Instacart Market Basket Analysis

Gareth Jobling, Suchetha Kadavgere, Xuanran Ji, Ananth GV

I. Introduction

Instacart has quickly become a leader in grocery delivery and pickup services. In the uncertain times we live in today with COVID-19, we can see just how valuable a service like Instacart is. That is what peaked our interest to dive in and explore some of the data Instacart has made available. This data can allow us to conduct a market basket analysis on frequently bought instacart products. The goal of this analysis is to answer 4 main data questions to help Instacart uncover trends to improve their product.

II. Identify Questions

A. Question 1

The first set of questions we aim to answer in our analyses revolves around finding what types of products get reordered the highest and lowest percent. Namely we hope to answer the following: Are there any products that get repeatedly bought more frequently than others? And do these products fall into a certain category? What about the products that do not get bought as frequently? Is there any reason as to why?

These questions are important to the Instacart development team. They can use the insights gained from these questions to better provide users with recommended products to buy. Knowing what products get reordered the most and why will help in creating an algorithm on the app for ordering products. For example, product types that would get reordered the most often should be listed first and so on.

B. Question 2

Question 2 was to conduct a market basket analysis. We wanted to answer questions like: At what day of the week do people order the most? How often do people use the app to reorder? How many products are usually ordered at once? What is the percentage of reorder in a particular order? Market basket analysis. These questions are to study how people order on an online portal and to decide which department can be placed next to each other to make it convenient for customers buying in the shop and to decrease the fetch time for people attending to deliver an order.

C. Question 3

Question 3 aims at answering questions with respect to market analysis. It gives an understanding of the firm's (Instacart) strategy and weak points. The analysis itself throws some light on:

1. Which products and in which department make more sales on reorders?
2. What time of the day is the instacart used the most by customers?

3. What could be the reason behind the produce department having more reorders?

D. Question 4

The aim of Question 4 was to focus on variable add to cart order. Start by asking what are some of the most ordered products, then go into detail of the question: Which products are more likely to be added to the shopping cart from which aisle, which department? For this question, we are going into detail to understand the aisle and department of popular products. Answering this question helps Instacart to better allocate human resources and manage the team. My third question is the relationship between add to cart order and reorders orders. Add to cart order is the sequence of adding a product in an order. Answering this question helps Instacart to understand the behavior of the customer in terms of reordering products.

III. Describe Dataset

Source:https://www.kaggle.com/c/instacart-market-basket-analysis/data

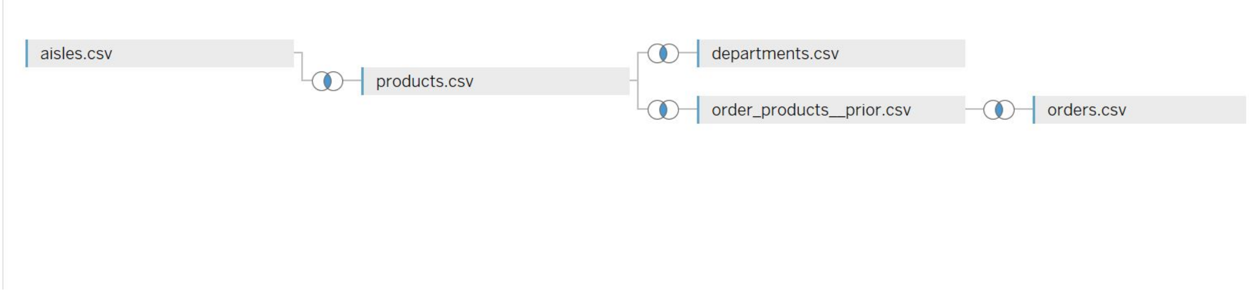
The data was obtained from kaggle and in CSV format. Totally there were 6 files namely:

Aisle.csv : 134 rows
Departments_csv : 21 rows
Order_products_prior.csv: 1048575 rows
products_csv : 49689 rows
Orders_csv : 1048575 rows

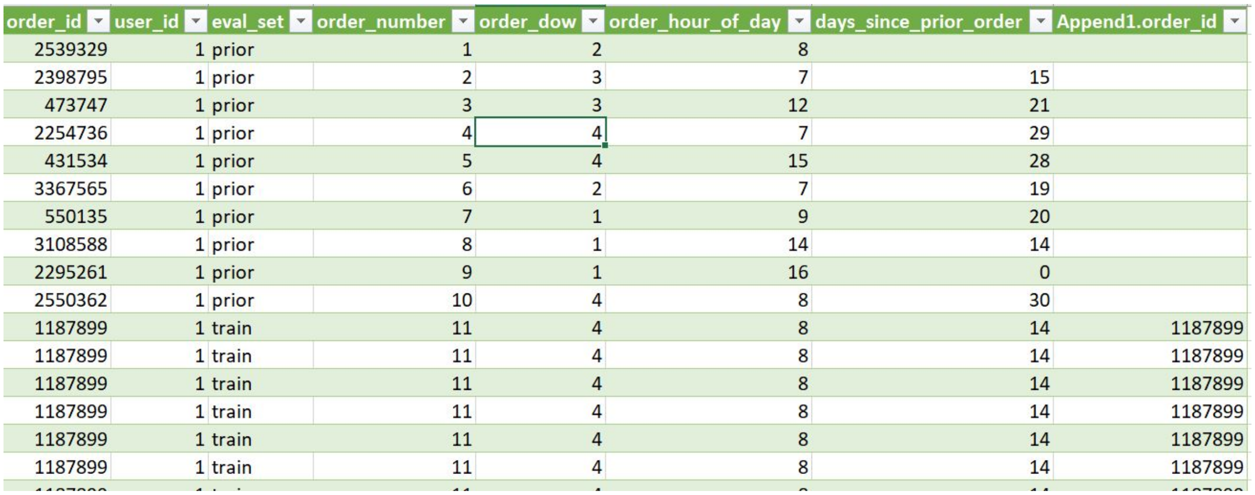
Table	Column	Data Type
Aisle.csv	aisle_id	int
	aisle	char
Departments_csv	department_id	int
	department	char
Order_products_prior.csv	order_id	int
	product_id	int
	add_to_cart_order	int
	reorder	int(binary)
Products.csv	product_id	int
	product_name	char

	aisle_id	int
	department_id	int
Orders	order_id	int
	user_id	int
	eval_set	char
	order_number	int
	order_dow	int
	order_hour_of_day	int
	days_since_prior_order	int

The data tables were joined using tableau/ excel



Data integrity issues - I performed a test to check if all order_ids from the orders table were present in the merged file of order_products_prior + order_products_train. But some orders_ids were not in the merged file. I used an excel query to perform this.



order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order	Append1.order_id
2539329	1	prior	1	2	8		
2398795	1	prior	2	3	7	15	
473747	1	prior	3	3	12	21	
2254736	1	prior	4	4	7	29	
431534	1	prior	5	4	15	28	
3367565	1	prior	6	2	7	19	
550135	1	prior	7	1	9	20	
3108588	1	prior	8	1	14	14	
2295261	1	prior	9	1	16	0	
2550362	1	prior	10	4	8	30	
1187899	1	train	11	4	8	14	1187899
1187899	1	train	11	4	8	14	1187899
1187899	1	train	11	4	8	14	1187899
1187899	1	train	11	4	8	14	1187899
1187899	1	train	11	4	8	14	1187899
1187899	1	train	11	4	8	14	1187899

Since we do not have access to anything more we planned to work with it as it is. The data was clean and did not require cleaning.

IV. Tests Run

A. Tests Run for Question 1

The first test we ran for question 1 was an analysis by type of product (aisles) to display the percent of time that product type was reordered. To do this we created a calculated field in Tableau that counts the number of times the product was reordered from the previous order divided by the total number of orders. We then created a benchmark of 50% to signify that this product gets reordered most of the time it is ordered. This will allow us to see if there are certain kinds of products that get reordered and if there is any sort of relationship between them all. This analysis answers the question which products get reordered more frequently than others.

Using the same test we can also see which products get recorded less frequently. Again, using the same benchmark we can see by product type (aisles) gets ordered at a lower rate than others. These are product types that instacart might either want to advertise better or not advertise at all because they may not get bought as frequently because they do not need to be, helping to answer why these products do not get reordered.

The last type of analysis uses how often the product type is reordered on average. This will help us to identify if these products are staples and are bought most times someone places an order or if they are one time purchases. This can help the developers place these products near the top of the order page.

B. Tests Run for Question 2

Q1) On what day of the week do people order the most?

The first thing looking at orders data was to analyse at what time the traffic is higher and on which day. I used a histogram and bar graph to explore using Tableau.

Q2) How often do people use the app to order?

To see the buying pattern from an app I used days_since_prior_order column to build a treemap.

Q3) How many products are usually ordered at once?

Pivot table helped to get the number of products for each order_id and the frequency of the number of products with pivot charts gave an idea of the average number of products ordered at once.

Q4) Percentage of reorder in a particular order

To study the trends of customer behaviour with respect to ordering I grouped the data by order_id and then got sum(reordered) and count(product_id) to get number of reorders and number of total products for a single order_id. I then calculated the percentage of reorders and built a bar graph with bins.

Q5) Market basket analysis

To know which department should be placed where strategically, I used market basket analysis. I referred to videos on how to run a market basket analysis and created a dashboard to select different departments and see what was bought along with it the most number of times.

C. Tests Run for Question 3

Q1) Which products and in which department make more sales on reorders?

Test: In order to answer this question, I used products, orders, order_products__train, order_products__prior tables to do a group by and measure

Q2) What could be the reason behind the produce department having more reorders? Which other departments have low reorders and what additional attribute supports this answer?

Test: If plotting the graph of departments and reorders, we can clearly see that the produce department generally has a greater reorder rate. The most obvious reason for this is the expiration of produce related products. Hence people generally tend to buy it every few days or so.

Q3) What time of the day is the instacart used the most by customers?

Test: This is essential to answer, because the application is not supposed to experience any downtime. We can ensure this by checking the usage statistics among customers. We plot the orders and reorders count against the time of the day.

D. Tests Run for Question 4

I am using tableau for the question**.** By applying filters to product name and set up size to the sum of frequency that product name appeared. I am able to obtain this circle chart. We observed that Bananas, Bags of Organic bananas, and organic Strawberries are very popular among other products.

For the second question, I am using tableau to answer this question. By joining department.csv, aisles.csv, products.csv, order_products_prior.csv. As we can see, fresh vegetables from the produce department seem to be very popular. We also notice that Organic food is also very popular VS non-organic food. In order to verify that, I will try to plot the % of organic food VS % of non-organic food from add to cart order. I am doing it by using a filter in tableau to filter all product names that contain organic. There are 49,667 types of different product names. 5,036 of them have organic in their product name. By using analysis-total-row and column in tableau I obtained the following statistics. Number of non-organic_add to cart orders is 184,864,556. Number of organic_add to cart orders is 85,998,311. Next, plot the pie chart to show the proportion organic and non-organic related add to cart order.

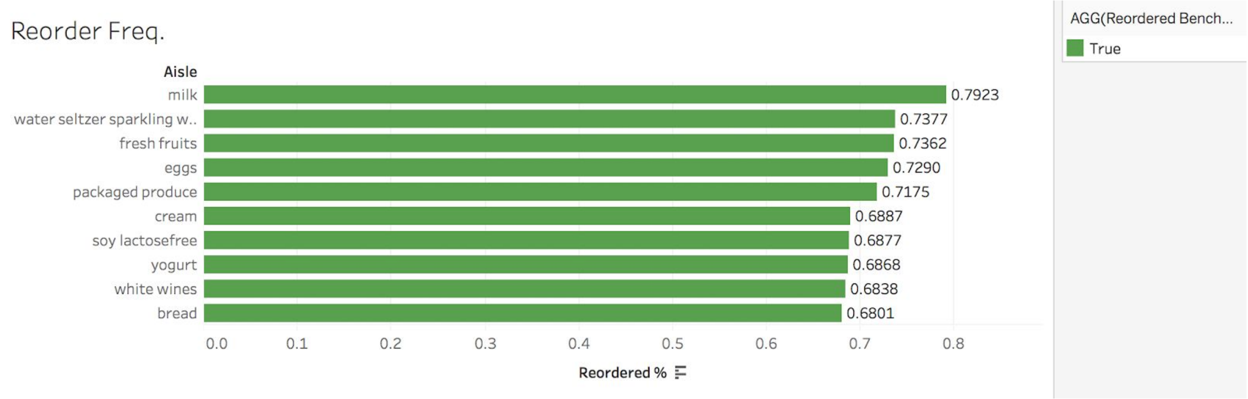
For question 3, I am using Excel to plot the Number of Add to cart order VS Reordered Ratio. What we did is to first group add to cart order by their number. Calculated reorderer ratio based by using sum of reordered divided by total number of orders.

V. Results

A. Question 1 Results

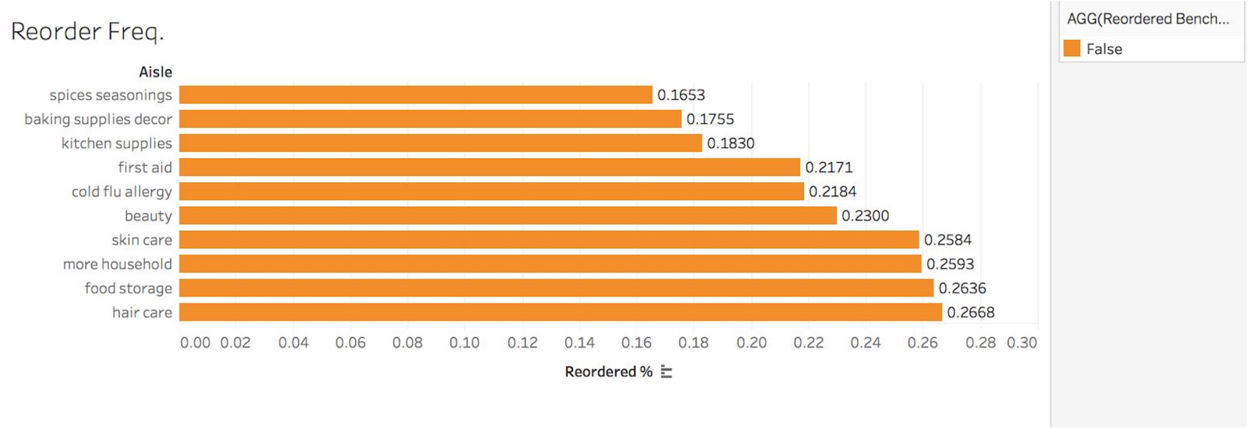
Figure 1 is the result from the first analysis, looking at the most frequently reordered products. This chart shows the top 10 most reordered products As we can see milk, water, fresh fruit, eggs, packaged produce, cream, lactose free dairy, yogurt, white wine, and bread all were reordered in 68% of orders. Excluding white wine, all of these products are items that expire faster than other products and are widely considered to be staples so it would make sense that these get reordered the most. If someone was to have ordered these items in a previous order then it is a good bet they will order them again. I would recommend that Instacart add in an algorithm that puts these items at the top of the recommended items to add to cart.

Figure 1: Products with the highest Reorder Frequency



On the reverse side, as shown in Figure 2, are the items that are reordered the least amount. These items include spices & seasonings, baking supplies, kitchen supplies, first aid, cold & allergy items, beauty items, skin care products, household items, food storage, and hair care. These are all items that are not regularly on a grocery list. These items either last a long time, such as spices or supplies, or do not need to be purchased because they aren't always used, such as first aid items. Included in Instacart's algorithm should be a way to not put these items near the top of the recommended list.

Figure 2: Products with the lowest Reorder Frequency

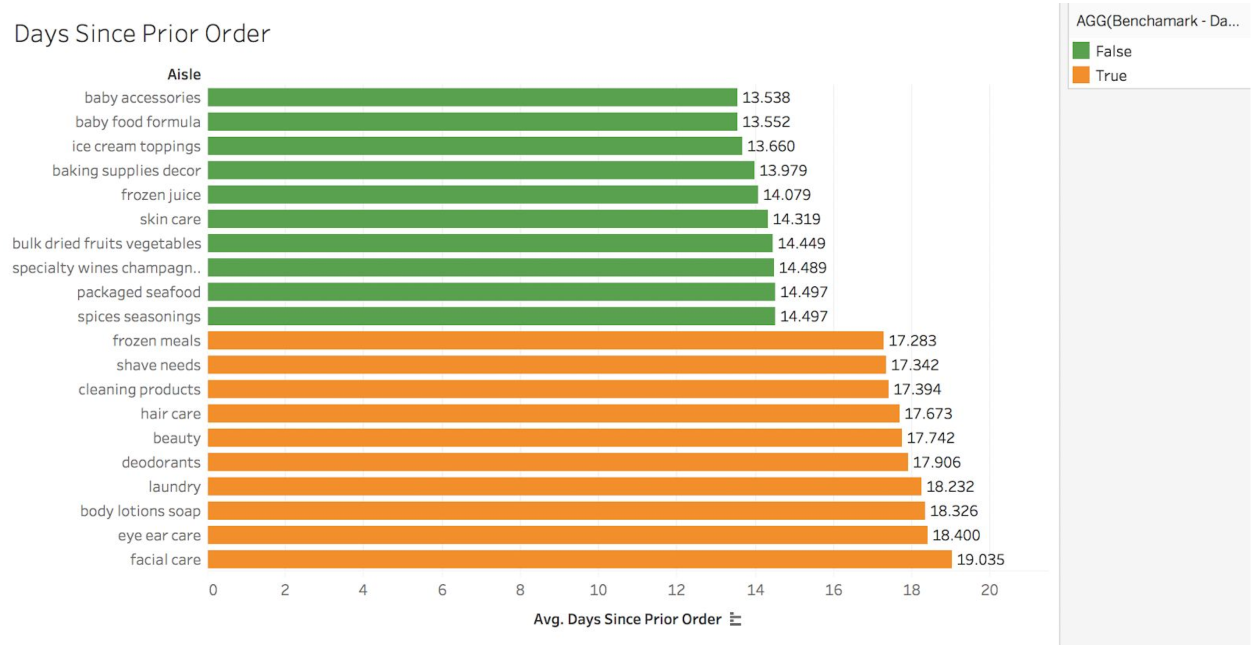


The final analysis of the frequency of items reordered looks at the average number of days between reorders for these items, shown in Figure 3. Items such as baby accessories, baby formula, ice cream toppings, baking supplies decor, frozen juice, skin care, bulk fruits & vegetables, specialty wines, packages seafood, and spice & seasonings get reordered the fastest. Some of these items, such as baby items, can be explained very easily since those who have babies would need these items purchased often. Other items such as bulk fruits and vegetables or specialty wines seem to not have such an easy explanation. One such reason could be that they are not as reordered often as other items and have a low sample size relative to other items that get reordered.

Using the same graph we can see that facial care items, eye & ear care items, soap, laundry items, deodorant, beauty items, hair care, cleaning products, shave needs, and frozen meals get reordered with the longest time in between reorders. These can all be explained by them having long shelf lives and don't get used as quickly as other items such as bread or eggs.

Knowing how many days it takes on average for items to be reordered is a valuable statistic. Instacart can use these numbers to help make better recommendations for users on the app depending on how long it has been since users last ordered their previously ordered items.

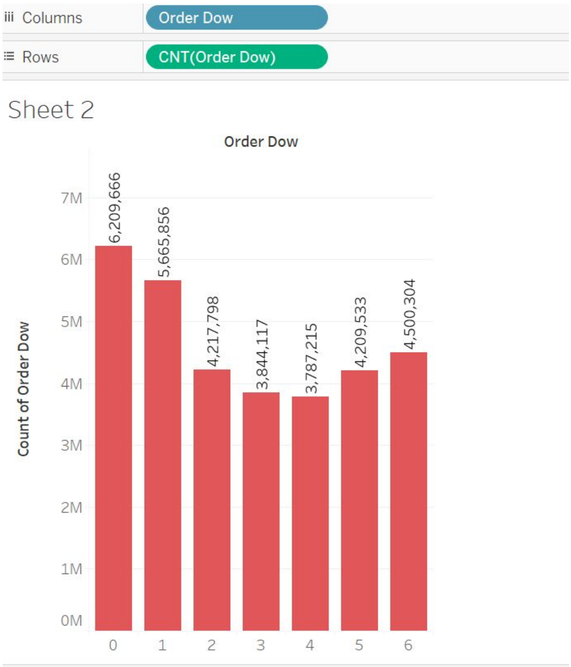
Figure 3: Number of days since prior Order



B. Question 2 Results

Most of the orders are made on 0 and 1 day of the week. Information about which day 0 is was not given but we can assume it’s sunday. A higher traffic can be expected on a sunday.

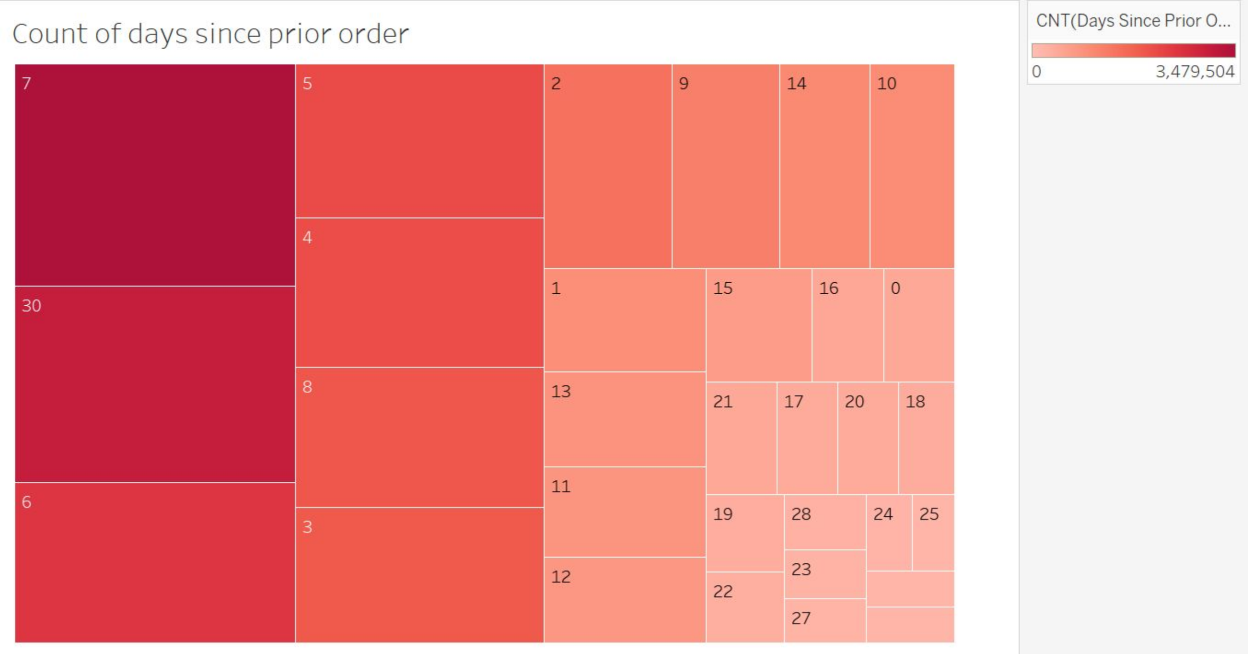
Figure 4: Day of the week versus the frequency count of the of order


Columns Order Dow

Rows CNT(Order Dow)

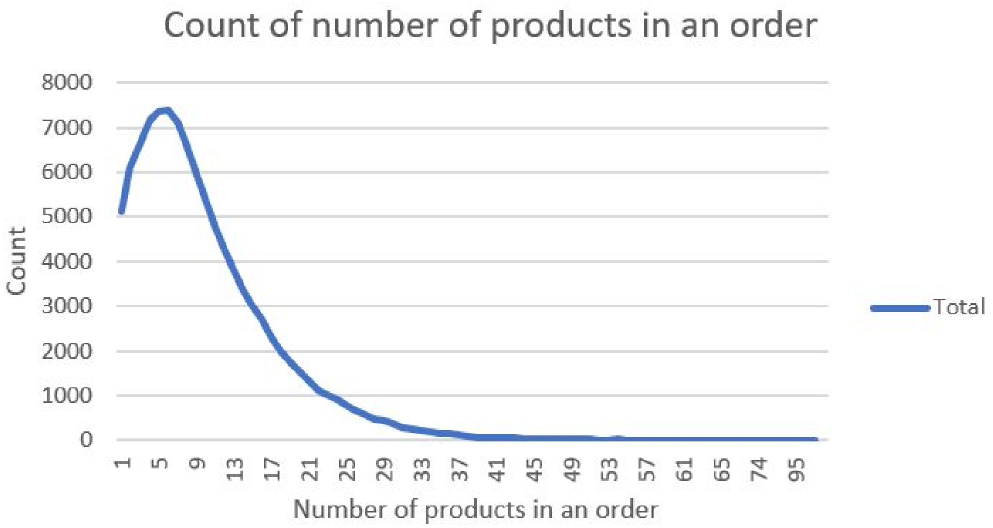
Looks like people reorder within 7 days more frequently followed by 30 days being the next frequent time period. Recommendation or reminder can be sent to customers weekly or monthly with offers.

Figure 5: Count of days since prior order



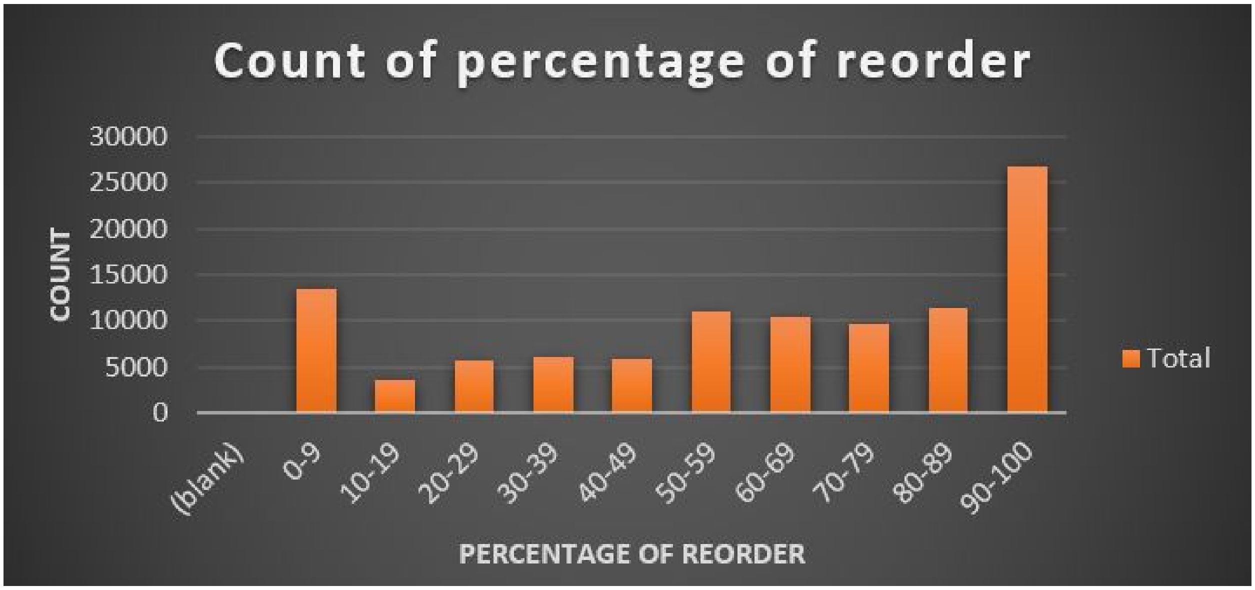
Mostly 7 to 9 products are ordered at once. If the ordered products are more than 9 then customers can be segmented into groups like premium buyers etc.

Figure 6: Count of number of products in an order



A lot of orders have 90-100% reordered products. So in an order 90-100% products are reordered products. Discounts, combo offers can be given looking at the reordered products combinations.

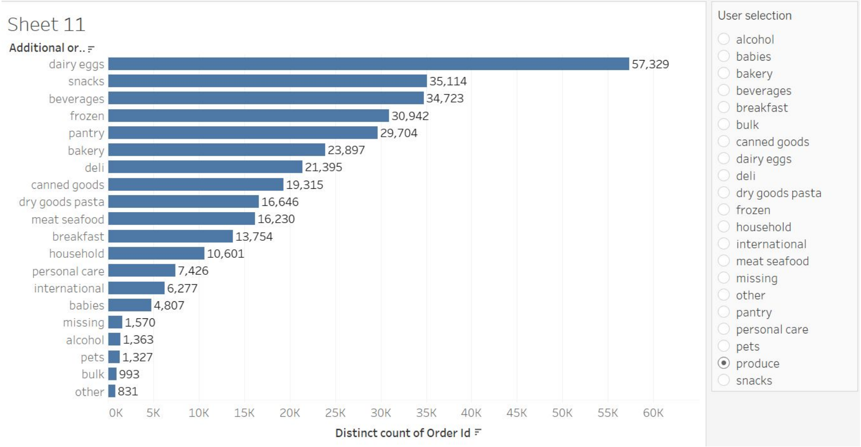
Figure 7: Count of percentage of reorder



Market Basket analysis

Dairy eggs seem to be in the basket a lot of times along with produce.

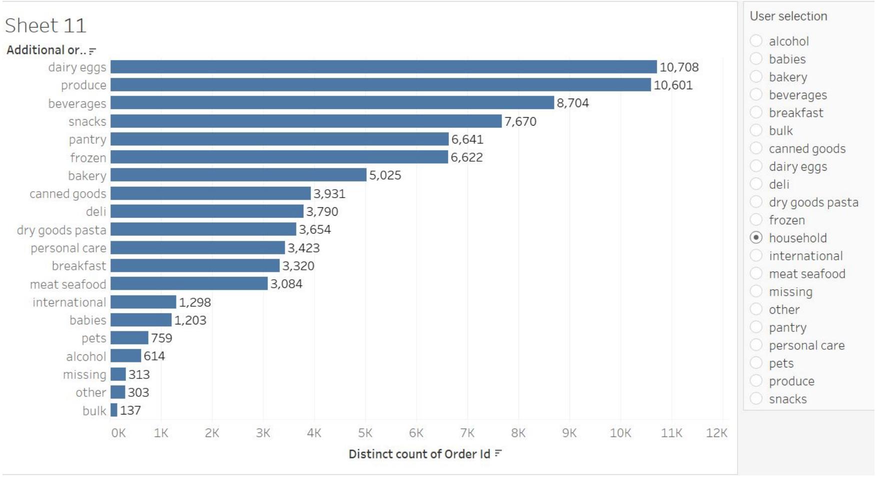
Figure 8: Market Basket Analysis



User selection
○ alcohol
○ babies
○ bakery
○ beverages
○ breakfast
○ bulk
○ canned goods
○ dairy eggs
○ deli
○ dry goods pasta
○ frozen
○ household
○ international
○ meat seafood
○ missing
○ other
○ pantry
○ personal care
○ pets
● produce
○ snacks

Similarly households will mostly be bought along with eggs,veggies,beverages. With this we can decide which department to place next to each other.

Figure 9: Market Basket Analysis



C. Question 3 Results

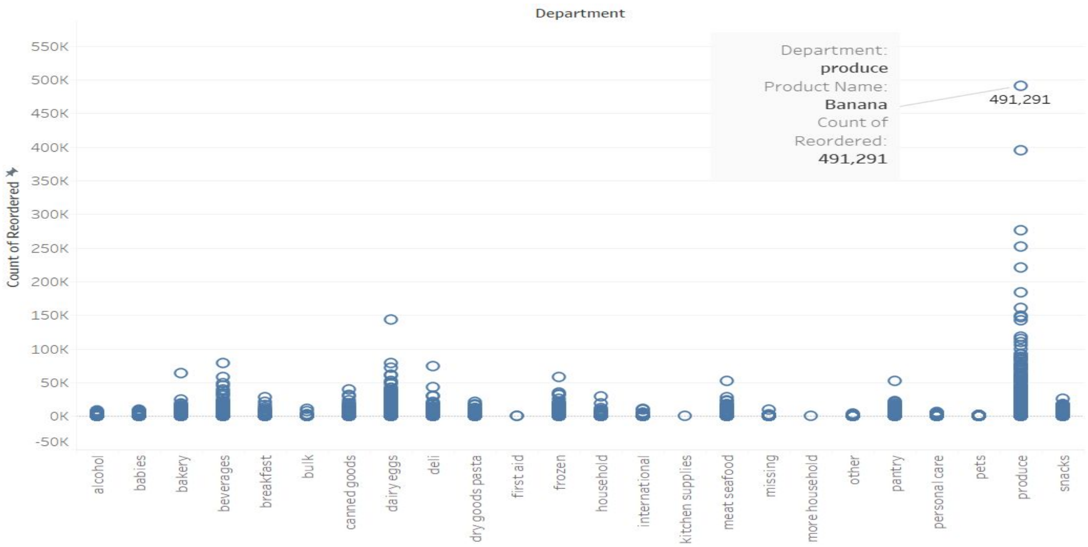
Tests for Question 3:

To understand what products were reordered the most and from which departments and time of the day, I joined all the secondary tables to form a primary table for the analysis.

Q1)Which products and in which department make more sales on reorders?

In order to answer this question, I used products, orders, order_products__train, order_products__prior tables to do a group by and measure. We can see that the banana in produce is sold the most - 491,291.

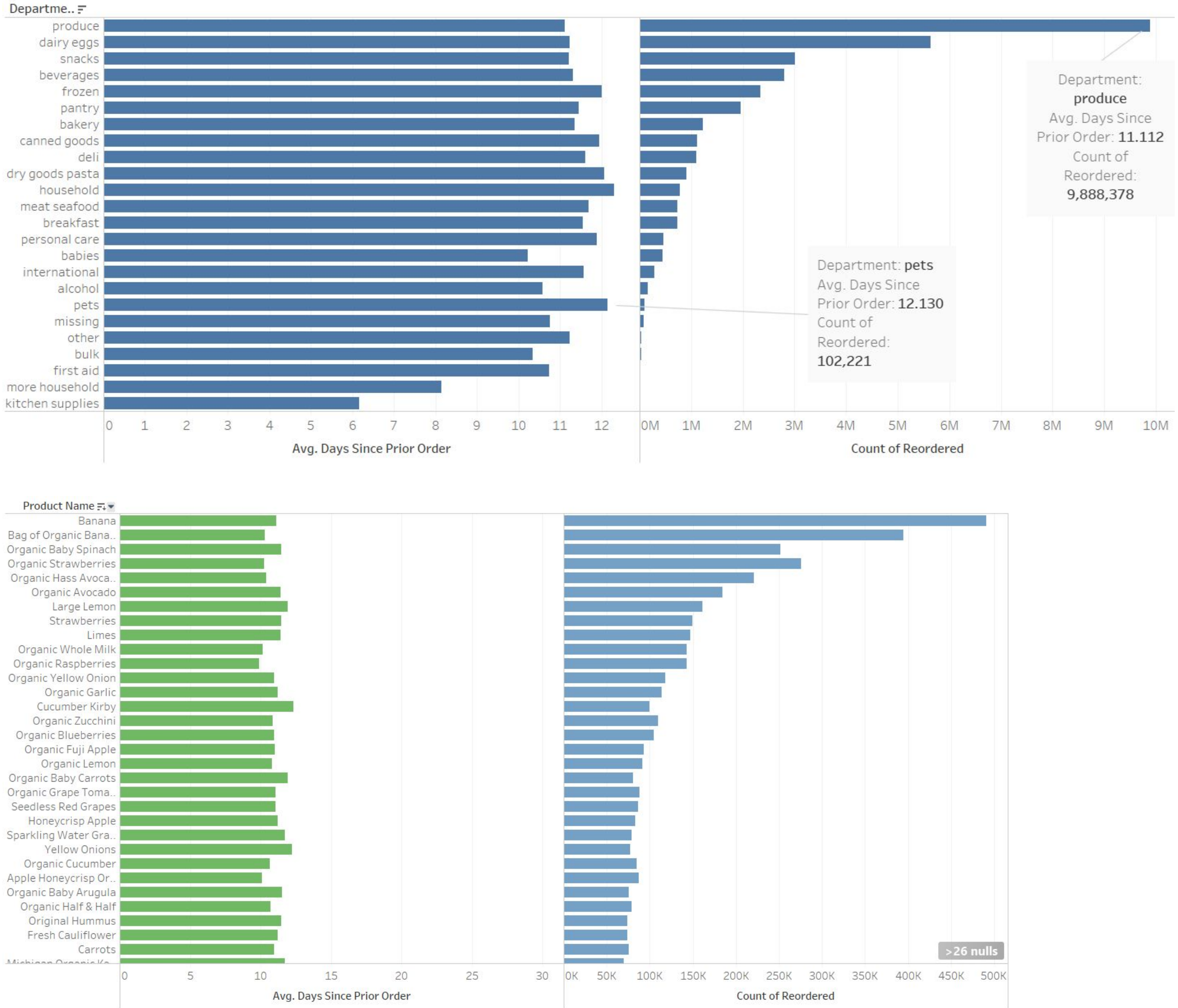
Figure 10: Sales Reorder by Department

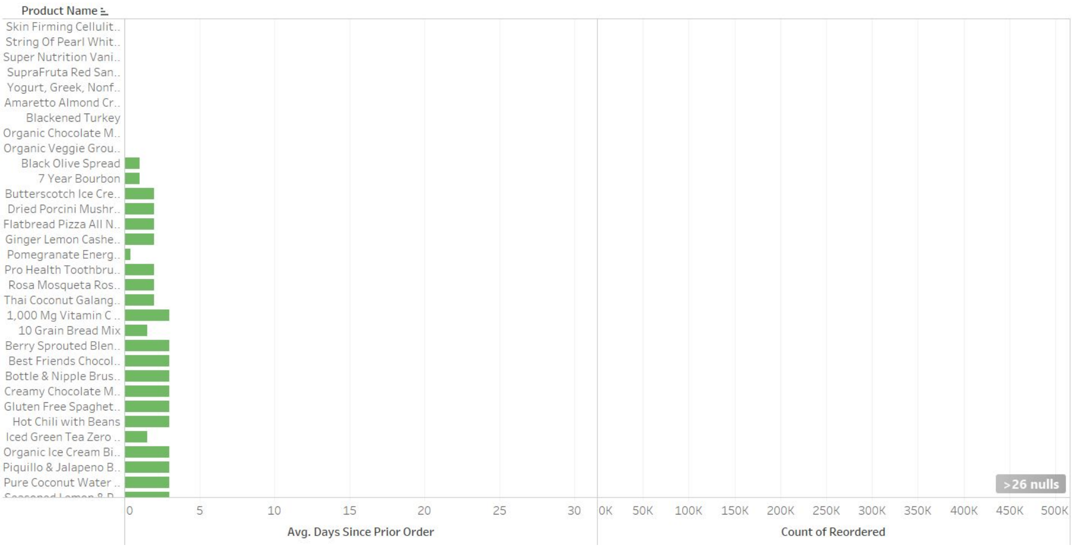


Q2) What could be the reason behind the produce department having more reorders? Which other departments have low reorders and what additional attribute supports this answer?

If plotting the graph of departments and reorders, we can clearly see that the produce department generally has a greater reorder rate. The most obvious reason for this nature is the expiration of produce related products. Hence people generally tend to buy it every few days or so. The attribute that supports this answer is the average_days_since_prior_order. If we plot this we can see that alcohol has a reorder count of 102,221 but the average days since prior reorder is 13. But produce has most reorders (9888,378) but average days since prior order is 10.

Figure 11

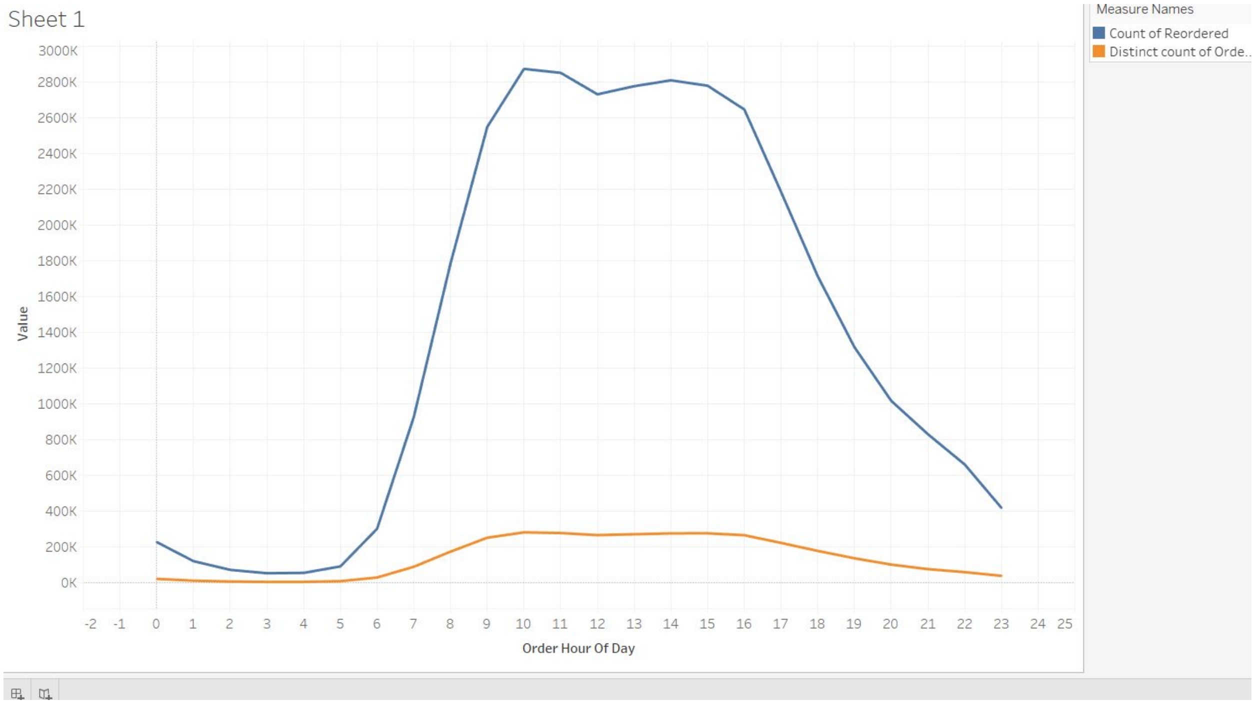




Q3) What time of the day is the instacart used the most by customers?

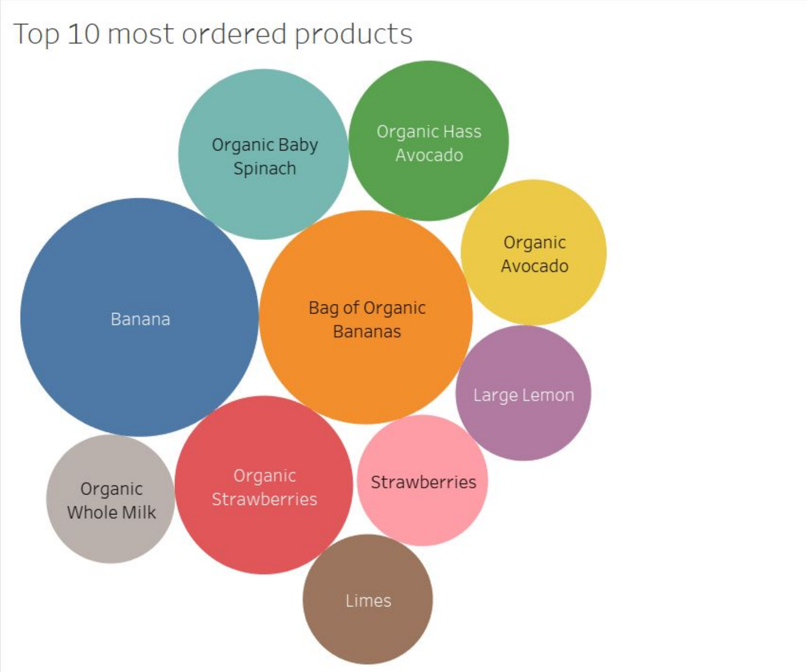
This is essential to answer, because the application is not supposed to experience any downtime. We can ensure this by checking the usage statistics among customers. We plot the orders and reorders count against the time of the day. We can see that most people order products on instacart from 6am to 11pm. To be more specific, the usage is the most from 10m to 2pm.

Figure 12: Instacart Usage by Hour



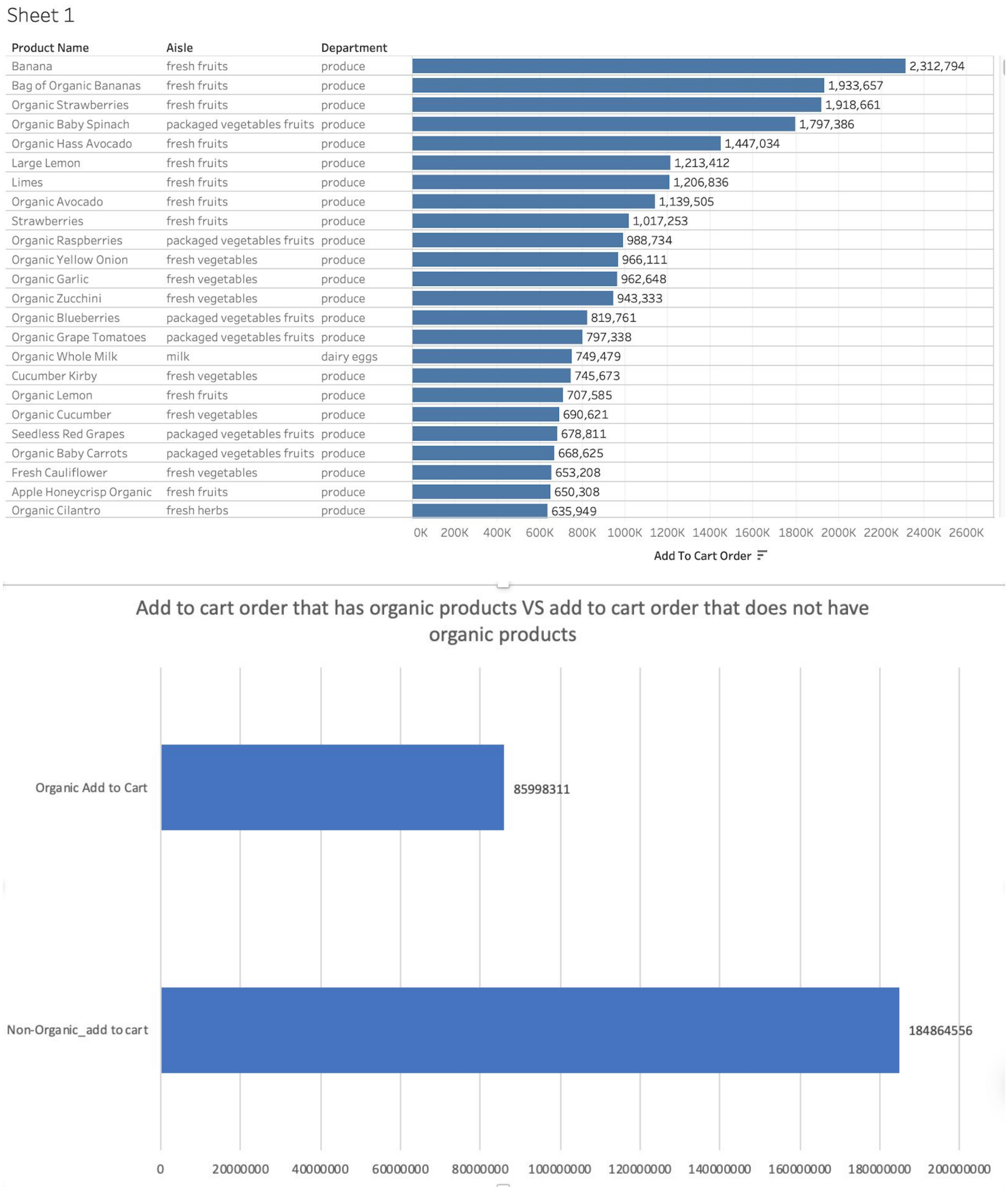
D. Question 4 Results

Figure 13: Top 10 Most Reordered Products



We observed that Bananas, Bag of Organic bananas, organic Strawberries are very popular among other products.

Figure 14



Sheet 1

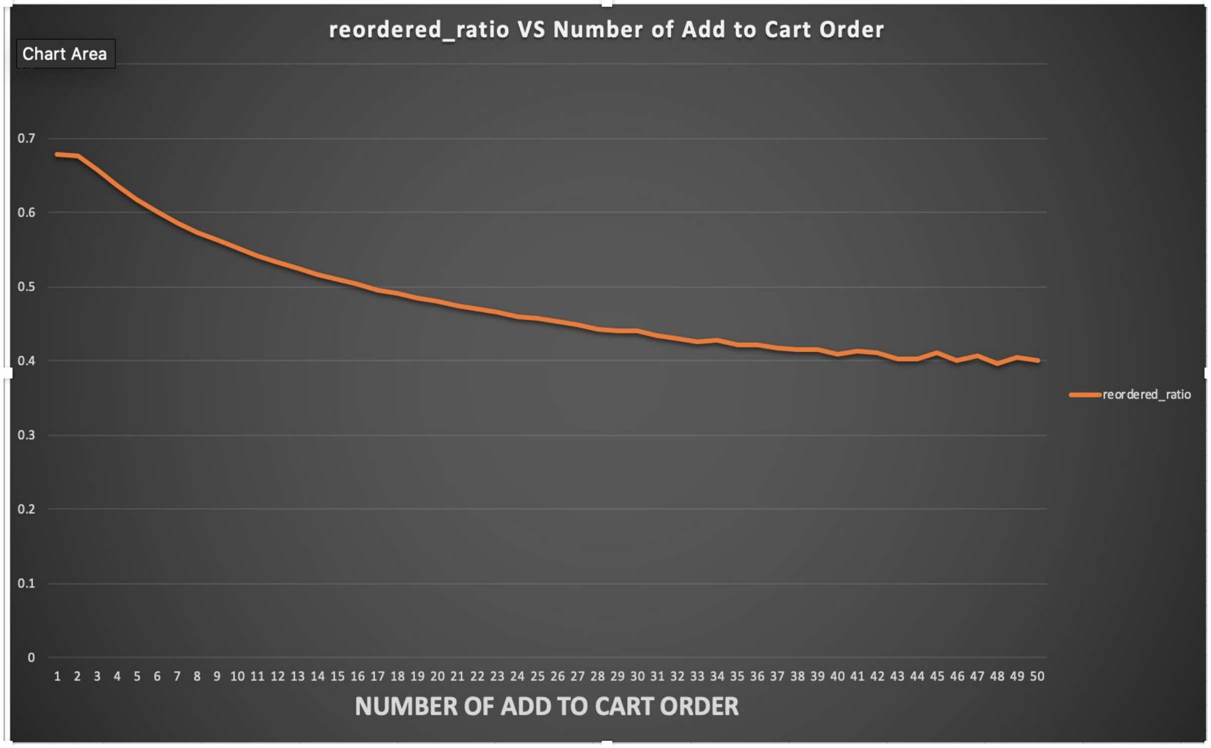
Product Name	Aisle	Department	Add To Cart Order
Banana	fresh fruits	produce	2,312,794
Bag of Organic Bananas	fresh fruits	produce	1,933,657
Organic Strawberries	fresh fruits	produce	1,918,661
Organic Baby Spinach	packaged vegetables fruits	produce	1,797,386
Organic Hass Avocado	fresh fruits	produce	1,447,034
Large Lemon	fresh fruits	produce	1,213,412
Limes	fresh fruits	produce	1,206,836
Organic Avocado	fresh fruits	produce	1,139,505
Strawberries	fresh fruits	produce	1,017,253
Organic Raspberries	packaged vegetables fruits	produce	988,734
Organic Yellow Onion	fresh vegetables	produce	966,111
Organic Garlic	fresh vegetables	produce	962,648
Organic Zucchini	fresh vegetables	produce	943,333
Organic Blueberries	packaged vegetables fruits	produce	819,761
Organic Grape Tomatoes	packaged vegetables fruits	produce	797,338
Organic Whole Milk	milk	dairy eggs	749,479
Cucumber Kirby	fresh vegetables	produce	745,673
Organic Lemon	fresh fruits	produce	707,585
Organic Cucumber	fresh vegetables	produce	690,621
Seedless Red Grapes	packaged vegetables fruits	produce	678,811
Organic Baby Carrots	packaged vegetables fruits	produce	668,625
Fresh Cauliflower	fresh vegetables	produce	653,208
Apple Honeycrisp Organic	fresh fruits	produce	650,308
Organic Cilantro	fresh herbs	produce	635,949

For question 2, we can see that the fresh fruit and vegetable aisle and produce department are really busy. One interesting food is that most of those products are organic. In our Instacart dataset, there are 49,667 different products. 5,036 (10%) of products contain organic 44,631 (90%) of products that do not contain organic.

About 32% of add to cart order has organic products in it. About 68% of add to cart order does not have organic products in it. So, organic products are a big part of the order with only 10% of the product to be organic but 32% in sales.

What we observed is that organic food is really popular. With about 10% of different product names (5036/49667=0.10). Organic product names were in almost 30% of the total add to cart order. That information is useful for companies to target their customers towards products.

Figure 15



We can see a decreasing trend in terms of add to cart order and reorder ratio that means Products that are added to the cart in the beginning, are more likely to be reordered compared to the product that was added later.

VI. Conclusion and Recommendations

There are many conclusions and recommendations we can draw from the analyses performed in this study. From the first set of analyses pertaining to the questions about what kinds of products get reordered the most frequently and least frequently we can draw 3 main conclusions. The first is that the type of products that get reordered the most frequently are items that are staples such as milk, water fruits, and eggs. The developers can use this information to better recommend to users what items they may want to add to their shopping cart. On the other hand, we can see that items that get reordered the least often are items that only are bought when needed such as seasonings, kitchen supplies, and skin care products. Again, the app developers would want to keep this information in mind when making recommendations. The last conclusion we can draw from this set of questions is that when items get reordered, the items that get reordered with the least amount of time in between are baby products. It would be a good idea for the algorithm that recommends products to Instacart users takes this into account that if a user purchases baby formula that they recommend this product the next time they make a purchase.

We can also recommend what product section to place next to each other with the market basket analysis like produce, dairy, egg, beverage, pantry, snack and household closer to each other, when can we expect a lot more orders to arrange for quick on time delivery like we saw sunday has the most traffic, when to send offers or reminders for eg weekly or monthly and further segment the customers depending on the quantity of products bought into categories like premium customers, loyal customers etc..