1. INTRODUCTION

Understanding consumer behavior is more crucial than ever in the quickly evolving world of e-commerce. Because of the growing number of digital platforms, businesses are relying more and more on data-driven strategies to enhance customer satisfaction, loyalty, and retention. Apart from significantly affecting the conventional retail sector, the emergence of e-commerce has presented businesses with numerous prospects to leverage customer information for their benefit. This dissertation advances e-commerce strategies through advanced customer segmentation and behavioral analysis techniques using a combination of simulated data and modern analytical models.

Customer segmentation, or dividing of a customer base into distinct groups based on common attributes, is an important part of marketing strategy. In the past, segmentation was done with basic demographic or transactional data, which allowed businesses to concentrate their marketing efforts on specific client segments. However, the drawbacks of conventional approaches have become clear as customer data has become more varied and complex. More accurate segmentation techniques have been developed through the application of machine learning and data analytics. By employing these techniques, businesses can create more tailored and effective marketing campaigns and obtain a deeper understanding of consumer behavior.

Recency, Frequency, and Monetary analysis, or RFM analysis, is a widely used technique for customer segmentation. It evaluates customers based on their previous purchases. To identify high-value customers, RFM analysis examines how much, how often, and how recently customers have spent. It's an easy method of locating customers who buy things. RFM analysis is a helpful tool for many businesses, but it often lacks the capacity to handle complex and varied data. Experts have begun combining multiple clustering algorithms with traditional RFM analysis to overcome these obstacles. Through the integration of multiple algorithms, these methodologies enhance the precision of segmentation and yield more valuable and precise customer insights.

Predicting the Customer Lifetime Value (CLV), in addition to RFM analysis, has become an essential part of today's e-commerce strategy. By using CLV prediction to estimate the future value that a customer will bring, businesses can maximize their marketing budgets and prioritize high-value customers.

Understanding consumer behavior has additionally increased the significance of integrating data from various sources. E-commerce platforms today have access to a huge quantity of data from various sources, including transactional data, social media activity, online browsing habits, and customer reviews. Businesses can integrate these diverse data sources to generate comprehensive client profiles that provide a detailed understanding of customer behavior. This multi-source data integration is particularly helpful for creating targeted advertisements because it enables businesses to understand not only what their customers are buying, but also why they are buying it. By utilizing insights from multiple data sources, companies can create more successful and tailored marketing strategies that connect with their target audience.

K-Means and DBSCAN are powerful clustering algorithms that complement each other in customer segmentation tasks. K-Means partitions data into pre-defined clusters by reducing the distance between data points and their centroids, making it useful for structured datasets with spherical clusters. But it has issues with noise and needs us to indicate how many clusters we want. DBSCAN addresses these limitations by identifying clusters of any random shapes and detecting noise without needing a predefined cluster count. Businesses can achieve more accurate segmentation by combining the strengths of K-Means and DBSCAN. This allows them to handle diverse and noisy customer data effectively and gain deeper insights.

Another very popular area in e-commerce analytics is geospatial analysis. Learning about the shopping habits and geographical spread of clients can be very helpful in determining local trends and preferences. With the use of geospatial analysis tools, we can display customer data on a map. This geographic perspective is crucial for businesses looking to expand into new markets or target specific regions with their marketing campaigns. Additionally, companies can develop more insightful marketing campaigns and a deeper understanding of customer behavior by fusing geospatial data with other behavioral metrics like RFM scores or CLV predictions.

In summary, I will improve e-commerce strategies by incorporating behavioral analysis and advanced customer segmentation techniques. Using RFM analysis combined with K-Means and DBSCAN, I will improve the customer segmentation for better accuracy in identifying high-value customers. In addition, in order to prioritize marketing efforts, I will estimate Customer Lifetime Value (CLV). Regional trends can be found through geospatial analysis, and a good understanding of customer behavior can be obtained through the integration of multiple data sources. Using simulated datasets created with the Python Faker library, I will be guaranteeing ethical and privacy compliant analysis that will help businesses increase customer satisfaction, retention, and engagement.

* 1. AIM

This dissertation intends to improve customer segmentation and behavioral analysis through the development of a complete structure that incorporates the modern data analytics techniques. This research aims to give businesses actionable insights for more individualized, focused, and successful marketing strategies while maintaining adherence to ethical and privacy standards through the use of simulation, predictive modelling, clustering techniques, geospatial analysis, and practical applications.

* 1. OBJECTIVES

This dissertation's objective is to create a complete structure that uses data analytics methods to improve customer segmentation and behavioral analysis in order to improve e-commerce strategies. The goal is to integrate various methodologies in order to obtain deeper insights into consumer behavior and help businesses develop more focused and successful promotional strategies. In order to accomplish this, the dissertation sets out a number of particular goals intended to fill in the gaps in the field of e-commerce analytics and offer feasible suggestions for practical applications.

* + 1. Simulation of data using the python "faker" library

It is crucial to use datasets that closely resemble actual customer data in order to develop and test the suggested methodologies. The creation of simulated datasets that imitate different behavioral metrics, and customer transactions details, and other customer interactions is necessary. The "Faker" library for Python will be used to accomplish this, as it enables the creation of realistic synthetic data while upholding data security and privacy regulations. This dissertation attempts to provide a secure setting for testing various analytical techniques by simulating diverse datasets, ensuring that the suggested approaches are flexible and applicable to a broad range of e-commerce scenarios.

* + 1. Conducting RFM Analysis and CLV Prediction to identify High-Value Customers

The next goal is to identify high-value clients by finding customer lifetime value (CLV) and recency, frequency, and monetary (RFM) analysis. RFM analysis will be conducted by calculating three key metrics for each customer: Recency (time since last purchase), Frequency (number of purchases), and Monetary value (total amount spent). RFM analysis is a well-established technique for assessing customers based on their purchasing patterns, offering a simple approach to divide clients into various categories. However, future customer value is not taken into account by RFM analysis.

RFM metrics (Recency, Frequency, Monetary Value) and behavioral characteristics are combined to create a linear regression model that is used to predict Customer Lifetime Value (CLV) with accuracy. By applying a train/test split to the dataset and using Linear Regression to fit the model, our aim is to give businesses actionable insights to optimize customer segmentation, forecast future revenue, and enhance the allocation of marketing resources. To measure prediction accuracy and model fit, the Mean Absolute Error (MAE) and R-squared (R²) metrics are used to evaluate the performance of the model. By combining these two strategies, companies will be able to better target their marketing efforts and give priority to their most valuable clients, which will increase long-term profitability.

* + 1. Implementation of Hybrid Clustering Techniques

In order to achieve more precise and dependable customer segmentation, the first goal is to apply hybrid clustering techniques, which combine the advantages of multiple clustering algorithms, specifically DBSCAN and K-Means. Conventional clustering techniques, like K-Means, are frequently employed for splitting clients into groups according to similarity, yet, they are open to outliers and assume spherical clusters. On the other hand, DBSCAN is more immune to noise and can detect clusters of different sizes and shapes, but it requires precise adjustment of parameters like the epsilon distance and minimum number of points. This goal is to develop a more adaptable and reliable segmentation strategy that can better manage the diversity of customer data in e-commerce by combining these algorithms into a hybrid model. The silhouette score and Davies-Bouldin index, two cluster validity indices, will be used to assess the hybrid clustering model's efficiency. By measuring the compactness and separation of clusters, these indices will aid in making sure that the segmentation process successfully discerns between various customer behavior and preferences. Businesses will be able to recognize different customer groups more precisely thanks to this improved segmentation accuracy, which will result in more individualized and powerful marketing campaigns.

* + 1. Performing Geospatial Analysis to Understand Regional Trends and Preferences

Using geospatial analysis to investigate and visualize the geographic distribution of clients and their purchasing patterns is the next goal. Businesses hoping to enter new markets or target particular geographic areas with their strategies must have a thorough understanding of the geographic distribution of their customers and the regional variations in their preferences. Using Geographic Information Systems (GIS) tools, this goal involves mapping customer data and spotting regional trends. In order to provide a broader understanding of customer behavior that takes into account both behavioral and geographic dimensions, this dissertation integrates geospatial data with other customer metrics, such as RFM scores and CLV predictions. Businesses can use the observations collected from geospatial analysis to create more focused marketing strategies that suit local consumer preferences and trends.

* + 1. Development of a Prototype Application to Demonstrate Practical Applications

The creation of a prototype application displaying the real-world implementation of the advanced customer segmentation and behavioral analysis methods this dissertation suggests is our last goal. This prototype will be constructed with React on the front end and Django on the back end to offer a stable and expandable framework for applying the techniques discovered during this study. Users will be able to enter customer data into the application, and it will use the integrated models to produce reports on behavioral analysis and segmentation. The functionality of the prototype will be evaluated by making sure the application successfully displays the real-world uses of the developed methodologies.

* + 1. Addressing Ethical and Privacy Considerations

While some privacy concerns are eliminated when working with synthetic data, it is important to make sure that the methodologies developed adhere to ethical standards and data privacy regulations such as General Data Protection Regulation (GDPR). This goal focuses on recognizing and resolving potential ethical problems with behavioral analysis and customer segmentation, such as biases in models and data, fairness in segmentation, and openness in data use. This dissertation seeks to promote responsible use of advanced data analytics in e-commerce and to establish trust by making sure that the developed techniques are both ethically sound and privacy-compliant.

* 1. **SUMMARY OF THE EXISTING LITERATURE**

Businesses are depending more and more on data-driven strategies to enhance customer segmentation, forecast Customer Lifetime Value (CLV), and boost marketing initiatives as e-commerce expands. Conventional approaches, such as RFM analysis and simple clustering techniques, frequently fail to capture the complexity of consumer behavior. By combining hybrid clustering techniques, utilizing synthetic data to protect privacy, and combining multi-source data, including behavioral and geospatial information, for a thorough analysis, this dissertation aims to overcome these constraints. This research intends to improve customer engagement and profitability by optimizing e-commerce strategies through the use of machine learning models. These models will improve customer segmentation accuracy, CLV predictions, and overall marketing efficiency.

* + 1. Data Generation

Real customer data is often used for e-commerce analytics, which raises security and privacy concerns. Therefore, a lot of researchers use tools for creating synthetic data. A popular library for creating realistic but anonymized datasets in Python is called "Faker" as highlighted by Anthony Carrola (2022). These datasets allow for the evaluation and verification of data models without sacrificing security since they replicate real-world consumer behaviors while following privacy regulations. The Faker library was used in a study by Addin (2022), to create telecom customer datasets that allowed for the analysis of behavioral and demographic data. These methods are being used more and more in a variety of industries, such as e-commerce and healthcare. (Tang et al., 2023)

For my dissertation, I'm creating artificial datasets that replicate actual consumer behavior by using the "Faker" library. These datasets will contain purchase history, demographic information, and behavioral data such as time spent on-site and engagement metrics.

By creating realistic yet anonymized datasets, the complexities of customer behavior for precise predictions while adhering to security and privacy regulations is achieved (Wei 2023). The synthetic data makes it possible to test advanced segmentation and predictive models by resembling customer transactions and behavioral interactions. This guarantees that the dissertation's conclusions can be applied to actual e-commerce settings. This method accomplishes the objectives of data-driven customer behavior analysis while also guaranteeing ethical compliance.

* + 1. Multi-Source Data Integration

Conventional customer segmentation models mainly use transactional data, which means they frequently overlook insightful information from other sources like social media, browsing habits, and geographic data. As noted by Shen (2021), most segmentation techniques are constrained by the amount of data they can analyze, which frequently leaves out important details about the complexity of modern consumer behavior.

I'm suggesting in this dissertation that several data sources be integrated, such as transactional, behavioral, and geographic data. This thorough method provides an extensive knowledge of customer behavior by utilizing data from multiple sources to create a comprehensive customer profile. Mandal (2022).

The integration of multiple data sources will enhance the effectiveness of customer segmentation. A study by Yıldız (2023) creates hyper-personalized product recommendations by combining extra behavioral data with customer segmentation based on RFM analysis. Research has demonstrated that elements like perceived value and self-efficacy have a big impact on consumer engagement in e-commerce. (Cao 2022). By taking into account extra metrics like time spent on-site and customer engagement, this approach will enable more sophisticated segmentation models and offer useful insights for customized marketing strategies. Customers are prone to make repeat purchases and spread the word about the brand as a result, which improves customer engagement (Maria).

* + 1. 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 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 et al. (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).

I’m suggesting combining 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 of 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 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 2020)

* + 1. 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. (Marmol 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 2022).

In my dissertation, I will add behavioral data like the frequency of logins, length of sessions, and pages visited to 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 2023).

* + 1. 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, s and han 2023). While effective in certain scenarios, K-Means also has limitations, its assumption of spherical clusters and its sensitivity to outliers. Study by Siagian (2021) have 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 (R Y H 2021). 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 y (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 y 2023). It has been demonstrated that the K-Means and DBSCAN algorithms increase segmentation accuracy in scenarios involving complex customer data. (John 2023). Businesses can create more individualized and focused marketing campaigns thanks to this improved segmentation, which raises customer engagement and profitability.

* + 1. Geospatial Analysis

Although geospatial analysis can yield valuable insights into regional trends and preferences, its application in customer segmentation has been underutilized (Ghahramani 2022). 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 demonstrate 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, I will combine behavioral, transactional, and geographic data in this dissertation. I will be mapping the locations of my customers using Geographic Information Systems (GIS) and other geospatial tools, and 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.

1. **METHODOLOGY**

This dissertation's methodology focuses on creating a solid framework for improving e-commerce strategies using behavioral analysis and customer segmentation. This section describes the particular procedures, tools, methods, and issues that must be taken into account in order to accomplish the objectives of the dissertation.

**2.1. Data Generation**

Simulated Data Creation Using the "Faker" Library

Given the constraints of using real customer data due to privacy and security concerns, this dissertation employs synthetic data generation using the Python "Faker" library. This approach allows us to create realistic, anonymized datasets that closely mimic real-world customer behavior and demographics without compromising privacy.

2.1.1. Customer Data Creation

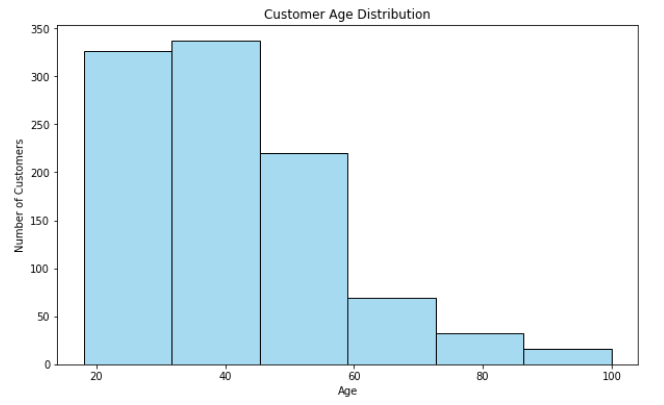
Creation of accurate synthetic customer profiles that reflect the complexity and diversity of actual customer bases by including demographic and geographic data.

1. Customer ID Generation:

The function generate\_unique\_customer\_id from the code creates a unique 5-digit customer ID for each customer by randomly generating an integer between 10000 and 99999. The uniqueness of each ID is ensured by checking against a list of already generated existing IDs (existing\_ids).

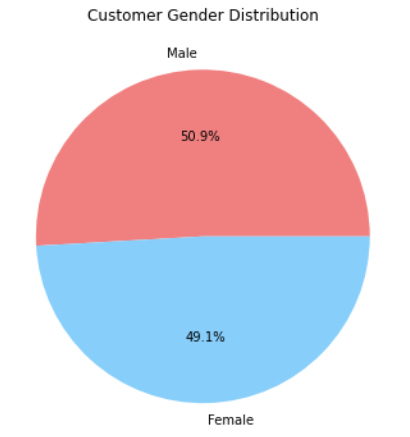
1. Customer Demographic Attributes:

* Age Distribution:



Each customer's age is assigned by using a weighted probability distribution over six age ranges. The probabilities reflect the typical age demographics of online shoppers.

* Gender Distribution:



A weighted random choice to assign a gender (Male or Female) to each customer is done based on a predefined distribution (46.9% female and 53.1% male), which is consistent with data from the general population.

* Location Data: To provide geographic context, latitude, and longitude are generated for each customer using the Faker library with a UK locale using the function “fake\_uk”. This resembles the variation in customers' locations.
* 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.

1. Data Storage:

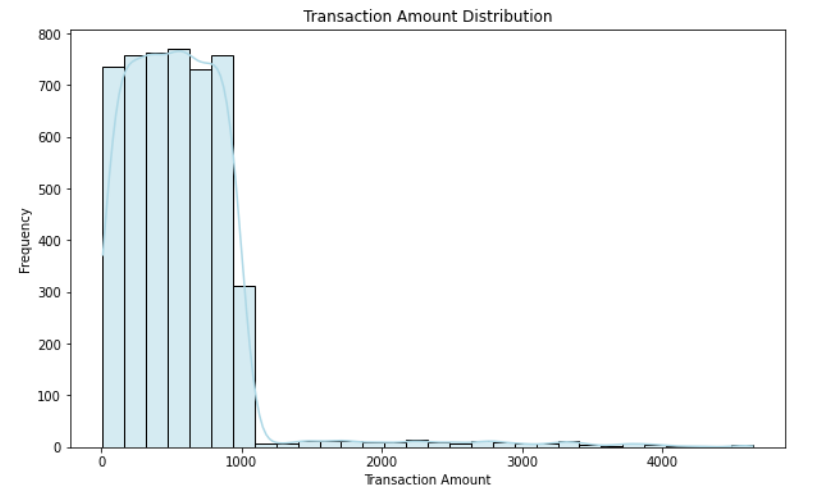
After being created, the customer data is saved in a Python list called customer\_data and subsequently transformed into a pandas Dataframe called customer\_df. For later use, this Dataframe is saved as a CSV file called "customers.csv."

* + 1. Transaction Data Creation:

Generation of accurate transaction records for every customer, including 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:



A uniform 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, simulating outliers.

* 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. The inclusion of a variety of attributes, including product category, payment method, and discount, allows for a multifaceted analysis of customer purchasing patterns and preferences.

1. Data Storage: After being created, the transaction data is saved in a Python list called transaction\_data and subsequently transformed into a pandas DataFrame. For later use, this DataFrame is saved as a CSV file called "transaction.csv."
   * 1. Behavioral Data Creation

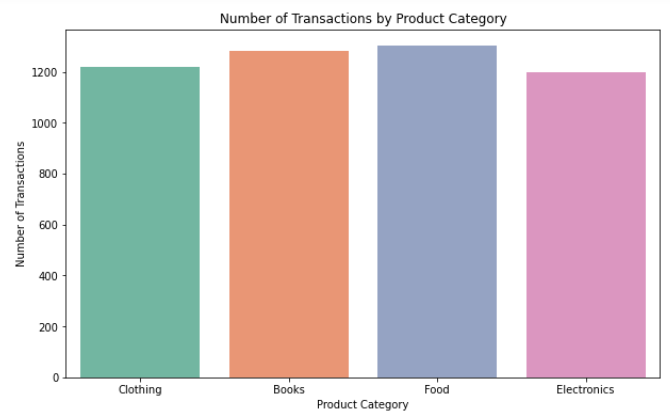
Production of behavioral data, such as frequency of logins, page views, and time spent on the site, that mimics user interactions with the e-commerce platform.

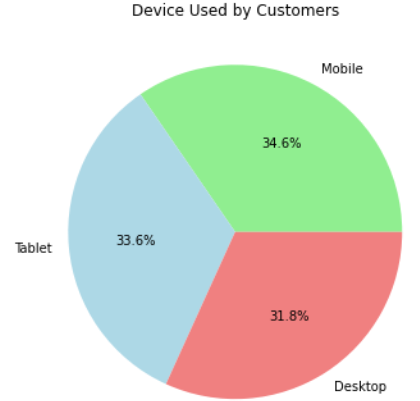
Steps and Techniques:

1. Behavioral Attributes:

* Login Data: Attributes like last\_login\_date, total\_login\_count, and pages\_visited are generated to simulate a customers usage of an 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.

Examples of how the data is distributed:





1. Data Storage: After being created, the behavioral data is saved in a Python list called behavioral\_data and subsequently transformed into a pandas Dataframe. For later use, this Dataframe is saved as a CSV file called "behavioral.csv."
   * 1. Data Validation and Matching

With a particular focus on customer IDs to confirm that every customer has corresponding entries in the transaction and behavioral datasets, we need to make sure that the generated data is consistent and prepared for analysis.

* Intersection of Customer IDs: Sets are created from the customer IDs in each DataFrame (customer\_ids\_customers, customer\_ids\_transactions, customer\_ids\_behavioral). To find out how many customers have corresponding entries in the transaction and behavioral data, the intersections of these sets are computed.
* Consistency Check: To make sure the data aligns properly, the number of common customer IDs between the transaction, behavioral, and customer datasets is computed and printed. To ensure accurate analysis and preserve data integrity, this step is essential.
  1. **Recency, Frequency, and Monetary (RFM) Analysis**

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.
  + 1. Data Preparation and steps for RFM analysis
* Convert Transaction Dates: pd.to\_datetime() is used to convert the transaction\_date column in the transaction DataFrame (transaction\_df) to datetime format. To calculate dates correctly, this conversion is required.
* Determine Latest Transaction Date: transaction\_df['transaction\_date'].max() is used to determine the dataset's latest transaction date. This date is used as a reference to determine how recent each customer's transactions are.
* Recency Calculation: The code calculates recency by determining the difference between the current (latest) transaction date and the last transaction date for 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 to bring all three metrics (R, F, M) to the same scale:

Standardized Value = Value – Mean Standard Deviation

This ensures that no single metric disproportionately influences the results during clustering.

* + 1. RFM Scoring

Tools:

* The seaborn and matplotlib library is used for generating and display the plot.
* The pandas library is used to manipulate and prepare the data.

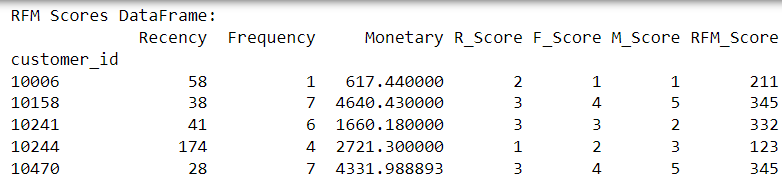
RFM scoring is used to convert the RFM values into quantile-based scores, which are then used to segment customers. This is done using the pd.qcut() function, which divides each RFM metric into quantiles and assigns a score.

* R Score: Customers with lower recency values (i.e., 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



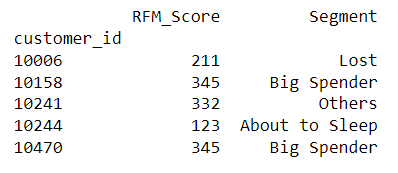
These scores are combined into an overall RFM score, such as "555" for the most valuable customers and "111" for the least valuable ones.

* + 1. Customer Segmentation Using RFM Scores

The function dynamic\_segment\_customer uses the RFM scores to segment customers into predefined groups based on their Recency, Frequency, and Monetary scores.

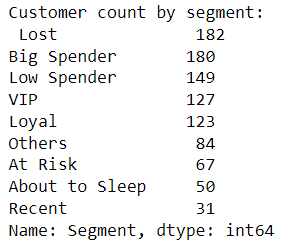
Segmentation Logic:

* 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), regardless of spending or frequency
* 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: Catch-all for customers not fitting the above categories



Customers are assigned an R\_Score, F\_Score, and M\_Score based on their RFM values. Higher scores indicate better customer behavior (e.g., recent purchases, frequent transactions, and higher spending). This segmentation helps in identifying which customers are most valuable, at risk, or recently acquired.

* + 1. Calculating the count of Customers in each segment



* + 1. 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

These features provide deeper insights into customer engagement beyond RFM analysis.

Loyalty and Churn Risk Scoring

Loyalty Score and Churn Risk Score are calculated to know customer insights.

* Loyalty Score:

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

This score weighs various customer activities to identify the most engaged and loyal customers.

* Churn Risk Score:

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

This score identifies customers at risk of churning based on their recent inactivity, low frequency of purchases, and low engagement.

Monetary Value and Purchase Consistency Scoring

* Monetary Value Score:

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

This score provides a holistic view of the customer’s financial value.

* Purchase Consistency Score:

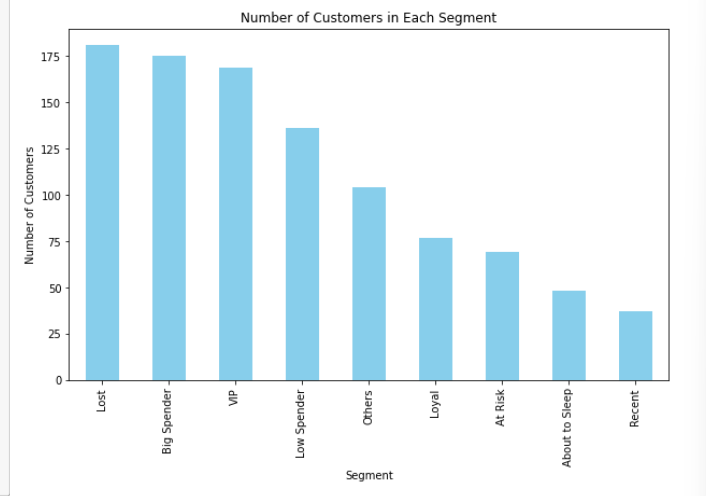
Purchase Consistency Score = (Frequency × 0.5) + (Number of Transactions × 0.3) + (1/ (Recency×0.2))

This score evaluates how consistently customers make purchases, combining both frequency and recency.

* + 1. 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. The bar plot represents the number of customers that fall into each of the predefined customer segments such as VIP, Big Spender, Loyal, Lost, and others.



The resulting bar chart helps businesses quickly assess the distribution of their customer base. For instance, segments with a high number of customers, such as "Lost" or "Big Spender," can be targeted with specific retention or promotional strategies, while segments with fewer customers, like "Recent" or "Loyal," may require further analysis to improve engagement.

**Customer Cohorts and Segment Distribution**

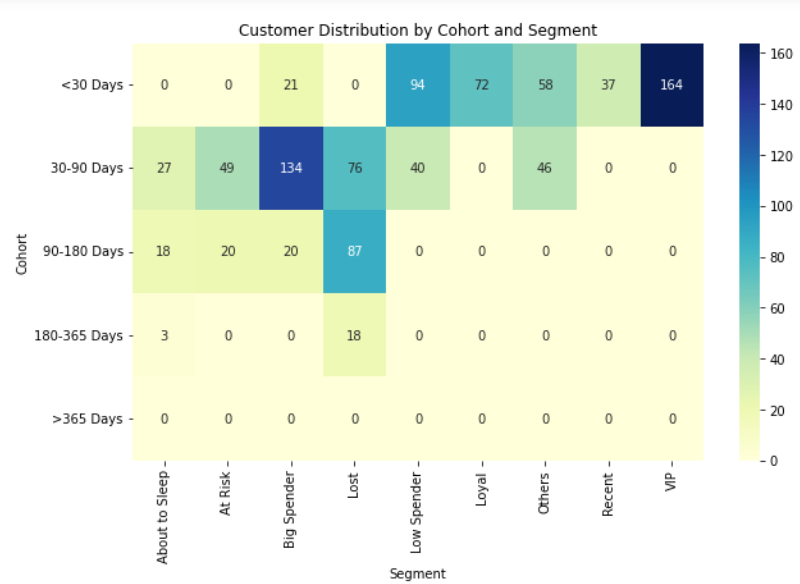
Cohort analysis is used to group customers based on their recent activity (Recency). Customers are divided into cohorts based on how many days have passed since their last purchase.

Cohort Binning Logic:

Customers are divided into cohorts using pd.cut() function, with bins for recency, These bins represent different levels of recency.

<30 Days, 30-90 Days, 90-180 Days, 180-365 Days, >365 Days

A heatmap is generated using seaborn.heatmap(), color-coding the customer counts across cohorts and segments, providing a visual understanding of customer distribution. This helps businesses understand how customer engagement changes over time.



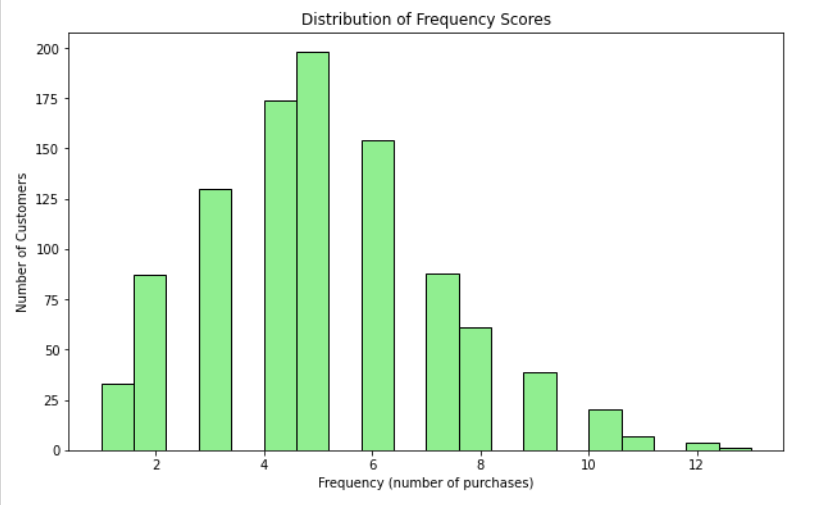
The plot shows that many customers in the "VIP" or "Big Spender" segments belong to the <30 Days cohort, indicating frequent and high-value transactions. Conversely, the "Lost" segment might be more prevalent in the >365 Days cohort, highlighting customers who have not interacted with the business in a long time.

By analyzing the distribution across cohorts, businesses can tailor marketing strategies to re-engage at-risk or lost customers, while also nurturing high-value customers.

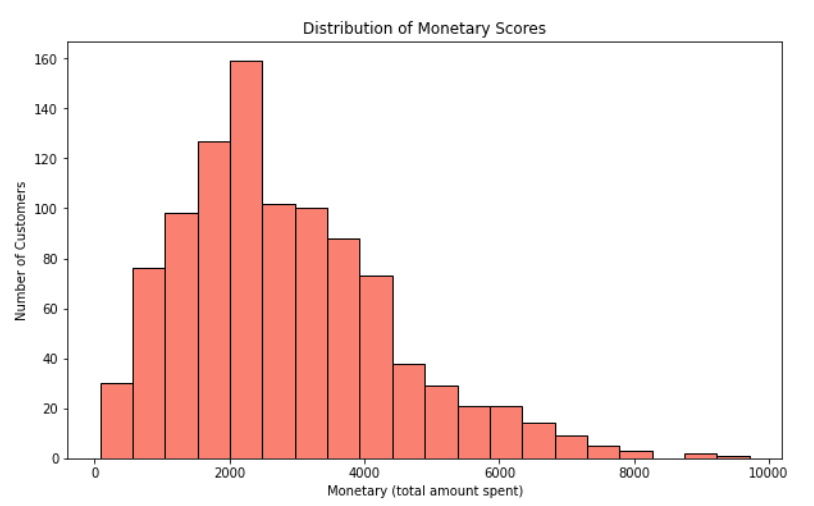
**Visualization of the distribution of customers based on Frequency and Monetary**

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.



The histogram provides insight into how frequently customers make purchases. Typically, you will see that a large number of customers have made fewer purchases, with a tapering off as the frequency increases. This information helps to identify whether the majority of the customer base consists of repeat buyers or one-time purchasers.



The monetary histogram helps identify how much money customers spend over time. This distribution often shows that most customers spend within a particular range, and only a few customers may be contributing to significantly higher revenues.

ROC Curve and AUC Analysis

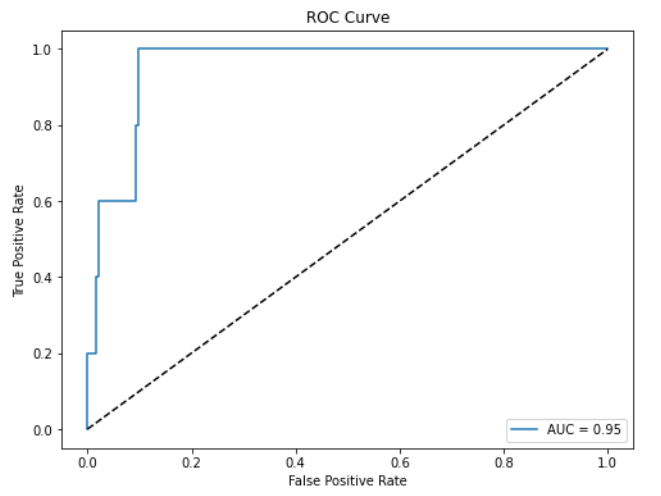
The ROC Curve (Receiver Operating Characteristic) and AUC (Area Under the Curve) are critical tools for evaluating the performance of binary classification models. In the context of RFM (Recency, Frequency, Monetary) analysis, these metrics help predict customer behaviors like churn and loyalty by assessing the model’s ability to classify customers effectively.

RFM scores along with additional behavioral data serve as features for a binary classification model to predict customer churn or loyalty.

* ROC Curve: Plot showing the performance of a binary classifier.
* AUC: Represents the model's ability to distinguish between classes.

Plot summary:

* 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.



* 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 invaluable for evaluating how well RFM and behavioral data predict customer behaviours such as churn or loyalty, helping businesses refine their strategies.

* 1. **Customer Lifetime Value (CLV) Prediction**

The primary goal of this methodology is to predict Customer Lifetime Value (CLV) using Linear Regression, based on RFM scores and additional behavioral metrics of customers. The code is designed to predict Monetary value (used as a proxy for CLV) and evaluate the model's performance using Mean Absolute Error (MAE) and R-Squared (R²).

* + 1. Data Preparation

The independent variables include a combination of Recency, Frequency, and various behavioral metrics such as total login count Pages visited, time spent on site, number of transactions, promo clicks, sessions per day, and social shares.

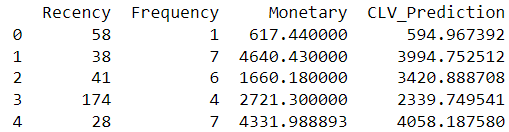
The target variable for the model is the Monetary value, which represents the total amount spent by the customer and is treated as a proxy for CLV. In real-world scenarios, this could be replaced with other parameters.

* + 1. Train-Test Split

The dataset is split into 80% training data and 20% test data using the train\_test\_split() method. This ensures that the model is trained on a portion of the dataset and tested on unseen data to measure its generalization ability.

* + 1. 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.



The performance of the model is measured using two key metrics:

Mean Absolute Error (MAE): Measures the average difference between the actual CLV (Monetary) and the predicted values.

R-Squared (R²): Measures how well the independent variables explain the variance in the target variable (Monetary). An R² value of 1 indicates a perfect fit, while a value of 0 means that the model does not explain any variance.



The mean absolute error indicates that, on average, the predicted CLV deviates by around 818 pounds from the actual value. The lower value could be because of the synthetic data does not adequately simulate the nuances of customer behavior, or as the model is not being tested against real-life variances in spending patterns.

The R-Squared value of 0.4987 suggests that approximately 50% of the variance in the target variable (Monetary) is explained by the features included in the model. This is a relatively low value as the variability in features might be limited or too simplistic compared to real-world data and potentially due to the lack of realistic correlations in the data. Real-world datasets often have more variance due to diverse customer behaviors.

* 1. **K-Means and DBSCAN Clustering** 
     1. Merging RFM and Behavioral DataFrames

The first step involves merging two datasets, RFM data and behavioral data. This step combines customers’ purchasing behaviors with additional behavioral metrics like login counts, time spent on the website, and pages visited using a function pd.merge(). Merging these datasets allows for a more comprehensive understanding of customer activity.

The next step selects relevant features for clustering analysis. The selected features include both RFM metrics (Recency, Frequency, Monetary) and behavioral data such as total\_login\_count, pages\_visited, and time\_spent\_on\_site. These features represent customer behaviors and spending habits.

* + 1. Standardizing the Features:

Since the features have different scales, standardization is performed using StandardScaler to normalize the feature values to have a mean of 0 and a standard deviation of 1. This ensures that no single feature dominates the clustering process due to scale differences.

Function Used: StandardScaler()

Formula: Xscaled = (X−μ)/ σ

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

* + 1. Dimensionality Reduction Using PCA:

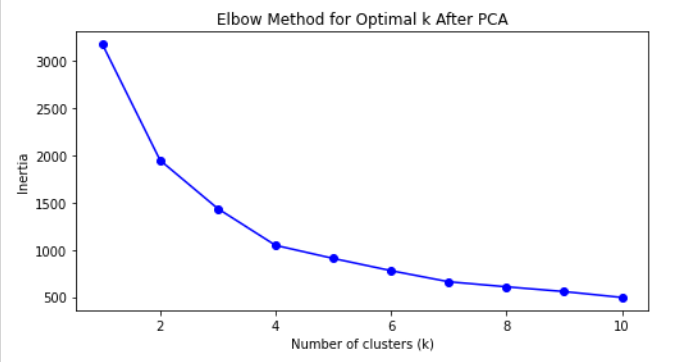
Principal Component Analysis (PCA) is applied to reduce the dimensionality of the feature space to two components. Reducing the dimensions simplifies the clustering process and helps visualize the data in a 2D space without losing much information.

Function Used: PCA(n\_components=2)

Formula: PCA works by projecting the data onto new axes (principal components) that maximize variance, essentially capturing the most informative dimensions.

* + 1. Elbow Method for Optimal k:

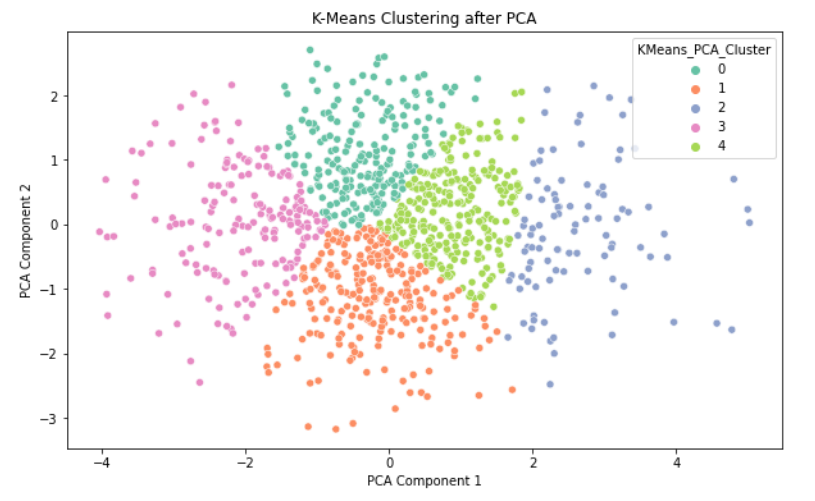
The Elbow Method is used to determine the optimal number of clusters, k, for K-Means clustering. Inertia is calculated for different values of k and plotted to visualize the "elbow" point, which represents the optimal number of clusters. This is done using a function KMeans() with kmeans.inertia\_. After determining the optimal number of clusters, each customer is assigned to a cluster based on proximity to the nearest centroid.



The plot shows inertia (how much the data points differ from their cluster centers) on the y-axis and the number of clusters (k) on the x-axis. The "elbow" point happens at k = 5, where the decrease in inertia becomes less noticeable. This point shows that increasing the number of clusters beyond 5 doesn’t improve the model much. Therefore, using 5 clusters offers a good balance between keeping the model simple and capturing important patterns in the data.

* + 1. Visualizing Clusters After PCA:

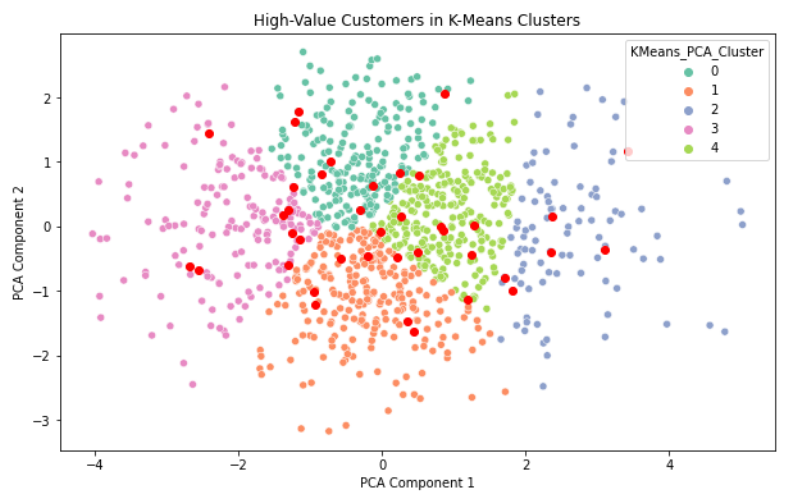
The clustered data is visualized in 2D using the two principal components. Each point represents a customer, and the color corresponds to the cluster assignment.



This plot shows the K-Means clustering results after applying PCA (Principal Component Analysis) to reduce the data to two dimensions. Each color represents a different cluster, and the points are grouped based on similarities in the data. The x-axis and y-axis represent the two main components from the PCA, which capture the most important information from the original features. The clear separation of colors indicates that the K-Means algorithm was able to effectively divide the data into 5 clusters, as identified by the earlier Elbow Method

* + 1. High-Value Customer Visualization:

High-value customers are visualized within the clustering space, this allows for a deeper understanding of how high-value customers are similar or differ from the general customer base.



This plot displays the high-value customers (highlighted in red) within the K-Means clusters, based on Principal Component Analysis (PCA). These high-value customers were identified by analyzing their RFM scores (Recency, Frequency, Monetary). Customers who scored highly across all three metrics, typically represented by an RFM score of '555', were considered high-value. This score indicates that they made recent purchases, do so frequently, and spend significant amounts of money.

The distribution of high-value customers across multiple clusters is important because it reflects the diversity in their behavior. For instance, some might make frequent smaller purchases, while others might make fewer but larger ones. Understanding this variation allows businesses to create targeted engagement strategies for each cluster. Additionally, by analyzing their placement across clusters, businesses can optimize resource allocation, ensuring that each type of high-value customer receives the appropriate attention to maximize their lifetime value.

* + 1. Silhouette Score and Davies-Bouldin Index for K-Means:

Two metrics are used to evaluate the quality of the clustering:

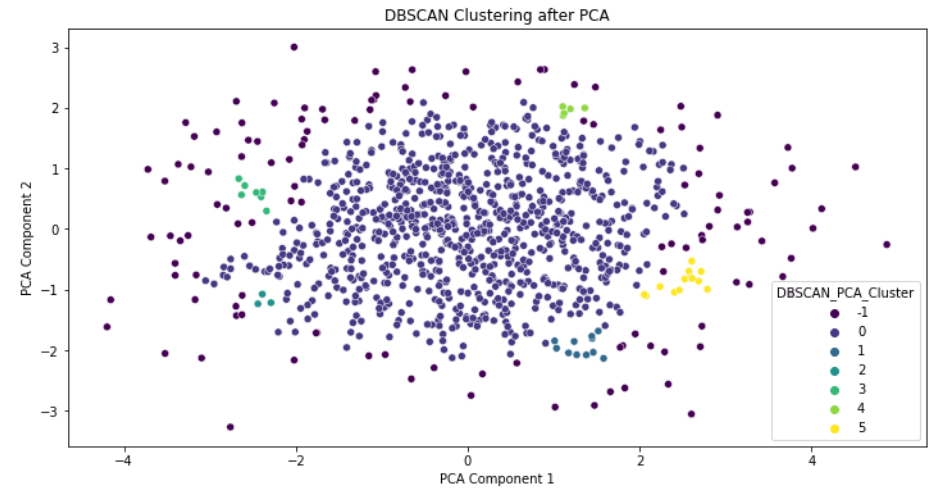
Silhouette Score: Measures how well-separated the clusters are, with higher scores indicating better-defined clusters.

Davies-Bouldin Index: Measures the average similarity ratio of each cluster with the cluster that is most similar to it, where lower values indicate better clustering.

* + 1. DBSCAN Clustering:

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is applied to detect outliers (noisy points) and dense clusters. Unlike K-Means, DBSCAN does not require the number of clusters to be specified and can detect arbitrarily shaped clusters. It also identifies outliers, which are labeled as "-1".

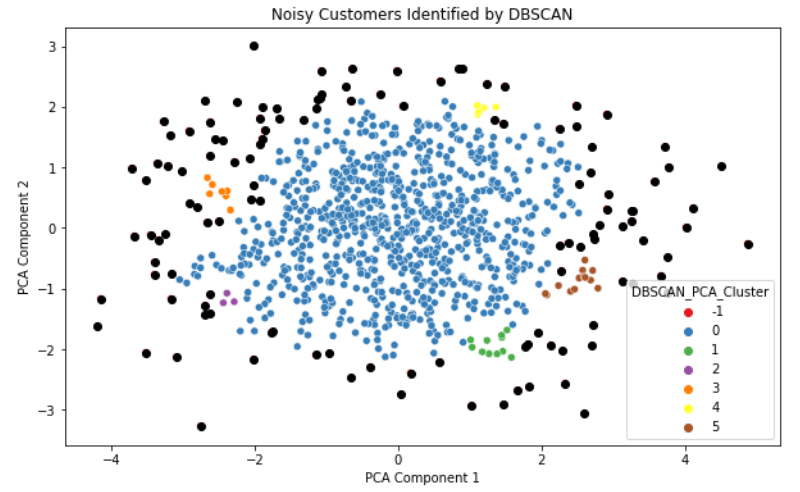
Function Used: DBSCAN(eps=0.25, min\_samples=5)



This plot shows the results of DBSCAN clustering after applying Principal Component Analysis (PCA) for dimensionality reduction. The two axes represent the first two principal components, allowing the data to be visualized in a 2D space. Different colors represent clusters identified by DBSCAN, with a total of five clusters labeled from 0 to 4. Cluster 0, shown in blue, contains the majority of points, while smaller clusters (1 to 4) are more spread out in the plot.

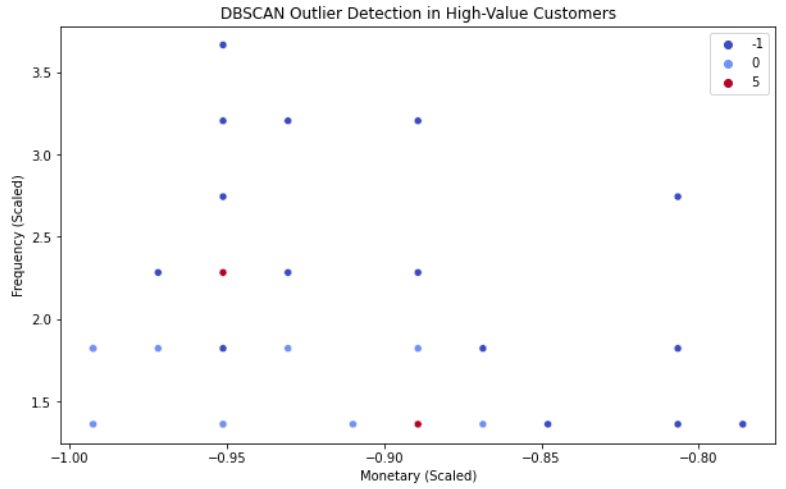
* + 1. Outlier Detection with DBSCAN:

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



The points that are labeled as noise, indicated by the -1 label are highlighted in the plot. These outliers may indicate customers with unusual or extreme behaviors. These insights can be valuable for identifying potential fraudulent transactions or anomalous customer behaviors that may not fit into typical customer segments.

* + 1. DBSCAN Outlier Detection



This plot shows DBSCAN outlier detection among high-value customers, focusing on Monetary (total amount spent) and Frequency (number of purchases). DBSCAN is used to find customers whose spending and purchasing habits stand out from the norm.

Outliers (marked as -1) are identified as customers whose behavior is significantly different from typical patterns, while regular customers are grouped into clusters. The decision to plot Monetary and Frequency is important because these metrics reveal key customer behaviors:

* Monetary shows how much customers spend overall, identifying those with unusually high or low spending.
* Frequency reflects how often customers make purchases, distinguishing between regular and occasional buyers.

By analyzing these two metrics, DBSCAN can uncover valuable insights, such as customers who spend a lot but purchase infrequently or frequent buyers with low spending. Understanding these outliers helps businesses refine customer segmentation, personalize marketing, and identify potential risks or opportunities in customer behavior.

* + 1. Silhouette Score and Davies-Bouldin Index for DBSCAN:

The quality of DBSCAN clustering is also evaluated using the silhouette score and Davies-Bouldin index, focusing only on non-noise points. If there are sufficient non-noise clusters, these metrics provide insights into how well DBSCAN performed in separating dense clusters.

* 1. **Geographic Visualization**

Here, we aim to visualize customer locations with a focus on high-value customer segments, providing insights for data-driven marketing and strategic planning.

* + 1. 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) coordinate system ensures global accuracy for spatial data.

* + 1. 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.

* + 1. Data Validation

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

* + 1. Interactive Map Creation

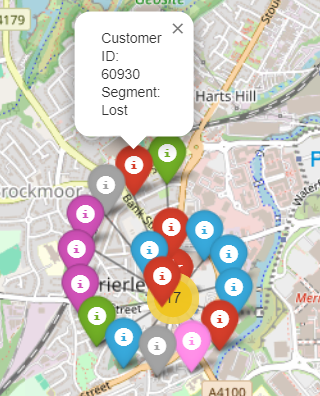
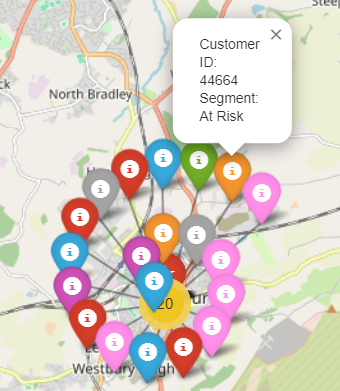


The map highlights regions with a dense concentration of high-value customers (VIPs), the clustering of VIP customers in certain regions provides a clear indication of where the most loyal and profitable customers are located, helping to optimize logistics, sales, and customer support.

Areas with fewer high-value customers may represent untapped market opportunities, where businesses could focus on improving customer engagement or expanding their presence.

* + 1. Highlighting High-Value Customers

VIP customers are visually emphasized with pink markers and icons, allowing quick identification of key segments. This provides clear insights into where the most valuable customers are geographically concentrated.

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

Each customer segment is represented by a unique color, making it easy to visually distinguish between different types of customers. For instance, VIP customers are highlighted in pink as they are high-value, while At Risk customers are represented in orange and Lost customers in yellow. Other segments like Loyal, Big Spenders, and Low Spenders are marked with distinct colors such as green, purple, and blue respectively, ensuring clear visual differentiation.

Each marker on the map is interactive. By clicking on a marker, a pop-up appears with key details about the customer, such as their customer ID and segment. This allows you to explore individual customers in detail without cluttering the map with too much information upfront.

The geographical data for each customer (their latitude and longitude) is managed using GeoPandas library, which makes it easy to plot their locations on the map. GeoPandas ensures that the data is formatted correctly for mapping.

The map allows for real-time interaction. You can zoom in and out of different regions and click on individual customer markers to explore more information. This interactive nature makes it a valuable tool for exploring customer data.

Geographic Insights: The color-coded map provides clear insights into where different customer segments are concentrated. For example, a business can quickly identify regions where VIP customers are located, which can guide decisions on where to focus marketing efforts or special promotions. Similarly, areas with a high concentration of Lost or At Risk customers can signal where retention efforts are needed.

* + 1. Exporting and Sharing the Map

The map is saved as an HTML file, making it easily accessible and shareable for collaborative strategic decisions.