Project 2 - Instacart Market Basket Report

Faris Haddad, Greg Tully, Sam Temlock
W200, Section 1

Background Context

Instacart has emerged as a leading grocery delivery platform in North America and, as avid food enthusiasts, we wanted to understand what people are ordering and how their behaviour differs. After exploring the dataset and the different data points it contains, we identified the opportunity to focus on user and order behaviour when buying produce (fruits, vegetables, etc.).

Research Questions

1. *What is the breakdown of customers that buy organic versus ones that never buy organic when given the option?*
2. *Does purchasing behavior differ between organic and never organic users?*
3. *Does user purchasing behavior change as they make more orders on Instacart? And is buying "organic" a stable or dynamic behavior?*

Research Dataset

The source dataset we leveraged was from a machine-learning based Instacart Kaggle competition that contained a relational set of files describing Instacart customers' orders over time. For the purpose of the competition, the ordered products were split into the prior, train and test datasets, of which the test dataset was not publicly available.

Source datasets: Instacart Market Basket Analysis (available on Kaggle)

- Datasets leveraged for this analysis - "aisles.csv", "departments.csv", "order_products__prior.csv", "order_products__train.csv", "orders.csv", "products.csv"

We note that the "orders.csv" contains information pertaining to each order as a whole, while each "order_products__*.csv" contains information about the basket of products within each order (one order can have multiple products).

Key metrics from the source datasets:

Number of Orders	Number of Users	Total products ordered in prior dataset	Total products ordered in train dataset
3,421,082	206,209	32,434,489	1,384,617

Data Preparation & Exploration

Before we dove into our research questions, we explored the source data to run sanity checks and perform transformations in order to create a clean, produce-only dataset that could be used as the baseline for our analysis.

As we aim to explore how organic buying behavior differs between users, we needed to create a dataset where users made purchases on products that offered both an organic and non-organic alternative, i.e., like-for-like products are available between organic and non-organic. Cleary Produce is an ideal category to do this analysis as many products have like-for-like options. For example, here are the different banana products in the dataset: *Baby Bananas, Bag of Organic Bananas, Banana, Bananas, Green Bananas, Manzano Banana, Organic Banana, Plantain Bananas, Red Banana*

Aside from offering like-for-like products, we also selected Produce because these products make up the largest department in the dataset and almost half of all orders are organic (see charts below).

Some key assumptions that were made during our data preparation and exploration process include:

- We assume that the users had a choice between organic vs non-organic produce
- We assume Instacart “recommendations” do not influence customer behavior
- We assume that each order is what the user added to the shopping cart vs what they actually received (Instacart has an auto replacement feature if the product is unavailable)
- We assume that items that contain ‘missing’ and ‘other’ department are non-produce items

First, the following sanity checks were conducted to validate the integrity of the source data (see key variables in Appendix A):

1. ID numbers for orders, departments, aisles, and products were checked for uniqueness
2. Datasets were checked to ensure that total number of values matched the amounts declared in the source (as listed in the table above) and there were no NaN values
3. Each aisle was checked to make sure there was a unique mapping to a single department (many-to-one relationship)

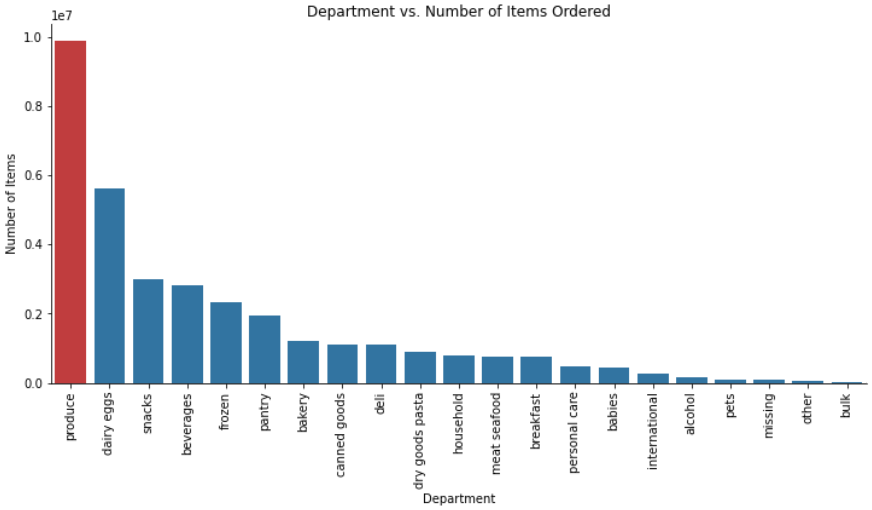
Next, we joined the source datasets into a single dataset containing only produce (as defined by the ‘produce’ value by department). The following steps were taken to create this new dataset:

1. First, we removed any orders that had been assigned to the test dataset, given that this dataset was not available to us.
2. Then, we combined the prior and train orders into one dataset, and checked to make sure no products were dropped.
3. This new combined dataset was then joined with the orders dataset so there was an association between each order and the basket of products within it.

4. Next, the products, aisles, and departments datasets were joined with the new dataset to associate the respective information with each product ordered (available as “complete_dataset.csv”).

Next, we wanted to understand the different departments and how many items are purchased in each department and how many users ordered products in that department.

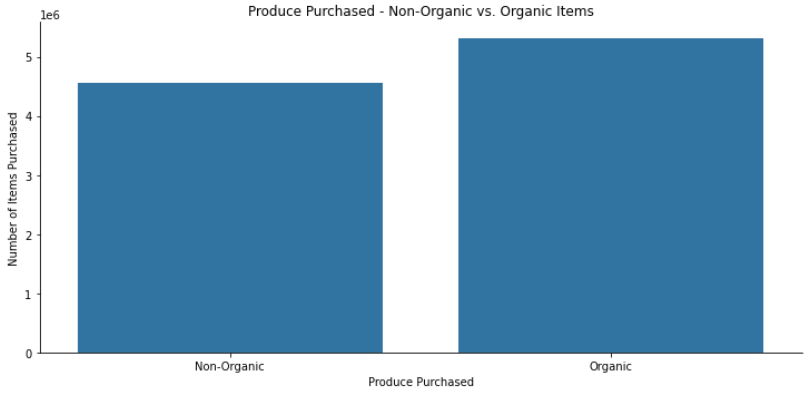
Figure: Number of items purchased per department



Produce is the largest department, both by number of orders and number of users that have ordered.

Within produce as a department, we want to understand the breakdown of the nearly 10 million items of produce, how many were organic and how many were non-organic.

Figure: Number of organic vs non-organic produce items purchased



We can see that within the produce department, 54% of the produce purchased was organic, amounting to over 5,000,000 items.

Finally, we filtered the dataset to only include produce (available as “produce_data.csv”). Below are some key metrics from the new produce dataset:

Number of Orders	Number of Users	Total produce ordered	Total organic produce ordered
2,506,247	194,331	9,888,378	5,323,624

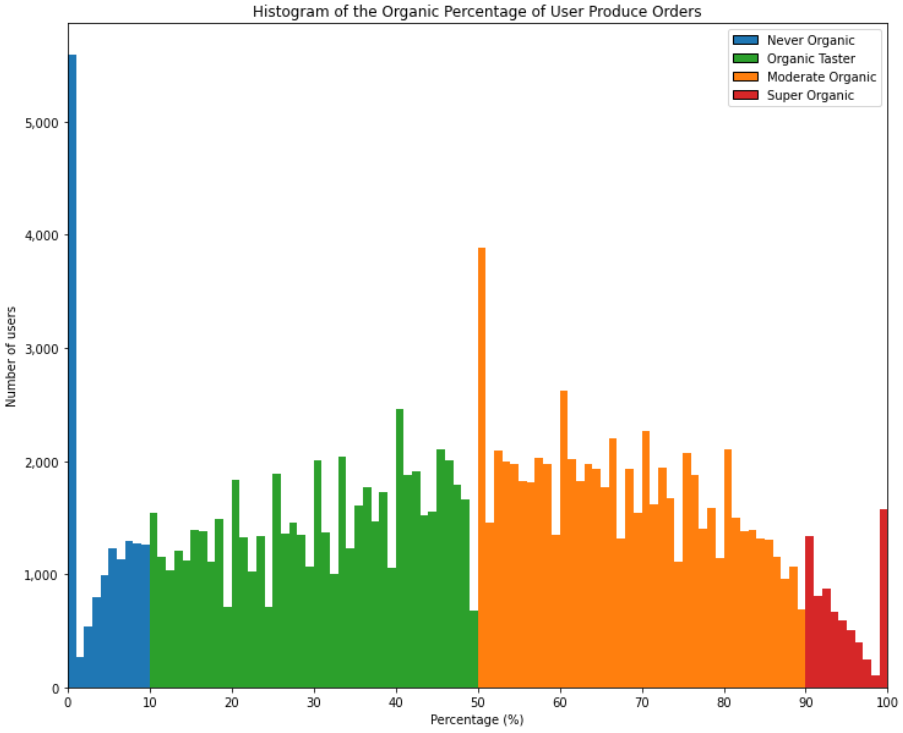
Main Insights

Question (1): What is the breakdown of customers that buy organic versus ones that never buy organic when given the option?

To further investigate the breakdown of customers that buy organic vs those that don't, we decided to come up with four user "segments" that would enable us to broadly categorize these users into four distinct groups. We believed that this segmentation would allow us to better understand the organic purchasing behavior of different types of Instacart users.

To first determine the cutoffs we plotted a distribution of the percentage of organic purchases per user. As seen in the figure below, we determined that each segment should have enough samples to pull meaningful insights out, while also being reasonably allocated so that the more extreme segments would have a smaller number of users.

Figure: Distribution of percentage of organic purchases per user

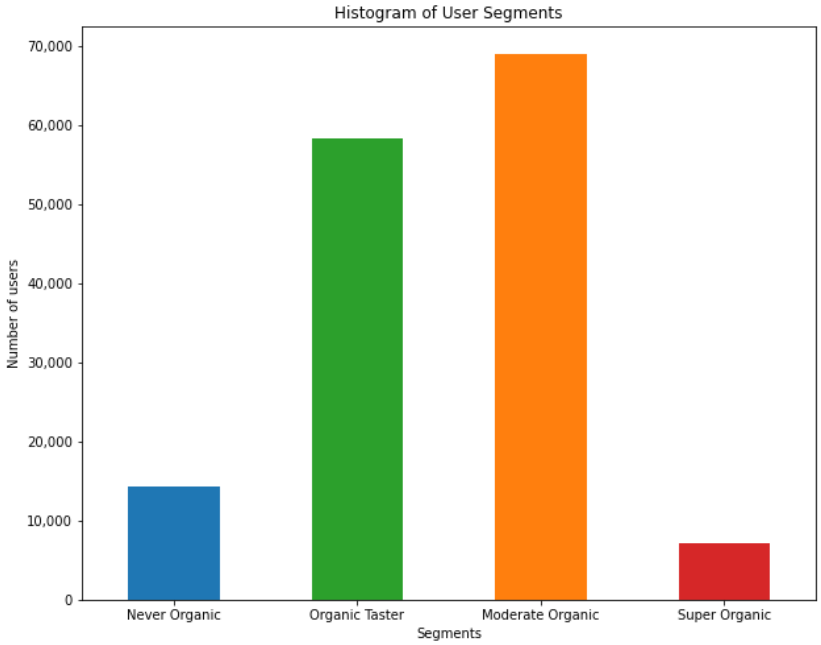


As can be seen from the above distribution, we segmented customers into the four organic segments based on the percent ranges. The segments and their respective ranges are shown in the table below.

Segment	Percent of total organic produce purchased
Never Organic	Below 10%
Organic Taster	Between 10% and 50%
Moderate Organic	Between 50% and 90%
Super Organic	90% or above

These segments were then plotted against the number of users to observe their relative sizes (see figure below).

Figure: Number of users per segment



As can be seen in the segmentation above, the majority of users are Organic Tasters and Moderate Organics. The segments and their respective figures are shown in the table below.

	Never Organic	Organic Taster	Moderate Organic	Super Organic
Number of Users	14,367	58,326	69,074	7,102
Number of Orders	188,169	859,853	1,206,072	117,531
Total Produce Purchased	527,700	3,316,590	5,398,847	427,346
Total Organic Produce Purchased	24,006	1,109,260	3,706,362	400,502

Question (2): Does produce purchasing behavior differ between organic and never organic users?

We hypothesized that between customer segments, purchasing behaviour differs, especially in how often and how much produce each segment generally purchased.

Not surprisingly, we also observed that between segments there are no obvious differences in how customers behaved. Especially with regard to when they shopped online. There was no difference in what day and time each segment buys on InstaCart.

Figure: People appear to mostly order on the weekend but also do it during the week as well.

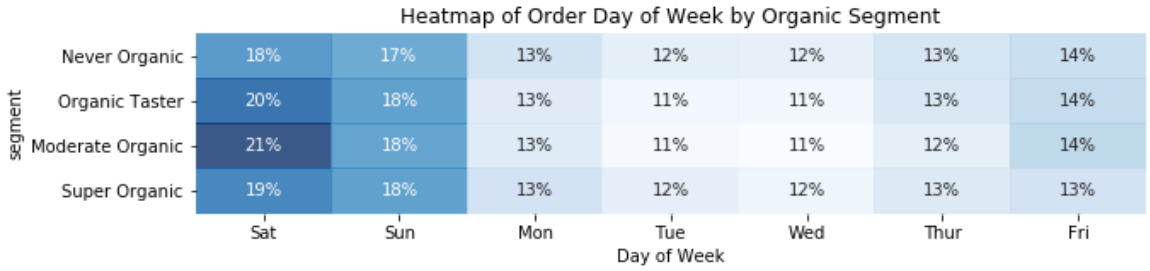
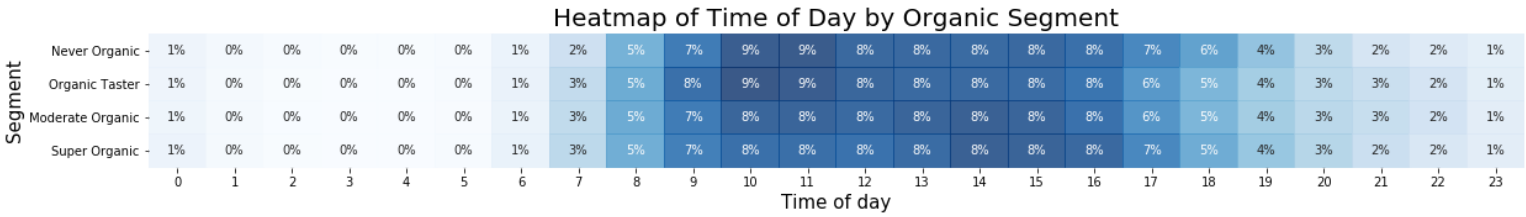


Figure: People appear to mostly order during working hours, between 9AM - 5PM, peaking around mid-morning (10-11AM)



We next examined how purchasing differs between customer segments. The segment that appears to shop in a manner different than others are the Never Organic shoppers compared to shoppers that buy organic. In general, Never Organics seem to shop less often (fewer number of orders per user and shop with more days between each order) and, somewhat surprisingly, Never Organics, on average, buy less items of produce per order.

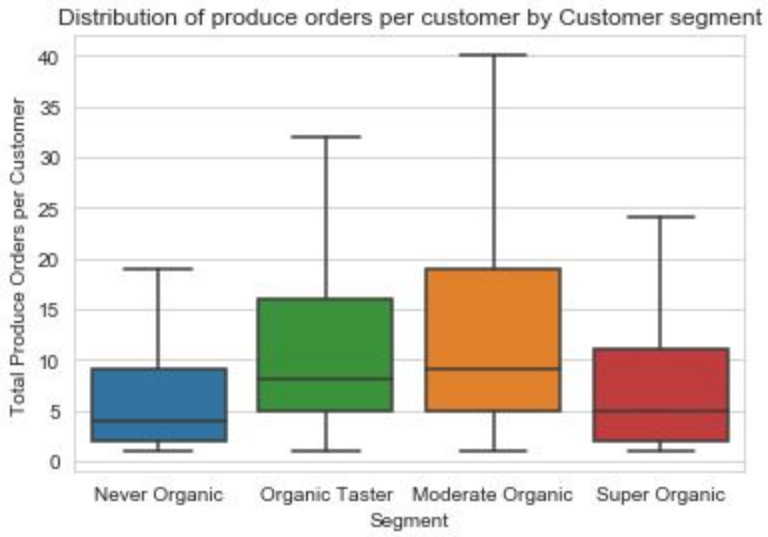
Figure: Table showing purchasing behavior metrics between different segments

Segment	Average # of orders	Average # of days since last order	Average order size
Never Organic	7.6	12.5	2.6
Organic Taster	12.8	11.8	3.8
Moderate Organic	15.4	10.6	4.4
Super Organic	10.0	10.2	3.4

Zooming in on the number of orders per customer, Never Organics have the lowest followed by Super Organics, with Moderate Organics as the shoppers who order more often.

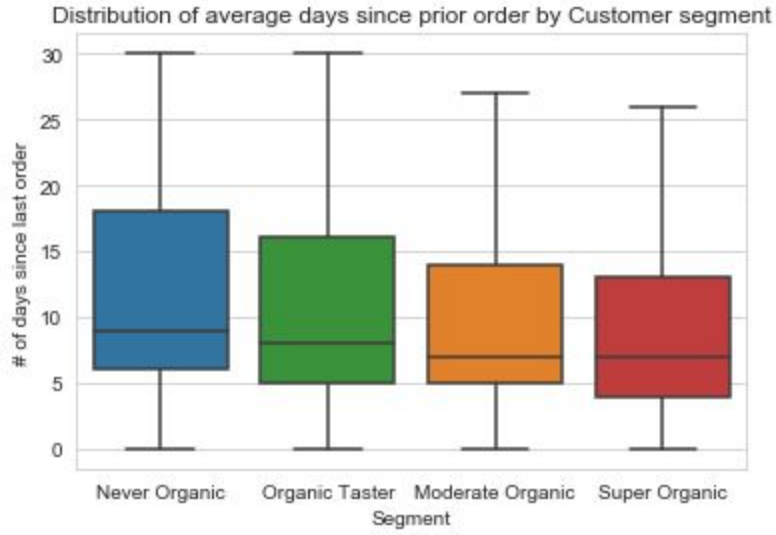
** The following three box plots do not show the "heavy tail" of the dataset, representing ~1% of data*

Figure: Box plot of distribution of produce orders per customer.



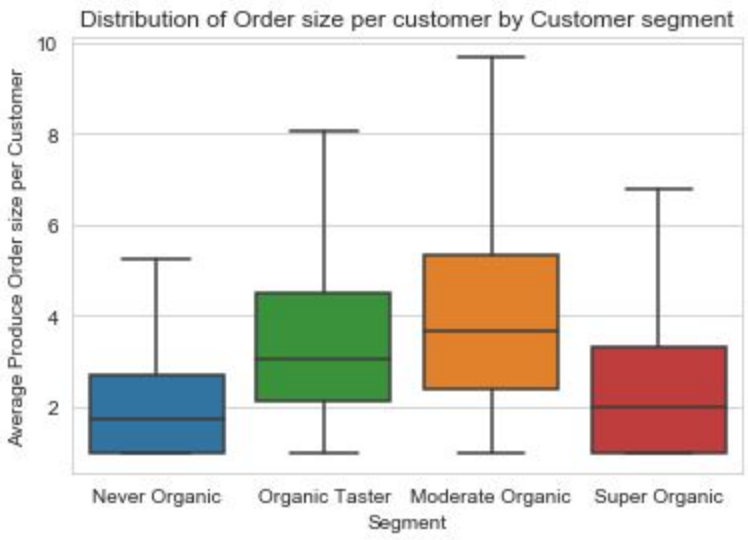
Similarly, Never Organics seem to have a longer lag time between orders compared to other segments.

Figure: Box plot of distribution days since prior order per customer



In looking at shopping cart size, both Never Organics and Super Organic have the lower produce items in their carts. This could be because Never Organics just buy less produce while Super Organics are more picky about what produce they do buy.

Figure: Box plot of produce order size per customer.



We also observed that the trends shown above are similar when users buy any product (not just produce). Below is the same summary table showing purchasing behavior metrics between different segments but for all products. For this analysis, we created a new segment “No Produce” (~12k users or ~6% of all users) that did not buy any produce. Again, Never Organic (and even more so No Produce) seems to buy less often, with few orders and few items per order than Organic segments.

Figure: Table showing purchasing behavior metrics (for all products) between different segments

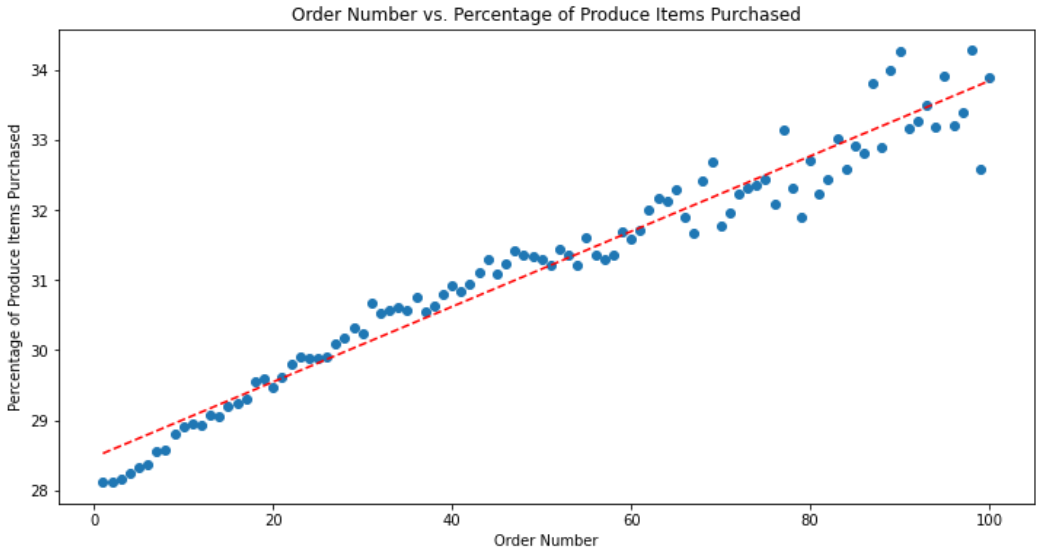
segment	Average # of orders	Average # of days since last order	Average order size
No Produce	9.1	15.0	4.5
Never Organic	13.1	12.7	8.6
Organic Taster	16.5	11.8	10.0
Moderate Organic	18.5	10.7	11.1
Super Organic	14.6	10.9	8.8

Question (3): Does user purchasing behavior change as they make more orders on Instacart? And is buying “organic” a stable or dynamic behavior?

The motivation behind question 3 was to investigate the overall produce purchasing behavior trends of customers and, more specifically, focus on users in their respective segments over their order history. Our theory was that users who order more often (as order number increases) tend to buy more produce. In addition, we anticipated that customers in the extreme categories (Never Organic and Super Organic) were more likely to exhibit stable behavior, while those in between (Organic Taster and Moderate Organic) were more likely to fluctuate in their purchasing behavior.

First we wanted to understand how the percentage of orders that contained produce items changed as users made more orders.

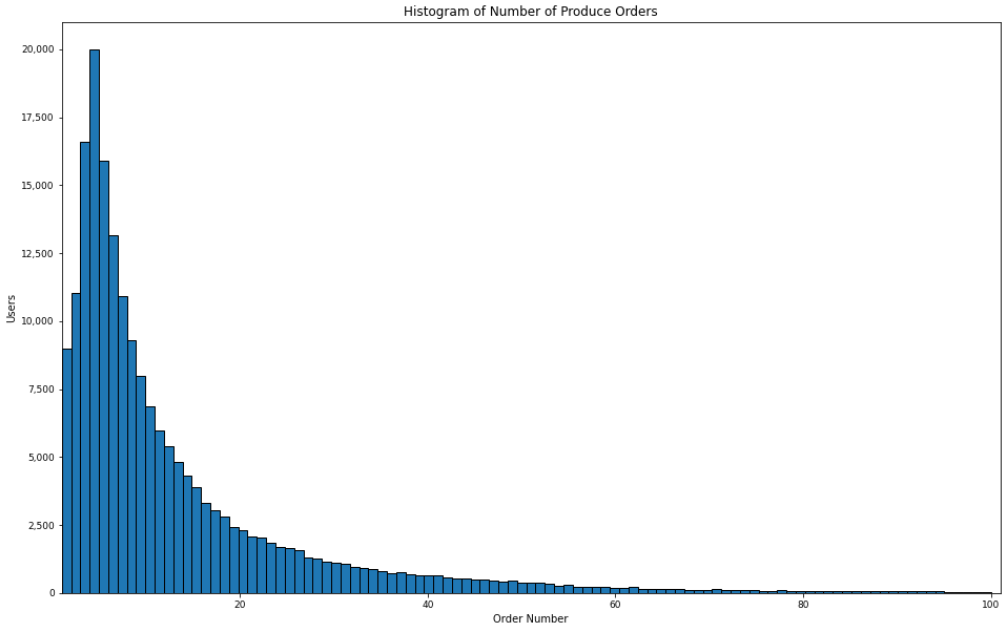
Figure: Percentage of produce items purchased out of total order by order number



As users ordered more, we see that more orders contain at least one produce item.

We then delved into the produce purchasers. We examined the total produce orders count per user to determine how many orders should be used as a starting point to begin to measure user behavior.

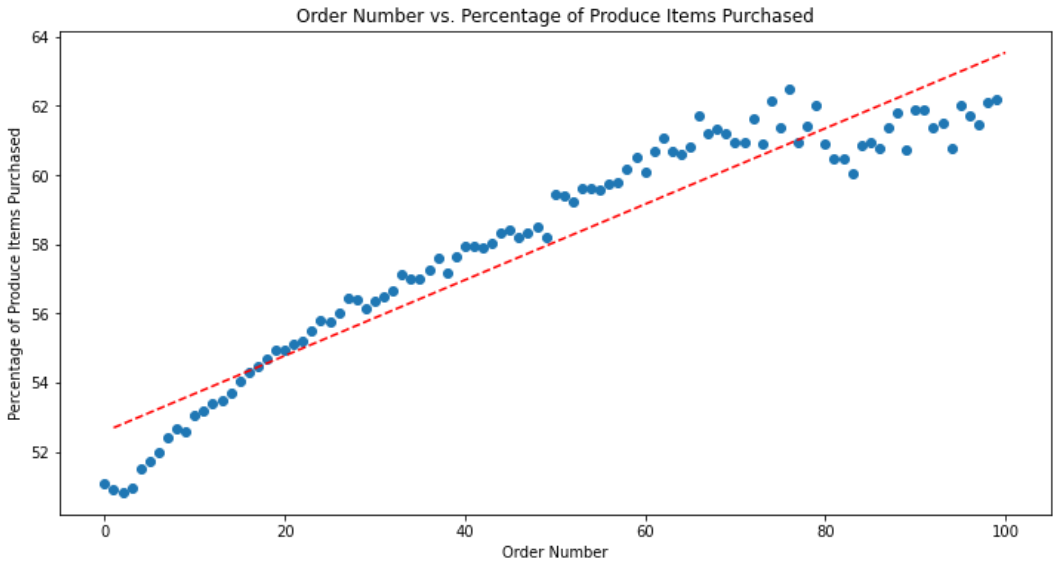
Figure: Distribution of number of produce orders by user



Upon examination of the histogram above, we see that there is a heavy tailed distribution for the produce order count of users.

We analyzed users that purchased produce and what percentage of these items were organic as users made more orders. From the scatter plot below, we see that the percentage of organic produce within the orders increases.

Figure: Percentage of organic produce items purchased out of produce orders by order number

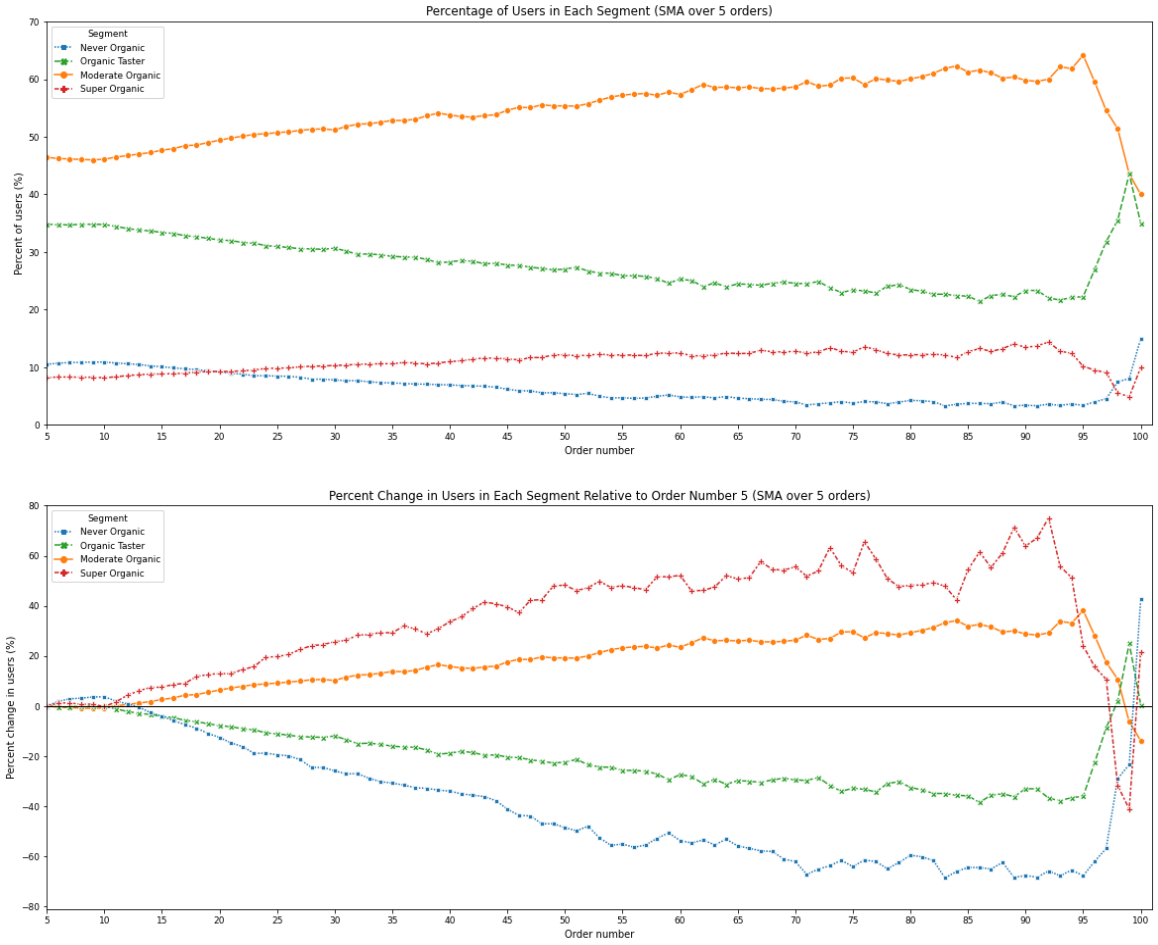


We then checked the mean order count for users, which was 12.9. Thus, we felt safe in deciding to only consider users who had ordered at least 10 times, and to start tracking their behavior from the 5th order. We reasoned that this would provide enough data to show some variation in order behavior, and our sample size still had 80,459 users to work with.

Next, we determined that the best way to track user behavior across orders would be to use weighted moving averages (factoring in size of orders). The two kinds of averages we used were simple moving averages (SMAs) and cumulative moving averages (CMAs). An SMA calculates the rolling average over a set “window” of orders (e.g., averages over 1-5, 2-6, 3-7), while a CMA continues to expand while including all previous orders before it (e.g., averages over 1-5, 1-6, 1-7). We set both the window for the SMAs and the minimum for the CMAs to 5.

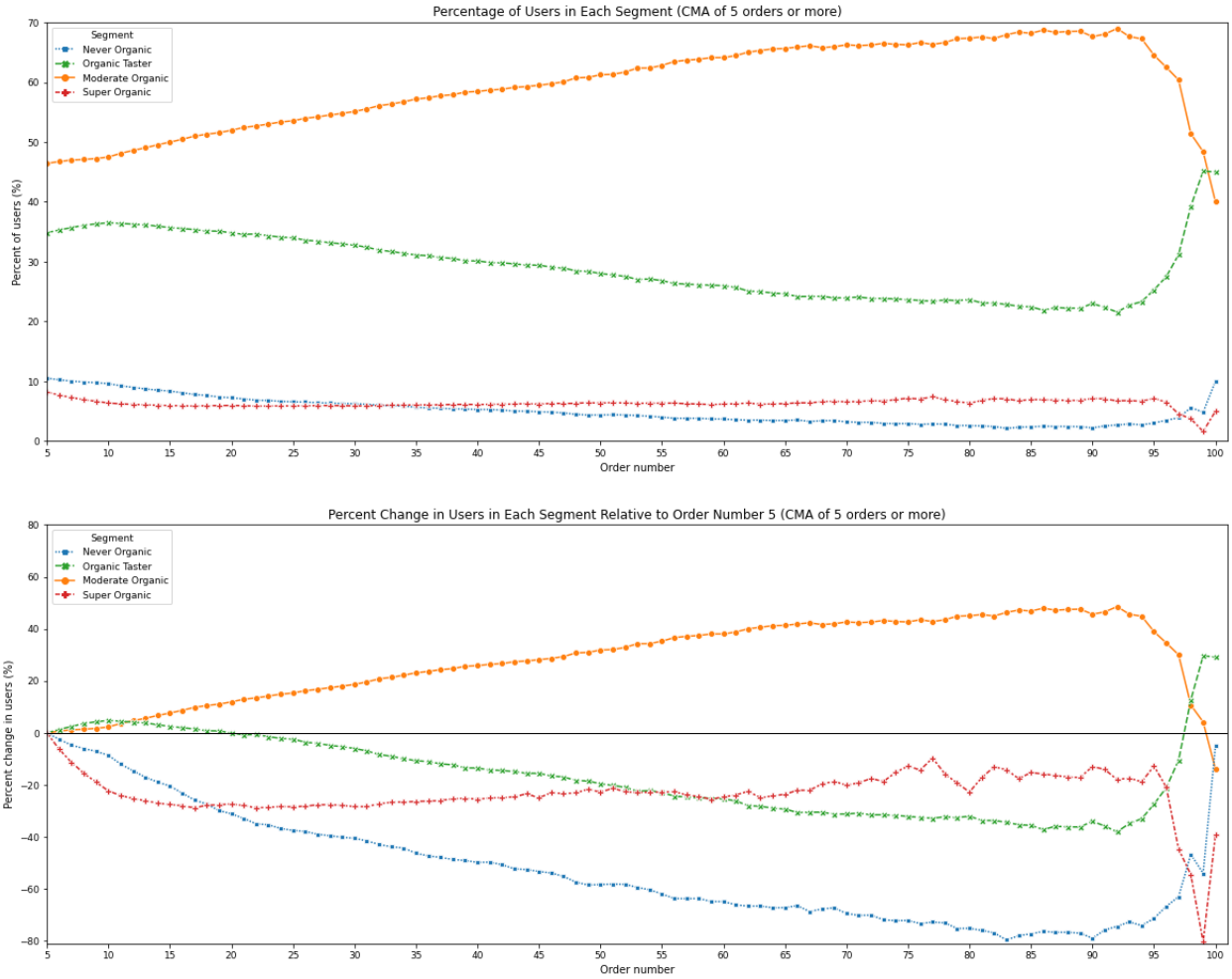
Using these calculations, we plotted the percentage distribution of users across the segments as the number of orders increased, as well as the percentage changes for each segment relative to the initial percentage distribution. Note that for all plots, there is a certain degree of noise after the 95th order, given that the number of users is low (as seen from the histogram above).

Figure: Percentage distribution change per segment across orders (calculated using a SMA)



From the first plot using the SMA above, we see that the percent distribution of Moderate Organic & Super Organic users increases while Never Organic & Organic Tasters users decrease as more orders are made. From the second plot of the SMA relative percentage change, we can see that the degree to which the segmental distribution changes is larger for the Super Organic and Never Organic segments and smaller for the Moderate Organic and Organic Taster segments.

Figure: Percentage distribution change per segment across orders (calculated using a CMA)

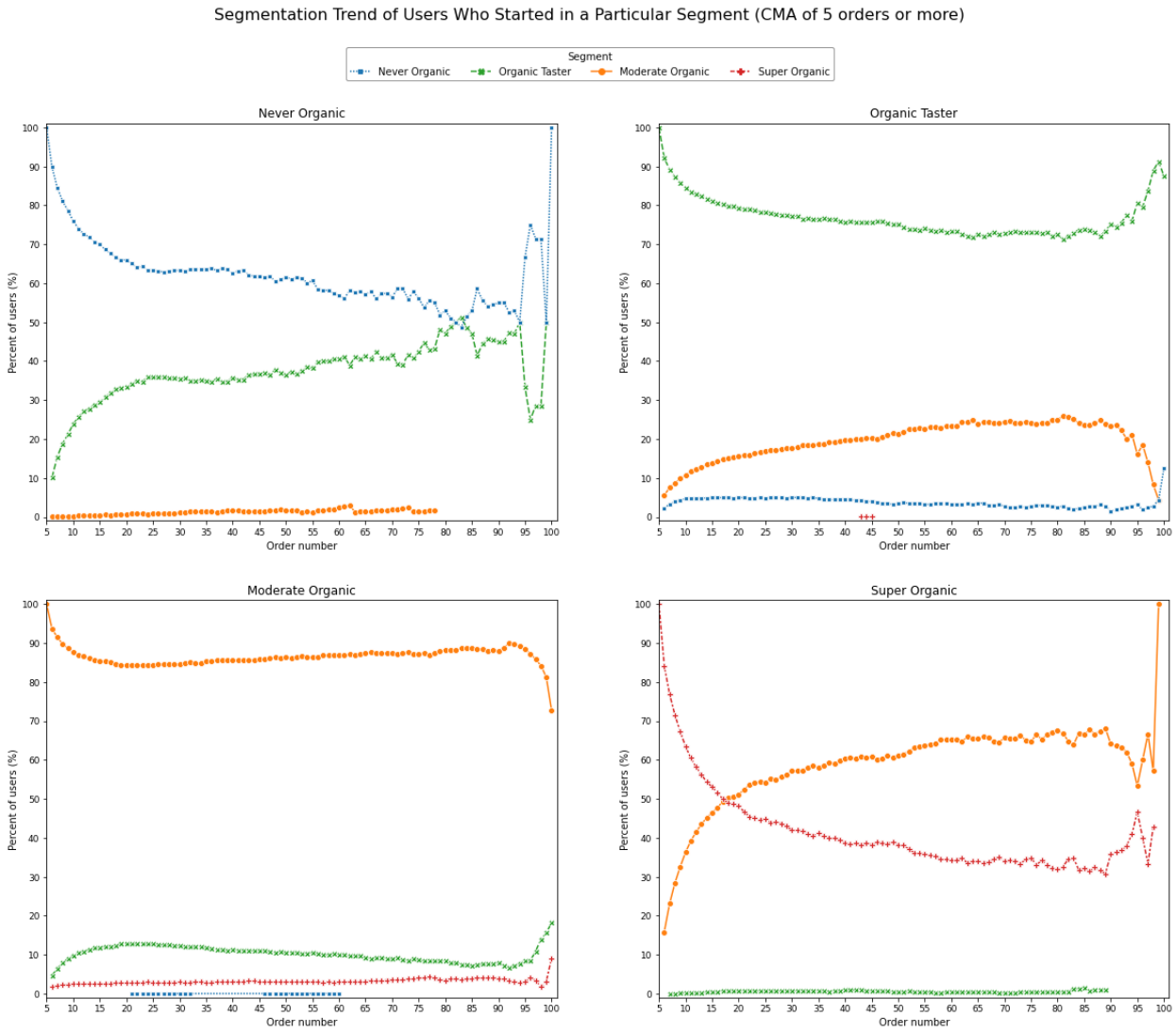


A similar observation can be made using the CMA. From the first plot using the above, we see that the percent distribution of Moderate Organic & Super Organic users seem to increase while Never Organic & Organic Tasters users decrease as more orders are made. However, we see from the second plot that there is a large initial drop (around 30%) for the Super Organic segment before the percentage of users starts to increase again, while the other segments have roughly the same percentage changes as seen from the SMA plot.

As part of a deeper analysis, we then chose to investigate user behavior on a per segment basis in order to better understand the evolution across order numbers for each segment.

As seen in the figure below, we chose to observe the trends of users per segment based on which segment the user started in (calculated using the CMA of the first 5 orders). From the quadrant of plots, we see that up to 50% of the users who were initially Super Organic and Never Organic fall out of those segments as they put in more orders and instead fall into the middle segments of Moderate Organic and Organic Taster. Conversely, we see that more of the users who started off as Moderate Organic and Organic Taster seem to stay that way (roughly 90% and 80% respectively). This would suggest that these two segments are more stable than their dynamic counterparts.

Figure: Segment trends based on a user's first 5 orders across orders (CMA)



Limitations

The following is a list of key limitations we identified in our dataset that prohibit us from deriving stronger insights from our analysis:

- We do not have enough context of the data captured to understand what the real baseline is for the user behavior (i.e., is order number 1 the user's first ever order on Instacart, or just when Instacart started collecting the data for this dataset?)
- There are no dates captured within the time data to conduct any detailed time-series analysis (e.g., purchasing behavior based on season, trends, shortage of certain produce). For example, the date of order number 1 may differ greatly between users.
- There is no data regarding the availability of organic produce at the time of order (e.g., some users may not have had the option of buying organic produce depending on where they shop, the grocery stores that are within delivery range, and whether the organic produce was in stock)

Key Takeaways

In conclusion, given the analysis of our three research questions, we believe that the following insights can help us understand the purchasing behavior of produce among Instacart users:

- **Organic customers make up a large portion** of users (>50% of purchased produce is organic)
- Users that **never buy organic produce (i.e., Never Organics) seem to shop differently** than users that buy organic; they order less, more sporadically and when they do order, they buy less produce than organic buyers
    - Super Organic users also order less produce per order, we hypothesize that they might be “pickier” buyers
- As users make more orders, they **seem to include produce items more often** and **to shift towards organic produce** as well
- Users in the **Super Organic and Never Organic segments appear less stable** than those in the Moderate Organic and Organic Taster segments. Super Organics migrate towards Moderate while Never Organics migrate towards Tasters

Appendix A - Key Variables

- **order_id** & **user_id**: links the order number with the user that made the order.
- **order_id** & **product_id**: links the products that are in each order.
- **product_id** & **product_name** & **department_id** & **aisle_id**: links the type of product (e.g., is it Organic? and the Department; is it Produce? And the Aisle; is it Milk?) that is in each order.
- **order_number**: provides the order_id sequence number for each user (1 = first, n = nth). We will use this to see how user behavior in buying organic changes (or not) as they order more.
- **days_since_prior_order**: # of days since the last order. We will use this variable to assess user behavior to see if certain segments order more or less frequently.
- **reordered**: this variable informs whether or not the user has previously ordered the product (1 if it has been ordered in the past, 0 otherwise). We will use this variable to assess re-order frequency
- **order_dow** & **order_hour_of_day**: the day of the week and hour of the day of the order (day 0 = Saturday, time in 24-hour clock). We will use these variables to determine the order day/time for sets of users.