Uncovering Customer Purchasing Patterns in Retail Transaction Data

Ulrich Batanado
Ubatanad@students.kennesaw.edu
CS 4412
2-8-2026

***Abstract*—This project will explore a retail transaction dataset to discover patterns in how customers shop. Using the Knowledge Discovery in Databases (KDD).The results provide insight into customer purchasing habits and product relationships, while also discussing the limitations of the data and the discovered patterns.**

I. Dataset Description

The dataset used in this project is the Retail Transaction Dataset, obtained from Kaggle : (https://www.kaggle.com/datasets/fahadrehman07/retail-transaction-dataset/data)

It's a comprehensive dataset captures the essence of transactions within a retail environment. The dataset contains approximately 100,000 transaction records with 10 attributes. The file size is small enough to be easily handled on a standard personal computer.

The dataset represents individual retail transactions, where each row corresponds to a customer purchase. It includes information about customers, products, transaction details, pricing, payment methods, and time of purchase. The data is suitable for analyzing purchasing behavior, product relationships, and customer activity patterns.

Key attributes include: "CustomerID", "ProductID", "Quantity", "Price", "TransactionDate", "PaymentMethod", "StoreLocation", "ProductCategory", "DiscountApplied", and "TotalAmount"

Some data quality issues include possible outliers such as unusually large quantities or transaction amounts, and limited contextual information such as customer demographics

TABLE I
Transaction Dataset Description

Column	Description
CustomerID	Unique identifier for each customer.
ProductID	Unique identifier for each product.
Quantity	Number of units purchased for a product.
Price	Unit price of the product.
TransactionDate	Date and time of the transaction.
PaymentMethod	Payment method used by the customer.
StoreLocation	Location where the transaction occurred.
ProductCategory	Category of the product.
DiscountApplied (%)	Discount percentage applied.
TotalAmount	Total amount paid for the transaction.

.

II. Discovery Questions

A. Which products are frequently purchased together?

Why it's valuable: Recognizing relationships between products enables retailers to better understand customer purchasing behavior, plan targeted promotions, and create effective product bundles. This represents a typical case of pattern discovery in transactional data.

B. Are there distinct customer segments based on purchasing behavior?

Why it's valuable: By categorizing customers based on how much they spend, how often they buy, and the size of their orders, retailers can customize marketing efforts, loyalty schemes, and stock planning for each distinct group.

C. Which transactions or customer behaviors are unusual or anomalous?

Why it's valuable: Identifying anomalies can uncover data entry mistakes, fraudulent activities, or distinctive buying behaviors that may signal previously unnoticed opportunities or potential risks.

III. Planned Techniques

In this project, I will use a mix of data mining techniques to explore the retail transaction dataset and uncover interesting patterns. To start, I will use association rule mining, like the Apriori or FP-Growth algorithm, to find products that are often purchased together. This will show which items tend to appear in the same shopping basket, revealing natural connections between products and potential opportunities for promotions or bundling.

Next, I plan to use clustering, specifically K-Means, to group customers based on their shopping behavior. By looking at features like total spending, how often they shop, and the average size of their baskets, I can identify distinct types of customers—such as frequent low-value shoppers, occasional high-value shoppers, or bulk buyers. Alongside clustering, I will also apply a classification technique, such as decision trees, to better understand what features distinguish high-value customers from the rest. Here, the focus isn't on prediction accuracy but on understanding which factors matter most.

Finally, I will explore anomaly detection to find unusual transactions or atypical customer behavior. Techniques like

Isolation Forest or Local Outlier Factor will help spot transactions that deviate from normal patterns, which could indicate rare buying behavior, errors in the data, or even potential fraud. Together, these approaches will provide a well-rounded view of the dataset, allowing me to uncover meaningful patterns in customer behavior, product relationships, and unusual transactions—all of which will be interpreted and evaluated in the final analysis.

IV. Preliminary Timeline

The project will follow the KDD process and is organized around the three main milestones. For Milestone 2 (due March 8), the focus will be on dataset selection, preprocessing, and initial exploration. During this stage, I will clean the retail transaction dataset, handle missing or duplicate values, and perform basic feature engineering such as calculating transaction totals, basket sizes, and customer-level metrics. A potential challenge in this phase is dealing with inconsistent or incomplete data, which may require careful filtering and validation to ensure quality for later analysis.

For Milestone 3 (due April 5), the emphasis will shift to applying data mining techniques. Association rules will be used to identify frequent product combinations, clustering will segment customers based on shopping behavior, and preliminary anomaly detection will highlight unusual transactions. At this stage, tuning algorithm parameters and selecting meaningful features may be challenging, as these decisions can significantly impact the patterns discovered.

Finally, Milestone 4 (due May 4) will focus on interpreting and evaluating the results. I will analyze the patterns found, relate them back to the discovery questions, and assess their significance and limitations. Creating clear visualizations and effectively communicating insights will be important, and a potential challenge is ensuring that interpretations are valid and not overgeneralized from the dataset. Overall, this timeline ensures that each stage of the KDD process is completed systematically, allowing for thorough analysis and meaningful pattern discovery.

References

[1] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
Comprehensive reference on the KDD process, association rules, clustering, classification, and anomaly detection.

[2] Tan, P. N., Steinbach, M., & Kumar, V. (2018). *Introduction to Data Mining* (2nd ed.). Pearson.
Covers data preprocessing, feature engineering, and practical data mining applications.

[3] Kaggle. (2023). *Retail Transaction Dataset*. Retrieved from https://www.kaggle.com/datasets/fahadrehman07/retail-transaction-dataset/data
The dataset used in this project, containing transaction-level retail purchase records.

[4] Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2), 207–216.
Original paper describing association rule mining and Apriori algorithm.

[5] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD '96)*, 226–231.
Reference for clustering and anomaly detection methods.

[6] Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 413–422.
Reference for anomaly detection techniques applied in this project.

[7] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.
Practical guide for implementing data mining techniques, including feature engineering, clustering, classification, and evaluation.

[8] OpenAI ChatGPT. (2026). *ChatGPT: Large Language Model*.