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

Project 2 Proposal - Instacart Market Basket

Team GitHub repository: Project2_Haddad_Tully_Temlock

Primary dataset: *Instacart Market Basket Analysis (available on Kaggle)*

- *Key metrics from original datasets:*
  - Number of orders: 3,421,082
  - Number of users: 206,209
  - Number of total products ordered in prior dataset: 32,434,489
  - Number of total products ordered in train dataset: 1,384,617

- *Key metrics from new cleaned dataset:*
  - Number of orders: 3,346,083
  - Number of total product orders: 33,819,106
  - Number of "Produce" product orders: 9,888,378
  - Number of organic "Produce" orders: 5,323,624

Research questions:

1. *What is the breakdown of customers that buy organic versus ones that never buy organic when given the option?*
   - Out of the customers who buy organic, can they be further segmented into "super organic" that choose an organic option the majority of the time, "light organic" that sometimes choose organic and "organic tasters" that choose organic sporadically?
   - Due to product heterogeneity, we propose to do this analysis only on customers that purchased "Produce" products (fruits and vegetables) as like-for-like comparisons are possible between organic and non-organic. "Milk" is another category to analyze. Both "Produce" and "Milk" are high volume in the dataset.

2. *Does purchasing behavior differ between organic and never organic users?*
   - Within the organic customers, does purchasing behavior differ between "super organic" , "light organic" and "organic taster" segments?
   - Some proposed metrics to assess purchasing behavior are:
     - Re-order frequency (we hypothesize there is a difference between organic vs never organic)
     - # products per order (we hypothesize there is a difference)
     - Order day/time (we hypothesize there is not a difference)

3. *[OPTIONAL QUESTION] - Is "buying organic" a stable or dynamic behavior? (i.e., Do organic customers always behave the same in how much organic they buy or do they fluctuate between different Organic segments?)*

- If organic buyers change their behavior as they order more on Instacart, do the customers become more or less organic in their purchasing? We hypothesize that the "super organic" are stable while the "light" and "tasters" fluctuate.

Initial plots, figures, or tables:

- A population distribution of segmented organic (super organic, light organic, organic tasters) vs non-organic customers
- A set of histograms detailing purchasing behavior metrics within customer segments.
- A table detailing stable vs dynamic behavior of customers
- [OPTIONAL PLOT] Regression plot of correlation of type of customer (dependent variable) vs purchasing behavior (independent variable)

  **The appendix has a few mock figures to be built for analysis*

Variables to explore and insights to derive:

- The dataset is provided in seven different files.To begin the analysis, we need to build a master dataset that connects each order made with each user and the products that are in each order. The key variables to do this are:
  - **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.
  - **New variable:** We propose to create a variable that measures the percentage of organic produce purchased. Based on the percentage, each user_id will be assigned a segmentation:
    - % of organic produce/total produce (over entire user purchase history)
    - Segmentations: Never Organic = 0%; Organic Taster = 1-25%; Light Organic = 25-50%; Super Organic = +50%

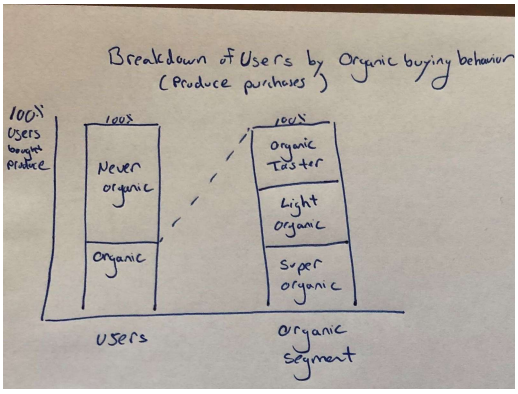
Supplemental datasets:

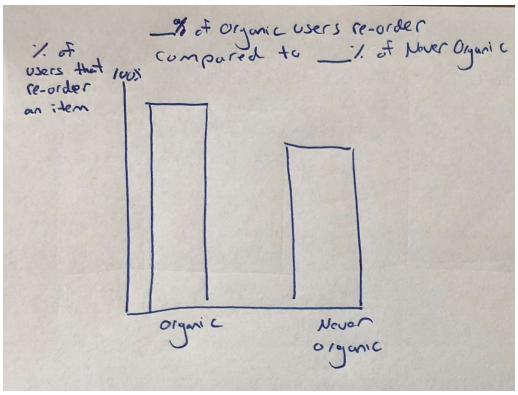
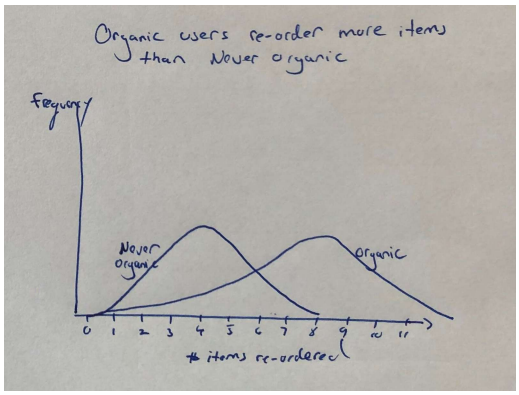
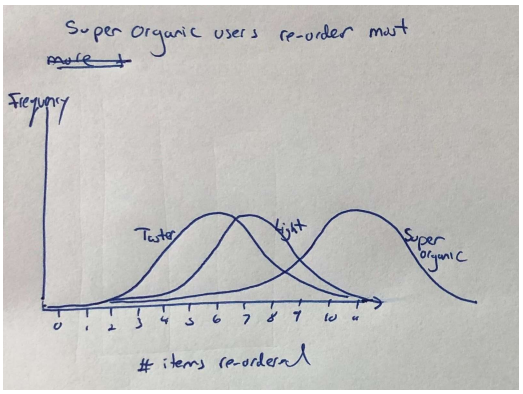
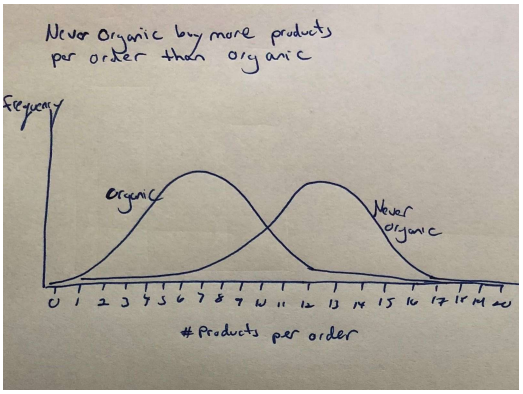
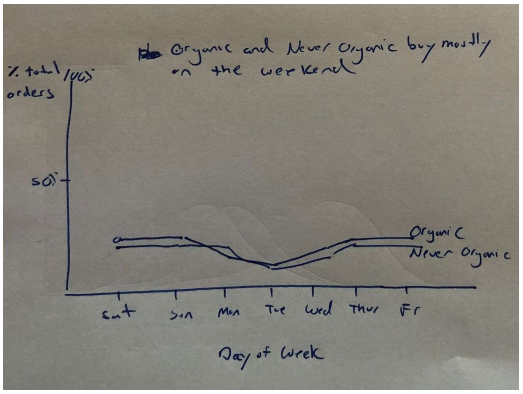
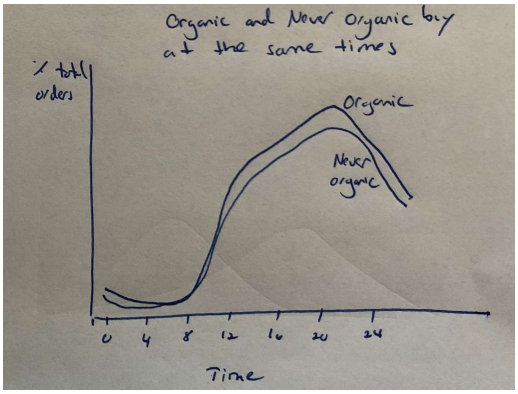
- We do not plan to incorporate supplemental datasets. When Instacart goes public (expected 2021) it would be interesting to see profitability for product lines (or departments) to analyze which organic user segments are more/less profitable.

Final report plan:

- For our final report, we intend to document the data exploration and analysis journey from the initial questions posed, to the data cleaning, to the statistical analysis conducted, and finally to the insights that were derived (structured in that order). We will detail the different challenges we faced at each step of the analysis, and the justification of decisions and assumptions made throughout the process. The insights will be supported by contextual stories and figures that provide the statistical backing for our conclusions. As part of the conclusion, we also plan to address any further analysis that could be conducted in the future to build and expand upon the insights derived in the report.

Appendix: Mock figures to be used for analysis

1. What is the breakdown of customers that buy organic versus ones that never buy organic when given the option?	
User Breakdown	
2. Does purchasing behavior differ between organic and never organic users?	

Re-order analysis	  
Products per order	
Orders by Day of the week & Time	 

3. [Optional Question] Is “buying organic” a stable or dynamic behavior?	
Stability Chart