Rotman

Master of
Management
Analytics

OPTIMIZING THE 'INSTABASKET' AISLE:
A DATA-DRIVEN APPROACH

Harnessing Data to Elevate MM&A's Customer Experience

Team 3:

- Aijia Yuan
- Billy Luong
- Julian Oppedisano
- Mohamed Emran

October 6, 2023

Rotman School of Management
UNIVERSITY OF TORONTO

Agenda

- **Background**
  - Business Objectives & Constraints
  - Defining Success: Metrics Overview
- **Analysis**
  - Data Sources & Key data Tendencies
  - Product Selection
  - Product Substitution
- **Conclusion**
  - Business Implication
  - Future Work
- **Acknowledgements**
- **Q&A**

Background

Business Objectives & Constraints

Main Objective

Enhancing the MM&A Experience: Responding to the high demand from Instabasket personal shoppers by populating a specialized aisle, optimizing both product selection and substitution based on data-driven insights.

Two Specific Aims

Product Configuration

Purpose: Populate the "Instabasket" aisle with top-performing products.

Based On: Historical purchasing data, focusing on products that have shown consistent high demand.

Efficient Substitution

Purpose: Seamlessly replace out-of-stock or unavailable items without disrupting the shopping experience.

Strategy: Identify and group products that can be appropriate substitutes for popular items, thus minimizing order disruptions.

Constraints

Product Capacity:

- Total products: 1000 max.
- Refrigerated items: 100 max.
- Frozen items: 100 max.

Substitution Parameters : Unrestricted substitution recommendations per product. However, the quality and relevance of substitution are considered.

Background

Defining Success: Metrics Overview

Metric #1: Product & Substitution Utilization:

Product Definition: Percentage of orders containing products from the Instabasket aisle.

Substitution Definition: Percentage of orders with at least one item from the recommended substitutes.

Target: A high percentage in both metrics signals the effectiveness and relevance of the Instabasket aisle and its substitute recommendations. ↑

Metric #2: Product & Substitution Aisle Utilization:

Product Definition: Average percentage of items in each order sourced from the Instabasket aisle.

Substitution Definition: Average percentage of substituted products in each order.

Target: Achieve a high percentage for both, reflecting the importance of aisle products and the acceptance of substitute suggestions. ↑

Metric #3: Store Flow Efficiency

Definition: Average number of other aisles visited outside the specialized aisle.

Target: Minimize, ensuring a seamless shopping experience and reduced disruptions. ↓

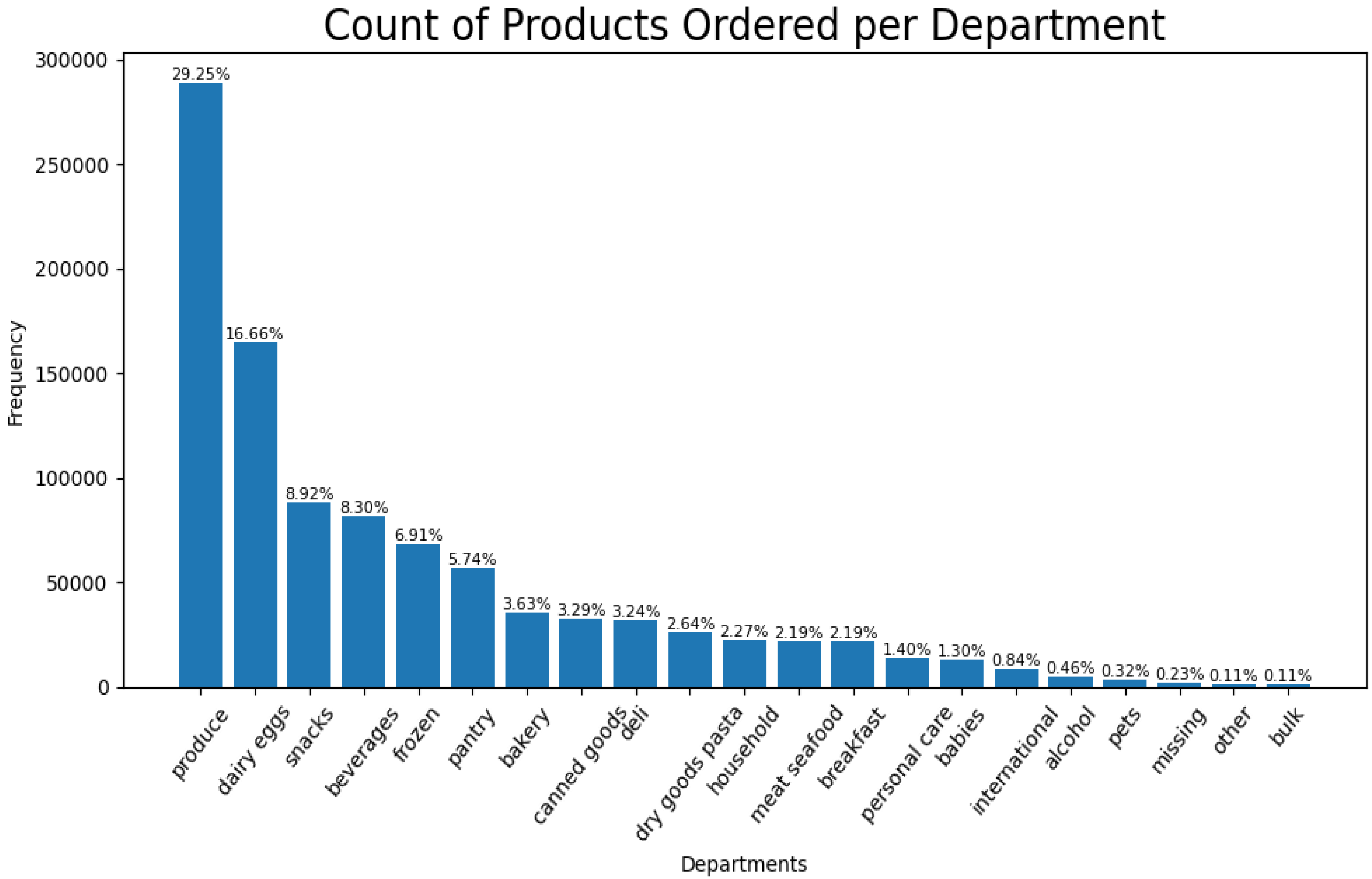
Analysis

Data Source & Key Data Tendencies

- Dataset used was from "[The Instacart Online Grocery Shopping Dataset]"
- The data collected was from 2017

- The dataset encompasses **987,000** individual records, representing about **98,000** distinct customer orders.
- Data spans **21** departments and **134** distinct product aisles
- Each order contains an average of **10** products

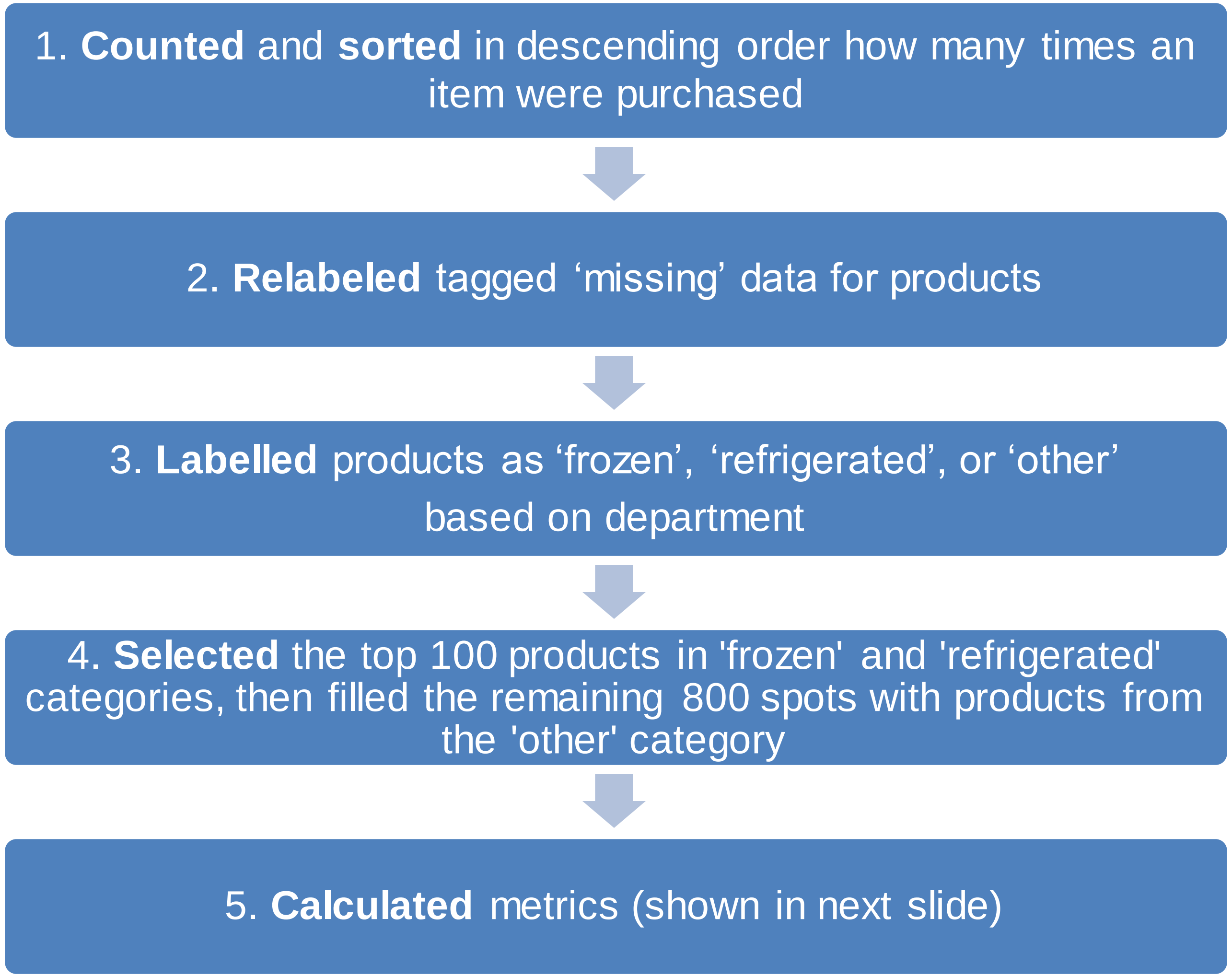
- Over **50%** of products ordered were from just 3 departments: Produce, Dairy & Eggs, and Snacks.
- The dataset contains some incomplete entries, notably with 'missing' department labels for certain products.



Count of Products Ordered per Department

Analysis

Product Selection: Methodology



1. **Counted** and **sorted** in descending order how many times an item were purchased

2. **Relabeled** tagged ‘missing’ data for products

3. **Labelled** products as ‘frozen’, ‘refrigerated’, or ‘other’ based on department

4. **Selected** the top 100 products in 'frozen' and 'refrigerated' categories, then filled the remaining 800 spots with products from the 'other' category

5. **Calculated** metrics (shown in next slide)

Key Considerations

- Our method was designed to maximize the utilization of the Instabasket aisle spaces allocated for 'frozen' and 'refrigerated' products.
- Products were determined as ‘frozen’, ‘refrigerated’, or ‘other’ (in-aisle shelves) based on department
- Some products were listed as missing or others which needed to be assigned to a more descriptive aisle before determining allocation to in-aisle, fridge, or freezer space
- Manually performed as there were few products that were highly purchased as will be shown below

Analysis

Product Selection: Model Accuracy & Expected Performance

Metric	Target*	Result
Orders that utilize the in-aisle items	97833	89846 **(91.84%)**
Average number of items in each order that utilize in-aisle items	10.09	5.23 **(51.86%)**
Median number of items in each order that utilize in-aisle items	6.0	4.0 **(66.67%)**

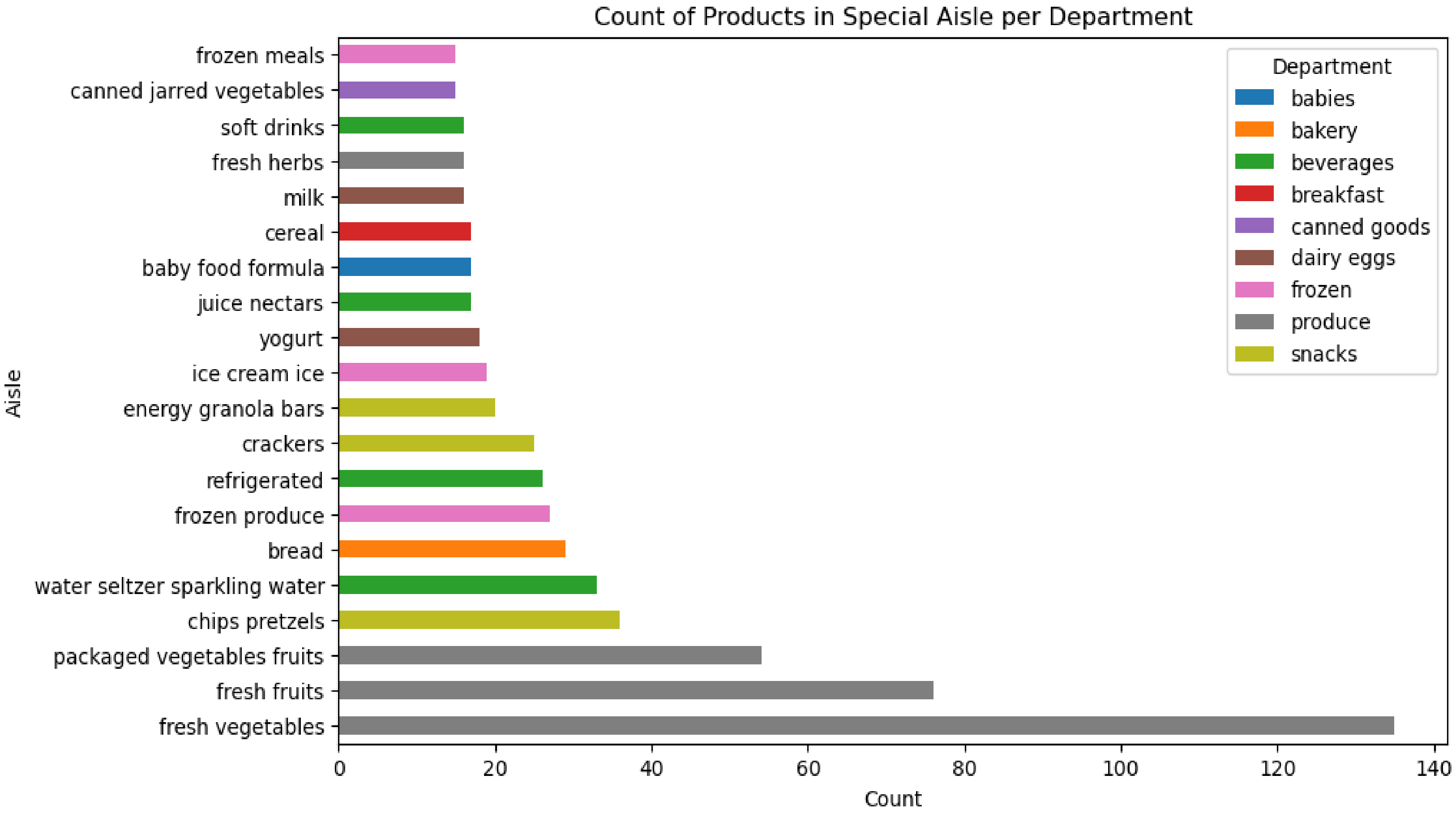
Note:

- We decided not to use a train-test split since the task isn't to build a predictive model but to identify a trend for the most frequently ordered products, so the entire dataset was used for the metrics.
- The benchmark for evaluating target outcomes is based on metrics assuming the absence of an InstaBasket aisle.

Analysis

Product Selection: Key insights

- **Produce Dominance:** The 'produce' department, encompassing fresh fruits and vegetables, significantly outnumbers other categories, emphasizing its role in customer preference.
- **Change in Dairy & Eggs**: Despite dairy and eggs showing popularity among shoppers, their prominence diminishes after filtering indicating a wide variety but fewer high-frequency items in this department.



Rotman

Analysis

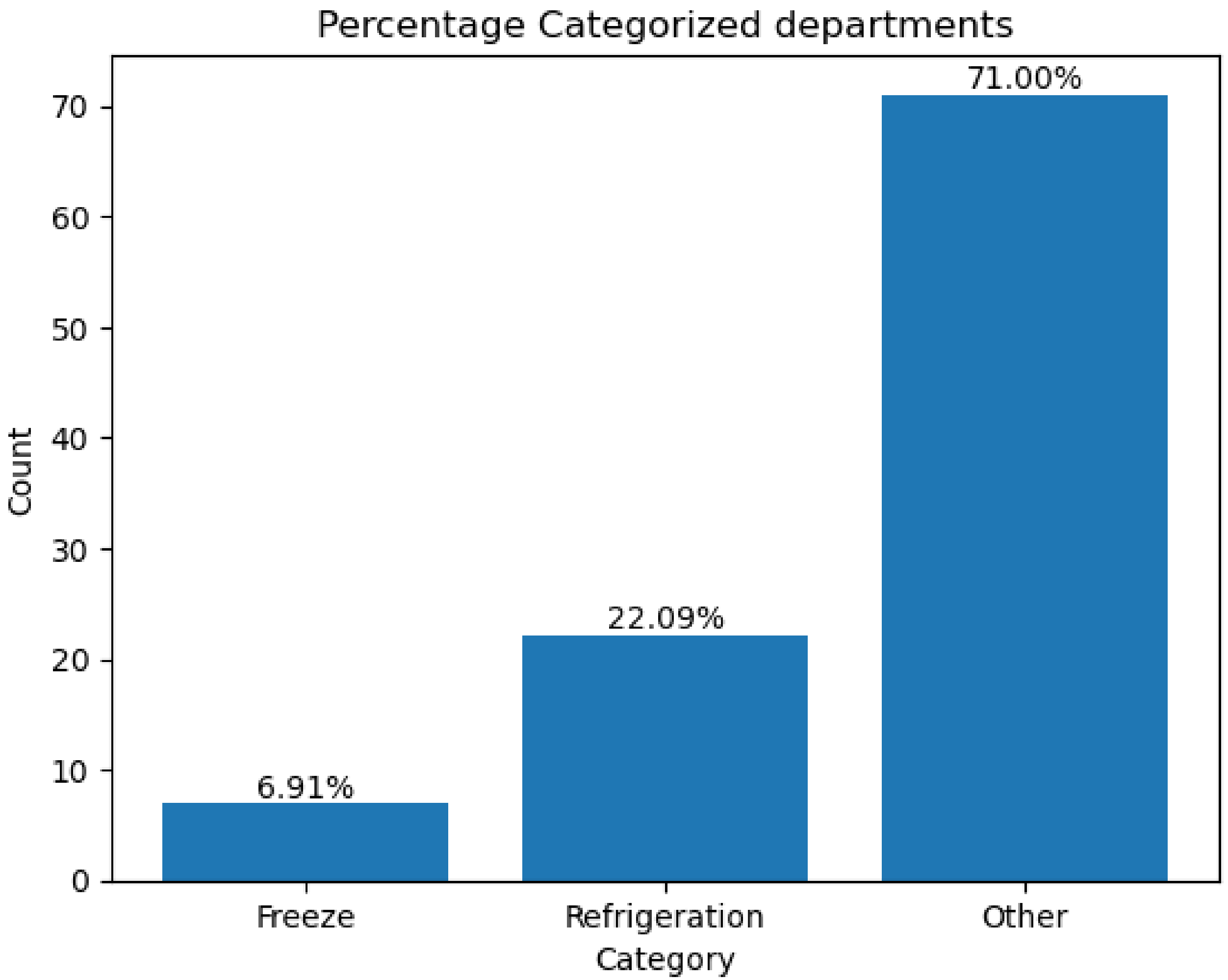
Product Selection: Limitations and Assumptions

- Given the popularity of '**Produce**' and limited refrigeration space, we decided NOT to include refrigeration categorization and classify as 'Other'.

- **Refrigeration** categorization is defined as products from the dairy eggs, meat seafood, or deli department

- **Frozen** categorization is defined as products from the frozen department

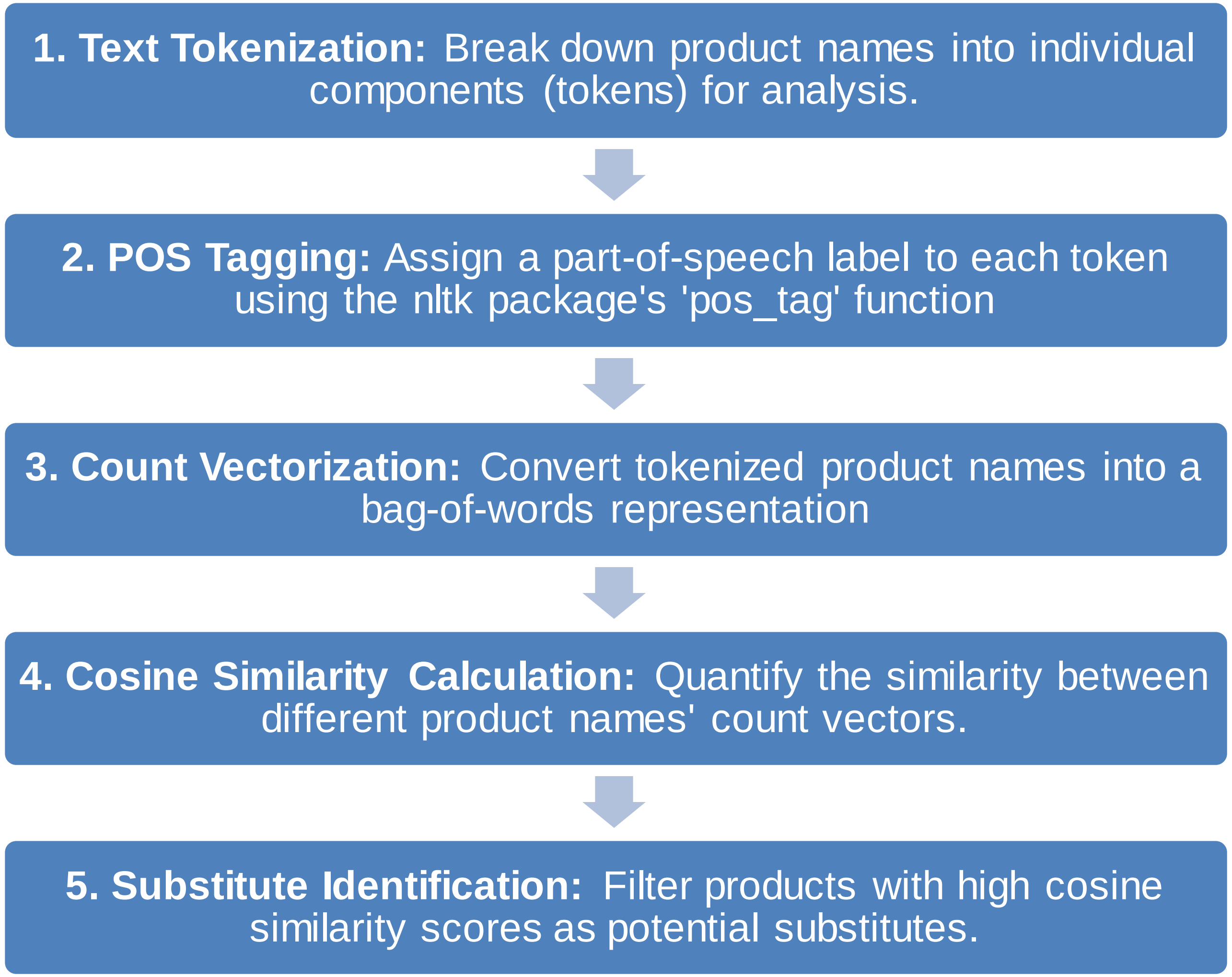
- All other products were categorized as '**Other**'



Rotman

Analysis

Product Substitution: Methodology - Natural Language Processing (NLP)



1. Text Tokenization: Break down product names into individual components (tokens) for analysis.

2. POS Tagging: Assign a part-of-speech label to each token using the nltk package's 'pos_tag' function

3. Count Vectorization: Convert tokenized product names into a bag-of-words representation

4. Cosine Similarity Calculation: Quantify the similarity between different product names' count vectors.

5. Substitute Identification: Filter products with high cosine similarity scores as potential substitutes.

Key Considerations:

- **Preprocessing Filters:** Prior to word matching, excluded meaningless words, removed adjectives, omitted words with fewer than 2 characters, and converted plural nouns to singular form.
- **Quality Assurance:** Implement multiple filters to remove low similarity scores and to restrict the number of potential substitutes, to enhance the relevancy and accuracy of suggestions.
- **Accuracy Metrics:** Utilized a test-train split as well as a Precision@K metric to measure the accuracy of recommended product substitutes.

Analysis

Product Substitution: Model Accuracy & Expected Performance

Metric	Target	Result
Orders that utilize the in-aisle items	97833	92217 **(94.26%)**
Average number of items in each order that utilize in-aisle items	10.09	6.02 **(59.63%)**
Median number of items in each order that utilize in-aisle items	6.0	4.32 **(72%)**
Number of Product Substitution Recommendations	N/A	2745

Analysis

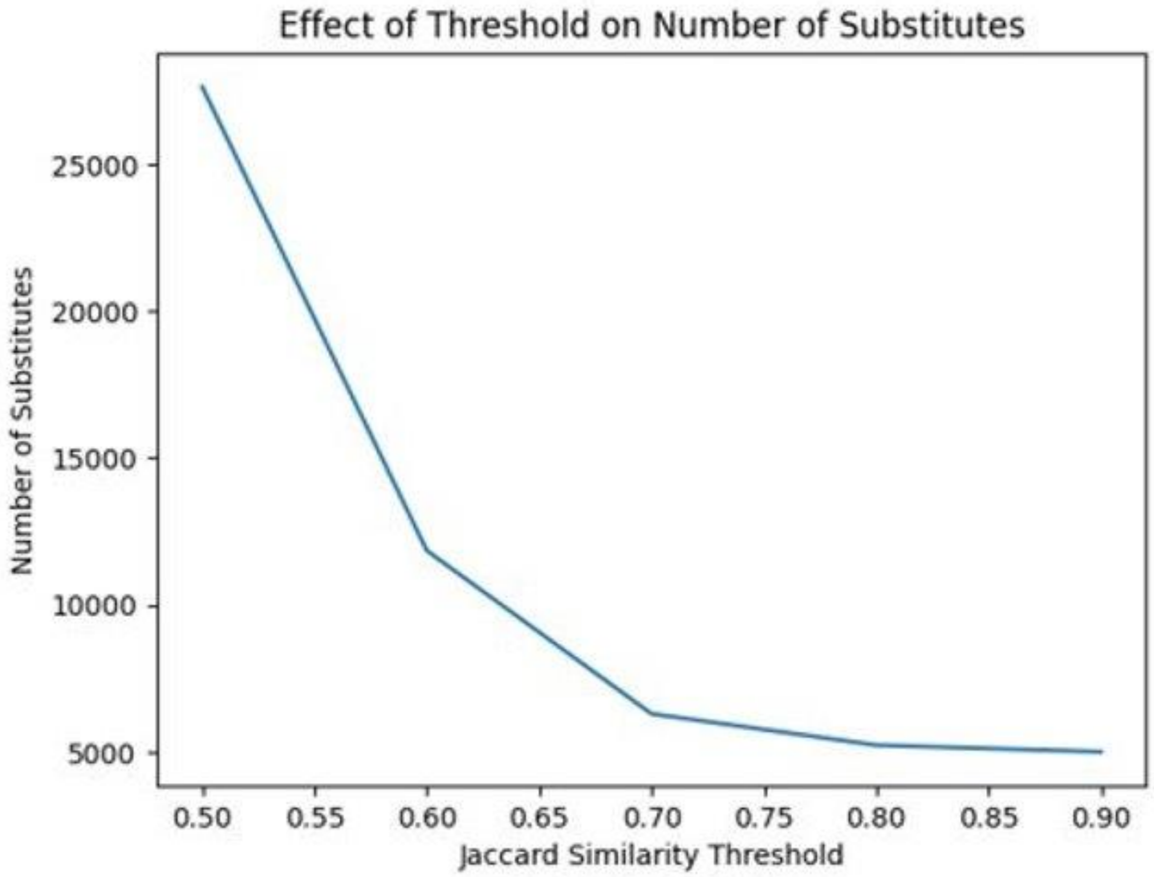
Product Substitution: Accuracy of Substitution Recommendations using Precision@K

Balancing Precision with Product Coverage

- Precision@K is a method to validate the accuracy of our substitution recommendations by comparing them against actual purchase patterns.
- By focusing on substitutes above the median threshold, our model suggests only top-tier substitutes.
- This strategy enhanced our Precision@K score, indicating a closer match with customer preferences, but reduced the overall number of product recommendations, as illustrated in the accompanying chart.

Achieved Precision@K:

- **Performance Data: 38.04%**
- **Interpretation:** Out of the product substitutes our model recommended, approximately 38.04% were actually bought together with the target product in real orders.



Analysis

Product Substitution: Assumption and Limitations

Assumptions

- **Acceptance of Substitutes:** The model presumes a full acceptance (100%) of all substitutions. Actual acceptance rates may differ, thus affecting the accuracy of the substitution recommendations.
- **Descriptiveness of Product Names:** Product names contain enough words to describe their nature and usage. Generic or vague product names affect the accuracy of substitutes.
- **Similarity Basis:** Products with similar names are assumed to be good substitutes. There is no consideration for brand loyalty.
- **Categorization Decision:** 'Produce' was classified as 'other' due to its popularity and restricted refrigeration space.

Limitations

- **Word Ordering:** Our model disregards the sequence in which words appear. Context is lost when the words are broken up.
- **Static Nature of the Model:** The NLP method misses nuances in customer buying behaviors, potentially leading to less accurate product suggestions.
- **Independence of Words:** The bag-of-words (BoW) model overlooks the interplay of words within product names.
- **Ignorance of Discounts:** The model did not focus on revenue, so the 5% discount offered for accepting substitutes wasn't considered. This may affect the actual impact of product substitutes.

Analysis

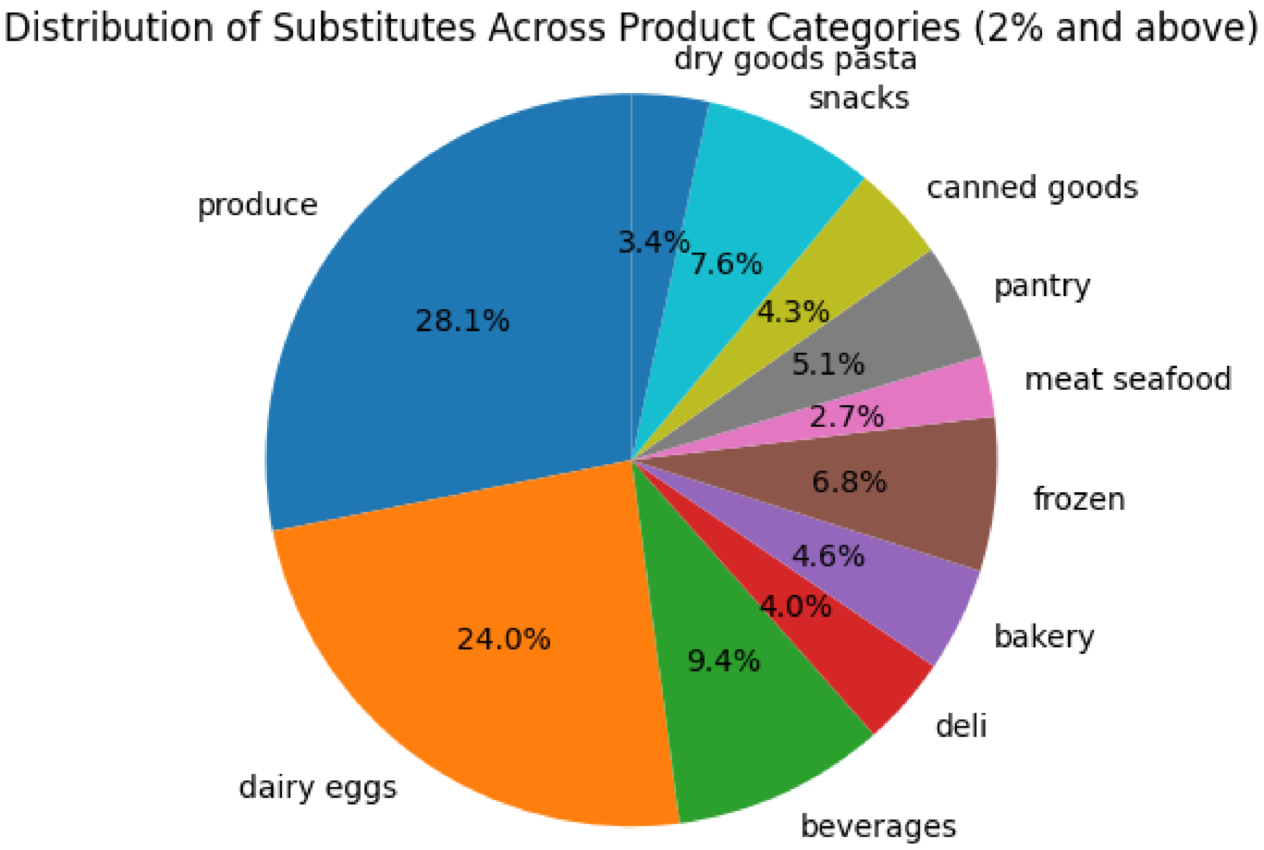
Product Substitution: Key Insights

- **Produce & Dairy Dominance**: Our model highlighted that a significant portion of substitutions pertained to the 'produce' (28.1%) and 'dairy & eggs' (24.0%) categories.

- **Tailored Recommendations:** Our model's focus on the descriptiveness of product names resulted in a nuanced distribution of substitutions across categories.

- **Substitution Opportunities:** Categories like 'meat seafood' and 'frozen' have a lower percentage of substitutions. These might represent areas where product naming might not be as descriptive, or where there are fewer available substitutes.

- **Revenue Considerations:** Categories with high substitution rates might benefit from targeted discounts or promotions.



Rotman

Conclusion

Future Work

Data Expansion: Integrate newer and more comprehensive datasets to capture changing shopping patterns.

Product Tagging: Implement auto-categorization of products to speed up selection and remove manual checks.

Enhanced Substitution Model Integration: Combine NLP and behavioural data for a holistic recommendation system. This could improve accuracy in product recommendations and lead to an increase in sales.

User Feedback Integration: Incorporate direct feedback from users and personal shoppers to get real-time enhancement of product recommendations and personalize shopping experiences.

Other Substitution Approaches Explored

Item-to-Item Based Collaborative Filtering:

- Uses customer behavior and cosine similarity metric to determine product substitution. For technical details on the Collaborative Filtering Approach, see Appendix A
- **Outcome**: Produced suboptimal results compared to our chosen NLP method.

FP-Growth & Apriori Method:

- Analyze frequent item sets in transaction data to uncover products often purchased together. For technical details on these methods, see Appendix A
- **Outcome**: Technical implementation issues and produced suboptimal results compared to our chosen NLP method.

Conclusion

Business Implications

Increased Store Aisle Efficiency: Our identification of the top 1000 and placement within an optimized "Instabasket" aisle reduces the need for shoppers to traverse multiple aisles, streamlining operations and reducing in-store customers.

Improved Customer Experience through Relevant Substitutes: Our NLP methodology offers customers top-tier product substitutes, potentially reducing out-of-stock disappointments. Suggesting high-quality product substitutions can lead to increased basket sizes and overall revenue growth.

Revenue Implications: Offering discounts for accepted substitutions may impact revenue. However, it could also enhance customer loyalty and satisfaction by offering value.

Acknowledgements

We extend our appreciation to our judges for their invaluable insights and collaborative spirit. Thank you for your engagement, feedback, and role in shaping our work.

Dmitry Krass, Academic Co-Director

Gerhard Trippen, Academic Co-Director

Brian Keng, Data Scientist in Residence

Meghan Chayka, Data Scientist in Residence

Jay Cao, MDL Research Associate

Rotman

Questions & Answers

We're open to feedback, questions, or any further discussions.

Here's
where it
changes.

Appendix A

Links to Relevant Codebases

Product Selection Code

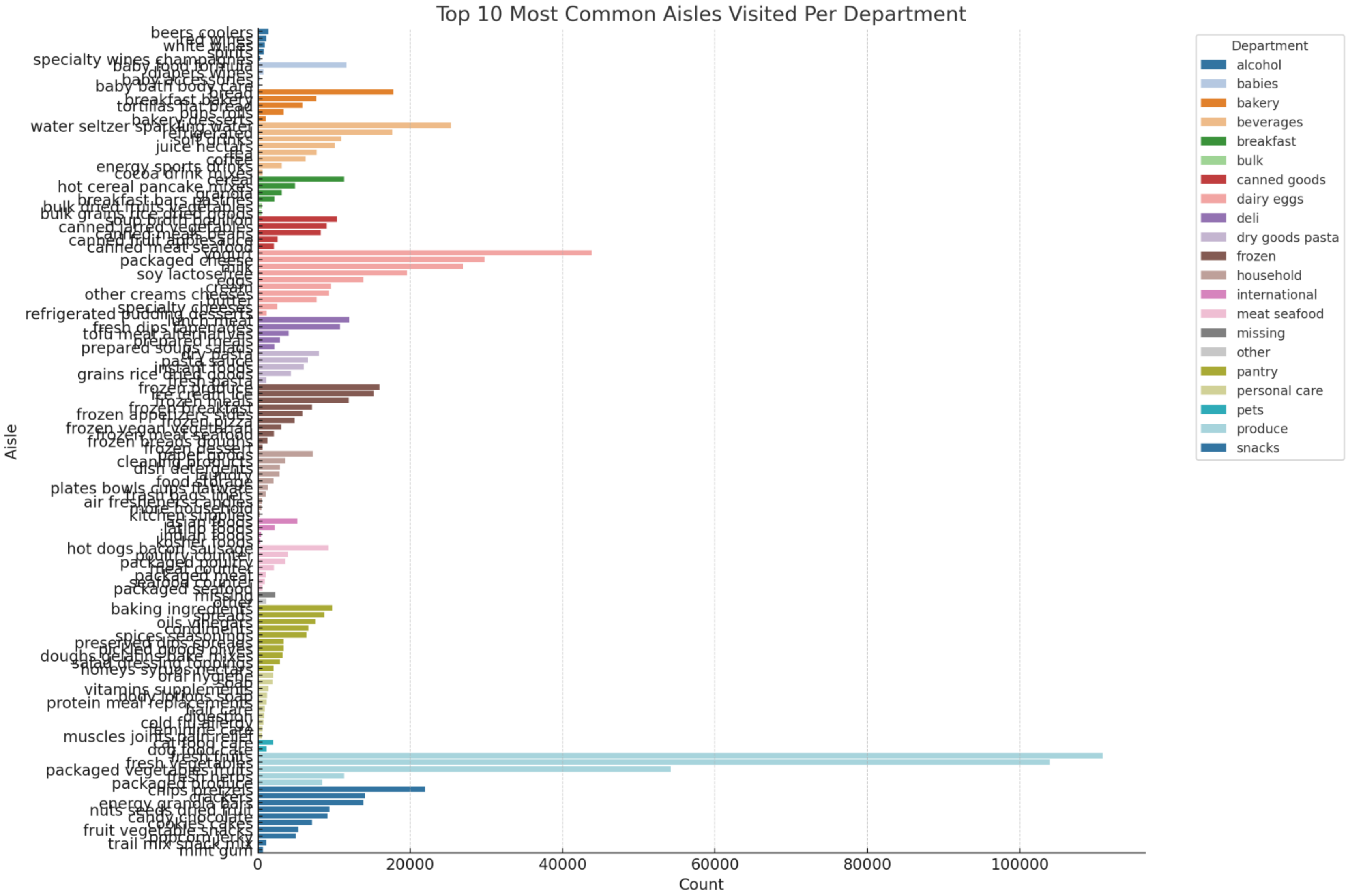
Product Substitution Code

Item-Item Collaborative Filtering Code

Apriori Method Code

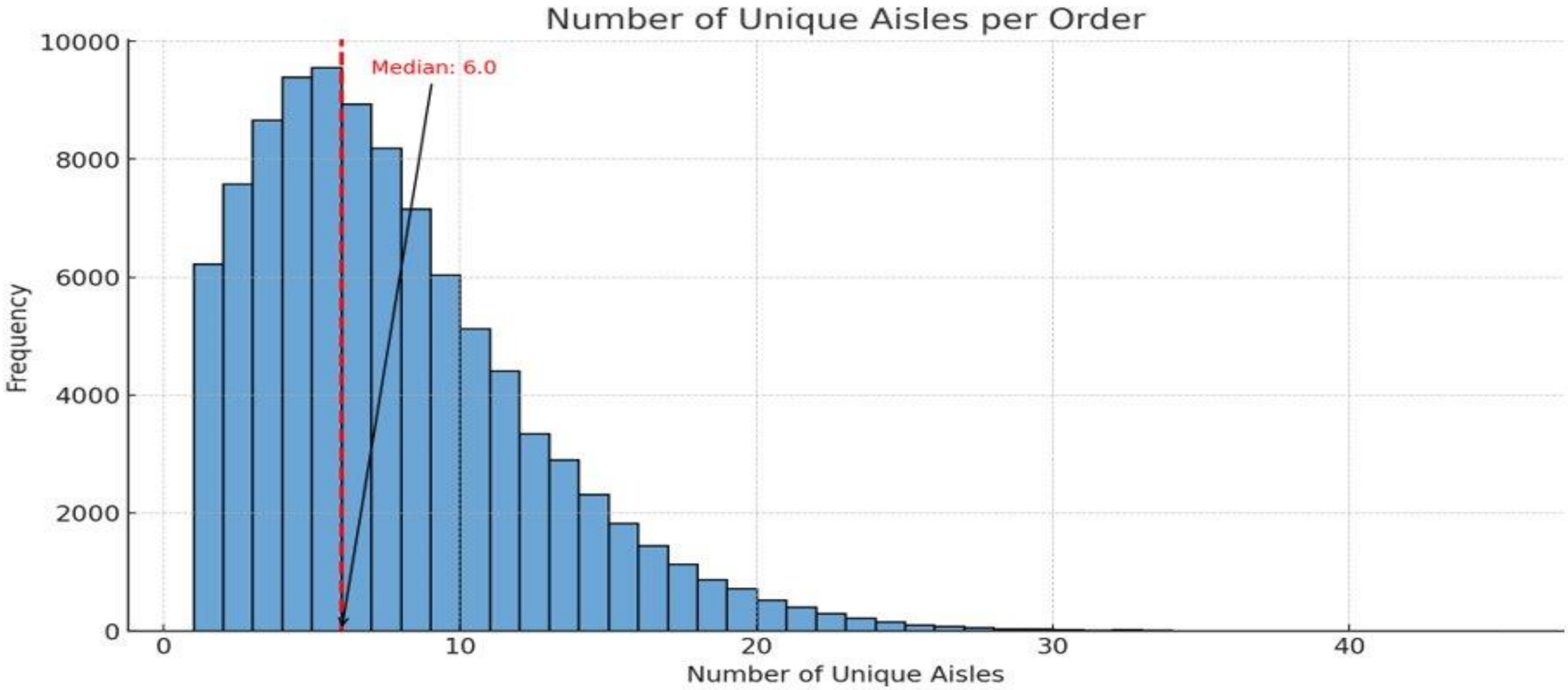
Appendix B

Data Source & Key Data Tendencies



Appendix C

Data Source & Key Data Tendencies


Number of Unique Aisles per Order
Median: 6.0
Frequency
Number of Unique Aisles