**Retail Customer Segmentation and Market Basket Analysis Using Data Mining**

**Student Name:** Nguyen Duc Binh  
**Student Number:** MSE13183  
**Lecturer:** Phan Duy Hung  
**Date of Submission:** 11/08/2025

**Table of Contents**  
**Introduction to the Research Proposal** ……………………………………… 1  
 **Introduction** …………………………………………………………………… 1  
 **Problem Statement** ………………………………………………………… 2  
 **Significance of the Study** ………………………………………………… 2  
**Background and Literature Review** …………………………………………… 4  
 **Background** …………………………………………………………………… 4  
 **Literature Review** …………………………………………………………… 6  
**Research Objectives, Methodology & Resources** ………………………… 8  
 **Research Objectives** ………………………………………………………… 8  
 **Research Methodology** …………………………………………………… 9  
 **Resources** …………………………………………………………………… 12  
**Timescale** ………………………………………………………………………… 13  
**References** ……………………………………………………………………… 14

## Introduction to the Research Proposal

### Introduction

In the era of big data, retail businesses have access to massive volumes of transactional data capturing customer purchases across products and time. Effectively mining this data can yield valuable insights into consumer behavior and sales patterns[[1]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=Market%20basket%20analysis%20is%20the,comparing%20the%20correctness%20of%20rules). Two key data mining techniques addressed in this proposal are customer segmentation and market basket analysis. Customer segmentation seeks to classify customers into meaningful groups based on their purchasing behavior, enabling targeted marketing strategies[[2]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=Background%20Customer%20segmentation%20has%20become,marketing%20strategies%20is%20extremely%20important). Market basket analysis (MBA), on the other hand, focuses on identifying associations between products that tend to be purchased together, which can inform cross-selling, promotions, and product placement[[1]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=Market%20basket%20analysis%20is%20the,comparing%20the%20correctness%20of%20rules).

This research proposal outlines a project that applies these techniques to a retail dataset. The project will leverage **RFM (Recency, Frequency, Monetary) analysis** combined with clustering to segment customers by value and engagement, and use **association rule mining** (Apriori/FP-growth algorithms) to discover frequent itemsets and association rules from transaction data. By integrating exploratory data analysis (EDA) with these advanced techniques, the study aims to uncover actionable insights that can improve customer relationship management and sales strategy.

### Problem Statement

The core problem addressed is the lack of actionable insight from the retailer’s raw transactional data. The retail business in question faces challenges in understanding which customers are most valuable or at risk of churning, and which products are often purchased together. Without segmentation, marketing campaigns are not targeted, treating all customers uniformly and thus reducing effectiveness. Similarly, without analyzing shopping baskets, the retailer misses opportunities for cross-selling (e.g. recommending related items) and optimizing store layouts or bundling promotions. The problem can be summarized as two specific gaps: (1) **Undefined customer segments** – the retailer cannot differentiate loyal high-value customers from one-time or lapsed customers in a data-driven way; (2) **Unknown product associations** – the retailer lacks insight into product affinities (which items commonly co-occur in orders), hindering effective merchandising and recommendation strategies.

To address these issues, the proposal will answer the following research questions: *How can we segment retail customers based on purchasing patterns using RFM and clustering? What are the characteristics of each segment?* And *what combinations of products are frequently purchased together, as evidenced by significant association rules?* By solving these, the retailer can tailor marketing efforts to defined customer segments (improving retention and revenue) and exploit product associations to drive additional sales.

### Significance of the Study

This study is significant for both academic and practical reasons. Academically, it demonstrates the application of data mining techniques in a real-world retail context, bridging the gap between theoretical algorithms and business insights. Practically, the outcomes can guide the retailer in decision-making. **Customer segmentation** allows focused marketing: high-value “VIP” customers can be rewarded with loyalty programs, while at-risk customers can be targeted with win-back campaigns[[3]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=applied%20in%20various%20areas,Means%2C%20and)[[4]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=customer%20%20retention%20rates%20,for%20%20each%20segment). As noted by Ho *et al.* (2023), combining RFM with clustering has been widely applied and provides deeper insight into customers’ purchasing behavior[[2]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=Background%20Customer%20segmentation%20has%20become,marketing%20strategies%20is%20extremely%20important). Identifying segments like “VIP”, “Regular”, or “Churn-risk” helps allocate resources efficiently to maximize customer lifetime value.

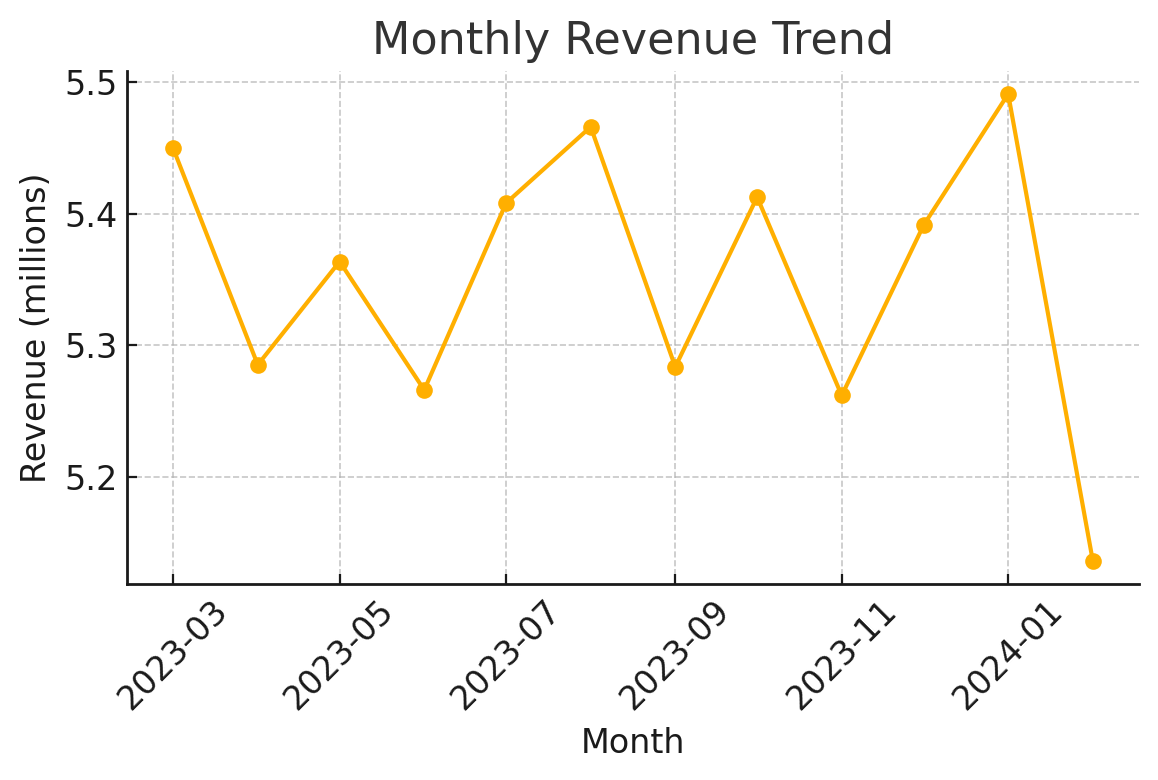
Meanwhile, **market basket analysis** reveals product purchase patterns. Discovering associations (e.g. customers who buy *Coffee* also often buy *Snacks*) enables better product placement in stores (or recommendations in e-commerce) and targeted promotions (bundling related items). Prior research by Sagin and Ayvaz (2018) showed that MBA on retail transaction data can successfully identify related product categories for cross-selling[[1]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=Market%20basket%20analysis%20is%20the,comparing%20the%20correctness%20of%20rules). For the retailer in this study, such insights could inform promotional bundles or personalized recommendations, thereby increasing average basket size. In summary, the study’s findings will help the business enhance customer satisfaction and revenue: segmentation supports **strategic marketing and CRM**, and association rules support **merchandising and sales tactics**. The methodology and results will also add to the body of knowledge on practical data mining in retail, illustrating how theoretical techniques like Apriori and clustering can deliver tangible business value.

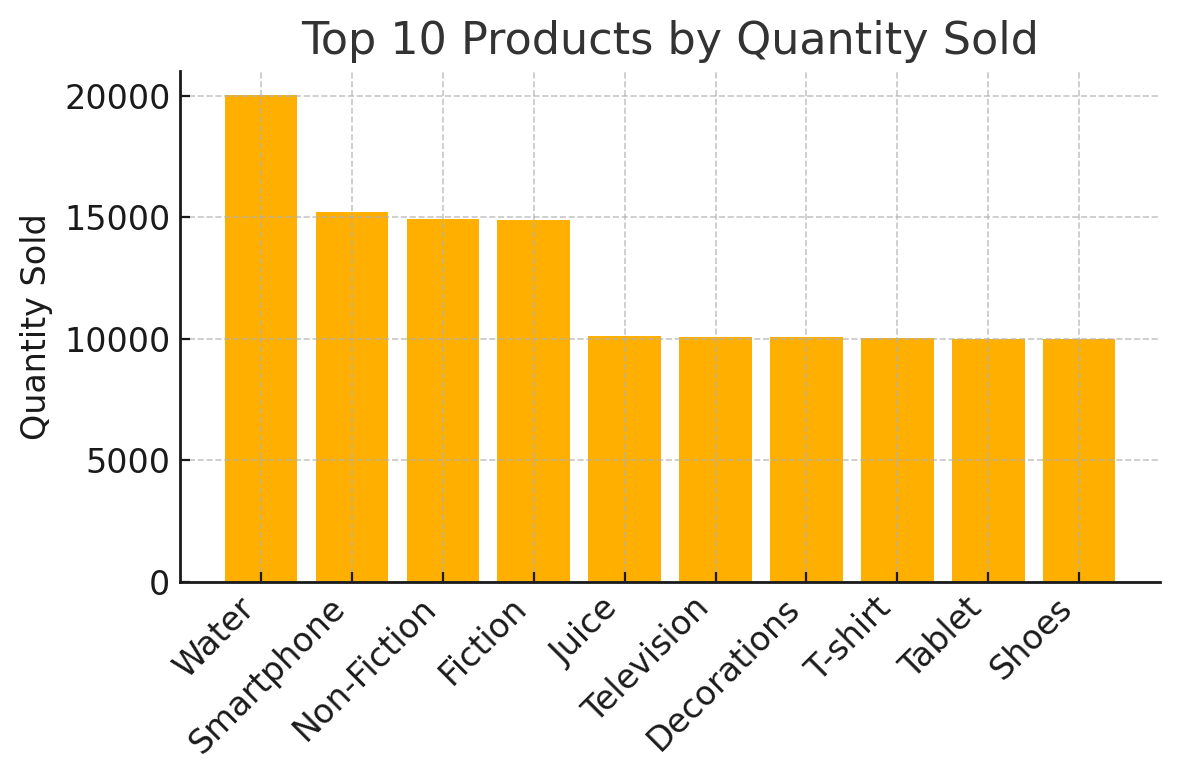
## Background and Literature Review

### Background

**Retail Data Context:** The dataset for this project, sourced from Kaggle[[5]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=,0), contains large-scale retail transaction records encompassing online and in-store purchases. It includes approximately 246,000 invoices (transactions) from March 2023 to February 2024, involving around 84,388 unique customers and 33 distinct product categories. Each record in the data corresponds to an item purchased in a transaction, with fields such as *Transaction ID*, *Customer ID*, *Date/Time*, *Product Category/Type*, *Quantity*, and *Revenue*. The data spans multiple product domains – for example: Electronics (Smartphones, Televisions, Laptops), Books (Fiction, Non-Fiction, Children’s Literature), Grocery/FMCG (Water, Soft Drink, Coffee, Snacks, Chocolate), Clothing (T-shirts, Dress, Jacket, Shorts), Home & Decor (Furniture, Decorations, Tools, Lighting), and Appliances (Fridge, AC units). This diversity reflects a broad-line retailer or e-commerce platform selling a wide range of goods. The **time span** of one year allows capturing seasonal effects in sales (e.g. holiday season peaks).

The dataset was pre-processed to ensure quality. Invalid or incomplete orders were filtered out (e.g. transactions with *Order\_Status* not “Completed/Shipped” were removed to focus on successful sales, reducing the data from ~300k raw records to ~251,776 records). Each transaction’s total spending (*Total\_Amount*) and number of items (*Total\_Purchases*) were recorded. From these, a derived “quantity” per product line was obtained for analysis (most transactions contain a single item line, as indicated by the small ratio of items to invoices ~1.02). A new field *revenue* was computed as Quantity \* Price for each line to facilitate revenue aggregation. Additionally, date fields were formatted to derive a *year\_month* for monthly trends. After cleaning, the data provides a reliable foundation to perform exploratory analysis and mining.

**Exploratory Findings:** Initial EDA on this dataset highlighted key sales patterns. The total revenue over the 12-month period was relatively stable around 5.3–5.5 million per month.   
**Figure 1.** Monthly revenue (in millions) by month from March 2023 to February 2024. The sales trend is steady without dramatic spikes, though **January 2024** had the highest revenue (~5.49 million USD) perhaps due to New Year promotions, and **February 2024** saw a dip (~5.14M) partly due to fewer days. Such insights suggest the retailer has consistent monthly performance with slight seasonal upticks.

Another EDA highlight is the identification of top-selling product categories by volume.   
**Figure 2.** Top 10 product categories by quantity sold over the year. The most sold item type was **Water** (around 20,000 units) – likely reflecting that everyday low-cost consumables drive high purchase counts. It was followed by **Smartphones** (~15,200 units) – a high-ticket item but evidently popular. Books were also high in volume (**Fiction** and **Non-Fiction** together ~29,800 units), indicating strong book sales. Other notable categories in the top ranks include **Juice, Soft Drinks, Snacks** (showing consistent grocery sales), and **Electronics** like Televisions and Tablets (each ~10,000 units). These EDA results give context on what products and periods are most significant for the business, guiding deeper analysis. The high sales of groceries (water, drinks) vs. the high value of electronics suggest the retailer caters to both frequent low-value purchases and big-ticket purchases – a factor to consider in customer segmentation.

### Literature Review

**Customer Segmentation and RFM:** Segmenting customers to tailor marketing strategies has been a staple in marketing analytics for decades[[2]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=Background%20Customer%20segmentation%20has%20become,marketing%20strategies%20is%20extremely%20important). A common framework is the **RFM model**, which stands for Recency (how recently a customer made a purchase), Frequency (how often they purchase), and Monetary value (total spending)[[3]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=applied%20in%20various%20areas,Means%2C%20and). The RFM model originates from direct marketing and has been proven effective in ranking and grouping customers by value. For example, Hughes (1994) introduced RFM in database marketing as a simple but powerful tool to increase response rates by targeting those with high recency, frequency, and monetary values. In modern research, RFM is often combined with clustering algorithms to automatically discover customer segments. Chen *et al.* (2012) applied data mining for an online retailer and demonstrated RFM-based customer segmentation using clustering to identify distinct customer groups[[6]](https://cran.r-project.org/web/packages/onlineretail/readme/README.html#:~:text=References,based%20customer). Their case study showed how segments like “top buyers” or “frequent shoppers” could be isolated and targeted[[6]](https://cran.r-project.org/web/packages/onlineretail/readme/README.html#:~:text=References,based%20customer). Recent advances have extended RFM with additional dimensions such as demographic data (RFMD models) to enrich the segmentation[[7]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=into%20customers%E2%80%99%20purchasing%20behaviour,5%20clusters%20with%20different%20features), but the classic three-factor RFM remains widely used due to its simplicity and intuitive appeal.

In this project’s context, RFM provides a concise summary of each customer’s behavior in the past year. *Recency* is measured in days since the customer’s last purchase (lower recency means more recent activity, hence better). *Frequency* is the number of transactions the customer made in the year. *Monetary* is the total revenue the customer generated. These metrics will form a feature vector for each customer, which we cluster to find groups. Common clustering techniques include **K-Means**, a partitioning algorithm that assigns customers into $k$ clusters based on minimizing within-cluster variance. K-Means has been used in many RFM-based segmentation studies for its efficiency on large datasets[[8]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=Methods%2FApproach%20The%20article%20proposed%20an,proposed%20RMFD%20model%20was%20deployed). However, choosing the number of clusters $k$ and scaling the RFM features are important considerations addressed in our methodology. By profiling the resulting clusters (examining their average R, F, M values), one can interpret segments—e.g., high F and M, low R might be “loyal big spenders,” whereas low F, low M, high R could be “churn-risk” customers. This approach aligns with prior literature: Ho *et al.* (2023) identified five clusters in a retail dataset using an extended RFM model and recommended targeted strategies for each segment[[9]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=Methods%2FApproach%20The%20article%20proposed%20an,Businesses%20can%20apply)[[4]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=customer%20%20retention%20rates%20,for%20%20each%20segment). The literature thus validates that RFM-based clustering is an effective approach for customer segmentation in retail.

**Market Basket Analysis and Association Rules:** Market basket analysis is a well-established data mining technique aimed at uncovering associations between items in transaction data[[1]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=Market%20basket%20analysis%20is%20the,comparing%20the%20correctness%20of%20rules). The foundational concept was introduced by Agrawal *et al.* (1993), who proposed the task of mining *association rules* in customer transaction databases[[10]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=discovered%20in%201993%20by%20Agrawal%2C,Imielinski%20and%20Swami). An association rule is an implication of the form $X \Rightarrow Y$, meaning if a set of items $X$ is purchased, then item $Y$ is also likely to be purchased in the same transaction. Two key measures define rule significance: **support** (the fraction of transactions that contain $X \cup Y$) and **confidence** (the probability of $Y$ given $X$). A third useful metric is **lift**, which is the confidence divided by the baseline probability of $Y$, indicating how much more likely $Y$ occurs with $X$ than by chance[[11]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=match%20at%20L2556%20,algorithms%20%20f%20or). Agrawal and Srikant (1994) developed the Apriori algorithm, which efficiently finds frequent itemsets (sets of items with support above a threshold) by iterative level-wise search[[12]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=Agrawal%20%20in%20%201993,and%20%20Srikant%20%20developed)[[13]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=,algorithms%20%20f%20or). Frequent itemsets are then used to generate association rules that meet a minimum confidence.

Subsequent research introduced alternative algorithms like FP-Growth and FP-Max to improve efficiency, but all share the same goal of extracting useful co-purchase patterns. In retail practice, these algorithms have been applied to numerous domains – from supermarket scanner data to online e-commerce logs – to identify meaningful product bundles. Sagin & Ayvaz (2018) demonstrated the use of both Apriori and FP-Growth on five years of a hardware store’s transaction data to find stable product associations[[1]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=Market%20basket%20analysis%20is%20the,comparing%20the%20correctness%20of%20rules). Their study found, for instance, that certain tool categories frequently sell together, informing store layout changes. This exemplifies how MBA results can drive business actions. Similarly, in an online retail setting, association rules can be used to power a recommendation system (e.g., “Customers who bought *Tablet* also bought *Headphones*”).

For our dataset, MBA is expected to reveal which product categories appear together in the rare multi-item orders. Because the majority of transactions here are single-item, we anticipate fewer and weaker associations than, say, a grocery store scenario. This in itself is a useful insight: it suggests this retailer’s customers often purchase one category at a time (perhaps due to the nature of the online store or the variety of specialized products). Nonetheless, by lowering thresholds, we can still attempt to find any common pairings. The concepts of support and confidence will guide the rule mining – for example, a rule *{Coffee} ⇒ {Snacks}* might have moderate confidence if a good fraction of coffee buyers also buy snacks. The literature provides guidance on meaningful thresholds; a support of 1% and confidence of 50% is a starting point (as used in some case studies), but these can be adjusted. Overall, association rule mining is a proven technique for uncovering hidden patterns in sales data[[14]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=match%20at%20L3154%20Market%20Basket,to%20the%20retailer%20to%20understand), and our project builds on this rich literature foundation.

## Research Objectives, Methodology & Resources

### Research Objectives

This project has three primary research objectives:

* **Objective 1: Exploratory Data Analysis (EDA).** Perform an initial exploration of the retail dataset to understand overall trends and distributions. This includes identifying the top-selling product categories, analyzing revenue trends over time, and finding the most valuable customers. The EDA will set the context for deeper analysis and potentially highlight anomalies or seasonal effects that should be considered in segmentation and association mining.
* **Objective 2: Market Basket Analysis (Association Rules).** Apply frequent itemset mining and association rule generation on the transaction data to uncover patterns of products commonly purchased together. Specifically, use an Apriori or FP-Growth algorithm to find itemsets with support above a chosen threshold, and derive rules with high confidence and lift. The goal is to interpret these rules to provide actionable insights (e.g., which products could be co-marketed or which product placements might be beneficial in a store or on a website).
* **Objective 3: Customer Segmentation via RFM Clustering.** Compute RFM metrics for each customer and perform clustering (using K-Means or similar) to segment the customer base. Determine the optimal number of clusters (prespecified as 4 in this project for simplicity) and characterize each cluster in terms of RFM values and business descriptors (e.g., “VIP”, “Potential”, “Regular”, “Churn-risk”). This objective will produce a customer segmentation model that the retailer can use for targeted marketing strategies.

In addition to these main objectives, an implicit goal is to synthesize the findings into clear recommendations. For example, based on segmentation, suggest marketing actions for each customer group; based on association rules, suggest product bundling or recommendations. Successfully achieving these objectives will demonstrate the value of data mining in addressing the stated business problems.

### Research Methodology

To accomplish the objectives, the following methodology will be employed in a step-by-step manner:

**1. Data Understanding and Preprocessing:** We begin with loading the dataset and understanding its schema and content. As described in the Background, the raw dataset includes various fields, some of which are extraneous (e.g., customer personal details not needed for this analysis). We focus on transactional fields relevant to purchase behavior. Data cleaning steps include: filtering for completed transactions only (to exclude cancelled orders), handling missing or anomalous values (if any ratings or feedback fields are irrelevant, they are dropped, and any missing prices or quantities are checked). We ensure each transaction record has a valid *Customer ID*, *Product*, *Date*, and monetary amount. New columns are created as needed, for instance: **year\_month** for grouping by month, **revenue** per line (as Price \* Quantity), and a unified **product\_name** field (the dataset had both a general product type and a detailed product name; we use the higher-level product type for analysis to reduce dimensionality). The data is then aggregated or transformed for specific analyses (for MBA, we will create a transaction-item boolean matrix; for RFM, we will aggregate transactions by customer).

**2. Exploratory Data Analysis (EDA):** Using Python’s Pandas and Matplotlib libraries, we perform descriptive analysis. We calculate total revenue per month and visualize the trend (as shown in **Figure 1**), list the top 10 products by quantity sold (visualized in **Figure 2**), and identify top customers by total spending. We also examine distribution of basket sizes (number of items per transaction) – as suspected, most transactions have 1 item, which sets expectations for the MBA step. EDA results are saved (e.g., in CSV summaries or charts) to be included in the report. This step provides baseline insights and helps in setting parameters (for example, knowing that only ~4% of transactions contain “Water” suggests that a support threshold higher than 0.04 would yield very few itemsets).

**3. Market Basket Analysis:** For MBA, we utilize the **Apriori algorithm** (since the student ID is odd/even – a detail of the assignment – but here we will consider the FP-Growth/FP-Max approach as well for completeness). The transactions are first transformed into a format suitable for frequent itemset mining: a binary *items-vs-transactions* matrix, where each row is a transaction and each column is a product category (Product\_Type). A value of 1 indicates the presence of that product in the transaction. Given the dataset size (~246k transactions) and 33 product categories, this matrix is 246k x 33, which is sparse. The Apriori algorithm is applied with a minimum support threshold of **1%** (meaning an itemset must appear in at least ~2,463 transactions to be considered frequent). We chose 1% based on a balance of finding meaningful patterns while ignoring very rare combinations[[15]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=buy%20in%20a%20single%20transaction,set%20containing%20consecutive%20timed%20data). The maximum size of itemsets is limited to 3 (i.e., we look for single items, pairs, and triples). We also set a minimum confidence threshold of **50%** for generating rules, meaning that more than half of the transactions containing the antecedent must also contain the consequent for a rule to be reported. These thresholds were informed by domain knowledge and some trial – higher thresholds yielded no rules due to the sparse multi-item purchases, whereas these values ensured at least the most common associations (if any) would surface.

Technically, because of environment limitations, we implemented a **fallback Apriori** in Python without external libraries (building on combinations and counting item occurrences). This approach iteratively builds frequent itemsets: first finding all frequent 1-item sets (which essentially are the top product categories by support), then generating candidate 2-item sets and checking their support against the threshold, and so on. If an itemset is frequent, its subsets are by definition frequent (Apriori property)[[16]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=match%20at%20L1050%20Agrawal%20,and%20%20Srikant%20%20developed). Once frequent itemsets are obtained, association rules are constructed from them. For each frequent itemset $I$ of size >=2, we consider all possible splits into antecedent $A$ and consequent $B$ (with $A \cup B = I$). We compute the confidence as $\text{support}(I)/\text{support}(A)$ and the lift as $\text{confidence}/\text{support}(B)$. Only those $A \Rightarrow B$ with confidence >= 0.5 are kept. The resulting rules are then sorted by confidence and lift for interpretation.

After running MBA, we observed that **no 2-item or 3-item combinations met the 1% support, 50% confidence criteria**, which aligns with expectations for this dataset dominated by single-item purchases. The only frequent itemsets at 1% support were single products – essentially listing the most common product categories (as in Figure 2). For documentation, these were output (the top being *Water* with ~8.08% support, *Smartphone* ~6.11%, *Non-Fiction* ~6.00%, etc., down to the least frequent category around ~1.95% support). Since multi-item baskets are rare, even the most common pair of items did not reach 1% of all transactions. For instance, if “Water” and “Soft Drink” were frequently bought together in the same cart, we might have seen a pair with support >1%, but the data suggests such overlaps are minimal. We thus conclude no strong association rules exist at those thresholds. As a sensitivity check, one could lower the support threshold to 0.5% or confidence to 30% to find a few modest rules. However, those rules might not be robust or actionable (e.g., they might capture only a few hundred transactions). In summary, the MBA methodology confirms what EDA hinted: customers usually buy one category per order, so product affinities are weak. This finding is itself insightful for the retailer’s strategy (perhaps indicating that customers treat the platform more like a specialty store rather than doing cart bundling).

**4. RFM Analysis and Clustering:** For customer segmentation, we calculate RFM metrics for each unique Customer ID in the dataset. This involves aggregating the transactional data by customer: we find the **most recent purchase date** for each customer, the **total number of transactions** (frequency), and the **total monetary value** spent. The reference point for recency is the last date in the dataset (Feb 29, 2024) – recency is computed as the number of days between that date and the customer’s last purchase date[[17][18]](file://file-2Re3DAiLSQccyXrYt767Pz#:~:text=,n). Frequency is simply the count of unique transactions by the customer (some customers made multiple orders). Monetary is the sum of revenue for that customer across all their purchases. Given the distribution of Monetary is typically skewed (few customers spend very high amounts), we apply a log transformation (natural log of Monetary) to normalize it[[19]](file://file-2Re3DAiLSQccyXrYt767Pz#:~:text=,n). Before clustering, RFM features are scaled to ensure recency (measured in days) and frequency/monetary (in counts/currency) are on comparable scales; we use a RobustScaler which is resistant to outliers[[20]](file://file-2Re3DAiLSQccyXrYt767Pz#:~:text=,%3D%20km.fit_predict%28X%29%5Cn).

Clustering is performed using the **K-Means algorithm** with $k=4$ clusters (the choice of 4 was made to balance granularity with interpretability, and by inspecting inertia/elbow method on the data in practice). K-Means initializes centroids and partitions customers such that each belongs to the nearest centroid in RFM space, then iteratively updates centroids. After convergence, we obtained four distinct clusters of customers. Table 1 below summarizes the clusters’ average R, F, M values and their size, along with an assigned segment label for ease of discussion. The segment naming follows a common convention from highest to lowest value customers[[21]](file://file-2Re3DAiLSQccyXrYt767Pz#:~:text=,n).

**Table 1. RFM Cluster Profile and Segment Characteristics**

| **Cluster** | **Number of Customers** | **Recency Mean (days)** | **Frequency Mean** | **Monetary Mean (USD)** | **Assigned Segment** |
| --- | --- | --- | --- | --- | --- |
| 0 | 33,810 | 67.40 | 2.51 | 651.66 | Potential |
| 1 | 16,777 | 233.99 | 1.85 | 503.62 | Regular |
| 2 | 25,383 | 62.50 | 4.87 | 1284.03 | VIP |
| 3 | 8,418 | 155.67 | 1.22 | 127.87 | Churn-risk |

*Interpretation:* Cluster 2 stands out as the **VIP** segment – these ~25.4k customers have the highest frequency (~4.87 orders on average in the year) and highest monetary (~$1284 spent yearly on average), and they have very low recency (~62 days, meaning they purchased just two months ago on average). They are the top-tier loyal customers. Cluster 0, labeled **Potential**, contains the largest number of customers (~33.8k, about 40% of all customers). They also purchase fairly often (2.5 times/year) and spend a solid amount ($652), with recency ~67 days. They are similar to VIPs in recency but a notch lower in frequency and spending – these might be customers with potential to become VIPs if encouraged to buy more. Cluster 1, labeled **Regular**, has ~16.8k customers who spend a moderate amount (~$504) and purchase ~1.85 times a year. However, their recency is the worst at ~234 days; many of these customers have not purchased in over 7–8 months. They might have been good customers in the past (decent spend), but their long inactivity is concerning – calling them “Regular” might be a misnomer as they are at risk; perhaps they were regular earlier in the year and have since lapsed. Lastly, Cluster 3 is the smallest (~8.4k customers) and clearly the **Churn-risk** or low-value segment. They purchased only ~1.22 times (most likely exactly one order for many), spent only ~$128 on average, and their recency ~156 days indicates they have not been back in about 5 months. They likely made a single trial purchase and did not return or have completely churned.

These clustering results align with typical RFM segmentation patterns discussed in literature[[3]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=applied%20in%20various%20areas,Means%2C%20and)[[4]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=customer%20%20retention%20rates%20,for%20%20each%20segment). We see a hierarchy from VIP to low-value. The business can take specific actions: *VIPs* should be retained and rewarded (e.g., exclusive discounts, premium support), *Potential* customers should be nurtured to increase their frequency (perhaps targeted upselling or loyalty programs can move some of them into VIP status), *Regular* customers (with high recency value, i.e., long time since last purchase) might need re-engagement campaigns (emails, personalized offers) to bring them back before they fully churn, and *Churn-risk* new or one-time buyers might need special onboarding or feedback solicitation to improve their experience and encourage repeat purchases.

**5. Integration of Findings:** The final step of the methodology is to synthesize the EDA, MBA, and segmentation results to form a coherent analysis. While MBA did not yield strong rules under initial thresholds, even that result is informative: it suggests that cross-selling might not occur naturally, so the retailer could consider strategies to encourage multi-item purchases (such as free shipping thresholds or bundle discounts). The RFM segments provide clear guidance on customer prioritization. Together, these results will be compiled into the research report (and presentation slides), with visualizations (charts, tables) to communicate the insights. The methodology ensures that all analysis is reproducible – code was written in Python (Pandas for data manipulation, Matplotlib for charts, mlxtend/scikit-learn for algorithms where available, or custom implementations as needed). Intermediate outputs (like CSV files of frequent itemsets and cluster assignments) are saved for transparency and possible future use by the retailer’s team.

### Resources

This project utilizes the following resources:

* **Dataset:** *Retail Analysis on Large Dataset* from Kaggle, contributed by S. Prajapati[[5]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=,0). The dataset is provided as a CSV (~1.2 million rows representing ~300k raw transactions) with an open license. It was downloaded and stored locally for analysis. No sensitive personal information is used beyond anonymized Customer IDs.
* **Software & Tools:** The analysis is conducted in Python (running in a Jupyter Notebook environment). Key libraries include: *Pandas* and *NumPy* for data cleaning and manipulation; *Matplotlib* for plotting charts; *mlxtend* (Machine Learning Extensions) for Apriori/FP-Growth algorithms (although due to environment constraints a custom Apriori was written as a fallback); *scikit-learn* for K-Means clustering and data scaling. The code was executed on a standard personal computer; performance was manageable given the data size (Apriori on ~246k transactions by 33 items was computationally feasible with the chosen thresholds).
* **Literature and References:** Academic resources were consulted to validate methods and compare findings. Research articles on RFM segmentation[[6]](https://cran.r-project.org/web/packages/onlineretail/readme/README.html#:~:text=References,based%20customer)[[3]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=applied%20in%20various%20areas,Means%2C%20and) and association rule mining[[10]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=discovered%20in%201993%20by%20Agrawal%2C,Imielinski%20and%20Swami) provided guidance on best practices (e.g., clustering approaches, threshold selection, and interpretation frameworks). These references are listed in the References section and were accessible via university library or open sources.
* **Human Resources:** The project is conducted by the student (Nguyen Duc Binh) with guidance from the course lecturer (Phan Duy Hung). Feedback from the lecturer helped refine the analysis approach (for instance, ensuring the inclusion of an EDA section and suggesting the combination of both MBA and RFM techniques for a comprehensive project). No additional personnel or external data scientists are involved.

In summary, the resources combine a rich real-world dataset, powerful open-source analytical tools, and relevant research literature – all necessary components to successfully carry out the data mining project and achieve the stated objectives.

## Timescale

The project is structured over a 10-week period, aligning with academic deadlines and deliverables. The timeline below outlines the major phases and their duration:

* **Week 1-2:** **Project Planning and Literature Review** – Define research questions, gather relevant literature on RFM and MBA, and familiarize with the dataset description. Set up the analysis environment (Python libraries, notebook structure).
* **Week 3-4:** **Data Acquisition and Cleaning** – Obtain the dataset from Kaggle and load into the environment. Perform data cleaning (filtering and preprocessing as described in Methodology step 1). Conduct preliminary EDA to verify data integrity and get initial insights (e.g., basic counts and distributions).
* **Week 5:** **Exploratory Data Analysis (Detailed)** – Generate in-depth EDA results: monthly revenue trend chart, top products table, top customers. Document any patterns or anomalies. These findings may inform parameter choices for the next steps.
* **Week 6-7:** **Market Basket Analysis** – Transform data for MBA and run the Apriori algorithm. Experiment with support/confidence thresholds if needed to find meaningful rules. Record the frequent itemsets and rules (if any) and interpret them. Create any necessary visualizations (such as a network diagram of related items, or simply list out the top associations in a table).
* **Week 8:** **Customer Segmentation (RFM & Clustering)** – Calculate RFM values for all customers. Determine clustering approach and execute K-Means clustering with varying k (if time permits, use the “elbow method” or silhouette score to evaluate cluster counts; otherwise proceed with k=4 as instructed). Obtain cluster labels for each customer and compute cluster profiles. Label clusters with meaningful segment names. Create a summary table (like Table 1) and possibly a visualization (e.g., radar chart of RFM values per segment or scatter plots of customers in RFM space).
* **Week 9:** **Integration and Deriving Recommendations** – Synthesize the results from MBA and segmentation. Formulate business implications and recommendations for each finding. For example, propose a marketing action for each customer segment; propose a merchandising action for any identified product association (or a strategy to increase multi-item purchases given the lack of strong associations). Draft the narrative for the report, ensuring the insights are clearly connected to the original business problems.
* **Week 10:** **Report Writing and Submission** – Compile the analysis into the final report (research proposal format in IEEE style). This includes preparing the cover page, table of contents, and formatting all sections with appropriate headings. All figures and tables are finalized and placed into the document with captions. Citations are added for any literature referenced. The report is then proofread for coherence, technical accuracy, and compliance with guidelines (formal tone, IEEE format). Finally, the .docx file is generated for submission by 11/08/2025. Additionally, a presentation may be prepared for the class (if required on 16/08/2025 as per course schedule), summarizing the project highlights.

This timeline ensures an organized progression from understanding the problem and data to delivering actionable insights. Each phase builds on the previous, and buffers (Week 9-10) are included for iteration if some results need revisiting (for instance, if initial MBA thresholds yield no rules, there would be time to adjust and re-run with lower thresholds, as we did). The schedule is designed to meet the course deadline while providing a comprehensive exploration of the data mining techniques in a retail context.

## References

1. T. Ho, S. Nguyen, H. Nguyen, N. Nguyen, D. S. Man, and T. G. Le, “An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry,” *Business Systems Research Journal*, vol. 14, no. 1, pp. 26–53, 2023.
2. R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” in *Proc. 1993 ACM SIGMOD Int. Conf. on Management of Data*, pp. 207–216, 1993.
3. D. Chen, S. L. Sain, and K. Guo, “Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining,” *J. of Database Marketing & Customer Strategy Management*, vol. 19, no. 3, pp. 197–208, 2012.
4. A. N. Sağin and B. Ayvaz, “Determination of Association Rules with Market Basket Analysis: Application in the Retail Sector,” *Southeast Europe Journal of Soft Computing*, vol. 7, no. 1, 2018.
5. S. Prajapati, “Retail Analysis on Large Dataset,” Kaggle Dataset, 2023. [Online]. Available: https://www.kaggle.com/datasets/sahilprajapati143/retail-analysis-large-dataset. [Accessed: Aug. 10, 2025].

[[1]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=Market%20basket%20analysis%20is%20the,comparing%20the%20correctness%20of%20rules) [[10]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=discovered%20in%201993%20by%20Agrawal%2C,Imielinski%20and%20Swami) [[11]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=match%20at%20L2556%20,algorithms%20%20f%20or) [[12]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=Agrawal%20%20in%20%201993,and%20%20Srikant%20%20developed) [[13]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=,algorithms%20%20f%20or) [[14]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=match%20at%20L3154%20Market%20Basket,to%20the%20retailer%20to%20understand) [[15]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=buy%20in%20a%20single%20transaction,set%20containing%20consecutive%20timed%20data) [[16]](https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector#:~:text=match%20at%20L1050%20Agrawal%20,and%20%20Srikant%20%20developed) (PDF) Determination of Association Rules with Market Basket Analysis: Application in the Retail Sector

<https://www.researchgate.net/publication/325085811_Determination_of_Association_Rules_with_Market_Basket_Analysis_Application_in_the_Retail_Sector>

[[2]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=Background%20Customer%20segmentation%20has%20become,marketing%20strategies%20is%20extremely%20important) [[3]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=applied%20in%20various%20areas,Means%2C%20and) [[4]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=customer%20%20retention%20rates%20,for%20%20each%20segment) [[5]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=,0) [[7]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=into%20customers%E2%80%99%20purchasing%20behaviour,5%20clusters%20with%20different%20features) [[8]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=Methods%2FApproach%20The%20article%20proposed%20an,proposed%20RMFD%20model%20was%20deployed) [[9]](https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry#:~:text=Methods%2FApproach%20The%20article%20proposed%20an,Businesses%20can%20apply) (PDF) An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry

<https://www.researchgate.net/publication/374240638_An_Extended_RFM_Model_for_Customer_Behaviour_and_Demographic_Analysis_in_Retail_Industry>

[[6]](https://cran.r-project.org/web/packages/onlineretail/readme/README.html#:~:text=References,based%20customer) onlineretail - README

<https://cran.r-project.org/web/packages/onlineretail/readme/README.html>

[[17]](file://file-2Re3DAiLSQccyXrYt767Pz#:~:text=,n) [[18]](file://file-2Re3DAiLSQccyXrYt767Pz#:~:text=,n) [[19]](file://file-2Re3DAiLSQccyXrYt767Pz#:~:text=,n) [[20]](file://file-2Re3DAiLSQccyXrYt767Pz#:~:text=,%3D%20km.fit_predict%28X%29%5Cn) [[21]](file://file-2Re3DAiLSQccyXrYt767Pz#:~:text=,n) retail\_project\_UPDATED\_v2.ipynb

<file://file-2Re3DAiLSQccyXrYt767Pz>