National College of Ireland

Enhancing Customer Segmentation and Behaviour Analysis with RFM Clustering: A Machine Learning Approach

MSc Research Project
Data Analytics

Nandheeswari Rajendran
Student ID: X22132210

School of Computing
National College of Ireland

Supervisor: Prof. Cristina Hava Muntean

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……...Nandheeswari Rajendran…….

Student ID: ……X22132210…….

Programme: ……..MSc in Data Analytics…….. **Year:** 2023-2024

Module: ……MSc Research Project……

Supervisor: …….. Prof. Cristina Hava Muntean……

Submission Due Date: ……31/01/2024…….

Project Title: Enhancing Customer Segmentation and Behaviour Analysis with RFM Clustering: A Machine Learning Approach

Word Count: ……7752…… **Page Count:** ……26……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Nandheeswari Rajendran……

Date: ……31/01/2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Customer Segmentation and Behaviour Analysis with RFM Clustering: A Machine Learning Approach

Nandheeswari Rajendran
X22132210

Abstract

This research project examines to tackle the challenge of customer segmentation and clustering through the utilization of Recency, Frequency, and Monetary (RFM) analysis with an extensive transactional dataset. RFM analysis is a potent marketing approach that categorizes, and groups customers based on recency, frequency, and monetary metrics. Additionally, time series analysis is conducted on both a monthly and daily basis to gain insights into customer interactions at various times of the day. In the clustering phase several distinct algorithms such as K-Means, Agglomerative, and Mean-shift are employed, using standardized RFM scores as input features. The research evaluates performance using the metrics Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. The research methodology encompasses data collection, preprocessing, feature engineering, detailed exploratory data analysis to extract meaningful customer attributes. These findings highlight valuable customer segments that can be targeted for specific marketing strategies. The comprehensive analysis of this study reveals that the Agglomerative clustering model consistently outperforms both K-Means and Mean-shift models, showcasing its superiority in effectively grouping customers based on their transactional behaviours, thus highlighting its importance in the online retail industry's optimisation of customer segmentation.

Keywords: Clustering, Monetary, Frequency, Recency, K-Means, Agglomerative, Mean-shift, Customer Segmentation, RFM Analysis, Distribution

1 Introduction

In current business environment, understanding and catering to the diverse requirements of customers is very essential for organizational success. This study report explores into how machine learning method of clustering can improve the way of customers segmentation, aiming to refine and improve its processes. Given the ample availability of customer data, our primary objective is to move beyond traditional demographic-based segmentation and explore deeper customer behaviour is for more informed decision-making. Through analysing of all available data, businesses may determine what must be changed to improve the experience of their customers. By revealing valuable customer segments, companies can better satisfy to the unique needs of these groups, refine their marketing efforts, and personalize product offerings to align with every segment’s preference. The research findings

underscore the significant role played by RFM analysis and clustering techniques in enhancing our comprehension of customer habits and improving promotional approaches in the evolving online retail sector.

The research question addressed by this study is: What are the significant insights associated with employing clustering algorithms such as Agglomerative, K-Means, and Mean-shift for customer segmentation based on RFM analysis of transnational data?

The key purpose of this research is to focus on the distinctive characteristics and interests of each consumer category to improve customer segmentation by the interpretability of resulting clusters scalability with extensive dataset(Varadarajan, 2020). The goal was to first assure the data's dependability by carefully processing and cleaning it, and then to improve its depth according to their transactional behaviours by utilising RFM analysis with Clustering approaches, more specifically was to focus on the frequency, recentness, and monetary value of their transactions. By focusing on the distinctive characteristics and interests of each consumer category, this strategy attempted to help firms better target their marketing campaigns.

The structure of this document follows comprehensive summary of existing studies is provided in section 2, with a focus on studies conducted on the topic of discussion in section 1. In section 3, the research approach and machine learning techniques that be employed to tackle the research topic are covered. The implementation's underlying architectural structure is expounded upon in section 4. The models built for each clustering model, and the data transformation activities are covered in section 5. The evaluation results of this research methodology are presented in section 6. The findings are summarised in section 7 together with the future scope.

2 Related Work

The two main sections of this research review are dedicated to improving customer segmentation using RFM analysis and integrating of clustering techniques. This literature review seeks to uncover different clustering algorithms to be applied in this research for fully comprehending consumer behaviour and purchasing habits. In addition, many clustering models will be employed in this study to do a comparison analysis and select the best model in terms of all factors considered and future prediction.

2.1 Customer Segmentation and RFM Analysis

The literature (Kabasakal, 2020) underscores Customer Relationship Management (CRM) pivotal role in sustaining business competitiveness through customer-centric strategies. The effectiveness of segmentation methods, notably the RFM Model, receives praise for its simplicity and effectiveness in categorizing customers. Remarkable strengths in the literature encompass detailed discussions on RFM's practical application in e-retail. Consequently, the study advocates for more accessible CRM tools, promoting the development of compact

software options. Furthermore, while studies support the potential advantages of RFM analysis, a recurrent call for future research urges a deeper exploration of the long-term effects of segment-based decisions, highlighting a gap in comprehensive outcome assessment within strategies. Additionally, even while studies demonstrate the potential benefits of RFM analysis, there is a persistent need for more study to fully examine the long-term impacts of segment-based decisions, pointing to a lack of thorough outcome assessment in CRM strategies.

The work done by (Abbasimehr and Shabani 2021) enhanced our ability to predict consumer behaviour by merging time series techniques with RFM analysis, addressing the limitations of conventional segmentation methods. They demonstrated that combining these approaches produced more accurate forecasts, particularly when tested with sales data from a bank. However, most of the testing was done in the banking industry, therefore its general applicability is unknown. Their approach looks promising, but more testing in many industries and with various types of data is necessary to ensure its applicability worldwide.

Based on the project study customer segmentation and profiling using RFM Analysis, which was done by (Sabuncu, Turkan, and Polat 2020) provides a thorough analysis on utilising the RFM model to categorise customers. The study used statistical techniques, starting with demographic data, to score and divide the consumer base into five parts using SPSS program by cluster analysis. It then used sophisticated analytical tools to create client profiles. The paper presents RFM score analysis as a critical instrument in gasoline station marketing and customer relationship management, highlighting the efficacy of scientific segmentation methodologies. The study (Ernawati 2021) data mining techniques in RFM based consumer segmentation examined the integration of RFM models and data mining techniques for the analysis of consumer behaviour. The research, which covered the years 2015–2020, emphasised the importance of clustering and visualisation as critical techniques in this field for successful market targeting and strategy development. Recognising the growing significance of visualisation, it put out a novel framework that combines data mining and Geographic Information Systems (GIS) to reveal customer attributes, allowing businesses to improve their market objectives and tactics.

The study done by (Smaili and Hachimi 2023) examines better understanding and forecasting consumer behaviour for efficient marketing with RFM-D Model, which is being undertaken by traditional techniques such as RFM segmentation are shallow because they ignore the variety of consumer preferences. This study improves the RFM model and generates more precise customer clusters by adding the Diversity (D) parameter, which considers the diversity of products people purchase. Comparisons with machine learning methods demonstrate how much more accurate and effective RFM-D model. To develop more accurate segmentation based on product-specific loyalty, the conclusion emphasises the necessity to include other criteria such as customer membership duration and offers future research directions. In addition, the study suggests investigating consumer groups with a range of tastes to guarantee more precise forecasts and inclusive targeting in advertising efforts. (Rahim, Mushafiq, Khan, and Arain 2021) The study emphasises how crucial it is to

carefully consider different factors and how they relate to one another to develop more useful segmentation models, especially in retail sectors where a wide range of consumer preferences are important in their repurchase behaviour. To make more accurate forecasts and prevent the exclusion of potential clients from marketing campaigns, it also promotes further research into customer clusters that allow membership in several segments.

The study of (Aggarwal and Yadav 2020) centres on customer segmentation in online shopping to calculate lifetime value. They cluster data into eight parts using fuzzy AHP and RFM characteristics to prioritise variables. Their research, which emphasises the importance of monetary value in CLV, helps marketers develop strategies for high CLV clusters. It does, however, recognise the limits of dataset specificity and recommends further study into various businesses, various RFM priorities, and alternative clustering methods such as Partitioning Around Medoids and Clustering Large Applications. Overall, it highlights how important customer segmentation is for customised strategies that maximise client lifetime value for online firms.

2.2 Integration of Clustering Techniques

The research (Firdaus and Utamas, 2021) presents the RFM+B model, a breakthrough in client segmentation in the banking sector. Recency-Frequency-Monetary analysis has historically been essential to comprehending customer behaviour. However, due to its shortcomings, customer balance (B) was included in the research. This model implements K-Means clustering to incorporate the aspects of balance, money, frequency, and recency. When the model was used on 147 thousand transaction records and 60,000 customers, it identified several consumer groups with high frequency, balance, recency, and monetary values. The study concludes that RFM+B is a useful tool for segmentation and could improve marketing strategies, increase corporate revenues, and promote the rise of Total Processing Fee. The study also highlights future opportunities for optimizing the model using alternative or combined clustering methods, contributing to the banking industry's analytical framework.

In (Abdulhafedh, 2021) study, customer segmentation for a credit card company's marketing plan was examined by the combination of clustering techniques: K-means, Hierarchical clustering, and PCA. To classify precise consumer categories, 8950 active cardholders' transaction histories were analysed. K-means excelled over Hierarchical clustering in terms of effectiveness, as demonstrated by metrics such as the Davis-Bouldin, Silhouette, and Dunn index. The study highlighted how PCA may reduce dimensions, reduce multicollinearity, and improve clustering by finding more customer clusters. Notably (Zhu, 2019) K-means performance was much improved by modifying the ideal K value considering PCA insights. To achieve better clustering results, the conclusion emphasised the significance of careful data preprocessing, investigating a variety of clustering algorithms, and suggesting future research directions involving increasingly complex techniques like Distribution-Model Based clustering, Density-Based, and Fuzzy C-means along with ongoing PCA integration.
The work presented by (Wu, 2020) investigates the online sales data to classify consumers according to their purchasing habits through employing RFM model and K-means clustering

algorithm. To improve customer happiness, the study divides consumers into four different groups and suggests customised CRM tactics. Substantial gains in key measures, such as a 529% increase in active customers, a 279% increase in total purchase volume, and a noteworthy 101.97% increase in total consumption amount, demonstrate the efficacy of this strategy. Through a methodical analysis of consumer behaviour, the study (Anitha and Patil, 2022) offers businesses a useful foundation for improving overall performance and CRM strategy optimisation using K-means algorithm. Several dataset clusters are verified by the Silhouette Coefficient calculation. In addition to highlighting the need for more study, the conclusions stress the need of integrating algorithms into CRM systems to support managerial decision-making for ongoing boosting performance and the necessity of updating algorithms to fit updated datasets for theoretical analysis.

The research by (Mamashli and Zolfani's 2022) project investigates the combination of data analysis and customer behaviour modelling using the RFMT mode in Iranian private banks. Based on transaction history, frequency, and spending patterns, the study (monetary (M), recency (R), Time(T) and customer frequency (F)) divided mobile banking users into six clusters. This highlights how crucial it is to comprehend consumer behaviour by RFMT analysis. In the research paper given by (Mensouri, 2022) a new model called RFMTS was developed for improved customer segmentation by adding the satisfaction (S) value to the existing RFMT model. Any dataset from which the five variables (recency, frequency, money, time, and satisfaction) can be extracted can be used with their newly developed model. By classifying customers into distinct groups according to the suggested RFMTS model, decision makers better able to recognise market segments and create marketing and sales plans that increases client loyalty. The RFM+B model was creatively established by (Firdaus and Utama 2021), completely changing the way that banks segment their customer base. The RFM+B model extended previous RFM approaches' focus on transactional behaviour (recency, frequency, and monetary factors) by incorporating customer balance. This model successfully divided 65 thousand consumers from 147 thousand transactions over a six-month period into four different clusters by utilising K-Means clustering. (Parikh and Abdelattah 2020) The study focus of RFM analysis and clustering techniques to identify consumer purchase patterns with relatively high balances, frequency, expenditure, and recentness. It also establishes a new standard for research in related fields in the future. The paper recommends more research into different or combination clustering techniques to maximise the effectiveness of the model in upcoming applications.

Using the DBSCAN algorithm (Monalisa, 2023) investigates customer segmentation by combining demographic data and RFM measures. Their objective was to identify possible clients by enhancing and organising data. They identified 5 clusters and 31 noisy data points through careful analysis using RStudio tools, and they were able to create optimal parameters resulting in a strong silhouette index values 0.42. Based on RFM analysis, Cluster 1 was identified as a potential customer, whilst Clusters 2–5 were classified as devoted customers. And the paper (Hossain, 2017) illustrated to employ centroid-based algorithms added to density-based clustering models respectively, this work presented the concept of utilising density-based algorithms for consumer segmentation has an extra option to get satisfactory

clustering results in grouping and analysing the customer behaviour. The research paper authored by (Peker, Kocyigit, Eren 2017) presents a pioneering concept known as the Length, Recency, Frequency, Monetary, and Periodicity (LRFMP) model, which offers a fresh perspective on customer segmentation within the grocery retail sector. This research merges the LRFMP model with clustering techniques, utilizing real-world data from a Turkish grocery chain. Furthermore, the paper has practical implications, as it provides a structured guide for researchers and practitioners to effectively profile customers using the LRFMP model.

In summary, while existing literature has underscored the efficacy of RFM analysis and clustering techniques in customer segmentation, limitations persist in their adaptability to dynamic market conditions, and interpretability of results with performance metics. A more flexible strategy that combines RFM measures with a sophisticated comprehension of clustering methods such as K-Means, Agglomerative, and Mean-shift are required to meet the dynamic nature of consumer behaviour in contemporary markets. However, they emphasise that for these models to be broadly applicable, further testing and ongoing development are required. Overall, these studies highlight the need for continued improvements in segmentation techniques to strengthen marketing plans and optimise consumer value.

3 Research Methodology

This research follows a structured approach, using a technique called Knowledge Discovery in Databases, to create and test a model that divides customers into different groups using clustering machine learning methods. The approach of Knowledge Discovery in Databases in shown in Figure 1, a systematic way to derive important insights gathered from extensive sets of data, which is especially useful when dealing with the complexities of customer segmentation. By thoroughly analysing and comparing various clustering algorithms, this study improves how accurately and effectively customers are grouped together. The results of this research offer valuable insights into how effective clustering algorithm are highlighting its potential to enhance marketing strategies and make customer experiences better while also strengthening business profits.

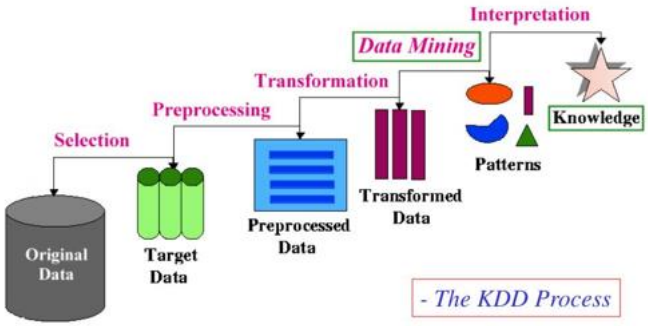

Figure 1: Knowledge Discovery Database Steps

3.1 Data Collection

This study makes use of transactional data from an online gift shop located in the United Kingdom used to provide a diversified dataset (Online UCI ML Repository). The retail company offers a range of distinctive gift items, and according to the dataset, most of its customers are also wholesale merchants. The dataset captures all the transactions between the duration of years 2010 and 2011 that comprises total 541909 observations and as given in Table 1, dataset has eight variables in total, five of which are categorical and three of which are numerical variables.

InvoiceNo	a unique 6-digit code assigned for every individual transaction
StockCode	a unique 5-digit code assigned for every individual product
Description	product name
Quantity	the quantities of every product purchased in a transaction
InvoiceDate	the time and date of creation of each transaction
UnitPrice	cost of every individual product
CustomerID	a unique 5-digit whole number that is assigned to each individual customer
Country	the country name where the customer lives

Table 1: Dataset Description

The research technique comprised a methodical framework from data gathering to analysis. Python libraries such as scikit-learn, and pandas helped with data pre-processing, cleaning, and outlier identification. Python was used in feature extraction to calculate total spend and derive temporal information. Scikit-learn, a Python package, was used for RFM analysis to scale features and perform analysis. Exploratory visualisations and grouping were made easier by Python packages Matplotlib, Seaborn, and scikit-learn. Various clustering methods were used for segmentation and in-depth analysis included evaluation of algorithm performance through metrics.

3.2 Data Preprocessing and Cleaning

To assure the integrity and dependability of the dataset, pre-processing and data cleaning was started from the beginning of the project. Handling missing values was an essential aspect of preparing data. Two columns in the dataset have missing values such as CustomerID and Description of the product. Dropping of those missing entries are done and after removing them there exist total 406829 observations. This guaranteed the completeness of the dataset, allowing for analysis without sacrificing its integrity. The dataset was made more accurate and in line with the desired analytical objectives by eliminating irrelevant entries. Statistical techniques like the interquartile range (IQR) were utilized for identification of outliers, or data points that differed drastically separate from the rest. Since outliers had the ability to skew analyses, handling them required either removing or adjusting them. By ensuring that the dataset was not overly influenced by extreme values, this process helped to produce analyses that were more reliable and accurate.

3.3 Feature Extraction

Feature extraction aimed to enrich the dataset by deriving new attributes that provided deeper insights into customer behaviour. By taking time-related information out of transaction dates, it was possible to determine how customer behaviour changes over time. Transaction timestamp patterns were identified by dissecting them into relevant elements, which highlighted seasonal patterns, peak activity periods, and frequency of purchases. Strategies that matched the activity patterns of the customers were made possible by this insight.

Finding the cumulative sum spent by each customer offered a thorough understanding of their financial involvement. This statistic identified high-value customers or those in need of extra incentives for more involvement by indicating a customer's overall monetary value to the company. A thorough understanding of the financial impact on various consumer segments was provided by adding a new feature spend at the end column as shown in Figure 2.

	InvoiceNo	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Invoice_Year	Invoice_Month	Invoice_Day	Invoice_WeekDay	spend
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	2010	12	1	2	15.30
1	536365	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	2010	12	1	2	20.34
2	536365	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	2010	12	1	2	22.00
3	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	2010	12	1	2	20.34
4	536365	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	2010	12	1	2	20.34

Figure 2: Addition of New Feature - Spend

And customers cancelled transactions are identified using regex symbols and removed them as they don’t appear as final transaction. These factors increased the depth and complexity of the dataset. This upgraded dataset proven to be an invaluable tool for well-informed decision-making. In essence, feature extraction broadened the dataset's scope by incorporating temporal and financial insights, enabling deeper understanding, and facilitating informed strategic decisions in subsequent project phases. Following all this steps, the dataset now includes roughly 20,000 transactions and 4339 distinct customers.

3.4 RFM Analysis

It was essential to assess customers using monetary, frequency, and recency indicators as RFM analysis a marketing tool to segment customers. These RFM metrics recorded the frequency, volume, and recentness of the purchases made by customers. Recency was calculated by knowing customers recent visit subtracting with their last active visit, and frequency calculated by counting customer number of invoices engaged in transactions followed monetary calculated analysing total amount spent of each customer. By giving these measures numerical values shown in Figure 3, it was possible to evaluate each customer's interaction with the company numerically. In the experimental stage, the distributions of the

RFM metrics provided a thorough understanding of consumer involvement by highlighting trends that some customers are very frequent, recent and spend on high while some customers are not however there existing lots of variation in customer data. RFM analysis provided an organised way to classify clients, which set the groundwork for strategic decision-making. This analysis led to an improvement in customer happiness, retention, and overall business performance.

	monetary	frequency	recency	r_quartile	f_quartile	m_quartile	RFM_Score
count	4339.000000	4339.000000	4339.000000	4339.000000	4339.000000	4339.000000	4339.000000
mean	105.041484	91.708689	2053.307905	2.999539	3.021434	3.006684	3.011869
std	100.007757	228.792852	8988.247311	1.415028	1.429696	1.413138	0.651848
min	13.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.100000
25%	30.000000	17.000000	307.000000	2.000000	2.000000	2.000000	2.600000
50%	63.000000	41.000000	674.000000	3.000000	3.000000	3.000000	3.000000
75%	154.500000	100.000000	1661.000000	4.000000	4.000000	4.000000	3.400000
max	386.000000	7847.000000	280206.000000	5.000000	5.000000	5.000000	5.000000

Figure 3: RFM Metrics Calculation

3.5 Exploratory Data Analysis

To obtain a thorough understanding of customer behaviour, the exploratory phase of our investigation made use of a wide range of analytical techniques and visualisations. Spending behaviour analysis, customers monthly invoice, quantity distributions, and word clouds emphasising frequently occurring phrases in product descriptions were among the visualisations. Comparative studies between other nations and in-depth time series analyses conducted on a monthly and daily basis exposed complex patterns, trends, and variances in consumer behaviour. On analysing the countries order distributions depicted in Figure 4, it is explicit as the transaction data is from UK base company, most orders (89%) came from the UK customers. Analysing quantity distributions was another important investigation that revealed trends in purchase volumes and consumer purchasing behaviour. Identification of these distributions helped in better understand customer behaviour regarding product quantities purchased.

	Country #	Country %
United Kingdom	354345	0.890484
Germany	9042	0.022723
France	8342	0.020964
EIRE	7238	0.018189
Spain	2485	0.006245
Netherlands	2363	0.005938
Belgium	2031	0.005104
Switzerland	1842	0.004629
Portugal	1462	0.003674
Australia	1185	0.002978
Norway	1072	0.002694
Italy	758	0.001905
Channel Islands	748	0.001880
Finland	685	0.001721
Cyprus	614	0.001543
Sweden	451	0.001133

Figure 4: Country Wise Distribution of Orders

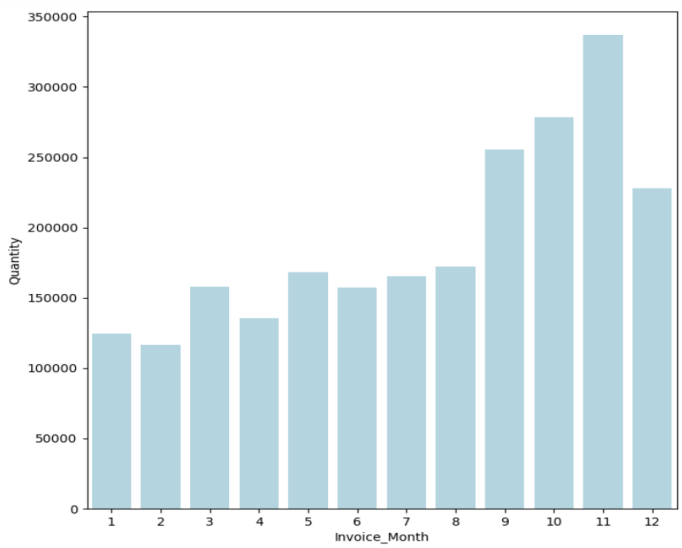

Figure 5: Customers Monthly Invoices Distribution

The significant finding is that November month has the customers highest transaction volume of the year as illustrated in Figure 5. Furthermore, plotting the numerical features of spend and quantity, unit price made it easier to evaluate the normality of metric distributions, which helped to better comprehend their statistical properties. It is explicit from Figure 6, the spend has more variations than the quantity and unit price. The amount of a product acquired in a purchase and its unit price are negatively correlated and it makes sense to assume that when a product's unit price rises, less of that thing have been sold and it is significant because it illustrates how customers tend to purchase fewer items in greater quantities when prices rise.

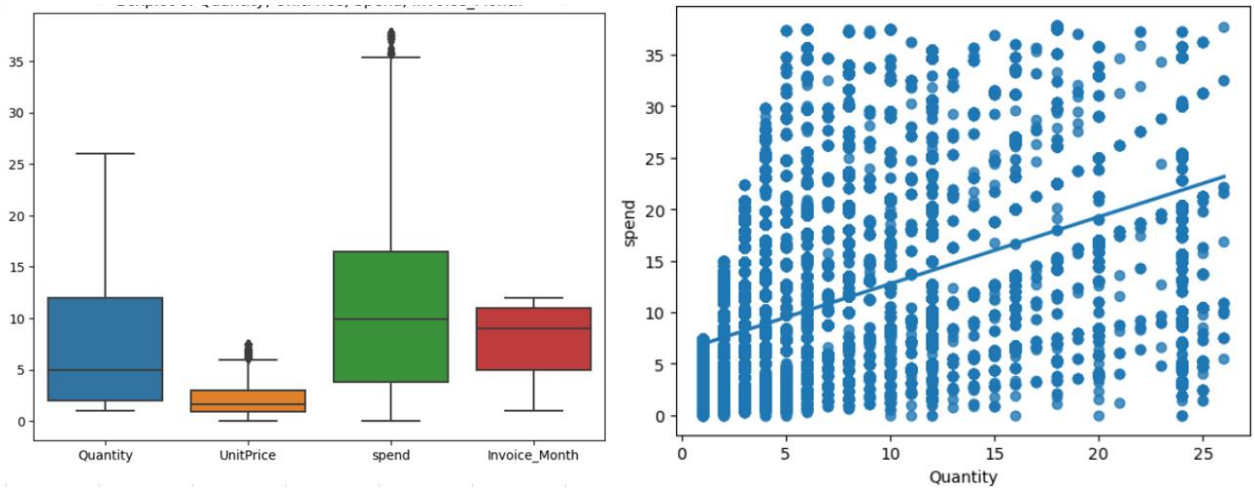

Figure 6: Analysis on Spend and Quantity

Moreover, word clouds were utilised to summarise commonly used terms that were taken from product descriptions, providing quick insights into words or phrases that are frequently utilised in relation to our items. Also, the analysis helped in understanding what was most popular in November, a month in when sales boomed, and a significant number of Christmas-related products are sold since it is likely that customers begin to prepare for the holiday in

November. As indicated by Figure 7, word clouds offered a summary of important terms from product descriptions and may have suggested popular features or categories.

Figure 7: Word Cloud - November Month Sold Products

Followed customer spending behaviour, the most and the least expensive product that are purchased are identified and illustrated in Figure 8, examining how different countries shop reveals interesting trends in purchasing habits across regions. This analysis helps us understand if certain regions tend to buy in similar ways, offering clues about how purchasing habits might relate to maximum and minimum limits. Differentiated purchase behaviours and temporal trends were found by performing in-depth time series analyses on various scales and analysing data on a country-by-country basis. These insights provide a basis for well-informed decision-making and provide an overview of the extensive exploratory analyses carried out in this stage.

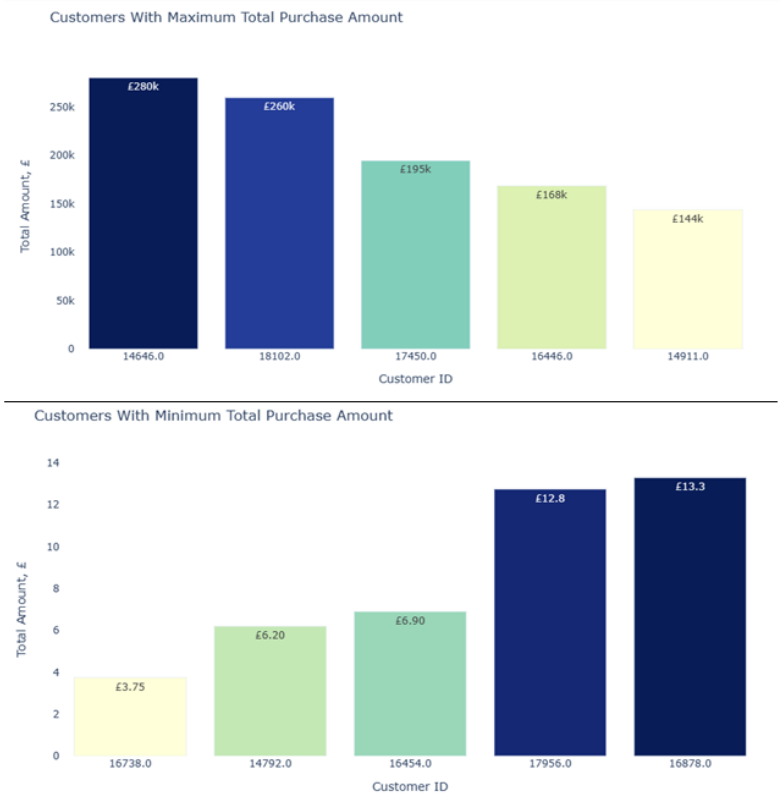

Figure 8: Analysis of Customers Total Purchase Amount

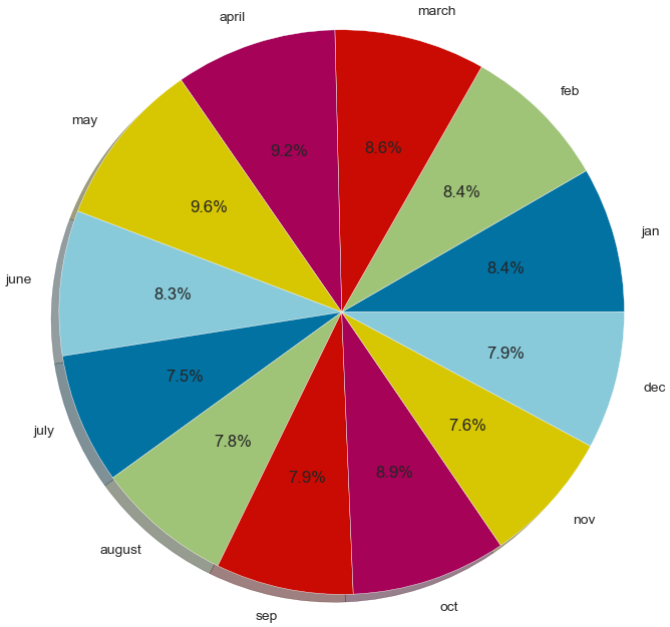

Figure 9: High Sales Proportion of Each Month

In time series analysis, examining monthly purchasing trends provided a more detailed view of consumer patterns by comprehending how purchasing habits change over the course of a given month. By analysing the biggest portion of total customer purchases of every month in the year illustrated in Figure 9 shows monthly fluctuations in consumer buying patterns.

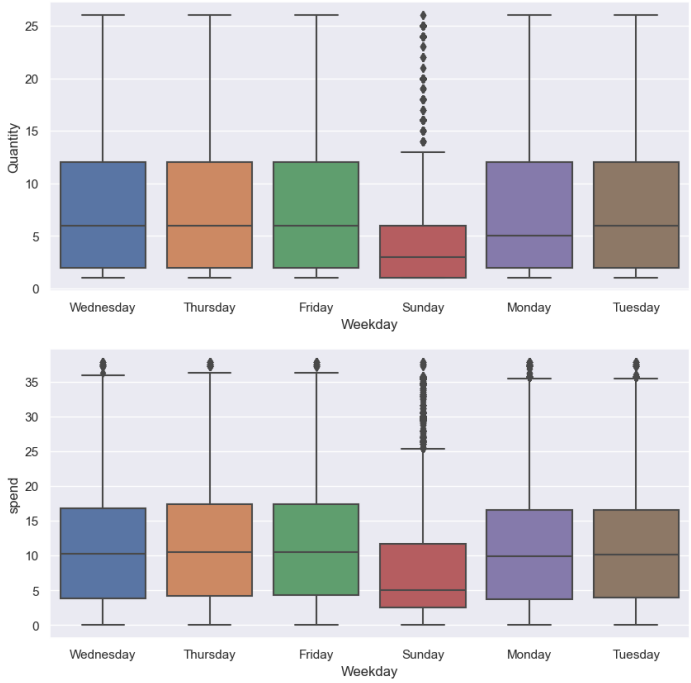

Figure 10: Weekly Transaction Analysis

As depicted in Figure 10, it is explicit that on the weekly transaction chart on Sundays there exist a notable decrease in both quantity of items sold and the total transaction amounts even though there are a few exceptional cases, it gives us interesting insight that the pattern indicates a consistent and significant decline in sales and spending specifically observed on Sundays, prompting the need for deeper analysis.

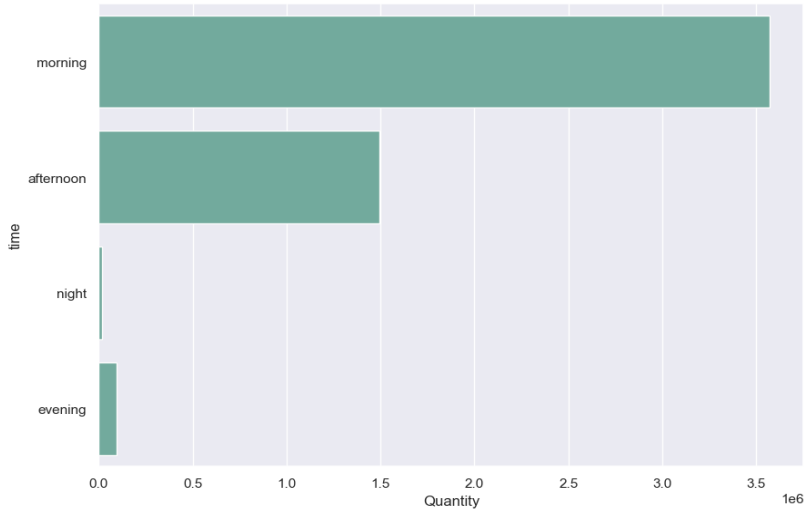

Figure 11: Parts of the Day Analysis

Analysis on classifying parts of the day as illustrated in Figure 11 reveals that most transactions around 63% of the total, happened in the morning, with a concentration between 8 and 12 AM. This is noteworthy since it shows a clear preference for making purchases in these early hours, suggesting a significant tendency among consumers to choose morning transactions.

4 Design Specification

This section clarifies the prerequisites by outlining the fundamental machine learning techniques and architecture used in the implementation process shown in Figure 12. This research follows two-layer design architecture with data persistence layer and business logic layer.

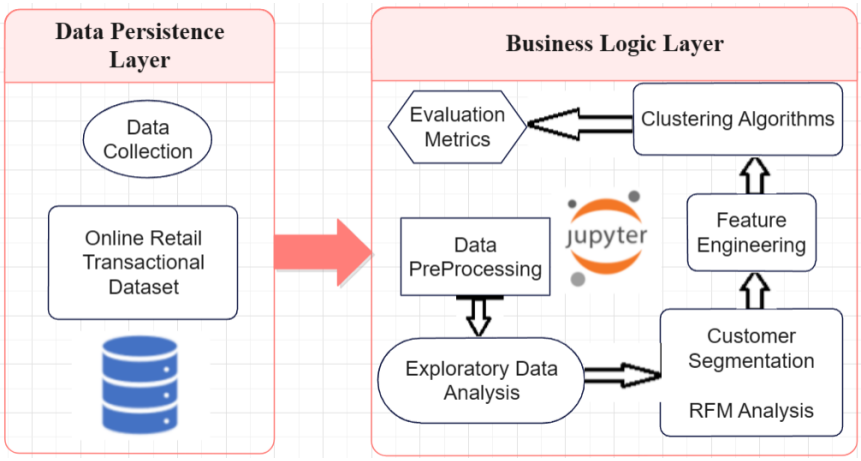

Figure 12: Design Architecture

Tier 1: Data Persistence Tier

The first layer, which is concerned with data durability, entails gathering and getting ready the dataset. The study is based on a publicly available online dataset that may be accessed as an excel file through the UCI repository. In this layer, transactional and customer data are extracted and then imported into a Jupyter Notebook for in-depth analysis and next steps.

Tier 2: Business Logic Tier

The next layer, referred to as the Business Logic Tier, includes the modelling and data processing stages that are crucial. Here, obtained data is meticulously cleaned to address missing values and guarantee consistency. New features time-related, total-spend features are developed to maximise model performance. Clustering techniques such as Agglomerative, K-Means, and Mean-shift are used in the segmentation of customers into clusters. Metrics like the Davies-Bouldin Index, Calinski-Harabasz Index, and Silhouette Score are employed to assess the effectiveness of the models and identify the best clustering model for consumer segmentation.

5 Implementation

This section focuses on the implementation details that explores deeply into the fundamentals of data engineering, including different clustering models and their segmentation of customers. To facilitate data manipulation and data analysis procedures, the project made use tools and technology stack of Jupyter Notebook for Python code execution, Python was also utilised along with a variety of libraries such as NumPy, Pandas, Seaborn, Matplotlib, and Pandas Profiling.

5.1 Clustering Based on RFM

The preparation procedures involving outlier analysis, feature scaling, visualisations, and the use of RFM attributes for clustering analysis. Using the Interquartile Range approach, outliers eliminated for the monetary, frequency, and recency characteristics to ensure more representative data for clustering analysis by removing extreme values. As indicated in Figure 13, the potential for these outliers to skew visualisations and affect grouping accuracy made their discovery which is essential for statistical analysis and predictive modelling. After the transformation of RFM attributes, the monetary feature showcased the distribution of customer spending, the frequency of customer purchases and the distribution of time since the last purchase allowing more accurate insights.

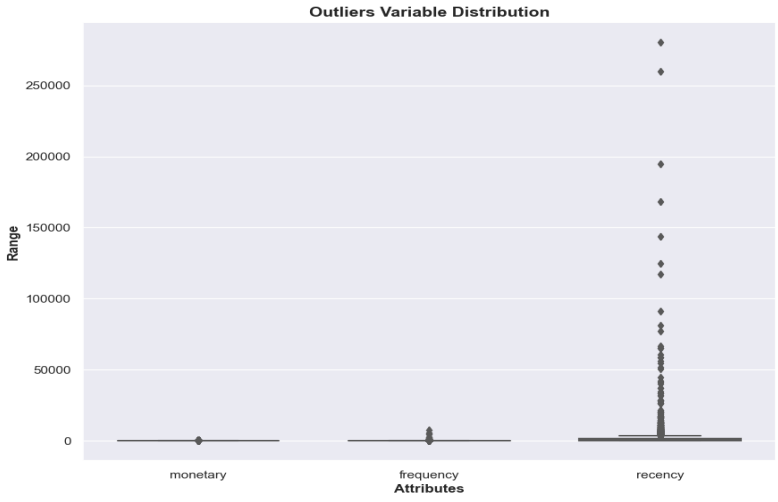

Figure 13: Outlier Analysis of RFM Attributes

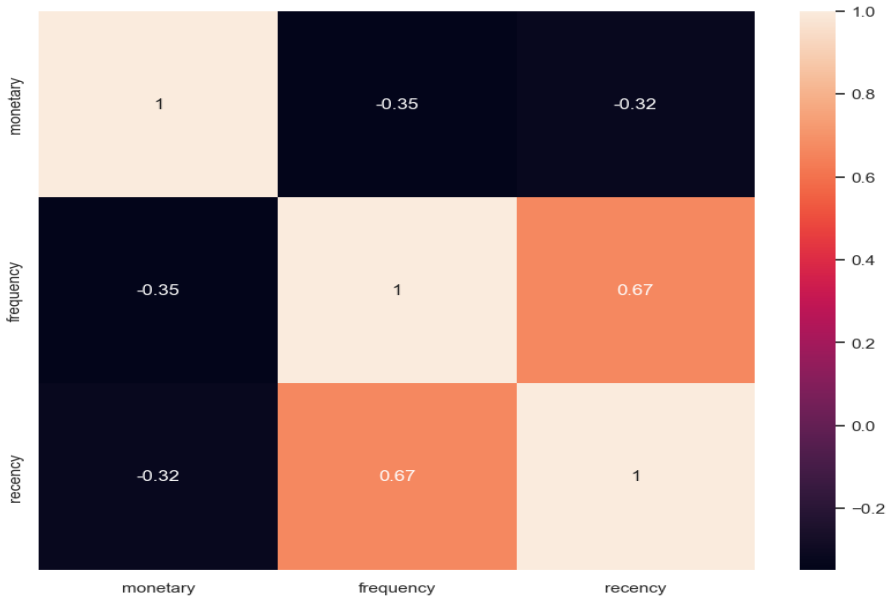

Figure 14: Heatmap of RFM Attributes

The heatmap visualization depicted in Figure 14, shows there exist moderate to strong positive correlations between monetary and frequency attributes, as well as frequency and recency values 0.67, while it shows a weak negative correlation between monetary and recency -0.32. Stronger positive connections are indicated by darker shade that are closer to 1.0 and weaker positive connections are indicated by lighter shade that were closer to 0.2. This heatmap provided insightful information about the connections between RFM measures, directing future research and segmentation strategies based on possible clustering patterns and interrelationships among the metrics.

5.2 Clustering Models

The clustering phase implemented three different clustering techniques used in unsupervised machine learning to find distinct consumer groups within the dataset. These techniques included Agglomerative, which combined comparable data points into bigger clusters, Mean-shift, which identified dense areas as clusters, and K-Means, which divided data according to centroid similarity. These techniques are chosen by following the guidelines provided on the learning website (scikit-learn) to identify inherent patterns or similarities in customer behaviour, preferences, or purchasing tendencies which are critical for defining distinct consumer segments. And the performance evaluation criteria are used to evaluate each clustering algorithm's performance in defining consumer segments to determine how effective every technique offered consumer categorization.

5.2.1 Agglomerative Clustering

The Agglomerative clustering algorithm was implemented to identify natural groupings within the customer dataset based on RFM attributes. Using a hierarchical clustering

technique, this clustering model was constructed with the help of Python modules. This technique creates a hierarchical tree-like structure (dendrogram) by first addressing every data point as a separate cluster afterwards merging clusters based on how similar they are to one another. The dendrogram was examined, considering the vertical line that crosses the longest vertical distance without crossing any horizontal lines, to establish the ideal number of clusters.

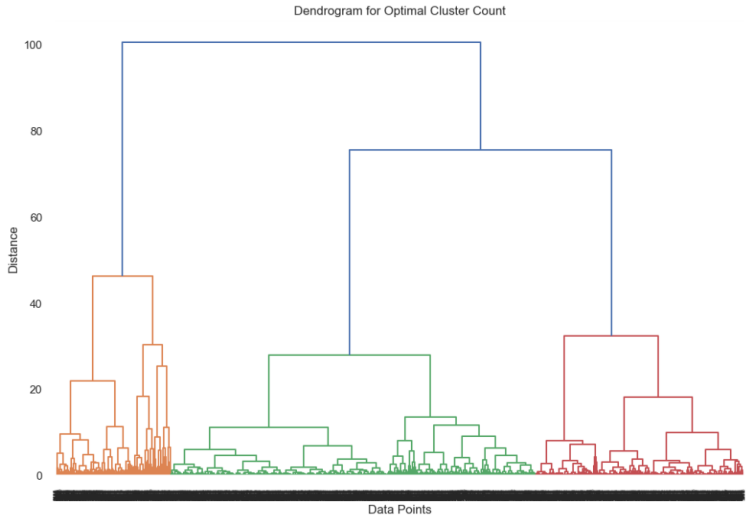

Figure 15: Dendrogram Analysis of Agglomerative Clustering

To determine the optimal cluster count, it became evident that the dataset has four clusters in Agglomerative modelling as depicted in Figure 15. Each data point was given a cluster label using the pre-processed dataset, enhancing it with cluster knowledge. Graphical depictions were produced to demonstrate the distribution of the clustered data. The clusters were visualised through a 3D scatter plot portrayed in Figure 16, that shows four clusters in different colours. Different customer groups can be identified from these plotted clusters. For example, the top-left cluster, which is high in frequency and recency but low in monetary value, represents engaged customers who might benefit from encouragement to spend more. Meanwhile, customers in the bottom-right cluster are valuable but not frequent that they may require re-engagement efforts.

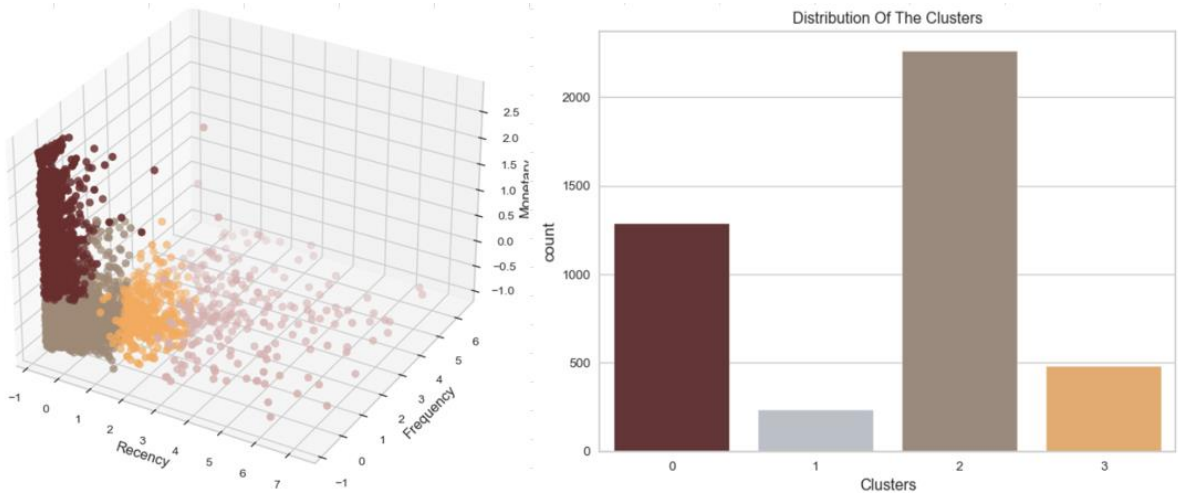

Figure 16: 3D Plotting of Clusters and RFM Dimensions

Furthermore, the clustering results were visualised using a scatterplot shown in Figure 17, which displays the distribution of data points based on monetary and frequency aspects. This graphic provides a thorough knowledge of the segmentation that the agglomerative clustering algorithm was able to accomplish by helping to comprehend the various expenditure and income characteristics of the clusters to meet the unique requirements of every consumer group.

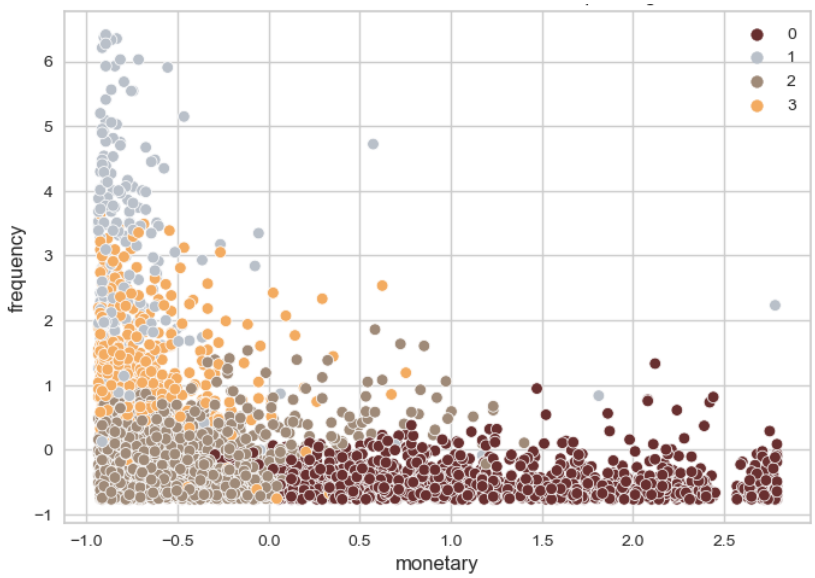

Figure 17: Clusters Profile Based on Income and Spending

5.2.2 K-Means Clustering

K-Means clustering describes an algorithm that organizes data points into clusters based on similarities in their features. It's crucial for customer segmentation by facilitating businesses to categorize their customer base dividing into distinct clusters with similar behaviours, preferences, or characteristics. In this section, an initial arbitrary number of clusters (k=3) was chosen to explore inherent patterns using Elbow method and the sum of squared distances (SSD) was computed for varying numbers of clusters aiding in identifying an optimal cluster count depicted in Figure 18. After the SSD analysis, silhouette analysis was performed to evaluate clustering quality for each k value.

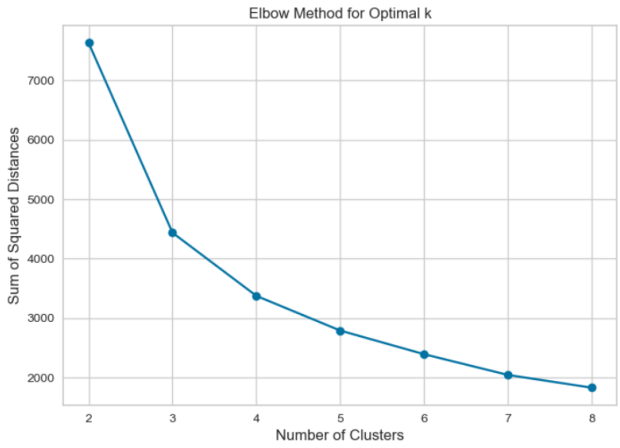

Figure 18: Elbow Method Analysis

For tuning the model, K-Means clustering algorithm was used in conjunction with silhouette analysis to discover the ideal number of clusters in the dataset. By using this technique, every cluster arrangement was evaluated, and the silhouette score was calculated. More distinct clusters are indicated by higher scores. With an emphasis on improved intra and inter cluster separation for improved clustering quality, this iterative evaluation made it easier to determine the optimal cluster count.

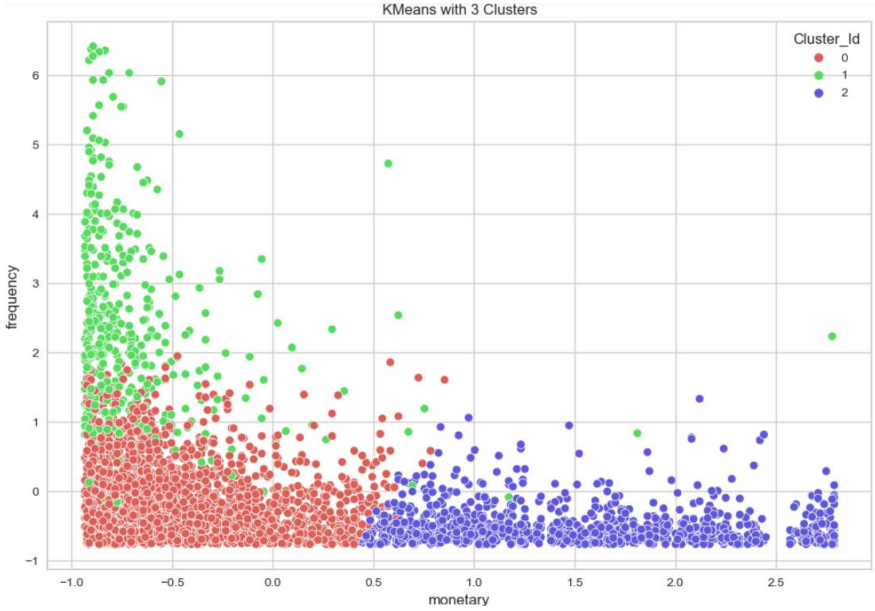

Figure 19: K-Means Clustering Results

Visual representations of the K-Means clustering outcomes were depicted in Figure 19, showcasing relationships between scaled RFM features colour coded by three different cluster assignment. The K-Means clustering analysis facilitated the identification and understanding of inherent structures within the dataset, enabling the discovery of distinct customer segments.

5.2.3 Mean-shift Clustering

The Mean-shift clustering algorithm was implemented to identify clusters within the dataset based on density estimation. In customer segmentation, Mean-shift clustering serves as a powerful tool to categorize customers using their distinctive attributes or purchasing habits. Based on density estimation, Mean-Shift is a distribution independent clustering method that works well in situations where the number of clusters is unknown because it doesn't involve predetermining the number of clusters. In the implementation process, Mean-shift was applied by iteratively shifting data samples towards the peak density within their local distribution until convergence, thereby delineating clusters based on high-density regions in the feature space. Cluster labels were applied to the data points in a scatter plot illustrated in Figure 20, which allowed the monetary and frequency dimensions to be seen clearly and without the requirement for predetermined cluster numbers.

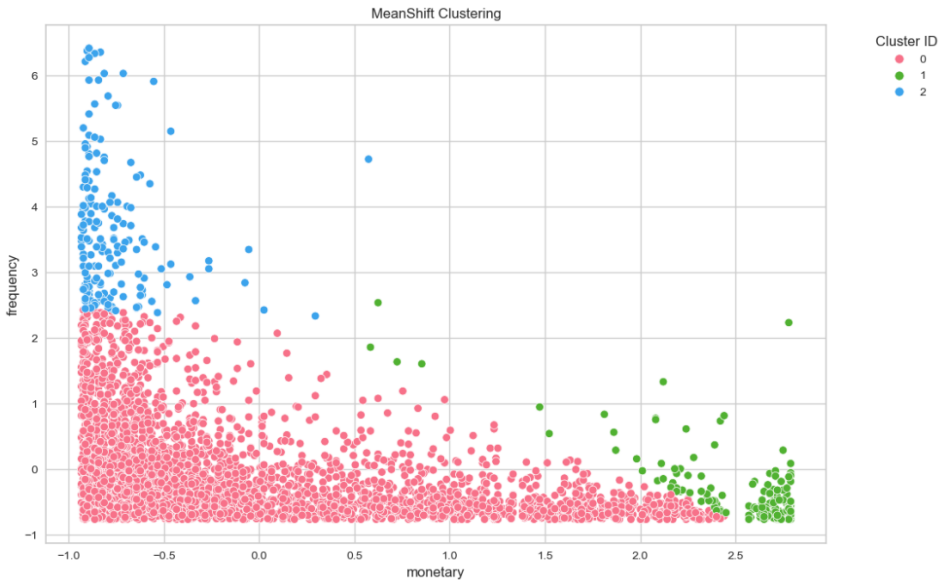

Figure 20: Mean-shift Clustering Results

6 Evaluation

This section of research evaluates the clustering models by comprehensive analysis of the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index as metrics. These metrics offer a strong, label-independent evaluation that is essential for unsupervised RFM segmentation. They also provide information about the applicability of individual data points and the general quality of the clustering.

6.1 Silhouette Score

This metric evaluates how well customers within segments share similarities while maintaining distinctions from other segments. The Silhouette score solutions cluster coherence and segregation, creating internally consistent and individual customer groups formed by transactional patterns. Scores range from -1 to 1, where greater values signify more defined clusters.

6.2 Calinski-Harabasz Index

Calculating the variance relationship between clusters, this index assesses the clarity of boundaries between customer segments. A higher score indicates well-defined and separate segments, crucial for creating distinct groups reflective of unique transactional attributes. Higher scores indicate clearer cluster boundaries and better clustering quality.

6.3 Davies-Bouldin Index

Measuring average correlation of clusters, Davies-Bouldin index evaluates the balance within clusters coherence and separation. Lower scores indicate clear and internally cohesive segments with minimal overlap in transactional behaviours, essential for describing distinct customer groups. Lower scores signify better cluster separation and distinctiveness.

6.4 Discussion

The Table 2 given below presents an insightful evaluation of three clustering models, Agglomerative, K-Means, and Mean-shift based on their performance across Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index metrics. Impressively, the Agglomerative clustering model is the clear winner, demonstrating remarkable strength with the greatest Calinski-Harabasz Index (5064), highest Silhouette Score (0.58) and Davies-Bouldin Index (0.78). These results highlight how well this clustering model can classify related data points and identify different clusters within the dataset. In contrast, even though the K-Means clustering model performs admirably, it is not as effective as the Agglomerative model. As seen by its Silhouette Score of 0.51, Calinski-Harabasz Index of 3574, and Davies-Bouldin Index of 0.86, the K-Means model is reasonably effective at grouping data points into clusters and differentiating between them, though not as effectively as the Agglomerative model. Meanwhile, the Mean-shift clustering model has the lowest results (Silhouette Score of 0.35, Calinski-Harabasz Index of 937, Davies-Bouldin Index of 0.71), indicating significantly worse performance on all criteria.

Clustering Model	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
Agglomerative	**0.58**	**5064**	**0.78**
K-Means	**0.51**	**3574**	**0.86**
Mean-shift	**0.35**	**937**	**0.71**

Table 2: Performance Evaluation Metrics

In the interpretation of results, the Agglomerative clustering model notably spotlight the influence of monetary features, potentially overshadowing recency, and frequency metrics. This clustering model identified distinct customer clusters with unique behavioural traits depicting customers with similar spending behaviours. The clustering results indicate interesting customer segments for which a range of marketing strategies could be used, one group of consumers shops more often but spends less. Presumably, they consistently choose inexpensive goods. Another group of consumers shops infrequently and spends little money. Conversely, the final group consists of consumers who spend a lot of money but are not regular shoppers.

This summary highlights the Agglomerative clustering model's tendency to heavily consider the customers spending feature in forming clusters, while also acknowledging the strengths of the K-Means model in describing distinct customer segments based on shopping behaviours. Overall, on the thorough analysis the Agglomerative clustering model is the best option for clustering this dataset, consistently outperforming both K-Means and Mean-shift models.

7 Conclusion and Future Work

This research used RFM analysis on transactional data obtained from a UK-based company to investigate consumer segmentation in an online retail dataset. The investigation started with extensive data preprocessing and feature engineering, which established a solid basis for clustering analysis. The subsequent exploratory data analysis revealed subtle trends, demonstrating significant variations in sales throughout the day and unique transaction quantities, which were especially noticeable on Sundays and in the morning. Utilising a range of visualisation methods, such as word cloud analyses of product descriptions and RFM feature representations, provided deep insights into client preferences and changing sales trends. After applying various clustering algorithms, Agglomerative clustering was the most effective approach for clustering this dataset. However, it showed a strong preference for the monetary feature, which may have overshadowed the importance of frequency and recency metrics in the clustering process. By identifying high-value customer categories, personalising interactions, and optimising marketing efforts, firms can ultimately encourage increased customer engagement and enhanced business outcomes with this method.

As part of future research, to refine the clustering process future endeavours could focus on mitigating the dominance of the monetary feature observed in Agglomerative clustering. Exploring techniques for feature selection or employing algorithms that balance the influence of singular dominant features could enhance the accuracy and reliability of segmentation. Furthermore, advanced clustering methods like ensemble clustering or subspace methods might offer deeper insights into complex customer behaviours.

References

Kabasakal, İ., 2020. Customer segmentation based on recency frequency monetary model: A case study in E-retailing. *Bilişim Teknolojileri Dergisi*, *13*(1), pp.47-56.

Abbasimehr, H. and Shabani, M., 2021. A new methodology for customer behavior analysis using time series clustering: A case study on a bank's customers. *Kybernetes*, *50*(2), pp.221-242.

Sabuncu, İ., Turkan, E. and Polat, H., 2020. Customer segmentation and profiling with RFM analysis. *Turkish Journal of Marketing*, *5*(1), pp.22-36.

Ernawati, E., Baharin, S.S.K. and Kasmin, F., 2021, April. A review of data mining methods in RFM-based customer segmentation. In *Journal of Physics: Conference Series* (Vol. 1869, No. 1, p. 012085). IOP Publishing.

Smaili, M.Y. and Hachimi, H., 2023. New RFM-D classification model for improving customer analysis and response prediction. *Ain Shams Engineering Journal*, p.102254.

Rahim, M.A., Mushafiq, M., Khan, S. and Arain, Z.A., 2021. RFM-based repurchase behavior for customer classification and segmentation. *Journal of Retailing and Consumer Services*, *61*, p.102566.

Aggarwal, A.G. and Yadav, S., 2020, June. Customer segmentation using fuzzy-AHP and RFM model. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 77-80). IEEE.

Firdaus, U. and Utama, D., 2021. development of bank's customer segmentation model based on rfm+ b approach. *Int. J. Innov. Comput. Inf. Cont*, *12*(1), pp.17-26.

Abdulhafedh, A., 2021. Incorporating k-means, hierarchical clustering and pca in customer segmentation. *Journal of City and Development*, *3*(1), pp.12-30.

Zhu, C., Idemudia, C.U. and Feng, W., 2019. Improved logistic regression model for diabetes prediction by PCA and K-means techniques. *Informatics in Medicine Unlocked*, *17*, p.100179.

Wu, J., Shi, L., Lin, W.P., Tsai, S.B., Li, Y., Yang, L. and Xu, G., 2020. An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm. *Mathematical Problems in Engineering*, *2020*, pp.1-7.

Anitha, P. and Patil, M.M., 2022. RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), pp.1785-1792.

Mamashli, Z. and Zolfani, S.H., 2022. Customer Segmentation Based on Mobile Banking User's Behavior. *Int. J. Mechatron. Electr. Comput. Technol*, *12*, pp.5267-5276.

Mensouri, D., Azmani, A. and Azmani, M., 2022. K-Means customers clustering by their RFMT and score satisfaction analysis. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *13*(6).

Barus, O.P., Nathasya, C. and Pangaribuan, J.J., 2023. The Implementation of RFM Analysis to Customer Profiling Using K-Means Clustering. *Mathematical Modelling of Engineering Problems*, *10*(1).

Monalisa, S., Juniarti, Y., Saputra, E., Muttakin, F. and Ahsyar, T.K., 2023. Customer segmentation with RFM models and demographic variable using DBSCAN algorithm. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *21*(4), pp.742-749.

Hossain, A.S., 2017, December. Customer segmentation using centroid based and density based clustering algorithms. In *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)* (pp. 1-6). IEEE.

Peker, S., Kocyigit, A. and Eren, P.E., 2017. LRFMP model for customer segmentation in the grocery retail industry: a case study. *Marketing Intelligence & Planning*, *35*(4), pp.544-559.

Plotnikova, V., Dumas, M. and Milani, F., 2020. Adaptations of data mining methodologies: A systematic literature review. *PeerJ Computer Science*, *6*, p.e267.

Varadarajan, R., 2020. Customer information resources advantage, marketing strategy and business performance: A market resources based view. *Industrial Marketing Management*, *89*, pp.89-97.

Wan, S., Deng, J., Gan, W., Chen, J. and Philip, S.Y., 2022, October. Fast Mining RFM Patterns for Behavioral Analytics. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 1-10). IEEE.

https://www.slideserve.com/moana-herring/overview-of-data-mining-the-knowledge-discovery-process

https://archive.ics.uci.edu/ml/datasets/online+retail

https://scikit-learn.org/stable/modules/clustering.html

Examiner Question and Answers:

Q1. Can you elaborate on which RFM features (Recency, Frequency, Monetary Value) individually or combined contribute most to distinct customer segments?

Answer: Using RFM analysis and clustering methods to investigate customer segmentation, the analysis revealed a notable inclination towards the significance of Monetary Value over Recency and Frequency features in forming distinct customer groups. This analysis revealed that spending behaviour held greater way in forming clusters and therefore, spending patterns had a greater influence on clusters than the holistic RFM spectrum. The customer segments that were found, such as regular yet low spending customers, infrequent spenders, and big spenders but infrequently, highlighted the importance of Monetary value in creating these groups. While Monetary value was the focus, Recency and Frequency features are still crucial to understanding consumer loyalty even though their impact on clustering is less noticeable. Future efforts should focus on normalising the importance of these RFM features, equalising their influence, or investigating other clustering approaches for more focused marketing strategies to generate more perceptive and balanced customer segments. Ultimately, within the RFM framework, Monetary Value emerged as the most contributor to identifying distinct customer segments.

Q2. How do customer segments identified through RFM-added clustering techniques using different algorithms evolve over time? Are there seasonal or periodic changes in customer behaviour within these segments?

Answer: The customer groups identified through RFM-based clustering methods using different algorithms demonstrate unique evolutionary trends over time. Customer behaviour in these categories varies seasonally and on a regular basis. For example, all segments show a noticeable decrease in transactions on Sundays, suggesting a general decline in activity during that day. Additionally, there is a notable increase in sales in the morning hours, indicating a common behavioural characteristic between these segmented groups. Fascinatingly, customer sections exhibit a range of behaviours for example, a group that spends less but frequently exhibits consistent behaviour over time. However, the cluster of high spending but less frequent buyers display fluctuations, notably increasing in certain periods, possibly influenced by seasonal promotions or specific events. When examined over time, these unique patterns reveal suggestive differences in consumer behaviour among different customer segments.

Q3. Briefly present the limitations of your research work.

Answer: While the research work on customer segmentation and clustering using RFM analysis is comprehensive and insightful, a few limitations encountered throughout its process. Firstly, even if the dataset was well cleaned and pre-processed, the removal of outliers using the interquartile range might have impacted the dataset, potentially eliminating important transactional information that could have provided insights into customer

behaviour at the extremes. Additionally, focusing solely on transactional data limits the understanding of customer behaviour beyond purchase patterns that could enhance segmentation accuracy. The focus on the RFM analysis may have unintentionally introduced bias, particularly during the clustering stage when the monetary feature appeared to outweigh the recency and frequency features. Although the clustering algorithms performed well, the dominance of the monetary feature in the agglomerative clustering suggests a potential bias of this feature on the segmentation process. Lastly, while the segmentation offers valuable insights, implementing these segments in marketing strategies might require further validation through targeted campaigns to truly assess their effectiveness in driving customer engagement and sales.

Q4. Justify why those models were investigated in your research.

Answer: The decision to explore Agglomerative, K-Means, and Mean Shift clustering models in this research was driven by their unique strengths and significance in the context of customer segmentation using RFM analysis within an online retail dataset. Agglomerative clustering was selected for its hierarchical nature and flexibility in handling diverse cluster shapes, aligning well with the goal of pinpointing specific customer segments through RFM analysis. Initially, it displayed potential in crucial assessment criteria such as the Silhouette Score and Calinski-Harabasz index. K-Means, known for its simplicity and ease of interpretation, was valued for its ability to create easily understandable customer segments based on shopping frequency and spending habits, resulting in three distinct clusters. Mean Shift, while not as prominent in performance, was included for its capability to detect clusters without predefined shapes or numbers, offering potential insights into underlying structures within the data. The justification behind exploring these models was to thoroughly evaluate their capacity in capturing involved customer behaviour patterns, with the aim of identifying robust and actionable segments essential for targeted marketing strategies in the online retail domain.