data mining project

Freddy Erazo

February 2026

1 Introduction

1.1 Name and Source

In this project, we use the *Retail Transaction Dataset*, a synthetic grocery retail dataset published on Kaggle for market basket analysis and retail pattern discovery. The dataset is publicly available as a CSV file on the Kaggle

1.2 Size

The dataset contains approximately 30,000 unique retail transactions, with each row corresponding to a single shopping basket in a simulated grocery store environment. It is stored as a single CSV file with 4 columns and has a total file size of about 2.37 MB. The four columns are a transaction identifier, a customer identifier, a comma-separated list of products, and the transaction date.

1.3 Data Description

Each record in the dataset represents one purchase event at a simulated grocery store, capturing what a customer bought during a single trip. The data was generated with realistic product combinations and purchase patterns, making it suitable for association rule mining, recommendation tasks, and general market basket analysis. Products include common grocery categories such as beverages, snacks, dairy, household items, fruits, vegetables, and frozen foods.

1.4 Key Features and Attributes

We plan to use the following key attributes from the dataset:

- **Transaction ID**: A unique identifier for each transaction, used to define individual baskets.
- **Customer ID**: An anonymous identifier that links multiple transactions made by the same customer, allowing customer-level aggregation and profiling.

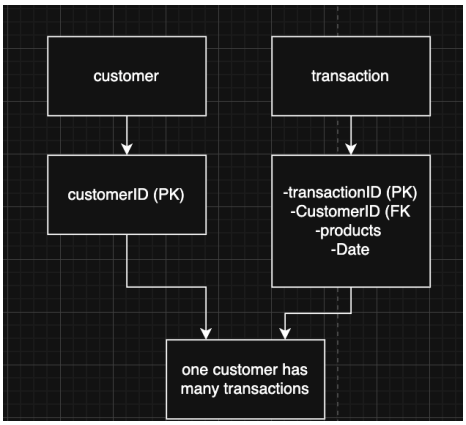

Figure 1: Enter Caption

- **Products**: A comma-separated list of all products purchased in the transaction, which will be transformed into itemsets for association rule mining and other basket-level analyses.
- **Date**: The date on which the transaction took place, which can be used to study temporal patterns such as changes in shopping behavior over time.

1.5 Data Quality Considerations

The dataset is entirely synthetic and does not contain real user information, so there are no privacy-related missing values or redactions. However, because the data is generated, there may be artifacts such as extremely rare products, unusually large baskets, or highly regular temporal patterns that differ from real-world noise. In addition, product names are stored as free-text strings, so minor inconsistencies in naming or spelling may require standardization during preprocessing.

1.6 Illustrative Schema

Figure **??** shows a simple logical schema of the dataset, where each customer can have many transactions, and each transaction contains a list of products. A sample data table (Table **??**) illustrates typical values for the main attributes.

2 Discovery Questions

2.1 Research Questions

In this project, we focus on discovery-oriented questions about shopping behavior in the Retail Transaction Dataset, rather than prediction tasks. We propose

the following research questions:

1. **RQ1:** Which groups of products are most frequently purchased together in this grocery retail dataset, and what are their associated support, confidence, and lift values?

2. **RQ2:** Are there natural segments of customers based on their purchasing behavior (for example, typical basket size and product category preferences)?

3. **RQ3:** Which transactions appear anomalous relative to typical baskets, and what characteristics (such as basket size or unusual product combinations) distinguish these anomalies?

2.2 Motivation and Value

RQ1 is valuable because identifying frequently co-purchased product sets is the core goal of market basket analysis and directly supports decisions such as store layout, product placement, and cross-selling strategies. Retailers can use such patterns to design product bundles and targeted promotions that encourage customers to add complementary items to their baskets. Understanding measures such as support, confidence, and lift also helps prioritize the most meaningful associations rather than trivial co-occurrences.

RQ2 is important because clustering customers by their purchasing behavior reveals hidden segments that are not defined by any explicit labels. For example, the analysis may uncover groups such as "large weekly shoppers," "frequent small-basket shoppers," or "category-focused shoppers" (e.g., mostly snacks or mostly fresh produce). These segments provide actionable insights for personalized marketing, loyalty programs, and inventory planning without requiring any prior assumptions about customer types.

RQ3 is interesting because anomalous transactions can highlight unusual shopping behavior or potential data issues that would be missed by simple summary statistics. Very large baskets, highly atypical product combinations, or irregular patterns over time may correspond to special events, bulk purchases, or errors in the data. Studying these anomalies can improve data quality and also suggest edge cases that are important for understanding real-world retail operations.

2.3 Discovery, Not Prediction

All three questions are framed as pattern discovery problems rather than prediction problems. We are not trying to predict a future label (such as whether a customer will churn or what rating they will give a product); instead, we aim to uncover structure that is already present in the transaction data. Association rules, customer clusters, and anomalies are all unsupervised patterns that emerge from the data itself and align with the emphasis on discovery in data mining.

3 Planned Techniques

In this project, we will apply data mining techniques from three categories: association rule mining, clustering, and anomaly detection. These techniques directly address the discovery questions defined in the previous section and are all unsupervised or pattern-discovery oriented rather than prediction-focused.

3.1 Association Rule Mining (Apriori / FP-Growth)

To answer **RQ1** (frequently co-purchased products and their support, confidence, and lift), we will apply association rule mining to the transaction-level baskets. We plan to use the Apriori or FP-Growth algorithm to discover frequent itemsets and generate association rules of the form $X \Rightarrow Y$, where X and Y are sets of products that tend to appear together in the same basket [**?**, **?**]. For each rule, we will compute standard interestingness measures such as support, confidence, and lift to prioritize patterns that are both frequent and non-trivial in the retail context [**?**, **?**]. These rules will reveal meaningful product combinations (e.g., bread and milk, or chips and soda) and provide a direct, interpretable answer to RQ1.

3.2 Customer Clustering (K-Means / Hierarchical)

To address **RQ2** (discovering natural segments of customers), we will construct customer-level feature vectors summarizing purchasing behavior and then apply clustering algorithms. Example features include average basket size, total number of transactions, and relative frequency of purchases across major product categories, analogous to behavioral segmentation features such as frequency, monetary value, and recency used in customer analytics [**?**, **?**]. We will experiment with K-Means and potentially Hierarchical clustering to group customers into segments with similar shopping behavior, and then interpret each cluster by examining its characteristic feature values [**?**]. This approach supports discovery of unlabeled customer types and directly answers RQ2.

3.3 Anomaly Detection on Transactions or Customers

For **RQ3** (identifying anomalous transactions), we will perform simple anomaly detection based on the distributions and clusters learned in the previous steps. One approach is to compute distance-based anomaly scores, such as the distance of a transaction or customer to its assigned cluster centroid, and flag points with unusually high distances as potential anomalies [**?**]. Alternatively, we may use simple rule-based thresholds on features such as basket size or number of distinct product categories to identify extremely unusual baskets. This unsupervised approach follows the general idea of detecting deviations from typical transaction patterns in financial and payment systems, adapted here to grocery retail data [**?**]. The resulting anomalies will be examined qualitatively to understand what makes them unusual and how they relate to RQ3.

Figure 2:

1. **Data Ingestion and Cleaning:** Load the CSV file, parse the product lists, remove duplicate or invalid records, and standardize product names.
2. **Exploratory Data Analysis (EDA):** Compute basic statistics (e.g., distribution of basket sizes, number of unique products, transaction counts per customer) and visualize them.
3. **Association Rule Mining:** Transform transactions into a suitable format (e.g., one-hot encoding) and run Apriori or FP-Growth to find frequent itemsets and association rules.
4. **Customer Feature Engineering:** Aggregate transactions by customer to compute behavioral features such as average basket size and category-level proportions.
5. **Clustering and Segment Interpretation:** Apply K-Means or Hierarchical clustering to the customer features and interpret the resulting segments.
6. **Anomaly Detection and Inspection:** Compute anomaly scores for transactions or customers, flag outliers, and inspect them to understand what makes them unusual.

This pipeline will be documented with flowcharts and diagrams as part of the visual documentation required by the project.

4 Preliminary Timeline

This section outlines a rough plan for Milestones M2, M3, and M4, as well as anticipated challenges for the project. The timeline may be adjusted as we gain a deeper understanding of the dataset and tools.

4.1 Milestone M2: Data Understanding and Preparation

For M2, the focus will be on fully understanding and preparing the Retail Transaction Dataset for analysis.

- Download the dataset and set up the project environment (Python, Jupyter notebook, version control).
- Load the CSV file, inspect column types, and perform basic cleaning (remove duplicates, standardize product name formatting, handle missing or malformed entries).
- Compute basic descriptive statistics (e.g., distribution of basket sizes, number of unique products, top products by frequency).
- Create initial visualizations such as histograms of basket size and bar charts of popular products.
- Define and document the exact transformations needed to convert raw product lists into itemsets suitable for association rule mining.

By the end of M2, the goal is to have a clean, well-understood dataset and a clear feature representation for both transaction-level and customer-level analyses.

4.2 Milestone M3: Core Pattern Discovery

For M3, the main objective is to apply the planned data mining techniques to address the discovery questions.

- Implement association rule mining (Apriori or FP-Growth) on the transaction-level data to discover frequent itemsets and association rules.
- Tune minimum support and confidence thresholds to obtain a manageable set of interpretable rules.
- Engineer customer-level features (e.g., average basket size, transaction frequency, product category proportions) and standardize them for clustering.
- Apply clustering algorithms such as K-Means or Hierarchical clustering to discover customer segments and interpret their characteristics.

- Begin preliminary anomaly detection by computing simple anomaly scores (e.g., distance to cluster centroids or threshold-based rules on basket size) and flagging unusual transactions or customers.

By the end of M3, we expect to have a first set of association rules, customer clusters, and candidate anomalies that directly relate to RQ1–RQ3.

4.3 Milestone M4: Refinement, Evaluation, and Reporting

For M4, the focus shifts to refining the discovered patterns, evaluating their usefulness, and preparing the final deliverables.

- Refine association rule mining results by focusing on the most meaningful rules (e.g., high lift, domain-relevant product combinations) and discarding redundant or uninteresting rules.
- Evaluate and refine clustering results (e.g., experiment with different numbers of clusters, compare cluster quality metrics, and improve interpretability).
- Refine anomaly detection criteria and provide clear examples of anomalous transactions or customers with qualitative explanations.
- Create final visualizations: rule tables, cluster profile plots, anomaly examples, and the overall analysis pipeline flowchart.
- Finalize the LaTeX project report (including all sections and figures) and ensure that the GitHub repository is organized with code, documentation, and the proposal in the `docs/` folder.

4.4 Anticipated Challenges

Several challenges are anticipated over the course of the project:

- **Parameter selection:** Choosing appropriate thresholds for association rule mining (support, confidence, lift) so that the results are neither too sparse nor overwhelmed by trivial rules.
- **Cluster interpretability:** Selecting the number of clusters and feature transformations that produce segments that are both statistically reasonable and easy to interpret.
- **High-cardinality product space:** Managing the large number of distinct products when encoding transactions, which may lead to sparse high-dimensional representations.
- **Anomaly definition:** Defining "anomalous" behavior in a way that is meaningful for this dataset, rather than flagging points that are only slightly unusual due to noise.

- **Time management and iteration:** Ensuring there is enough time for iterative refinement of models and visualizations while meeting the deadlines for M2, M3, and M4.

5 GitHub Repository

https://github.com/freddyrerazo-ctrl/cs4412-retail-transactions