University of Strathclyde Glasgow

CS989: BIG DATA FUNDAMENTALS

A Study on E-commerce Customer Segmentation

Registration No. 202352860

Contents

List of Figures and Tables

Chapter-1 Introduction

The rise of e-commerce has fundamentally transformed business operations and customer engagement, offering businesses a unique opportunity to extract valuable insights from the substantial amounts of data generated through online transactions. Customer segmentation, a crucial aspect of data analysis, enables businesses to tailor their marketing strategies, enhance customer satisfaction, and optimize overall operations.

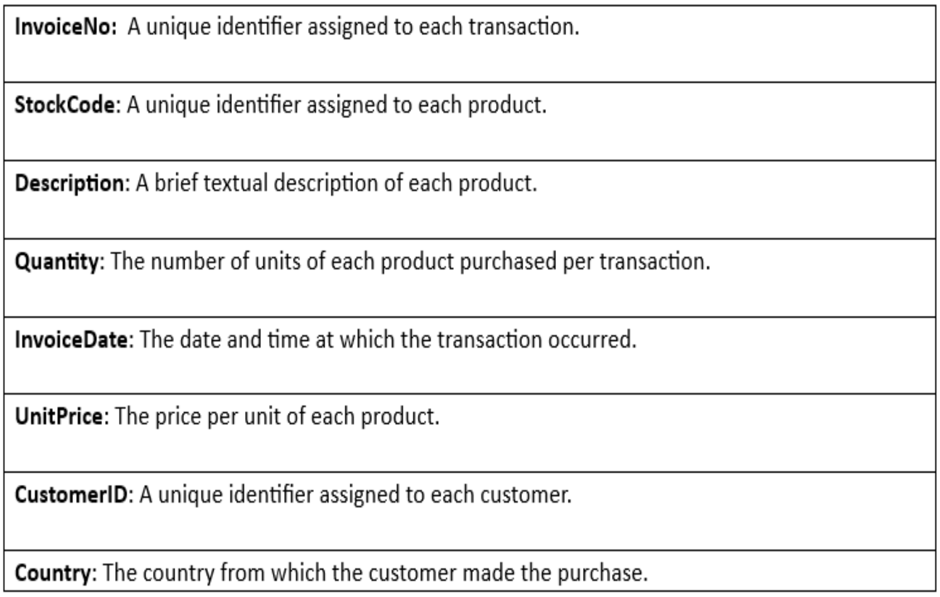
This project aims to analyse customer purchase behaviours to create distinct segments for an e-commerce platform. By identifying and understanding these segments, businesses can develop targeted marketing strategies that address specific customer needs and preferences. This approach enhances customer engagement, drives higher sales, and fosters greater loyalty, ultimately leading to improved business efficiency and profitability.

Chapter-2 Data Description and Processing

2.1 Dataset Information

The dataset used for the analysis is the Online Retail Dataset from Kaggle. This dataset comprises transactional data from a UK-based online retail store, encompassing various attributes that describe each transaction.

The dataset contains the following key attributes:



InvoiceNo: A unique identifier assigned to each transaction.

StockCode: A unique identifier assigned to each product.

Description: A brief textual description of each product.

Quantity: The number of units of each product purchased per transaction.

InvoiceDate: The date and time at which the transaction occurred.

UnitPrice: The price per unit of each product.

CustomerID: A unique identifier assigned to each customer.

Country: The country from which the customer made the purchase.

The dataset consists of 541909 records and 8 attributes, providing a comprehensive overview of customer transactions and their associated details.

2.2 Data Cleaning

To ensure the data's suitability for analysis, several preprocessing steps were undertaken.

2.2.1 Handling Missing Values:

The first step in data cleaning is to identify and handle missing values.

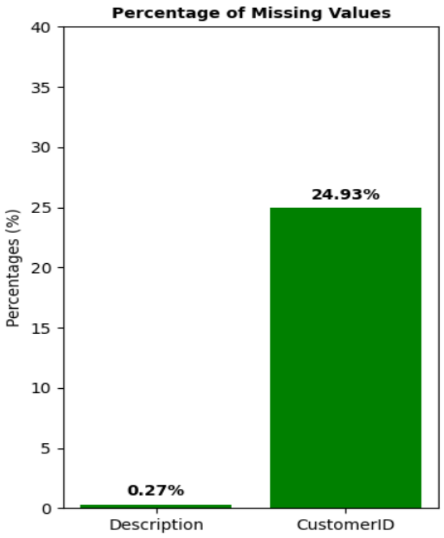


Figure 1: Percentage of Missing Values

The above bar chart(Figure1) shows that 24.93% of CustomerID values are missing, while only 0.27% of Description values are missing. By identifying rows with missing values in the CustomerID or Description columns, we further refined the dataset. By removing these rows, we aim to construct a cleaner and more reliable dataset, which is essential for achieving accurate clustering and creating an effective recommendation system.

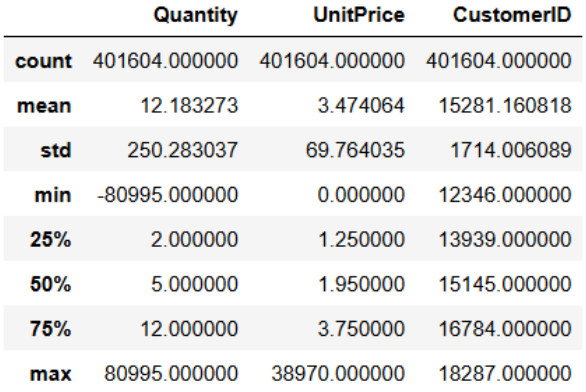
2.2.2 Removing Duplicates:

Duplicate entries can skew the analysis and lead to incorrect insights. Therefore, it was essential to identify and remove any duplicate records in the dataset. The dataset contained 5,225 duplicate rows. These duplicates were removed to ensure the accuracy of the analysis. After removing duplicates and null values 401604 is new number of rows available in dataset for analysis.

Chapter-3 Exploratory Data Analysis (EDA)

EDA is essential for ensuring data quality, understanding customer behaviour, selecting appropriate models, and improving model performance. It lays the foundation for effective customer segmentation, enabling businesses to implement targeted marketing strategies and enhance customer satisfaction.

3.1 Overview of Dataset



	Quantity	UnitPrice	CustomerID
count	401604.000000	401604.000000	401604.000000
mean	12.183273	3.474064	15281.160818
std	250.283037	69.764035	1714.006089
min	-80995.000000	0.000000	12346.000000
25%	2.000000	1.250000	13939.000000
50%	5.000000	1.950000	15145.000000
75%	12.000000	3.750000	16784.000000
max	80995.000000	38970.000000	18287.000000

Figure 2: Overview of Dataset

The summary statistics table(Figure2) provides an overview of the Quantity, UnitPrice, and CustomerID columns in the dataset.

All three columns have 401604 non-missing values.

Quantity: The wide range and high standard deviation indicate variability in the number of items purchased per transaction.

UnitPrice: The significant variation in unit prices, as indicated by the high standard deviation and range, reflects a diverse product catalogue.

CustomerID: The range of CustomerIDs shows the extent of unique customers in the dataset.

3.2 Distribution of Customer ID's

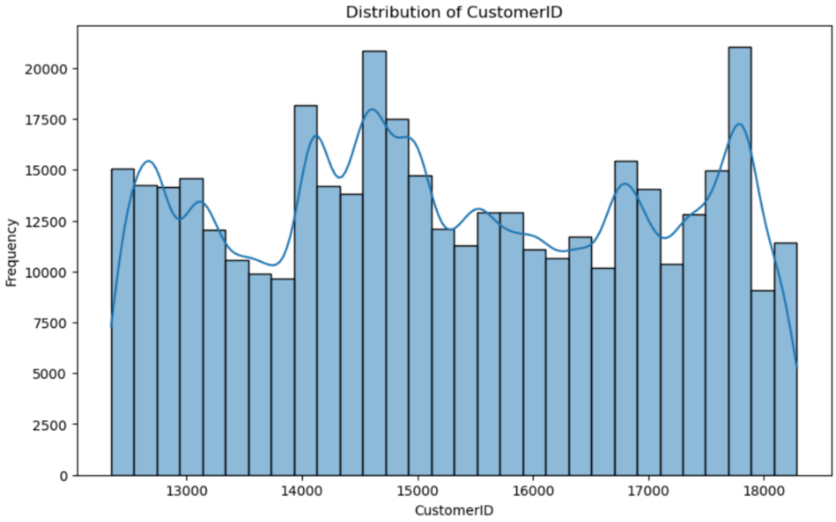


Figure 3 Distribution of CustomerID

Figure3 i.e., histogram shows the distribution of CustomerID values in the dataset. The frequency of transactions for each CustomerID is relatively uniform, with no significant skewness or extreme peaks, indicating a diverse customer base. This uniform distribution suggests that customer interactions are well-spread across the dataset, which is beneficial for performing reliable segmentation analysis.

3.3 Top 10 Most Frequent Stock Codes

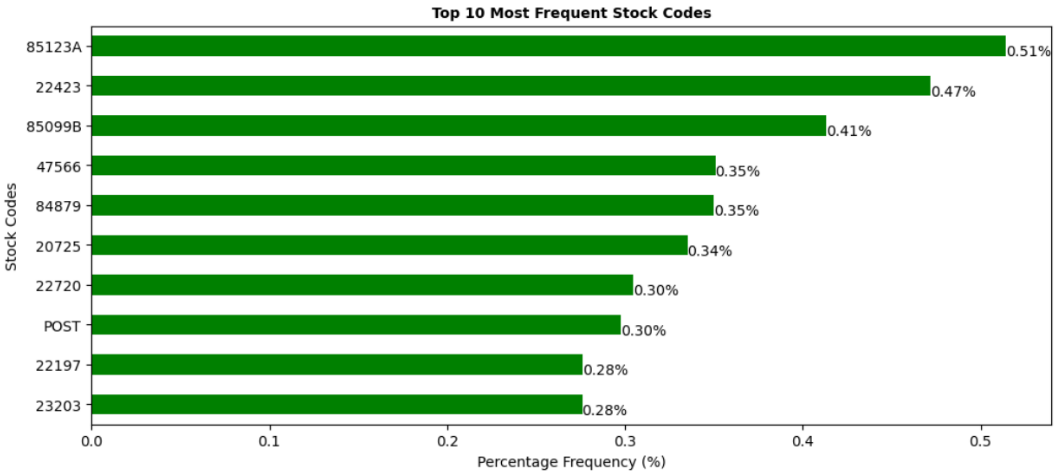


Figure 4 Top 10 Most Frequent Stock Codes

The bar chart(Figure4) displays the top 10 most frequent stock codes in the dataset, representing the percentage frequency of each stock code among all transactions. Stock code 85123A is the most frequent, accounting for 0.51% of all transactions. This chart highlights the most popular products in the dataset, which can be useful for inventory management and targeted marketing strategies.

3.4 Top 30 Most Frequent Descriptions

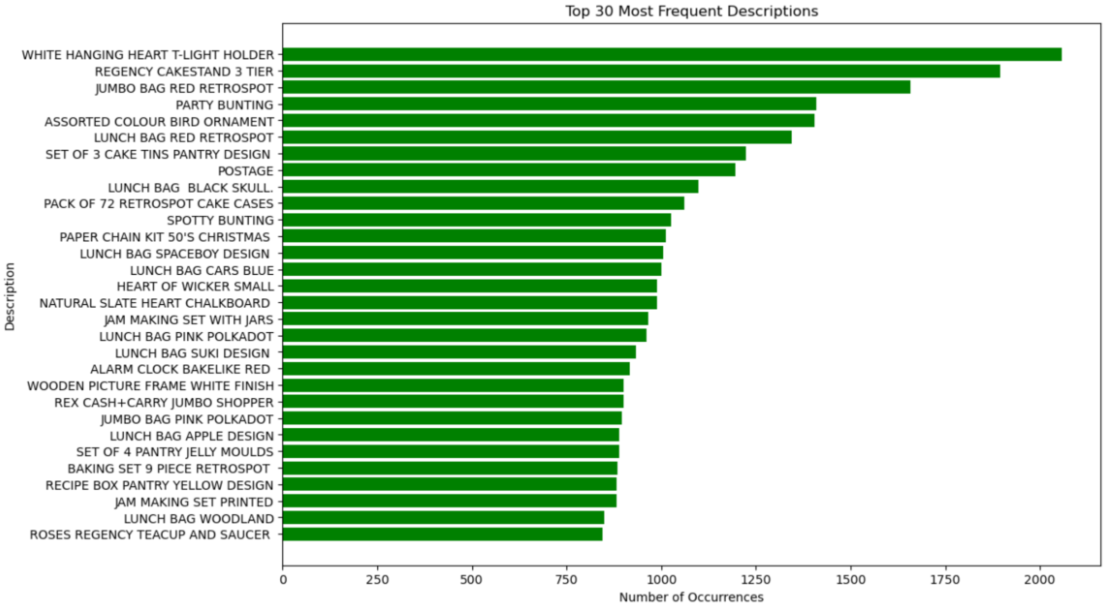


Figure 5 Top 30 Most Frequent Descriptions

The bar chart(Figure5) displays the top 30 most frequent product descriptions in the dataset, ranked by the number of occurrences. The most frequent item, "WHITE HANGING HEART T-LIGHT HOLDER," appears approximately 2,000 times. Other popular items include various types of bags, ornaments, and household goods. This chart highlights the products that are most commonly purchased, which can inform inventory management and marketing strategies by focusing on these high-demand items.

Chapter-4 Feature Engineering

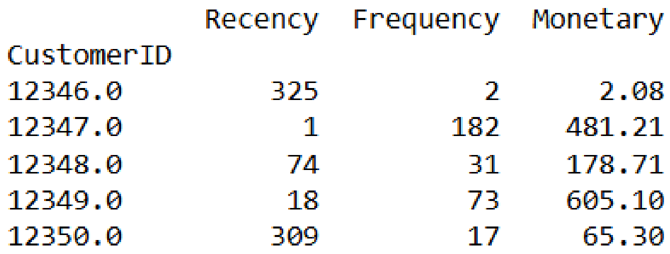
RFM analysis is a powerful method used for analysing customer value and segmenting the customer base. RFM stands for:

Recency (R): Measures how recently a customer made a purchase

Frequency (F): Measures how often a customer makes a purchase within a specific period.

Monetary (M): Measures the total amount of money a customer has spent over a specific period.

Together, these metrics provide crucial insights into customer behaviour, helping to personalize marketing strategies and enhance customer retention.



```
            Recency  Frequency  Monetary
CustomerID
12346.0         325          2      2.08
12347.0           1        182    481.21
12348.0          74         31    178.71
12349.0          18         73    605.10
12350.0         309         17     65.30
```

Figure 6 RFM Metric

Figure6 shows RFM metrics for five customers. Customer 12346.0 has low engagement (325 days since last purchase), low activity (2 purchases), and minimal expenditure (2.08), indicating low value. Customer 12347.0 has very recent engagement (1 day), high activity (182 purchases), and significant expenditure (481.21), marking them as high value. Customer 12348.0 shows moderate engagement (74 days), activity (31 purchases), and expenditure (178.71), indicating potential risk. Customer 12349.0 has recent engagement (18 days), high activity (73 purchases), and high expenditure (605.10), identifying them as high value. Customer 12350.0 shows low engagement (309 days), moderate activity (17 purchases), and low expenditure (65.30), suggesting a risk of churn. These metrics aid in segmenting customers for targeted marketing, focusing on retaining high-value customers and re-engaging those at risk.

4.1 Distribution of Recency

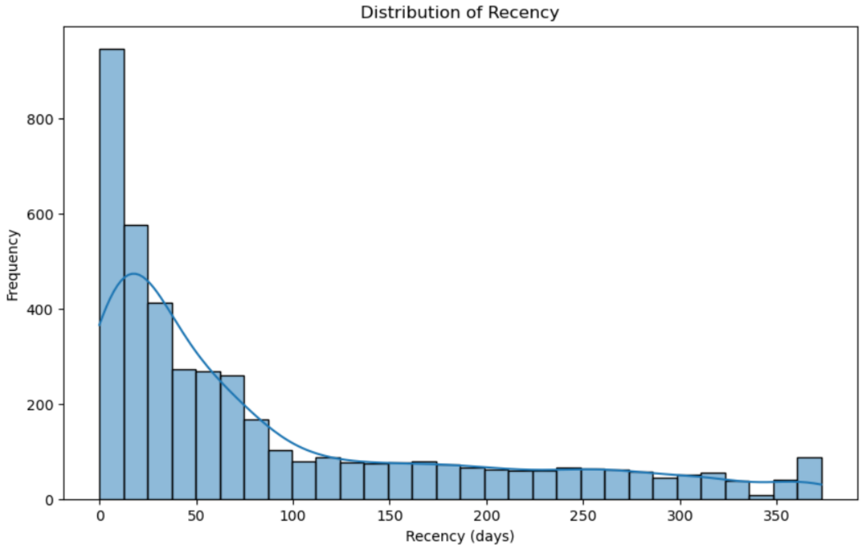


Figure 7 Distribution of Recency

The histogram(Figure 7) shows that most customers made purchases recently, within the last 50 days. The number of purchases declines steadily as the recency increases, indicating fewer customers with longer gaps since their last purchase. This suggests high recent engagement but highlights the need to re-engage less active customers.

4.2 Distribution of Frequency

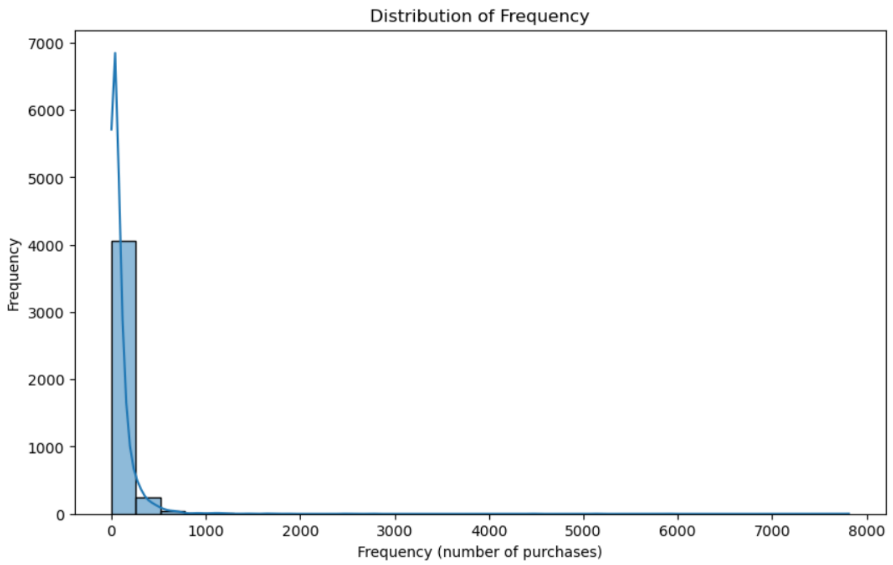


Figure 8 Distribution of Frequency

The histogram(Figure 8) shows the distribution of the Frequency metric. Most customers have made very few purchases, with a sharp decline in frequency as the number of purchases increases. This suggests that a small number of customers are highly active, while the majority make purchases infrequently.

4.3 Distribution of Monetary

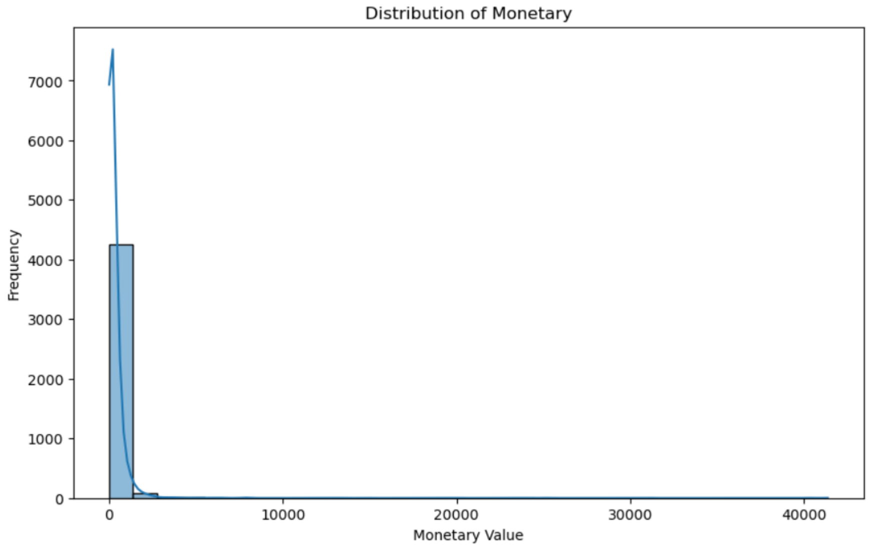


Figure 9 Distribution of Monetary

The histogram(Figure 9) shows the distribution of the Monetary metric. Most customers have spent relatively small amounts, with a sharp decline in frequency as the monetary value increases. This indicates that a small number of customers contribute significantly to total revenue, while the majority spend modest amounts.

4.4 Recency vs Frequency

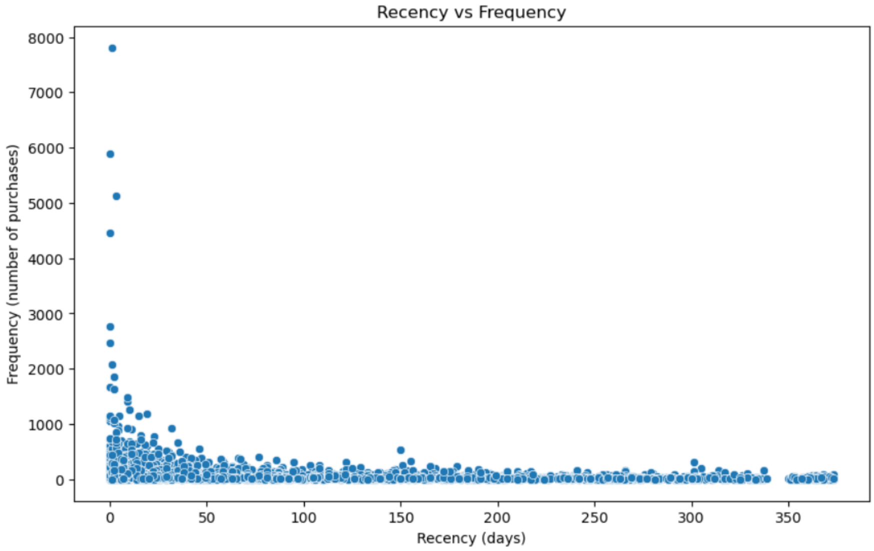


Figure 10 Recency VS Frequency

The scatter plot(Figure 10) shows the relationship between Recency and Frequency. Most customers cluster in the lower left corner, indicating they have made frequent purchases recently. As recency increases, the frequency of purchases drops significantly, with very few customers making frequent purchases after a long time since their last purchase. This pattern highlights that frequent buyers are typically recent buyers, while those who have not purchased in a long time tend to buy less frequently.

4.5 Recency vs Monetary

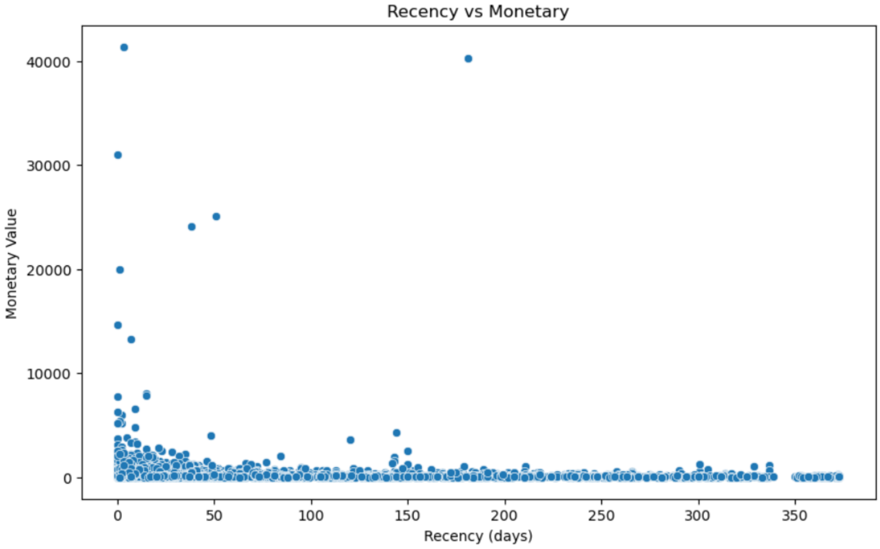


Figure 11 Recency VS Monetary

The scatter plot(Figure 11) shows the relationship between Recency and Monetary Value. Most customers cluster near the origin, indicating low recency and low monetary values. However, a few high-spending customers are visible at the top left, showing significant spending with recent purchases. As recency increases, the monetary value generally decreases, indicating that customers who have not purchased recently tend to spend less. This highlights the importance of recent high spenders to the business.

4.6 Frequency vs Monetary

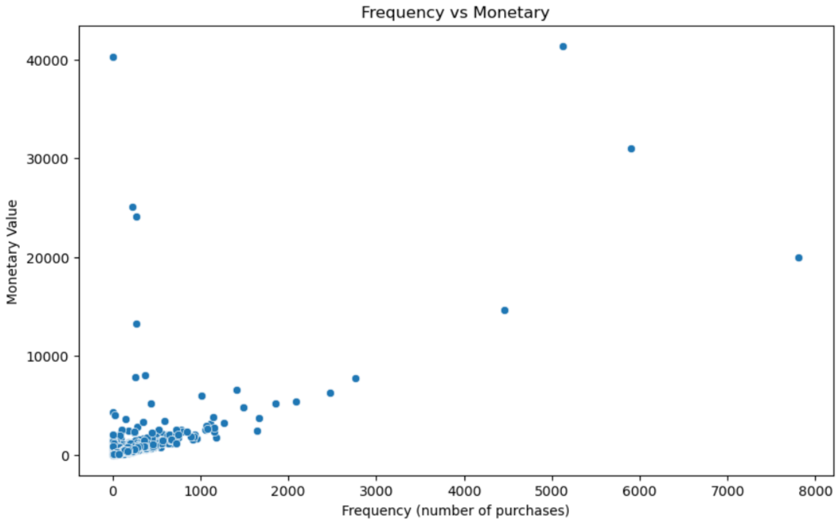


Figure 12 Frequency VS Monetary

The scatter plot(Figure 12) depicts the relationship between Frequency and Monetary Value. Most customers are clustered at the lower end of both axes, indicating low frequency and low spending. However, a few customers with high frequency exhibit substantial monetary values, suggesting that frequent buyers tend to spend more overall. This positive correlation highlights the importance of frequent purchasers as key contributors to revenue.

4.7 Correlation of RFM Features

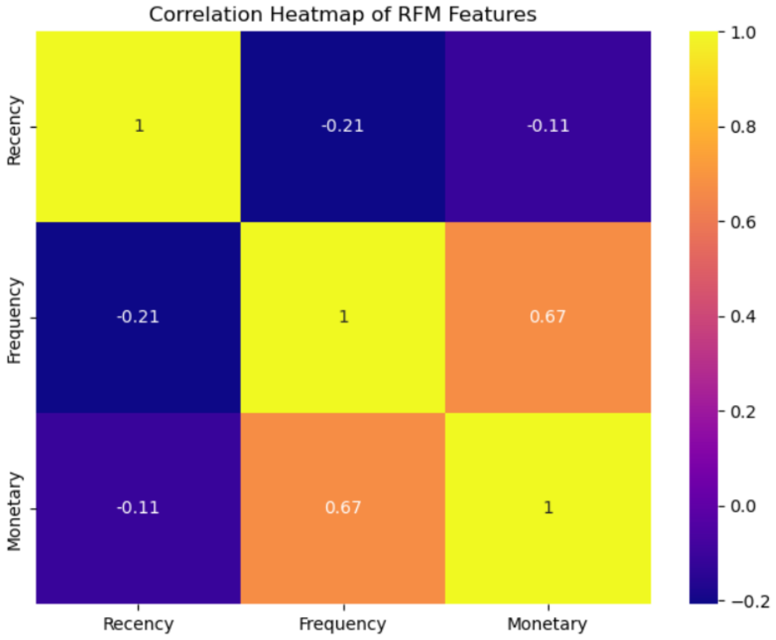


Figure 13 Correlation Heatmap of RFM Features

The correlation heatmap(Figure 13) shows relationships between RFM features. A negative correlation between Recency and Frequency (-0.21) indicates that frequent buyers tend to purchase more recently. The weak negative correlation between Recency and Monetary (-0.11) suggests recent purchasers spend slightly more. A strong positive correlation between Frequency and Monetary (0.67) shows frequent buyers also spend more overall. These insights help in effective customer segmentation and targeted marketing strategies.

Chapter-5 Unsupervised Learning

Unsupervised learning is a type of machine learning where the algorithm is trained on unlabelled data, meaning the data has no predefined labels or categories. The goal is to identify hidden patterns or intrinsic structures within the data. In customer segmentation, it can group customers with similar behaviours without prior knowledge of their categories, helping to tailor marketing strategies and improve customer insights.

5.1 K-Means Clustering

K-means clustering is an unsupervised learning algorithm used to partition a dataset into K distinct, non-overlapping subsets or clusters. The algorithm works by initializing K centroids, then iteratively assigning each data point to the nearest centroid and updating the centroids based on the mean of the assigned points. This process continues until the centroids stabilize, minimizing the variance within each cluster.

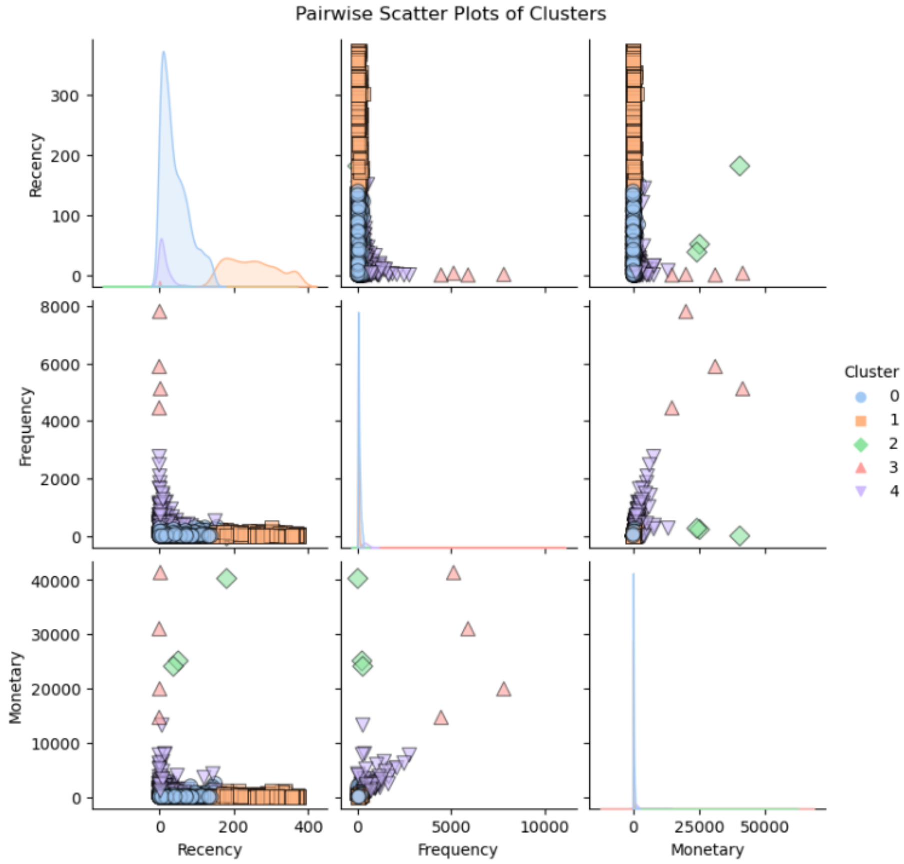


Figure 14 Pairwise Scatter Plots of Clusters

Figure14 illustrate the relationships between Recency, Frequency, and Monetary values for the five clusters identified through K-means clustering. Cluster 0 (blue circles) consists of customers with low recency, frequency, and monetary values, indicating recent but infrequent and low-value purchases. Cluster 1 (orange squares) also has low monetary values but slightly higher frequency than Cluster 0. Cluster 2 (green diamonds) represents moderate recency and frequency with high monetary values, indicating valuable but less frequent buyers. Cluster 3 (red triangles) includes high-frequency, high-spending customers, showing very high engagement and value. Cluster 4 (purple triangles) features frequent buyers with high monetary values but varying recency, suggesting valuable yet not always recent buyers. These visualizations help understand the distinct characteristics of each segment, supporting targeted marketing strategies to enhance engagement and retention.

5.2 Hierarchical Clustering

Hierarchical clustering is an unsupervised learning method that builds a hierarchy of clusters. It starts by treating each data point as a single cluster and then successively merges pairs of clusters until all points are contained in one cluster.

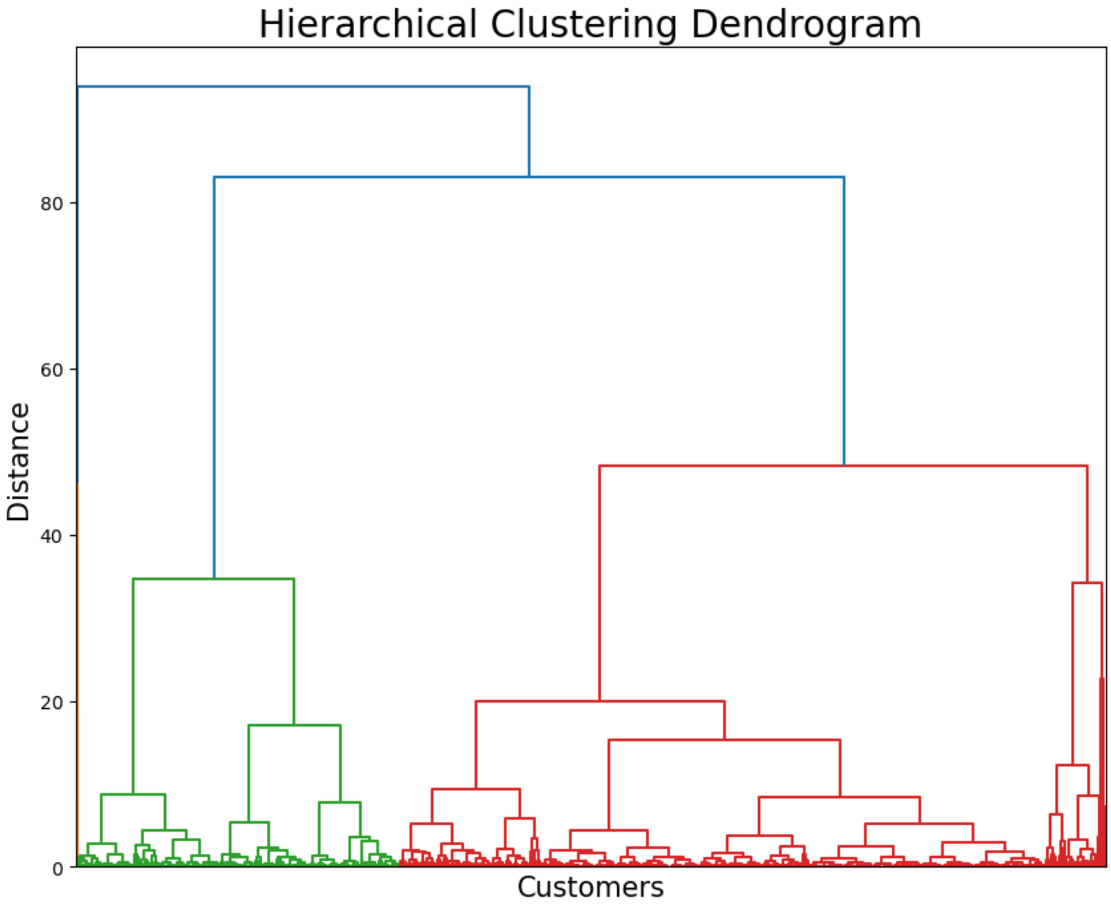


Figure 15 Hierarchical Clustering Dendrogram

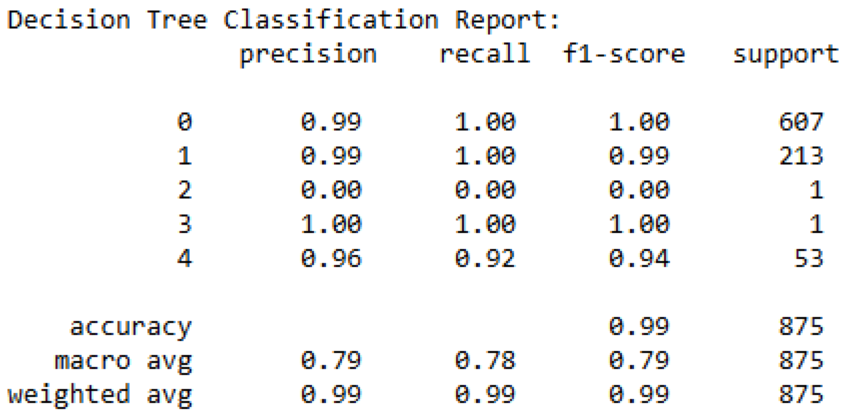
The dendrogram from hierarchical clustering(Figure 15) visualizes customer segments based on RFM (Recency, Frequency, Monetary) values. Each merge represents combined clusters, with vertical distance indicating dissimilarity. Three main clusters are revealed: Cluster 0 (green) with moderate to low RFM values (infrequent, low-value purchases), Cluster 1 (red) with varying RFM values (diverse behaviours), and Cluster 3 (blue) with high-frequency, high-value customers (high engagement).

The silhouette score of 0.619 indicates well-defined clusters, validating effective segmentation for targeted marketing strategies.

Chapter-6 Supervised Learning

Supervised learning is a type of machine learning where the model is trained on labelled data, meaning each training example includes input-output pairs. The algorithm learns to map inputs to outputs and can make predictions on new, unseen data. In this project, it is used to predict customer segments based on their RFM features. This allows for automated classification of new customers into predefined segments, enabling personalized marketing and better resource allocation to enhance customer engagement and retention.

6.1 Decision Tree Classification

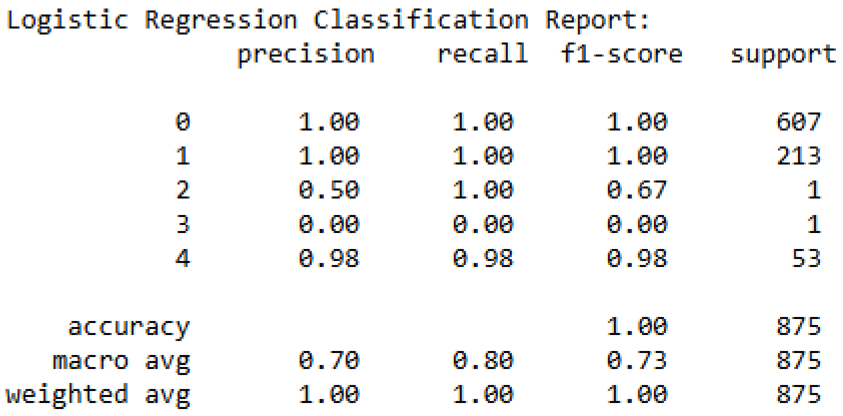


```
Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       607
           1       0.99      1.00      0.99       213
           2       0.00      0.00      0.00         1
           3       1.00      1.00      1.00         1
           4       0.96      0.92      0.94        53

    accuracy                           0.99       875
   macro avg       0.79      0.78      0.79       875
weighted avg       0.99      0.99      0.99       875
```

Figure 16 Decision Tree Classification report

Figure16 shows high precision, recall, and f1-scores for most clusters, particularly clusters 0, 1, and 3, indicating accurate predictions. It fails to predict cluster 2, resulting in zero scores for that segment. With an overall accuracy of 0.99 and a high weighted average f1-score of 0.99, the model effectively classifies customers for targeted marketing strategies.

6.2 Logistic Regression Classifier

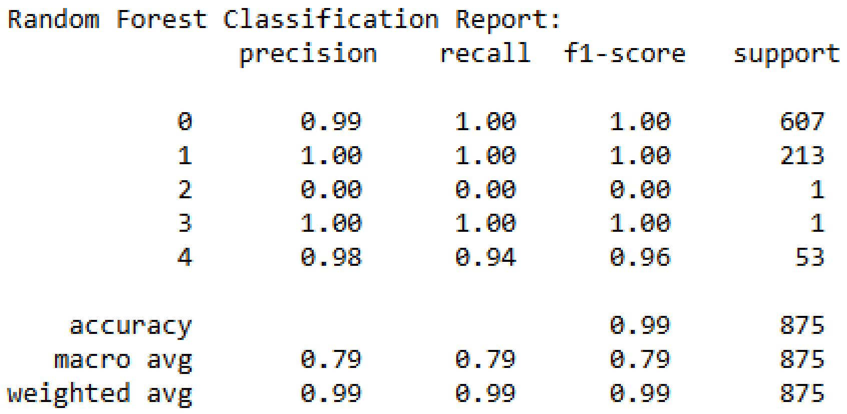


```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       607
           1       1.00      1.00      1.00       213
           2       0.50      1.00      0.67         1
           3       0.00      0.00      0.00         1
           4       0.98      0.98      0.98        53

    accuracy                           1.00       875
   macro avg       0.70      0.80      0.73       875
weighted avg       1.00      1.00      1.00       875
```

Figure 17 Logistic Regression Classification report

Figure17 shows high precision, recall, and f1-scores for clusters 0, 1, and 4, indicating accurate predictions. Cluster 2 has a precision of 0.50 and an f1-score of 0.67, while cluster 3 has zero scores, reflecting no correct predictions. The overall accuracy is 1.00, with a weighted average f1-score of 1.00, demonstrating excellent performance across most segments. This highlights the effectiveness of logistic regression in classifying customers for targeted marketing efforts.

6.3 Random Forest Classifier



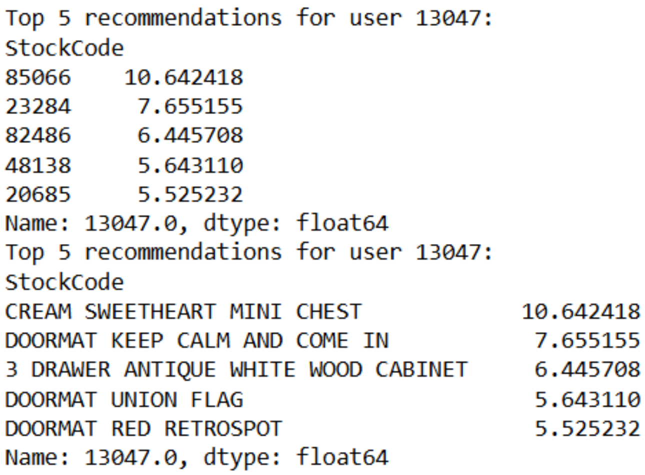
```
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       607
           1       1.00      1.00      1.00       213
           2       0.00      0.00      0.00         1
           3       1.00      1.00      1.00         1
           4       0.98      0.94      0.96        53

    accuracy                           0.99       875
   macro avg       0.79      0.79      0.79       875
weighted avg       0.99      0.99      0.99       875
```

Figure 18 Random Forest Classification report

Figure18 shows high precision, recall, and f1-scores for clusters 0, 1, and 3, indicating strong predictive accuracy. Cluster 2 has zero scores, and cluster 4 has slightly lower but high metrics, with an f1-score of 0.96. The overall accuracy is 0.99, with a weighted average f1-score of 0.99, reflecting excellent performance across most segments. This demonstrates the random forest model's effectiveness in classifying customers for targeted marketing strategies.

Chapter-7 Recommendation System

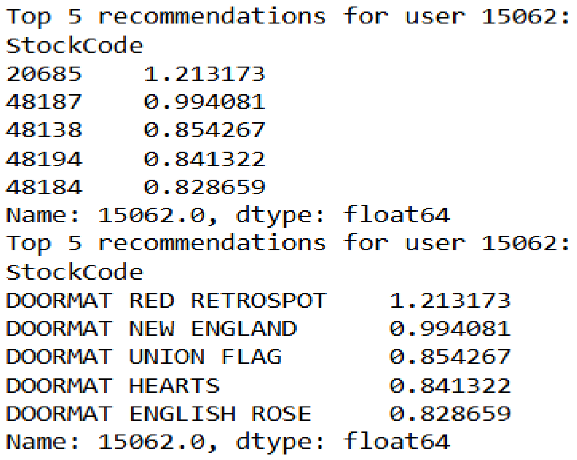
A recommendation system in customer segmentation aims to suggest products to customers based on their past behaviour and the behaviour of similar customers. By leveraging the insights gained from customer segmentation, such as RFM analysis and clustering, a recommendation system can personalize marketing efforts, enhance customer experience, and increase sales.



```
Top 5 recommendations for user 13047:
StockCode
85066    10.642418
23284     7.655155
82486     6.445708
48138     5.643110
20685     5.525232
Name: 13047.0, dtype: float64
Top 5 recommendations for user 13047:
StockCode
CREAM SWEETHEART MINI CHEST           10.642418
DOORMAT KEEP CALM AND COME IN          7.655155
3 DRAWER ANTIQUE WHITE WOOD CABINET    6.445708
DOORMAT UNION FLAG                     5.643110
DOORMAT RED RETROSPOT                  5.525232
Name: 13047.0, dtype: float64
```

Figure 19 Recommendation for user 13047

Figure19 displays the top 5 product recommendations for user 13047. It include both stock codes and product descriptions, with corresponding scores indicating the predicted relevance or likelihood of purchase. The "CREAM SWEETHEART MINI CHEST" has the highest score of 10.64, suggesting it's the most relevant recommendation for this user.



```
Top 5 recommendations for user 15062:
StockCode
20685    1.213173
48187    0.994081
48138    0.854267
48194    0.841322
48184    0.828659
Name: 15062.0, dtype: float64
Top 5 recommendations for user 15062:
StockCode
DOORMAT RED RETROSPOT    1.213173
DOORMAT NEW ENGLAND      0.994081
DOORMAT UNION FLAG       0.854267
DOORMAT HEARTS           0.841322
DOORMAT ENGLISH ROSE     0.828659
Name: 15062.0, dtype: float64
```

Figure 20 Recommendation for user 15062

Figure20 shows the top 5 product recommendations for user 15062. The recommended products are primarily various types of doormats, each with a corresponding relevance score. The "DOORMAT RED RETROSPOT" has the highest score of 1.213173, indicating it is the most relevant recommendation for this user. The other products have slightly lower scores but are still considered highly relevant.

Chapter-8 Conclusion

In this project we analysed customer purchase behaviours using a dataset from a UK-based online retail store, encompassing 401,604 transactions. RFM analysis provided essential insights into customer value, revealing patterns such as high engagement from recent buyers.

Unsupervised learning methods, including K-means and hierarchical clustering, effectively segmented customers. K-means identified five clusters, highlighting high-frequency, high-value customers and low-value buyers. Hierarchical clustering confirmed these segments, achieving a silhouette score of 0.619, indicating well-defined clusters.

Supervised learning models, such as Decision Tree (99% accuracy), Logistic Regression (100% accuracy), and Random Forest (99% accuracy), accurately predicted customer segments, particularly in dominant clusters. These models demonstrated their utility in automating customer classification for targeted marketing.

The recommendation system, based on RFM insights, provided personalized product suggestions, enhancing customer experience and boosting sales. For example, user 13047 received top recommendations like "CREAM SWEETHEART MINI CHEST" with a relevance score of 10.64. Overall, this project equips businesses with data-driven tools to optimize marketing strategies, improve customer retention, and increase profitability.

Chapter-9 Reflection

Working on this project taught me the significance of data cleanliness and analysis. I learned how to deal with missing values, remove duplicates, and apply RFM analysis to understand customer behaviour. K-Means and Hierarchical clustering assisted in identifying consumer segments, while decision trees, logistic regression, and random forests increased classification accuracy. Developing a recommendation system demonstrated practical advantages. However, depending on historical data has limitations, and future studies could incorporate real-time data and other elements to provide further insights. Overall, this project helped me get better grasp data-driven approaches for business improvement.

Chapter-10 References

Bose, I. and Chen, X. (2009) 'Quantitative models for direct marketing: A review from systems perspective', *European Journal of Operational Research*, 195(1), pp. 1-16.

Rygielski, C., Wang, J.C. and Yen, D.C. (2002) 'Data mining techniques for customer relationship management', *Technology in Society*, 24(4), pp. 483-502.

Kaggle (n.d.) *Online Retail Dataset*. Available at: https://www.kaggle.com/datasets (Accessed: 25 June 2024).

McKinsey & Company (2016) *Advanced analytics in retail: Building value in an increasingly competitive landscape*. Available at: https://www.mckinsey.com (Accessed: 1 July 2024).

Parthasarathy, S. (2006) 'Customer Relationship Management: A Data Mining Approach', in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*. Chicago, IL, 27-29 June. New York: ACM Press, pp. 1-10.

Liang, Y., Zhou, Y., Han, Z., Zhao, L. and Li, W. (2023) 'Customer segmentation in e-commerce: A data-driven approach', *Information Technology and Management*, Available at: https://link.springer.com/article/10.1007/s10257-023-00640-4 (Accessed: 27 June 2024).

Chen, Y., Zhang, S., Yuan, X. and Huang, Z. (2017) 'Customer segmentation in e-commerce: A comprehensive study', *Journal of Business Research*, 83, pp. 1-11. Available at: https://www.sciencedirect.com/science/article/pii/S0148296317304939 (Accessed: 30 June 2024).