A Clustering-Based Approach for Customer and Product Segmentation in Online Retail

LB
Master of Science in Data Science
University of Colorado - Boulder
Boulder, Colorado, USA

Abstract graphic meant to convey Data Science/Clustering concepts, AI-Generated.

Abstract

A key component of ensuring an online retail business thrives is its ability to move beyond high-level metrics and gain deeper insights into customer behavior and product relationships. Using a data-science-driven approach, it is possible to capture nuanced patterns from a foundational dataset of itemized transactions.

This project presents a methodology centered around a two-phase unsupervised learning framework for deriving business intelligence from raw transactional data. The first phase tackled product categorization through a hybrid methodology consisting of hierarchical clustering and NLP techniques, which revealed the primary revenue drivers of the business through categorical sales analysis. The insights gained on the purchasing habits of top customers were then used in the next phase.

The second phase focused on customer segmentation, which was successfully accomplished by expanding upon the standard RFM (Recency, Frequency, and Monetary) model to generate features for Gaussian Mixture Models. Despite lower scores on common clustering-centered metrics, the GMM still proved to be superior in comparison to a baseline K-Means model, as the former was able to identify complex structures within the data that directly represented distinct customer personas.

Both phases were successful in their goals of being able to generate actionable insights in a novel but replicable manner, demonstrating a clear process of transforming raw data into integral assets for business systems and processes enhanced with machine learning solutions.

CCS Concepts

• **Applied computing** → **Decision analysis**; • **Computing methodologies** → **Cluster analysis**; **Model verification and validation**.

Keywords

data mining, clustering, customer segmentation, product categorization, e-commerce, RFM analysis

ACM Reference Format:
LB. 2025. A Clustering-Based Approach for Customer and Product Segmentation in Online Retail. In . ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The digital era has led to the rise of online retail as a dominant force in the global economy. Through ubiquitous connectivity, businesses can provide products around the clock to customers in an

always-online environment, generating vast amounts of transactional data. Data mining aims to address the situation often described as "drowning in data but starving for knowledge." Due to the continuous generation of information, many retailers rely solely on high-level analytics, such as total sales or top customers by revenue; however, a deeper look at these massive datasets can unlock profound business insights beyond the surface-level view of traditional metrics.

The application of data science through data mining allows us to uncover complex underlying patterns, such as customer purchase behaviors and latent relationships between product offerings. Analyzing such patterns is a critical endeavor for any business that wants to thrive in a fast-paced and global e-commerce market. Customer segmentation enables targeted marketing campaigns, personalized recommendations, and proactive retention strategies, while product categorization allows for more efficient inventory management, improves cross-selling opportunities, and enables a clearer view of customer demand that can be obscured by a high volume of individual stock codes. Gaining knowledge of the underlying operations of a business and making decisions based on actionable insights is critical for achieving sustained profitability and higher customer retention rates.

For businesses that have successfully transitioned to online sales but lack the data infrastructure to generate meaningful insights, traditional methods are insufficient for handling the scale and complexity of e-commerce data. For this project's dataset (which only contains stock codes and item descriptions), manually categorizing thousands of individual SKUs would be a slow and painstaking process. Furthermore, a simple ranking of customer accounts by monetary value fails to explain why a customer is valuable. Ultimately, simple and manual methods are not only unscalable but also lacking in explanatory power. As such, there is a need for automated and sophisticated methodologies that can transform raw data into meaningful, actionable segmentations for both products and customers.

To fill this gap, this project implemented a two-phase clustering framework. First, hierarchical clustering was used to automatically create the base coherent categories for inventory items by leveraging intrinsic features within the dataset, and was further refined through a mix of NLP and rule-based association techniques. In the second phase, new features were engineered from raw transactional data by applying RFM (Recency, Frequency, and Monetary) concepts. These served as the dimensions for the clusters produced by Gaussian Mixture Models, which resulted in distinct customer personas suitable for segmentation. Both phases proved successful in utilizing different clustering techniques for different purposes, showcasing how businesses can leverage unsupervised learning techniques to uncover critical insights within their data, which could then be used to modernize processes or improve current operations through actionable insights.

2 Related Work

In the world of retail, product knowledge is an integral source of information that greatly dictates the day-to-day operations of a company. How other processes, such as sourcing, production, marketing, and financial planning, are executed can be heavily dependent on the kind of goods being sold. Just as having a solid grasp of this dynamic provides the foundational knowledge of internal operations, analyzing product portfolio performance in the context of consumer behavior and market trends uncovers key opportunities for strategic growth.

The concepts of frequent itemsets and association rules laid out by Agrawal et al. [2] provide a framework for beginning this kind of analysis. In their foundational paper, the framework was explained in full detail, from the concepts of support and rule-based generation of higher-order itemsets to suggestions on how to improve efficiency through pruning and an account of a successful application in a retail environment. Its usefulness, however, is contingent on the data containing either highly distinct items or well-defined categories for meaningful aggregation, both of which are lacking in the chosen dataset. In this e-commerce scenario, where product variety is in the thousands, the focus must first be on establishing these categories, since the categorical characteristics of a product segment and their subsequent relationships to other archetypes of goods carry more weight than the importance of individual itemsets. Establishing this will allow us to better understand the higher-level structure of the company's inventory, which is more informative to a wholesale business that caters to other companies.

Another vital aspect of a successful e-commerce endeavor is a company's ability to leverage customer segmentation strategies to improve both internal operations (managing churn) and external outreach (targeting new consumers). One established framework is the Recency, Frequency, and Monetary (RFM) model, which provides a clear picture of consumer behavior by evaluating transactional history. Part of the appeal of RFM is that it is intuitive and easily generalized. This, of course, also comes with the downside of being subjective, potentially leading to implementations that fail to capture the nuances of consumer habits. To improve upon these weaknesses, researchers have extended the RFM framework by integrating it with unsupervised machine learning algorithms. For instance, Wei et al. [3] demonstrated a case study where they modified the RFM framework to include Length (customer tenure) and applied a Self-Organizing Map algorithm. Through this approach, they generated distinct customer segments that proved valuable when devising marketing strategies. This project accomplished a similar goal, though with the added complexity of a diverse international market and thousands of SKUs. Therefore, it is an expansion on the framework proposed by Wei et al. [3] by engineering additional features to extend the RFM model further.

Ultimately, the approach taken aligned with the two primary goals of this project. Establishing item categorization and customer segmentation profiles for this unique e-commerce context meant that a higher level of specificity was needed in choosing the correct clustering methods. The diversity stemming from the products and the wide range of customers increased the variability within both data structures; as such, the inventory and consumer structures were dissimilar and operate in different dimensions. therefore, this project required more robust clustering models that are suitable for each data type.

3 Methodology

3.1 Dataset and Tools

The dataset for this data mining project is the "Online Retail Dataset," a public transactional dataset made available by Chen et al. [1] via the UCI Machine Learning Repository. It is a transactional dataset containing all business conducted by a UK-based online retailer between December 1, 2010, and December 9, 2011. The company sells unique, all-occasion gifts and primarily caters to wholesalers. The dataset contains over 540,000 rows, with each row representing an itemized line of a transaction. It contains eight columns: invoice number, stock code, merchandise description, quantity purchased, invoice date, unit price, customer ID, and the customer's country of residence.

This project was implemented in Python within a Jupyter Notebook environment. It leveraged core data science libraries such as pandas and Numpy for data preprocessing and EDA, nltk for NLP techniques, scikit-learn and Scipy for machine learning, statsmodels for statistical analysis, Plotly for visualizations, and Optuna for hyperparameter optimization.

3.2 Data Preprocessing and Management

3.2.1 Data Loading and Initial Exploration. The raw dataset was loaded, and its shape was checked, revealing that it contained 541,909 transactional records where each row represented a single line item. It also contained 8 columns: InvoiceNo, StockCode, Description, InvoiceDate, Country, UnitPrice, CustomerID, and Quantity. An initial feature, TotalPrice, was engineered by multiplying Quantity and UnitPrice to represent the value of each transaction. After checking for missing values and duplicates, the process moved past the initial exploration phase.

3.2.2 Handling Cancelled and Negative Orders. As mentioned in the UCI Machine Learning Database, InvoiceNo contains cancelled transactions, which are entries prefixed with a C. A total of 9,288 cancelled transactions were identified and filtered out to focus only on completed transactions. Preliminary statistical analysis using pandas also revealed the presence of negative values in the Quantity and UnitPrice columns. Further inspection confirmed these were related to non-sale administrative entries such as returns and bad debt adjustments. These records were also removed to isolate positive sales transactions.

3.2.3 Filtering Non-Product Administrative Codes. For StockCode to be a useful feature for clustering, the 5-digit numerical convention needed to hold true for the majority of items. Analysis confirmed that this was mostly the case, besides for some codes that were purely alphabetic (e.g., 'POST', 'AMAZONFEE', 'M' for Manual). These were identified as administrative charges (e.g., postage, bank fees) rather than actual products and were removed from the dataset.

3.2.4 Final Cleaned Dataset. The result of this comprehensive cleaning pipeline was a final, reliable dataset containing only completed, positive sales of actual products. This cleaned dataset formed the foundation for all subsequent analysis and modeling.

3.3 Exploratory Data Analysis and Preliminary Findings

3.3.1 Statistical and Distributional Analysis. After preprocessing, exploratory data analysis was performed to understand the characteristics and patterns of the cleaned data. Descriptive statistics and visualizations showed that numerical features such as Quantity, UnitPrice, and TotalPrice were heavily right-skewed and contained a large number of extreme outliers. This provided critical information for modeling, as many statistical models require normally distributed data. It indicated that there might be a need for data transformations. Additionally, to show the underlying distributions of the majority of the data, we would also need to filter out these outliers.

3.3.2 Geographic and Temporal Analysis. The month and hour were extracted from InvoiceDate to analyze sales patterns. The key insights showed a significant upward trend in monthly sales volume, peaking in November for the holiday season. For daily sales, most of the volume occurred during standard UK business hours, with peaks in the early afternoon. Both findings suggest a cyclical pattern consistent with most retail operations. For geographic analysis, a choropleth map was used to visualize the distribution of orders. As expected, most sales came from the United Kingdom, though analysis showed that a significant amount of volume also came from surrounding European countries. Consumers from Japan, Australia, and America were also present, highlighting the company's global reach.

3.3.3 Customer and Inventory Analysis. An analysis of the top 50 customers showed that they accounted for a significant amount of revenue, accounting for around 28% of total sales, displaying the Pareto Principle in action. Visualizing their purchase behaviors through stacked bar charts revealed that while most buy a wide variety of items, a few focus on purchasing only a handful of SKUs, or sometimes even a single product.

3.4 Phase 1: Product Categorization

The main goal of this phase was to employ machine learning and data mining techniques for the purpose of quickly establishing an inventory structure to address the lack of categorization for products. As seen in the customer purchases stacked bar chart, the thousands of StockCode made it nearly impossible to gain any sort of practical information from the chart, except for the observation that some customers only buy a few items. If categorized data were available, an analysis that is focused on aggregates would provide better insight. This was the initial motivation for the first phase, which was accomplished through a hybrid methodology discussed below.

3.4.1 Feature Engineering from Stock Codes. To leverage the hypothesized latent structure within the alphanumeric system of labeling items, a novel feature named StockCodeFloat was engineered. The need for this feature occurred after validating the lengths of all the entries under the StockCode column, which revealed that some SKUs contained one or two alphabetical suffixes after the 5-digit code. To use these codes in a clustering algorithm, these alphanumeric characters had to be translated to numerical values

in a manner that kept their relative placement among other 5-digit numerical inventory items in their vicinity. A mapping system that converts suffixes into unique decimal values was created, which preserved the inherent numerical ordering and allowed for the use of clustering algorithms.

3.4.2 Algorithmic Clustering. The StockCodeFloat and UnitPrice columns were plotted on a scatter plot to confirm the hypothesis of latent structure. This revealed distinct, vertically oriented, and markedly separated groups of items. Due to the shapes of these clusters, Hierarchical Agglomerative Clustering was chosen as the main algorithm. The ability of this algorithm to group points based on proximity without the constraints of assuming spherical clusters made it ideal for this situation. The fcluster method produced an initial set of 21 distinct product clusters that served as the blueprint for grouping other items.

3.4.3 Iterative, Rule-Based Refinement. While the initial clustering provided a solid base structure for the rest of the categories, a large series of StockCodes were densely packed in a very small range that appeared to be used for general labeling, lacking a defining collective characteristic. Critically, items in this range accounted for 75% of the total revenue, so a solution was needed to incorporate these into more well-defined fcluster categories based on their item descriptions.

To reallocate most of these items, a rule-based, NLP-assisted recategorization process was developed. A helper function was created that transformed the description column into lowercase, tokenized the text, and extracted the nouns. Counting the frequency of each noun within the group revealed the presence of underlying themes that could not be captured by the initial clustering.

A priority-based system was used for final reassignment: first, they were to be allocated towards the base fcluster categories, followed by the creation of a new functional cluster if needed, and lastly, they were assigned to a general stylistic category if the aesthetic itself was the most defining and frequent characteristic. With each iteration, only the most thematically coherent groups were extracted and reallocated until all that was left from this generic cluster was a set of truly miscellaneous items.

3.5 Phase 2: Customer Segmentation

3.5.1 Feature Engineering Through RFM. Characterizing customer populations was a prerequisite for the main goal of phase 2, Customer Segmentation. To generate a customer-centric view from raw transactional data, RFM-style metrics such as num_orders, days_since_last, and total_purchases were added as features alongside another metric (total_unique_stock) that tracked purchasing variety. A subsequent feature called customer_activity_score was engineered through a combination of scaled frequency (num_orders) and recency (days_since_last) calculations to capture a more holistic view of individual customer engagement.

3.5.2 Data Transformation and Statistical Validation. Initial EDA on the new features revealed extreme outliers. To ensure the clustering algorithm focused on the primary population distribution, these outliers were filtered out. Afterwards, statistical tests were performed on the engineered features to determine their suitability for Gaussian Mixture Model Clustering. Bartlett's test confirmed that the features had unequal variances,and the Henze-Zirkler test for multivariate normality confirmed that the distribution was not normal. To address this, QuantileTransformer was applied, which reshaped the skewed data to Gaussian-like distributions. Immediately prior to modeling, a likelihood-ratio test was conducted to determine if there was statistical evidence supporting a multi-group model over a single-group model. A high Likelihood Ratio and p-value < 0.0001 provided strong statistical justification that the dataset's structure would be better captured by a multi-group model.

3.5.3 Modeling and Comparative Setup. As previously referenced, the Gaussian Mixture Model (GMM) was chosen as the primary model for customer segmentation due to its flexibility in identifying clusters of various sizes, shapes, orientations, and for modeling soft, overlapping cluster boundaries. To ensure a meaningful segmentation outcome, the Optuna library was used to optimize the hyperparameters by searching for the values that minimized the Bayesian Information Criterion. To ensure reproducibility, the results of the best trial, along with the starting means, were saved and directly used to initialize the main GMM model. For a comparative analysis, a baseline K-Means clustering model was also trained on the same data and was initialized with the same number of clusters.

4 Evaluation and Results

Each phase of this project was assessed through both quantitative and qualitative approaches in order to ensure a multifaceted approach. Additionally, the outcomes were also contextually evaluated with regard to the specific problem each phase aimed to address, primarily in its effectiveness and interpretability as a solution to finding actionable business insights.

4.1 Phase 1: Product Categorization Results

Figure 1 shows the latent structure inherent within the StockCodeFloat numbering system along with the segmentation produced by fcluster. The dense green cluster represents the generic category that had to be further reallocated.

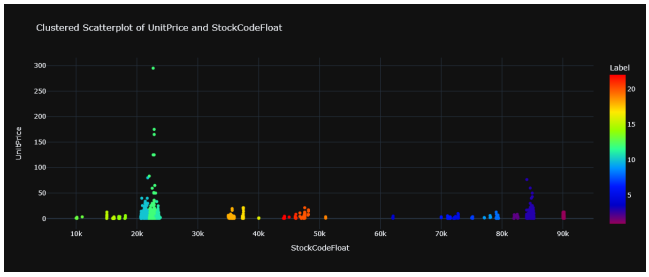

Figure 1: Segmentation results of the initial grouping attempt through clustering. The hues represent all of the different categories automatically segmented by fcluster.

4.2 Insight into Top Customer Behavior

As seen in Figure 2, the aggregated version of the stacked bar chart of the top customer purchases better depicts purchasing behavior in terms of item variety. The stacked bars in the first visual were

too dense to be useful, while the banded segments in the updated visual clearly reveal the types of products and their value. It also reveals that sales cannot just be attributed to a few kinds of items. Instead, we see a healthy distribution of the main categories, which suggests that the company is proficient at identifying and fulfilling the needs of a wide range of consumers.

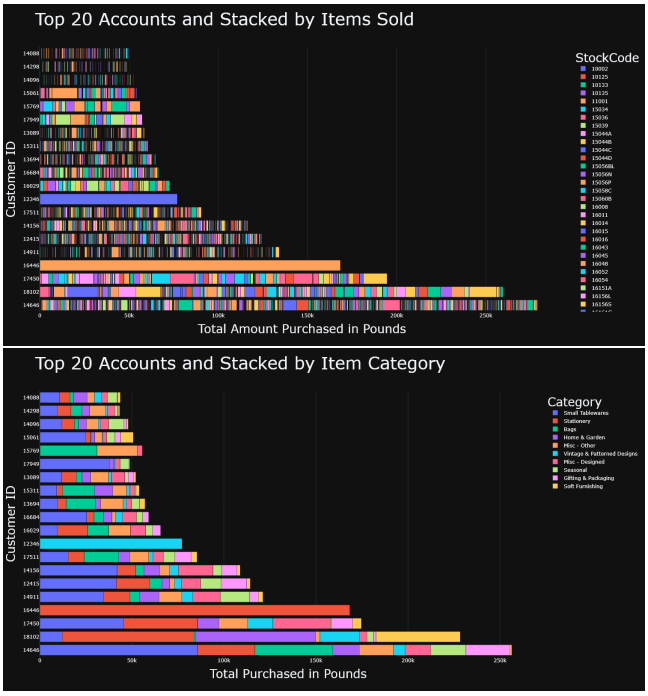

Figure 2: Breakdown of sales by inventory for the top 20 customer accounts before (top) and after (bottom) the iterative refinement process. The dense stacks on the top provide little insight into purchasing behavior, while the refined version allows for better actionable insights.

4.3 Final Output of Hybrid NLP and Rule-Based Allocation

The two pie charts in Figure 3 definitively show the impact of the rule-based, multi-pass process in further segmenting the main generic cluster. The first pie chart shows how items were initially distributed based on the fcluster output, where a large majority of unrelated products were in the same category. The second chart represents the final output after the application of NLP techniques and iterative reallocation. This visual provides a much more actionable set of information by having well-defined segments and tangible percentage numbers.

4.4 Phase 2: Customer Segmentation Results

Table 1: Comparison of Clustering Algorithm Performance Metrics.

Metric	K-Means	GMM
Silhouette Score (Higher is Better)	**0.261**	0.020
Davies-Bouldin Score (Lower is Better)	**1.053**	1.902

4.4.1 Quantitative Evaluation. : Quantitatively, the results of the Gaussian Mixture Model were compared against a baseline K-Means model output using two standard cluster validation metrics, the Silhouette Score and Davies-Bouldin Score, both of which measure intracluster cohesion and intercluster separation. As seen in Table 1, taking the raw scores at face value implies that the K-Means clustering result is quantitatively better, but further evaluation reveals that the scores themselves do not tell us the whole story.

4.4.2 Visual and Qualitative Evaluation. : Analyzing the clusters made by GMM (top) and K-Means (bottom) visually in Figure 4 shows why GMM is the better model for customer segmentation than K-means. The first thing to note is the high degree of uniformity present among the K-Means clusters. The equal and homogeneous clusters created by K-means are the exact opposite of what is needed for customer segmentation; if population segments are mostly indistinguishable from one another, then this inhibits us from identifying the different behavioral purchasing habits that characterize one group from another.

On the other hand, the variety displayed by the clusters GMM produced in terms of size, shape, density, and direction allow us to characterize the main differences between each segment just from looking at the graphs. For example, the top right graph represents the fringe consumer groups; the placement of the clusters dictates the general purchasing behavior, specifically the direction of where their specific deficiencies are. For the core consumer clusters in the top-left quadrant, we see a clear positive linear trend, suggesting that most of their business comes from consumers who score well on multiple metrics.

4.4.3 Final Customer Personas. : While visualizing the clusters provides evidence for the existence of distinct population segments, it is only by analyzing the descriptive statistical summaries that we can turn our data into actionable insights. Table 2 and Table 3 show the averages of each cluster on the features that were engineered for this phase, along with the mean of the whole consumer population as context.

The core customer base can be summarized by the following personas:

- **Cluster 5 (Core Active Customers):** The largest and most consistent segment, scoring higher than the population average on all metrics and purchasing a wide variety of products.
- **Cluster 1 (High-Value VIPs):** A smaller but highly significant segment that drives revenue through high-value orders of specific products.

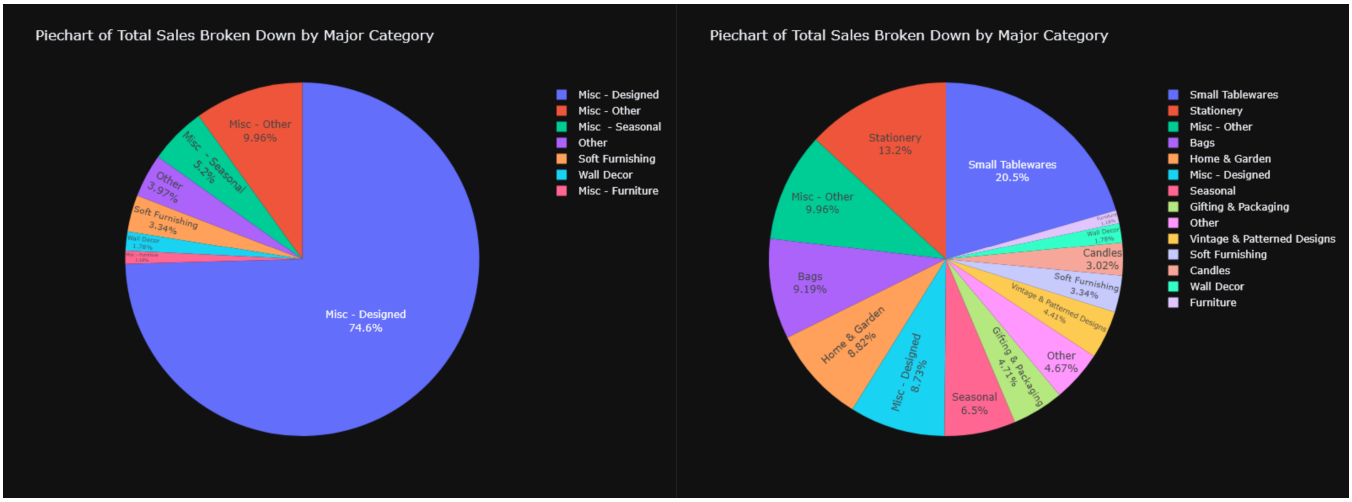

Figure 3: Comparison of sales distribution by category. Left: The initial result from algorithmic clustering, dominated by a single generic category (74.6%). Right: The final result after iterative, rule-based refinement, revealing a much clearer and more actionable distribution of the true product pillars.

Table 2: Statistical Profiles of the Core Customer Segments, compared against the overall population average.

Metric	Overall Population	Cluster 0 (At-Risk)	Cluster 1 (VIPs)	Cluster 5 (Core Active)
Count	4328	1045	337	2263
Mean Orders	4.10	2.70	5.93	5.32
Mean Days Since Last	92.35	101.17	66.18	73.74
Mean Unique Stock	59.93	44.57	14.61	89.17
Mean Total Purchase (£)	1839.06	901.45	4937.31	2247.97
Median Purchase (£)	665.98	689.57	928.06	1006.98

Table 3: Statistical Profiles of the Fringe Customer Segments, compared against the overall population average.

Metric	Overall Population	Cluster 2 (Single Product)	Cluster 3 (Churned)	Cluster 4 (Niche)
Count	4328	94	14	575
Mean Orders	4.10	1.37	1.00	1.25
Mean Days Since Last	92.35	188.54	373.00	142.36
Mean Unique Stock	59.93	1.00	29.14	9.73
Mean Total Purchase (£)	1839.06	1622.74	267.92	191.50
Median Purchase (£)	665.98	170.10	260.31	169.55

- **Cluster 0 (At-Risk Customers):** A large segment with lower-than-average metrics, representing a key opportunity for re-engagement and churn prevention.

Clusters 0, 1, and 5 in Table 2 can be described as the Core Customer Segments that account for most of the revenue generated by the business as an online retailer. Among these, Cluster 5 is the only group that scored higher on all of the metrics relative to the mean of the entire population while also being the largest. This group as a whole purchases a wide variety of products in a consistent manner, acting as a constant source of income. Cluster 1 is considerably smaller in terms of population but is just as financially significant due to higher transaction totals with an order frequency on par with that of Cluster 5. Doing so while purchasing only a specific set of products means that this population tends to only purchase either higher-priced goods or order popular items in much larger quantities than most consumers. Ensuring that the company can maintain the business of these two segments is critical if they wish to be successful long-term. Identifying consumers that might be at a high risk of churn, such as the consumers that make up Cluster 0, in order to create renewed business opportunities, can be just as

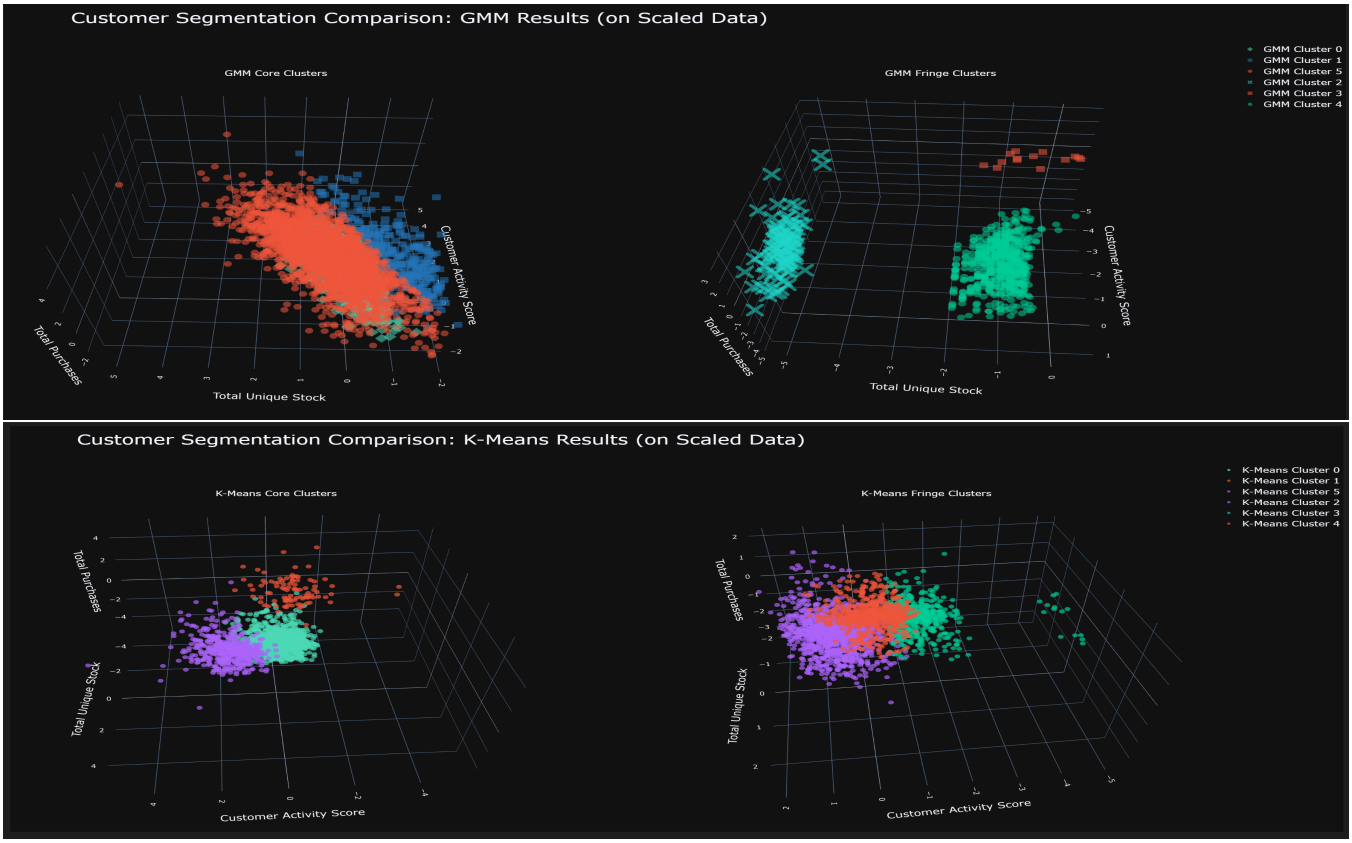

Figure 4: Visualization of the clusters created by GMM and K-Means. Top: Gaussian Mixture Model Outputs. The left graphic represents the Core Clusters, and the right contains the Fringe Clusters. Bottom: K-Means output given the same number of means found by GMM. All clusters are of about the same size and placement.

important, especially in online retail, as the landscape can change rapidly. Understanding what factors contribute to the lower metrics in this cluster compared to the other core segments can not only bring business back but also help establish new ones as well.

The fringe segments represent customers with more extreme or niche purchasing behaviors:

- **Cluster 3 (Churned Customers):** The smallest segment, characterized by the longest time since their last order.
- **Cluster 2 (Single Product Wholesalers):** A segment defined by high-value, bulk purchases of a single unique product, indicated by a large discrepancy between its mean and median purchase value.
- **Cluster 4 (Niche Consumers):** The largest fringe group, representing a smaller, but well-rounded, version of the core customer base.

Clusters 2,3, and 4 in Table 3 represent the Fringe Customer Segments that represent various extremes of purchasing behavior. The smallest segment, Cluster 3, most likely represents churned customers, considering the average date of last order is significantly higher than all other clusters. Clusters 2 and 4 might be comprised of non-commercial transactions since the business did generate direct-to-consumer sales, even if most of their buyers are mainly commercial wholesalers.

Still, there are interesting things to note about the statistics derived from the GMM clustering, such as the parallels between Clusters 2 and 4 to Clusters 1 and 5, respectively. Cluster 4 has similar characteristics to Cluster 5 in the context of other Fringe Clusters: it has the biggest population among fringe clusters and is well-rounded with regard to consumer metrics when compared to Clusters 2 and 3. Cluster 2 is similar to Cluster 1 in terms of purchasing habits, which are centered around specific products. For Cluster 2, it was a much narrower scope of a single type of item, purchased mainly in bulk. This seems to be a minor purchase in most cases, though the difference between the mean and median total purchase amounts signals the presence of possible outliers. The labels were traced back from the CustomerID to the itemized Dataframe and identified the most extreme transaction responsible for the difference. The transaction in question was one massive order of over 74,000 units of ceramic jars for a grand total of £77,000. While this was undoubtedly the highest sale total for transactions in Cluster 2, it is also a one-time purchase of a single product type, which is what characterizes the cluster as a whole.

While the fringe segments (Clusters 2, 3, and 4) do not generate as much in product demand as the core segments, analyzing their unique behaviors can still present significant opportunities for growth and operational improvement. Each cluster raises strategic questions for the business. For the "Single Product Wholesalers" (Cluster 2), the key question is how to replicate their high-value, bulk-order behavior with other products or customers? For the "Niche Consumers" (Cluster 4), if these are indeed non-commercial accounts, how can the business better understand their purchasing habits in order to capture more direct-to-consumer sales? For the "Churned Customers" (Cluster 3), what factors link their purchasing history to their inactivity? The ability to propose these questions, which is the direct result of the clustering results, directly encapsulates the value of actionable insights that result from the data mining process.

This scenario has shown that even a preliminary discussion of summary-level statistics, derived from an industry-specific implementation of unsupervised learning, can lead to many possible paths for progress and expansion. This phase has also shown that while a deep analysis on use-case and practicality of implementation is much more involved than a simple quantitative numerical check, it also provides us a greater understanding of why one model might be better suited than another, even if the math does not agree on the surface.

5 Discussion

5.1 Timeline

This project was completed within the proposed four-week schedule, successfully accomplishing all planned weekly tasks, including the full implementation of both product categorization and customer segmentation, as well as all remaining work for the final report and presentation slides.

Week 1 Focused on data understanding, completed exploratory data analysis, comprehensive data cleaning, and other preprocessing needs.

Week 2 Dedicated to implementing hierarchical clustering, created NLP-extracted category labels, engineered features for customer segmentation, and prepared check-in presentation materials.

Week 3 Dedicated to customer segmentation, including feature normalization, statistical modeling tests, and GMM implementation.

Week 4 Focused on final model evaluation, result interpretation, discussion, and the creation of the final report and presentation.

5.2 Lessons Learned

The first major challenge was the incomplete interpretability of the initial clusters provided by Hierarchical Clustering on the 'StockCode' scatterplot. While the initial group of categories provided a good base, a significant amount of inventory items remained in a dense cluster that was too large and generic to be useful. This was overcome by developing a rule-based, NLP-assisted hybrid technique for further reallocation. The key lesson was that a purely algorithmic approach is not always sufficient, but in such cases, knowing how to creatively apply other powerful data science techniques and processes can lead to finding the right solutions.

The other major challenge was understanding how to approach the fact that statistical tests suggested that the baseline K-Means theoretically produced better results than the GMM model, when an initial visual inspection of the clusters immediately showed a much greater range of variety and separation among the GMM output clusters. This suggested that the GMM had a greater chance of characterizing each segment in a truly distinctive manner, and instead of simply forgoing the GMM models, I made sure to see the process through first while relying on my past retail management experiences and understanding of the metrics that were engineered for this phase. This challenge provided the opportunity for me to experientially understand what it means to not only implement the right kind of model for the specific task at hand, but also identify instances where integrating different sources of knowledge can be greatly beneficial to the entire process of data mining.

6 Conclusion

In today's digital era, success in e-commerce requires a proactive approach. The speed at which trends change and the demands of the market mean that a business must not only keep its operations up-to-date but also understand how to retain existing customers and target new ones. For established businesses with vast amounts of data but potentially outdated information structures, the best way to do so is by extracting actionable insights from existing data.

This project demonstrated how a data-driven framework can modernize business analysis processes. The most significant findings are:

- **A Hybrid "Human in the Loop" Methodology:** A process combining the automation of hierarchical clustering with the nuance of human-in-the-loop, rule-based refinement proved highly effective. This hybrid approach transformed raw stock codes into clearly defined categories, enabling a clear analysis of the business's true revenue drivers.

- **Enhanced Segmentation through a Holistic Process:** The customer segmentation phase showed how a comprehensive process of thoughtful feature engineering, data transformation, and correct model selection enhances the interpretability of an RFM-style analysis. The resulting GMM successfully identified distinct customer personas characterized by distinct purchasing habits. Understanding these habits is a critical factor for success in retail operations as it not only informs businesses of their own demographic but also helps direct inquiries that stimulate internal progress.

- **The Primacy of Qualitative Evaluation:** The multifaceted evaluation process proved that quantitative metrics alone can be misleading. For this dataset, visual evidence and qualitative analysis were critical in validating the GMM as the superior model over a K-Means baseline, highlighting the necessity of aligning evaluation techniques with the specific problem and data geometry.

Ultimately, this project provides a replicable framework for extracting value from transactional data, where practical insights give way to deeper questions and concrete steps towards business growth.

7 Future Work

Generalizing the Hybrid Framework The methodology used for product categorization—leveraging machine learning for the bulk of the work and incorporating domain knowledge for fine-tuning—serves as a general framework. This approach could be adapted to modernize other business information systems, creating fast, organized, and use-case-specific processes that outperform manual efforts.

Enhanced Customer Segmentation The current RFM-style features could be extended by engineering more advanced metrics. Using higher granularities or composite measures could capture specific business-relevant information. Additionally, more advanced clustering algorithms could be explored to represent the segments in different ways.

Dynamic Segmentation and Monitoring This research could be extended into a dynamic process. By storing the model, segmentation results, and customer metrics in a database that is periodically updated, the business could identify and analyze how customers migrate between segments over time. This would provide powerful insights for sales and marketing strategies.

Sub-clustering for Targeted Campaigns For core customer clusters that are high in density and population, further sub-clustering could be performed with different segmentation outcomes in mind and with a different set of preprocessing techniques. An example would be using different sets of contextual features (e.g., specific products purchased) in order to provide a higher-level of specificity needed for things like targeted ad campaigns or consumer outreach services.

References

[1] Daqing Chen, Sai Liang Sain, and Kun Guo. 2012. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management* 19, 3 (Sept. 2012), 197–208. https://doi.org/10.1057/dbm.2012.17

[2] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD '93)*. ACM, New York, NY, USA, 207–216. https://doi.org/10.1145/170035.170072

[3] Jo-Ting Wei, Shih-Yen Lin, Chih-Chien Weng, and Hsin-Hung Wu. 2012. A case study of applying LRFM model in market segmentation of a children's dental clinic. *Expert Systems with Applications* 39, 5 (April 2012), 5529–5533. https://doi.org/10.1016/j.eswa.2011.11.066