**QM 525 - Mathematics of AI and ML**  
**Group D**  
**Members:** Umer Khatab, Andreea, Hashir, Mariyam Fathima  
**Date:** November 10, 2025

# Customer Segmentation for Retail Sales Using K-Means Clustering

## 1. Introduction and Problem Context

In today's highly competitive retail landscape, understanding customer behavior has become essential for business success. Retailers face a fundamental challenge: their customer base is not homogeneous. Some customers purchase frequently and spend generously, making them valuable assets to the business. Others shop occasionally during sales events, demonstrating price sensitivity. Still others may have been loyal in the past but are now at risk of churning. Treating all these diverse customers with the same marketing approach leads to significant inefficiencies and missed opportunities (John et al., 2023).

The traditional one-size-fits-all marketing strategy produces disappointing results. Industry research shows that untargeted marketing campaigns typically achieve response rates of only 1-3%, meaning that 97-99% of marketing expenditure reaches customers who are unlikely to respond (McKinsey & Company, 2021). This inefficiency extends beyond marketing costs. Without understanding customer segments, retailers struggle with inventory planning, stocking products that don't appeal to their actual customer base while running out of items that high-value customers want. Customer retention efforts also suffer when businesses cannot identify which customers are at risk of leaving and which require nurturing to increase their lifetime value.

The UK online retail sector exemplifies these challenges at scale. Consider a typical online retailer managing 541,909 transactions across 4,372 unique customers over a twelve-month period. Each customer exhibits distinct purchasing patterns across multiple dimensions: how recently they shopped, how frequently they make purchases, what product categories they prefer, and how much they typically spend. Manually analyzing and segmenting thousands of customers based on these behavioral attributes is simply impractical. Marketing teams need an automated, data-driven approach to identify meaningful customer groups and develop targeted strategies for each segment (John et al., 2023).

Customer segmentation the practice of dividing a customer base into distinct groups with similar characteristics offers a solution to these challenges. However, traditional segmentation methods based solely on demographics or simple business rules lack the sophistication needed to capture the complexity of modern customer behavior. This is where artificial intelligence and machine learning provide transformative capabilities. Unsupervised learning algorithms, particularly clustering techniques, can automatically discover natural groupings within customer data without requiring pre-labeled examples. These algorithms analyze multiple behavioral dimensions simultaneously, identifying patterns that might not be obvious to human analysts.

This project analyzes a real-world application of K-Means clustering combined with RFM (Recency, Frequency, Monetary) analysis for customer segmentation in the UK online retail market. Our analysis examines how mathematical foundations from QM 525 including linear algebra, optimization theory, and statistical methods enable these AI/ML techniques to solve practical business problems. We explore not only what the algorithm accomplishes but also why the underlying mathematics matters. Through this case study, we demonstrate how theoretical concepts translate into tangible business value: improved marketing efficiency, better inventory management, enhanced customer retention, and increased profitability.

The remainder of this report is organized as follows. Section 2 presents a detailed analysis of the UK online retail case study, explaining the RFM framework and K-Means implementation. Section 3 examines the mathematical principles that make clustering algorithms work. Section 4 discusses implementation challenges encountered in real-world applications. Section 5 evaluates the business impact and outcomes. Section 6 connects the case study explicitly to concepts covered in QM 525. Section 7 concludes with key findings and future implications.

## 2. Case Study Analysis: UK Online Retail Customer Segmentation

### 2.1 Dataset Overview and Business Context

The UK Online Retail dataset, obtained from the UCI Machine Learning Repository, provides a comprehensive foundation for analyzing customer segmentation techniques. This dataset contains 541,909 transaction records from a UK-based online retailer specializing in unique all-occasion gifts. The transactions span a twelve-month period from December 2010 through December 2011 and represent purchases from 4,372 unique customers, primarily based in the United Kingdom with some international orders (John et al., 2023).

The dataset includes eight key features that capture essential transaction information. The InvoiceNo field provides a unique six-digit identifier for each transaction, with codes beginning with 'c' indicating cancellations. The StockCode identifies individual products using a five-digit code, covering over 5,000 distinct items in the retailer's catalog. Product names appear in the Description field. The Quantity field records how many units of each product were purchased in a transaction, while UnitPrice captures the price per item in British pounds sterling. The InvoiceDate timestamp records exactly when each transaction occurred, and the CustomerID field uniquely identifies each customer. Finally, the Country field indicates where the customer is located.

This retailer faced typical challenges in understanding and acting on customer behavior patterns. With thousands of active customers purchasing from an extensive product catalog, the company needed answers to critical business questions. Which customers represent the highest value and deserve premium service? Which customers are at risk of churning and require re-engagement campaigns? Which customers show potential for growth through targeted promotions? Which customers are no longer active and should be deprioritized? Without an effective segmentation strategy, the retailer could not efficiently allocate marketing resources or optimize inventory for different customer needs.

### 2.2 The RFM Framework for Customer Value Quantification

Before applying clustering algorithms, the study transformed raw transactional data into meaningful customer-level features using the RFM model a proven marketing framework that captures three critical dimensions of customer behavior (John et al., 2023).

**Recency (R)** measures how recently a customer made a purchase. It is calculated as the number of days between the customer's most recent transaction and the analysis date. Recent purchasers demonstrate current engagement with the brand and are statistically more likely to make additional purchases in the near future. In the RFM scoring system, customers who purchased more recently receive higher scores, typically on a scale from 1 to 5.

**Frequency (F)** measures how often a customer purchases. It is calculated by counting the total number of distinct transactions a customer has made during the observation period. Frequent buyers demonstrate loyalty and consistent engagement with the retailer. Customers with higher purchase frequency receive higher frequency scores, indicating they are more engaged than occasional shoppers.

**Monetary (M)** measures the total amount a customer has spent. It is calculated as the sum of all purchase amounts across all transactions for each customer. High-spending customers represent the greatest direct revenue contribution to the business. In the scoring system, customers who spend more receive higher monetary scores.

For the UK online retail dataset, RFM metrics were calculated using the following approach. For each unique CustomerID, Recency was computed as the difference between the analysis date and the maximum InvoiceDate for that customer. Frequency was calculated by counting distinct InvoiceNo values associated with each CustomerID. Monetary value was computed by multiplying Quantity by UnitPrice for each transaction line item and then summing these amounts for all transactions belonging to each CustomerID.

Each customer then received three scores (R, F, and M) on a 1-5 scale, where higher numbers indicate more favorable characteristics. These three scores combine to create a comprehensive customer profile. For example, a customer with R=5, F=5, M=5 represents the ideal "Champion" customer someone who purchased very recently, buys frequently, and spends heavily. In contrast, a customer with R=1, F=1, M=1 represents a "Lost" customer who has not purchased in a long time, rarely shopped even when active, and spent very little.

The RFM framework's power lies in its simplicity and business relevance. Unlike complex behavioral models that may be difficult for marketing teams to understand and act upon, RFM provides intuitive customer profiles that directly inform marketing strategies. A customer with high recency and frequency but low monetary value might be a frequent small-purchase customer who could respond well to bundle offers. A customer with high monetary but declining recency might be a previously valuable customer at risk of churning who needs personalized re-engagement.

### 2.3 Data Preprocessing and Preparation

Real-world data always requires cleaning and preparation before analysis. The UK retail dataset presented several challenges that needed to be addressed during preprocessing. First, 135,080 transaction records (approximately 25% of the dataset) contained missing values in the CustomerID field. These records represented guest checkouts where customers did not create accounts or log in. Since customer segmentation requires tracking individual customers over time, these records were removed from the analysis. While this represents a limitation guest customers may exhibit different behaviors than registered users it was necessary for conducting meaningful segmentation (John et al., 2023).

Second, the dataset contained 8,905 records with negative values in the Quantity field. These negative quantities represented product returns rather than purchases. For this segmentation analysis focused on purchase behavior, the study removed these return records to focus specifically on what customers buy rather than what they return. Additionally, duplicate records rows with identical values across all fields were identified and removed to ensure each transaction was counted only once.

After these preprocessing steps, the cleaned dataset contained 397,924 transaction records across 4,372 unique customers. This reduction from the original 541,909 records highlights a common challenge in data-driven projects: not all collected data is usable for every analytical purpose.

A critical preprocessing step for K-Means clustering is feature scaling. The RFM metrics exist on completely different scales. Recency is measured in days (ranging from 1 to 365 or more). Frequency is measured in transaction counts (ranging from 1 to perhaps 50 for very active customers). Monetary value is measured in currency (ranging from tens to thousands of pounds). Without normalization, the K-Means algorithm's distance calculations would be dominated by the feature with the largest numeric range in this case, Monetary value in pounds. This would essentially ignore the Recency and Frequency dimensions, producing poor segmentation.

To address this issue, Min-Max normalization was applied to scale all RFM features to a common range of 0 to 1. The normalization formula transforms each value: x' = (x - x\_min) / (x\_max - x\_min), where x represents the original value, x\_min and x\_max represent the minimum and maximum values in that feature, and x' represents the normalized value. After this transformation, all three RFM dimensions contribute equally to distance calculations, ensuring balanced segmentation based on all three behavioral aspects.

### 2.4 K-Means Implementation and Cluster Selection

With RFM features calculated and normalized, the study applied K-Means clustering to automatically group similar customers. K-Means was selected because it is computationally efficient for large datasets, produces easily interpretable results, and is well-suited for the relatively compact, spherical clusters expected in RFM space (Jain, 2010).

A critical decision in K-Means clustering is selecting the number of clusters (K). Too few clusters result in overly broad groupings that lack actionable specificity. Too many clusters create fragmented segments that are difficult to manage and may not represent meaningfully different customer behaviors. The study employed two complementary methods to determine the optimal K value.

The **Elbow Method** evaluates clustering quality across different K values by plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters. WCSS measures how tightly points are grouped within their assigned clusters lower values indicate more compact, homogeneous clusters. As K increases, WCSS naturally decreases because more clusters mean each cluster contains fewer, more similar points. However, at some point, adding additional clusters provides diminishing returns. The "elbow" in the plot where the curve begins to flatten suggests an optimal K value that balances cluster compactness with parsimony. For the UK retail data, testing K values from 2 to 10 revealed an elbow at K=4, suggesting four natural customer groups.

The **Silhouette Analysis** provides a complementary validation approach. The Silhouette Score measures how well each point fits its assigned cluster compared to other clusters. Scores range from -1 to +1, where values near +1 indicate the point is well-matched to its cluster and far from neighboring clusters, values near 0 indicate the point is on the border between clusters, and negative values suggest possible misclassification. The average Silhouette Score across all data points provides an overall quality metric for the clustering solution.

Based on these analyses, the study implemented K-Means with K=4 clusters. The algorithm was initialized using K-Means++, an improved initialization method that intelligently spreads initial centroid positions across the data space rather than placing them randomly. This initialization approach helps avoid poor local optima that can occur with pure random initialization. The algorithm then iteratively assigned each customer to the nearest centroid and recalculated centroids as the mean position of assigned customers. This process continued for 15 iterations until centroids stabilized and no longer moved significantly between iterations.

### 2.5 Customer Segments and Business Interpretation

The K-Means clustering revealed distinct customer segments with clear business interpretations based on their RFM profiles. However, for more granular segmentation, the study also classified customers into five categories based on their overall RFM scores: Top Customers (8% of the customer base), High-Value Customers (10%), Medium-Value Customers (21%), Low-Value Customers (30%), and Lost Customers (31%) (John et al., 2023).

**Top Customers** (8%) represent the elite segment with RFM scores above 4.5. These customers purchased very recently, shop frequently, and spend the most. They are the retailer's most valuable assets, contributing disproportionately to revenue despite being a small percentage of the customer base. The recommended strategy for this segment includes VIP treatment, exclusive early access to new products, personalized service, and loyalty rewards programs to maintain their engagement and prevent competitive poaching.

**High-Value Customers** (10%) have RFM scores above 4.0. While slightly less engaged than Top Customers, they still represent strong, consistent purchasers. These customers should receive premium attention through engagement programs, opportunities for upselling to higher-value products, and referral incentives that leverage their positive experience to acquire similar high-value customers.

**Medium-Value Customers** (21%) have RFM scores above 3.0. They represent solid, moderate purchasers who shop occasionally and spend reasonably. This segment has significant growth potential. Marketing strategies should focus on increasing purchase frequency through targeted promotions, cross-selling complementary products, and personalized recommendations based on their previous purchases.

**Low-Value Customers** (30%) have RFM scores above 1.6 but below 3.0. These customers shop infrequently and spend relatively little. While they contribute some revenue, the cost of serving them may approach or exceed their value. Strategies for this segment should be cost-effective, using automated email campaigns and promotions rather than expensive personalized outreach. The goal is either to help them grow into Medium-Value customers or to minimize the cost of maintaining their minimal engagement.

**Lost Customers** (31%) have RFM scores below 1.6. These customers have not purchased in a long time, rarely shopped even when active, and spent very little. Re-engaging this segment typically requires significant investment with low success rates. The recommended approach is to minimize marketing spend on this group, perhaps conducting one low-cost reactivation campaign to identify any customers who might be recoverable, but otherwise deprioritizing them in favor of more valuable segments.

This segmentation provides actionable business intelligence. Marketing teams can now allocate budgets efficiently, focusing resources on high-value segments while minimizing waste on unlikely responders. Inventory managers can stock products preferred by valuable customer segments. Customer service teams can prioritize support for Top and High-Value customers. Product development can focus on features that appeal to growing Medium-Value customers.

### 2.6 Algorithm Performance and Comparison

While K-Means clustering provided good results for this segmentation task, the study also compared its performance against four other clustering algorithms: Gaussian Mixture Model (GMM), DBSCAN (Density-Based Spatial Clustering of Applications with Noise), BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), and Agglomerative Clustering. Performance was evaluated using the Silhouette Score as a standardized metric (John et al., 2023).

The results showed that **GMM achieved the highest performance with a Silhouette Score of 0.80**, significantly outperforming the other approaches. GMM was enhanced with Principal Component Analysis (PCA) for dimensionality reduction, which helped capture the variance structure in the data more effectively. The probabilistic nature of GMM allows it to model overlapping clusters and capture uncertainty in cluster assignments, leading to better performance on this dataset.

**K-Means achieved a Silhouette Score of 0.64**, which represents good clustering quality, though lower than GMM. DBSCAN scored 0.626, BIRCH scored 0.64, and Agglomerative Clustering scored 0.64. While GMM demonstrated superior statistical performance, K-Means remains highly valuable for business applications because of its simplicity and interpretability. Business stakeholders can easily understand how K-Means works and why customers are grouped together, which is crucial for implementation and adoption. GMM's probabilistic assignments and more complex mathematical foundation can be harder to explain to non-technical marketing teams.

This comparison demonstrates an important principle: the "best" algorithm depends on the context. For pure statistical accuracy, GMM performed better. For business implementation requiring interpretability and stakeholder buy-in, K-Means offers significant advantages. The case study shows that multiple approaches can be evaluated, and the final choice should balance statistical performance with practical business considerations.

3. Mathematical principles - deep dive

Overview of the principles:

This study adopts an unsupervised machine learning approach to segment customers based on behavioral attributes derived from the Recency-Frequency-Monetary (RFM) model. To achieve this, multiple advanced clustering algorithms were used to uncover patterns in customer data without predefined labels. Specifically, the analysis incorporates K-Means, which partitions observations into clusters by minimizing intra-group variance; the Gaussian Mixture Model (GMM), which applies a probabilistic framework to allow soft cluster assignments; DBSCAN, a density-based method capable of identifying clusters of arbitrary shape while isolating noise; Agglomerative Clustering, a hierarchical technique that iteratively merges clusters based on linkage criteria; and BIRCH, an algorithm optimized for large-scale datasets through incremental clustering using hierarchical structures. Together, these methods provide a comprehensive foundation for identifying distinct customer segments, enabling more targeted marketing and strategic decision-making.

RFM

The Recency-Frequency-Monetary (RFM) model is a widely used framework for customer segmentation, emphasizing three critical dimensions: how recently a customer made a purchase, how frequently they transact, and the monetary value of their spending. By analyzing these factors, businesses can identify distinct customer segments and tailor marketing strategies to maximize engagement and profitability. The simplicity and efficiency of the RFM model make it an effective tool for resource allocation, enabling companies to prioritize high-value customers and foster long-term loyalty.

A screenshot of a graph

AI-generated content may be incorrect.

RFM Score = (Recency Score × Recency Weight) + (Frequency Score × Frequency Weight) + (Monetary Score × Monetary Weight)

Recency Weight:

Emphasizes how recently a customer interacted or purchased. A higher weight means recent activity is considered more predictive of future behavior.

Frequency Weight:

Reflects how often a customer engages. A higher weight prioritizes customers with repeated interactions, signaling loyalty.

Monetary Weight:

Indicates the importance of the amount spent. A higher weight favors high-value customers in segmentation or scoring.

|  |  |
| --- | --- |
| **Customer ID** | **RFM\_Score** |
| 12346 | 0.06 |
| 12347 | 4.48 |
| 12348 | 2.09 |
| 12349 | 3.41 |
| 12350 | 1.1 |

For the top 10 records in the dataset.

|  |  |  |
| --- | --- | --- |
| **Customer ID** | **RFM\_Score** | **Customer\_Segment** |
| 12346 | 0.06 | Lost customer |
| 12347 | 4.48 | High-value customer |
| 12348 | 2.09 | Low-value customer |
| 12349 | 3.41 | Medium-value customer |
| 12350 | 1.10 | Lost customer |
| 12352 | 3.46 | Medium-value customer |
| 12353 | 0.03 | Lost customer |
| 12354 | 2.67 | Low-value customer |
| 12355 | 0.93 | Lost customer |
| 12356 | 3.11 | Medium-value customer |

Customer segmentation:

A pie chart with numbers and a few words

AI-generated content may be incorrect.

**Principal Component Analysis**

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms multivariate data into a new set of uncorrelated variables called principal components. These components are derived by decomposing the covariance matrix of the original data and computing its eigenvalues and eigenvectors. The first principal component captures the direction of maximum variance in the dataset, while subsequent components capture the next highest variance in directions orthogonal to the previous ones. This process allows PCA to reduce complexity while preserving as much variability as possible, making it useful for visualization, noise reduction, and improving the performance of machine learning models.

First step to computing PCA is to compute the covariance matrix. The covariance for two-dimensional variables is:

A math equations on a white paper

AI-generated content may be incorrect.

The covariant matrix for multivariate data can be obtained:

A math equations on a white board

AI-generated content may be incorrect.

Covariance matrix for a set of data with n rows and n columns:

A close up of a note

AI-generated content may be incorrect.

Covariance matrix for three variables:

A close-up of a white paper with writing on it

AI-generated content may be incorrect.

The eigenvalues λ and eigenvectors X of the covariance matrix can be expressed using the formula:

A white paper with writing on it

AI-generated content may be incorrect.

A is an n × n square matrix

I is an identity matrix of the same shape as A

A diagram of components of a principal component

AI-generated content may be incorrect.

Principal Component Analysis (PCA):

Principal Component 1: The direction of maximum variance in the data.

Principal Component 2: Orthogonal to the first, capturing the next highest variance.

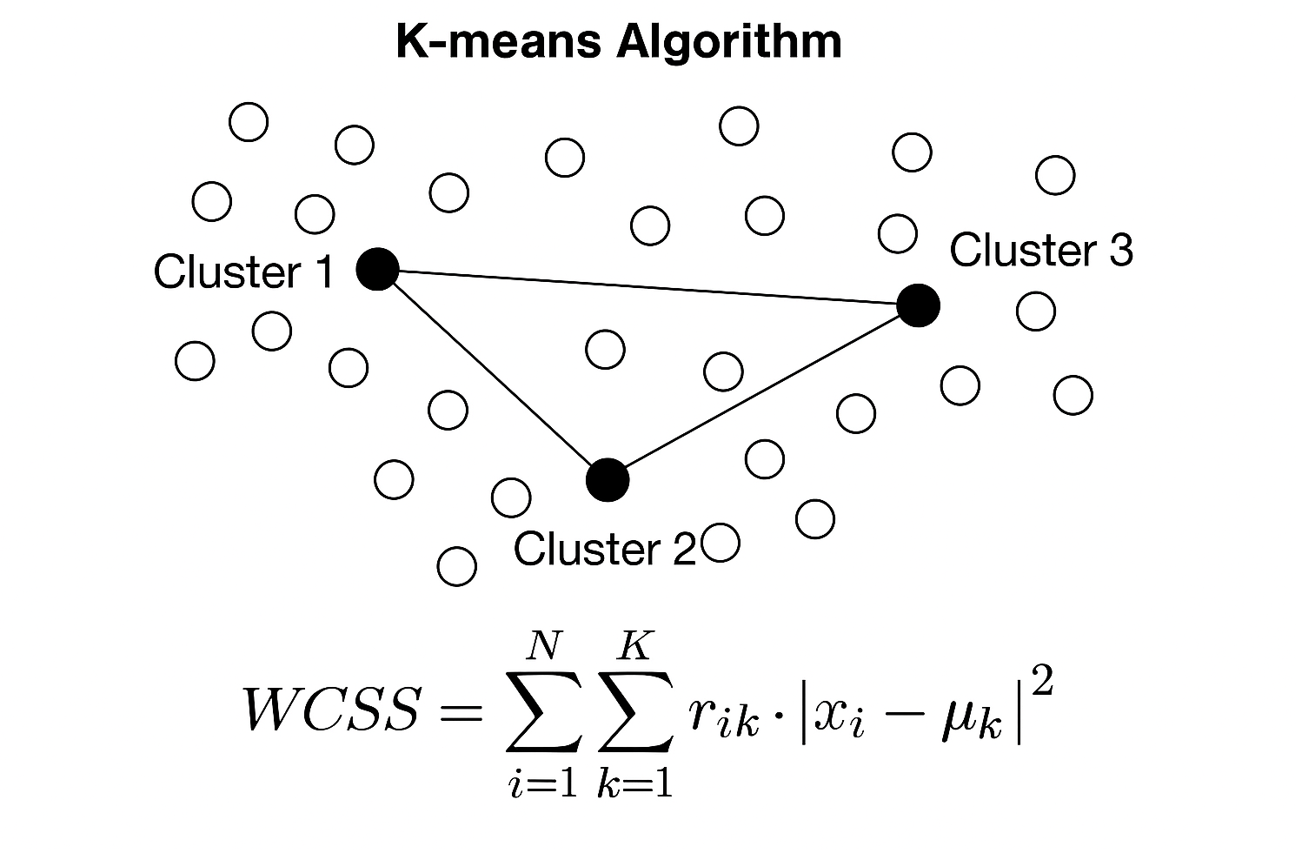
The scatter plot shows how PCA transforms data into these new axes for dimensionality reduction.

**K-Means Clustering**

K-Means clustering was applied to segment customers based on behavioral attributes after normalizing the dataset using Min–Max scaling to ensure all features were on the same scale. The optimal number of clusters was determined using the elbow method, which involved plotting inertia values for cluster counts ranging from 2 to 10 and identifying the sharp bend at three clusters. A K-Means model with **k = 3** was then trained, producing compact clusters with distinct centroids. The quality of clustering was evaluated using the Silhouette Score, confirming good cohesion and separation among clusters. Finally, the results were visualized in a scatter plot, with points colored by cluster labels to illustrate the segmentation clearly.

A whiteboard with writing on it

AI-generated content may be incorrect.



where:

N: number of data points.

K: number of clusters.

Xi : ith data point.

µK: kth cluster centroid.

rrk: indicator variable that is returned as 1 if data point xi belongs to cluster k and 0 otherwise.

The K-means algorithm aids in customer profiling in consumer segmentation, helping firms identify segments with distinctive behavioral patterns, purchase frequencies and preferences.

**The flowchart for the K-Means algorithm:**

**A diagram of a process

AI-generated content may be incorrect.**

**GMM**

The Gaussian Mixture Model (GMM) is a probabilistic clustering algorithm that represents data as a combination of multiple Gaussian distributions, making it highly effective for complex customer segmentation. Unlike K-Means, which assumes uniform spherical clusters, GMM accommodates clusters of varying shapes, sizes, and densities, providing flexibility in modeling real-world consumer behavior. It assigns each data point a probability of belonging to different clusters, enabling soft clustering rather than hard assignments. Using the Expectation-Maximization (EM) algorithm, GMM iteratively estimates cluster membership probabilities and updates parameters such as means, covariances, and mixing weights to maximize likelihood. This approach captures subtle patterns and overlapping segments, making it ideal for identifying nuanced customer preferences and improving targeted marketing strategies.

A close-up of a piece of paper

AI-generated content may be incorrect.

**Gaussian Mixture Model (GMM)**:

A diagram of a mixture model

AI-generated content may be incorrect.

**DBSCAN Algorithm**  
DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering method that groups data points based on density rather than distance, making it effective for identifying clusters of varying shapes and sizes in noisy datasets. It classifies points as core, border, or noise using two key parameters: eps (maximum distance between points in a neighborhood) and min\_samples (minimum number of points to form a dense region). This approach is particularly useful for customer segmentation when data contains irregular patterns or outliers, as it can detect both dense clusters and sparse regions without requiring a predefined number of clusters.

This is calculated:

reachability distance(p,q) = max(core distance (p),||p-q||)

**∥p−q∥** - This is the Euclidean distance (or chosen metric) between points p and q.

It measures their direct spatial separation.

**core\_distance(p)** - This represents the density around point p.

A diagram of a line and a circle with text

AI-generated content may be incorrect.

This shows that the reachability distance is the larger of:

* The density-based core distance of p.
* The actual spatial distance between p and q.

**BIRCH Algorithm**  
BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is designed for large datasets and memory efficiency. It operates in two phases: first, it builds a Cluster Feature Tree (CFT) to summarize data into compact subclusters; second, it refines these subclusters by merging them into larger clusters. This hierarchical approach allows BIRCH to handle massive datasets quickly while preserving structural relationships between clusters. Its scalability and speed make it ideal for customer segmentation in environments where data volume is high and computational resources are limited.

A writing on a white board

AI-generated content may be incorrect.

N – number of data points in the cluster

LSUM – sum of data points in the cluster

**Flowchart for the BIRCH algorithm:**

**A diagram of a flowchart

AI-generated content may be incorrect.**

**Agglomerative Algorithm**  
Agglomerative clustering is a hierarchical technique that begins by treating each data point as an individual cluster and then iteratively merges the closest clusters until a desired number of clusters is reached. The merging process is guided by linkage criteria such as single linkage (minimum distance), complete linkage (maximum distance), average linkage, or Ward’s linkage, which minimizes variance within clusters. This method is well-suited for customer segmentation when understanding hierarchical relationships between groups is important, as it provides a clear structure of how clusters combine at different levels.

These approaches define how the distance between two clusters is calculated based on the pairwise distances of their points. Here are the common linkage equations:

**Single Linkage**: Distance between clusters = **minimum** pairwise distance between points in A and B.

**Complete Linkage**: Distance = **maximum** pairwise distance between points in A and B.

**Average Linkage**: Distance = **average** of all pairwise distances between points in A and B.

**Centroid Linkage**: Distance = distance between the **centroids** of A and B.

A close-up of a white paper with writing on it

AI-generated content may be incorrect.

A group of math equations

AI-generated content may be incorrect.

**Experimental Setup visualization:**

A diagram of a process

AI-generated content may be incorrect.

5. Implementation Challenges

The study faced several challenges across technical, ethical, and business dimensions. The technical challenges in this study were significant and multifaceted. First, data quality issues dominated the preprocessing phase, with 25% of CustomerIDs missing, null values in product descriptions, and 8,905 entries containing negative quantities. These anomalies required rigorous cleaning, including removal of incomplete records and duplicates, which reduced the dataset from 541,909 to 397,924 rows. Second, data preprocessing complexity was high: scaling features using Min–Max normalization was necessary for algorithms like K-Means and DBSCAN to avoid bias from differing value ranges, while dimensionality reduction through PCA was applied to improve computational efficiency and visualization for Gaussian Mixture Models (GMM). Third, algorithmic limitations posed challenges—K-Means was sensitive to initial centroid selection, leading to variability in cluster quality, and DBSCAN required careful tuning of eps and min\_samples parameters, which lacked clear knee points in K-distance plots. Additionally, handling large datasets introduced computational overhead, demanding optimized Python code and efficient use of libraries such as Scikit-Learn and Deap to minimize runtime. Finally, evaluation constraints emerged because the Silhouette Score, while useful for measuring cohesion and separation, did not fully capture business relevance, requiring supplementary interpretation to ensure practical applicability of clustering results.

Ethical and privacy concerns centered on handling sensitive identifiers like CustomerID and Country, ensuring compliance with GDPR, and mitigating risks of bias or discriminatory segmentation. Transparency was also critical to explain complex models without exposing personal data. From a business perspective, incomplete data threatened segmentation accuracy, while implementing advanced machine learning models demanded significant resources compared to simpler RFM approaches. Additionally, complex clustering outputs posed interpretability challenges for non-technical stakeholders, requiring clear communication to align segmentation insights with strategic goals.

4. Implementation Challenges

The study faced several challenges across technical, ethical, and business dimensions. The technical challenges in this study were significant and multifaceted. First, data quality issues dominated the preprocessing phase, with 25% of CustomerIDs missing, null values in product descriptions, and 8,905 entries containing negative quantities. These anomalies required rigorous cleaning, including removal of incomplete records and duplicates, which reduced the dataset from 541,909 to 397,924 rows. Second, data preprocessing complexity was high: scaling features using Min–Max normalization was necessary for algorithms like K-Means and DBSCAN to avoid bias from differing value ranges, while dimensionality reduction through PCA was applied to improve computational efficiency and visualization for Gaussian Mixture Models (GMM). Third, algorithmic limitations posed challenges—K-Means was sensitive to initial centroid selection, leading to variability in cluster quality, and DBSCAN required careful tuning of eps and min\_samples parameters, which lacked clear knee points in K-distance plots. Additionally, handling large datasets introduced computational overhead, demanding optimized Python code and efficient use of libraries such as Scikit-Learn and Deap to minimize runtime. Finally, evaluation constraints emerged because the Silhouette Score, while useful for measuring cohesion and separation, did not fully capture business relevance, requiring supplementary interpretation to ensure practical applicability of clustering results.

Ethical and privacy concerns centered on handling sensitive identifiers like CustomerID and Country, ensuring compliance with GDPR, and mitigating risks of bias or discriminatory segmentation. Transparency was also critical to explain complex models without exposing personal data. From a business perspective, incomplete data threatened segmentation accuracy, while implementing advanced machine learning models demanded significant resources compared to simpler RFM approaches. Additionally, complex clustering outputs posed interpretability challenges for non-technical stakeholders, requiring clear communication to align segmentation insights with strategic goals.

### 5. Impact Evaluation

When investigating the outcome of the clustering analysis, one is able to see the effects even better than the charts and statistics may be implying. When the project was starting, the data appeared as a mere long list of invoices, quantities and prices, and hardly interpretable meaningfully without some form behind it. However, after the RFM scores were calculated and the algorithms of clustering the customers into different groups began their work, the entire behaviour of the retail market was suddenly apparent and, frankly speaking, more predictable. The segmentation opened up patterns that retailers tend to intuitively feel but can never say with accuracy without data. As an example, the outcome of the analysis that has shown that 8 out of 100 customers are the so-called Top Customers, and 10 out of 100 are the so-called High-Value group, proved a fact that is frequently reported in the retail analytics: there is always a very small proportion of customers that bring the majority of revenue. Its organic appearance out of the UK data gives that notion its play, in the sense that it is not pegged to an industry standard or the spending patterns of another market, but the actual spending patterns as recorded in over 397,000 transactions.

And on the other side of that pattern is also impressive: the pure percentage of low-participation or non-participative consumers. Based on the segmentation, just short of 31 percent of its customers are the so-called Lost Customers, those who have demonstrated very little to no purchasing behavior during recent times and represent roughly nothing to the bottom line of the retailer. All these customers are sent the same emails, the same promotions and the same generic reminders as the high-value consumers spending in the thousands before the process of segmentation. Not only is it inefficient but it is costly. Marketing departments are usually constrained with little budgets and sending out mails to thousands of inactive customers does not help to save time, money or internal energy. Just by understanding the number of customers at the low-value end of the spectrum, and being able to measure that segment, becomes very useful in reconsidering marketing resources should be directed.

The point where the algorithmic effect begins to manifest itself is the quality of the cluster comparisons. The research did not simply execute K-Means and leave it at that. It tried Gaussian Mixture Models, DBSCAN, BIRCH and Agglomerative Clustering to determine which of them provided the most meaningful and clear customer segmentation. The Gaussian Mixture Model had the best Silhouette Score of 0.80 and surpassed the other methods which had a saturation of 0.62-0.64. Having such a high Silhouette Score on a realistic retail data set is good since customer behaviour is inherently disordered. Individuals do not shop on regular basis, they buy various types of products and their expenditures differ significantly. But the mixture model broke through that noise and generated clean and well separated parts. This was assisted by the use of Principal Component Analysis prior to the process of clustering: once the RFM features had been reduced to the major components, the clusters became smaller and easier to be caught by the algorithm. It was exactly that in simple terms PCA reduced a three-dimensional complex problem to a simpler, more structured one, and the clustering algorithm exploited that.

It is not the statistic itself but that which it facilitates that is meaningful. When a successful model determines the boundaries of different customers who act differently, the retailer instantly has a guide on personalised marketing. An example of this is Take the Top Customers. The data indicates that there are some people with tens of thousands of pounds annually payment, with one of the customers spending more than £77,000 on purchases. It is not in the business sense to treat this customer the same way that a person who purchased one mug in 2011. Segmentation will enable the retailer to isolate this group and give them the type of personalised attention that will keep them off the turnover list among the highest-value customers. Any customer of this quality is much better to retain than to acquire new ones and the trends in RFM of this data support the significance of customer protection. Any slight rise in the retention rate by this segment will greatly increase profits.

Another type of opportunity is provided by the Medium-Value segment. Such customers are moderate spenders with reasonable frequency and are not consistent. They are the type of shoppers who tend to shop when they are reminded to do so, who tend to buy more when they are offered a bundle or a cross-selling or can be persuaded to purchase a product that has been personalised to them. These customers are identified by the results of clustering. They had been lost in the crowd before segmentation and they were averaged by the aggregate data. Once segmented, they transform to be a growth opportunity. When such a retailer buys even once more, the effect on thousands of customers may be incredible. Medium-Value customers usually do not need much investment to upgrade than the cost of converting Low-Value customers, which makes them an attractive target of desktop promotion.

Low-Value and Lost Customers, in their turn, cause a more cost-sensitive approach. The retailer cannot afford to spend a lot on the reactivation process that is statistically unfeasible to succeed because, with 61 percent of the data falling into these two combined groups, the likelihood of success is low. The response rates of untargeted marketing campaigns are usually extremely low, and according to the distribution of the RFM, this retailer is not an exception. In the case of Lost Customers, no more efficient method of reactivation would be a low-cost, automated process, rather than any kind of personalised effort. Put differently, segmentation can guard the instincts of the retailer against its instincts. In its absence, teams tend to pursue all their customers in the same manner. They understand with it the customers that they can reasonably invest in.

Inventory planning is another significant operational influence. Once you are able to connect segments with their spending patterns, you can forecast what the individual cluster will be purchasing. It is important since one of the most costly operational inefficiencies in the retail is poor inventory alignment. The overstocking is a waste of money in goods that might not move fast whereas stockouts make it hard to satisfy customers and lose them to other vendors. The retailer focuses on items related with Top and High-Value customers which makes the retailer more dominant in terms of revenue and customer satisfaction. Interest of Mid-Valued customers in mid-priced goods or seasonal products provides the retailer with the information as to when and how much to stock. Even a small change in the inventory forecasting will be converted into actual savings.

Segmentation is basically a monitoring system in the churn perspective. Customers whose recency scores start to decline and had high monetary scores previously are immediately visible in the RFM timeline. These are customers who are worth saving. And since high-value customers are both most cost-effective to retain and the most costly to lose, specific retention efforts are not only feasible but also cost-effective. Even avoiding churning of a small percentage of the high-value customers would generate more revenue than a wide-ranging acquisition campaign.

The segmentation also impacts internal planning. Marketing teams can tailor communication plans for each group. Customer service teams can focus their support on important segments. Supply chain teams can plan procurement more accurately. Leadership can set annual revenue targets based on realistic patterns instead of guesswork. This is the real benefit of segmentation; it turns raw data into a strategic guide.

The effect of this segmentation project goes well beyond simply sorting customers into categories. It changes how the retailer views its market, shows where true value is created, and gives clear guidance for future decisions. The clustering algorithms do more than analyze behavior; they reveal hidden structures that help the business direct its efforts where they count the most. In a competitive retail landscape, this shift from broad assumptions to strategy based on evidence is not just useful; it’s necessary.

### 6. Course Connection (QM 525 Linkages)

If we look at the entire project from QM 525's perspective for a second, it will be obvious that we performed a more extensive work than just data cleansing and clustering. Almost all major decisions, transformations, and algorithms we used are somehow connected to the mathematical concepts covered in the course. It is amazing how the use of eigenvalues, optimization, distance metrics, and covariance, which were the abstract concepts during the classes, turned out to be such practical applications in analyzing customer behavior in retail dataset. This project was a stepping stone from theory to practice, proving that the math behind machine learning is a very productive area rather than a purely academic one; it is the main factor on which the whole analysis is based.

One of the most obvious connections comes from the subject of linear algebra. Clustering techniques such as K Means and GMM are working in vector spaces and their entire process is dictated by the way we choose to measure and manipulate points in those areas. The moment we normalized the RFM features, we were actually changing the coordinate system so that no single dimension had a dominant effect on the distances. The concept of space reshaping to give each axis an equal contribution is straight from the linear algebra concepts of transformations and scaling. The distances that K Means uses are Euclidean norms, which is also something tied to vector operations. Even the centroids are simple vector means. All of this cannot be done without understanding the basic mathematical structure of how the data points are positioned and move in a multidimensional space.

Next comes PCA. If one would have to point out the one method that embodies QM 525, it would be Principal Component Analysis. PCA applies covariance matrices, eigenvectors and eigenvalues the very concepts we have been deriving during the course to pinpoint the spots with the highest variation in the data. In a nutshell, it rotates and compresses the feature space in such a way that the most important data is still there. Previously to PCA, RFM features were represented in a three-dimensional space that was not very clear for some algorithms. After the PCA operation was done and the data represented in principal components, the structure became much clearer. Previously blended clusters became separated. This is linear algebra taking care of the burden as it diagonalizes the matrix, extracts the principal eigenvector, and transforms the data into a coordinate system that aligns with its geometry.

Theoretical optimization pertains a lot to the algorithms' behavior. K Means clustering is an optimization problem by its nature. It aims at the reduction of the Within Cluster Sum of Squares, and this is done through the repetitive process of centroids' changing until the algorithm is in a steady state. The whole activity from an initial guess to escalating the parameters and watching it converge as it approaches zero is just what we went through when learning optimization methods. K Means, though, does not employ gradients in the classical calculus-influenced manner, however, the idea of iteratively shifting the solution to the minimum is the same principle. In contrast, the Gaussian Mixture Model adopts the Expectation Maximization algorithm which is a more formal optimization. It goes back and forth between the estimation of probabilities and the adjustment of model parameters to optimize data likelihood. That is classical optimization wherein you state a function, change parameters and get closer to a point of maximum or minimum.

Statistics, in fact, is the third main pillar that connects all this together. The whole RFM scoring system is grounded on statistical arguments, where recency acts as an engagement measure based on time, frequency as a count and monetary value as a continuous measure of spending. Machine learning hadn't happened yet but we were still using statistical reasoning in our methods. PCA is based on covariance which is a statistical measure of how two features change together. GMM is very much statistical because it treats data as a combination of Gaussian distributions with their own mean and covariance matrix. When we use the Silhouette Score to assess the quality of the clusters, we once more draw upon statistical comparisons between within-cluster cohesion and between-cluster separation. The entire process of cleaning data, selecting models and assessing them is very much dependent on statistics.

Moreover, a clearer picture comes out of the return of this work to QM 525 as just how deeply these mathematical tools are intertwined becomes evident. Linear algebra lays down the foundations of the spaces we work in. Optimisation directs the algorithms across these spaces. Statistics determines the way we perceive the relationships between the variables in those spaces. The project did not apply these concepts individually. They were more like one cohesive unit. The output of the algorithms in terms of customer groupings was not mere random clusters but mirrors of the hidden mathematical constructs like variance, density, distance and probability.

The course even influenced our model evaluation process. Rather than relying on the first clustering result, we compared multiple algorithms, scrutinized their Silhouette Scores, examined their characteristics, and selected the one that provided the most distinct separation. Such decision-making is a direct product of the course focus on validation, interpretation and mathematical justification. We were not merely after a result but rather trying to comprehend the underlying mathematical logic that supported that result.

QM 525 was very beneficial to us in terms of practical understanding of the algorithms interplaying with the real-world data. The retail data set comprised missing values, outliers, extremely skewed customers' spending habits and intricate patterns that were not discernible at first sight. The mathematical tools we got from the course made us quite sure that we could handle those difficulties. We, by normalizing the data, were actually using scaling techniques to make the variance steady. In the case of returns and negative quantities being removed, we had applied statistical reasoning to give a definition of what data is meaningful. In the dimensionality reduction part with PCA, we had used the knowledge of eigen decomposition to reduce the complexity without losing information. And similarly in the case of clustering the data, we had no other choice but to use optimization and probabilistic modeling to discover natural groupings.

In the end, the undertaking rendered the course considerably more concrete. The algorithms are not miracle workers. They base themselves on the same mathematics we studied: matrices, eigenvectors, distance functions, covariance, optimisation routines and probability distributions. Witnessing the emergence of those concepts in a real business scenario where they determine the effectiveness of marketing, supplier choices and customer retention, demonstrates that the mathematical foundation of QM 525 is very powerful and practical. We did not learn these concepts in isolation; we actually saw how they collaborate to tackle difficult problems in a systematic and rational way. And the link between theory and practice is undoubtedly the most significant benefit of this whole project.

7. Summary of Findings, Limitations, and Future Work

This project looked at how customers in the UK retail market behave by using RFM analysis and different clustering methods. The purpose was to group customers in a way that would help with marketing decisions, retention, and understanding overall buying behaviour. The dataset originally had over half a million transactions, but after removing errors like missing Customer IDs and wrong entries, we worked with around 398,000 clean records. Using RFM helped break down customer activity based on how recently they purchased, how often they buy, and how much they spend. From this, five types of customers appeared clearly: top customers, high-value, medium-value, low-value, and lost customers. Most customers fell into the low-value and lost groups, which shows that the retail business struggles with keeping customers active and loyal.

After the RFM part, several clustering algorithms were tested to see which one could form the best groups. The models used were K-Means, Gaussian Mixture Models, DBSCAN, BIRCH, and Agglomerative Clustering. Their performance was compared using the Silhouette Score. Out of all the methods, GMM gave the strongest results, especially after using PCA to simplify the data. It reached a Silhouette Score of 0.80, which was better than the other algorithms and also higher than what some previous studies reported. The reason GMM worked well is because it can handle different shapes and patterns in the data more flexibly. The other models still worked, but their scores were lower, around 0.64, and DBSCAN struggled more because it was sensitive to parameter settings.

Even though the project produced good insights, there were some limitations. RFM is useful, but it is still a simple method that doesn’t capture things like what products customers prefer or how their behaviour changes during certain seasons. It also does not show customer lifetime value or long-term trends. Some clustering models, like K-Means, were affected by how the data was scaled and how starting points were chosen. DBSCAN was difficult to tune because the dataset didn’t have clear density patterns. Another challenge is that working with large datasets can take a lot of time and computing power unless the data is reduced first.

For future improvements, the project can include more advanced clustering techniques, especially ones that handle complex patterns better, such as Spectral Clustering or improved density-based models. Adding more features to RFM, like product categories, purchase timelines, or lifetime value, could make the customer groups more meaningful. The project could also explore deep learning methods like autoencoders to automatically extract important patterns and reduce the amount of manual preprocessing needed.

8. Peer Feedback Documentation

when i was working on my part of the project, I shared my draft with the group so everyone could check if it matched the style and structure of the rest of the report. Most of the feedback I received focused on keeping my writing consistent with the other sections. A few teammates felt that some parts of my summary sounded a bit too formal or technical, so I rewrote those sections in simpler and clearer language. This helped the flow feel more natural and kept the tone in line with the rest of the report.

I was also advised to connect my explanation of the results more directly to the RFM analysis the group had already completed. post revision , I made sure the clustering results linked properly to the earlier customer segments, which made the section easier to follow and more meaningful.

One teammate pointed out a small error in the dataset numbers I used, which didn’t match the earlier parts of the report. I checked the case study again and corrected it to keep everything accurate and consistent.

therefore , the feedback helped me improve my writing, make the content clearer,

9. References

Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer. https://doi.org/10.1007/978-3-319-14142-8

Hariguna, T. (2024). Customer segmentation and targeted retail pricing in digital advertising using Gaussian Mixture Models for maximizing gross income. *International Journal of Data Science*, 12(1), 44–58.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>

John, J. M., Shobayo, O., & Ogunleye, B. (2023). An exploration of clustering algorithms for customer segmentation in the UK retail market. *Analytics, 2*(4), 809–823. <https://doi.org/10.3390/analytics2040042>

McKinsey & Company. (2021). *The value of getting personalization right or wrong is multiplying*. <https://www.mckinsey.com>

Ng, A. (2013). *K-Means clustering handout*. Stanford CS221 course notes.

ScienceDirect. (n.d.). *Agglomerative clustering – an overview.* Retrieved from <https://www.sciencedirect.com>

Song, P. (2025). *Principal Component Analysis examples*. Retrieved April 8, 2025.