Instacart

Jesus Casas and Luis Cedeno
Comp541 Data Mining
College of Engineering & Computer Science, California State University Northridge

Abstract. This report presents a comprehensive analysis of the Instacart Online Grocery Shopping Dataset, focusing on uncovering product association patterns, understanding customer ordering, and extracting actionable insights through data mining techniques. We describe the preprocessing workflow applied to raw transactional data, present exploratory data analysis (EDA) to characterize order and product distributions, and detail the implementation of both Apriori and FP-Growth algorithms to discover frequent itemsets and association rules. The results highlight key cross-department and intra-department purchasing patterns, displaying recommendations for targeted promotions and inventory placement.

1 Introduction

The growth of e-commerce has generated vast volumes of transactional data, enabling retailers to utilize data mining techniques to optimize shelf placement, cross-sell strategies, and inventory management. Instacart, a leading online grocery platform, provides anonymized order-level records, product metadata, and departmental hierarchies. By mining these data sets, we aim to identify frequent purchase patterns that can inform recommendation engines and store layouts.

This project leverages a subset of the Instacart dataset, comprising over 3 million orders and more than 49,000 unique products. First we preprocess and integrate disparate tables, including orders, products, aisles, and departments to create an analysis-ready dataset. Next we conduct exploratory data analysis to summarize order sizes, examine product popularity, and assess departmental distributions. Building on these insights, we apply association mining techniques (Apriori and FP-Growth) to discover high confidence co-purchase rules. Finally, we interpret the resulting rules to formulate business recommendations with the findings from our analysis.

2 Data Preprocessing and Exploratory Data Analysis

To explore cross-department purchasing behavior, specifically we limited only the Apriori algorithm to the top 15,000 items for better performance and clearer results. Otherwise this algorithm would fail during testing yielding a “Failed not enough ram” error message. We set a minimum support threshold of 0.005 and a minimum confidence threshold of 0.3 to ensure that the patterns we identified were both frequent and reliable.

To characterize overall ordering behavior and department distributions, we generated four key visualizations (Appendix Figures 1–4):

Figure 1. Top 10 Most Popular Products A descending bar chart of the top 10 products by total order count reveals that staples like Banana (≈19,000 orders) and Bag of Organic Bananas (≈15,500 orders) dominate customer baskets, followed by organic fruits and spinach (Figure 1).

Figure 2. Distribution of Products per Transaction The histogram of products per order shows a right-skewed distribution with a median of 9 items and a long tail extending to 80+ items, indicating most orders contain fewer than 15 products but large grocery runs occur infrequently (Figure 2).

Figure 3. Temporal Order Patterns Line plots of orders by day of week and by hour of day highlight peak demand on Sundays (≈600,000 orders) and Saturdays, with troughs midweek (Wednesday/Thursday). Hourly patterns show sharp increases starting at 7 AM, peaking between 10 AM–4 PM (≈280,000 orders/hour), then tapering off in the evening (Figure 3a & 3b).

Figure 4. Items Sold by Department A bar chart of total items sold per department confirms that Produce (≈410,000 items) and Dairy & Eggs (≈215,000) lead sales, followed by Snacks, Beverages, and Frozen. Less popular categories (e.g., Bulk, Pets) have under 5,000 items each (Figure 4).

Collectively, these analyses reveal customer preferences for fresh produce and dairy products, typical basket sizes around 9 items, and strong weekend and daytime ordering trends, informing both marketing and operational planning.

3 Data Analysis

For this portion of our project, we focused on uncovering association patterns within the dataset. Specifically, we analyzed association patterns across items between departments and patterns across entire departments. To achieve this, we applied two popular association rule mining algorithms: Apriori and FP-Growth. These methods allowed us to identify frequent itemsets and generate meaningful rules that highlight purchasing trends and relationships within the data.

For the association patterns across items between departments, one interesting association we found was with customers who purchased organic whole milk, who had a strong connection with and bought a bag of organic bananas, which led us to believe there is a strong association between the dairy eggs and produce departments (Appendix 6). Another notable pattern showed that users who bought blueberries often purchased bananas as well, connecting the frozen and produce departments (Appendix 6). Lastly, we also observed a strong two-way relationship between organic whole milk and organic strawberries, both of which span the dairy eggs and produce categories.

For the association patterns across departments, the strongest relationship we identified was that 87.3% of customers who purchased from canned goods also bought produce, suggesting a clear link between these two departments (Appendix 7). Another significant connection showed that 52.8% of those who bought from canned goods also bought pantry items, which represents the strongest lift at 1.45, highlighting a very strong association between these departments (Appendix 7). Lastly, 52.3% of canned goods shoppers also purchased snacks, showing the central role of canned goods in multi-department shopping behavior. Canned goods by far had the highest values for the association patterns among all departments.

4 Conclusion

Our analysis of the Instacart Dataset we uncovered meaningful association patterns. These patterns revealed how customers tend to purchase certain products together across and within departments. Using the Apriori and FP-Growth we identified items which have a strong co-purchasing relationship such as the frequent pairing of organic whole milk with organic whole bananas and strawberries which highlights the strong connection between the dairy eggs and produce departments. Lastly we want to highlight the department with the highest number of cross-department associations canned goods with cross-department relationships with produce, pantry items, and snacks suggesting canned goods has a central role in shopping behavior.

Finally we want to highlight these findings are valuable for real-world applications in enhancing the personalized item recommendation systems at checkout. By leveraging these association rules, online grocery platforms like Instacart can suggest relevant items to customers in real time, improving user experience and increasing the likelihood of additional purchases.

5 Extra Credit

For our extra credit work, we first ran a Python script, now included in the project folder alongside the newly generated products_with_prices.csv that fed all 50,000 entries from products.csv into the Ollama LLM to assign each product a realistic price. This process required approximately twenty hours of local computation, but it allowed us to treat price as a genuine weight rather than a placeholder, enabling more meaningful market-basket analysis.

Using this price-weighted dataset and limiting to the top 1,000 most frequently ordered products, we performed an Apriori analysis whose results are summarized in Figure 7. The highest weighted-lift bundle is Organic Red Onion paired with Limes; customers who buy red onions are over three times more likely than average to also purchase limes, and with an average price of \$10.99 this yields a weighted-lift of about 37.9. Other notable high-value cross-sells include Organic Cilantro to Limes (weighted-lift ≈ 34.1), and combinations of strawberries and bananas upselling raspberries (weighted-lift ≈ 27.4). Beverage pairs such as Grapefruit Sparkling Water with Lime Sparkling Water also surface prominently, reflecting extremely strong lift despite modest unit prices. These insights suggest clear opportunities for targeted bundle promotions and in-app upsell recommendations that focus on high-lift, high-revenue combinations.

Appendix

Colab: project-1-datamining.ipynb

Figure 1

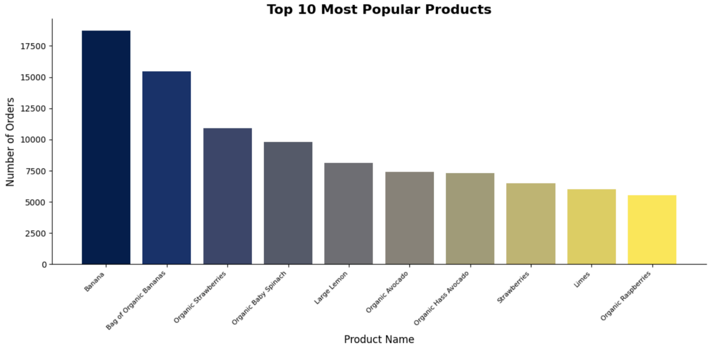


Figure 2

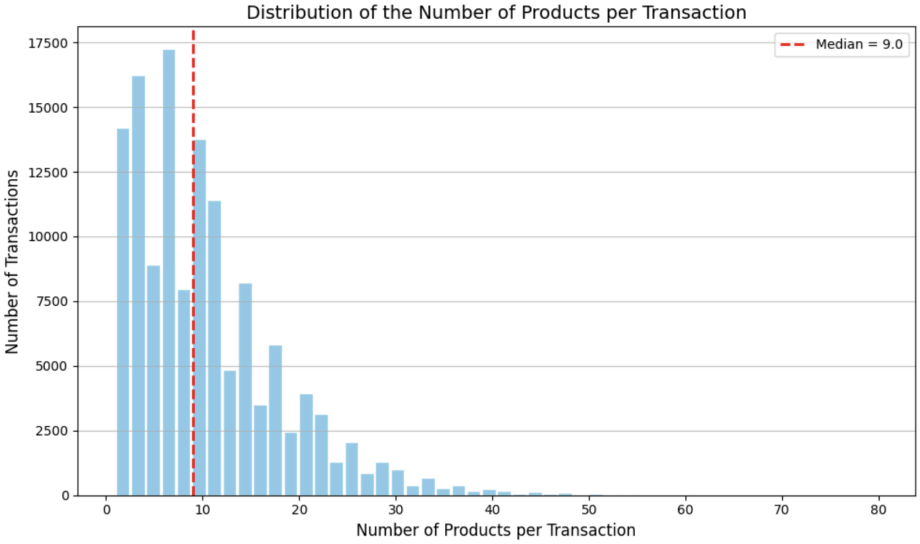


Figure 3

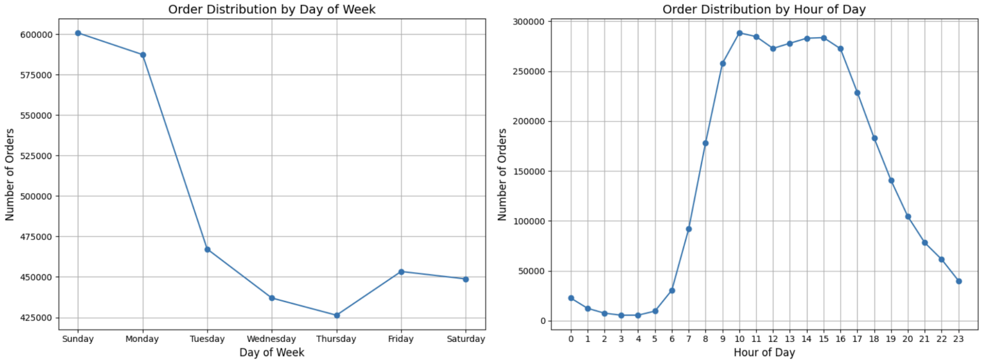


Figure 4

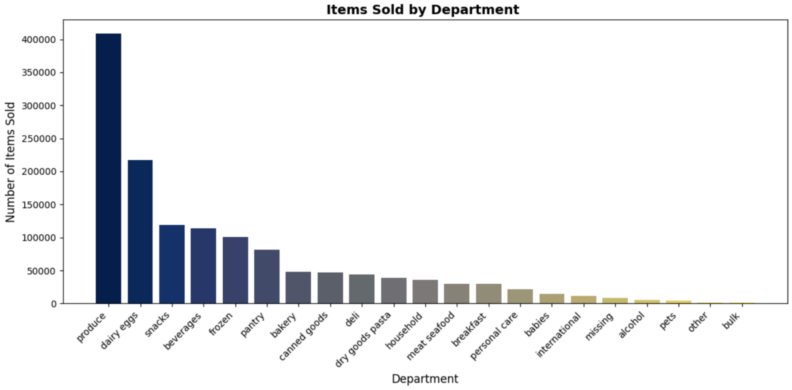
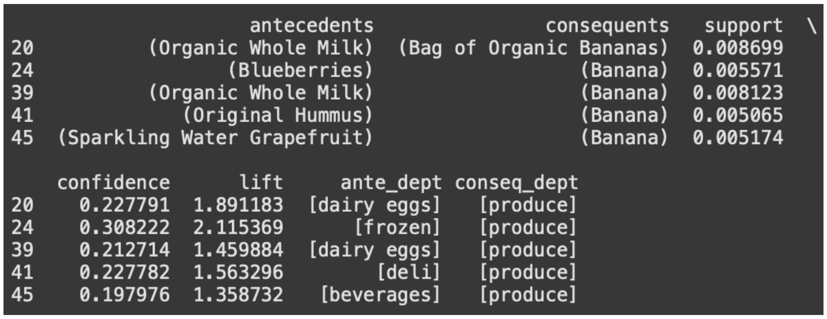
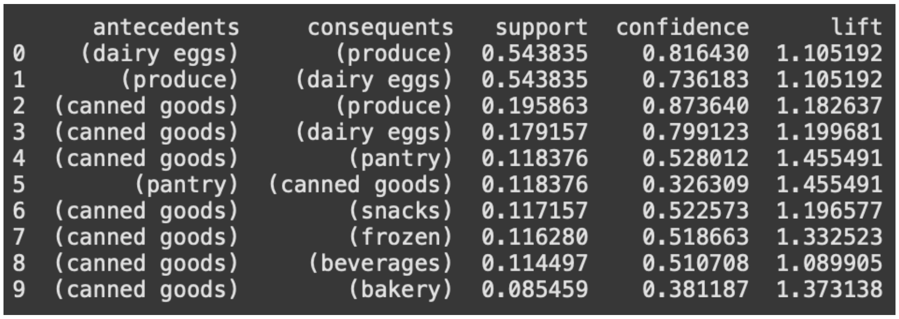


Figure 5



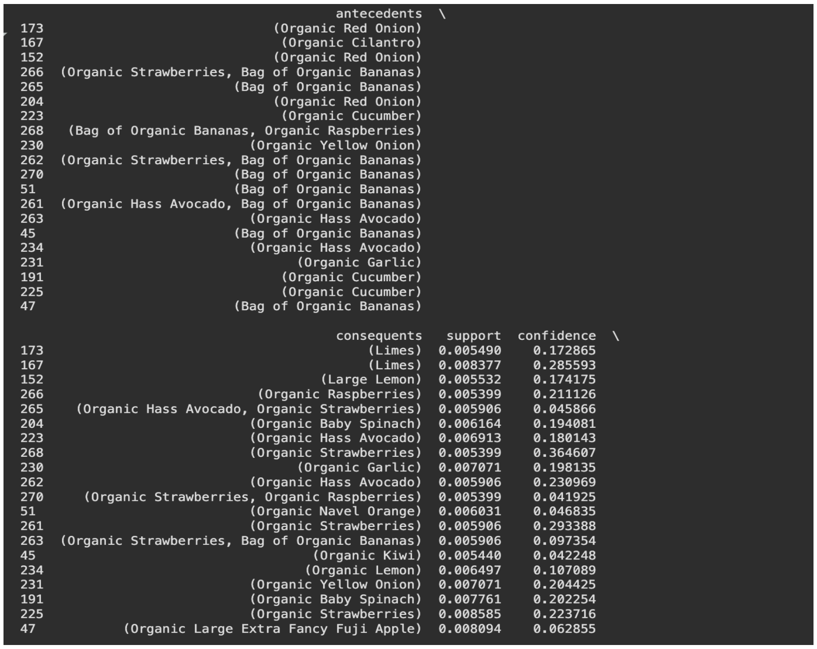
```
                     antecedents               consequents   support  \
20          (Organic Whole Milk)  (Bag of Organic Bananas)  0.008699
24                 (Blueberries)                  (Banana)  0.005571
39          (Organic Whole Milk)                  (Banana)  0.008123
41             (Original Hummus)                  (Banana)  0.005065
45  (Sparkling Water Grapefruit)                  (Banana)  0.005174

    confidence      lift     ante_dept conseq_dept
20    0.227791  1.891183  [dairy eggs]   [produce]
24    0.308222  2.115369      [frozen]   [produce]
39    0.212714  1.459884  [dairy eggs]   [produce]
41    0.227782  1.563296        [deli]   [produce]
45    0.197976  1.358732   [beverages]   [produce]
```

Figure 6



```
      antecedents     consequents   support  confidence      lift
0    (dairy eggs)       (produce)  0.543835    0.816430  1.105192
1       (produce)    (dairy eggs)  0.543835    0.736183  1.105192
2  (canned goods)       (produce)  0.195863    0.873640  1.182637
3  (canned goods)    (dairy eggs)  0.179157    0.799123  1.199681
4  (canned goods)        (pantry)  0.118376    0.528012  1.455491
5        (pantry)  (canned goods)  0.118376    0.326309  1.455491
6  (canned goods)        (snacks)  0.117157    0.522573  1.196577
7  (canned goods)        (frozen)  0.116280    0.518663  1.332523
8  (canned goods)     (beverages)  0.114497    0.510708  1.089905
9  (canned goods)        (bakery)  0.085459    0.381187  1.373138
```

Figure 7



```
                                        antecedents  \
173                             (Organic Red Onion)  
167                              (Organic Cilantro)  
152                             (Organic Red Onion)  
266  (Organic Strawberries, Bag of Organic Bananas)  
265                        (Bag of Organic Bananas)  
204                             (Organic Red Onion)  
223                              (Organic Cucumber)  
268   (Bag of Organic Bananas, Organic Raspberries)  
230                          (Organic Yellow Onion)  
262  (Organic Strawberries, Bag of Organic Bananas)  
270                        (Bag of Organic Bananas)  
51                         (Bag of Organic Bananas)  
261  (Organic Hass Avocado, Bag of Organic Bananas)  
263                          (Organic Hass Avocado)  
45                         (Bag of Organic Bananas)  
234                          (Organic Hass Avocado)  
231                                (Organic Garlic)  
191                              (Organic Cucumber)  
225                              (Organic Cucumber)  
47                         (Bag of Organic Bananas)  

                                        consequents   support  confidence  \
173                                         (Limes)  0.005490    0.172865  
167                                         (Limes)  0.008377    0.285593  
152                                   (Large Lemon)  0.005532    0.174175  
266                           (Organic Raspberries)  0.005399    0.211126  
265    (Organic Hass Avocado, Organic Strawberries)  0.005906    0.045866  
204                          (Organic Baby Spinach)  0.006164    0.194081  
223                          (Organic Hass Avocado)  0.006913    0.180143  
268                          (Organic Strawberries)  0.005399    0.364607  
230                                (Organic Garlic)  0.007071    0.198135  
262                          (Organic Hass Avocado)  0.005906    0.230969  
270     (Organic Strawberries, Organic Raspberries)  0.005399    0.041925  
51                           (Organic Navel Orange)  0.006031    0.046835  
261                          (Organic Strawberries)  0.005906    0.293388  
263  (Organic Strawberries, Bag of Organic Bananas)  0.005906    0.097354  
45                                   (Organic Kiwi)  0.005440    0.042248  
234                                 (Organic Lemon)  0.006497    0.107089  
231                          (Organic Yellow Onion)  0.007071    0.204425  
191                          (Organic Baby Spinach)  0.007761    0.202254  
225                          (Organic Strawberries)  0.008585    0.223716  
47           (Organic Large Extra Fancy Fuji Apple)  0.008094    0.062855  
```