# Abstract

With the rapid growth of e-commerce, businesses must adopt more advanced strategies to better understand customer behavior and improve segmentation. This dissertation investigates the application of techniques such as Recency, Frequency, and Monetary (RFM) analysis, Customer Lifetime Value (CLV) prediction, and hybrid clustering models to improve customer segmentation in e-commerce. Due to privacy concerns, real customer data was not available, so synthetic data was generated using the Python Faker library. This helped the creation of a dataset that closely matched real-world e-commerce behaviours, such as customer demographics, transactions, and behavioural metrics.

Customers were segmented using K-Means and DBSCAN clustering algorithms based on RFM scores and behavioral data. This hybrid approach helped identify distinct customer segments while successfully dealing with outliers and noise. Furthermore, geospatial analysis was used to map customer locations, revealing regional trends that can be used in more targeted marketing strategies. To tie everything together, a prototype application was created using Django and React, allowing users to interactively visualize customer segments and geospatial insights, demonstrating how these methods can be used in real-world settings.

Performance metrics such as the Silhouette Score, Davies-Bouldin Index, and regression evaluation metrics were used to evaluate how well the clustering models and CLV predictions worked. The results of this research provide valuable insights for e-commerce businesses looking to improve their marketing strategies, increase customer retention, and increase profitability through advanced customer segmentation and behavioral analysis techniques. The use of synthetic data ensures that these methods can be tested in a privacy-preserving manner, with the possibility of future real-world applications.

# Acknowledgment

I would like to express my heartfelt gratitude to my supervisor, Dr David Lightfoot, for his support, guidance, and patience throughout this dissertation. His encouragement during our weekly meetings played an important role in keeping me motivated. His expertise, along with his attention to detail, and his feedback, especially the one I got during the project presentation helped me with great insights to complete my dissertation. I truly couldn’t have done it without him.

I also want to extend a huge thank you to my friends and course mates. Their constant support, encouragement, and help made my journey so much easier. Whether they offered technical advice, or motivation, or were simply there when I needed someone to lean on, made a lot of difference.

Thank you all for being a part of this achievement, it means the world to me.

# Table of Content

[List of Figures 6](#_Toc178121687)

[Abbreviation Used 7](#_Toc178121688)

[1. Introduction 8](#_Toc178121689)

[1.1. Aim 8](#_Toc178121690)

[1.2. Objectives 8](#_Toc178121691)

[1.2.1. Simulation Of Data Using the Python "Faker" Library 9](#_Toc178121692)

[1.2.2. Conducting RFM Analysis to Identify Customer Segments 9](#_Toc178121693)

[1.2.3. Predicting Customer Lifetime Value (CLV) 9](#_Toc178121694)

[1.2.4. Implementation Of Hybrid Clustering Techniques 9](#_Toc178121695)

[1.2.5. Performing Geospatial Analysis to Understand Regional Trends 9](#_Toc178121696)

[1.2.6. Development Of a Prototype Application 10](#_Toc178121697)

[1.3. Methods And Tools Used 10](#_Toc178121698)

[2. Literature Review 12](#_Toc178121699)

[2.1. Data Generation 12](#_Toc178121700)

[2.2. Multi-Source Data Integration 12](#_Toc178121701)

[2.3. RFM Analysis 13](#_Toc178121702)

[2.4. Customer Lifetime Value (CLV) Prediction 14](#_Toc178121703)

[2.5. Hybrid Clustering Techniques 14](#_Toc178121704)

[2.6. Geospatial Analysis 15](#_Toc178121705)

[3. Methodology 17](#_Toc178121706)

[3.1. Approach 17](#_Toc178121707)

[3.2. Data Generation 18](#_Toc178121708)

[3.2.1. Customer Data Creation 18](#_Toc178121709)

[3.2.2. Transaction Data Creation: 20](#_Toc178121710)

[3.2.3. Behavioral Data Creation 21](#_Toc178121711)

[3.2.4. Data Validation and Matching 21](#_Toc178121712)

[3.3. Tools, Technologies, And Software Used for Development Process 22](#_Toc178121713)

[3.3.1. Development And Data Handling 22](#_Toc178121714)

[3.3.2. Machine Learning and Clustering 22](#_Toc178121715)

[3.3.3. Data Visualization 23](#_Toc178121716)

[3.3.4. Backend Development 23](#_Toc178121717)

[3.3.5. Frontend Development 23](#_Toc178121718)

[3.3.6. Development Environment 24](#_Toc178121719)

[3.4. Version Management 24](#_Toc178121720)

[3.4.1. Github For Code Versioning: 24](#_Toc178121721)

[3.4.2. Google Drive and Docs for Document Management: 24](#_Toc178121722)

[4. Development Process, Results & Evaluation 25](#_Toc178121723)

[4.1. Recency, Frequency, And Monetary (RFM) Analysis 25](#_Toc178121724)

[4.1.1. Data Preparation and Steps for RFM Analysis 25](#_Toc178121725)

[4.1.2. RFM Scoring 26](#_Toc178121726)

[4.1.3. Customer Segmentation Using RFM Scores 27](#_Toc178121727)

[4.1.4. Customer Scoring 28](#_Toc178121728)

[4.1.5. Data Visualization of The RMA Analysis 29](#_Toc178121729)

[4.2. Customer Lifetime Value (CLV) Prediction 33](#_Toc178121730)

[4.2.1. Data Preparation 34](#_Toc178121731)

[4.2.2. Train-Test Split 34](#_Toc178121732)

[4.2.3. Prediction And Evaluation 34](#_Toc178121733)

[4.3. K-Means and DBSCAN Clustering 35](#_Toc178121734)

[4.3.1. Merging RFM And Behavioral Dataframes 35](#_Toc178121735)

[4.3.2. Standardizing The Features: 36](#_Toc178121736)

[4.3.3. Dimensionality Reduction Using PCA: 36](#_Toc178121737)

[4.3.4. Elbow Method for Optimal K: 36](#_Toc178121738)

[4.3.5. Visualizing Clusters After PCA: 37](#_Toc178121739)

[4.3.6. High-Value Customer Visualization: 38](#_Toc178121740)

[4.3.7. Cluster Distribution Based on Recency, Frequency, And Monetary 39](#_Toc178121741)

[4.3.8. DBSCAN Clustering: 41](#_Toc178121742)

[4.3.9. Outlier Detection With DBSCAN: 42](#_Toc178121743)

[4.3.10. DBSCAN Outlier Detection for High-Value Customers Only 43](#_Toc178121744)

[4.3.11. Silhouette Score and Davies-Bouldin Index: 43](#_Toc178121745)

[4.4. Geographic Visualization 45](#_Toc178121746)

[4.4.1. Converting Customer Data into Geospatial Points 45](#_Toc178121747)

[4.4.2. Identifying High-Value Customers 45](#_Toc178121748)

[4.4.3. Data Validation 45](#_Toc178121749)

[4.4.4. Interactive Map Creation 45](#_Toc178121750)

[4.4.5. Highlighting Customer Segmentation in The Map 46](#_Toc178121751)

[4.5. Creation Of Prototype Application 48](#_Toc178121752)

[4.5.1. Backend Functionality 48](#_Toc178121753)

[4.5.2. Frontend Functionality: 49](#_Toc178121754)

[4.5.3. Deployment And Testing: 53](#_Toc178121755)

[5. Evaluation and Discussion 54](#_Toc178121756)

[5.1. Dissertation Results and Achievement 54](#_Toc178121757)

[5.2. Assumptions, Limitations, and Recommendations 55](#_Toc178121758)

[5.2.1. Assumptions 55](#_Toc178121759)

[5.2.2. Limitations And Recommendations 55](#_Toc178121760)

[5.3. Academic Achievement 57](#_Toc178121761)

[5.4. Project Management 57](#_Toc178121762)

[5.5. Relevant Legal, Social, Ethical, And Professional Issues 58](#_Toc178121763)

[5.6. Risk Management 59](#_Toc178121764)

[5.7. Future Work 60](#_Toc178121765)

[6. Conclusion 62](#_Toc178121766)

[7. References 63](#_Toc178121767)

# List of Figures

[Figure 1: Methodology Flowchart 17](#_Toc177929317)

[Figure 2: Customer Age Distribution 22](#_Toc177929318)

[Figure 3: Customer Genderr Distribution 22](#_Toc177929319)

[Figure 4: Transaction Amount Distribution 23](#_Toc177929320)

[Figure 5: Rfm Scores Dataframe 26](#_Toc177929321)

[Figure 6: Segment Distribution 27](#_Toc177929322)

[Figure 7: Customer Segment Distribution 29](#_Toc177929323)

[Figure 8: Customer Grouping Based On Recency 30](#_Toc177929324)

[Figure 9: Frequency Distribution 31](#_Toc177929325)

[Figure 10: Monetary Distribution 31](#_Toc177929326)

[Figure 11: Roc Curve And Auc Analysis 32](#_Toc177929327)

[Figure 12: Clv Prediction Values 34](#_Toc177929328)

[Figure 13: Mae And R-Squared 34](#_Toc177929329)

[Figure 14: Elbow Method For Optimal K 36](#_Toc177929330)

[Figure 15: Visualization Of Clusters 37](#_Toc177929331)

[Figure 16: High-Value Customer Visualization 37](#_Toc177929332)

[Figure 17: Recency Vs K-Means 38](#_Toc177929333)

[Figure 18: Frequency Vs K-Means 39](#_Toc177929334)

[Figure 19: Monetary Vs K-Means 39](#_Toc177929335)

[Figure 20: Dbscan Clustering 40](#_Toc177929336)

[Figure 21: Noise Identified By Dbscan 41](#_Toc177929337)

[Figure 22: Outliers From High-Value Customers 41](#_Toc177929338)

[Figure 23: Silhouette Score And Davies-Bouldin Index For K-Means 42](#_Toc177929339)

[Figure 24: Silhouette Score And Davies-Bouldin Index For Dbscan 43](#_Toc177929340)

[Figure 25: Snapshot Of Interactive Map 44](#_Toc177929341)

[Figure 26: Highlighting Customer Segmentation 45](#_Toc177929342)

[Figure 27: Frontend Dashboard 48](#_Toc177929343)

[Figure 28: Frontend Dashboard - Customer Search 48](#_Toc177929344)

[Figure 29: Frontend Dashboard - Customer Details Displayed 49](#_Toc177929345)

[Figure 30: Frontend Dashboard - Customer Segmentation Bar Plot 50](#_Toc177929346)

[Figure 31: Frontend Dashboard - Map Of Customer Location Distribution 50](#_Toc177929347)

# Abbreviation Used

RFM: Recency, Frequency, and Monetary

CLV: Customer Lifetime Value

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

PCA: Principal Component Analysis

MAE: Mean Absolute Error

R²: R-squared (Coefficient of Determination)

API: Application Programming Interface

KPI: Key Performance Indicator

ROC: Receiver Operating Characteristic

AUC: Area Under the Curve

ORM: Object-Relational Mapping

GIS: Geographic Information Systems

DRF: Django REST Framework

CSV: Comma-Separated Values (file format)

HTML: Hypertext Markup Language

GDPR: General Data Protection Regulation

Keywords: Hybrid Clustering, Geospatial Analysis, Synthetic Data, RFM Scoring, Regression Modelling, Silhouette Score, Davies-Bouldin Index, Elbow Method, Geospatial Points, Interactive Map.

# Introduction

This dissertation focuses on using advanced, data-driven e-commerce strategies to better understand and segment customers based on their behavior. Yıldız, Güngör Şen, and Işık (2023) argue that traditional approaches to customer segmentation, such as using basic demographics or purchase history, are insufficient to capture the complexities of modern consumer behavior. To address this, the research employs techniques such as RFM (Recency, Frequency, Monetary) analysis, machine learning models that predict Customer Lifetime Value (CLV), and a combination of clustering methods such as K-Means and DBSCAN. Geospatial analysis is also used to reveal regional shopping or behavioural trends.

The ultimate goal of this dissertation is to not only improve segmentation accuracy but also assist businesses in allocating marketing resources more efficiently, increasing customer engagement, and ensuring privacy compliance through the use of synthetic data. Finally, these strategies seek to increase customer satisfaction, retention, and profits for businesses.

## Aim

This dissertation aims to develop and apply advanced, data-driven techniques to improve customer segmentation and behavior analysis in e-commerce. By combining methods such as RFM analysis, Customer Lifetime Value (CLV) prediction, hybrid clustering algorithms, and geospatial analysis, the study hopes to provide businesses with deeper insights into their customers, allowing for more targeted and effective marketing strategies. The project will also involve creating a prototype to demonstrate how these techniques work in reality, all while adhering to ethical and privacy standards.

## Objectives

The main goal of this dissertation is to develop a clear framework for using modern data analytics to improve customer segmentation and better understand customer behavior, thus improving e-commerce strategies. The goal of combining different methods is to gain deeper insights into consumer behavior, allowing businesses to develop more targeted and effective marketing plans.

### Simulation Of Data Using the Python "Faker" Library

To test the proposed methods, realistic customer data will be simulated using the Python Faker library. This helps in the creation of datasets that mimic real-world scenarios while maintaining privacy and security. It also provides a flexible testing environment that's applicable to different e-commerce situations.

### Conducting RFM Analysis to Identify Customer Segments

RFM (Recency, Frequency, Monetary) analysis will be carried out to group customers based on their buying behavior. By looking at how recently customers made a purchase, how often they shop, and how much they spend, businesses can categorize customers into different segments, leading to more focused marketing efforts.

### Predicting Customer Lifetime Value (CLV)

Customer Lifetime Value (CLV) will be predicted using a linear regression model to estimate how much each customer is likely to spend in the future. This method combines RFM metrics with behavioral data to forecast future spending, helping businesses better allocate their marketing resources. The model's accuracy will be assessed using Mean Absolute Error (MAE) and R-squared (R²) metrics.

### Implementation Of Hybrid Clustering Techniques

A hybrid clustering model, combining K-Means and DBSCAN, will be developed to achieve more accurate customer segmentation. K-Means works well for structured data, while DBSCAN excels at detecting clusters of irregular shapes and handling noise in the dataset. The performance of this model will be evaluated using cluster validity indices like the silhouette score and Davies-Bouldin index, ultimately improving the accuracy of customer segmentation.

### Performing Geospatial Analysis to Understand Regional Trends

Geospatial analysis will be used to map customer data, highlighting regional trends and visualizing how preferences vary by location. By integrating geospatial data with RFM and CLV metrics, this research aims to develop marketing strategies that are tailored to specific regions based on local customer behaviours and preferences.

### Development Of a Prototype Application

A prototype application will be built using React and Django to demonstrate the practical application of the proposed methods. Users will be able to input customer data and receive detailed reports on segmentation and behavioral analysis, highlighting the real-world potential and value of these techniques for businesses.

## Methods And Tools Used

Python & Libraries (pandas, NumPy, Scikit-learn): Python was chosen for its flexibility in handling large datasets and its powerful libraries, which are ideal for data manipulation, machine learning, and statistical analysis. (Pedregosa *et al.*, 2011)

1. RFM Analysis: A straightforward but effective method for segmenting customers based on their purchasing behavior. It gives businesses a clear view of their high-value and at-risk customers while requiring minimal computational resources. (Yıldız, Güngör Şen and Işık, 2023)
2. K-Means Clustering: Selected for its efficiency in segmenting large datasets by grouping customers with similar traits. It’s great for identifying distinct groups, like frequent buyers. (Joung and Kim, 2023)
3. DBSCAN: Used alongside K-Means to manage noisy or irregular customer data. DBSCAN detects outliers and identifies clusters without needing a predefined number of groups, making it a useful complement to K-Means. (John, Shobayo and Ogunleye, 2023)
4. CLV Prediction (Linear Regression): This is a predictive method used to estimate future customer spending. Linear regression was chosen for its simplicity and easy interpretability, which makes it practical for business decision-making. (Sharma, Chakraborty and Sanyal, 2019))
5. Geospatial Analysis (GeoPandas, Folium): These tools allow for mapping customer data and analyzing regional trends. This helps businesses tailor their marketing strategies to specific locations based on customer distribution.
6. Prototype Development (React, Django): React was selected for building an interactive and user-friendly interface, while Django provides a secure and scalable backend for managing customer data effectively.

The methodologies used in this dissertation are rooted in well-established theories from both marketing and data science. RFM analysis, a core component of customer segmentation in this study, is derived from Customer Relationship Management (CRM) theory, which highlights the importance of monitoring customer behavior over time. The Pareto Principle (80/20 rule) further justifies this approach by recognizing that a small percentage of high-value customers often account for the majority of a company’s revenue. (Singh and Singh, 2017)

As per (Singh and Singh, 2017), while traditional CRM methods are effective in simpler cases, they struggle to manage and analyse large, complex datasets. This is where machine learning techniques like K-Means and DBSCAN come into play. These algorithms are integrated into the analysis to handle more extensive and complicated customer data, providing deeper insights and more accurate segmentation than what can be achieved with conventional CRM tools alone.

# Literature Review

As e-commerce continues to expand, businesses are turning more and more to data-driven strategies for improving customer segmentation, predicting Customer Lifetime Value (CLV), and refining their marketing efforts. Traditional methods, like RFM analysis and basic clustering, often fall short when it comes to capturing the complexity of modern consumer behavior. This dissertation tackles those limitations by incorporating hybrid clustering techniques, using synthetic data to ensure privacy, and integrating multiple data sources, including behavioral and geospatial information.

## Data Generation

Real customer data is commonly used in e-commerce analytics, but this brings up significant security and privacy concerns. As a result, many researchers turn to tools that generate synthetic data. One of the most popular libraries in Python for creating realistic but anonymized datasets is "Faker," as highlighted by Carrola (2022). These synthetic datasets allow researchers to evaluate and validate data models without compromising security, as they mimic real-world consumer behaviours while adhering to privacy regulations. For example, Sharaf Addin (2022) used the Faker library to create telecom customer datasets, enabling the analysis of both behavioral and demographic data. These approaches are increasingly being adopted in various industries, including e-commerce and healthcare (Tang et al., 2023).

For my dissertation, datasets are being generated that mimic real consumer behavior using the "Faker" library. These datasets will include purchase history, demographic details, and behavioral metrics such as time spent on-site and engagement levels. By creating realistic datasets, the complexities of customer behavior can be captured that are needed for accurate predictions, all while ensuring compliance with security and privacy regulations (Wei et al., 2023). This synthetic data allows for testing advanced segmentation and predictive models by closely resembling real customer transactions and interactions. As a result, the conclusions of this dissertation can be confidently applied to real-world e-commerce environments. This approach achieves the goals of data-driven customer behavior analysis while maintaining strict ethical standards.

## Multi-Source Data Integration

Conventional customer segmentation models primarily rely on transactional data, often missing valuable insights from other sources such as social media activity, browsing habits, and geographic information. As Shen (2021) points out, many segmentation techniques are limited by the amount of data they can process, which means they often overlook crucial details that capture the full complexity of modern consumer behavior. I'm suggesting integrating multiple data sources such as transactional, behavioral, and geographic data is being suggested, which will offer deeper insights by drawing on various types of data, enabling the creation of detailed customer profiles that go beyond what traditional segmentation methods can achieve (Mandal, 2022)

A study by Yıldız (2023) demonstrates how combining additional behavioral data with customer segmentation based on RFM analysis can lead to hyper-personalized product recommendations. Research has shown that factors such as perceived value and self-efficacy significantly influence consumer engagement in e-commerce (Cao et al., 2022). By incorporating extra metrics like time spent on-site and customer engagement, this approach enables the development of more sophisticated segmentation models and provides valuable insights for tailored marketing strategies. As a result, customers are more likely to make repeat purchases and recommend the brand to others, which in turn boosts overall customer engagement (Maria, Wijaya and Keni, 2021).

## RFM Analysis

RFM analysis is easy to use and good at classifying customers according to their past transactions or purchase behavior, it has long been the preferred technique for customer segmentation (Safari, Safari and Montazer, 2016). RFM analysis involves scoring customers based on three factors, recency (how recently a customer made a purchase), frequency (how often they purchase), and monetary value (how much they spend). Studies by Pengfei Li (2022) show how algorithms such as K-Means can be added to RFM analysis to increase segmentation accuracy. But by concentrating only on purchase metrics, traditional RFM analysis oversimplifies customer behavior and ignores other crucial elements like engagement and location (Bachtiar 2018). The RFM analysis in this dissertation with extra behavioral data, like the number of logins, pages viewed, and amount of time spent on the website, etc. The RFM model's segmentation becomes more accurate in customer engagement patterns by integrating these behavioral metrics. Furthermore, the fusion of RFM with machine learning algorithms such as K-Means and DBSCAN improves the accuracy of customer segmentation based on multi-dimensional behavior (Monalisa and Kurnia, 2019).

By providing a more comprehensive understanding of consumer behavior, this method enables companies to create more individualized marketing plans. By incorporating proficient clustering algorithms, businesses can achieve more precise and useful segmentation, which improves their ability to target high-value clients and maximize their advertising budgets. (Parikh and Abdelfattah, 2020 )

## Customer Lifetime Value (CLV) Prediction

CLV prediction has been used to estimate the future value that a customer will provide to the company, assisting businesses in allocating their marketing budgets. (Marmolet al., 2021). Research by Ankit Kumar (2023) shows how RFM analysis combined with Pareto/NBD models can forecast CLV based on previous purchasing behavior. However, the majority of current models are constrained by their exclusive focus on purchase history and neglect to include behavioral engagement metrics, which have the potential to enhance the precision of these predictions (Chen, Reibman and Arora, 2022). This dissertation will add behavioral data like the frequency of logins, length of sessions, and pages visited the conventional CLV model. Furthermore, machine learning models such as ensemble approaches are utilized to enhance the accuracy of CLV prediction. Ejgerdi and Kazerooni (2023) demonstrated that machine learning models, especially ensemble methods, significantly performed better than traditional models in CLV prediction. By using a combination of RFM metrics and behavioral data, we would be able to see the improvement in the prediction of customer value over time.

The addition of behavioral data in CLV prediction provides a more precise and accurate estimation of customer value. This gives companies a more complete picture of customer profitability by enabling them to identify both highly engaged and high-spending customers. These new methods are essential for enhancing e-commerce marketing effectiveness and resource allocation (Su et al., 2023).

## Hybrid Clustering Techniques

For customer segmentation, conventional clustering methods like K-Means have been widely used. By dividing the dataset into K clusters, each data point is assigned to the cluster with the closest mean in a process known as K-Means clustering. K-Means is a well-liked option because of its reliability and simplicity as demonstrated by (Li and Han, 2023). While effective in certain scenarios, K-Means also has limitations, its assumption of spherical clusters and its sensitivity to outliers. A study by Siagian (2021) has shown that while these limitations can be reduced by including additional dimensions in the data, like length (in LRFM), these models are still unable to handle complex and noisy data. This dissertation will use a hybrid clustering approach, combining K-Means with DBSCAN. DBSCAN is a good addition to K-Means because it is good at identifying randomly shaped clusters and detecting noise. DBSCAN can identify randomly shaped clusters and, unlike K-Means, does not require the specification of the number of clusters beforehand, making it appropriate for complex e-commerce datasets where customer behavior does not follow a clear pattern (John, Shobayo and Ogunleye, 2023). Researchers like Bhupender Singh Rawat (2023) have proposed hybrid models like these to improve segmentation accuracy, especially in e-commerce environments where customer behavior data is frequently noisy and non-spherical.

In a study conducted by Li (2023), to overcome non-spherical clusters and noisy data, hybrid particle swarm optimization was combined with K-Means clustering to enhance segmentation in e-commerce settings. In a similar way, combining K-means with DBSCAN would allow for better handling of noisy data and outliers, providing more accurate customer groupings (Li et al., 2023). It has been demonstrated that the K-Means and DBSCAN algorithms increase segmentation accuracy in scenarios involving complex customer data (John, Shobayo and Ogunleye, 2023). Businesses can create more individualized and focused marketing campaigns thanks to this improved segmentation, which raises customer engagement and profitability.

## Geospatial Analysis

Although geospatial analysis can yield valuable insights into regional trends and preferences, its application in customer segmentation has been underutilized (Ghahramani et al., 2021). Ghahramani also looked at how geospatial data could be integrated with RFM analysis, showing how mapping customer locations could be used to determine regional purchasing trends. To provide a complete picture of consumer behavior, the majority of studies, however, fall short in their attempts to fully integrate geospatial data with transactional and behavioral data. Acharya (2022) conducted a thorough study that demonstrates the potential of using geospatial and employs geospatial analysis to find regional patterns in consumer behavior, improving business strategies. In order to present a more comprehensive picture of customer behavior, behavioral, transactional, and geographic data will be combined in this dissertation. The mapping of the locations of my customers using Geographic Information Systems (GIS) and other geospatial tools will correlate those locations with metrics related to engagement and purchases.

Because geospatial data adds a geographic component to customer segmentation that traditional methods frequently lack, it improves the process (Griva 2022). This enables companies to better target high-value client segments and optimize delivery logistics by customizing their marketing strategies based on regional preferences. Greater customer satisfaction and loyalty are the result of better regional marketing initiatives made possible by an understanding of the locations of high-value customers.

# Methodology

The methodology in this dissertation is all about laying out a strong foundation to improve e-commerce strategies through better behavioral analysis and customer segmentation. In this section, I’ll Walk through the specific tools, and techniques that will be used to reach the goals of my dissertation.

## Approach

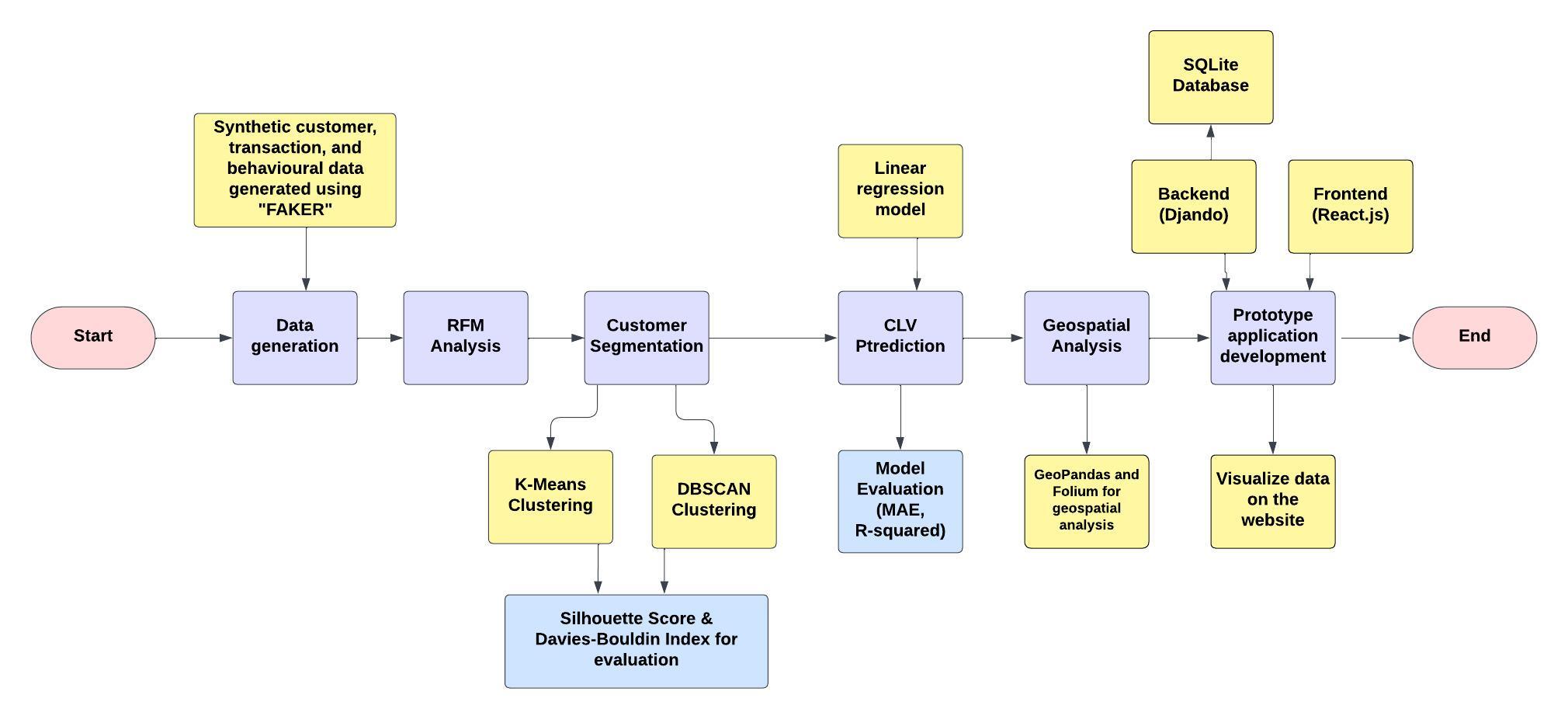


Figure 1: Methodology Flowchart

The methodology starts by generating synthetic datasets with the Python Faker library, creating realistic customer behavior while ensuring privacy. First, customers are segmented using RFM analysis. Next, predictive models are applied to estimate Customer Lifetime Value (CLV) through regression modelling, with the model's accuracy assessed using metrics like MAE and R-squared. To enhance segmentation, a hybrid clustering technique is used, combining K-Means for structured grouping and DBSCAN to handle outliers and irregular patterns.

Geospatial analysis is then integrated, mapping customer distribution to reveal regional trends by combining geographic insights with behavioral data. Finally, a prototype application is built to show how these techniques can be used in practice, allowing users to interact with and visualize segmentation and analysis results in real-time.

## Data Generation

Due to privacy and security concerns around using real customer data, this dissertation uses synthetic data generated through the Python "Faker" library. It is important to ensure that the data closely mimics real-world e-commerce behavior. For instance, the age distribution of customers has been created based on data sourced from [Statista](https://www.statista.com/topics/871/online-shopping/#topicOverview), ensuring that the distribution aligns with industry standards. The age ranges and their respective probabilities are as follows:

* 18 to 24 years: 16%
* 25 to 34 years: 26%
* 35 to 44 years: 23%
* 45 to 54 years: 18%
* 55 to 64 years: 10%
* 65 to 100 years: 7%

Additionally, the gender distribution as per Statista’s data is 53.1% of customers being male and 46.9% being female. By incorporating these realistic distributions, the synthetic data offers a sufficiently accurate testing environment for clustering and prediction models.

This method enables the creation of realistic, anonymized datasets that closely resemble real-world customer behavior and demographics, all while maintaining privacy safeguards (Dankar and Ibrahim, 2021).

### Customer Data Creation

1. Customer ID Generation:

Function is used to creates a unique 5-digit customer ID for each customer by randomly generating an integer between 10000 and 99999.

1. Customer Demographic Attributes in the generated data:

* Age Distribution:

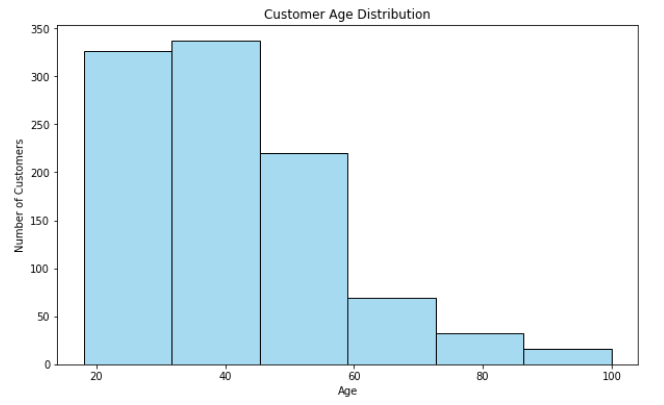


Figure 2: Customer Age Distribution

Each customer's age is assigned by using a weighted probability distribution over six age ranges as discussed in section 3.4.1 and shown in Figure 2.

* Gender Distribution:

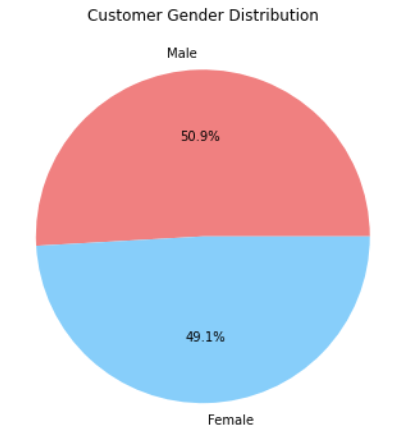


Figure 3: Customer Gender Distribution

* The assigning of a gender (Male or Female) to each customer is done based on a predefined distribution as discussed in section 3.4.1.
* Location Data: To provide geographic context, latitude and longitude are generated for each customer using the Faker library with UK locations using the function “fake\_uk”. This resembles the variation in customers locations inside of the UK.
* Join Date: To represent the first time a customer interacted with the e-commerce platform, the function is used to generate a random date within the last 15 years. This makes it possible to simulate long-term customer engagement and conduct more dynamic long-term customer behavior analysis.

### Transaction Data Creation:

Multiple transaction records for every customer will be generated such as information on the product category, payment method, transaction amount, and other elements of purchasing behavior.

1. Data Creation: Transaction data is generated in such a way that every generated transaction is linked to an existing customer ID generated earlier in the customer data. This ensures that each customer completes at least one transaction, enabling a thorough examination of customer behavior.
2. Transaction Attributes:

* Amount:

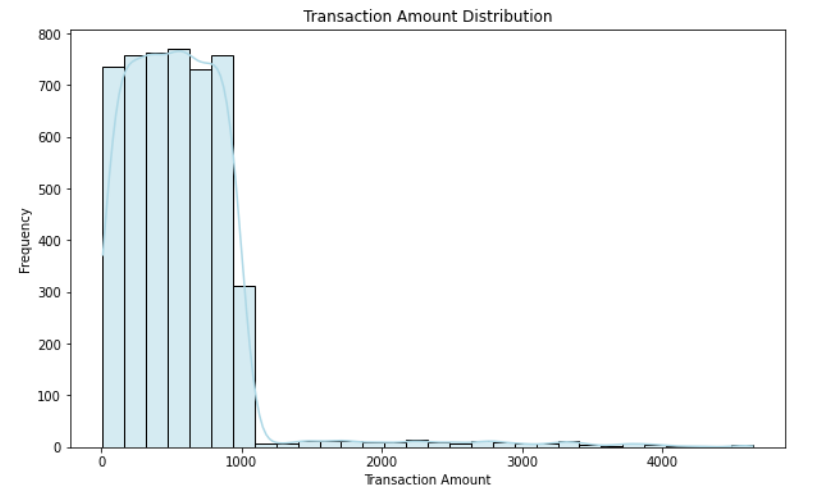


Figure 4: Transaction Amount Distribution

Random distribution within the range of 10 and 1,000 pounds is used to generate the transaction amount. Larger amounts are occasionally generated to represent scenarios where customers might make exceptionally large purchases.

* Other Attributes: Additional attributes such as transaction\_id (a unique UUID), transaction\_date (a random date within the current year), product\_category (e.g., Electronics, Clothing), payment\_method (e.g., Credit Card, PayPal), quantity, unit\_price, discount, transaction\_status (e.g., Completed, Pending) are also generated to provide a complete picture of each transaction.

### Behavioral Data Creation

Generation of behavioral data, such as frequency of logins, page views, and time spent on the site, that imitates user interactions with the e-commerce platform.

Steps and Techniques:

1. Behavioral Attributes:

* Login Data: Attributes like last\_login\_date, total\_login\_count, and pages\_visited are generated to simulate a customer’s usage of the e-commerce platform. The last\_login\_date is randomly selected from the past two months, indicating recent activity.
* Engagement Metrics: Metrics such as time\_spent\_on\_site (in minutes), num\_transactions, referral\_source (e.g., Search Engine, social media), device\_used (e.g., Mobile, Desktop), average\_order\_value, promo\_clicks, sessions\_per\_day, and social\_shares provide detailed insights into customer behavior. These metrics provide a deeper understanding of the devices that customers use, how they interact with the platform, and the success of various marketing channels.

### Data Validation and Matching

This is to ensure that each customer ID has matching entries in both the transaction and behavioral datasets. To find out how many customers have corresponding entries in the transaction and behavioral data, all three datasets are intersected. The number of common customer IDs between the transaction, behavioral, and customer datasets is checked, and ensured that every customer ID has corresponding transactional and behavioural data.

## Tools, Technologies, And Software Used for Development Process

This section outlines the tools and technologies used, focusing on their key functions within the dissertation.

### Development And Data Handling

* Python: Used as the core programming language for data generation, processing, analysis, and visualization due to its versatility and great choice of libraries (Pedregosa et al., 2011).
* Jupyter Notebook: Used for exploratory data analysis and documenting code, offering a user-friendly environment for development.
* pandas: Used for data manipulation, transforming datasets, and calculating RFM metrics.
* NumPy: Supported numerical operations essential for synthetic data generation and mathematical computations.
* Faker Library: Generated synthetic customer profiles, transaction data, and behavioral information for privacy-compliant and realistic data simulation.
* CSV: Used as the standard format for storing and accessing customer, transaction, and behavioral data.

### Machine Learning and Clustering

* Scikit-learn: Provided machine learning tools for customer segmentation (K-Means, DBSCAN) and prediction (Linear Regression), and to calculate performance metrics such as Mean Absolute Error (MAE), R², Silhouette Score, and Davies-Bouldin Index.
* StandardScaler: Used to normalize features to ensure uniform scaling before clustering for better performance and accuracy.
* Principal Component Analysis (PCA): Reduced dimensions of generated data for easier visualization and improved clustering results.
* Elbow Method: Used to identify the optimal number of clusters for K-Means by analyzing the relationship between cluster count and inertia.
* ROC Curve & AUC: Used to evaluate model performance for customer churn predictions, by analyzing the relationship between true positive and false positive rates.

### Data Visualization

* Matplotlib/Seaborn: Used to create visualizations for customer segmentation, RFM distribution, and other insights wherever necessary.
* GeoPandas: Used to handle geospatial data and convert customer latitude and longitude into map points for geographical analysis.
* Folium: Helped to generate interactive maps to visualize customer distribution and trends using MarkerCluster to aggregate nearby customer points.

### Backend Development

* Django: Managed backend operations, database interactions, and provided API endpoints for seamless frontend integration.
* Django ORM: Managed database operations such as storing and retrieving customer, transaction, and behavioral data.
* SQLite: Used as a database for efficient data storage and retrieval during development.
* Django REST Framework (DRF): Used to create API endpoints that allowed smooth data exchange between the backend and frontend.
* Django Static Files: Served static content such as customer maps generated by Folium for improved application performance.

### Frontend Development

* React.js: Managed the development of an interactive and dynamic user interface for the e-commerce dashboard.
* Axios: Managed HTTP requests to fetch data from the Django backend, ensuring smooth data communication between the frontend and backend.
* HTML5 & CSS3: Used to structure and style the frontend to ensure responsiveness and usability across the devices.
* React useState & Conditional Rendering: Used to enable toggling between dashboard components (RFM chart, customer map), improving interactivity and user experience.

### Development Environment

* Visual Studio Code (VSCode): Used as development environment for both backend and frontend code development.
* Node.js & npm: Managed frontend dependencies and handled the React development environment.
* Google Chrome: Used to test API endpoints to ensure correct data flow between the frontend and backend.
* Django Management Commands: Automated backend operations such as generating and updating customer maps dynamically.

## Version Management

For this dissertation, version management was handled using a combination of GitHub and Google Drive, ensuring efficient tracking of code development and documentation.

### Github For Code Versioning:

Repository Setup: All code, including data generation, machine learning models, and the prototype application, was stored and managed in a GitHub repository.

Update Frequency: Regular updates were made with updated code, ensuring traceability of all changes. This approach allowed for easy access to previous versions if needed.

### Google Drive and Docs for Document Management:

All drafts, notes, and documentation were stored in Google Docs. This allowed for real-time viewing of my work whenever required.

Google Drive for File Storage: Analysis code, figures, presentations, and other non-code materials were organized and stored in structured folders within Google Drive, ensuring easy access and backup if needed.

# Development Process, Results & Evaluation

This section takes a deep dive into the entire development process, it covers everything from preparing the data for analysis, designing and implementing the prototype to generating and analyzing the outputs. Key results from important stages, like customer segmentation and prediction models, are presented, along with an evaluation of their performance. This makes the section essential for understanding the project’s strengths, weaknesses, and overall outcomes.

## Recency, Frequency, And Monetary (RFM) Analysis

As per Yıldız (2023), Recency, Frequency, and Monetary (RFM) analysis is a customer segmentation technique that evaluates customers based on their purchase history using the following criteria:

* Recency (R): Number of days since the customer's last transaction.
* Frequency (F): Total number of transactions made by the customer.
* Monetary (M): Total amount spent by the customer.

### Data Preparation and Steps for RFM Analysis

* Convert Transaction Dates: pd.to\_datetime() function is used to convert the date column in the transaction dataframe to date-time format. To calculate dates correctly, this conversion is required.
* Recency Calculation: Determining the difference between the latest transaction date among the transactions in the dataset and the last transaction date of each customer.

Recency (R) = Latest Transaction Date − Last Transaction Date

This metric tells us how recent a customer's last transaction was.

* Frequency Calculation: Frequency is computed as the count of transactions for each customer:

Frequency (F) = Number of Transactions

A higher frequency suggests a more loyal or engaged customer.

* Monetary Calculation: The total amount spent by each customer is calculated by summing the transaction amounts:

Monetary (M) = ∑ Transaction Amounts

This value shows the total revenue generated by each customer.

* Normalization: RFM values are normalized using StandardScaler function to bring all three metrics (R, F, M) to the same scale:

Standardized Value = Value – Mean Standard Deviation

This is to avoid the scale to influence the results during clustering.

### RFM Scoring

RFM scoring is used to convert the RFM values into quantile-based scores using the pd.qcut() function, which further helps in segmentation.

* R Score: Customers with lower recency values (more recent transactions) get higher R scores.

R\_Score = 5 if most recent purchase  
 1 if least recent purchase

* F and M Scores: Customers with higher frequency and monetary values get higher F and M scores.

F\_Score, M\_Score = 5 if highest value  
 1 if lowest value

The thresholds used for RFM scoring (score of 5 for the most recent purchase) were chosen based on best practices in the e-commerce industry. These thresholds allow for quick and effective segmentation of customers into groups. They are also flexible and can be adjusted based on the specific characteristics of a dataset or business goals. For example, a business with high transaction volumes might reduce the recency threshold to capture more recent customers, while a luxury goods business might extend the monetary threshold to reflect higher spending habits.

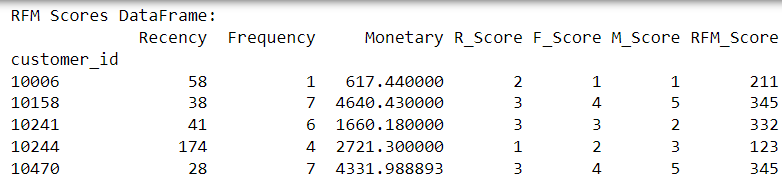


Figure 5: RFM Scores Dataframe

These scores are combined into an overall RFM score as shown in Figure 5, such as "555" for the most valuable customers and "111" for the least valuable customer.

### Customer Segmentation Using RFM Scores

Logical function is used to get the RFM scores to segment customers into predefined groups based on their generated Recency, Frequency, and Monetary scores.

Segmentation logic used:

* VIP: High Recency, Frequency, and Monetary (R >= 4, F >= 4, M >= 4)
* Loyal: Moderate Frequency and Monetary (R >= 4, F >= 3, M >= 3)
* Recent: High Recency (R >= 4)
* Big Spender: High Monetary (M >= 4), moderate Frequency (F >= 3)
* Lost: Low Recency, Frequency, and Monetary (R <= 2, F <= 2, M <= 2)
* At Risk: Low Recency but moderate Frequency (R <= 2, F >= 3)
* Others: Customers not belonging to the above categories

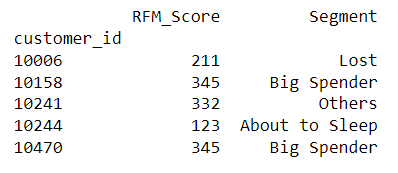


Figure 6: Segment Distribution

The segmentation of customers (Figure 6) based on RFM scores allows businesses to tailor their marketing strategies more effectively. For instance, 'VIP' customers (high Recency, Frequency, and Monetary scores) can be prioritized for loyalty programs or personalized promotions. Meanwhile, 'At Risk' customers (low Recency but high Monetary) might benefit from re-engagement campaigns, such as special offers or targeted ads, to prevent them from leaving.

### Customer Scoring

RFM data is merged with behavioral data to enhance customer scoring. Behavioral data includes metrics such as total login count, pages visited, time spent on site, number of transactions, and so on. These features could offer more in-depth insights into customer engagement, going beyond what traditional RFM analysis can capture (Van Doorn *et al.*, 2010).

1. Loyalty Score = (Total Login Count × 0.2) + (Pages Visited × 0.3) + (Time Spent on Site × 0.5)

This score measures customer engagement, with more weight given to time spent on the site, indicating loyalty.

1. Churn Risk Score = (Recency × 0.5) + (1 / Frequency × 0.3) + (1 / Total Login Count × 0.2)

This score identifies customers likely to leave, focusing on inactivity, low purchase frequency, and minimal logins.

1. Monetary Value Score = (Monetary × 0.7) + (Average Order Value × 0.2) + (Number of Transactions × 0.1)

This score evaluates a customer’s financial contribution, giving more weight to total spending.

1. Purchase Consistency Score = (Frequency × 0.5) + (Number of Transactions × 0.3) + (1 / Recency × 0.2)

This score tracks how consistently a customer makes purchases, rewarding frequent and recent buyers.

### Data Visualization of The RMA Analysis

#### Visualizing the customer segments: The purpose of this plot is to provide a visual summary of the customer segmentation based on the RFM scoring system and the dynamic segmentation process.

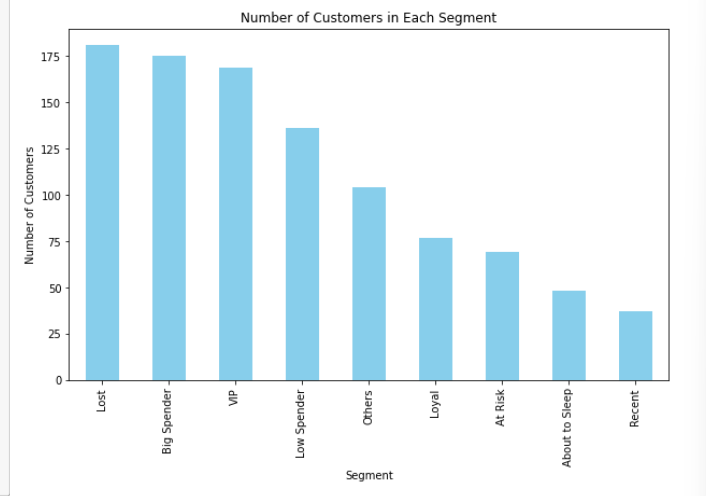


Figure 7: Customer Segment Distribution

The bar chart makes it easy for businesses to get a quick look of their customer distribution. For instance, larger segments like "Lost" or "Big Spender" might be great opportunities for targeted retention efforts or promotions. On the other hand, smaller segments like "Recent" or "Loyal" may need a closer look to figure out how to improve customer engagement and keep them coming back.

#### Customer Cohorts and Segment Distribution

Cohort analysis is used to group customers based on their recent activity (Recency).

Cohort Binning Logic: Customers are divided into cohorts using pd.cut() function, with bins for recency, These bins represent different levels of recency as shown in Figure 8. A heatmap is generated, color-coding the customer counts across segments, providing a visual understanding of customer distribution.

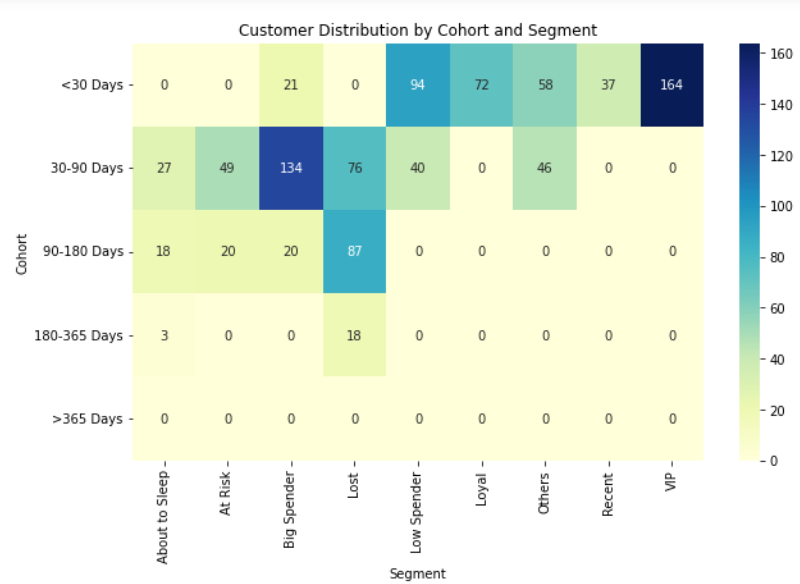


Figure 8: Customer Grouping based on Recency

The plot reveals that a lot of customers in the "VIP" or "Big Spender" segments fall into the <30 Days cohort, meaning they’ve made frequent and high-value purchases recently. On the other hand, the "Lost" segment is more common in the >365 Days cohort, showing customers who haven’t interacted with the business in a long time. By looking at how customers are spread across these cohorts, businesses can plan their marketing strategies working to re-engage at risk or lost customers.

#### Visualization of the distribution of customers based on Frequency and Monetary

The frequency histogram shows the number of customers who made a certain number of purchases, and the monetary histogram shows the number of customers who have spent certain total amounts. The range of values is divided into 20 bins. Each bin represents a range of frequency and monetary values, and the height of the bar represents the number of customers whose frequency falls within that range.

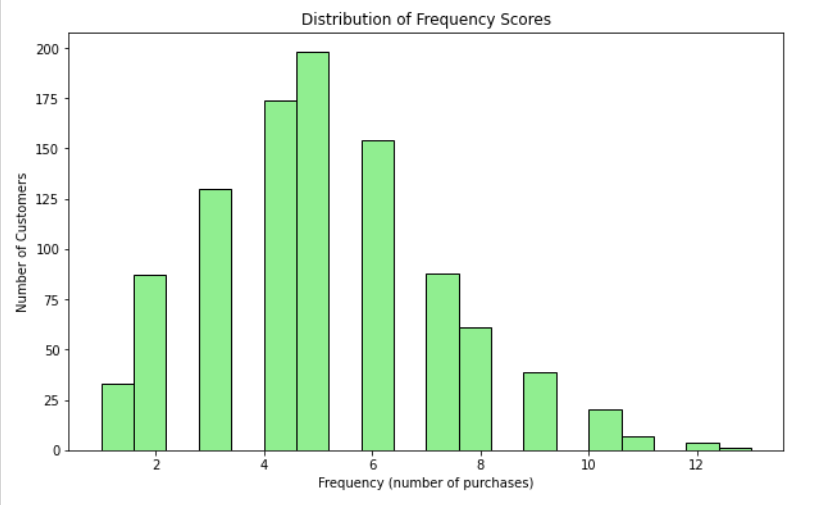


Figure 9: Frequency Distribution

The histogram (Figure 9), gives a clear picture of how often customers make purchases. We can notice that most customers tend to make only a few purchases, with fewer people buying more frequently. This helps businesses understand whether their customer base is mostly made up of repeat buyers or if they have a lot of one-time purchasers.

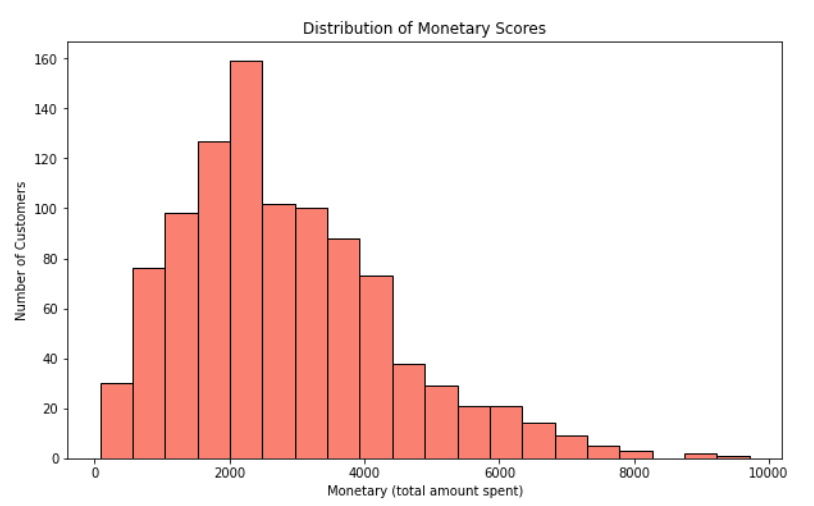


Figure 10: Monetary Distribution

The monetary histogram (Figure 10), gives insight into how much customers are spending over a time period. We can see that most customers fall within a certain spending range, while only a small group tends to contribute to much higher revenues. This helps businesses understand their spending patterns and identify those top-spending customers.

#### ROC Curve and AUC Analysis

The ROC Curve (Receiver Operating Characteristic) and AUC (Area Under the Curve) are important tools for evaluating how well a binary classification model performs. When applied to RFM (Recency, Frequency, Monetary) analysis, these metrics help predict customer behaviors, such as churn or loyalty, by measuring how accurately the model can classify customers into different categories.

In Figure 11 shown below,

* The X-axis represents the False Positive Rate (FPR), which measures the proportion of incorrectly classified negatives.
* The Y-axis represents the True Positive Rate (TPR), measuring the proportion of correctly classified positives.

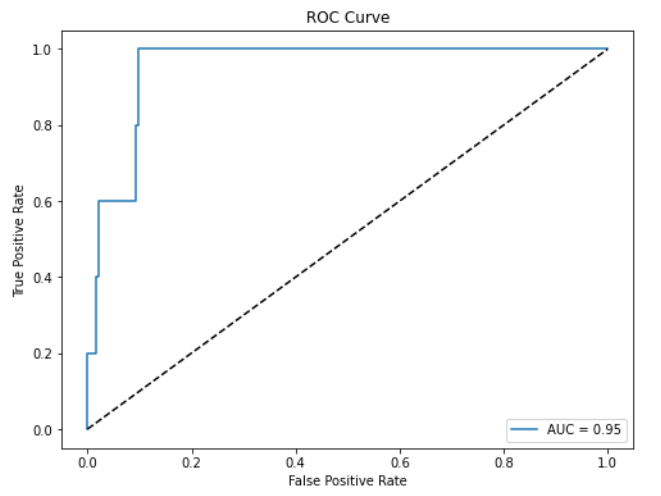


Figure 11: ROC Curve and AUC Analysis

* A diagonal line (dashed) represents a random classifier with AUC = 0.5, while a model that curves towards the top-left corner represents a good classifier.
* The AUC (Area Under the Curve) score of 0.95 suggests that the model is very effective at differentiating between the positive and negative classes.

ROC Curve and AUC are useful for evaluating how well RFM and behavioral data predict customer behaviours such as churn or loyalty, helping businesses refine their strategies.

## Customer Lifetime Value (CLV) Prediction

The main goal of this approach is to predict Customer Lifetime Value (CLV) using Linear Regression, leveraging both RFM scores and additional customer behavior metrics. Linear regression was chosen for its simplicity and interpretability, which are crucial for businesses that require clear and actionable insights. The model is set up to estimate the Monetary value (as a proxy for CLV) and assess its accuracy by using performance measures like Mean Absolute Error (MAE) and R-squared (R²).

### Data Preparation

The model uses a combination of factors as independent variables, including Recency, Frequency, and various behavioral metrics like total logins, pages visited, time spent on the site, number of transactions, promo clicks, sessions per day, and social shares. The target variable is the Monetary value, which reflects how much the customer has spent and acts as a stand-in for CLV. In practical situations, this could be replaced with other metrics depending on the specific needs of the business.

### Train-Test Split

The dataset is split into 80% training data and 20% test data. This ensures that the model is trained on a portion of the dataset and tested on unseen data to measure its generalization ability. (Rácz, Bajusz and Héberger, 2021)

### Prediction And Evaluation

The model's predictions are made for both the training and test datasets. The predictions for the entire dataset are stored in a new column, CLV\_Prediction in Figure 12.

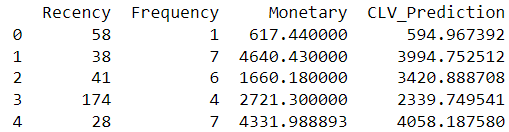


Figure 12: CLV Prediction Values

The performance of the CLV model is measured using two key metrics (Ramos and Silva, 2023):

1. Mean Absolute Error (MAE): Measures the average difference between the actual CLV (Monetary) and the predicted values.
2. R-Squared (R²): Shows how well the independent variables account for the changes in the target variable, which in this case is Monetary value. If the R² is 1, it means the model perfectly fits the data, while a value of 0 indicates that the model doesn't explain any of the variance at all.



Figure 13: MAE and R-Squared

From Figure 13, The Mean Absolute Error shows that, on average, the predicted CLV is off by about £818 from the actual value. This relatively low accuracy could be due to the synthetic data not fully capturing the complexities of real customer behavior, or because the model hasn’t been tested against real-world variations in spending patterns.

The R-Squared value of 0.4987 means that about 50% of the variation in the target variable (Monetary) is explained by the features in the model. This is a fairly low value, which could be because the features are too limited or simplistic compared to real-world data. It might also be due to the synthetic data lacking realistic correlations. In real-world datasets, we would probably see more variability because of the wide range of customer behaviours.

## K-Means and DBSCAN Clustering

While K-Means clustering is effective for segmenting structured data based on purchasing patterns, it assumes that all clusters are spherical and evenly distributed, which may not accurately reflect real-world customer behavior. By integrating DBSCAN, we are able to detect irregularly shaped clusters and handle noise, which is particularly useful in e-commerce datasets where outliers such as infrequent but high-value customers may exist. DBSCAN’s ability to identify these outliers complements K-Means' structure, providing a more holistic approach to customer segmentation.

### Merging RFM And Behavioral Dataframes

The process starts by merging two key datasets RFM data and behavioral data. This step brings together information about customers’ purchasing habits with additional behavioral metrics. By combining these datasets, we get a more complete picture of customer activity (Van Doorn et al., 2010).

Next, we focus on selecting the most relevant features for clustering analysis. This includes both RFM metrics (Recency, Frequency, Monetary) and behavioral data such as total login count, pages visited, and time spent on site. Together, these features offer a well-rounded view of customers behaviours and spending patterns, setting the stage for more accurate analysis.

### Standardizing The Features:

Function Used: StandardScaler()

Formula: Xscaled = (X−μ)/ σ

Where μ is the mean and σ is the standard deviation.

Because the features are on different scales, standardization is applied using StandardScaler. This step normalizes all the feature values so they have a mean of 0 and a standard deviation of 1. By doing this, we ensure that no single feature influences the clustering process simply because of its scale.

### Dimensionality Reduction Using PCA:

Function Used: PCA(n\_components=2)

Principal Component Analysis (PCA) is used to reduce the number of features down to two components, making the clustering process simpler and easier to visualize in a 2D space without losing much information (Marukatat, 2023).

### Elbow Method for Optimal K:

The Elbow Method helps determine the best number of clusters (k) for K-Means clustering. By calculating the variance for different values of k and plotting the results, we can spot the "elbow" point where the reduction in inertia slows down, indicating the ideal number of clusters. Once the optimal k is found, each customer is assigned to the cluster closest to the respective centroid (Marisa *et al.*, 2023).

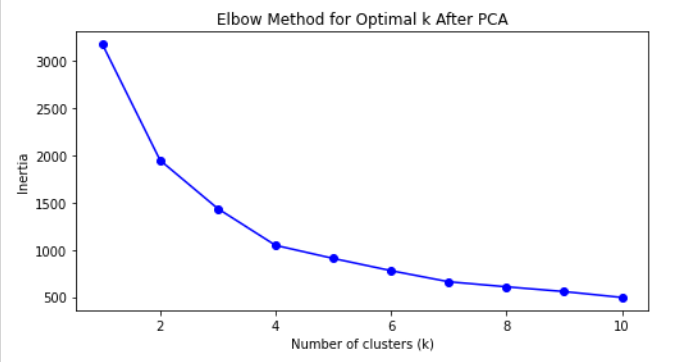


Figure 14: Elbow method for optimal k

The "elbow" appears at k = 5 as we can see in Figure 14, where the drop in inertia starts to level off. This suggests that adding more than five clusters doesn’t significantly improve the model. So, using 5 clusters strikes a good balance between keeping the model simple and capturing the key patterns in the data.

### Visualizing Clusters After PCA:

The clustered data is visualized in 2D using the two principal components.

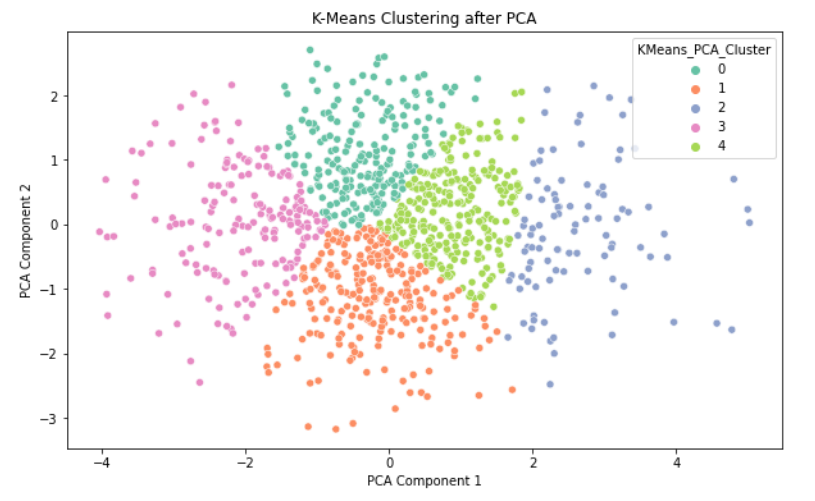


Figure 15: Visualization of Clusters

Figure 15 shows the results of K-Means clustering after reducing the data to two dimensions using Principal Component Analysis (PCA). Each colour represents a different cluster, with the points grouped according to similarities in the data. The x-axis and y-axis correspond to the two main components from PCA, capturing the most important information from the original features. The clear separation of the coloured clusters suggests that the K-Means algorithm successfully divided the data into 5 distinct groups, as determined by the earlier Elbow Method.

### High-Value Customer Visualization:

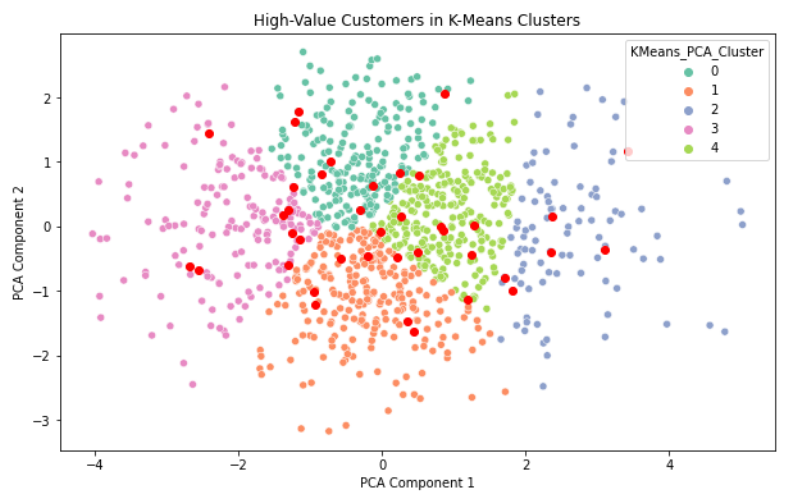


Figure 16: High Value Customer Visualization

This plot in Figure 16 highlights high-value customers (shown in red) within the K-Means clusters, based on the data reduced using Principal Component Analysis (PCA). These high-value customers were identified through their RFM scores (Recency, Frequency, Monetary), with those scoring '555' across all three metrics considered top-tier. This means they made recent purchases, buy frequently, and spend significant amounts.

The fact that these high-value customers are spread across multiple clusters is crucial, as it shows the diversity in their behavior. For example, some might make frequent, smaller purchases, while others may buy less often but spend more per transaction. By understanding their distribution across clusters, businesses can tailor their strategies and allocate resources more effectively, ensuring each type of high-value customer receives the right attention to boost their lifetime value.

### Cluster Distribution Based on Recency, Frequency, And Monetary

The box plots below show customer segmentation based on Recency, Frequency, and Monetary (RFM) metrics:

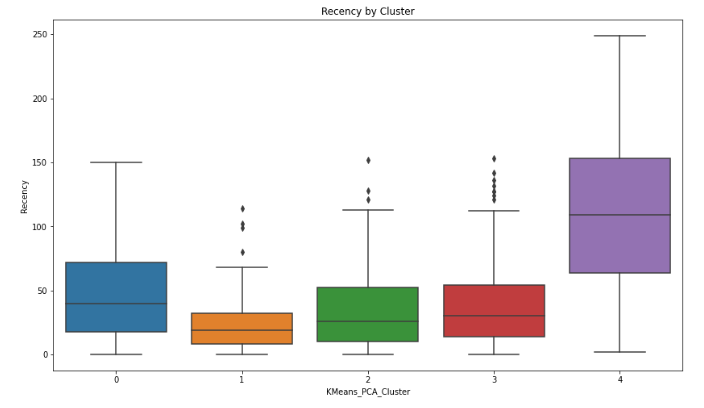


Figure 17: Recency vs K-Means

Recency: In Figure 17, Cluster 4 shows the highest recency, meaning these customers haven't made a purchase in a long time, while Cluster 1 has the lowest recency, representing the most recent buyers.

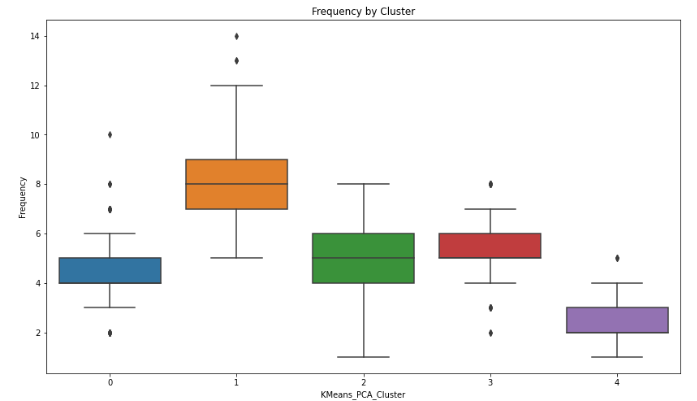


Figure 18: Frequency vs K-Means

Frequency: In Figure 18, Cluster 1 stands out with the highest frequency, indicating they are the most frequent buyers, while Cluster 4 has the lowest, pointing to less frequent purchases.

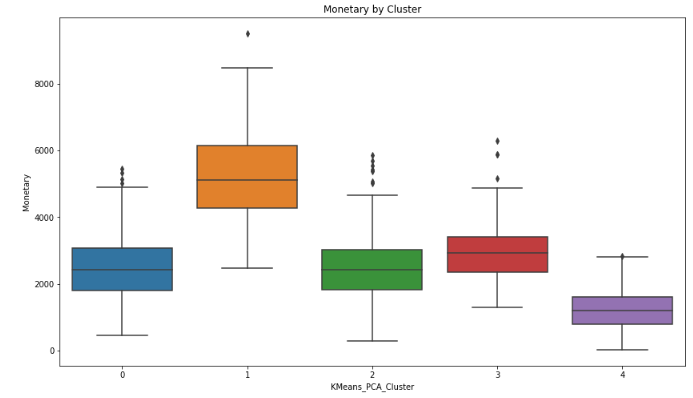


Figure 19: Monetary vs K-Means

Monetary: In Figure 19, Cluster 1 also leads in spending, with the highest monetary values, whereas Cluster 4 spends the least.

Overall, Cluster 1 represents high-value, frequent buyers, and Cluster 4 indicates low-engagement, low-spending customers.

### DBSCAN Clustering:

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is used to identify outliers (noisy points) and dense clusters in the data. ( John, J. M –1.3)) Unlike K-Means, DBSCAN doesn't require you to pre-define the number of clusters and can detect clusters of any shape. It also flags outliers, which are labeled as "-1," helping to isolate data points that don't fit well into any cluster".

Function Used: DBSCAN(eps=0.25, min\_samples=5)

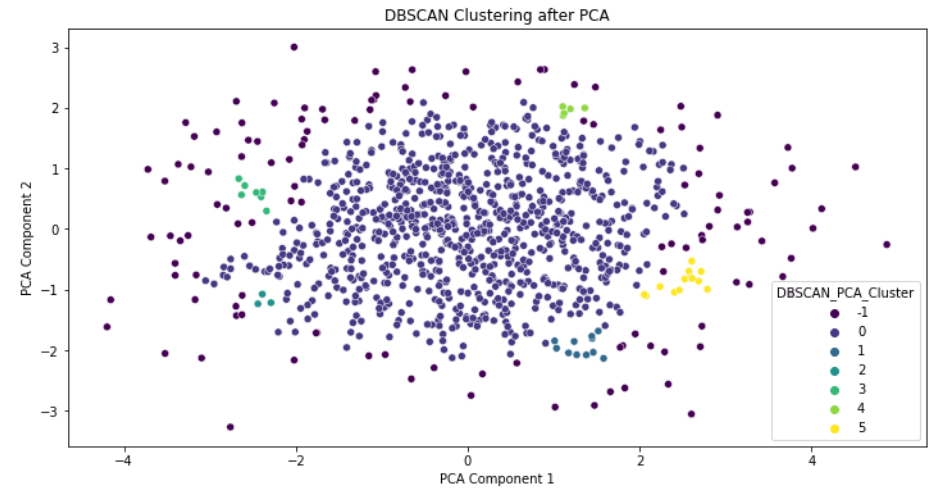


Figure 20: DBSCAN Clustering

This plot shows the results of DBSCAN clustering after reducing the data to two dimensions using Principal Component Analysis (PCA). The x and y axes represent the first two principal components, making it easier to visualize the data in a 2D space. Each colour represents a different cluster identified by DBSCAN, with a total of five clusters labelled from 0 to 4. Cluster 0, shown in blue, contains the majority of data points, while the smaller clusters (1 to 4) are more spread out across the plot.

### Outlier Detection With DBSCAN:

DBSCAN is used for detecting outliers in customer data. Customers who fall outside the core clusters will be labelled as -1 by DBSCAN and are identified as potential outliers.

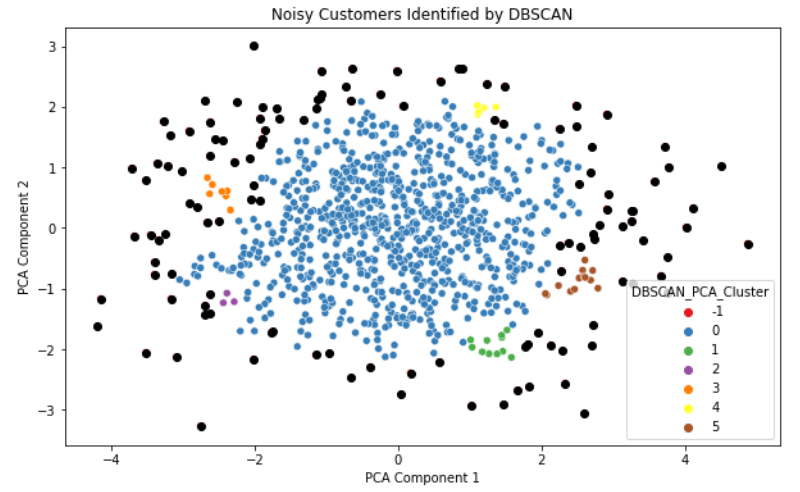


Figure 21: Noise identified by DBSCAN

The points labelled as noise, marked by the "-1" label (In Figure 21), are highlighted. These outliers often show unique customer behavior, like occasional big spenders or customers with unpredictable shopping habits. Identifying these cases allows businesses to tailor specific marketing strategies either to strengthen relationships or know about potential fraud. For instance, high-value outliers could benefit from personalized attention, such as dedicated customer service, while unusual spending patterns might trigger fraud detection measures to ensure security.

### DBSCAN Outlier Detection for High-Value Customers Only

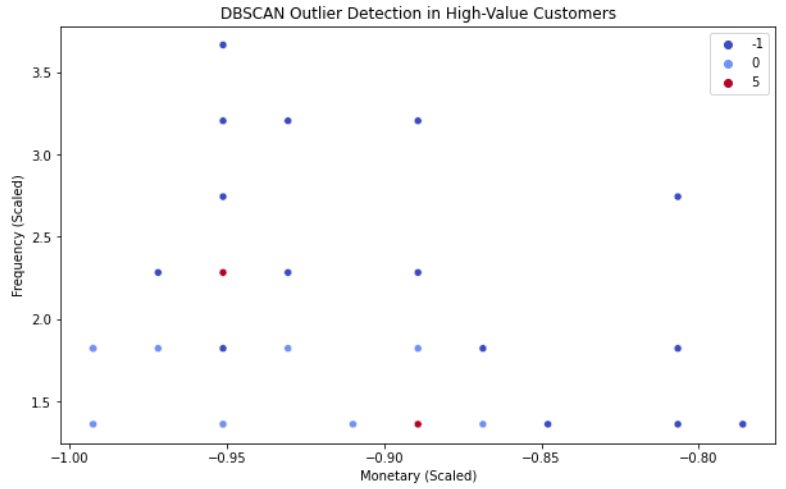


Figure 22: Outliers from High-Value Customers

This plot highlights DBSCAN's outlier detection among high-value customers, focusing on two key metrics: Monetary (total amount spent) and Frequency (number of purchases). DBSCAN is used here to spot customers whose spending and purchasing habits stand out from the rest. Outliers, marked as "-1," represent customers whose behavior is noticeably different from typical patterns, while regular customers are grouped into clusters.

Choosing to plot Monetary and Frequency is crucial because these metrics provide a clear picture of customer behavior. DBSCAN helps uncover insights such as customers who spend a lot but buy infrequently, or those who make frequent purchases but with lower spending. Understanding these outliers allows businesses to fine-tune their customer segmentation, personalize marketing strategies, and identify potential risks or opportunities in customer behavior.

### Silhouette Score and Davies-Bouldin Index:

The Silhouette Score evaluates clustering quality by measuring how similar an object is to its own cluster compared to other clusters, with higher scores indicating more distinct and well-separated clusters (Shahapure & Nicholas, 2020). Whereas The Davies-Bouldin Index assesses clustering performance by comparing the ratio of within-cluster scatter to between-cluster separation, where lower values signify better-defined clusters (Ros, Riad and Guillaume, 2023). A Silhouette Score above 0.5 and a Davies-Bouldin Index below 1.0 indicate well-defined clusters.

For K-Means



Figure 23: Silhouette Score and Davies-Bouldin Index for K-Means

* Silhouette Score (0.5223): This score indicates moderate separation and cohesion between clusters. K-Means generally works well with continuous data like RFM scores, but some overlap between clusters can occur because it assumes clusters are spherical.
* Davies-Bouldin Index (1.0586): A lower DB Index suggests that the clusters are compact and well-separated. This shows K-Means' strength in reducing variance within clusters, leading to tight, distinct groupings.

K-Means performed well because it's great at handling continuous data, like RFM, and focuses on minimizing the variance within clusters. This approach led to good separation and compact groupings, as reflected by the performance scores.

For DBSCAN,



Figure 24: Silhouette Score and Davies-Bouldin Index for DBSCAN

* Silhouette Score (0.3231): This lower score suggests weaker separation between clusters, possibly due to the challenges of density-based clustering or less-than-perfect parameter tuning. It indicates that some clusters may overlap or have unclear boundaries.
* Davies-Bouldin Index (0.9586): Despite the lower Silhouette Score, DBSCAN achieved a strong DB Index, meaning that the clusters it did detect are compact and well-separated. This reflects DBSCAN's strength in handling complex shapes and identifying outliers.

Overall, DBSCAN excelled at forming dense, well-separated clusters (as shown by the good DB Index), but it struggled to create clearly separated clusters in this particular dataset (indicated by the lower Silhouette Score). This could be due to the nature of the data or suboptimal parameter settings, but the algorithm still managed to identify meaningful dense clusters.

## Geographic Visualization

Understanding the geographic distribution of customers is critical for businesses looking to optimize regional marketing strategies and logistics. This also provides insights into supply chain optimization, allowing businesses to prioritize delivery routes to regions with the highest demand or customer value.

### Converting Customer Data into Geospatial Points

Using GeoPandas, customer data is transformed into GeoDataFrame with latitude and longitude mapped as geometric points. The WGS84 (EPSG:4326) geographic coordinate system that represents locations on the Earth's surface is used which ensures global accuracy for spatial data (Favretto, 2014).

### Identifying High-Value Customers

High-value customers, classified as "VIP" through RFM analysis, are filtered for special attention. This ensures that high-priority customers are visually distinct, helping focus business strategies on valuable segments.

### Data Validation

Latitude and longitude data are validated and corrected for any discrepancies. This ensures accurate plotting of customer locations, avoiding errors in visualization.

### Interactive Map Creation



Figure 25: Snapshot of Interactive Map

The map highlights regions with a dense concentration of high-value customers (VIPs), the clustering of VIP customers in certain regions provides a clear indication of where the most loyal and profitable customers are located, helping to optimize logistics, sales, and customer support.

Areas with fewer high-value customers may represent unnoticed market opportunities, where businesses could focus on improving their presence.

### Highlighting Customer Segmentation in The Map

MarkerCluster function is used to group customer locations based on proximity, helping visualize clusters of customers in different regions.

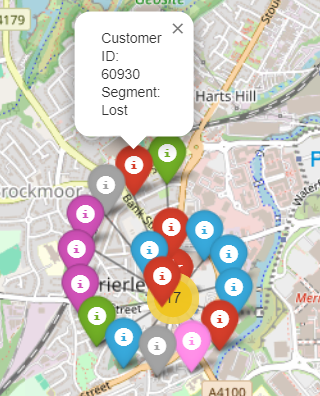
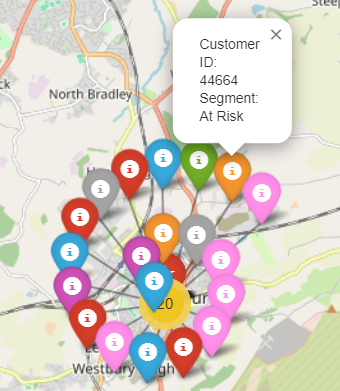
 

Figure 26: Highlighting Customer Segmentation

As shown in Figure 26, each customer segment is represented by a unique colour, making it easy to visually differentiate between customer types. For example, VIP customers are highlighted in pink, as they represent high-value clients, while "At Risk" customers are marked in orange, and "Lost" customers in yellow. Other segments, like "Loyal," "Big Spenders," and "Low Spenders," are also color-coded with shades of green, purple, and blue, ensuring a clear visual distinction.

Each marker on the map is interactive. Clicking on a marker brings up a pop-up with key details about the customer, such as their ID and segment. This allows us to dive deeper into individual customer profiles without cluttering the map with too much information at once.

The geographical data for each customer (latitude and longitude) is managed using the GeoPandas library, which simplifies the process of plotting customer locations on the map. GeoPandas ensures the data is properly formatted for mapping. The map itself is interactive, allowing us to zoom in and out of different regions and click on individual customer markers for more detailed information. This interactive feature makes it a powerful tool for exploring and understanding customer data.

Geographic Insights: The color-coded map offers a clear picture of where different customer segments are concentrated. For example, businesses can quickly spot regions with a high number of VIP customers, which can help guide decisions on where to focus marketing campaigns or special promotions. Likewise, areas with a large number of "Lost" or "At Risk" customers can highlight regions where stronger retention efforts are needed.

## Creation Of Prototype Application

An e-commerce dashboard will be designed to provide businesses with clear insights into customer behavior through interactive visuals. The dashboard features customer segmentation based on RFM (Recency, Frequency, Monetary) analysis, a geographic map showing where customers are located, and a detailed customer list for deeper analysis.

It’s all integrated into a user-friendly React.js frontend, supported by a Django backend to handle the data and functionality behind the scenes. The Django backend communicates with the React frontend through RESTful API endpoints, allowing real-time data exchange. Users can input customer data through the frontend interface and receive detailed insights. For example, a business user may query the system for customers with high RFM scores, view their geographic distribution on an interactive map, and drill down into individual transaction histories to tailor marketing strategies accordingly.

### Backend Functionality

Data generation:

Using the Faker library, synthetic data is generated for customers, transactions, and behavioral information, and the RFM data is generated by the analysis that is carried out to classify the customers into different segments.

Data Storage:

The backend uses SQLite as the database to store customer, transaction, and behavioral data by using the models that are created in Django to represent Customers, Transactions, and Behavioral data. The generated data is saved to the database using Django ORM (Object-Relational Mapping).

Customer Location Map

* Folium and GeoPandas are used to create a geographical distribution of customers based on their latitude and longitude.
* MarkerCluster from Folium is used to cluster customer locations on the map to provide a clear view of customer distribution.
* The generated map is saved as an HTML file and stored in Django's static folder for access via a URL.

API Endpoints:

* Django provides RESTful API endpoints for customer, transaction, and behavioral data, served using Django REST Framework (DRF).
* These endpoints allow the React frontend to fetch real-time data from the database on customers and their transactions.

### Frontend Functionality:

The React frontend is designed to provide an interactive and user-friendly interface for viewing customer information and visualizations.

#### Welcome dashboard:

First, the user accesses the React dashboard (Figure 27) where they are greeted with options to search for customers, view the RFM chart, and view the customer location map.

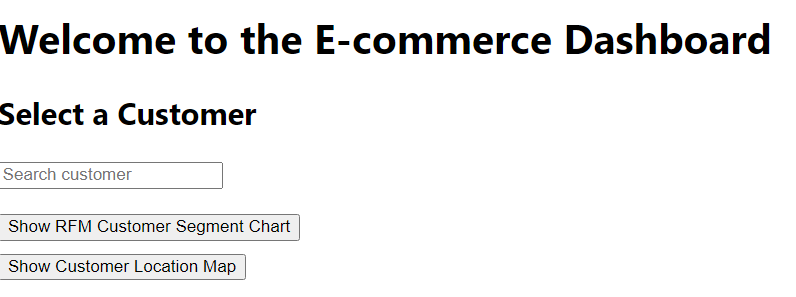


Figure 27: Frontend Dashboard

#### Customer Search Component:

A component where users can search for customers based on their names (Figure 28). This component communicates with the Django API to retrieve the customer data.

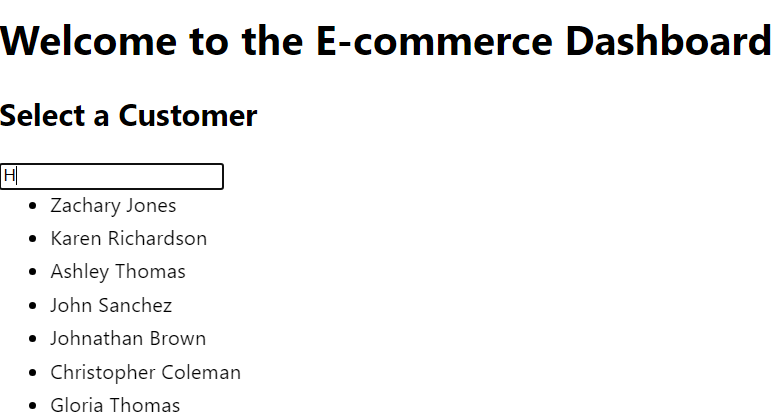
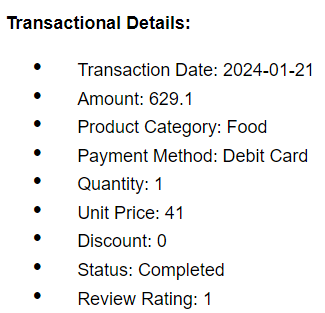
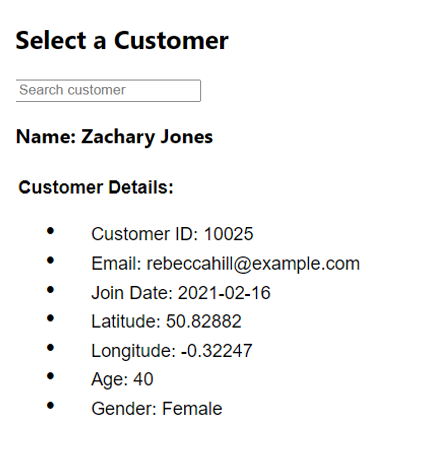
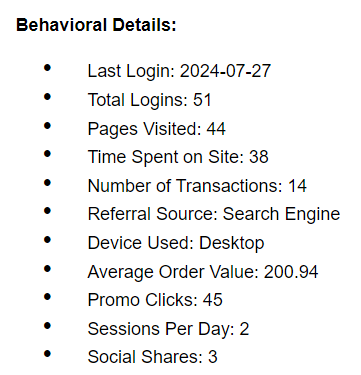


Figure 28: Frontend Dashboard - Customer Search

When the name of the desired customer is selected, customer, transactional, behavioral, and RFM details will be displayed, as shown in below Figure 29.





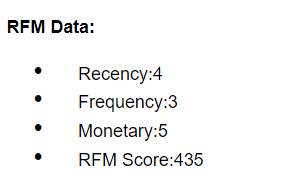


Figure 29: Frontend Dashboard - Customer Details Displayed

#### RFM Customer Segmentation:

As we can see in Figure 30, toggle button is provided in the UI to show or hide the RFM Customer Segment chart. When the user clicks on "Show RFM Customer Segment," the RFM chart is rendered, displaying customer segments based on the RFM analysis.

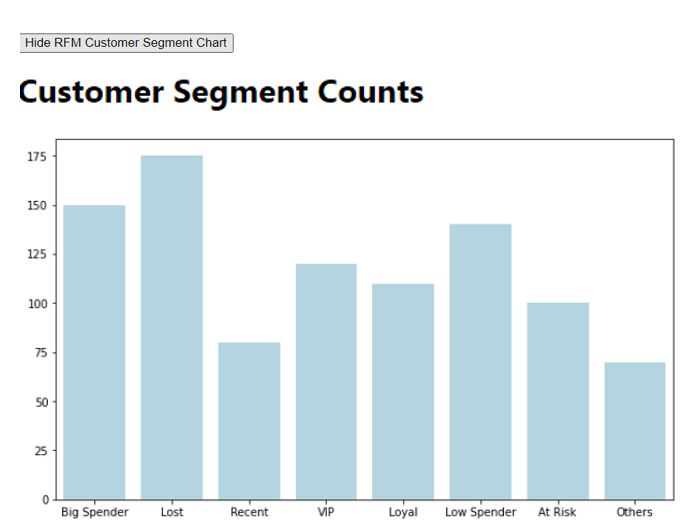


Figure 30: Frontend Dashboard - Customer Segmentation Bar plot

* The chart helps visualize customer distribution into different segments and provides actionable insights.

#### Customer Location Map:

* Another button toggle is provided to show or hide the Customer Location Map (Figure 31).
* Upon clicking the "Show Customer Location Map" button, an <iframe> is rendered, embedding the HTML map generated by Folium in the backend.

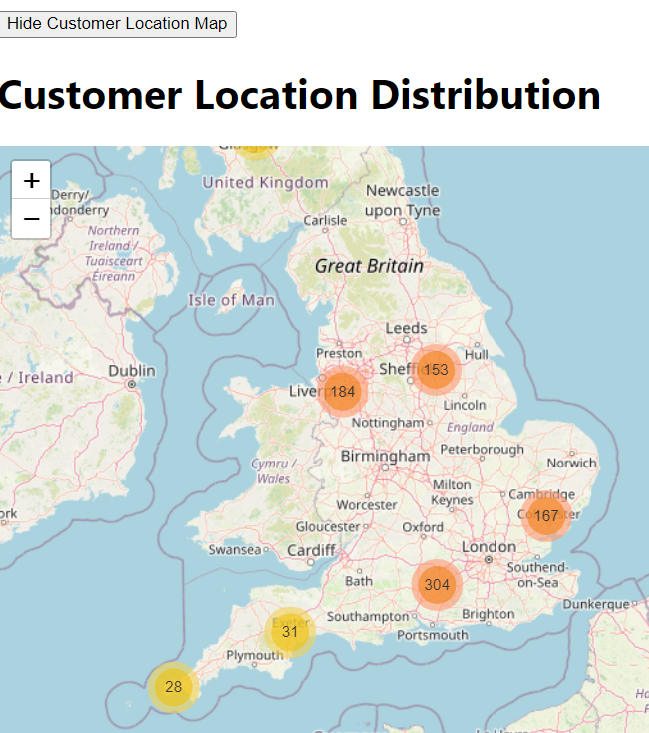


Figure 31: Frontend Dashboard - MAP of Customer Location Distribution

* This interactive map displays the geographical distribution of customers, allowing the business to understand where their customers are geographically concentrated.

#### Interactivity and Conditional Rendering:

* The React app makes heavy use of useState to toggle visibility between different components (RFM chart and Map).
* Conditional rendering ensures that the UI dynamically updates based on user input without reloading the page.

### Deployment And Testing:

Backend Testing:

* Run Django's management commands such as generate\_customer\_map to verify the customer distribution map is generated and stored in the static folder.
* Test API endpoints using Google Chrome browser to ensure customer, transaction, and behavioral data is correctly returned.

Frontend Testing: Manually interact with the dashboard to ensure that customer searches work correctly, the RFM chart is displayed when the button is clicked, and the customer map loads correctly within the iframe.

Integration Testing: Ensure that the frontend communicates smoothly with the Django API and that all components (RFM chart, customer map) are rendered without errors.

# Evaluation and Discussion

## Dissertation Results and Achievement

The results obtained from this research align with existing literature but also highlight areas where current e-commerce strategies can be improved. The hybrid clustering approach, which combines K-Means with DBSCAN, helped find more accurate customer segments compared to traditional methods like RFM analysis alone. This finding is consistent with studies by Rawat (2023) and Li (2023), who demonstrated that hybrid clustering models perform better in noisy and non-spherical data environments typical of e-commerce settings.

However, there are limitations in the model's ability to fully capture customer diversity. For example, while K-Means efficiently groups frequent buyers, it is sensitive to outliers, which DBSCAN addresses by detecting noise in the data. Yet, DBSCAN's reliance on manual parameter tuning can sometimes lead to suboptimal results, as demonstrated in the slightly lower silhouette score compared to K-Means (Silhouette Score: 0.3231 for DBSCAN vs. 0.5223 for K-Means). Future work could explore automated parameter tuning methods to improve DBSCAN’s performance as suggested by Rawat (2023).

The geospatial analysis revealed regional differences in customer purchasing patterns, a finding that aligns with the work of Mohamed (2023) on mapping regional trends in e-commerce. However, the reliance on synthetic data limits the generalizability of these findings, suggesting that future studies should use real-world data for more accurate regional predictions.

Lastly, the linear regression model used for CLV prediction showed moderate accuracy (R² = 0.4987). This outcome is in line with other studies that have found traditional linear regression models to struggle with the complexity of customer behavior in e-commerce (Zorina, 2019). Incorporating ensemble learning techniques could improve the model’s predictive power, an area for future development.

## Assumptions, Limitations, and Recommendations

This dissertation encountered a few challenges, limitations, and assumptions that may have influenced the scope and outcomes of the research. Despite these hurdles, businesses can still achieve better results by applying practical solutions to work around these limitations.

### Assumptions

1. Synthetic Data: This research assumes that synthetic data accurately mirrors real-world customer behavior. While it provides the benefit of ensuring privacy, synthetic data may not fully capture the complexity and variability of actual datasets, which could limit how well the results apply to real-world scenarios.
2. Linear Relationship in CLV Prediction: The CLV prediction model assumes a linear relationship between customer metrics (Recency, Frequency, and Monetary value) and future value, which may overlook non-linear patterns and complex interactions present in customer behavior.
3. DBSCAN Parameter: DBSCAN’s performance is highly dependent on the right parameter settings which is manually set, especially the epsilon value. Manually adjusting these parameters can sometimes lead to less ideal clustering results, particularly when dealing with noisy or complex data.
4. Customer Engagement Metrics: Metrics like session length and page views are used as predictors of customer behavior, but they may not capture external influences, such as seasonal trends or marketing campaigns, which could affect customer actions in ways these simplified metrics don’t reflect.

### Limitations And Recommendations

1. Synthetic Data:

Limitation: Synthetic data does not fully capture the richness and unpredictability of real-world customer data such as seasonal fluctuations or the influence of external factors, limiting the accuracy of insights.  
Recommendation: Future work should involve collaboration with businesses to access anonymized real-world data, improving model testing and validation.

1. Model Performance:

Limitation: The reliance on simpler models like linear regression may oversimplify customer behavior, missing out on complex patterns.  
Recommendation: Exploring advanced models such as decision trees, ensemble methods, or neural networks would likely enhance prediction accuracy, allowing businesses to better target high-value customers.

1. Computational Power:

Limitation: The analysis was constrained by the local computational resources, limiting the size and complexity of datasets that could be processed.  
Recommendation: Moving to cloud-based platforms like AWS or Google Cloud would allow for the analysis of larger datasets and real-time processing.

1. Parameter Tuning in DBSCAN:

Limitation: Manually assuming and tuning DBSCAN parameters, such as epsilon, can lead to suboptimal clustering results, especially in noisy datasets.  
Recommendation: Automating hyperparameter tuning through methods such as grid search or random search automation methods would optimize clustering performance, and improve segmentation accuracy.

1. Geospatial Analysis:

Limitation: The geospatial analysis in this dissertation relies on synthetic data, which may lack the differences between the real-world geographic patterns.  
Recommendation: Integrating real geospatial data will provide more precise and meaningful regional insights, allowing businesses to tailor marketing strategies to specific local preferences and customer behaviours.

1. Prototype Scalability:

Limitation: The current prototype lacks scalability and features and is not fully integrated with all the details or functions, limiting its practical application.  
Recommendation: Expanding the prototype’s infrastructure to cloud-based solutions would enable integration with real-world applications, improve operational efficiency, and support real-time customer segmentation and analysis.

## Academic Achievement

This dissertation helped my academic growth in several important areas, showcasing my readiness to apply these skills in professional settings and future research. I’ve gained valuable insights and developed key skills in advanced data analytics, and machine learning. I have learned how to apply RFM analysis and clustering techniques like K-Means and DBSCAN to carry out customer segmentation in e-commerce. I also got experience in building machine learning models using tools like Scikit-learn and NumPy. In addition, I have gained practical experience in developing scalable applications by working with React and Django, where I integrated customer segmentation and details into a working prototype. This experience helped me bridge the gap between theoretical models and practical software solutions, further strengthening my ability to bring data-driven insights into real-world applications.

## Project Management

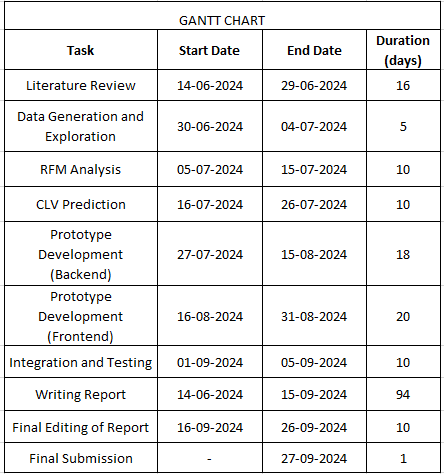


Table 1: GANTT Chart

Completing this dissertation required not only careful planning but also effective self-management to stay on top of the project timeline, resources, and potential risks. To keep things on track, the project was broken down into several phases, each with clear milestones like data generation, model development, prototype creation, and testing.

A Gantt chart was used to visualize tasks and deadlines, with each phase given a specific time frame to ensure a good balance between both the theoretical and practical components. Most of the time was dedicated to developing website prototype development and the final testing phase, which involved integrating the frontend and backend systems.

Version control through GitHub played a key role in saving the code throughout the different phases, making it easy to access between development stages. For document management, Google Drive provided a centralized platform for storing and sharing project materials. However, using additional tools like Slack or Trello for real-time task management could have further streamlined the process and allowed me to stay even more responsive to unexpected challenges.

## Relevant Legal, Social, Ethical, And Professional Issues

Legal Issues: The use of synthetic data avoided privacy concerns tied to regulations like GDPR (General Data Protection Regulation), but in real-world scenarios, handling actual customer data will require strict compliance with data protection laws. Businesses will need data anonymization, user consent, and secure handling protocols to avoid legal risks associated with customer data misuse.

Social Issues: Customer segmentation can lead to potential social concerns, such as misclassification or unfair targeting, which might negatively impact customer trust. The project focused on creating fair segmentation to prevent the alienation of certain customer groups. In practice, ensuring inclusive marketing strategies and avoiding exploitation of vulnerable segments will be critical to maintaining social responsibility.

Ethical Issues: Algorithmic bias presents a practical challenge when segmenting customers or predicting their behavior. Without careful bias testing, certain groups may receive unequal treatment, leading to unethical marketing practices. In future applications, it will be important to emphasize fairness in algorithmic decisions and ensure transparency around customer data use.

Professional Issues: Integrating real-time customer data requires maintaining data integrity and accountability throughout development. Adhering to professional standards such as proper documentation and version control helped ensure responsible development. Scaling the system will require strong data governance and continuous testing to meet industry best practices and professional ethics.

## Risk Management

The risk assessment chart (Table 2) outlines potential legal, social, and ethical risks associated with the project, evaluating their impact and likelihood on a 1-5 scale (1 being the lowest and 5 being the highest) and proposing mitigation strategies to address these concerns effectively.

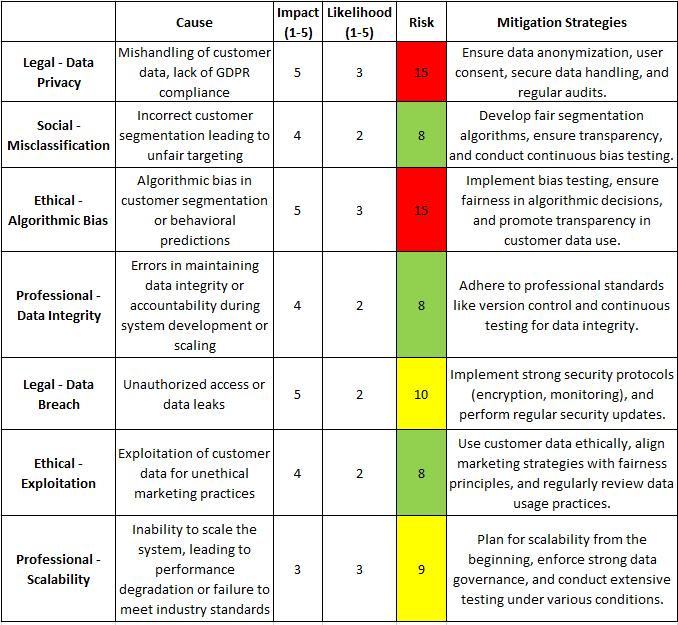


Table 2: Risk Assessment Chart

## Future Work

Collaboration with businesses to access anonymized real-world customer data will allow for more robust testing and validation of the models, resulting in more accurate and actionable insights. Additionally, implementing advanced models for CLV prediction, such as Gradient Boosting, Random Forests, or Neural Networks, could capture more complex customer behaviours, improving prediction accuracy beyond the limitations of simpler models.

To enhance responsiveness, the integration of real-time data processing through technologies like Apache Kafka or Spark Streaming would enable dynamic tracking of customer behavior, facilitating immediate marketing interventions. A move towards cloud infrastructure, leveraging platforms such as AWS or Google Cloud, would provide the necessary scalability to handle larger datasets, ensuring efficient computational performance.

Moreover, employing automated hyperparameter tuning methods, such as Grid Search or Bayesian Optimization, would streamline the clustering process for algorithms like DBSCAN and K-Means, eliminating the need for manual adjustments. Expanding the dataset to include additional behavioral metrics, such as social media activity, browsing habits, and customer reviews, would offer deeper insights into customer behavior and engagement.

The frontend interface could be upgraded for a better user experience and provide more dynamic visualizations, such as interactive dashboards that allow for deeper data exploration. Adding customizable filtering options for viewing details and segmenting customers based on various metrics like location would improve its usability.

Finally, exploring cross-industry applications of these methods in sectors like retail banking, telecommunications, or healthcare would demonstrate the flexibility of the models across diverse industries, allowing for broader adoption of data-driven customer insights. Also, expanding the prototype's reach by making it compatible with mobile devices would allow users to access insights on the go, increasing its flexibility and convenience of use.

# Conclusion

The tools and technologies that were used such as Python, pandas, and Scikit-learn were highly effective for managing data and performing machine learning tasks. K-Means and DBSCAN both provided strong customer segmentation results, with K-Means excelling at handling structured data and DBSCAN being great for dealing with noise and outliers. However, a challenge with DBSCAN was the manual parameter tuning, and automating this process could lead to better results. Principal Component Analysis (PCA) helped simplify data visualization, though some details may have been lost during the dimensionality reduction.

The prototype and analysis methods developed in this project mark a significant improvement in how e-commerce platforms can segment customers and analyse their behavior. By integrating RFM analysis, hybrid clustering techniques, and geospatial insights, this research gives businesses the tools to better understand their customer base and optimize their marketing strategies in a smarter, more data-driven way. The use of synthetic data also demonstrates how businesses can balance privacy concerns while still gaining useful insights, making these methods valuable for both research and real-world applications. In the future, applying these techniques to real customer data could help businesses fully harness the power of predictive analytics to drive growth and enhance customer satisfaction.

# References

1. Yıldız, E., Sen, C., & Işık, E. E. (2023). A Hyper-Personalized Product Recommendation System Focused on Customer Segmentation: An Application in the Fashion Retail Industry.
2. Pedregosa, F. et al. (2011) 'Scikit-learn: Machine learning in Python', the Journal of machine Learning research, 12, pp. 2825-2830.
3. Joung, J. and Kim, H. (2023) 'Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews', International Journal of Information Management, 70, pp. 102641.
4. John, J. M., Shobayo, O. and Ogunleye, B. (2023) 'An exploration of clustering algorithms for customer segmentation in the UK retail market', Analytics, 2(4), pp. 809-823.
5. Sharma, P., Chakraborty, A. and Sanyal, J. 'Machine learning based prediction of customer spending score'. 2019: IEEE, 1-4.
6. Singh, I. and Singh, S. (2017) 'Framework for targeting high value customers and potential churn customers in telecom using big data analytics', International Journal of Education and Management Engineering, 7(1), pp. 36-45.
7. Carrola, A. (2022) Synthesizing Realistic Substitute Data for a Law Enforcement Database using a Python Library. West Virginia University.
8. Sharaf Addin, E. H. et al. (2022) 'Customer mobile behavioral segmentation and analysis in telecom using machine learning', Applied Artificial Intelligence, 36(1), pp. 2009223.
9. Tang, R. et al. (2023) 'Does synthetic data generation of llms help clinical text mining?', arXiv preprint arXiv:2303.04360.
10. Wei, J. et al. (2023) 'Simple synthetic data reduces sycophancy in large language models', arXiv preprint arXiv:2308.03958.
11. Shen, B. 'E-commerce customer segmentation via unsupervised machine learning'. 2021, 1-7.
12. Mandal, P. C. (2022) 'Marketing information and marketing intelligence: generation of customer insights', International Journal of Technology Diffusion (IJTD), 13(1), pp. 1-14.
13. Cao, J. et al. (2022) '[Retracted] The Impact of Self‐Efficacy and Perceived Value on Customer Engagement under Live Streaming Commerce Environment', Security and Communication Networks, 2022(1), pp. 2904447.
14. Maria, I., Wijaya, V. and Keni, K. (2021) 'Pengaruh information quality dan service quality terhadap perceived value dan konsekuensinya terhadap customer engagement behavior intention (Studi pada social commerce Instagram)', Jurnal Muara Ilmu Ekonomi dan Bisnis, 5(2), pp. 321-334.
15. Safari, F., Safari, N. and Montazer, G. A. (2016) 'Customer lifetime value determination based on RFM model', Marketing Intelligence & Planning, 34(4), pp. 446-461.
16. Bachtiar, F. A. 'Customer segmentation using two-step mining method based on RFM model'. 2018: IEEE, 10-15.
17. Li, P. et al. 'An E-commerce customer segmentation method based on RFM weighted K-means'. 2022: IEEE, 61-68.
18. Monalisa, S. and Kurnia, F. (2019) 'Analysis of DBSCAN and K-means algorithm for evaluating outlier on RFM model of customer behaviour', Telkomnika (Telecommunication Computing Electronics and Control), 17(1), pp. 110-117.
19. Parikh, Y. and Abdelfattah, E. 'Clustering algorithms and RFM analysis performed on retail transactions'. 2020: IEEE, 0506-0511.
20. Marmol, M. et al. (2021) 'Maximizing customers' lifetime value using limited marketing resources', Marketing Intelligence & Planning, 39(8), pp. 1058-1072.
21. Kumar, A. et al. 'Customer Lifetime Value Prediction: Using Machine Learning to Forecast CLV and Enhance Customer Relationship Management'. 2023: IEEE, 1-7.
22. Chen, X., Reibman, A. and Arora, S. (2022) 'Sequential recommendation model for next purchase prediction', arXiv preprint arXiv:2207.06225.
23. Asadi Ejgerdi, N. and Kazerooni, M. (2024) 'A stacked ensemble learning method for customer lifetime value prediction', Kybernetes, 53(7), pp. 2342-2360.
24. Su, H. et al. 'Cross-domain adaptative learning for online advertisement customer lifetime value prediction'. 2023, 4605-4613.
25. Li, S. and Han, D. 'A Comparative Study on k-Means Clustering with Different Cluster Representations'. 2023: IEEE, 7959-7964.
26. Siagian, R., Sirait, P. S. P. and Halima, A. (2021) 'E-Commerce Customer Segmentation Using K-Means Algorithm and Length, Recency, Frequency, Monetary Model', Journal Of Informatics and Telecommunication Engineering, 5(1), pp. 21-30.
27. Rawat, B. S. et al. 'Hybrid Clustering Techniques for Optimizing Online Datasets Using Data Mining Techniques'. 2023: IEEE, 1-5.
28. Li, Y. et al. (2023) 'Customer segmentation using K-means clustering and the hybrid particle swarm optimization algorithm', The Computer Journal, 66(4), pp. 941-962.
29. Ghahramani, M. et al. (2021) 'Intelligent geodemographic clustering based on neural network and particle swarm optimization', IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52(6), pp. 3746-3756.
30. Acharya, A. et al. (2022) 'Geospatial analysis of geo-ecotourism site suitability using AHP and GIS for sustainable and resilient tourism planning in West Bengal, India', Sustainability, 14(4), pp. 2422.
31. Dankar, F. K. and Ibrahim, M. (2021) 'Fake it till you make it: Guidelines for effective synthetic data generation', Applied Sciences, 11(5), pp. 2158.
32. Van Doorn, J. et al. (2010) 'Customer engagement behavior: Theoretical foundations and research directions', Journal of service research, 13(3), pp. 253-266.
33. Rácz, A., Bajusz, D. and Héberger, K. (2021) 'Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification', Molecules, 26(4), pp. 1111.
34. Ramos, J. M. A. M. and Silva, F. A. 'Customer Lifetime Value Prediction: A Machine Learning Approach'. 2023: SBC, 486-500.
35. Marukatat, S. (2023) 'Tutorial on PCA and approximate PCA and approximate kernel PCA', Artificial Intelligence Review, 56(6), pp. 5445-5477.
36. Marisa, F. et al. (2023) 'POTENTIAL CUSTOMER ANALYSIS USING K-MEANS WITH ELBOW METHOD', JIKO (Jurnal Informatika dan Komputer), 7(2), pp. 307-312.
37. Shahapure, K. R. and Nicholas, C. 'Cluster quality analysis using silhouette score'. 2020: IEEE, 747-748.
38. Ros, F., Riad, R. and Guillaume, S. (2023) 'PDBI: A partitioning Davies-Bouldin index for clustering evaluation', Neurocomputing, 528, pp. 178-199.
39. Favretto, A. (2014) 'Coordinate questions in the web environment', Cartographica: The International Journal for Geographic Information and Geovisualization, 49(3), pp. 164-174.
40. Mohamed, S. et al. 'Implementation of a Geographical Information System (GIS) for E-commerce'. 2023: Springer, 471-485.
41. Zorina, K. (2019) 'Building segment based revenue prediction for CLV model'.