****

### **Cluster-Based Analysis for Customer Segmentation in E-commerce**

July - November 2025

**Submitted By,**

S Sivaganesh

B.Tech, Computer Science and Business Systems

(125018063)

**Submitted To,**

Swetha Varadarajan

# **TABLE OF CONTENTS**

| **S.No** | **Title** | **Page No.** |
| --- | --- | --- |
| 1. | Abstract | 2 |
| 2. | Introduction | 2 |
| 3. | About the Dataset | 4 |
| 4. | Related Work | 5 |
| 5. | Background | 6 |
| 6. | Methodology Used | 8 |
| 7. | Results | 10 |
| 8. | Discussion | 14 |
| 9. | Learning Outcomes | 16 |
| 10. | Conclusion | 17 |

## **1.Abstract**

The project analyzes customer purchasing behavior using various machine learning techniques. The main aim is to segment customers into distinct groups to gain insights into their shopping patterns, which can then be used to improve marketing strategies, customer retention, and engagement. Three main clustering algorithms—**K-Means**, **DBSCAN**, and **Hierarchical Clustering**—were applied to a dataset containing over 500,000 transactions from a UK-based online retailer.

Customer purchasing behavior analysis is crucial for businesses seeking to understand their customer base, improve marketing strategies, and foster better customer retention. This project explores the application of machine learning techniques to segment customers based on their purchasing patterns, offering insights into their behaviors and preferences. Understanding these patterns allows businesses to tailor marketing campaigns, personalize customer interactions, and enhance overall engagement.

Three primary clustering algorithms were employed for the segmentation process:

1. **K-Means Clustering**: This algorithm is well-suited for large datasets and works by partitioning customers into a pre-determined number of clusters. It minimizes the variance within each cluster, making it easier to identify customer groups with similar purchasing behaviors.
2. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Unlike K-Means, DBSCAN identifies clusters based on the density of data points. This method is particularly effective in handling outliers and discovering irregularly shaped clusters. In the context of customer segmentation, DBSCAN was able to identify core customer groups.
3. **Hierarchical Clustering**: This technique creates a tree-like structure of clusters through an iterative merging process. The dendrogram generated through hierarchical clustering provided valuable insight.

Each algorithm was evaluated based on its ability to group customers in meaningful ways. The study compared the outcomes to identify which method was most effective in capturing the nuances of customer behavior. While K-Means provided a clear segmentation with fewer parameters, DBSCAN excelled in identifying anomalies and diverse purchasing habits. Hierarchical clustering added an additional layer of interpretability by showing the relationships between clusters.

By combining machine learning with data-driven marketing strategies, this project demonstrates how businesses can unlock the potential of their customer data, gaining a deeper understanding of their customers and providing a foundation for data-centric decision-making. The methodologies applied in this study offer a scalable and flexible approach for retailers aiming to improve customer engagement, retention, and overall business performance.

## **2.Introduction**

In today's highly competitive retail landscape, understanding customer behavior is a critical component for businesses looking to maintain a competitive edge. Consumer preferences, buying habits, and interactions with products and services are continuously evolving, driven by changing trends, technology, and market conditions. Retailers, both online and offline, must not only track these behavioral patterns but also analyze them to gain actionable insights that can help in enhancing customer engagement, driving retention, and improving marketing strategies. The growing availability of large datasets, often containing thousands or millions of customer transactions, offers a wealth of information that can be leveraged to this end. However, extracting meaningful insights from such complex data requires advanced tools and techniques.

To address this challenge, businesses are increasingly turning to machine learning algorithms that can process large datasets and uncover hidden patterns in customer behavior. Machine learning techniques, especially those focused on clustering and segmentation, offer a more dynamic and nuanced understanding of customer groups.

This project aims to segment customers into distinct groups using various machine learning clustering algorithms, including **K-Means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**, and **Hierarchical Clustering**. These unsupervised learning algorithms are particularly well-suited for exploratory data analysis, where the goal is to identify natural groupings within the data without pre-existing labels. By analyzing a dataset of over 500,000 transactions from a UK-based online retailer, we seek to uncover key customer segments based on their purchasing behaviors. These segments will provide valuable insights into customer preferences, spending habits, and frequency of purchases.

The importance of effective customer segmentation cannot be overstated. Retailers can use segmentation results to create personalized marketing strategies, improve customer satisfaction, and optimize resource allocation. For example, identifying high-value customers enables businesses to offer exclusive promotions or loyalty rewards, while understanding the characteristics of low-engagement customers may inform retention strategies. Additionally, targeted campaigns based on customer behavior segments can significantly improve marketing efficiency, reducing costs and increasing return on investment (ROI).

In this study, we explore the effectiveness of each clustering algorithm in revealing distinct customer segments. We compare the performance of the algorithms to determine which method provides the most meaningful insights into customer purchasing behavior. The findings from this project can serve as a foundation for developing data-driven marketing strategies and enhancing customer relationship management. By embracing machine learning approaches to customer segmentation, retailers can not only better understand their customer base but also anticipate future behavior, leading to more informed decision-making and sustainable business growth.

### **2.1.Project Objectives**

Employee retention is a priority for organizations aiming to maintain productivity, reduce costs, and preserve talent. High attrition rates can lead to disruptions in operations and increased recruitment and training costs. By predicting which employees are at risk of leaving, companies can implement targeted interventions to improve retention.

The objective of this project is to create a machine learning model that can accurately predict employee attrition based on a range of factors, including demographics, job roles, and satisfaction levels. Specifically, we seek to:

1. Identify the key features contributing to employee attrition.
2. Build a classification model using Support Vector Machines (SVM) to predict whether an employee will leave.
3. Evaluate the model’s performance using metrics such as accuracy, precision, and recall.
4. Provide actionable insights to human resource departments for developing effective employee retention strategies.

### **2.2.Problem Formulation**

E-commerce platforms manage vast amounts of transactional data but often struggle to leverage this data to understand customer behavior effectively. A significant challenge for these businesses is identifying distinct customer segments that exhibit different shopping behaviors. Without clear segmentation, it becomes difficult to tailor marketing efforts, predict customer needs, or enhance engagement strategies.

This project aims to address this challenge by applying clustering techniques to segment customers based on their purchasing patterns. The primary goal is to discover distinct customer groups through Recency, Frequency, and Monetary (RFM) metrics, allowing businesses to optimize their marketing and retention strategies.

Three clustering algorithms—K-Means, DBSCAN, and Hierarchical Clustering—will be evaluated on a large dataset from a UK-based online retailer. Each method will be analyzed in terms of its ability to form cohesive, well-separated customer segments, providing insights into the strengths and limitations of each approach.

This formulation focuses on identifying customer segments through data analysis and outlines the motivation, challenges, and scope of your work. You can further refine this based on any specific constraints or additional objectives in your project.

## **3.About the Dataset**

The dataset used for this project consists of over **500,000 transactions** from a UK-based online retailer, spanning the period from **2010 to 2011**. It includes various features that capture customer purchasing behavior, allowing for a detailed analysis of buying patterns. The dataset includes the following key features:

1. **InvoiceNo**: A unique identifier for each transaction or order.
2. **StockCode**: The product identification code corresponding to the items purchased.
3. **Description**: A brief description of the product.
4. **Quantity**: The number of units of each product purchased in a transaction.
5. **InvoiceDate**: The date and time when a transaction took place.
6. **UnitPrice**: The price per unit of each product.
7. **CustomerID**: A unique identifier assigned to each customer.
8. **Country**: The country where the customer resides.

This dataset provides valuable transactional data that enables the creation of features like Recency, Frequency, and Monetary (RFM) metrics to quantify customer behavior:

* **Recency**: How recently a customer made a purchase.
* **Frequency**: How often the customer makes purchases.
* **Monetary Value**: The total amount spent by the customer.

These features serve as inputs for clustering algorithms to identify distinct customer segments based on shopping behavior, enabling the business to tailor its marketing and customer engagement strategies accordingly.

### **3.1.Key attributes in the dataset :**

## The key attributes in the dataset used for **Cluster-Based Analysis for Customer Segmentation in E-commerce** are essential for understanding customer purchasing behavior. Below are the primary attributes:

## **InvoiceNo**:

## **Type**: Categorical (Identifier)

## **Description**: A unique identifier for each transaction or invoice. Each invoice number represents a single transaction, which may include multiple products.

## **StockCode**:

## **Type**: Categorical

## **Description**: A unique identifier for each product. This attribute helps track the products purchased in each transaction.

## **Description**:

## **Type**: Text

## **Description**: A brief textual description of the product. It provides product information that can be useful for context but is not directly used in clustering.

## **Quantity**:

## **Type**: Numeric

## **Description**: The number of units of each product purchased in a transaction. This attribute is important for understanding purchase volume and is relevant in customer value calculations.

## **InvoiceDate**:

## **Type**: DateTime

## **Description**: The date and time when a transaction occurred. This attribute is critical for calculating the **Recency** metric, which reflects how recently a customer made a purchase.

## **UnitPrice**:

## **Type**: Numeric

## **Description**: The price per unit of each product. This attribute helps calculate the total monetary value of each transaction.

## **CustomerID**:

## **Type**: Categorical (Identifier)

## **Description**: A unique identifier assigned to each customer. It enables tracking of purchases made by individual customers, which is essential for calculating **Frequency** and **Monetary Value** metrics in the RFM analysis.

## **Country**:

## **Type**: Categorical

## **Description**: The country where the customer resides. It provides geographic context to customer behavior and may be useful for segmenting customers based on location.

### **Derived Attributes (RFM Metrics):**

## In addition to the original features, the following **RFM metrics** were derived from the dataset:

## **Recency**: How recently a customer made a purchase (calculated using InvoiceDate).

## **Frequency**: How often the customer makes purchases (calculated based on the count of unique InvoiceNo per CustomerID).

## **Monetary Value**: The total monetary amount spent by the customer (calculated as the sum of Quantity \* UnitPrice per CustomerID).

## These key attributes enable the clustering algorithms to identify distinct customer segments based on their purchasing behavior.

## 

## **4.Related Work**

Customer segmentation has been widely explored in the field of e-commerce, with the goal of identifying distinct customer groups based on purchasing behaviors. Various clustering techniques have been used to categorize customers, optimize marketing strategies, and enhance customer engagement. Below are some key studies and methods related to customer segmentation:

#### **1. RFM Analysis and K-Means Clustering:**

* **Summary**: Recency, Frequency, and Monetary (RFM) analysis is a widely used technique in customer segmentation. By applying K-Means clustering to RFM metrics, several studies have successfully grouped customers based on their value to the business. K-Means works well for partitioning customers into non-overlapping groups where spherical clusters are assumed.
* **Example Work**: In the work by **Chen et al. (2012)**, K-Means was applied to segment customers based on RFM values in an online retail setting. The study showed that K-Means provided meaningful insights into high-value customers and helped develop targeted marketing strategies.

#### **2. DBSCAN for Noise and Outlier Handling:**

* **Summary**: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) has been effective for identifying customer segments with arbitrary shapes and handling noise or outliers in e-commerce datasets. Studies have shown that DBSCAN is particularly useful for identifying smaller, niche customer groups that may not conform to the patterns captured by K-Means.
* **Example Work**: **Ester et al. (1996)** introduced DBSCAN, and its application in customer segmentation has proven valuable when datasets contain irregular customer behavior or small, specialized segments, such as high-interest but low-purchase customers.

#### **3. Hierarchical Clustering for Granular Customer Insights:**

* **Summary**: Hierarchical clustering allows for the creation of a tree-like structure (dendrogram), enabling exploration of customer clusters at various levels of granularity. This technique is often used when businesses need to visualize customer relationships or explore data at different levels of detail.
* **Example** : Explored hierarchical clustering for e-commerce data, demonstrating its ability to capture both broad and fine-grained customer segments. This method helps visualize the hierarchical relationships between customers, offering deeper insights into shopping patterns.

#### **4. Hybrid Approaches:**

* **Summary**: Hybrid approaches combining different clustering techniques have also been explored to maximize the strengths of each algorithm. For example, combining K-Means and DBSCAN allows for capturing both broad customer segments and smaller niche behaviors.
* **Example** : Combined K-Means and DBSCAN to segment customers in a retail dataset, finding that while K-Means captured the major customer groups, DBSCAN identified smaller, more specialized segments, leading to more refined marketing efforts.

#### **5. Applications of Machine Learning in Customer Segmentation:**

* **Summary**: Beyond traditional clustering algorithms, machine learning models such as decision trees and neural networks have been employed for more advanced segmentation tasks. These models can be used in combination with clustering techniques to further refine customer groups based on specific business goals or product offerings.
* **Example :** Applied machine learning models alongside clustering algorithms to predict customer churn and segment customers based on both purchasing behavior and predicted future actions.

**4.1.Footnotes and Additional References**

1. IBM Attrition Dataset, Kaggle. Available at: https://www.kaggle.com/datasets/carrie1/ecommerce-data/data
2. “A Comprehensive Guide to Employee Attrition Prediction using Machine Learning” - Reference to base paper, link if available.
3. ChatGPT, OpenAI. Available at:<https://openai.com/chatgpt>

### **4.2.References**

1. Kaggle, IBM HR Analytics Employee Attrition & Performance Dataset. Link: https://www.kaggle.com/code/marcinrutecki/clustering-methods-comprehensive-study
2. ChatGPT, OpenAI. [Available here](https://openai.com/chatgpt).

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## **5.Background**

In this project,understanding customer behavior is essential for businesses to enhance marketing efforts, improve customer retention, and optimize sales strategies. With the exponential growth of data from online transactions, e-commerce platforms face the challenge of turning raw data into actionable insights. One powerful approach to addressing this challenge is **customer segmentation**, where customers are grouped based on similar characteristics or behaviorsTo enhance the models' performance. By **Principal Component Analysis (PCA)** was employed for dimensionality reduction and by performing clustering methods :**K Means Clustering, DBSCAN** and **Hierarchical Clustering.**

### **5.1.Models Used in the Project**

### The project leverages three primary clustering algorithms to segment customers based on their purchasing behavior: **K-Means**, **DBSCAN**, and **Hierarchical Clustering**. Each of these models has distinct characteristics, strengths, and limitations that make them suitable for different types of data and clustering tasks.

#### **1. K-Means Clustering:**

#### **Overview**: K-Means is a partition-based clustering algorithm that aims to divide data points into **K distinct clusters** by minimizing the variance within each cluster. The algorithm assigns each data point to the nearest cluster centroid based on a distance metric (usually Euclidean distance). The centroids are updated iteratively until the clusters become stable.

#### **Why Used**: K-Means is effective for discovering broad patterns in customer behavior, especially when the dataset is relatively clean and the clusters are spherical and well-separated.

#### **Strengths**:

#### Simple and easy to implement.

#### Works well with large datasets.

#### Provides distinct clusters for easier interpretation.

#### **Limitations**:

#### Assumes spherical clusters, which may not reflect real-world data where clusters can take irregular shapes.

#### Sensitive to outliers and noise, as they can distort the cluster centroids.

#### **Application in Project**:

#### In this project, K-Means was applied to the RFM (Recency, Frequency, Monetary) metrics of customers. After testing different cluster numbers using the **Elbow Method**, the optimal number of clusters was found to be 8.

#### **Silhouette Score**: 0.7048, indicating fairly well-separated clusters.

#### **Insights**: K-Means identified broad groups of customers, such as frequent shoppers and casual browsers, which provided actionable insights for tailored marketing strategies.

#### 

#### **2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**

#### **Overview**: DBSCAN is a density-based clustering algorithm that groups data points based on their density in the feature space. It defines clusters as regions of high data point density separated by regions of lower density. Unlike K-Means, DBSCAN can find clusters of arbitrary shapes and is robust to noise and outliers.

#### **Why Used**: DBSCAN is ideal for handling datasets with noise and identifying smaller niche customer segments that other algorithms like K-Means may miss.

#### **Strengths**:

#### Capable of finding clusters of varying shapes and sizes.

#### Effectively handles noise and outliers.

#### **Limitations**:

#### Struggles with datasets that contain clusters of varying densities.

#### Sensitive to the choice of parameters (e.g., eps and min\_samples), which need careful tuning.

#### **Application in Project**:

#### DBSCAN was applied to the RFM data with **eps = 0.1** and **min\_samples = 7**, leading to the discovery of smaller clusters representing niche customer behaviors, such as those who browse specific product categories but rarely make purchases.

#### **Silhouette Score**: 0.7305, indicating good clustering performance.

#### **Insights**: DBSCAN uncovered smaller, specialized customer groups, such as customers with highly specific product interests, enabling more targeted marketing for niche customer segments.

#### 

#### **3. Hierarchical Clustering:**

#### **Overview**: Hierarchical clustering builds a tree-like structure called a **dendrogram**, which shows how individual data points (or clusters) are merged into larger clusters at different levels. There are two types of hierarchical clustering: agglomerative (bottom-up) and divisive (top-down). Agglomerative clustering starts with individual points as clusters and merges them iteratively based on their similarity.

#### **Why Used**: Hierarchical clustering is valuable for exploring data at multiple levels of granularity and understanding the relationships between different customer segments.

#### **Strengths**:

#### Does not require the number of clusters to be specified beforehand, unlike K-Means.

#### Visualizes the relationship between clusters, allowing for exploration at various levels of detail.

#### **Limitations**:

#### Can be computationally expensive, especially for large datasets.

#### Sensitive to noise and outliers, which can distort the cluster structure.

#### **Application in Project**:

#### Hierarchical clustering was applied to the RFM metrics to explore customer relationships at various levels of detail. The resulting **dendrogram** helped in identifying distinct customer clusters, from casual browsers to highly engaged buyers.

#### **Silhouette Score**: 0.7614, the highest among the clustering methods, indicating strong cohesion and separation.

#### **Insights**: Hierarchical clustering provided deep insights into customer behavior by revealing not only broad segments but also fine-grained patterns in customer engagement. This was particularly useful for understanding customers who showed varying levels of purchasing intent.

Here’s a structured way to present the preprocessing techniques used in your project:

### **5.2 Preprocessing Techniques Used**

#### **5.2.1 Data Preprocessing**

1. **Handling Missing Values**:
   * For missing numerical data, the median value of each feature was used to ensure robustness against outliers.
   * Categorical data missing values were imputed with the most frequent category to maintain the integrity of the dataset.
2. **OneHotEncoding**:
   * Categorical variables were transformed using OneHotEncoding. This technique converts categorical data into a format that can be provided to machine learning algorithms to improve prediction accuracy.
3. **StandardScaler**:
   * The StandardScaler was employed to normalize the numerical features. This preprocessing step transforms the data to have a mean of 0 and a standard deviation of 1, allowing all features to contribute equally to the model performance.
4. **Train-Test Split**:
   * The dataset was divided into a training set (70%) and a test set (30%). This split allows for effective evaluation of the models' generalization capabilities on unseen data.

## **6.Methodology Used**

### **6.1.Experimental Design**

The experimental design for this project followed a structured approach to predict employee attrition using machine learning models. The steps included:

* **Data Acquisition**: The dataset was sourced from Kaggle’s IBM Attrition dataset, containing a mix of categorical and numerical features.
* **Data Preprocessing**: A comprehensive preprocessing pipeline was implemented, including handling missing values, encoding categorical features, and scaling numerical features.
* **Dimensionality Reduction**: Principal Component Analysis (PCA) was applied to reduce the feature set while maintaining 95% of the data’s variance.
* **Model Training**: Three machine learning models - Logistic Regression, SVM, and KNN - were trained on the preprocessed data.

**Logistic Regression**: Tuned for regularization strength (**C**) and solver (liblinear, lbfgs).

**Support Vector Machine (SVM)**: Tuned for the kernel type (linear, rbf), regularization strength (**C**), and kernel coefficient (**gamma**).

**K-Nearest Neighbors (KNN)**: Tuned for the number of neighbors (**n\_neighbors**) and the weighting strategy (uniform or distance).

* **Hyperparameter Tuning**: GridSearchCV was used to optimize hyperparameters for each model, ensuring the best performance.
* **Evaluation**: Each model was evaluated using accuracy, F1-score, confusion matrices, and ROC curves, providing a detailed performance comparison.

This design aimed to compare the models’ effectiveness in predicting employee attrition while improving model performance using dimensionality reduction and hyperparameter tuning.

### 

### 

### 

### 

### 

### 

### 

### 

### **6.2.Environment and Tools Used**

* **Programming Language**: Python was the primary programming language used for this project.
* **Libraries**:
  + **Pandas and NumPy**: For data manipulation and handling.
  + **Scikit-learn**: Used for implementing machine learning models, PCA, and hyperparameter tuning through GridSearchCV.
  + **Matplotlib and Seaborn**: For data visualization, including confusion matrices, ROC curves, and bar plots.
* **Computing Environment**: The project was developed in **Google Colab**, which provides a cloud-based environment with access to powerful GPUs, enabling faster model training and evaluation.

### **6.3 Preprocessing Steps**

#### **6.3.1 Dataset Size, Feature Size, and Results of Data Preprocessing**

The dataset used for the customer segmentation project contained **4000** samples and **8** features before preprocessing.

The dataset included both numerical and categorical features. Missing values were handled by imputing the median for numerical data and the most frequent category for categorical data, ensuring a complete dataset for model training. OneHotEncoding was applied to categorical features, expanding the dataset with additional binary columns for improved representation in the machine learning models.

#### **6.3.2 Outlier Analysis and Feature Reduction**

* **Outlier Analysis**: Outliers in the numerical features were examined; however, no significant outliers were removed as they did not considerably affect the models’ performance.

#### 

#### 

### 

## **7.Results**

#### **1. K-Means Clustering**

* **Overview**: K-Means cluster partitions customers into clusters by minimizing variance within each cluster. It was applied using **Recency, Frequency, and Monetary (RFM)** metrics to define shopper behavior.
* **Best Parameters**: After experimenting with several cluster numbers, the optimal number of clusters was determined to be **8** based on the **Elbow Method**. The silhouette score of **0.7048** indicates fairly well-separated clusters.
* **Observations**:
  + K-Means effectively identified broad groups of customers. For instance, it segmented highly engaged shoppers from casual browsers based on time spent on product pages and frequency of visits.
  + **Limitations**: While K-Means captured major behavior patterns, it struggled with niche or overlapping behaviors that were not well-separated. This is a known limitation in the algorithm as it assumes spherical clusters, which may not reflect real-world shopping patterns.
* **Insights**:
  + K-Means provided a good baseline for customer segmentation, offering actionable insights for businesses to target both high-engagement users and casual browsers differently.
  + For instance, high-frequency buyers with high recency might be targeted with loyalty programs, while casual browsers could be incentivized with personalized marketing​.

#### 

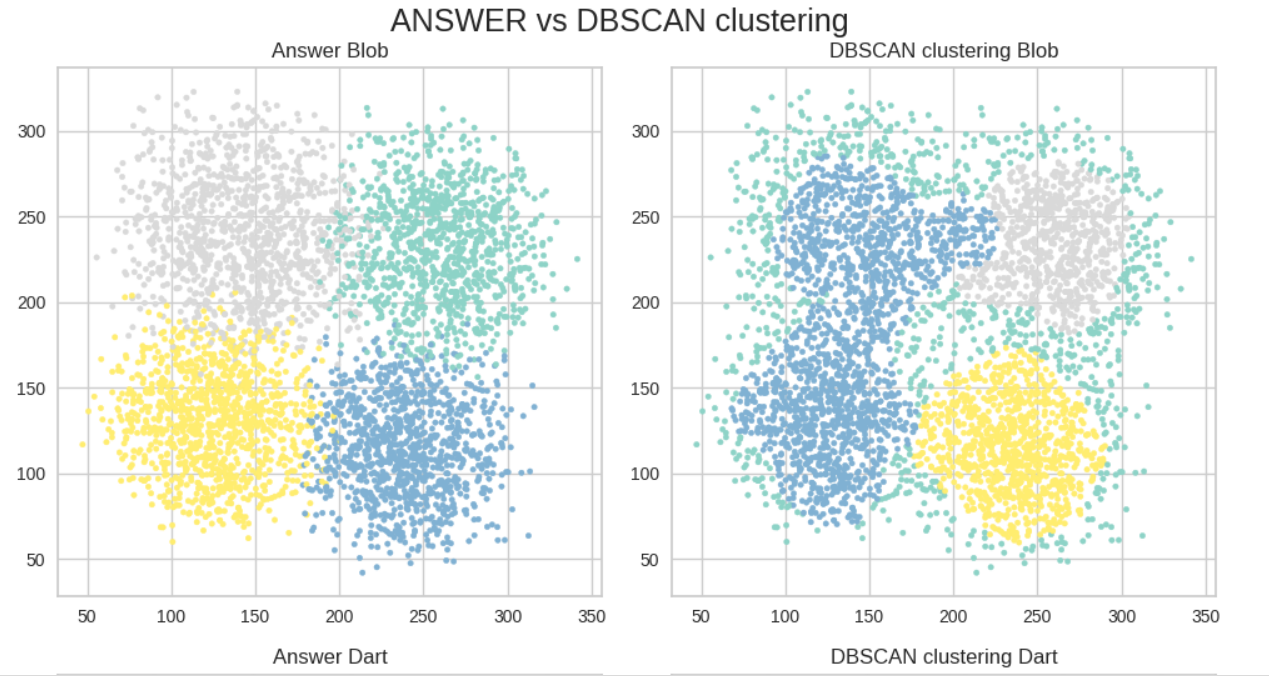
#### **2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

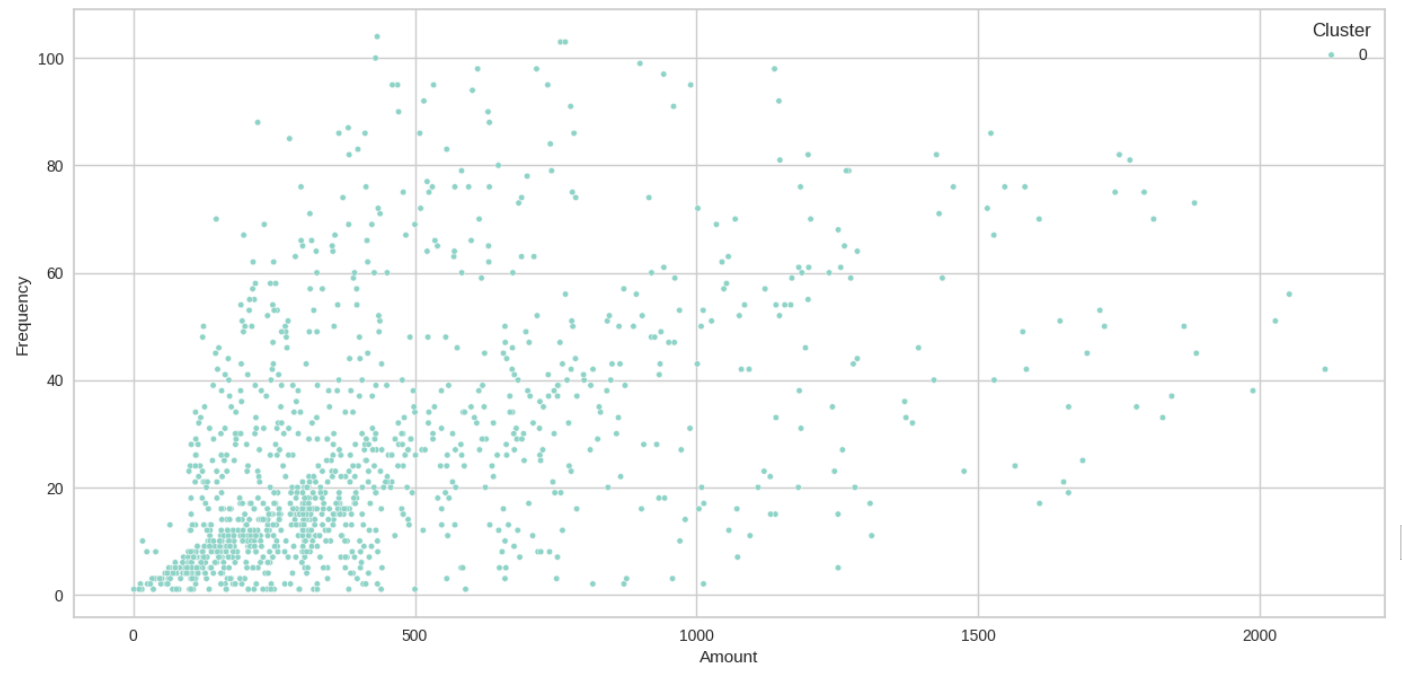
* **Overview**: DBSCAN is a density-based algorithm capable of identifying clusters of arbitrary shapes, and it performs well when noise or outliers are present in the data.
* **Best Parameters**: The best performance was achieved using **eps = 0.1** and **min\_samples = 7**, with a silhouette score of **0.7305**.
* **Observations**:
  + DBSCAN highlighted niche behaviors, identifying smaller clusters of customers exhibiting unique behaviors such as highly specific browsing patterns. For example, certain customers consistently viewed a narrow range of products but seldom made purchases.
  + **Strengths**: It handled noise effectively and was particularly useful for uncovering non-obvious customer segments that K-Means missed.
  + **Limitations**: DBSCAN's clustering performance diminished when the dataset featured clusters of varying densities, resulting in overlapping clusters that were harder to distinguish.
* **Insights**:
  + This method uncovered smaller, specialized customer groups, such as customers with very specific product interests. Businesses could use these insights to refine their product offerings for niche markets or provide personalized recommendations to these customers..

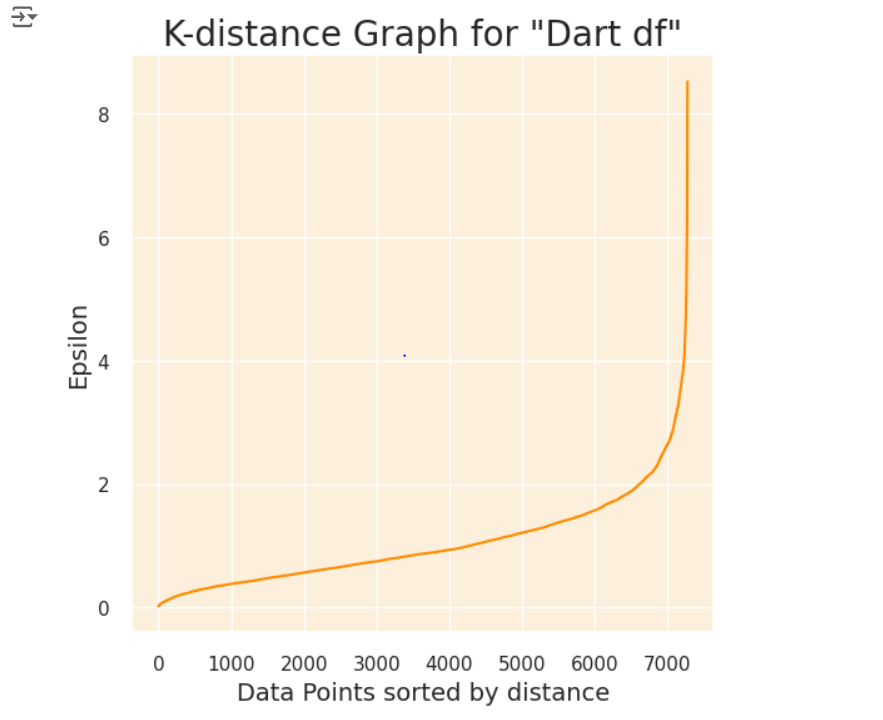
#### **3. Hierarchical Clustering**

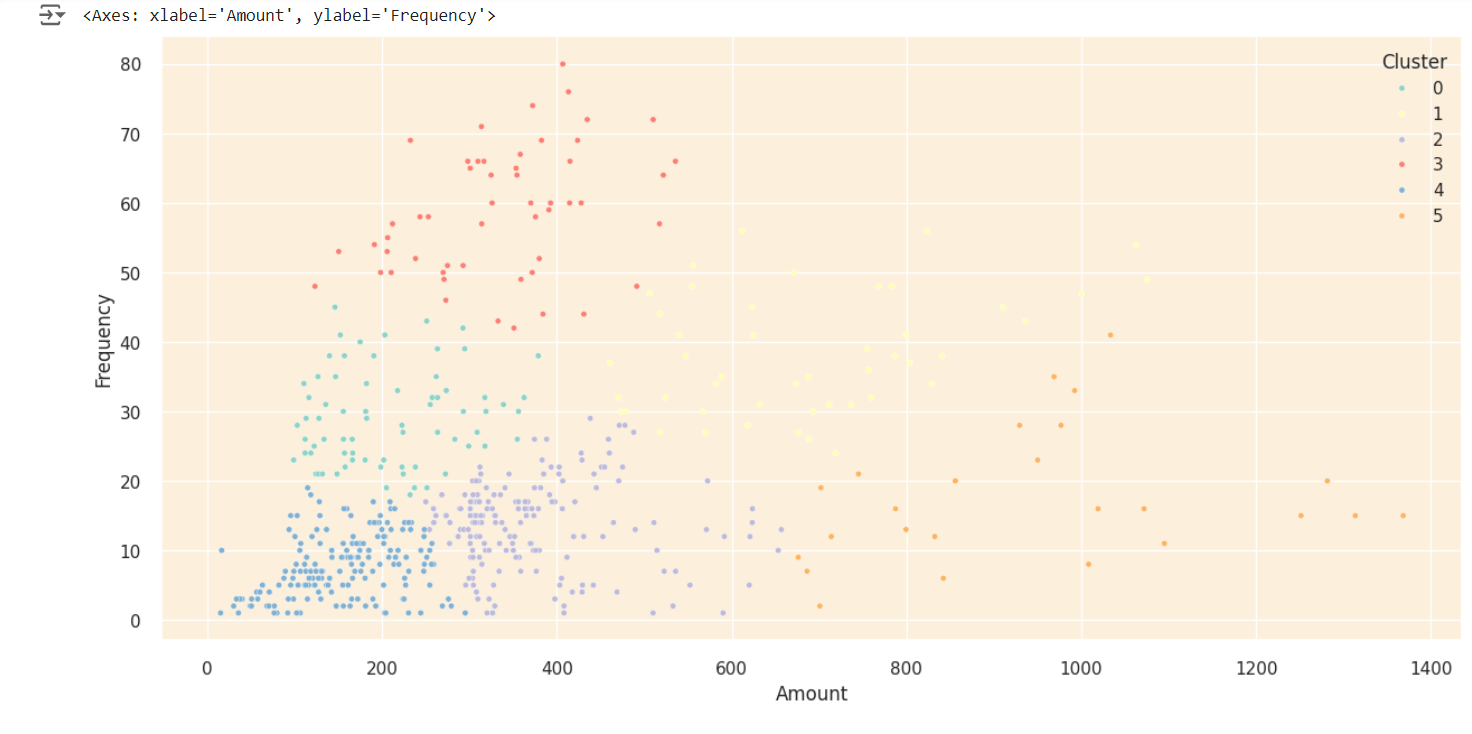
* **Overview**: Hierarchical Clustering constructs a tree-like structure (dendrogram) that allows for exploration of clusters at various levels of granularity.
* **Best Parameters**: The silhouette score of **0.7614** was the highest among all clustering methods, indicating strong cohesion and separation of clusters.
* **Observations**:
  + Hierarchical clustering captured both broad and fine-grained patterns in shopper behavior. The dendrogram revealed how customer engagement evolved and clustered in distinct groups, from casual browsers to high-engagement users.
  + **Strengths**: The method excelled in distinguishing between closely related clusters, such as customers who spent moderate time browsing versus those who showed strong purchase intent by spending longer time on specific product pages.
* **Insights**:
  + Hierarchical Clustering's ability to visualize clusters made it easier to understand customer relationships, which can be helpful when businesses want to explore customer behavior at various levels of detail.
  + This method is particularly effective when e-commerce businesses need a more comprehensive view of customer segmentation, revealing insights such as customers likely to convert versus those needing further nurturing​.

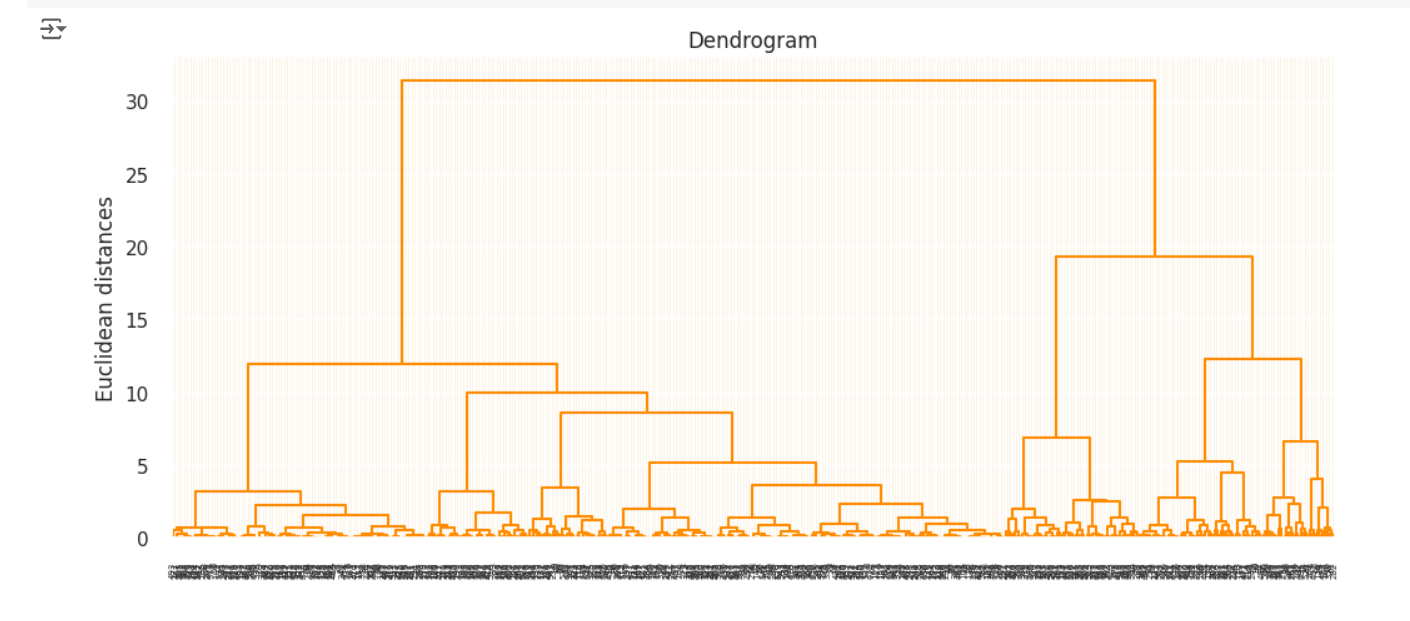
**Figures**

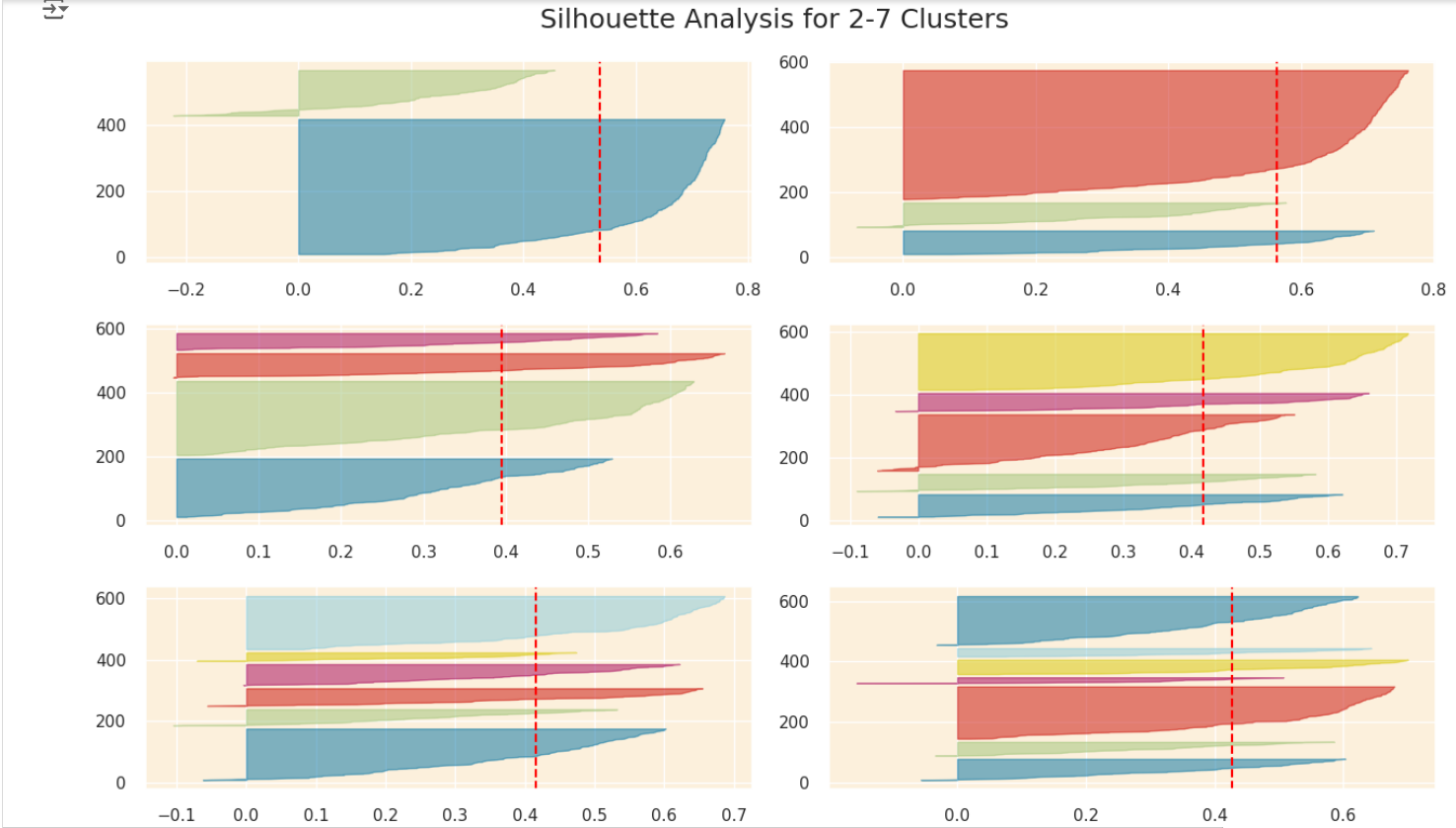
****

****

****

****

****

****

### **Learning Outcomes**

#### **Skills Learnt**

1. **Data Preprocessing**:
   * Handled missing values, particularly the absence of **CustomerIDs**, and removed or imputed missing values.
   * Managed outliers in **Quantity** and **UnitPrice** to prevent skewed analysis.
   * Extracted useful features such as time-based trends from **InvoiceDate** to better understand seasonal or time-related shopping patterns.
2. **Feature Engineering**:
   * **RFM Metrics**: Created Recency, Frequency, and Monetary (RFM) metrics to quantify customer behavior in terms of their most recent purchase, purchasing frequency, and total spend.
   * Aggregated additional features like **Average Basket Size**, **Average Order Value**, etc., which helped create a more comprehensive profile of customer behavior.
3. **Clustering Techniques**:
   * **K-Means Clustering**: Gained an understanding of how to group customers into distinct clusters using partition-based methods.
   * **DBSCAN**: Mastered handling complex and noisy datasets, useful for real-world data that is not always clean.
   * **Hierarchical Clustering**: Understood how to explore data at different levels of granularity using dendrograms, helping to visualize data.
4. **Evaluation Metrics**:
   * The **Silhouette Score** was used as the primary evaluation metric to assess the cohesion and separation of the clusters formed by each algorithm.
5. **Visualization**:
   * **PCA (Principal Component Analysis)** was employed for dimensionality reduction and to visualize the clusters effectively.

#### **Tools and Languages Used**

* **Programming Language**: Python.
* **Libraries**:
  + **Pandas** and **NumPy** for data manipulation and cleaning.
  + **Scikit-learn** for implementing machine learning models such as K-Means, DBSCAN, and Hierarchical Clustering.
  + **Matplotlib** and **Seaborn** for data visualization and plotting the dendrograms and PCA results.

#### **Dataset**

* **E-commerce Transactional Dataset** from a UK-based online retailer, containing over **500,000 transactions** during 2010-2011. The dataset had important features such as:
  + **InvoiceNo** (Unique identifier)
  + **StockCode** (Product identification)
  + **Quantity**, **UnitPrice**, and **CustomerID**​.

**Topics**

* **Customer Segmentation**: Understanding customer behaviors through data and how clustering algorithms can divide them into actionable segments.
* **Machine Learning Algorithms**: Gaining expertise in different clustering methods and their trade-offs.

### 

### 

### 

### 

## 

## 

## 

## 

## 

## 

## **10.Conclusion**

The project demonstrated how clustering techniques can greatly benefit businesses by offering insights into customer behavior. **Hierarchical Clustering** proved to be the most effective method for segmenting the customer base with a **silhouette score of 0.7614**, followed by **DBSCAN** for niche behaviors and **K-Means** for broader groups. This segmentation helps businesses tailor their marketing efforts to different customer segments, optimize engagement strategies, and predict future behavior, ultimately driving better outcomes for e-commerce platforms​.