**Capstone Project**

**Online Retail**

**(Customer Segmentation)**

By Chloe Yeo

Date: 2/11/24

Table of Contents

[Problem statement 3](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.30j0zll)

[Industry/ domain](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.3znysh7) 3

[Stakeholders 4](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.2et92p0)

[Business question 4](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.tyjcwt)

[Data question 5](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.3dy6vkm)

[Data 5](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.1t3h5sf)

[Data science process 6](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.4d34og8)

[Data analysis 6](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.2s8eyo1)

[Modelling 1](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.z337ya)0

[Outcomes 1](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.3j2qqm3)2

[Implementation 1](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.26in1rg)3

[Data answer 1](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.lnxbz9)3

[Business answer 14](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.35nkun2)

[Response to stakeholders 1](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.1ksv4uv)4

[End-to-end solution 1](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.44sinio)4

[References 15](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit#heading=h.2jxsxqh)

# Problem statement

This project focuses on customer segmentation for an online retail store to enhance personalized marketing strategies, product recommendations, and customer retention. By clustering customers based on their purchasing behaviors, the project seeks to identify key segments, such as high-value or frequent buyers, and tailor engagement strategies for each group. Effective customer segmentation enables online retailers to create personalized experiences that improve customer satisfaction and loyalty. Addressing segmentation challenges offers several benefits, including improved retention through targeted engagement, increased sales from personalized recommendations and promotions, and optimized marketing spend by focusing resources on high-potential customer groups. Ultimately, segmentation allows the business to understand its customers’ unique needs and behaviors, laying a foundation for improved customer experience and profitability.

Currently, the business may face challenges with generic marketing strategies that do not resonate with different customer types, potentially leading to lower satisfaction, missed engagement opportunities, and revenue loss. By implementing a data-driven segmentation model, the business can better categorize customers into meaningful groups based on their behavior. This approach allows for personalized strategies to improve customer satisfaction, identify and retain high-value or at-risk customers, and maximize engagement efforts for each segment. Various studies have shown that machine learning techniques such as RFM analysis, K-means, DBSCAN, and hierarchical clustering have demonstrated substantial benefits in multiple retail and e-commerce, including increased engagement, higher revenue, improved retention, and optimized resource allocation. Different research indicates that machine learning-based customer segmentation empowers businesses to make data-driven decisions that enhance profitability and loyalty in a competitive online retail environment.

# Industry/ domain

This project operates within the e-commerce and online retail industry, a rapidly growing sector fueled by increased digital adoption and the convenience of online shopping. As consumer preferences shift toward digital purchasing, competition has intensified, with established companies facing challenges from startups that leverage advanced technology and data analytics to deliver highly personalized experiences. The e-commerce value chain includes key stages: Product Sourcing and Inventory Management, Digital Marketing and Customer Acquisition, Customer Segmentation and Personalization, Order Fulfillment and Logistics, Customer Service and Retention, and Data Analytics and Feedback. Together, these functions support operational efficiency and strategic growth, enabling companies to remain responsive to customer needs and market trends.

Critical concepts within e-commerce, such as Customer Segmentation, Personalization, and Data-Driven Decision Making, are essential for business success. Customer Segmentation divides customers into targeted groups for more effective marketing, while Personalization enhances the customer experience through tailored recommendations and messages. Data-Driven Decision Making leverages analytics and machine learning to optimize marketing, improve customer engagement, and strengthen overall business performance. Although this project focuses on e-commerce, the principles of customer segmentation and personalized marketing are also vital in other industries, including traditional retail, hospitality, financial services, healthcare, and telecommunications, where they help improve marketing effectiveness, customer satisfaction, and retention.

# Stakeholders

The primary stakeholders in this project include the Marketing Team, Sales Team, Product Development Team, and Senior Management/Executives, each with a vested interest in leveraging customer segmentation to drive growth and optimize engagement strategies. The Marketing Team, responsible for creating targeted campaigns, relies on segmentation to craft messages and promotions that resonate with specific customer groups, enhancing both acquisition and retention efforts. The Sales Team uses segmentation insights to deepen customer relationships and focus on cross-selling and upselling opportunities, helping to drive revenue growth by concentrating efforts on high-potential customers.

Additionally, The Product Development Team benefits from segmentation by aligning product offerings with customer preferences, which informs product enhancements, bundling, and the development of new offerings tailored to specific segments. Senior Management and Executives, including the CEO and CFO, view segmentation as a strategic tool for maximizing customer lifetime value (CLV) and overall business growth.

Stakeholders aim to see improvements in key performance indicators (KPIs) such as customer retention, lifetime value, and revenue, while also gaining strategic insights that will inform high-level business decisions.

# Business question

The primary business question for this project is: "How can we identify distinct customer segments to tailor marketing, improve engagement, and maximize revenue?". This project focuses on creating meaningful customer segments that enable targeted marketing. Personalized marketing can significantly enhance engagement and conversion rates; for example, with an average of 100,000 monthly site visitors and an average order value of $50, even a 1% increase in conversion could generate an additional $50,000 per month, or $600,000 annually.

Traditional accuracy metrics are not directly applicable to clustering, so I rely on clustering quality measures such as the Silhouette Score (where a score above 0.5 generally indicates distinct clusters) and the Davies-Bouldin Index (DBI) (where values below 1.0 indicate well-separated clusters). Also an effective segmentation depends on managing the implications of over-segmentation (false positives) and under-segmentation (false negatives). Over-segmentation can lead to a misallocation of marketing resources and customer confusion, while under-segmentation risks missed revenue opportunities, increased customer churn, and reduced satisfaction among high-value customers.

# Data question

My data question is: "What is the best machine learning clustering model that can group customers based on purchasing behavior and spending patterns, enabling precise, data-driven engagement strategies?" To answer this question, I used transactional data that can be aggregated by customers. This includes information such as Customer ID to uniquely identify customers, Invoice No/Transaction ID to track individual purchases, Purchase Date and Time to calculate recency, Quantity Purchased to measure purchase frequency, and Unit Price to compute the monetary value of each transaction.

Using this data, we derive three primary features for clustering: Recency (days since last purchase), Frequency (total purchases over a given period), and Monetary Value (total customer spending), which capture customer engagement, loyalty, and value. To evaluate clustering quality and determine the most effective model, we use metrics like the Silhouette Score (with higher scores indicating more distinct clusters) and the Davies-Bouldin Index (DBI) (where lower values indicate compact, well-separated clusters). These metrics will guide us in selecting the model that best supports data-driven engagement strategies by providing clear, actionable customer segments.

# Data

The data was sourced from Kaggle, specifically from the dataset "Online Retail" provided by Vijayuv. The dataset can be accessed at<https://www.kaggle.com/datasets/vijayuv/onlineretail>.

**Volume**: The dataset contains approximately **541,909 rows** and **8 columns**.

**Attributes**:

* **InvoiceNo**: Unique invoice number for each transaction.
* **StockCode**: Product/item identifier.
* **Description**: Description of the product.
* **Quantity**: Number of units purchased for each transaction.
* **InvoiceDate**: Date and time of the transaction.
* **UnitPrice**: Price per unit of the product.
* **CustomerID**: Unique identifier for each customer.
* **Country**: Country of residence for each customer.

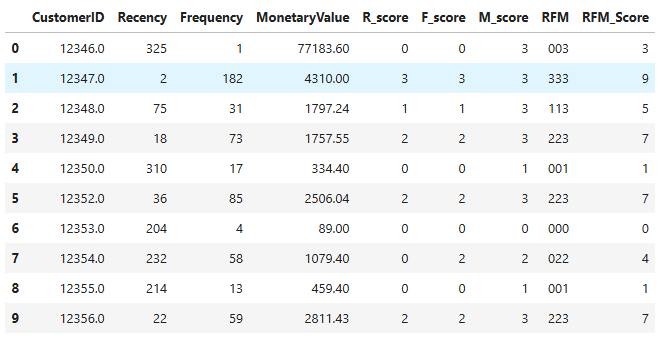
This dataset, widely used for customer segmentation and retail analytics, is considered reliable for exploratory and educational purposes. However, it lacks contextual details, such as the precise time frame of transactions, which limits its use for longitudinal analysis. Additionally, missing entries in the CustomerID field may impact the accuracy of customer-level segmentation. The raw data presents quality issues, including missing values in CustomerID, unitprice of 0, negative values in Quantity (potentially indicating product returns or errors), and duplicate entries that could distort analysis. Therefore, pre-processing steps, such as removing duplicates, handling missing values, and addressing negative quantities, are essential to improve data quality for analysis. This dataset was generated from actual transactions at an online retail store based in the UK, likely exported from a point-of-sale or order management system. It is a static snapshot of historical transaction data and does not update on an ongoing basis; for continuous analysis or production use, organizations would need access to live transactional data from their internal systems to enable ongoing customer segmentation.

# Data science process

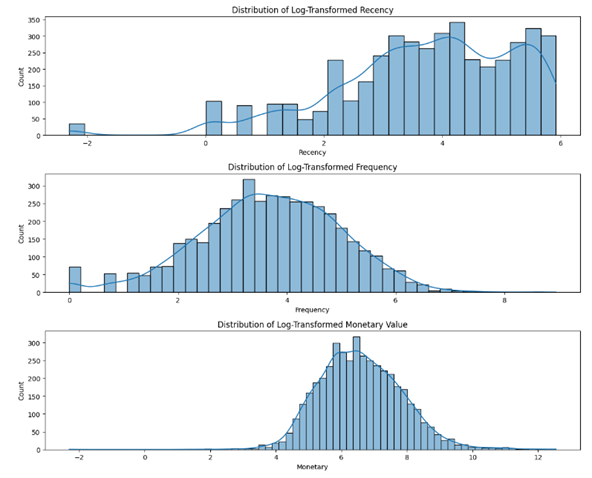
## Data analysis

The data pipeline for wrangling raw data consists of multiple steps:

* Data Loading: Import the raw transactional data, typically using a data manipulation library like pandas.
* Data Cleaning: Remove or handle missing values in key fields such as CustomerID, address duplicate entries, and remove or process rows with negative quantities to manage product returns or errors.
* Feature Engineering: Create derived features such as Recency (days since last purchase), Frequency (number of purchases), and Monetary Value (total spending) for each customer. Then applied log to normalise the RFM metrics and applied standard scaler.



**Table 1**. Group data by CustomerID to calculate RFM values and RFM score to produce a customer-level dataset ready for clustering analysis.

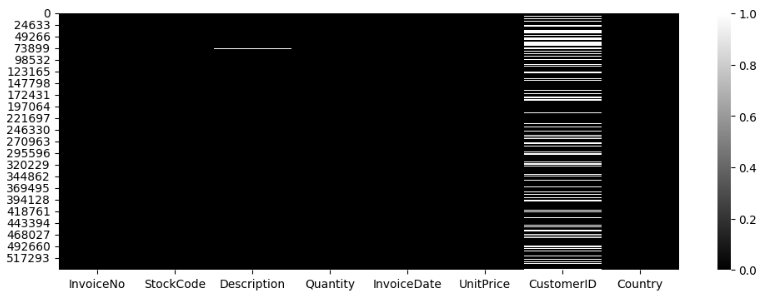


**Fig 1**. Apply log-transformation to normalise RFM features to ensure more effective clustering.

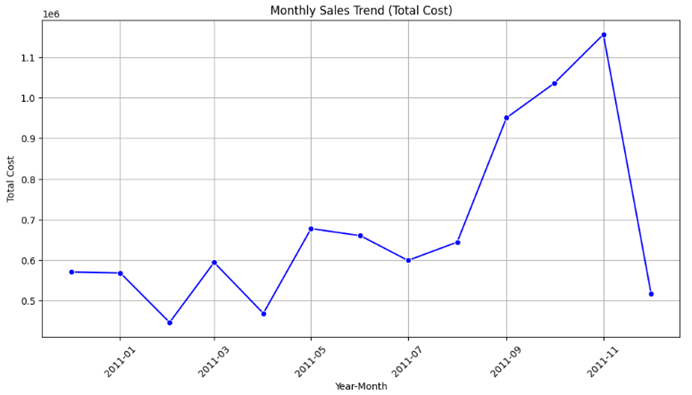
During Exploratory Data Analysis (EDA), trends in Total\_Cost for monthly and hourly were examined to gain insights into customer engagement. Also right-skewed distributions in Frequency and Monetary Value necessitated log transformation. This modular pipeline is designed for reuse with similar transactional data, adaptable with minor adjustments for new features or transaction formats.



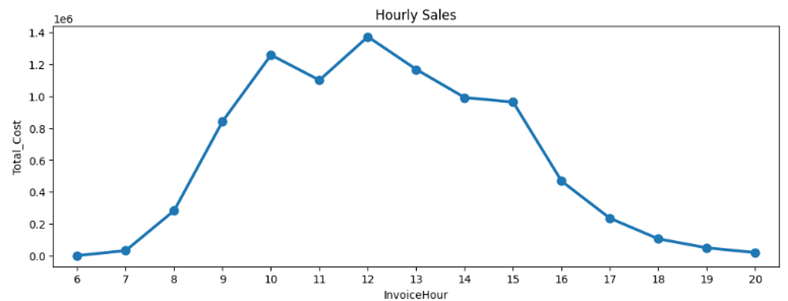
**Figure 2.** Heat map correlation analysis to explore the correlation between RFM

****

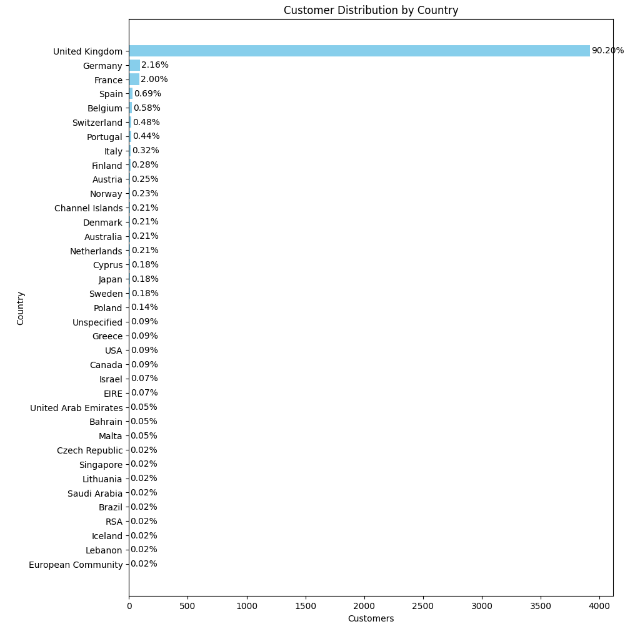
**Fig 3**. Lots of missing values in CustomerID, need to be removed



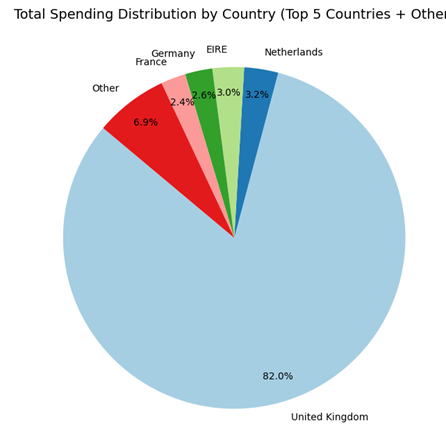
**Fig 4**. Monthly sales trend. Insufficient data for December 2012



**Fig 5**. Hourly sales trend



**Fig 6.** Number of customer distribution by Country



**Fig 7**. Distribution of the amount of Total spending by country

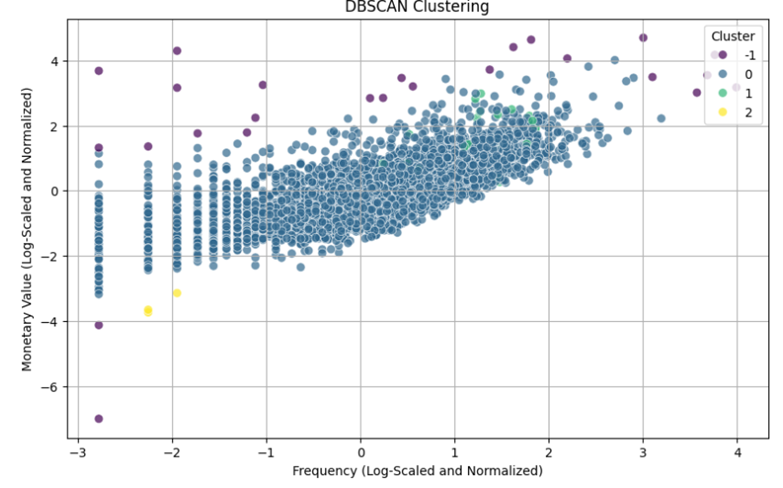
Intermediary data structures include a Cleaned DataFrame (containing only valid transactions) and a Customer-Level Aggregated DataFrame (grouped by CustomerID for RFM metrics), along with log-transformed features stored separately for clustering analysis. During hierarchical clustering, a dendrogram (a tree structure) is built as an intermediary structure to visualize and interpret the nested clustering of data.

## Modelling

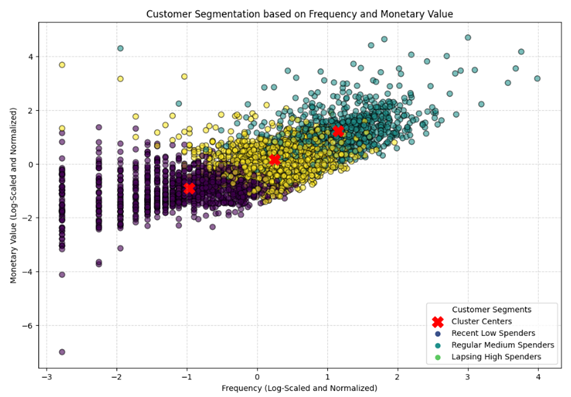
For customer segmentation, the primary features selected were Recency, Frequency and Monetary Value. Interaction between Frequency and Monetary Value revealed that frequent buyers generally have higher spending, indicating a positive correlation between purchase frequency and customer value. Additionally, customers with lower recency (i.e., those who purchased recently) tend to show higher frequency and monetary value, underscoring the importance of recent engagement for high-value customers.

Recency and Frequency were found to capture much of the clustering performance, effectively differentiating between active, loyal customers and occasional or inactive ones. Monetary Value further refined segmentation by highlighting high-spending segments, which is essential for identifying and engaging high-value customers. Features were selected based on the RFM model, a well-established approach for customer segmentation that provides a comprehensive view of customer behavior and value.

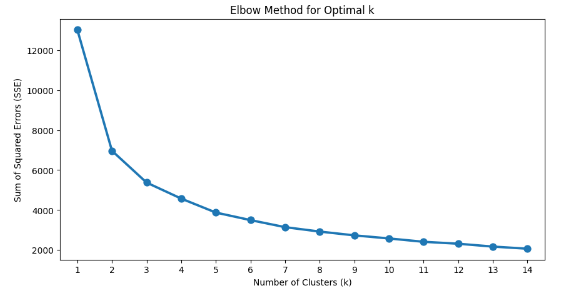
Feature engineering included log transformation to manage skewness in Recency, Frequency and Monetary Value, normalization to standardize feature scales, and RFM calculation to aggregate transaction data by customer. The project used three unsupervised clustering models—K-means, DBSCAN, and Hierarchical Clustering—to identify the most effective segmentation. Training times varied, with K-means and DBSCAN each requiring only a few seconds to a minute, while hierarchical clustering was longer due to pairwise distance calculations.



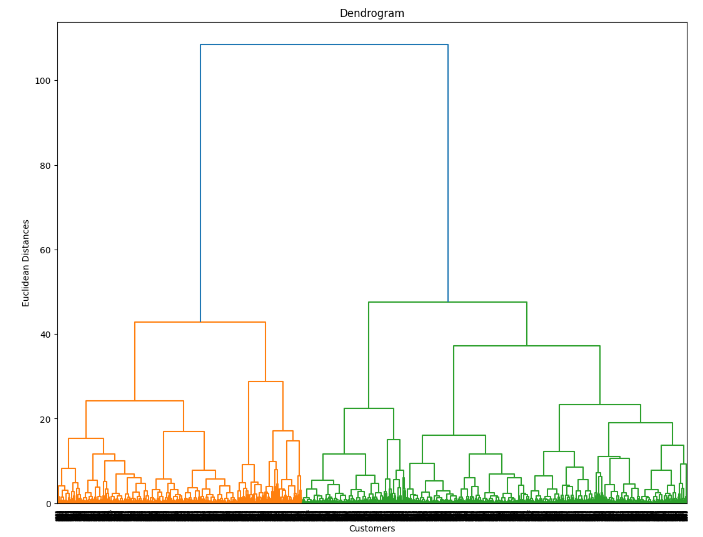
**Fig. 8** DBSCAN Clustering



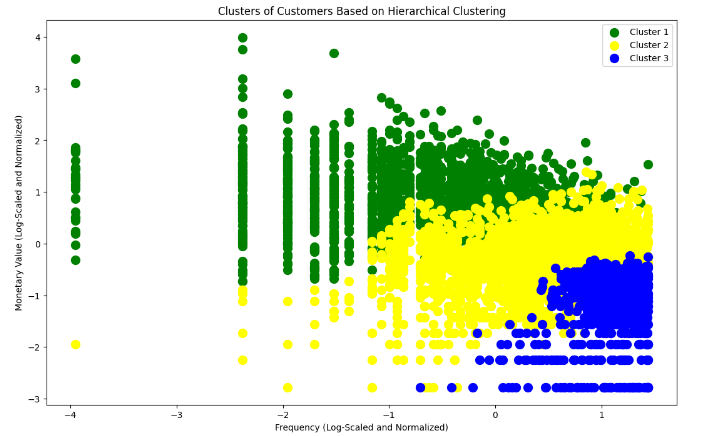
**Fig.9** K-mean clustering



**Fig 10** Elbow method for K-means



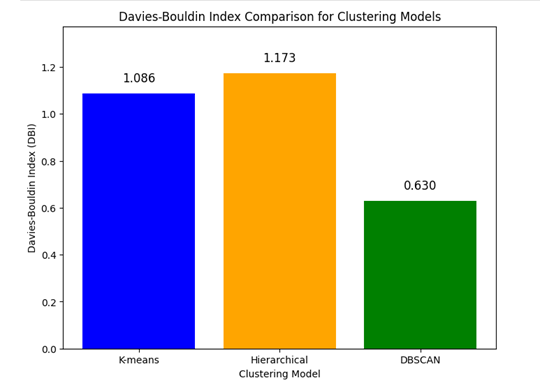
**Fig 11** Dendrogram for Hierarchical Clustering

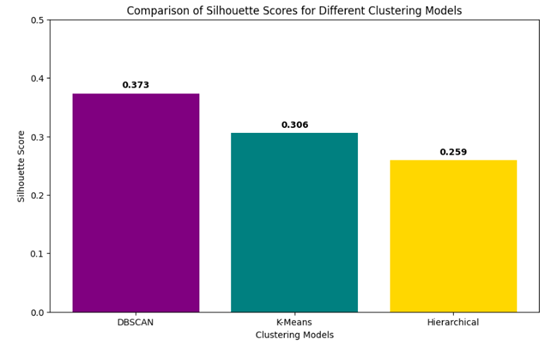


**Fig. 12** Hierarchical Clustering

## **Fig 13**. Comparison of Silhouette score for 3 different models

## **Figure 14.** Comparison of DBI for 3 different models



Python and libraries such as pandas, NumPy, scikit-learn, and Matplotlib were used for data manipulation, modeling, and visualization, and cloud platforms like Jupyter Notebook supported computation. Clustering performance was evaluated using the Silhouette Score and the Davies-Bouldin Index (DBI). DBSCAN was ultimately selected as the final model due to its superior performance, achieving the lowest DBI and highest silhouette score. Additionally, DBSCAN’s ability to handle noise and outliers resulted in well-defined, actionable customer segments.

## Outcomes

The data science process for customer segmentation successfully identified distinct customer segments using the RFM analysis, providing a foundation for targeted marketing and engagement strategies. Among the unsupervised machine learning models tested—K-means, DBSCAN, and Hierarchical Clustering—DBSCAN was selected as the optimal approach due to its superior clustering quality, as evidenced by the highest Silhouette Score and lowest Davies-Bouldin Index.

DBSCAN’s ability to handle noise and outliers led to clear, actionable segments. Key insights revealed a positive correlation between purchase frequency and customer value, as well as the association between recent engagement and high-value, loyal customers. The RFM metrics, supported by feature engineering techniques like log transformation and normalization, proved effective for clustering, with Recency and Frequency capturing much of the segmentation performance. A reusable data pipeline was developed to ensure consistency and adaptability for future data processing. Overall, the process equips the business with a data-driven approach to customer segmentation, enabling tailored marketing strategies that enhance customer retention, satisfaction, and revenue.

## Implementation

implementing the customer segmentation model in production requires careful planning to ensure its effectiveness and scalability. Key considerations include automating the data pipeline to regularly refresh the RFM metrics and retrain the model with updated data, maintaining accuracy and relevance.

Also monitoring model performance through metrics like the Silhouette Score and Davies-Bouldin Index, along with tracking business KPIs, is essential for assessing ongoing effectiveness and detecting data drift.

Finally, ensuring the model's outputs are interpretable and actionable for business stakeholders will facilitate effective decision-making, with regular collaboration to refine segmentation strategies based on feedback. These steps will ensure the model remains robust, secure, and aligned with business objectives in a production environment.

# Data answer

The data question was answered satisfactorily. After evaluating multiple models, DBSCAN was selected for its superior clustering quality, evidenced by the highest Silhouette Score and lowest Davies-Bouldin Index, indicating well-defined and actionable customer segments. DBSCAN’s effectiveness in handling noise and outliers further improved segmentation, yielding reliable insights for targeted marketing strategies. While a Silhouette Score of 0.373 suggests some limitations in cluster separation, the low Davies-Bouldin Index, meaningful business-relevant segments, and comparative advantages over other models justify a reasonable confidence level in DBSCAN’s clustering results.

The confidence level is not only supported by robust evaluation metrics but also the suitability of the RFM features for capturing customer behavior. While these clustering results are reliable, maintaining high confidence over time will require periodic data updates and model retraining to adapt to changing customer patterns.

# Business answer

The business question was also answered satisfactorily. By employing customer segmentation models, specifically DBSCAN, the project identified actionable customer groups based on purchasing behaviors and spending patterns, enabling targeted marketing and engagement strategies for distinct customer types, such as high-value and low-engagement customers. While DBSCAN’s Silhouette Score was moderate (0.373), its low Davies-Bouldin Index indicated compact, well-separated clusters, supporting confidence in the model’s quality. Additionally, DBSCAN’s ability to handle noise and outliers outperformed K-means and Hierarchical Clustering, further reinforcing its suitability for this task. The alignment of segments with business needs enables tailored marketing actions that can drive engagement and revenue, resulting in a moderately high confidence level in the business answer. This solution offers practical, data-driven insights that effectively address the company’s objectives.

# Response to stakeholders

The data-driven customer segmentation model provides stakeholders with actionable insights into distinct customer groups, enabling tailored marketing and engagement strategies to maximize revenue and customer loyalty. By using DBSCAN to identify high-value, frequent, and low-engagement segments, the business can apply precise, targeted campaigns—such as loyalty programs for top spenders and re-engagement offers for low-frequency customers. We recommend continuously updating the model with new transactional data to capture evolving customer behaviors and maintain segmentation relevance. Overall, this segmentation approach aligns with the business’s strategic goals, providing a solid foundation for personalized customer engagement that can drive long-term growth and profitability.

# End-to-end solution

The end-to-end solution involves implementing an automated pipeline that continuously processes transactional data to maintain up-to-date customer segments, enabling data-driven engagement strategies. This pipeline begins with data ingestion, where transactional data is cleaned, transformed, and aggregated to calculate RFM metrics for each customer. Using the DBSCAN model, customers are then segmented into meaningful groups, such as high-value, frequent, and low-engagement clusters. The segmented data is integrated with the marketing systems, allowing for targeted campaign execution based on each customer group’s profile. Regular monitoring and retraining of the model ensure its alignment with changing customer behaviors, while performance metrics like customer retention, engagement rates, and revenue uplift are tracked to measure the impact of the segmentation. This streamlined solution provides a scalable, actionable framework for delivering personalized marketing, optimizing resource allocation, and achieving long-term customer engagement and revenue growth.

# References

**Notebook**: EDA\_and\_Cleaning.ipynb, RFM\_Analysis\_and\_Segmentation.ipynb, Clustering\_Models.ipynb, Model\_Evaluation\_and\_Comparison.ipynb:, Deployment\_and\_Visualization.ipynb

**Dataset**: <https://www.kaggle.com/datasets/vijayuv/onlineretail>.

**Exported models**: kmeans\_model.pkl, dbscan\_model.pkl, hierarchical\_model.pkl, rfm\_segments.csv

**Libraries**: pandas, numpy,. matplotlib, seaborn, plotly, scikit-learn, scipy,

pickle, joblib

**Algorithms**: RFM Analysis, Clustering Models: K-Means, DBSCAN, and Hierarchical Clustering, Evaluation Metrics:

<https://seaborn.pydata.org/introduction.html>

<https://matplotlib.org/stable/tutorials/introductory/pyplot.html>

<https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-value-of-getting-personalization-right-or-wrong-is-multiplying>

<https://ieeexplore.ieee.org/abstract/document/9439459>

<https://www.statista.com/topics/871/online-shopping/>