Decoding Consumer Choices: An In-Depth Analysis of Instacart Market Baskets

Darien Diaz Rivera and Nancy Lopez
Knight Foundation School of Computing and Information Science
Florida International University, Miami, FL 33199 USA
Email: ddiaz169@fiu.edu, nlope003@fiu.edu
Project Website: https://instacart-mba-site/index.html

***Abstract*—The rapid growth of e-commerce and digital grocery platforms has generated an unprecedented volume of transactional data, creating new opportunities to understand customer purchasing behavior at scale. Such insights are essential for optimizing operations, enhancing recommendation systems, and improving the overall shopping experience in an increasingly competitive retail environment. In this study, the Instacart's online grocery dataset is analyzed using a combination of data mining techniques, including exploratory analysis, association rule mining, and clustering. These complementary approaches allow capturing behavioral patterns, uncovering relationships among co-purchased items, and identifying distinct customer segments. Preprocessing steps were applied to ensure data quality and support the analytical workflow. The findings highlight how different analytical perspectives contribute to a deeper understanding of customer behavior, offering practical implications for recommendation systems, store layout optimization, and cross-selling strategies.**

***Index Terms*—market basket analysis, Instacart, association rules, Apriori, consumer behavior**

I. Introduction

Understanding customer behavior has become a central focus in modern retail analytics, as purchasing decisions are shaped by complex and continually evolving factors, including the introduction of new products and shifts in customer preferences [1]. Prior research has demonstrated that analyzing large-scale transactional data can reveal valuable insights into preference formation, repeat purchasing tendencies, and product affinities that influence consumer decision-making [2]. Online grocery platforms, in particular, provide a unique environment for examining these patterns due to the high volume, fine-grained detail, and consistency of digital order records.

In this study, the publicly available Instacart dataset is utilized to investigate customer purchasing behavior through multiple data mining techniques. The dataset contains millions of orders across thousands of products, offering a comprehensive view of how customers interact with an online grocery service. To develop a foundational understanding of the data, an exploratory data analysis (EDA) is conducted to examine ordering trends across days and hours, contrast weekend and weekday behaviors, evaluate differences between new and loyal customers, and identify key products and departments with high purchase frequency and reorder rates. These descriptive findings provide context for broader behavioral tendencies observed within the platform.

To uncover deeper relationships, association rule mining using the Apriori algorithm is applied to identify products that are frequently co-purchased, revealing patterns that can inform cross-selling strategies and recommendation systems. The analysis is further extended by performing clustering based on reorder behavior to segment customers into groups with distinct purchasing tendencies. This combination of techniques enables the identification of both item-level relationships and customer-level behavioral differences.

Throughout the analysis, preprocessing steps are implemented to ensure data quality and computational feasibility, including filtering high-frequency products and structuring transactional records appropriately for each analytical method. Collectively, these perspectives offer a multifaceted understanding of Instacart customer behavior and generate insights relevant to recommendation systems, marketing strategies, and operational planning.

II. Goal

The goal of this project is to transform the massive and complex transactional data generated by the Instacart online grocery platform into a structured set of actionable business insights. Specifically, the project aims to develop a deep, data-driven understanding of customer purchasing behavior to inform strategic decision-making, improve customer satisfaction, enhance operational efficiency, and support revenue growth.

III. Objectives

The main objectives of this project are:

1) Understand Customer Purchasing Behavior: Apply Exploratory Data Analysis (EDA) to understand general shopping behavior, including key metrics such as average order size, reorder frequency, and the most popular products and departments.
2) Discover Item Relationships: Use the Market Basket Analysis (Apriori Algorithm) to identify statistically significant associations between different products, aisles, and departments (e.g., which products are frequently purchased together or which departments or aisles appear more often in the transactions). This type of

insight supports recommendation systems and cross-promotional strategies.

3) Identify Time-Based Shopping Patterns: Implement Sequential Pattern Mining to uncover common product sequences. These patterns help estimate the likelihood of purchasing a product given that a sequentially related item was bought in a previous transaction. Such insights serve as a foundation for recommendation systems, order fulfillment planning, and logistics optimization.
4) Segment the Customer Base: Apply Clustering Analysis (K-Means) to group customers into distinct segments based on their purchase profiles. This enables the development of highly customized marketing campaigns and personalized shopping experiences for each segment.
5) Generate Actionable Insights: Translate findings from all analytical methods into clear, practical recommendations that can be implemented to optimize inventory, enhance targeted marketing efforts, and improve the user interface.

IV. Motivation

The motivation for this project arises from the critical role of data analytics in maintaining a competitive advantage within the rapidly evolving e-commerce grocery sector.

1) Strategic Necessity: Grocery shopping remains an essential and routine consumer activity. As platforms like Instacart manage increasingly large volumes of transactional data, the failure to analyze this information represents a significant missed opportunity for revenue generation and service improvement.
2) Enhancing Customer Experience: Understanding the underlying structure of customer shopping habits is the foundation for personalization. By predicting purchasing trends and offering relevant recommendations, Instacart can significantly enhance the customer experience, fostering greater loyalty and satisfaction.
3) Operational Efficiency: Unlocking hidden patterns, particularly strong product associations and temporal sequences, has direct implications for internal operations. These insights allow for optimization of warehouse layouts, inventory management, and the overall efficiency of the order fulfillment process, leading to reduced costs.
4) Academic Contribution: From a data science perspective, this project serves as a case study demonstrating the effective application and integration of multiple unsupervised learning techniques, specifically EDA, association rules, sequential mining, and clustering, to solve a complex, real-world business challenge using a massive dataset.

V. Research Questions

This project focuses on answering the following questions:

1) What are some patterns in customer purchasing behaviors?
2) Which products are frequently purchased together?
3) What items are customers most likely to reorder?
4) How well can we predict future purchase behavior?

VI. Prior Works and Challenges

Prior works have successfully validated the use of techniques like Market Basket Analysis (MBA) to identify co-occurrence patterns; they have often faced critical methodological and contextual challenges that this integrated approach seeks to overcome.

Most established works on transactional data often leverage the foundational Apriori Algorithm, which is primarily centered on identifying strong association rules to inform basic cross-selling and store layout strategies. Studies previously applied to retail and e-commerce consistently yield rules involving high-volume, staple products, or findings that confirm the co-purchase of core items [3]. While these results are valuable for large-scale inventory management and general promotions, they typically lack the granularity required for highly personalized customer experiences. Also, operational studies often treat consumer behavior as a static input rather than a dynamic pattern requiring the deeper, sequential analysis this research provides.

The utility of traditional Market Basket Analysis (MBA) and related methods significantly diminishes when applied to high-volume, sparse datasets like the one provided by Instacart, where major limitations become evident. First, the massive size of e-commerce data creates a computational bottleneck; the Apriori algorithm requires repeated database scans, forcing researchers to employ high minimum support thresholds and by doing that, it filters out thousands of valuable low-frequency associations while also leading to a trivial rule saturation involving common staple items [4]. Second, traditional MBA is a static analysis that suffers from a lack of context. It ignores the temporal sequence of purchases, what is bought next, which is necessary for predictive modeling, and fails to integrate customer identity by analyzing a generalized average shopper. This methodological gap prevents the development of personalized recommendations and segment-specific marketing strategies that are essential for modern e-commerce success.

A critical shortcoming in the majority of the prior MBA literature is its narrow focus on a single, static transaction. This static analysis leads to the Neglect of Customer Segmentation; prior studies usually generalize findings to an average customer, failing to account for the heterogeneous purchasing habits of different groups. A promotional strategy based on general association rules will inevitably fail to resonate with distinct customer segments as it is not properly tailored for the specific segment. The integration of Sequential Pattern Mining and K-Means Clustering in this project is specifically designed to provide the necessary time-based and customer context to generate personalized insights; this level of detail is often missing in published works.

In summary, while prior works have successfully validated the tools for transactional analysis, they often sacrifice depth and personalization for computational expediency. This research addresses these deficiencies by adopting an integrated

analytical framework designed to extract more granular, contextual, and ultimately actionable insights.

VII. Data Sources and Description

The foundation for this study is the Instacart Online Grocery Basket Analysis dataset, publicly released by Instacart and made available through Kaggle [5]. This dataset provides a real-world, large-scale snapshot of consumer transactional data from Instacart, a leading online grocery delivery platform.

The data is highly granular, structured across 5 relational tables that collectively contain over 34 million records of purchasing activity. The vast amount of data presents both a computational challenge, as noted in the project overview, and a rich source of patterns essential for Market Basket Analysis (MBA) and sequential pattern mining.

The components of the dataset, their approximate sizes, and main features are summarized in table I.

TABLE I: Data sources information

File	Size	Features	Description
Orders	3.4 M	user_id, order_id, order_number, order_dow, order_hour_of_day, days_since_prior_order	Metadata describing when the order was placed and the time elapsed since the previous order
Order Products	32.4 M	order_id, product_id, add_to_cart_order, reordered	Core transactional data linking products to orders
Products	49,688	product_id, product_name, aisle_id, department_id	Descriptive data linking product IDs to names and hierarchical categories
Aisles	134	aisle_id, aisle	Descriptive table mapping aisle IDs to names
Departments	21	department_id, department	Descriptive table mapping department IDs to names

This comprehensive, relational structure allows for a full end-to-end analysis, from raw product purchases to categorized shopping patterns, enabling the discovery of strong association rules and sequential patterns critical for generating the actionable business insights detailed in this research.

VIII. Methodology

A. Data Preprocessing

This analysis incorporates a comprehensive data processing phase to ensure the reliability and usability of the Instacart dataset. The process involves cleaning and preparing multiple interrelated tables, including orders, order products, products, aisles, and departments. Data types are standardized, missing or inconsistent values are addressed, and duplicate entries are removed to establish a consistent analytical foundation.

After cleaning, the datasets are merged using shared primary keys such as order ID, product ID, aisle ID, and department ID, enabling the construction of an integrated dataset that captures product-level, user-level, and order-level information. Additional preprocessing steps include generating derived fields, such as aggregated customer metrics, to enrich the dataset and support subsequent pattern mining. The dataset is further filtered to reduce computational complexity while preserving meaningful patterns.

This structured preparation results in a high-quality dataset suitable for exploratory analysis, association rule mining, and customer segmentation.

B. Exploratory Data Analysis (EDA)

Following the data preparation step, an exploratory data analysis was conducted to gain a comprehensive understanding of the structure, distribution, and behavioral patterns within the Instacart dataset. This step involved examining key dimensions such as product popularity, ordering frequency, and trends across days of the week and hours of the day. Customer ordering patterns were analyzed, including the prevalence of reorders, typical order sizes, and the distribution of products across aisles and departments, to identify high-activity categories and consumption trends. Summary statistics, visualizations, and cross-tabulations were used to uncover initial insights about customer behavior and to highlight patterns that needed deeper investigation through association rule mining. The EDA also helped identify potential biases or sparsity issues in the dataset, informed the selection of appropriate support thresholds, and guided the segmentation of users and products for more targeted pattern discovery. Overall, the EDA phase established the foundational understanding necessary to interpret subsequent market basket analysis results accurately and meaningfully.

C. Association Analysis

The next phase of the analysis consists of performing association analysis to uncover meaningful relationships between products that are frequently purchased together. This phase applies the Apriori algorithm to identify frequent itemsets, defined as combinations of products that appear in customer orders at a frequency exceeding a predefined minimum support threshold [6]. From these itemsets, association rules are generated to quantify the strength of relationships between products using three key metrics: support, confidence, and lift [7]. Support measures how often an itemset appears in the dataset and is defined in Equation 1.

$$\text{Support, } s(X \to Y) = \frac{\sigma(X \cup Y)}{\text{N}} \tag{1}$$

Where sigma refers to the count of transactions where items X and Y appear together, and N is the total number of transactions or orders.

Confidence, shown in Equation 2 reflects the conditional probability of purchasing item Y when X is present.

$$\text{Confidence, } c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{2}$$

Equation 3 defines Lift and measures how strongly the presence of X increases the likelihood of purchasing Y.

$$\text{Lift}(X \to Y) = \frac{c(X \to Y)}{\sigma(Y)} \tag{3}$$

D. Clustering and Segmentation

To further explore trends in purchasing behavior beyond association patterns, a segmentation analysis was performed using clustering techniques. This study focused on reorder rate as the primary behavioral feature, as it provides insight into the degree of consistency in shopping habits. The reorder rate was computed with Equation 4.

$$\text{Reorder Rate} = \frac{\text{Number of reordered items}}{\text{Total number of items purchased}}. \quad (4)$$

After calculating this metric for all users, the k-means clustering algorithm was applied to group customers with similar reorder tendencies [8]. The k-means algorithm partitions observations into k clusters by minimizing the Euclidean distance between each data point and its corresponding cluster, as shown in Equation 5. Where μ_i denotes the centroid of cluster C_i [9]. The distribution of reorder rates was examined to guide the selection of an appropriate number of clusters and to ensure meaningful segmentation. The resulting clusters revealed distinct behavioral profiles, such as users with highly repetitive purchasing patterns and those who exhibit more exploratory behavior. These segments enhance the interpretation of association rules by illustrating how product affinities may vary across different types of shoppers.

$$\min_{C_1,\ldots,C_k} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2, \quad (5)$$

IX. Results

A. Time-Based Ordering Patterns

The heatmap in Figure 1 shows that online grocery orders peak on Sundays, especially between 9:00 a.m. and 5:00 p.m., suggesting that many customers use the weekend to prepare for the week. Weekdays exhibit a more moderate but steady pattern, with most orders placed during daytime hours from roughly 7:00 a.m. to 5:00 p.m. Across all days, overnight activity remains minimal. Overall, customers predominantly shop during typical waking hours, with a strong concentration on Sunday and Monday and early afternoons.

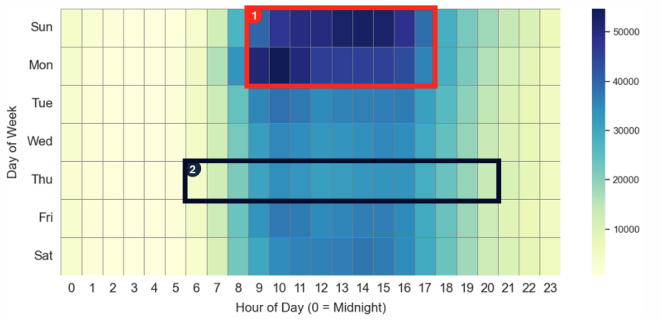

Fig. 1: Orders by Day and Hour

B. Weekend vs. Weekday Trends

The horizontal bar graph in Figure 2 shows that produce, canned goods, and frozen items are more common on weekends. Beverages, snacks, and breakfast items are more popular on weekdays. Items in the middle, such as bakery and bulk, are consistent through the week.

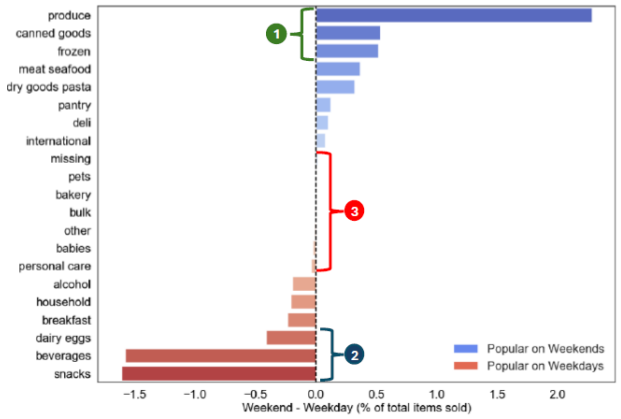

Fig. 2: Weekend vs. Weekday comparison

C. New vs. Loyal Customer Behavior

Customers were categorized into 3 groups based on the number of orders. Customers with less than 6 orders were assigned to the New group, customers with 6 to 15 orders were assigned to the Regular group, and customers with 16 or more orders were assigned to the Loyal group. By analyzing the basket size, no significant difference among the 3 groups was observed; basket size remained consisted with about 10 items. In terms of reorder rate, the customers in the Loyal group were 2.4 times more likely to reorder than New Customers (68% vs 28%) as shown on Figure 3.

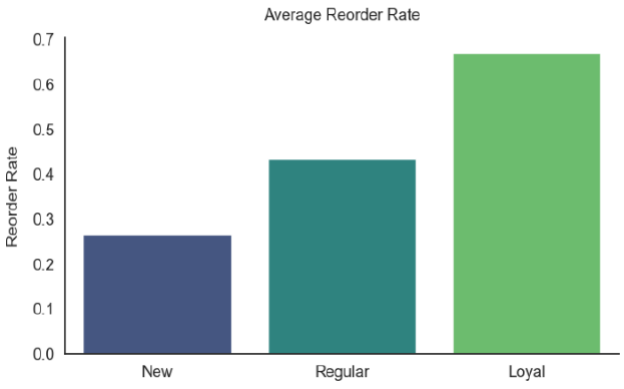

Fig. 3: Reorder Rate by Loyal group

D. Order Frequency

Weekly purchasing patterns were identified, along with broader monthly trends, as illustrated in Figure 4. These recurring cycles indicate that many customers shop on a

consistent schedule, suggesting routine replenishment behavior or enrollment in recurring delivery options. The presence of a clear 30-day peak further supports this interpretation, pointing to a subset of customers who place orders on a monthly basis, likely aligned with household replenishment habits or pay cycles. This regularity also aligns with the strong weekend trend observed in the data, where Sundays and Mondays consistently show the highest order volumes. Together, these patterns highlight predictable shopping tendencies influenced by weekly routines, monthly replenishment needs, and potentially subscription-based fulfillment services.

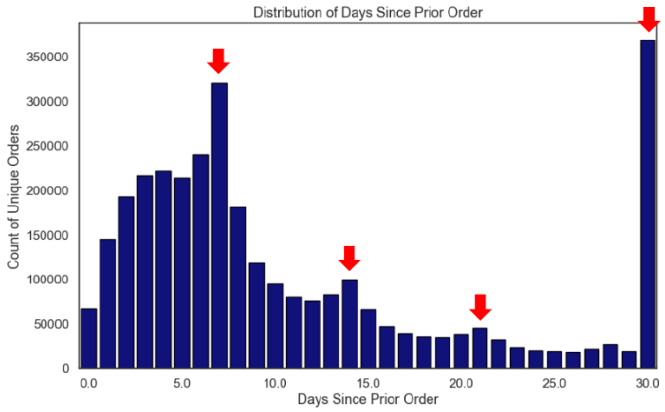

Fig. 4: Purchase patterns

E. Top 100 Products by Order Volume

To visualize the wordcloud on Figure 5, the product names were normalized by applying Natural Language Processing (NLP) techniques. The Dairy and Fresh Produce products dominate overall order demand and top categories like banana, avocado, and milk show high reorder rates, reflecting repeat consumption patterns.

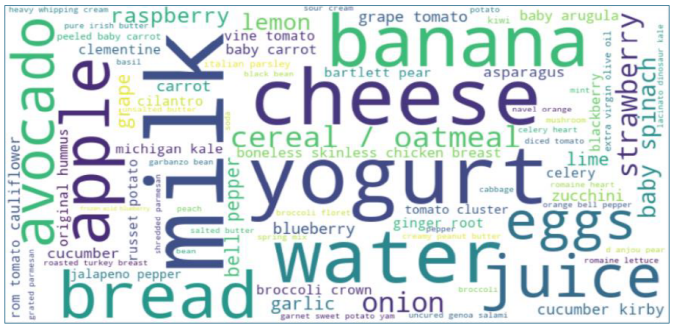

Fig. 5: Wordcloud Top 100 products

F. Data Sampling

The Figure 6 illustrates the initial data reduction step necessary for the analytical phase of the project. Due to the size of the Instacart dataset, which comprises millions of orders and over 34 million individual records, it presents a significant computational bottleneck for common algorithms like Apriori and K-Means. To enable the models to run efficiently and obtain timely results on conventional computers, a smaller random sample dataset was needed. This process involved drawing a statistically representative subset of the data for analysis, while allowing complex algorithms to execute effectively without sacrificing the integrity of the overall consumer behavior patterns. This strategic sampling was essential for resource management and computational feasibility during model development.

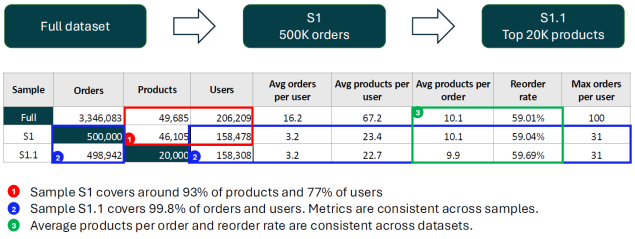

Fig. 6: Data Sampling

G. Association Rule Mining

Association Rule Mining (Apriori Algorithm) is used to identify predictable and hidden relationships between items purchased together within a single Instacart transaction. The analysis is driven by three core metrics that are used to evaluate the strength and importance of each rule.

- **Support:** Measures how frequently an item combination appears in the entire dataset.
- **Confidence:** Indicates the reliability of the rule (e.g., the probability of buying item Y given that item X has already been purchased).
- **Lift:** Quantifies the strength of the association. A Lift value greater than 1.0 means the two items are more likely to be bought together than if they were independent, confirming a positive relationship.

1) Products Frequently Purchased Together: Table II shows the top frequent itemsets for products frequently purchased together with parameters set for minimum support of 0.5% and minimum confidence of 30%. A total of 76 frequent itemsets with a maximum length of 2 were identified. The most frequent associations occur within produce, especially fruits.

TABLE II: Top Frequent Itemsets

Itemset	Count	Support
[Bag of Organic Bananas, Organic Hass Avocado]	9,558	1.92%
[Bag of Organic Bananas, Organic Strawberries]	9,522	1.91%
[Organic Strawberries, Banana]	8,717	1.75%
[Banana, Organic Avocado]	8,376	1.67%
[Organic Baby Spinach, Banana]	8,013	1.61%
[Bag of Organic Bananas, Organic Baby Spinach]	7,323	1.59%
[Strawberries, Banana]	6,443	1.29%
[Banana, Large Lemon]	6,405	1.28%
[Bag of Organic Bananas, Organic Raspberries]	6,347	1.28%
[Organic Strawberries, Organic Hass Avocado]	6,181	1.24%

Table III shows the 6 association rules generated by the model displaying a strong confidence and lift. Bananas are a strong anchor product, appearing in all rules.

TABLE III: Product-level Association Rules

Rule	Support	Confidence	Lift
[Organic Fuji Apple] → [Banana]	1.05%	38.01%	2.58
[Honeycrisp Apple] → [Banana]	0.98%	35.53%	2.41
[Cucumber Kirby] → [Banana]	0.59%	32.82%	2.23
[Organic Large Fuji Apple] → [Bag of Organic Bananas]	0.75%	31.86%	2.72
[Organic Avocado] → [Banana]	1.58%	30.44%	2.30
[Seedless Red Grapes] → [Banana]	0.79%	30.31%	2.06

2) Frequent Aisle Combination: On this section, the top frequent itemsets at the aisle level are analyzed. Because the number of ailes is considerably lower, a greater support is expected, therefore the threshold have to be adjusted, the minimum support was set to 1% while the minimum confidence is kept at 30%. This resulted in 2,704 frequent itemsets with a maximum length of 6. In addition, 6,517 rules were generated. Table IV shows the top frequent itemsets generated.

TABLE IV: Top Frequent Itemsets at Aisle Level

Itemset	Count	Support
[fresh fruits, fresh vegetables]	158,927	31.85%
[fresh fruits, packaged vegetables fruits]	135,534	27.16%
[fresh vegetables, packaged vegetables fruits]	117,341	23.52%
[fresh fruits, fresh vegetables, packaged vegetables]	93,677	18.78%
[fresh fruits, yogurt]	93,353	18.71%
[fresh fruits, milk]	81,824	16.40%
[packaged cheese, fresh fruits]	76,393	15.32%
[fresh vegetables, yogurt]	71,881	14.41%
[packaged cheese, fresh vegetables]	66,613	13.39%
[yogurt, packaged vegetables fruits]	63,695	12.77%

Table V chart shows the results of the Apriori analysis and highlights the sparse nature of grocery purchases. The number of popular combinations drops very quickly as the length increases from 2 items to 3 items and beyond. This pattern confirms that while customers have core items they buy regularly, the specific combination of many different products in a single basket is rare. Therefore, the most valuable and statistically reliable product relationships for generating cross-selling ideas are found in the shorter combinations, mainly those containing just two or three items.

TABLE V: Aisle-Level Stats

Len	Itemsets	Min Support	Max Support	Median Support	Std Dev
2	788	1.00%	31.85%	1.73%	2.60%
3	1,154	1.00%	18.77%	1.45%	1.28%
4	633	1.00%	7.57%	1.31%	0.72%
5	125	1.00%	3.30%	1.18%	0.40%
6	4	1.01%	1.62%	1.11%	0.28%
Total	**2,704**	**1.00%**	**31.85%**	**1.45%**	**1.72%**

The top rules generated for aisle combinations, displayed on Table VI, reveal that the purchase of fresh vegetables is consistently the primary outcome, the consequent, indicating its role as a strong anchor item in multi-aisle purchasing patterns. The antecedent items most frequently associated with this purchase are combinations involving fresh herbs, fresh fruits, and packaged vegetables fruits. All the rules demonstrate extremely high confidence, mostly above 92%, which means that if a customer buys the antecedent items, they are very likely to also purchase fresh vegetables, the consequent item. Furthermore, the lift value for almost every rule is greater than 2, signifying a very strong positive correlation and co-purchasing behavior between the specific combinations of items suggesting these items are purchased together more often than would occur by chance.

3) Aisle Network: The aisle-level network represented in Figure 7 visualizes the strongest associations between grocery aisles based on customers' co-purchasing behavior. Each node represents an aisle, and the size of the node reflects its relative frequency in the dataset, larger nodes correspond to aisles that appear more often in customer baskets. Edges between nodes indicate that items from those aisles frequently occur together within orders, and the thickness of each edge is proportional to the strength of the association. The percentages shown on the edges represent the corresponding support.

The graph displays several behavior patterns. For example, fresh fruits, fresh vegetables, milk, and packaged cheese form a dense cluster with multiple high-support connections. This indicates that customers who purchase from one of these aisles are highly likely to purchase from the others within the same trip. These aisles represent core staples in grocery shopping, and their strong interconnectedness highlights their role as anchor categories in weekly household stock-up missions.

Other notable associations include yogurt, packaged vegetables and fruits, and sparkling water, each showing moderately strong ties with the central fresh produce cluster. Conversely, aisles such as bread or chips and pretzels display fewer or weaker connections, suggesting that these categories are often purchased independently rather than as part of larger staple-based baskets.

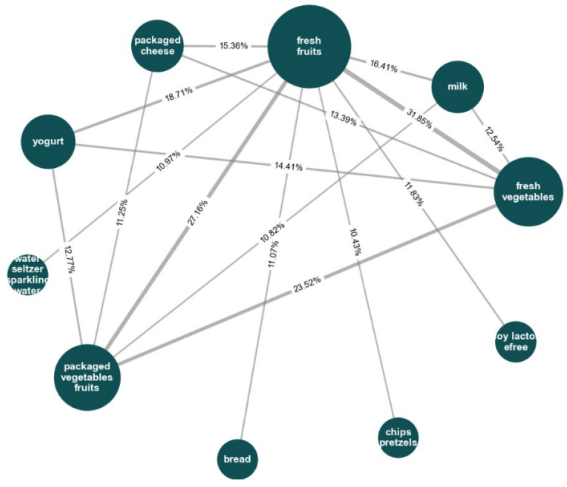

Fig. 7: Aisle Top Associations

TABLE VI: Aisle-Level Top Rules

Rule	Support	Confidence	Lift
[fresh herbs, fresh fruits, canned jarred vegetables] → [fresh vegetables]	1.25%	93.59%	2.10
[fresh herbs, fresh fruits, soy lactose free, packaged vegetables fruits] → [fresh vegetables]	1.14%	93.29%	2.10
[fresh herbs, fresh fruits, canned meals beans] → [fresh vegetables]	1.06%	93.22%	2.09
[fresh herbs, fresh fruits, soup broth bouillon] → [fresh vegetables]	1.13%	92.85%	2.09
[fresh herbs, fresh fruits, yogurt, packaged vegetables fruits] → [fresh vegetables]	1.55%	92.49%	2.08
[fresh herbs, frozen produce, packaged vegetables fruits] → [fresh vegetables]	1.10%	92.38%	2.07
[fresh herbs, packaged cheese, fresh fruits, packaged vegetables fruits] → [fresh vegetables]	1.49%	92.37%	2.07
[fresh herbs, eggs, packaged vegetables fruits] → [fresh vegetables]	1.07%	92.27%	2.07
[fresh herbs, soy lactose free, packaged vegetables fruits] → [fresh vegetables]	1.28%	92.21%	2.07
[packaged cheese, fresh vegetables, milk, yogurt, packaged vegetables fruits] → [fresh vegetables]	1.62%	91.92%	1.65

Overall, the network graph demonstrates that customer baskets tend to center around fresh produce and dairy aisles, and these core food groups drive many of the strongest co-purchasing patterns. This type of graph can help identify which aisles act as key basket drivers, inform product placement strategies, and reveal opportunities for cross-promotions or aisle adjacencies.

4) Frequent Department Combination: Analyzing the frequent department combinations, the strongest purchase patterns were observed. They revolve around Produce and Dairy Eggs, which are present in the combinations with highest support (55.07%). The top frequent itemsets show that Produce and Dairy Eggs are key anchors, consistently appearing alongside other departments like Snacks (33.18%), Beverages (33.00%), and Pantry (28.09%), as displayed on Table VII

TABLE VII: Department-Level Top Frequent Itemsets

Itemsets	Count	Support
[produce, dairy eggs]	274,779	55.07%
[produce, snacks]	165,551	33.18%
[produce, beverages]	164,638	33.00%
[beverages, dairy eggs]	158,041	31.68%
[dairy eggs, snacks]	157,951	31.66%
[frozen, produce]	147,348	29.53%
[produce, pantry]	140,161	28.09%
[frozen, dairy eggs]	140,127	28.08%
[produce, dairy eggs, snacks]	132,674	26.59%
[pantry, dairy eggs]	131,518	26.36%

The analysis generated 2,644 frequent itemsets, reaching a maximum length of 8 departments in a single combination as shown on Table VIII.

TABLE VIII: Department-Level Stats

Len	Itemsets	Min Support	Max Support	Median Support	Std Dev
2	123	3.32%	55.07%	11.46%	7.01%
3	422	3.41%	5.29%	4.70%	1.32%
4	765	1.75%	3.41%	2.35%	0.43%
5	784	1.53%	1.79%	1.24%	0.16%
6	441	1.04%	1.32%	1.12%	0.06%
7	106	1.00%	1.07%	1.00%	0.02%
8	3	1.00%	1.00%	1.00%	0.00%
Total	2,644	**1.00%**	**55.07%**	**1.78%**	**3.24%**

The strong presence of Produce and Dairy Eggs in the highest-supported combinations confirms their importance as the foundation of cross-departmental purchasing behavior.

The rules generated for department combinations highlights that Produce and Dairy Eggs are the strongest anchor departments in multi-department purchasing patterns, acting as the consequent in many of the top association rules. The top rules show very high confidence, all over 96%, particularly, those involving antecedents like snacks, deli, and produce as shown on Table IX. While Produce is the consequent in four of the top rules, Dairy Eggs is the consequent in six, indicating it is the most frequently co-purchased item when customers buy combinations of items from departments such as dry goods, pantry, canned goods, frozen, bakery, and deli. Despite the high confidence, the lift values are moderate, ranging from 1.28 to 1.42, suggesting that while the co-purchase is highly reliable, the items are only slightly more likely to be bought together than expected by chance.

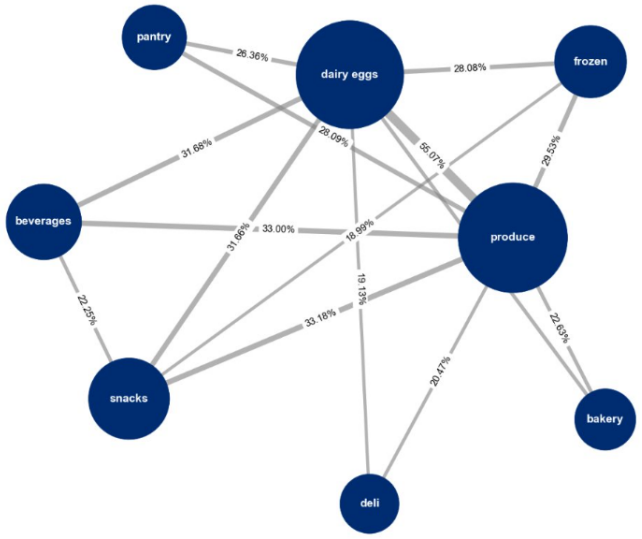

Fig. 8: Department Network Graph

5) Department Network: Figure 8 shows the network relationship across the 21 departments. The analysis was performed with a minimum support of 1% and a minimum confidence of 30%. This yielded 2,644 frequent itemsets and 27,262 association rules. The network graph helps visually highlight the key relationships between departments.

TABLE IX: Department-Level Top Rules

Top Rules	Support	Confidence	Lift
[dry goods pasta, pantry, canned goods, dairy eggs, snacks, deli] → [produce]	1.12%	96.41%	1.29
[meat seafood, pantry, canned goods, dairy eggs, deli] → [produce]	1.19%	96.21%	1.28
[frozen, bakery, produce, breakfast, snacks, deli] → [dairy eggs]	1.20%	96.18%	1.42
[bakery, produce, pantry, breakfast, snacks, deli] → [dairy eggs]	1.07%	96.18%	1.42
[dry goods pasta, pantry, canned goods, snacks, deli] → [produce]	1.19%	96.13%	1.28
[bakery, produce, beverages, dry goods pasta, pantry, deli] → [dairy eggs]	1.04%	96.10%	1.42
[frozen, bakery, produce, dry goods pasta, breakfast, snacks] → [dairy eggs]	1.07%	96.05%	1.42
[bakery, produce, beverages, breakfast, snacks, deli] → [dairy eggs]	1.18%	96.02%	1.42
[bakery, dry goods pasta, canned goods, dairy eggs, snacks, deli] → [produce]	1.00%	96.01%	1.28
[produce, dry goods pasta, pantry, breakfast, deli] → [dairy eggs]	1.00%	96.01%	1.42

For example, the Dairy Eggs department shows a very strong association with Produce, with 55.07% confidence, indicating that customers frequently purchase items from these departments together. The visualization also serves to identify the most potent co-purchase pairings, allowing for strategic merchandising and promotional decisions.

H. Customer Segmentation

1) K-means: The customer segmentation process began with a data preparation stage, which involved calculating the reorder rate for each user and constructing a user–department matrix in which the rows represent users, the columns represent departments, and the values correspond to each user's reorder rate. This matrix was then used as input for the clustering stage, where the K-Means algorithm was applied to group users into distinct clusters. Based on the elbow method, shown in Figure 9, a value of K = 5 was selected.

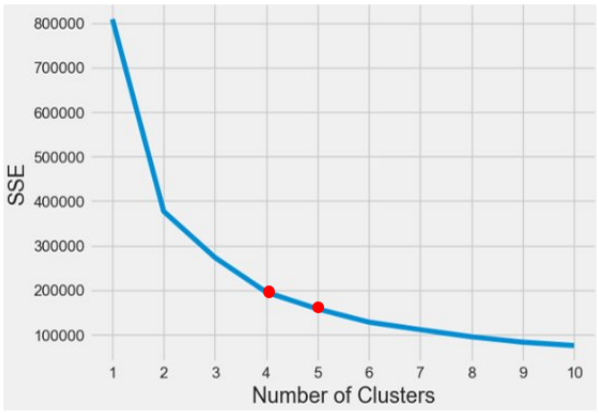

Fig. 9: Elbow Method

2) PCA: Principal Component Analysis (PCA) was used to reduce dimensionality [10]. The result is visualized as segments and their centroids, as seen in the scatter plot on Figure 10. Finally, the segment analysis stage takes these clusters to summarize each segment and provide insights, yielding distinct customer profiles such as "Core Grocery Essentials," "Bulk/Stock-up shoppers," and "Family-oriented". This methodology leverages reorder behavior and machine learning to move from raw user data to actionable, distinct customer segments.

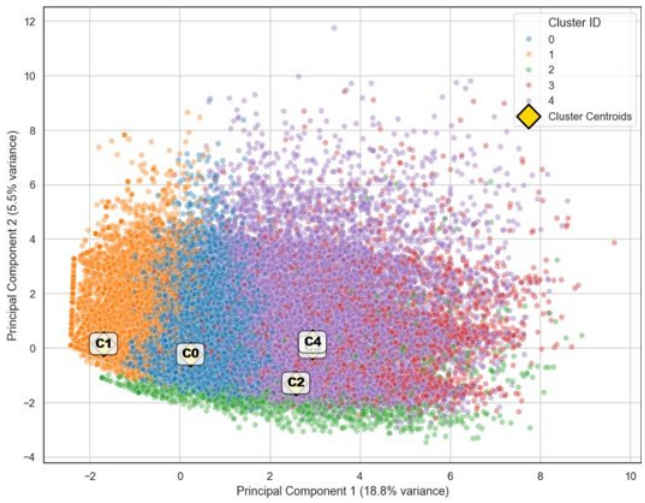

Fig. 10: Clusters

3) Segment Traits and Opportunities: The customer segmentation process, driven by reorder patterns, identified five distinct customer segments, each presenting unique traits and opportunities for targeted marketing.

- Segment 1: the Core Grocery Essentials, one of the largest group with about 56K customers; ideal to target them with strategies like category cross-selling so they can expand the items purchase footprint.
- Segment 2: characterized as Infrequent Buyers with low reorder behavior, is the largest group with 67K users, offering the primary opportunity for retention and activation campaigns.
- Segment 3: the Bulk/Stock-up Shoppers; this group exhibits a high reorder behavior despite being the smallest segment at 1.8K users, making them ideal for loyalty and subscription offers.
- Segment 4: the Family-oriented customers with about 6.7K users; they exibit high reorder behavior on deparments like Babies, Dairy Eggs, and Produce; this segment is ideal to be targeted with family bundles specials.
- Segment 5: the Premium loyalists with about 26K users in this group, demonstrates High and broad reorder behavior and is suited for VIP rewards and early access.

4) Products Frequently Purchased Together by Segment 4: This section focuses on the products frequently purchased together by Segment 4, the "Family-oriented" customers. The research showed that these customers have a strong co-purchase behavior compared to the overall population, with a +2.9 percentage point higher rate. This segment focuses heavily on premium, organic produce, as evidenced by the high frequency of items like organic strawberries, organic hass avocado, and organic baby spinach in the top itemsets as shown on Table X.

TABLE X: Segment 4 Frequent Itemsets

Itemsets	Count	Support
[Bag of Organic Bananas, Organic Hass Avocado]	1,876	4.86%
[Bag of Organic Bananas, Organic Strawberries]	1,766	4.58%
[Organic Strawberries, Banana]	1,390	3.50%
[Organic Strawberries, Organic Hass Avocado]	1,283	3.33%
[Bag of Organic Bananas, Organic Baby Spinach]	1,271	3.29%
[Bag of Organic Bananas, Organic Raspberries]	1,257	3.26%
[Banana, Organic Avocado]	1,078	2.79%
[Banana, Organic Whole Milk]	1,052	2.73%
[Organic Strawberries, Organic Baby Spinach]	1,023	2.65%
[Organic Strawberries, Organic Raspberries]	974	2.53%

5) Segment 4 Sequential Patterns at Aisle Level: Sequential pattern mining is a data mining technique used to discover statistically significant sequences of events or item purchases that occur in a specific order over time. Unlike association rule mining, which focuses on co-occurrence within a single transaction, sequential pattern mining identifies how behaviors unfold across multiple transactions, helping uncover trends such as what customers are likely to buy first, next, and later. This method is widely applied in retail analytics, recommendation systems, and customer behavior modeling [11]. The analysis of sequential pattern at the aisle level for the Segment 4 group shows the dominance of Produce categories in transition patterns, indicating a strong recurring replenishment behavior. The strongest observed transition shows that customers who purchase from fresh vegetables have a 34.3% probability of buying fresh fruits in their next order. This sequence is closely followed by the reverse, fresh fruits to fresh vegetables, and transitions involving packaged vegetables fruits.

TABLE XI: Segment 4 Aisle Transition

Transition (Aisle A → Aisle B)	Support	%
fresh vegetables → fresh fruits	13,231	34.30%
fresh fruits → fresh vegetables	13,122	34.02%
fresh fruits → packaged vegetables fruits	11,391	29.53%
packaged vegetables fruits → fresh fruits	11,377	29.50%
yogurt → fresh fruits	9,852	25.54%
fresh fruits → yogurt	9,821	25.46%
milk → fresh fruits	9,276	24.05%
fresh fruits → milk	9,230	23.95%
fresh vegetables → packaged vegetables fruits	8,904	23.08%
packaged vegetables fruits → fresh vegetables	8,863	22.98%

Dairy items such as milk and yogurt appear at the top sequences, showing key ties to produce items, see Table XI.

X. Discussions

This section focuses on discussing the insights derived from the analysis to address the four primary research questions.

To address the question "What are some patterns in customer purchasing behaviors?", the study utilized Exploratory Data Analysis (EDA) and identified patterns that support the hypothesis of scheduled replenishment in customer purchase activity. The analysis quantified a significant degree of predictability in customer transactions, revealing that orders consistently peak on Sundays and Mondays between 9:00 a.m. and 5:00 p.m. This recurring pattern indicates that the platform is predominantly used for large weekly stock-up orders as customers prepare for the upcoming week. Additionally, the presence of a clear 30-day peak in order volume suggests a dual temporal rhythm in purchasing habits, driven by both weekly cycles and monthly replenishment intervals. These findings reinforce the importance of aligning strategic workforce planning and inventory allocation with predictable peak demand periods.

In response to the question "Which products are frequently purchased together?", Market Basket Analysis was conducted using the Apriori algorithm. The results validated assumptions regarding strong product complementarity and the presence of key basket anchors. The analysis revealed quantitatively meaningful co-occurrences, such as the frequent itemset [Bag of Organic Bananas, Organic Hass Avocado]. Furthermore, the Produce and Dairy Eggs departments emerged as primary drivers of overall basket composition, collectively accounting for 55.07% support. These insights are directly actionable for optimizing digital merchandising strategies and enhancing cross-selling recommendation systems.

To explore the third question, "What items are customers most likely to reorder?", the study examined reorder patterns across customer groups. The findings indicate that Dairy and Fresh Produce items exhibit the highest likelihood of being repurchased. The analysis also revealed a meaningful correlation between customer loyalty and reorder frequency: loyal customers—defined as those with 16 or more orders—are 2.4 times more likely to reorder compared to new customers with fewer than 6 orders. This relationship underscores the strategic significance of customer lifecycle management, as reorderable items represent one of the most influential operational levers for reducing churn and increasing long-term customer value.

Finally, to address "How well can future purchase behavior be predicted?", K-Means clustering was employed to segment customers into behaviorally distinct groups. The resulting clusters demonstrated strong potential for predictive modeling. For instance, Segment 2, characterized as a large group of "Infrequent Buyers," enables precise identification of customers at high risk of low future engagement, supporting proactive retention initiatives. In contrast, customers in Segment 4—identified as family-oriented shoppers—are predicted to hold high future value and can be targeted with tailored offerings such as family-oriented bundles. This segmentation approach illustrates how future purchasing behavior can be inferred from past transactional patterns, enabling the development of personalized, data-driven engagement strategies.

XI. Contributions

This research helps to understand consumer behavior in the online grocery sector and the application of data mining techniques in e-commerce:

1) Integrated Analytical Framework: A significant contribution to this work is the use of a multi-faceted analytical framework. By combining Exploratory Data Analysis (EDA), association rule mining (Apriori), sequential pattern mining, and customer segmentation (K-Means Clustering), this study provides a more complete and detailed understanding of customer shopping habits than methods used in isolation.
2) Discovery of Actionable Product Relationships: This study successfully identified specific, high-lift association rules that offer direct utility for business strategy. For example, the strong link between fresh vegetables and fresh fruits (Lift: 2.113) can be used to inform cross-promotional campaigns, optimize product placement within the app, and increase the size of customer orders.
3) Defined Customer Segments for Targeted Marketing: Using K-Means clustering, the user base was grouped into five distinct segments based on their departmental reorder purchasing habits. This segmentation is a practical contribution, enabling Instacart to develop highly targeted marketing strategies, personalized offers, and curated shopping experiences specifically to the unique needs of each customer segment.
4) Extraction of Temporal Shopping Sequences: The application of sequential pattern mining identified common purchase paths, such as the frequent sequence of milk to fresh fruits. This temporal insight is crucial for building accurate real-time purchase prediction models, which directly supports the optimization of warehouse logistics and the efficiency of the order fulfillment process.

XII. Limitations

Although this study provides meaningful insights, multiple limitations should be acknowledged:

1) Sampling Effects: Due to computational constraints, association rules and clustering were performed on a sample of the dataset. While representative, sampling may obscure low-frequency but important patterns.
2) Lack of Demographic Attributes: The Instacart dataset does not include demographic or socioeconomic variables, limiting the ability to model deeper behavioral drivers.
3) Model Sensitivity: The segmentation results are constrained by the limited feature space used in the clustering process. With only the reorder rate driving cluster formation, the resulting segments represent a constrained behavioral perspective. Additional features may yield alternative or more robust segment structures.
4) Temporal Dynamics Not Modeled: Although purchasing cycles were observed, no predictive temporal model (e.g., Markov chains or seq2seq models) was implemented.

These limitations provide clear pathways for future work.

XIII. Conclusion

This project successfully transforms large-scale Instacart transactional data into actionable insights by applying exploratory data analysis, association rule mining, and clustering. The results reveal clear purchasing patterns, category-level behaviors, and significant differences in reorder tendencies across customer groups. Association rule mining highlights meaningful product affinities that can support recommendation systems and targeted promotions, while clustering uncovers distinct customer segments with unique behavioral profiles, providing direction for personalization and operational strategies.

Additionally, the purpose of this project was to demonstrate how different methodological approaches can be used to explore multiple analytical pathways within a complex dataset. The results presented here represent only a sample of the insights that can be uncovered. Given the richness of the Instacart dataset, this analytical framework has considerable potential for expansion, enabling deeper exploration across dimensions such as departments, aisles, time windows, customer cohorts, and product hierarchies. The opportunities for generating more granular and sophisticated insights are extensive.

The combined results from association rule mining and customer segmentation form the foundation for two key business applications:

- Optimizing Recommendations and Inventory: The strong product-to-product relationships identified through association rules provide a logical basis for building effective recommendation systems. These insights also support more strategic inventory planning by predicting which products are likely to be purchased together.
- Enhancing Targeted Marketing: The clustering analysis confirms that customers do not shop uniformly. This highlights the need for targeted marketing strategies, where promotions can be tailored to customer segments that exhibit high purchasing concentrations in specific high-value departments such as “Produce” and “Dairy Eggs.”

In summary, this work demonstrates the value of integrating multiple data mining techniques to understand consumer behavior from both item-level and customer-level perspectives. Building on these findings, future studies may incorporate sequential modeling, dynamic recommendation systems, more diverse feature engineering, or demographic attributes to further enhance predictive capabilities and business applications.

XIV. Acknowledgements

The authors would like to express their gratitude to Dr. Ananda Mondal for his guidance, support, and valuable feedback throughout the project. We also extend our appreciation to Dr. Agoritsa Polyzou for her mentorship and insightful

recommendations that significantly enhanced our methodology and analysis.

REFERENCES

[1] J. Kutz. (2025, Sep.) How to use retail data analytics to drive sales & revenue? [Online]. Available: https://airbyte.com/data-engineering-resources/retail-data-analytics

[2] T. Stylianou and A. Pantelidou, "A machine learning approach to consumer behavior in supermarket analytics," *Decision Analytics Journal*, vol. 16, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772662225000566

[3] R. Sethi, "Market basket analysis of instacart," 2023. [Online]. Available: http://ir.juit.ac.in:8080/jspui/jspui/handle/123456789/9972

[4] K. Tsoi, "Clustering and recommendation systems for instacart delivery orders," 2022. [Online]. Available: https://scholarworks.calstate.edu/downloads/pz50h311v

[5] M. Yasser, "Instacart online grocery basket analysis dataset," https://www.kaggle.com/datasets/yasserh/instacart-online-grocery-basket-analysis-dataset, accessed: 2025-01-25.

[6] D. Mwiti. (2025, Apr.) Apriori algorithm explained: A step-by-step guide with python implementation. [Online]. Available: https://www.datacamp.com/tutorial/apriori-algorithm

[7] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*, 2nd ed. Boston, MA: Pearson, 2019.

[8] N. G. Nivedhitha, M. and S. Monis, "Customer segmentation in retail using k-means clustering: A case study on shopping trends," 2025. [Online]. Available: https://ijrpr.com/uploads/V6ISSUE9/IJRPR52629.pdf

[9] Scikit-Learn, "Unsupervised learning clustering: K-means," https://scikit-learn.org/stable/modules/clustering.html#k-means, accessed: 2025-10-30.

[10] J. Shlens, "A tutorial on principal component analysis," 2014. [Online]. Available: https://arxiv.org/abs/1404.1100

[11] H. Phuc. (2025) Lab 3.3 mining sequential patterns. Accessed: 2025-11-15. [Online]. Available: https://www.kaggle.com/code/thbdh5765/lab-3-3-mining-sequential-patterns