FIU | Knight Foundation School of Computing and Information Sciences

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

Mentor: Dr. Agoritsa Polyzou
Darien Diaz (PID 3166508)
Nancy Lopez (PID 1485060)
IDC-6940 – Capstone Project
Computational Data Analytics
Florida International University
Fall 2025

Overview

Unlocking hidden patterns in customer purchase behavior

Problem

Understanding customer shopping habits is critical in the competitive grocery sector

Challenge

Massive transactional data without insights leads to lost revenue, poor experience, and inefficient operations

Solution

Apply data mining techniques to uncover patterns and generate actionable business insights

Motivation

- Grocery shopping is an essential routine
- Online platforms produce massive transactional datasets
- Unlocking hidden patterns in this data can:
  - Enhance customer experience
  - Drive revenue growth
  - Improve operational efficiency

Background - Instacart Overview

A Leader in Online Grocery Shopping

Instacart is an **online grocery delivery platform** that connects customers with personal shoppers who fulfill orders from local supermarkets

It partners with major retailers like:

- Costco
- Walmart
- Whole Foods
- Target

Goals, Objectives & Challenges

Goals / Objectives

- Understand customer purchasing behavior
- Discover product relationships
- Predict purchasing trends

Challenges

- Large dataset size
- High computational resources
- Data privacy and anonymity constraints

Key Research 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?

Methodology Overview

Data Pre-Processing

Exploratory Data Analysis (EDA)

Association Rule Mining (Market Basket Analysis)

Visualization and Interpretation

Clustering / Customer Segmentation

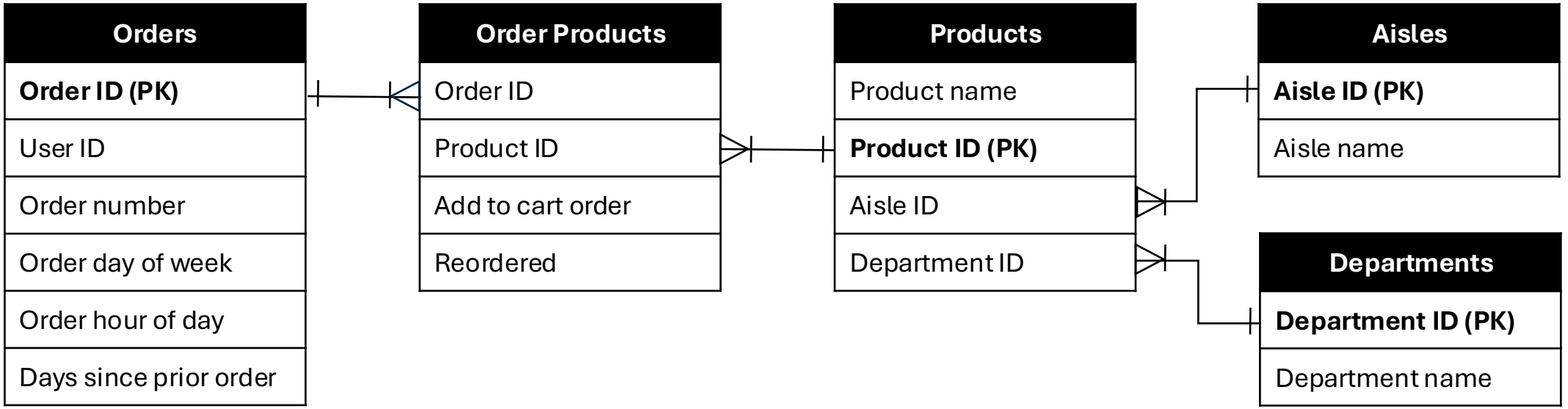
Dataset Overview

Source: Kaggle – *InstaCart Online Grocery Basket* dataset

Orders	Products	Customers	Aisles	Departments
				
3,421,083	**49,688**	**206,209**	**132**	**21**

***Total Records 33,819,106*

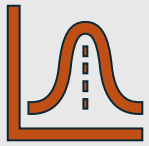
Entity Relationship Diagram



Exploratory Data Analysis (EDA)

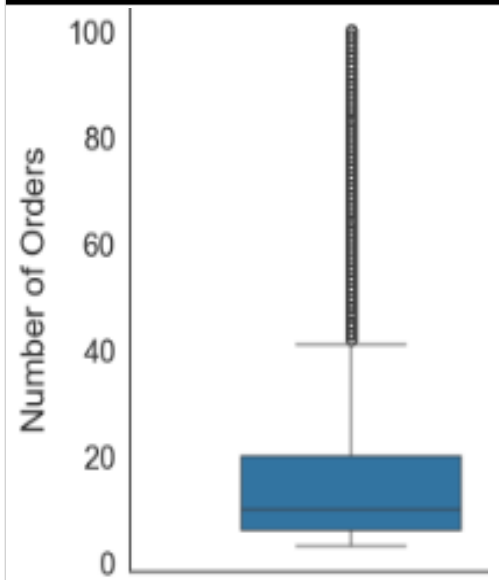
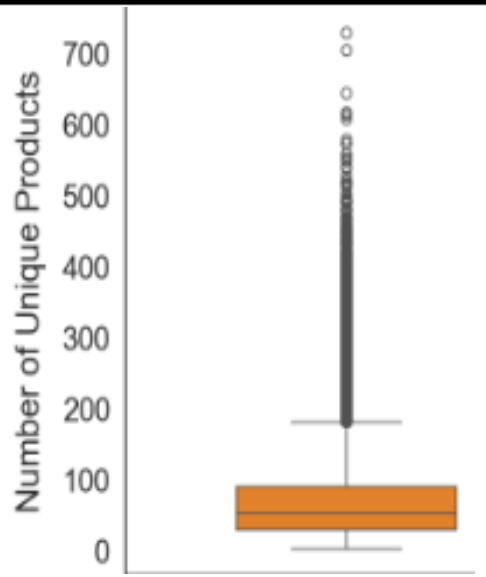
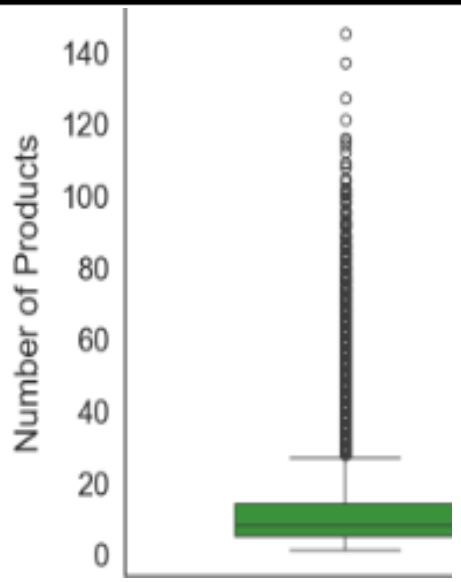
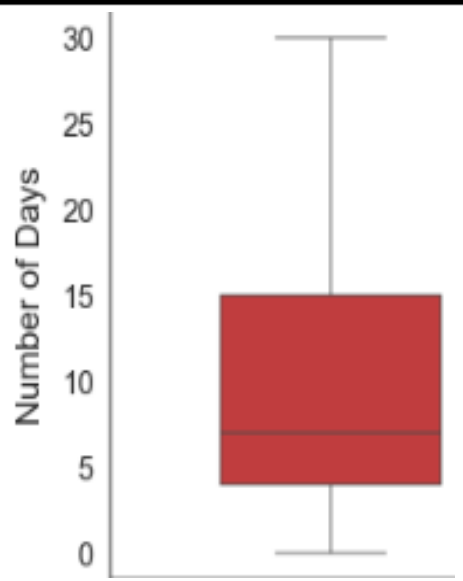
What patterns exist in customer purchasing behavior?

Explore the Instacart dataset to uncover hidden trends.

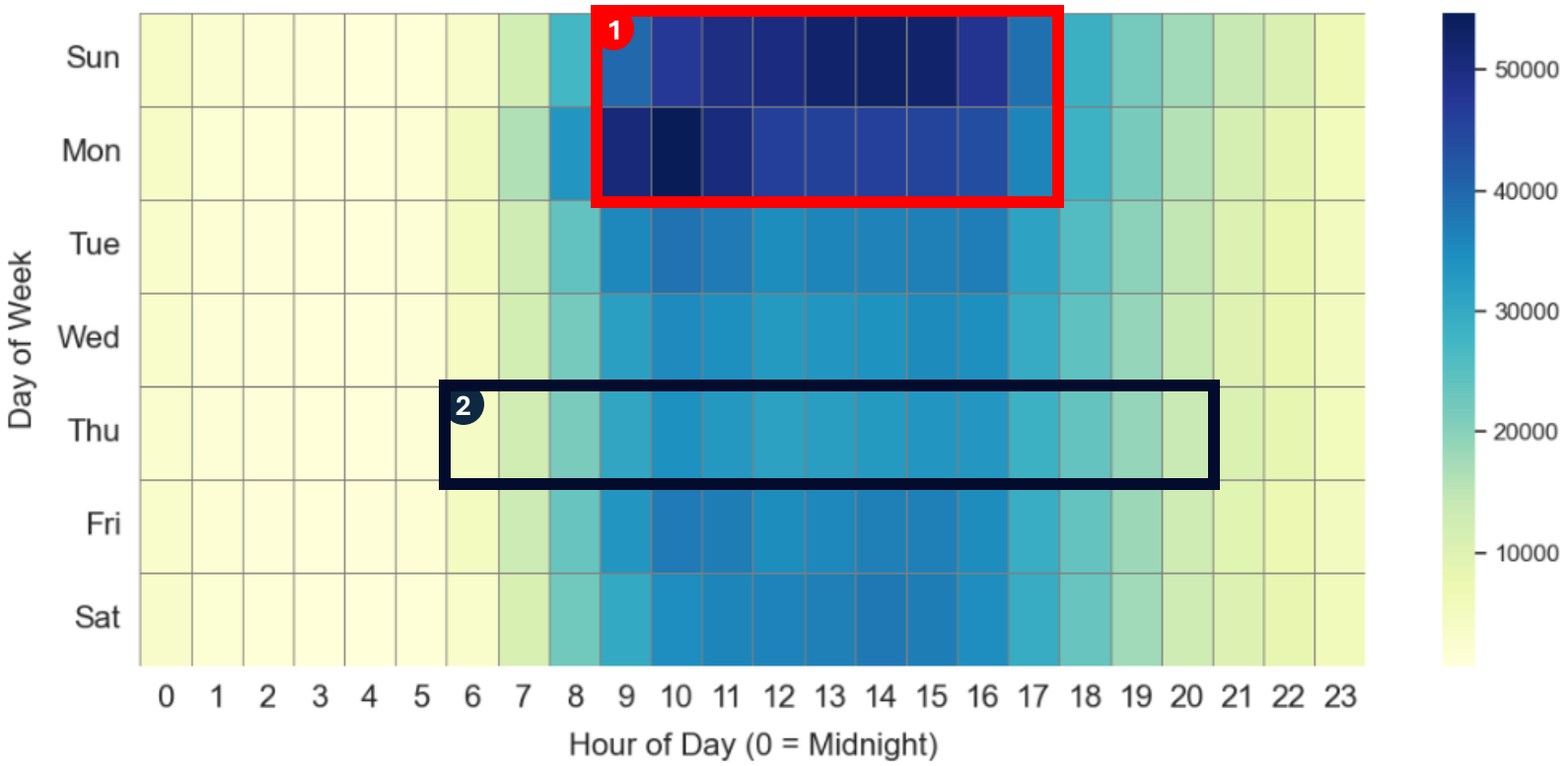


Identify distributions, trends, and product-level insights.

Summary Statistics of User and Order Behavior

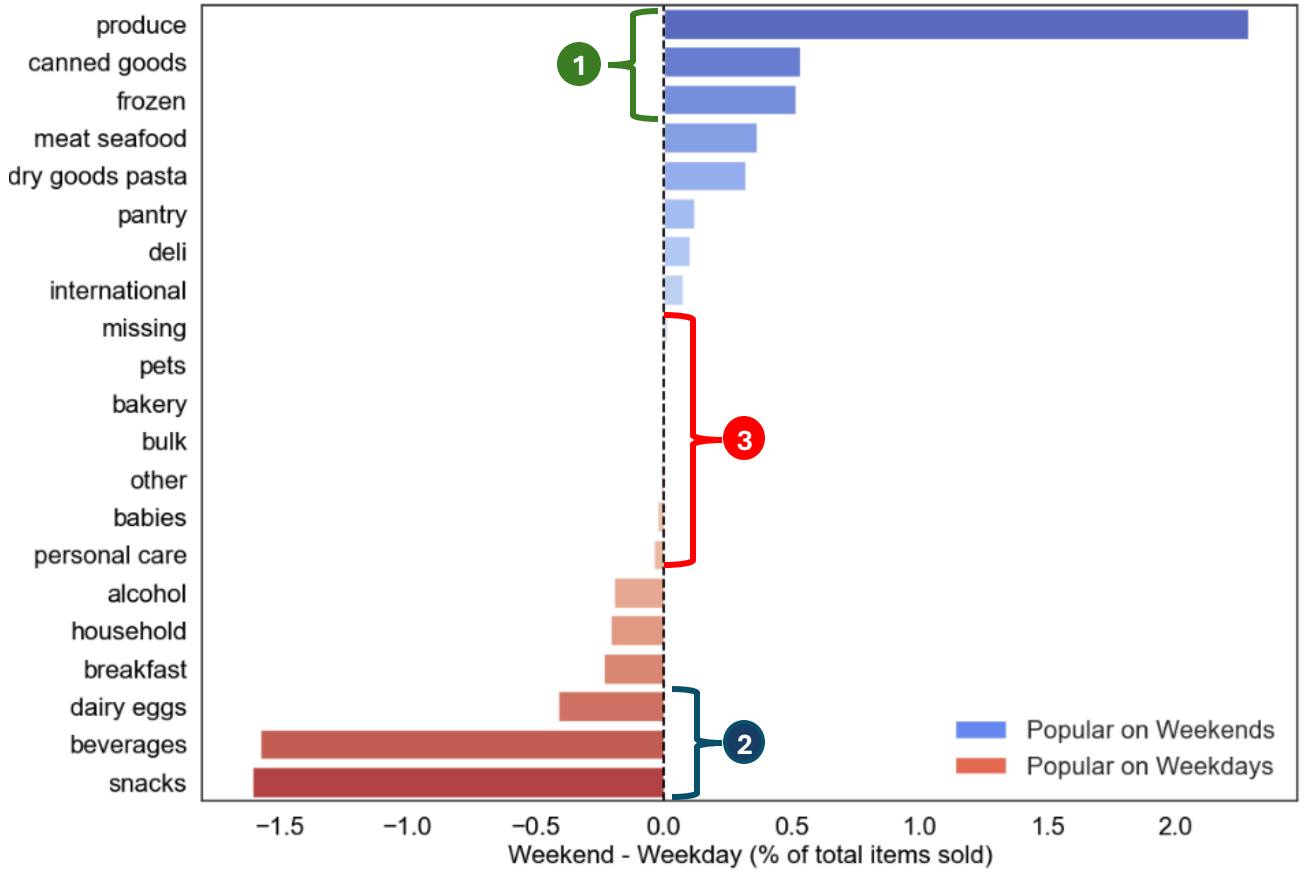
	Orders per User	Products per User	Products per Order	Days since Prior Order
				
Mean (std)	16.23 **(16.66)**	67.23 **(56.88)**	10.11 **(7.54)**	10.97 **(9.14)**
Min	3	1	1	0
25%	6	28	5	4
50%	**10**	**51**	**8**	**7**
75%	20	89	14	15
Max	100	728	145	30

When do customers purchase groceries?



1. Purchase activity is highest on Sundays and Mondays between 9AM and 5PM
2. Thursdays show the lowest order volume

Are certain product categories more popular on weekends vs. weekdays?



Insights:

1. Produce, canned goods, and frozen items are more common on weekends.
2. Beverages, snacks, and dairy items are more popular on weekdays.
3. Items in the middle are consistent through the week.

Customer Loyalty and Reorder Behavior

- **Loyalty Classification:**

  The Instacart customer base was segmented into 3 distinct loyalty groups:

  - **New** (less than 6 orders)
  - **Regular** (6–15 orders)
  - **Loyal** (16+ orders)

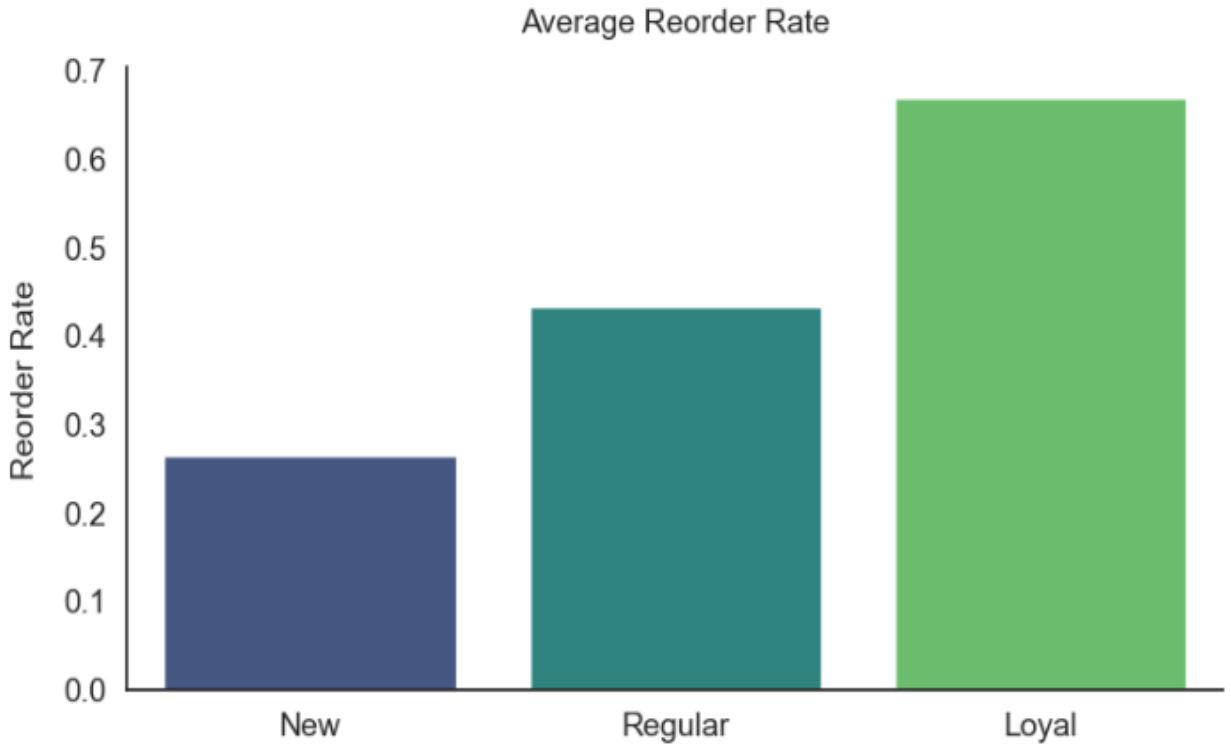
- **Similarity:**

  Basket Size: An initial analysis reveals that there is no significant difference in the average Instacart basket size between the 3 groups.

- **Difference:**

  Reorder Rate: Despite the similar basket size, as customers progress from New to Loyal, their propensity to place another order increases substantially, indicating that customer loyalty is a major driver of purchasing frequency.

Do new users have different buying patterns compared to loyal customers?



- The Reorder Rate increases as a customer shifts from New to Regular (28% to 44%)
- Loyal Customers are 2.4 times more likely to reorder than New Customers (68% vs 28%)
- No significant difference among the three groups in basket size

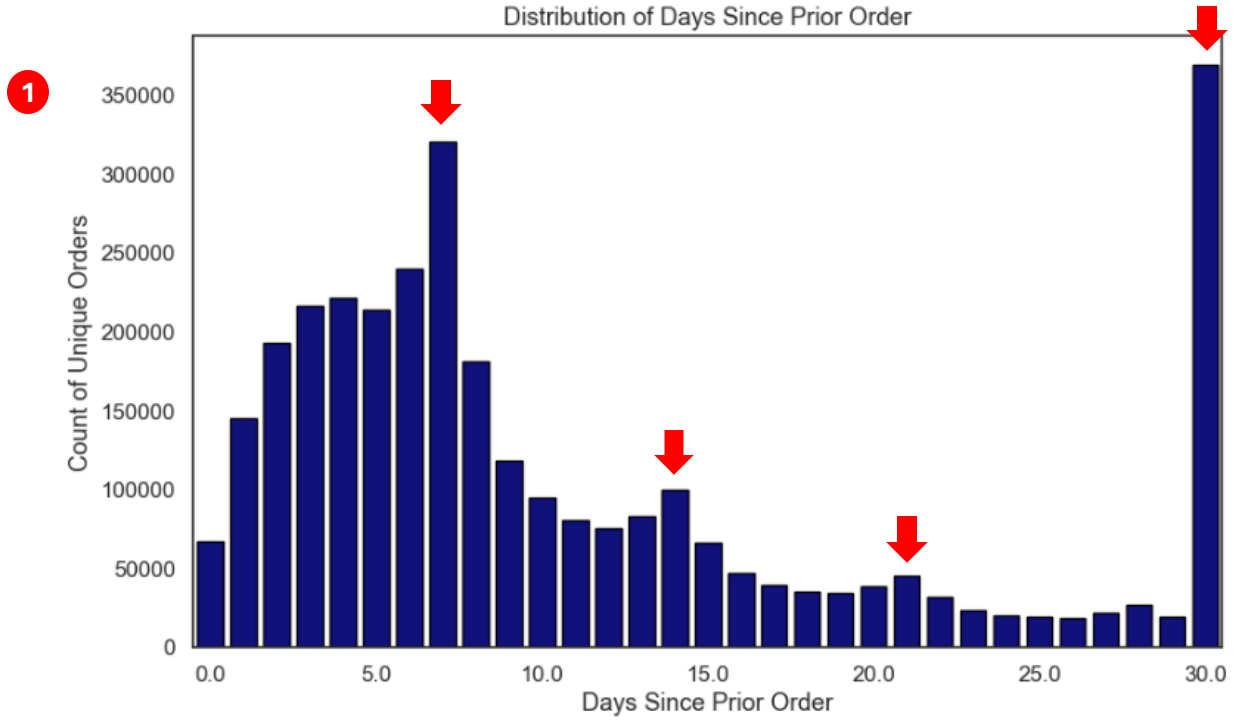
Customer Loyalty Group

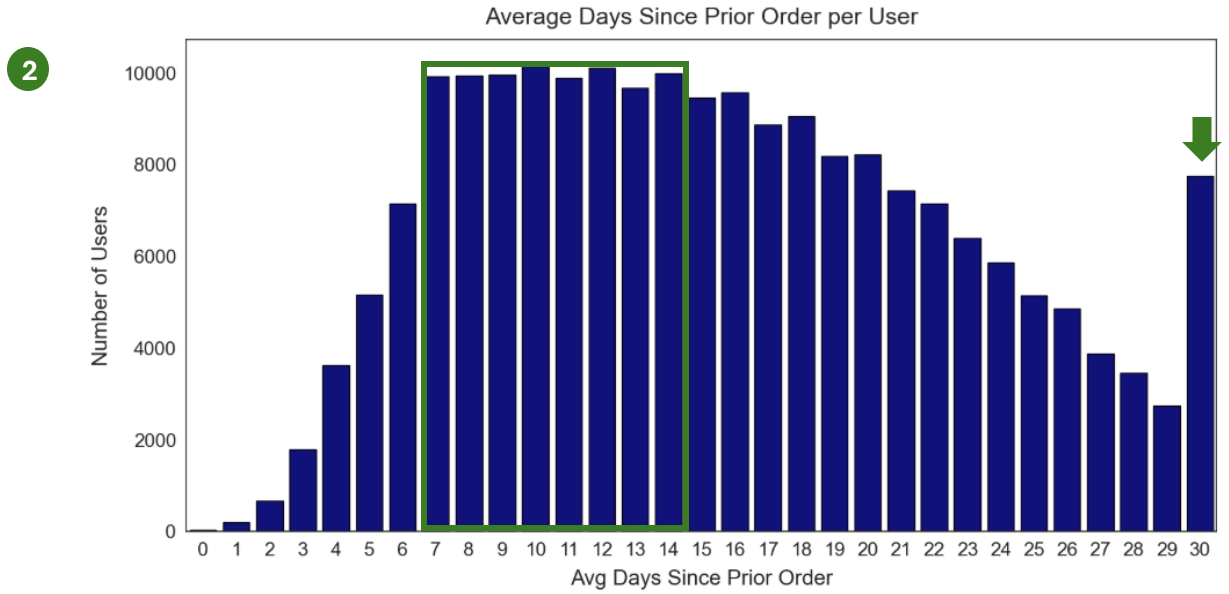
- New: <= 5 orders
- Regular: 6-15 orders
- Loyal: 16+ orders

How frequently do customers purchase groceries?

1. Order level:
   Weekly peaks are displayed.

2. Customer level:
   Avg is consistent from 7 to 14 days and then decreases.
   There are customers with a monthly trend.





Does the reorder cycle vary by department?

1. Products from babies and alcohol departments are more frequently ordered.
2. Household and pet products tend to be ordered less frequently.

department	Avg days between orders	Median days between orders
babies	9.3	7
alcohol	9.3	7
bulk	9.3	7
produce	10.2	7
pantry	10.3	7
other	10.3	7
snacks	10.4	7
dairy eggs	10.4	7
international	10.4	7
bakery	10.5	7
beverages	10.6	7
breakfast	10.6	7
meat seafood	10.7	7
deli	10.8	7
canned goods	10.9	8
personal care	10.9	7
dry goods pasta	11.1	8
frozen	11.1	8
household	11.4	8
pets	11.5	8

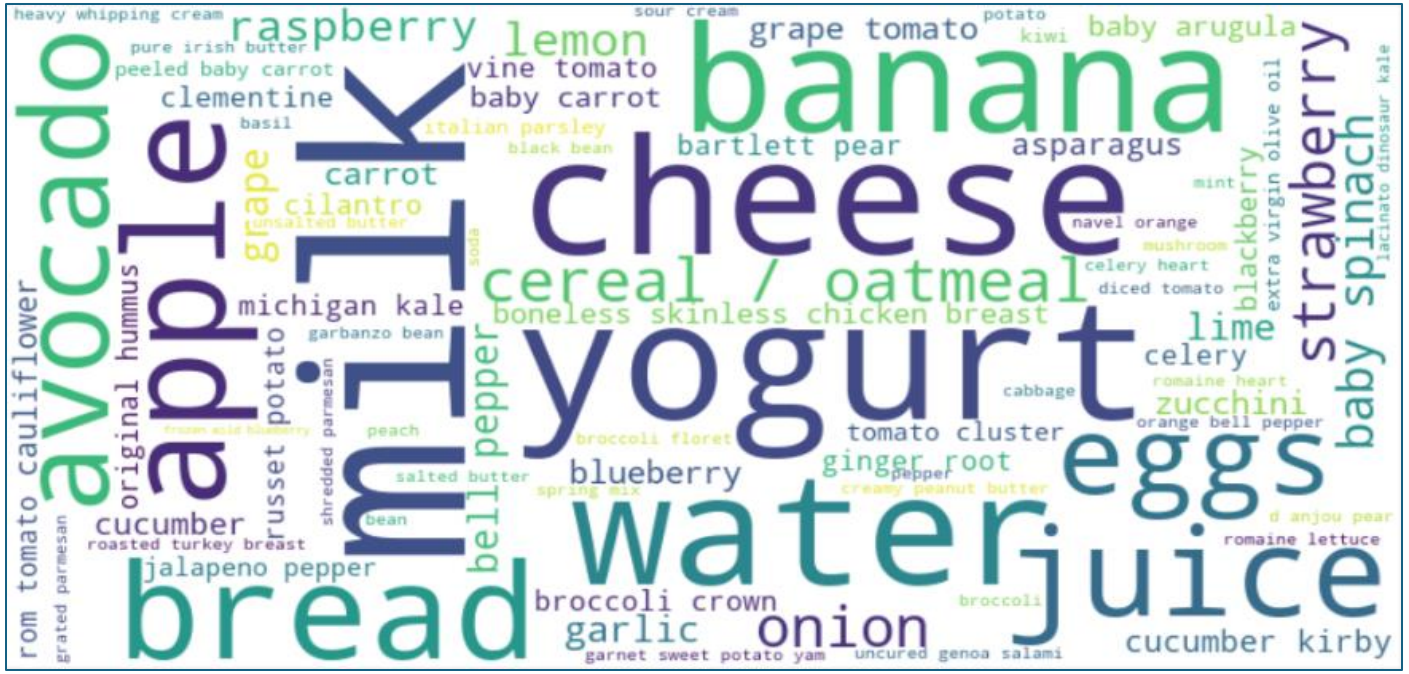
Top 100 Products by Order Volume

Sample of Top Ordered Products

Product	Order Volume	Reorder %
milk	1,997,589	73.9%
yogurt	1,164,719	68.4%
cheese	1,100,018	58.6%
water	1,010,320	72.5%
banana	913,028	83.4%
bread	727,266	64.8%
apple	723,459	69.6%
juice	614,502	61.4%
avocado	544,650	75.6%
eggs	480,035	70.2%

Total Orders

3,421,083



Key Highlights

- Product names were normalized by applying Natural Language Processing (NLP) techniques.
- Dairy and Fresh Produce products dominate overall order demand.
- Top categories like banana, avocado, and milk show high reorder rates, reflecting repeat consumption patterns.

Top 100 Products by Reorder Rate

Sample of Top Reorder-Rate Products

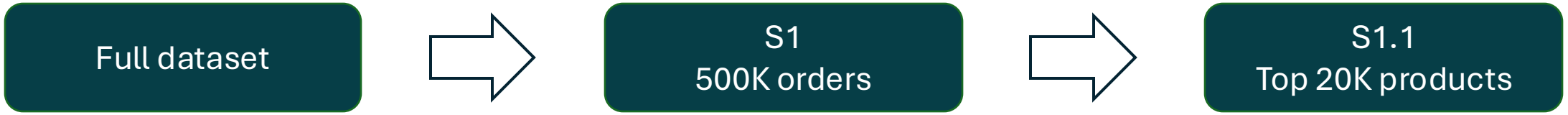
Product	Order Volume	Reorder %
serenity ultima...	90	93.3%
chocolate love ...	102	92.2%
simply sleep ni...	45	91.1%
bar peanut butt...	69	89.9%
soy crisp light...	67	89.6%
maca buttercup...	104	89.4%
benchbreak char...	111	89.2%
blueberry b meg...	99	88.9%
rare blended sc...	42	88.1%
fragrance free ...	131	87.0%

Total Orders
3,421,083

Key Highlights

- Products with the highest reorder rate are highly specific
- Despite having high reorder rates, these products show low order volume.

Data Sampling



Sample	Orders	Products	Users	Avg orders per user	Avg products per user	Avg products per order	Reorder rate	Max orders per user
Full	3,346,083	49,685	206,209	16.2	67.2	10.1	59.01%	100
S1	500,000	46,105	158,478	3.2	23.4	10.1	59.04%	31
S1.1	498,942	20,000	158,308	3.2	22.7	9.9	59.69%	31

1. Sample S1 covers around 93% of products and 77% of users
2. Sample S1.1 covers 99.8% of orders and users. Metrics are consistent across samples.
3. Average products per order and reorder rate are consistent across datasets.

Association Rule Mining

Discovering Hidden Connections

Itemsets

- Collection of items that appear together in the order
- Example:
  - (milk, bread)
  - (milk, eggs, cheese, banana)

Frequent Itemsets

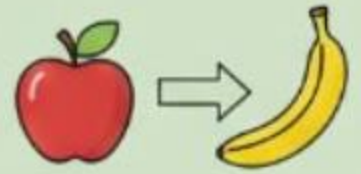
Itemsets with support above defined threshold.

Support = 2% (above Threshold = 0.5%)

Support = 0.4% (below Threshold = 0.5%)

Association Rules

Patterns showing how the presence of certain items predicts others.



Rule: Apple --> Banana

- **Support**: 1% (Both appear in 1% of orders)
- **Confidence**: Out of all orders containing apples, 38% also contain bananas.

Definitions adapted from Tan et al., 2021

Products Frequently Purchased Together

Parameters

- Transactions: 498,952
- Products: 20,000
- Thresholds used:
  - Min Support = 0.5%
  - Min Confidence = 30%

Results

- Total Frequent Itemsets: 76
- Max Itemset Length: 2
- Rules Generated: 6

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	1.68%
[Organic Baby Spinach, Banana]	8,013	1.61%
[Bag of Organic Bananas, Organic Baby Spinach]	7,923	1.59%
[Strawberries, Banana]	6,443	1.29%
[Banana, Large Lemon]	6,405	1.28%
[Bag of Organic Bananas, Organic Raspberries]	6,347	1.27%
[Organic Strawberries, Organic Hass Avocado]	6,181	1.24%

Association Rules

Rule	Support	Confidence	Lift
[Organic Fuji Apple] -> [Banana]	1.05%	38.01%	2.58
[Honeycrisp Apple] -> [Banana]	0.88%	35.53%	2.41
[Cucumber Kirby] -> [Banana]	0.98%	32.82%	2.23
[Organic Large Fuji Apple] -> [Bag of Organic Bananas]	0.75%	31.86%	2.72
[Organic Avocado] -> [Banana]	1.68%	30.44%	2.06
[Seedless Red Grapes] -> [Banana]	0.78%	30.31%	2.06

Key Highlights

- Most frequent associations occur within produce, especially fruits
- **Bananas** are a strong anchor product, appearing in all rules
- **Organic** shoppers show consistent patterns

Frequent Aisle Combinations

Parameters

- Transactions: 498,952
- Aisles: 132
- Thresholds used:
  - Min Support = 1%
  - Min Confidence = 30%

Results

- Frequent Itemsets: 2,704
- Max Itemset Length: 6
- Rules Generated: 6,517

Top Frequent Itemsets

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 fruits**]	93,657	18.77%
[fresh **fruits**, yogurt]	93,333	18.71%
[fresh **fruits**, milk]	81,894	16.41%
[packaged cheese, fresh **fruits**]	76,629	15.36%
[fresh **vegetables**, yogurt]	71,881	14.41%
[packaged cheese, fresh **vegetables**]	66,813	13.39%
[yogurt, **packaged vegetables fruits**]	63,695	12.77%

Frequent Itemsets Stats by Length

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%**

Key Highlights

- Short itemsets (2 & 3 aisles) are most frequent and most impactful
- Max support drops sharply from 31.85% (Len=2) to 18.77% (Len=3) and further for longer sets

Rules Generated for Aisle Combinations

Total Rules Generated: 6,517

Top Rules	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

Key Highlights

- **Fresh vegetables** is the strongest anchor aisle in multi-aisle purchasing patterns.
- Rules involving **fresh herbs**, **fresh fruits**, and **packaged vegetable fruits** show the highest confidence (>92%), indicating very reliable co-purchasing behavior.
- All top rules have a lift > 2.0, indicating strong relationship among these aisles.

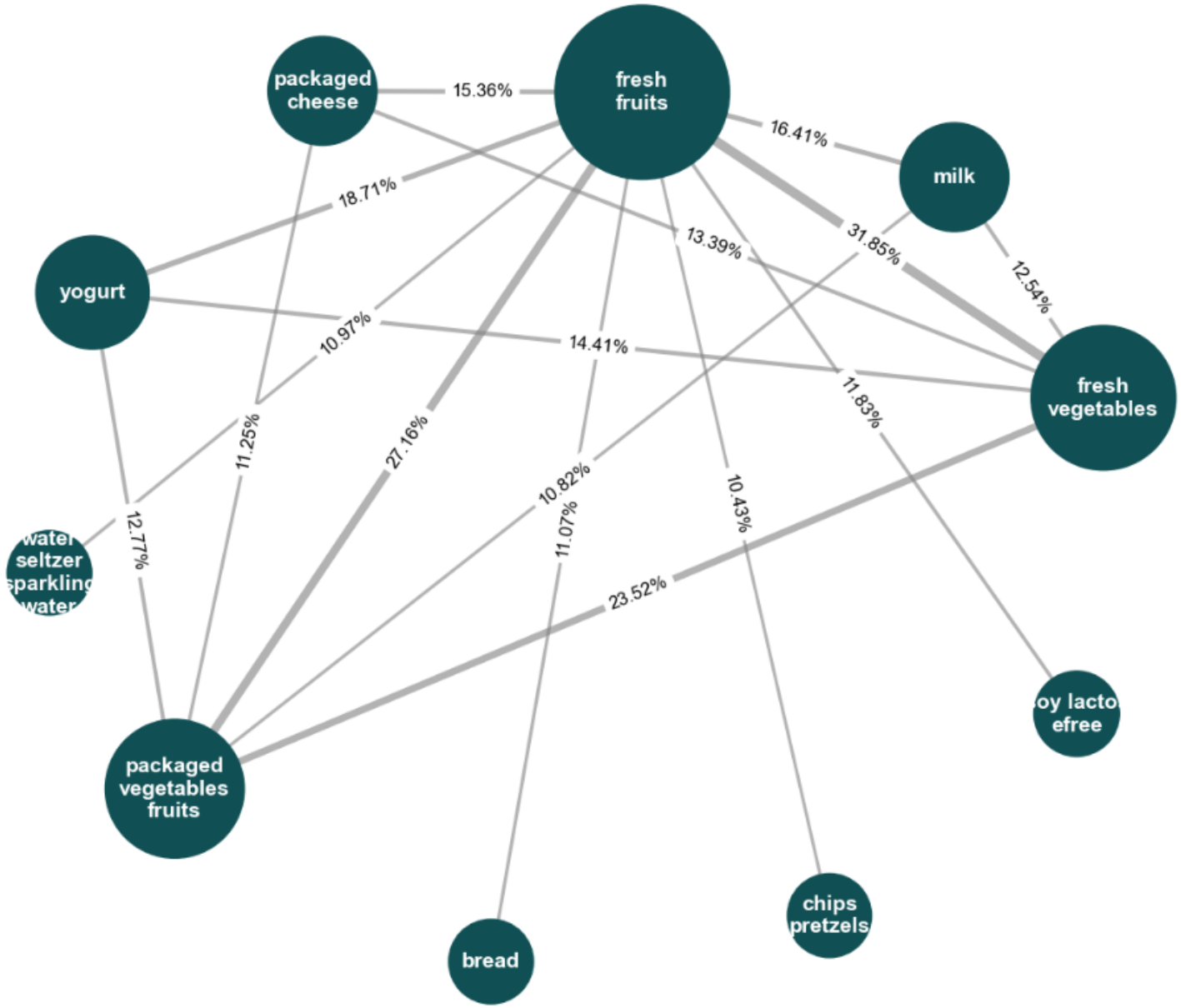
Aisle Network Graph

Parameters

- Transactions: 498,952
- Aisles: 132
- Thresholds used:
  - Min Support = 1%
  - Min Confidence = 30%

Results

- Frequent Itemsets): 2,704
- Max Itemset Length: 6
- Rules Generated: 6,517



Frequent Department Combinations

Parameters

- Transactions: 498,952
- Departments: 21
- Thresholds used:
  - Min Support = 1%
  - Min Confidence = 30%

Results

- Frequent Itemsets: 2,644
- Max Itemset Length: 8
- Rules Generated: 27,262

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%

Frequent Itemsets Stats by Length

Len	Itemsets	Min Support	Max Support	Median Support	Std Dev
2	123	8.32%	55.07%	11.69%	7.01%
3	422	3.41%	8.29%	4.70%	1.32%
4	765	1.79%	3.41%	2.35%	0.45%
5	784	1.23%	1.79%	1.46%	0.16%
6	441	1.04%	1.23%	1.12%	0.05%
7	106	1.00%	1.04%	1.02%	0.01%
8	3	1.00%	1.00%	1.00%	0.00%
Total	**2,644**	**1.00%**	**55.07%**	**1.78%**	**3.24%**

Key Highlights

- **Produce** and **dairy eggs** are present in the strongest department combinations

Rules Generated for Department Combinations

Total Rules Generated: 27,262

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%	96.06%	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

Key Highlights

- **Produce** and **dairy eggs** are the strongest anchor departments in multi-department purchasing patterns.
- Rules involving **snacks**, **deli**, and **produce** show the highest confidence (>96%), indicating very reliable co-purchasing behavior.

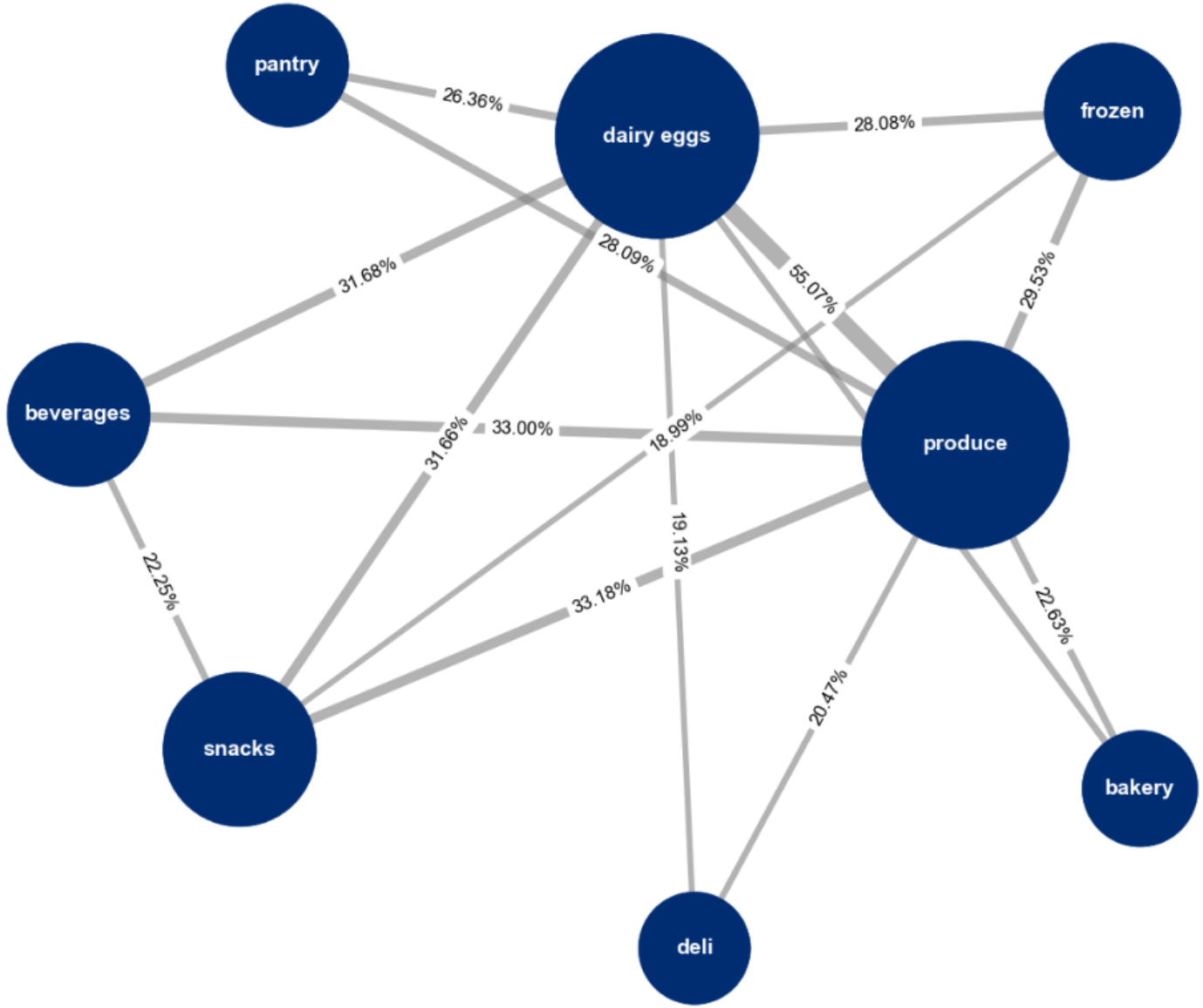
Department Network Graph

Parameters

- Transactions: 498,952
- Departments: 21
- Thresholds used:
  - Min Support = 1%
  - Min Confidence = 30%

Results

- Frequent Itemsets: 2,644
- Max Itemset Length: 8
- Rules Generated: 27,262



Customer Segmentation
Reorder-Driven Patterns

Data Preparation

- Reorder Rate per User
- User-Department matrix
  - Rows = Users
  - Columns = Departments
  - Values = Reorder Rate

Clustering

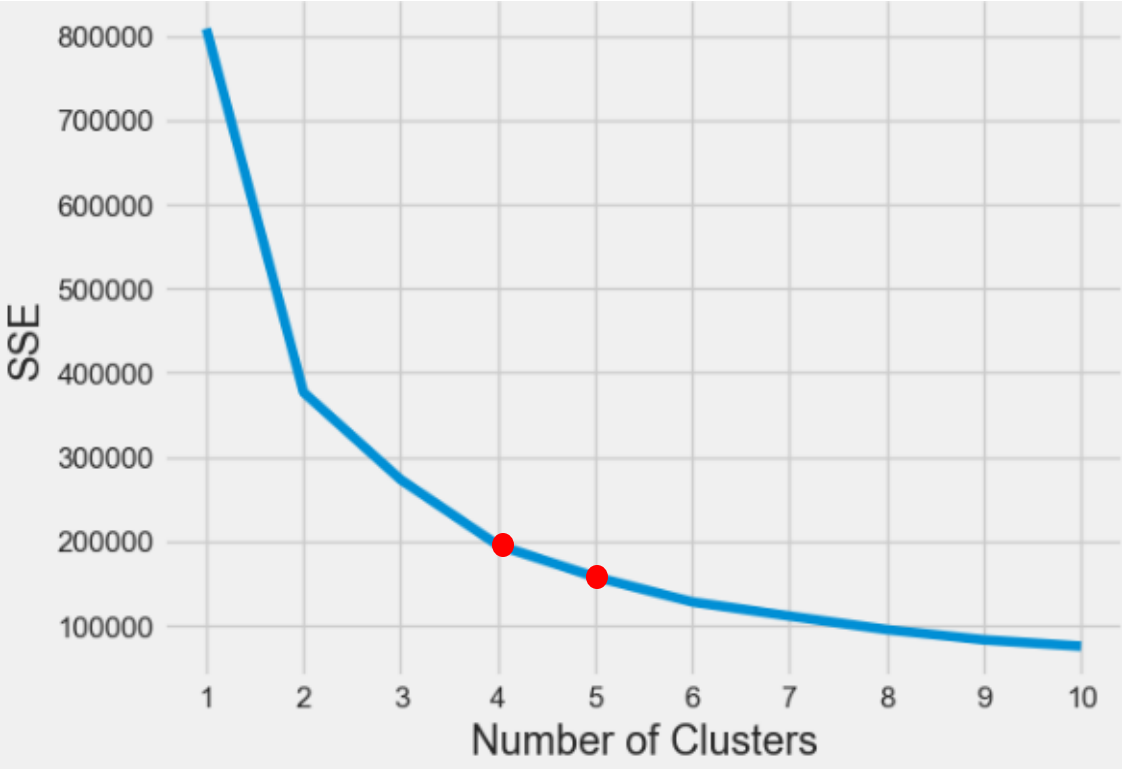
- Cluster users with K-Means
- Reduce Dimensions with PCA
- Visualize Segments and Centroids

Segment Analysis

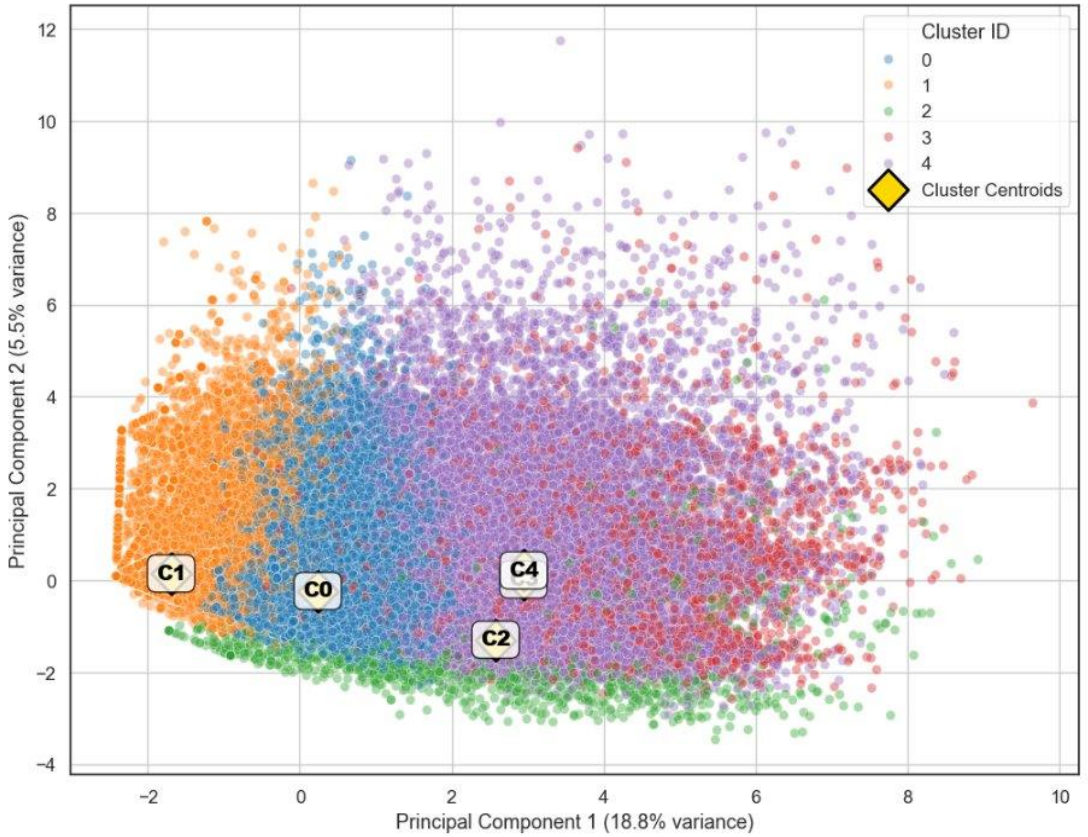
- Summarize Each Segment
- Provide insights

Clustering

Elbow Method



Principal Component Analysis (PCA)



Customer Segmentation based on Reorder Rate

Segment	Key Traits
1	Core grocery essentials
2	Infrequent buyers
3	Bulk/stock-up shoppers
4	Family-oriented
5	Premium loyalists

Department	Segment				
	1	2	3	4	5
frozen	29.0	7.6	44.5	50.9	57.3
other	0.5	0.2	2.0	1.9	2.1
bakery	25.0	5.1	42.5	48.2	55.8
produce	61.5	18.3	70.5	67.1	70.5
alcohol	1.4	1.9	0.5	1.4	2.8
international	2.1	1.0	13.4	11.4	17.8
beverages	42.5	15.5	49.6	54.4	60.8
pets	0.8	0.7	1.2	2.3	5.0
dry goods pasta	8.0	2.7	28.5	35.9	44.6
bulk	0.0	0.0	93.9	0.2	0.0
pantry	15.7	4.2	33.4	32.6	44.5
breakfast	12.2	3.1	29.4	34.1	39.8
canned goods	11.5	3.5	32.9	35.6	44.4
dairy eggs	64.6	12.9	69.0	70.8	73.5
household	6.2	3.4	15.1	22.0	25.5
babies	0.5	0.3	9.0	88.3	0.7
snacks	34.8	8.6	53.5	52.8	59.3
deli	22.0	4.4	40.6	42.6	50.7
missing	1.3	0.4	3.7	3.5	3.7
Total Users	**56,642**	**67,214**	**1,799**	**6,748**	**25,905**

Segment Traits and Opportunities

Segment	Key Traits	Size	Reorder Behavior	Opportunity
1	Core Grocery essentials	56K	Medium	Increase category cross-selling
2	Infrequent buyers	67K	Low	Retention & activation campaigns
3	Bulk/stock-up shoppers	1.8K	Very High	Loyalty & subscription offers
4	Family-oriented	6.7K	High in household and baby	Family bundles, household loyalty
5	Premium loyalists	26K	High and broad	VIP rewards, early access

Products Frequently Purchased Together by Segment 4

Parameters

- Transactions: 38,572
- Products: 17,889
- Thresholds used:
  - Min Support = 0.5%
  - Min Confidence = 30%

Results

- Total Frequent Itemsets: 347
- Max Itemset Length: 3
- Rules Generated: 103

Top 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	3.60%
[Organic **Strawberries**, Organic Hass **Avocado**]	1,283	3.33%
[Bag of Organic Bananas, Organic Baby Spinach]	1,264	3.28%
[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%

Top Association Rules

Rule	Support	Confidence	Lift
[Organic **Raspberries**, Organic Hass **Avocado**] -> [Bag of Organic Bananas]	1.11%	53.09%	2.64
[Organic Baby Spinach, Organic Hass **Avocado**] -> [Bag of Organic Bananas]	1.05%	47.04%	2.34
[Organic **Strawberries**, Organic Hass **Avocado**] -> [Bag of Organic Bananas]	1.41%	42.40%	2.11
[Organic D'Anjou Pears] -> [Bag of Organic Bananas]	1.49%	41.34%	2.05
[Cucumber Kirby] -> [Bananas]	1.25%	40.85%	2.08
[Organic Large Extra Fancy Fuji Apple] -> [Bag of Organic Bananas]	1.85%	40.74%	2.02

Key Highlights

- There a significantly stronger co-purchase behavior vs overall population, +2.9 pp higher.
- This segment display more complex frequent itemsets and stronger rules.
- These customers prefer premium, organic produce.

Aisle-to-Aisle Purchase Sequences (Segment 4)

Parameters

- Transactions: 38,572
- Aisles: 132

Results

- Aisle Transition Patterns: 16,599

Top Aisle Sequences

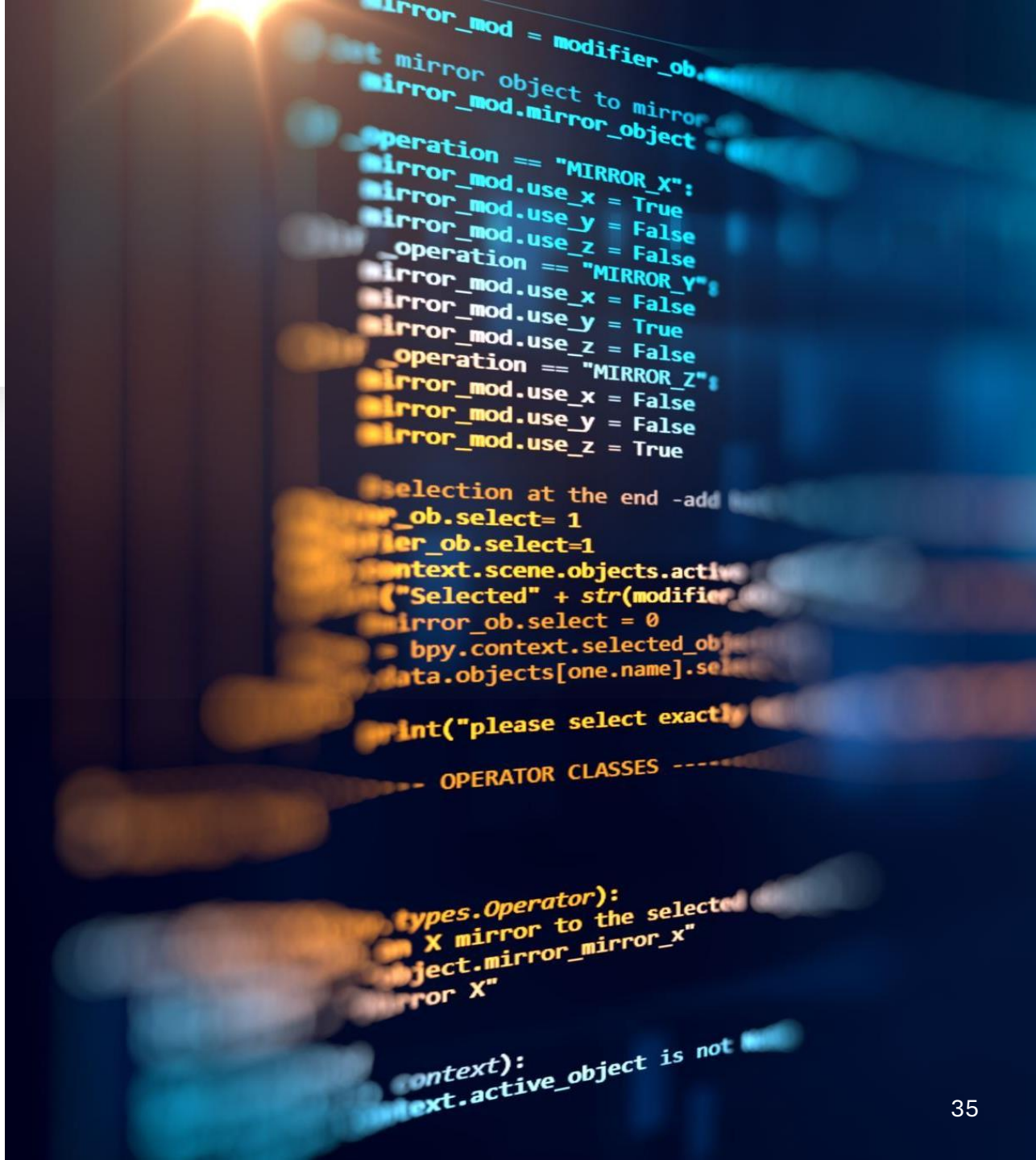
Transition (Aisle A → Aisle B)	Support	%
fresh vegetables --> fresh fruits	13,231	34.30%
fresh fruits --> fresh vegetables	13,122	34.02%
fresh fruits --> packaged vegetables fruits	11,391	29.53%
packaged vegetables fruits --> fresh fruits	11,377	29.50%
yogurt --> fresh fruits	9,852	25.54%
fresh fruits --> yogurt	9,821	25.46%
milk --> fresh fruits	9,276	24.05%
fresh fruits --> milk	9,230	23.93%
fresh vegetables --> packaged vegetables fruits	8,904	23.08%
packaged vegetables fruits --> fresh vegetables	8,863	22.98%

Key Highlights

- Customers who purchase from **fresh vegetables** have a 34.3% probability of purchasing from **fresh fruits** in their next order, the strongest transition observed.
- Produce categories dominate the top transitions, indicating habitual replenishment behavior.
- Dairy items such as milk and yogurt also show strong behavioral ties to produce categories.

Challenges

- Data Size and Computational Resources
- Data Sparsity
- Interpretability and Actionability of Insights
- Dynamic Nature of Consumer Behavior



Key Takeaways

Understanding Shopping Behavior: Data-driven insights reveal how, when, and what customers purchase.

Actionable Business Strategies: Market basket analysis enables personalized recommendations, targeted marketing, and inventory optimization.

Power of Data Science: Techniques like EDA, association rule mining, and clustering help transform raw data into valuable business decisions.

Future Opportunities: Insights from Instacart data can be extended to dynamic pricing, demand forecasting, and real-time recommendation systems.

References

Yasser H. (2021). InstaCart Online Grocery Basket Analysis Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/yasserh/instacart-online-grocery-basket-analysis-dataset

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22(2), 207–216. https://doi.org/10.1145/170036.170072

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). Introduction to Data Mining (2nd ed.). Pearson.

Herdiana, I., Kamal, M. A., Triyani, M. N. E., & Renny. (2025). A more precise elbow method for optimum K-means clustering. arXiv. https://arxiv.org/abs/2502.00851

scikit-learn. (2010). 2.3. Clustering — scikit-learn 0.20.3 documentation. Scikit-Learn.org. https://scikit-learn.org/stable/modules/clustering.html