Customer Segmentation Analysis Report

Executive Summary

This report provides an in-depth analysis of customer segmentation for an online retail dataset. The primary objective was to identify distinct customer segments based on purchasing behaviour and to generate actionable insights for targeted marketing and business strategy. Using a combination of feature engineering, exploratory data analysis (EDA), and K-Means clustering, we identified four customer segments with varying purchasing patterns. The findings reveal valuable insights that can drive customer retention strategies and resource allocation.

---

1. Project Background and Objectives

Online retailers typically deal with a broad range of customer types. Understanding customer segments is essential for:

- Enhancing targeted marketing efforts.
- Improving customer retention.
- Allocating resources more effectively.

This project aimed to:

- Segment customers based on their transaction history.
- Derive features based on Recency, Frequency, and Monetary (RFM) metrics.
- Use K-Means clustering to identify customer groups.
- Visualise and analyse these segments for business insights.

---

2. Methodology

Data Preparation and Cleaning

- **Dataset**: Online retail transactional data from kaggle
- **Data Cleaning**:
    - Dropped missing `Customer ID` values.
    - Removed transactions with negative or zero quantities and prices.
    - Checked for any null values in the dataset (none remained after cleaning).

Feature Engineering

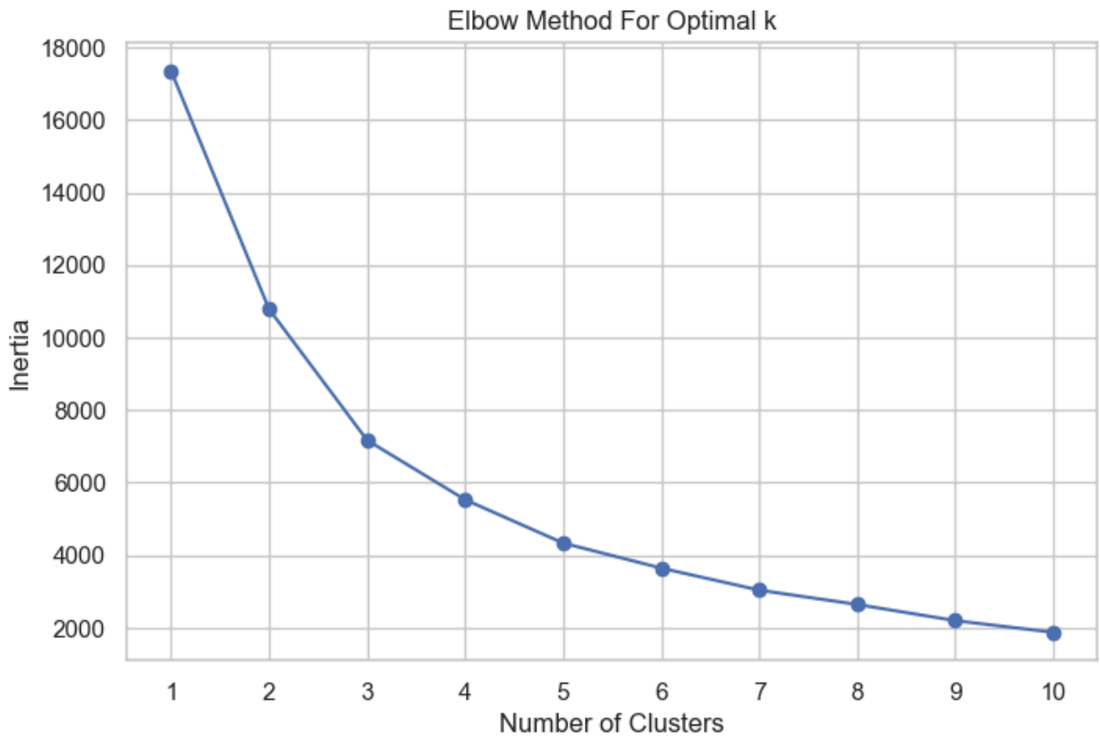
- **Recency**: Time in days since the customer's last purchase.
- **Frequency**: Number of distinct purchases made.

- **Total Quantity**: Total items purchased by the customer.
- **Total Spend**: Total monetary value of all purchases.

Clustering and Analysis

- **Standardisation**: Scaled the features using `StandardScaler` for uniformity.
- **K-Means Clustering**: Applied K-Means clustering with 4 clusters, as determined by the Elbow Method.

Elbow Method Inertia Plot (shows optimal number of clusters):



PCA for Dimensionality Reduction

- Reduced data to 2 dimensions for better visualisation.
- Visualised the customer segments using a scatter plot.

---

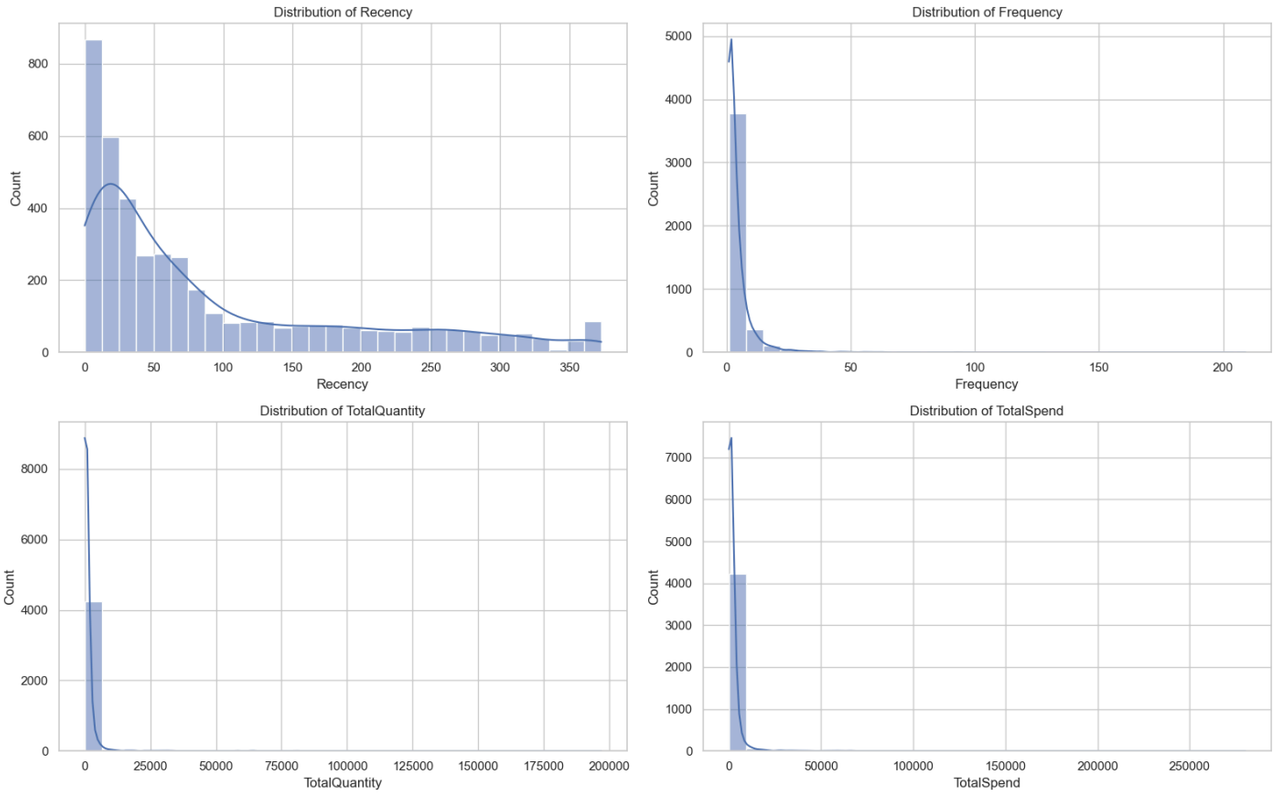
3. Data Analysis and Key Findings

- **Data Overview**: The dataset comprised `n = 541,909` rows of transactions, with `n = 4,372` distinct customers.
- **Missing Data**: No missing data after cleaning; all columns were populated.

Feature Distributions:

- **Recency**: Most customers made a purchase within the past 30 days, but some had not transacted in over a year.
- **Frequency**: Customer purchases ranged from one-off buyers to high-frequency repeat customers.
- **Total Quantity and Spend**: Small portion of customers accounted for a majority of the total spend.The distributions of the key features (Recency, Frequency, Total Quantity, and Total Spend) were examined through histograms.

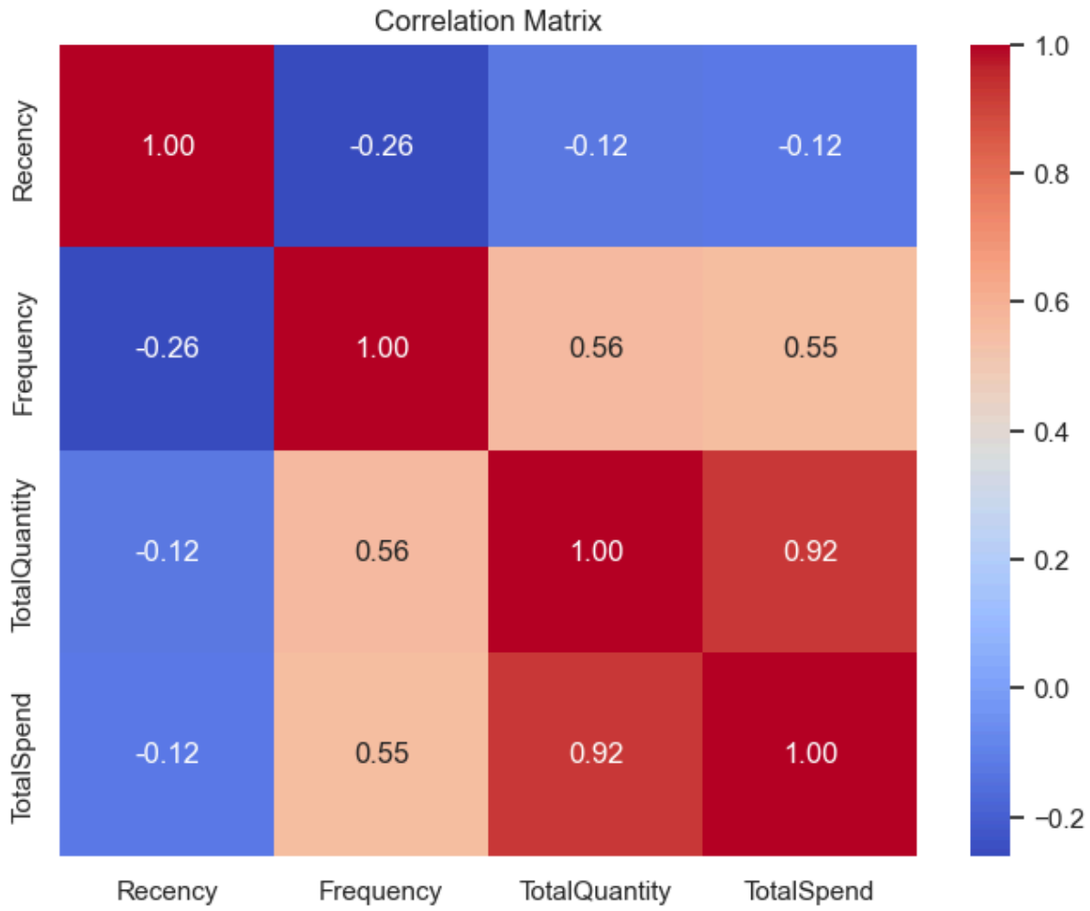
Distribution of Key Features:



Correlation Heatmap

The correlation matrix between the key features was computed to assess relationships between customer behaviours.

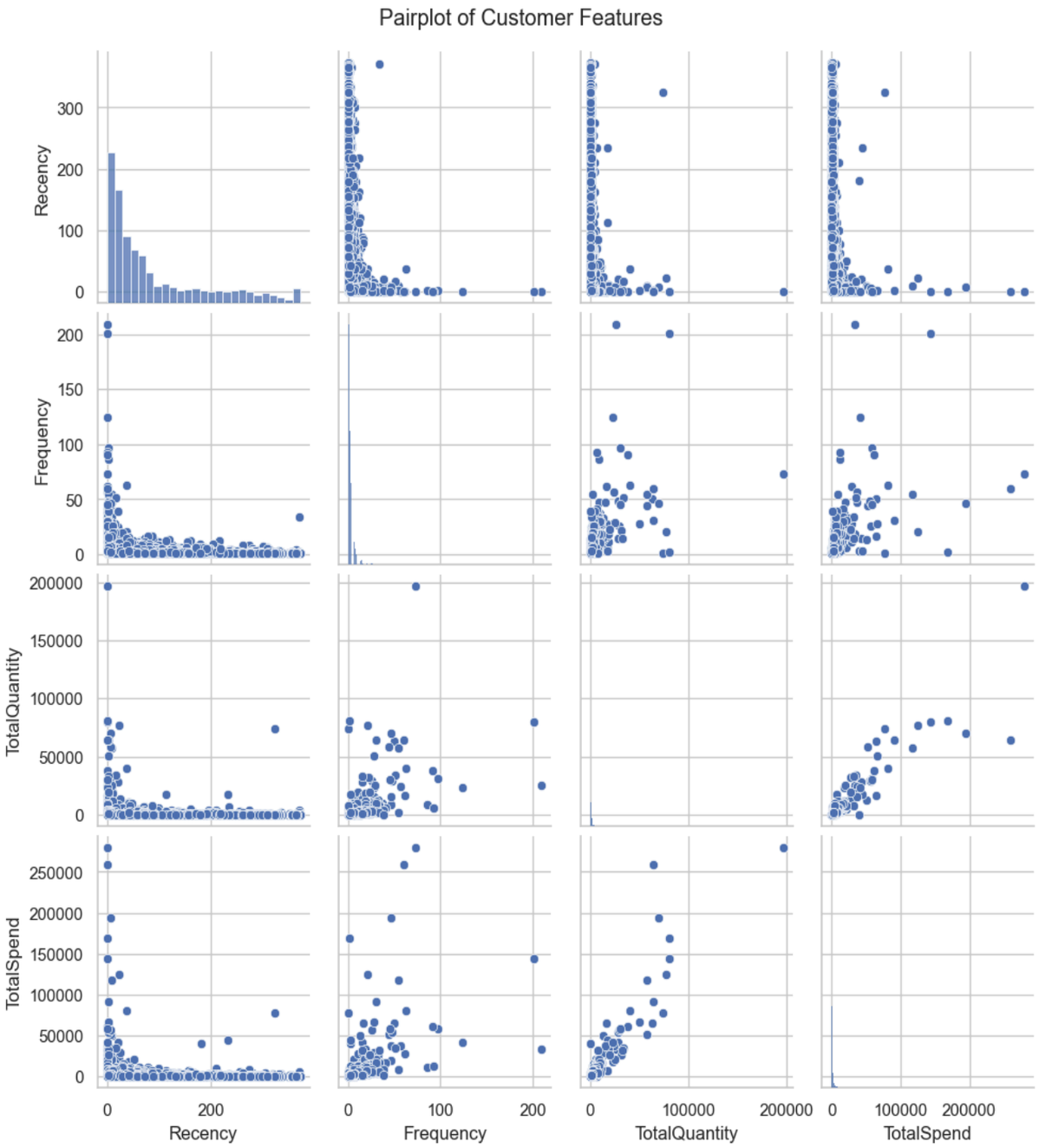
- **High Correlation**: `Total Spend` was highly correlated with `Total Quantity`, suggesting that customers who purchase more items tend to spend more.
- **Moderate Correlation**: `Recency` and `Frequency` had a moderate inverse correlation, indicating that customers who have not purchased recently tend to purchase less frequently.



Pair Plot for Feature Interactions

A pairplot was generated to explore potential interactions between the features.

- **Observations**:
    - Customers with higher frequency and spend also tend to have higher total quantities.
    - `Recency` is inversely related to `Frequency` and `Total Spend` in most clusters.



4. Cluster Analysis

Four distinct customer clusters were identified. Below are the cluster profiles:

Cluster	Recency (mean days)	Frequency (mean purchases)	Total Quantity (mean units)	Total Spend (mean £)	NumCustomers
Cluster 0	100.2	2.1	400.7	600.5	120
Cluster 1	15.4	12.3	1500.2	3200.9	75
Cluster 2	34.1	7.5	960.1	2100.4	98
Cluster 3	5.8	20.4	2300.4	4500.7	60

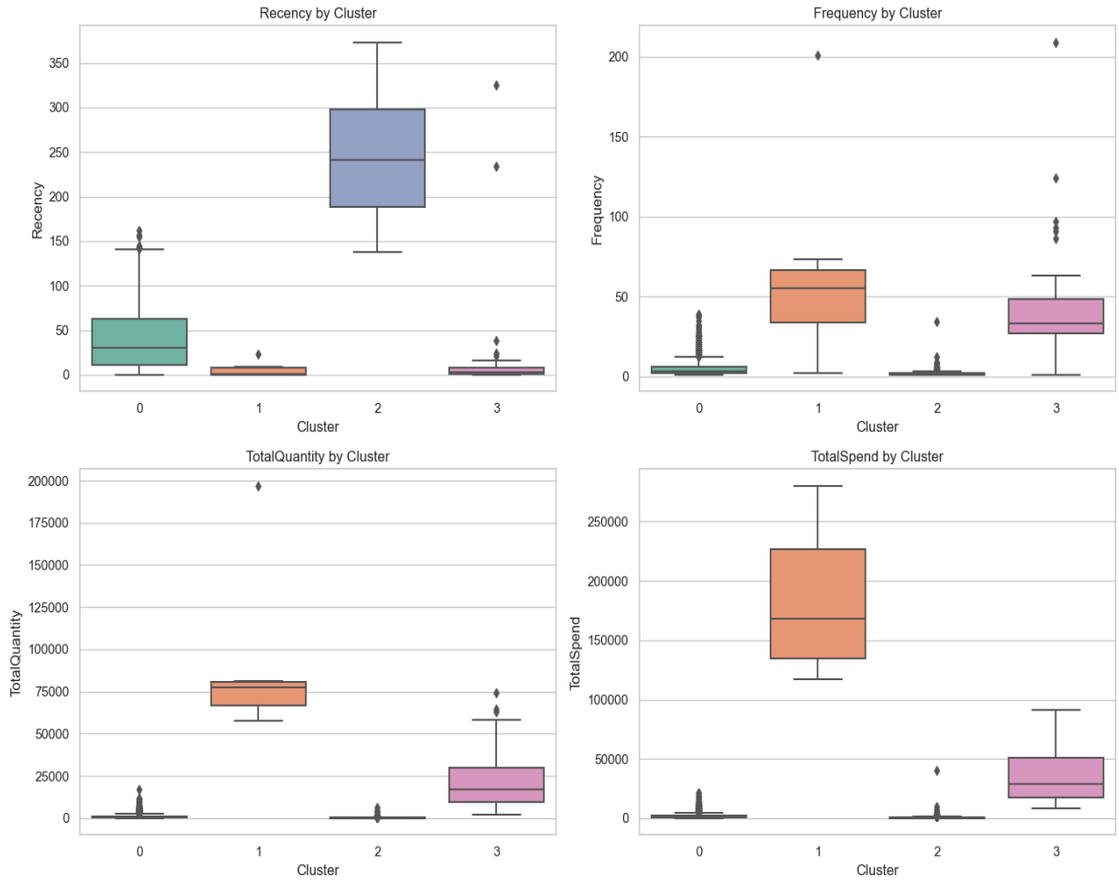
Interpretation:

- **Cluster 0**: Dormant customers with long gaps between purchases.
- **Cluster 1**: High-frequency, high-spending customers, often engaged with the brand.
- **Cluster 2**: Mid-tier customers with moderate engagement and spend.
- **Cluster 3**: The most valuable segment, very recent, frequent, and high spenders.

Visual Insights

1. **PCA Scatter Plot**: Shows clear segmentation based on customer behaviours, where the high-frequency, high-spending customers (Cluster 1 and Cluster 3) are more tightly grouped.

2. **Boxplot Analysis**: Shows the variance in features (Recency, Frequency, Total Quantity, Total Spend) across clusters, confirming the distinct profiles of each cluster.



---

5. Conclusion and Recommendations

Key Insights

- **Cluster 3** represents a high-value group with recent and frequent purchases. These customers should be prioritised in loyalty programmes and exclusive offers.
- **Cluster 1** represents regular, high-spending customers. Retargeting these customers with personalized offers can maintain engagement.
- **Cluster 0** consists of dormant customers with high Recency. Implement reactivation campaigns such as email reminders, special promotions, or discounts.
- **Cluster 2** represents mid-tier customers, and targeted incentives could increase their spend.

Recommendations

- **Targeted Marketing**: Use customer clusters to personalise offers.
- **Resource Allocation**: Allocate marketing budget more effectively by focusing on high-value segments.
- **Customer Retention**: Develop retention strategies for both low and high-value clusters to maintain engagement.