# Final Project Part 2

### Aaron Vo

### BIA 678

### July 29, 2024

# Problem Statement

Instacart is one of the world’s most popular online grocery delivery service. For Instacart to remain competitive while also adapting to the constant changing landscape of technology, they must utilize new data science tools to understand the popularity of different grocery items based on their ordering patterns. They aim to identify which products are most frequently ordered and determine if there are any noticeable trends or patterns in consumer purchasing behavior. With this knowledge, Instacart can use this insight to create new marketing strategies to make the customer experience better as well as increasing the maximum production of Instacart.

# Methodology and Tools

## Methodology

There are a series of key stops in the methodology when analyzing the Instacart Market Basket data. First, data is loaded into Python using the *pandas* library, which allows the user to perform various functions like cleaning messy data, make data readable, and make conclusions based on mathematical/statistical theories. The initial phase includes merging the various datasets, such as the product, department, and aisle dataset with the orders dataset. This allows the analysis to provide comprehensive Instacart metadata information on the product from orders The next step involves aggregating and summarizing data to identify key metrics. This includes, but not limited to, sales data, seasonal trends, and product popularity. This analysis includes computing mean and standard deviation, as well as visualizing distributions using histograms and box plots to uncover patterns and trends.

## Tools

The analysis primarily utilizes Python and the *pandas* library for data manipulation and analysis. Pandas is essential for data cleaning, merging, and aggregation tasks, providing powerful functions to handle large datasets efficiently. For data visualization, *matplotlib* and *seaborn* are used to create various plots and charts, such as histograms for distribution analysis and bar plots for visualizing product popularity.

# Insights

## Bestsellers

A graph of blue bars with white text

Description automatically generated

A graph of a number of blue bars

Description automatically generated

A graph of food items

Description automatically generated

Based on the product orders, the observations are:

1. Bananas is the most popular product amongst shoppers. The product with the highest number of orders is Banana with a total of 852,015 orders (organic and regular options combined).
2. Organic options are just as popular as the regular products. The customers prefer to get organic versions of their favorite products, as such the 4 out of the top 5 most popular categories are organic groceries.
3. Perishable items are most in demand. From the data, the perishable items are most in demand as they are all in the top 20 best-selling products.

A pie chart with different colored circles with Crust in the background

Description automatically generatedThe data suggests that fresh produce, particularly organic options, and dairy eggs, are highly favored by consumers. Bananas, both regular and organic, top the list, highlighting their popularity amongst shoppers. In addition, the appeal of organic strawberries, baby spinach, and Hass avocados also indicates a significant demand for organic and healthy food choices. This is further supported by the popularity of the department and isle as the perishable products is much more popular than the others. This trend underscores the importance of stocking a variety of fresh produce to meet consumer demand. Additionally, the significant orders for packaged vegetables and fruits suggest that convenience is also a key factor for consumers. Dairy products, particularly yogurt and cheese, remain important components of the consumer diet, indicating the need for a diverse and well-stocked dairy section.

## Seasonality Data

A blue graph with numbers

Description automatically generatedA blue graph with numbers and a white background

Description automatically generated

Based on the graph, one can see the seasonality trend of Instacart’s customers. Let’s look at the first chart, the days since prior order. This reveals an interesting observation. It is clear the customers frequently shop on the 7th day and the 30th day. This means that they mostly prefer to go shopping weekly, or monthly. Surprisingly, biweekly is not something customers often do as most customers who miss the first week time frame tends to wait till the end of the month. There were extra observations made on what customers would get after a certain day but there were no interesting observations there as it pretty much reflected the previous bestselling graphs.

## Customer Patterns

|  |  |
| --- | --- |
| Product Name | Count |
| Banana | 110916 |
| Bag of Organic Bananas | 78988 |
| Organic Whole Milk | 30927 |
| Organic Strawberries | 27975 |
| Organic Hass Avocado | 24116 |
| Organic Baby Spinach | 23543 |
| Organic Avocado | 22398 |
| Sprint Water | 16822 |
| Strawberries | 16366 |
| Organic Raspberries | 14393 |

The top items people add to their Instacart baskets reflect a strong preference for fresh and organic produce, as well as essential groceries. Bananas lead the list with 110,916 counts, highlighting their popularity due to their convenience, affordability, and versatility as a snack or ingredient. Following closely are bags of organic bananas, with 78,988 counts, which suggests a growing consumer inclination towards organic options. Organic whole milk, with 30,927 counts, is the first non-produce item on the list, indicating its importance in many households, whether for drinking, cooking, or as a staple for children.

Organic strawberries, avocados, and baby spinach also feature prominently, each with over 20,000 counts, showcasing a strong demand for organic fruits and vegetables. This trend extends to organic raspberries, which, although lower on the list, still have a significant count of 14,393. Regular strawberries also make the list, suggesting that while many prefer organic options, non-organic varieties are still popular. Additionally, spring water's presence on the list, with 16,822 counts, points to the importance of hydration and the preference for bottled water. Overall, the data indicates a consumer trend towards healthy, organic, and essential grocery items in Instacart baskets.

A blue graph with white background

Description automatically generatedThis graph shows the distribution of the total amount of products that each customers order. It represents how many products customers have in their cart per each order. It is clear that most customers prefer to have less than 10 items in their cart, and the amount of orders placed exponentially decreases the more items are added in their cart.

A graph with a line

Description automatically generated

We also see here that there is also a common trend. With the number of prior orders per customers. It is prevalent that there are always at least 3 prior orders for each customer when ordering on Instacart.

# Recommendations for Instacart

There are various tools and techniques Instacart can employ in order to create a solid recommendations system for their customers. Instacart can determine the order of departments for customers based on the add\_to\_cart\_order feature.

First, they would need to preprocess the data to ensure it is suitable for training. This would include organizing the dataset so that each row represents an individual customer's shopping trip, with features indicating the order in which the customer added items from various departments to their cart. The add\_to\_cart\_order feature provides the sequence of item additions, which can be mapped to their respective departments. For instance, if a customer adds items from the produce department first, followed by dairy and then beverages, these sequences can be encoded as features for the decision tree model. Instacart can train a model to predict the order of departments for a given customer. The model would learn from the patterns in the historical shopping data, identifying which departments are typically prioritized based on the add\_to\_cart\_order sequences. By training on a sufficiently large and diverse dataset, the decision tree can generalize these patterns to predict the department order for new customers. This approach can help retailers optimize the layout of their stores or the organization of their online shopping interfaces, making it easier for customers to find and purchase items in their preferred sequence, ultimately enhancing the shopping experience and potentially increasing sales.

Another example would be using collaborative filtering. This system can analyze patterns in customer behavior to recommend products that are often purchased together. For instance, if a customer has added bananas and organic milk to their cart, the system might recommend other products that are frequently bought with these items, such as organic baby spinach or organic strawberries. This method leverages the collective purchasing patterns of many users to generate recommendations. Using machine learning, a model can be trained on historical purchase data, with features representing the current items in the cart. The model learns associations between products and can suggest additional items based on these learned patterns. For instance, if the model identifies that customers who buy organic raspberries often also buy organic avocados, it can recommend avocados when raspberries are added to the cart. More advanced models, such as neural networks/classifications, can capture even more complex relationships between products, providing highly personalized and accurate recommendations.