40, reflecting the dispersion in transaction frequencies.
- Quantity purchased across the various categories varies, with mean quantities indicating typical purchase behaviour and standard deviations showing variability around the mean. Percentiles offer insights into the distribution of spending and transaction frequencies, helping understand customer behaviour at different spending levels.
- Basket spend averages at $14.80, with a standard deviation depicting the spread of transaction amounts. These statistics help in understanding customer behaviour, informing segmentation strategies, trend analysis, and marketing initiatives for better business decision-making which help in data driven recommendations.

Segmentation Methodology:

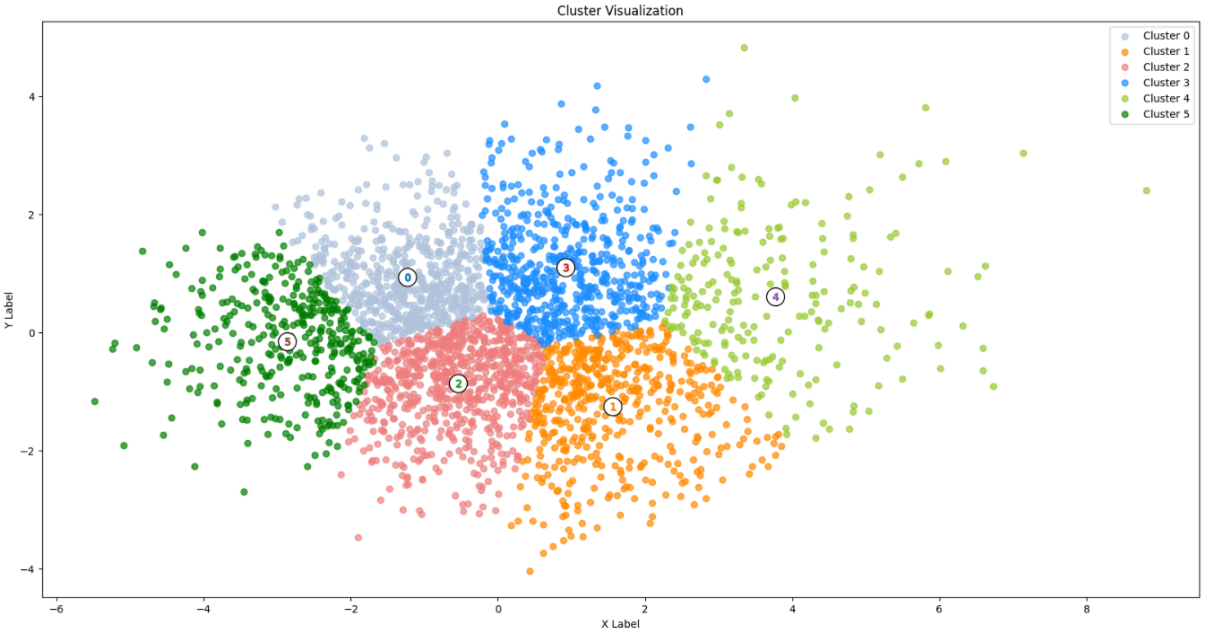
- **Initial PCA:**
  An initial PCA(Principal Component Analysis) was conducted on all the 20 category items to understand the variance each item had on the revenue. This information was then used to make an informed aggregation of the multiple groups into one category with similar or balanced opposing variance.
- **Preprocessing:**
  The data was also checked for null values and conversions were made to the table when necessary to get a suitable feature. Detailed information is given in the comments of the colab file.
- **Logging of merged dataset:**
  The selected feature dataset performed well when the data was logged. Since we are dealing with a lot of transactional data and it is used for customer behaviour, it is vital that we log the data before doing any further analysis.
- **PCA for 8 features:**
  Principal Component Analysis (PCA) is a technique used for dimensionality reduction in datasets. Its primary goal is to simplify complex datasets by transforming them into a lower-dimensional space while retaining most of the essential information. The PCA is done on the **8** engineered features as

mentioned in the Feature Engineering section. Generally we take PCA dimensions into account with 70% of the data as it gives us an overview of almost the whole dataset. While looking at the cumulative explained variance we see almost **71%** of the data being covered in **2 components**. This is used to visualise 2 PCA dimensions which are fitted to the logged data we obtained in the previous section.

- **Evaluating value of K**:
    1) **Silhouette Score:** This method is used to determine the value of K in the clustering that will be performed. While viewing the score we get a score of 0.378 for 2 clusters and a score of 0.3178 for 6 clusters. The scores after this have very little variance to the 6th cluster, thus concluding a total of 6 clusters used during the analysis.
    2) **Elbow Method:** While we did receive a value of K evaluation from the Silhouette score we will also be going for the Elbow method to get a more precise K for the clustering. In the below figure we observe that we have a strong elbow at 4 however there is still a larger variance dip at cluster number 5 and 6. Therefore, it is safe to say that we can consider 5 to 6 clusters for the final analysis.



- **Clustering:**
  The clustering algorithm used for visualisation is K-means. K-means is a simple and efficient algorithm for clustering. The value of K is defined from the Silhouette score and elbow method as 6 clusters. The figure below shows us the clusters formed with the dimensions from the PCA as 2.



Results:

Household Essentials and Regular shoppers (Segment 0): The customers in this segment show moderate high spending levels overall, but with a notably high expenditure on grocery items. They look like shoppers who are interested in Household essentials and groceries. Their frequency suggests that they are loyal customers.

Conscious Budget Shoppers (Segment 1): This segment has customers who shop less but are not the segment who has the lowest shopping and frequency value. The main focus is on Bakery and Meat & Dairy as there is a spike in these segments. They can be classified as customers who go for lower-priced groceries and drinks.

Pantry Shoppers (Segment 2): This segment appears to consist of customers who spend relatively high amounts across most categories, indicating a balanced shopping behaviour without any specific focus. However, there is a lot of focus on stocking up on pantry staples and household goods.

Balanced Basket Shoppers (Segment 3) : Customers in this segment are moderate to low shoppers. They spend moderately low amounts overall, with a relatively higher proportion spent on household items compared to other categories. Simple and clear

description of their spending across categories is seen. The basket is more balanced than the rest of the segments.

Essentials Shoppers (Segment 4): They seem to be customers that spend the least in every category. The lowest spend being on drinks and household items. Highlights their focus on core grocery needs with lower purchase frequency. They are minimalist shoppers and won't buy anything other than core basic needs.

Bulk Buyers (Segment 5): They are a High-Value All-Rounder consumers. This segment has the highest monetary value and frequency across all categories, indicating customers who spend significantly on various product categories like bakery, meat and dairy, drinks, grocery, and household items. The high spend on every category might suggest a huge family.

Summary:

Customer segmentation analysis reveals five key shopper segments. High-Value Regulars shop frequently for household essentials and groceries. Conscious Budget Shoppers focus on affordable bakery and dairy products. Pantry Shoppers buy a balanced amount across categories, emphasising pantry staples. Balanced Basket Shoppers are moderate consumers with a preference for household items. Essential shoppers are minimalist shoppers that prefer to focus on core items and Bulk Buyers spend the most across all categories, suggesting larger families.

Recommendations:

1) Household Essentials and Regular Shoppers:

The focus on this segment should be raised as they are potential high spending customers. Taking into account their frequency and their spend it is evident that they are loyal customers. Offering incentive on certain products like household items and grocery items can convert them into potential bulk buying customers. A loyalty card can be issued to them with a ranking of gold, silver or platinum and as the customer advances in the rank the incentives kick in. This is also an easy way for the store to track the customers from their end.

2) Balanced Basket Shoppers: While generating revenue is the main target of any organisation. More emphasis on the Balanced Basket Shoppers is needed. This segment shows a customer base which shops in every category. Their basket has a moderate size which can increase if given the right incentives. They should be allowed to collect points each time they shop. This will increase their frequency of visiting. Since they also have high consumption of household items we must send them alerts by sms or via email when those products drop in prices. This will make them engage with the store.

Reference list

[1] Anish Nair (2017). *RFM Analysis For Successful Customer Segmentation - Putler*. [online] Putler. Available at: https://www.putler.com/rfm-analysis/.

[2] Takle, A. (2017). *Segmentation in Python to find your best Customers*. [online] Medium. Available at: https://medium.com/@takleakshar/segmentation-in-python-to-find-your-best-customers-9c407c59f656.

PCA results of the final 8 features: