Intro

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 3 main data questions to help Instacart uncover trends to improve their product.

1. Identify Questions

Question 1:

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?

Question 2:

What percentage is usually reordered?
Which products are sold the most?
How often do people reorder?
At what time and day of the week do people order the most?
Market basket analysis

Question 3:

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

Question 4

Which products are sold the most?
Among all products, Which products are more likely to be added to the shopping cart from which Aisle, which department?
What's the relationship between add_to_cart order and reorder order? Any specific pattern?If there is, please explain.

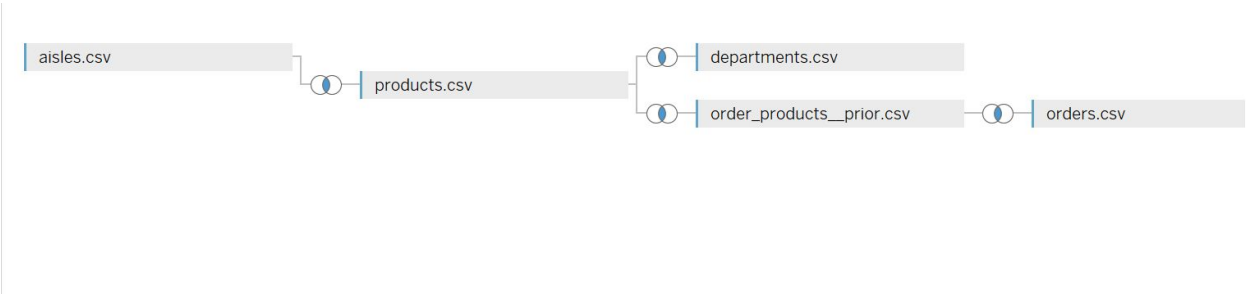
2. Describe Dataset

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 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.

3. Explain the Tests Run

Tests for Question 1:

The first test I ran was an analysis by type of product (aisles) to display the % of time that product type was reordered. To do this I 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. I then created a benchmark of 50% to signify that this product gets reordered most times it is ordered. This will allow us to see if there are certain kinds of products that get reordered and if there are any sort of relationship between them all. This analysis answers the question of 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.

Tests for Question 2:

To check if there is any data integrity problem I decided to use excel to append tables and merge them. There were missing order details in other tables but we just had that to work with.

Q1)At 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 on 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)Which products are sold the most?

To get to know what products were frequently bought I used the joins to join tables with product_id and got a table of product, count(product_id) from order_products_prior table. With this I built a packed bubble diagram for top 10 and bottom ten.

Q5)If the most sold products were most reordered products?

To get this answer I used the sum of reordered for each product and made a horizontal bar graph to compare with most sold products. They were the same except for raspberries in most reordered products.

Q6)Top 10 and bottom 10 department

I used tableau to join two tables to get the department and the count of orders for each department to build charts.

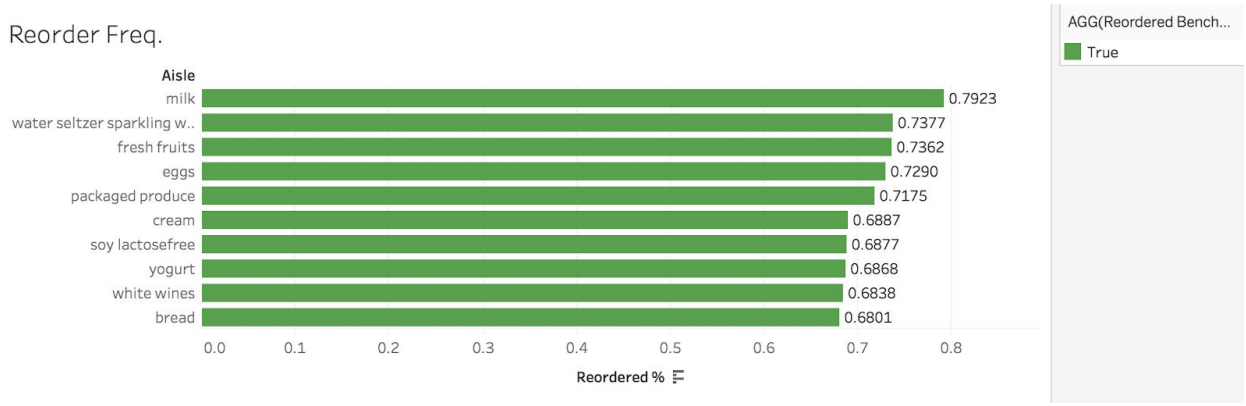
Q7)Percentage of reorder in a particular order

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.

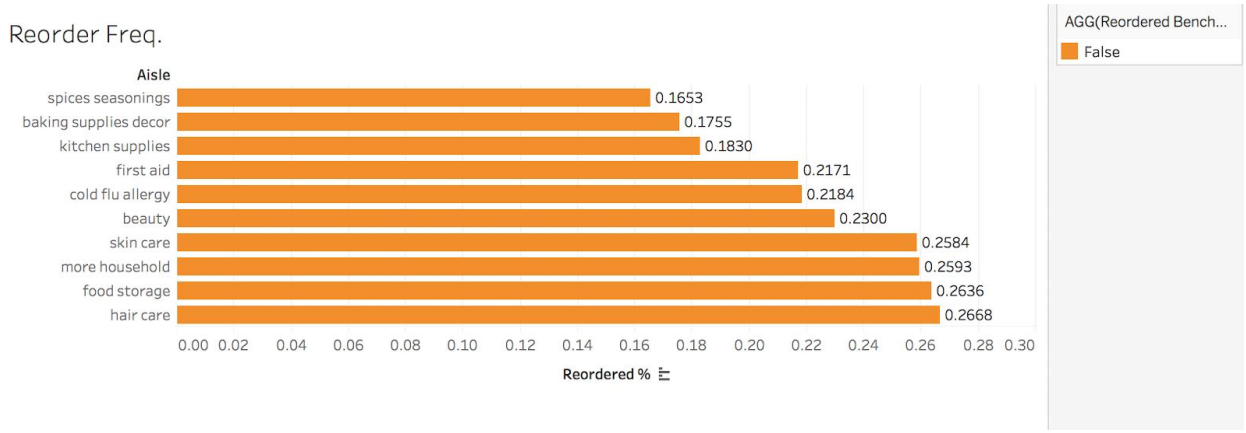
Q8)Market basket analysis
To know which department should be placed where strategically, I used market basket analysis. I referred to videos on how to run a market basket analysis and created a dashboard to select different departments and see what was bought along with it the most number of times.

Tests for Question 3:

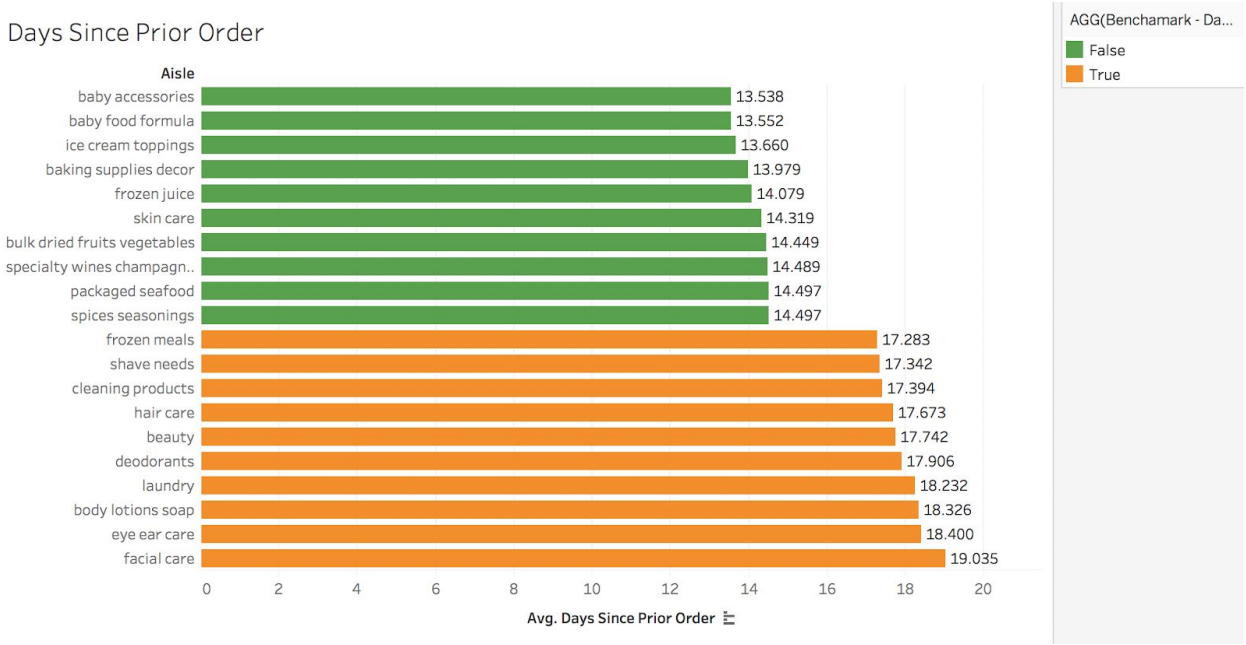
4. Summarize Results - Each explain results for our own question
Question 1 Results:



Above 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 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.



On the reverse side, the above graph shown in orange, 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.



The final analysis on the frequency of items reorder looks at the average number of days between reorders for these items, shown above. 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.

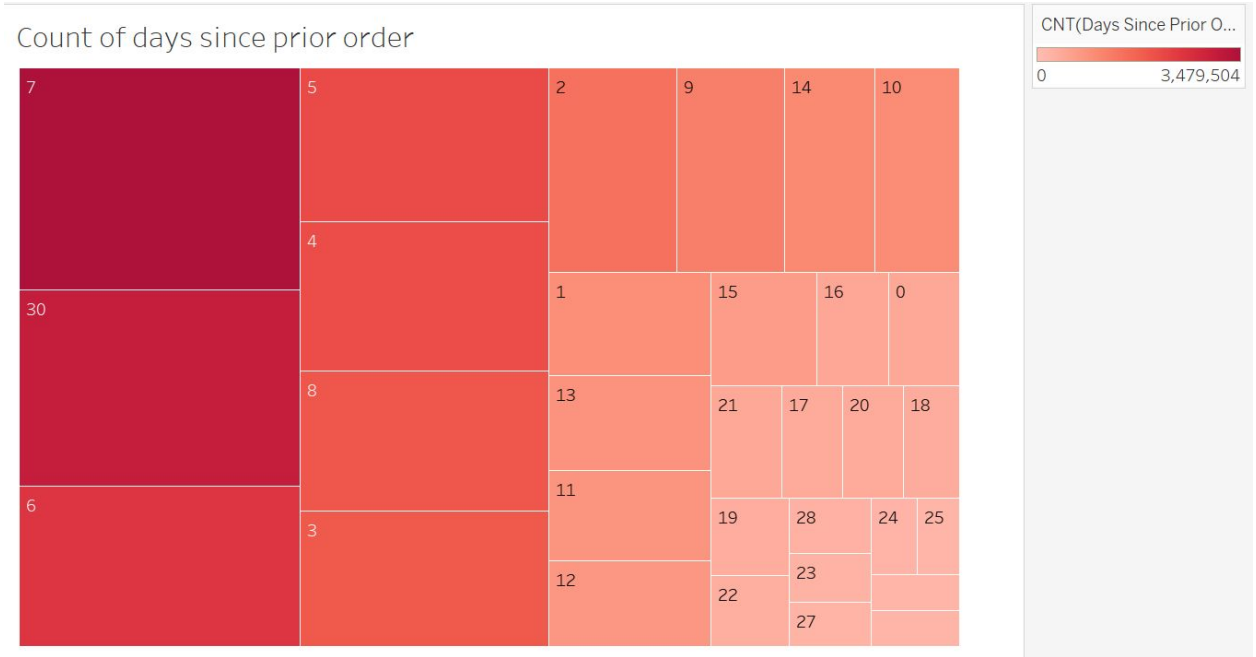
Using the same graph we can see that facial care items, eye & ear care items, soap, laundry items, deodorant, beauty items, hair care, cleaning products, shave needs, and frozen meals get reordered with the longest time in between reorders. These can all be explained by them having long shelf lives and 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.

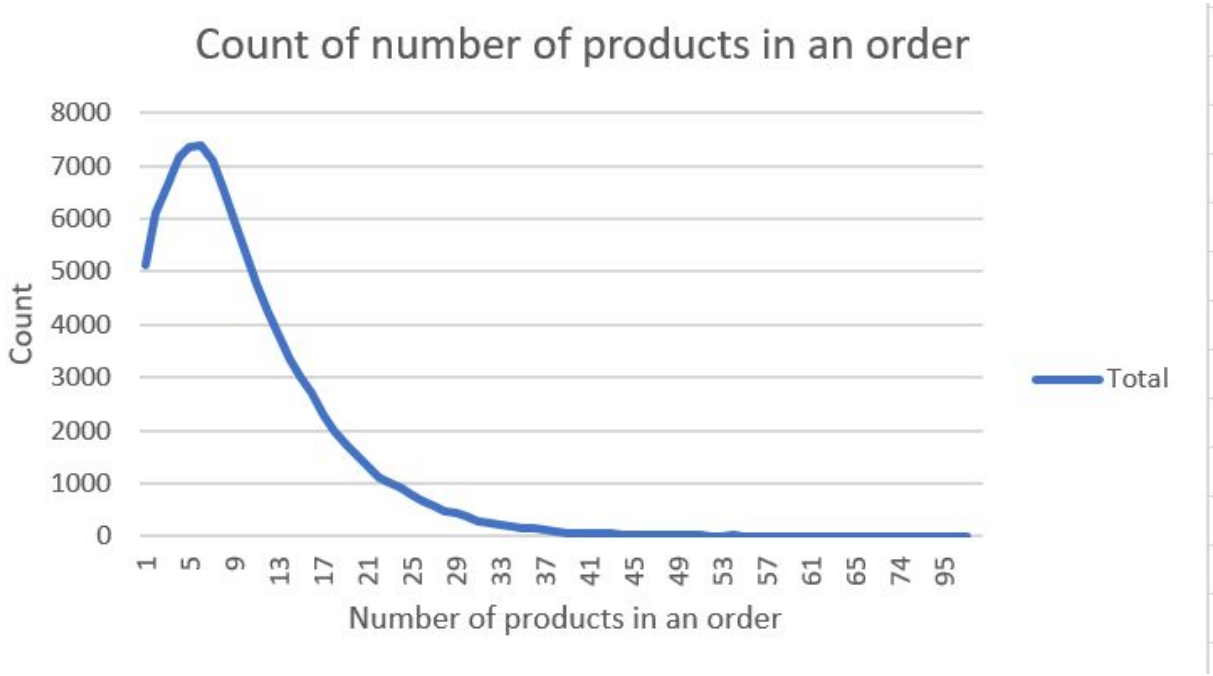
Question 2 Results:

At what time and day of the week do people order the most?
Most of the orders are placed from 10 am to 11am and 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.

How often do people reorder?
Looks like people reorder within 7 days more frequently followed by 30 days being the next frequent time period.

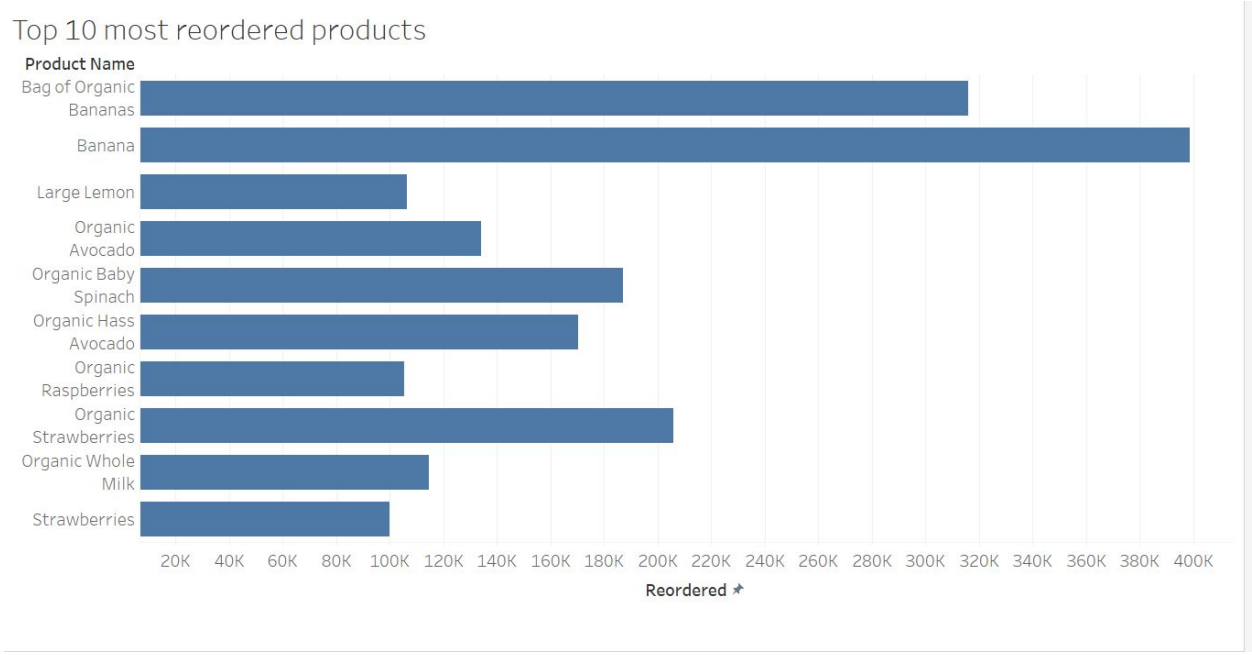


How many products are usually ordered at once?
Mostly 7 to 9 products are ordered at once.

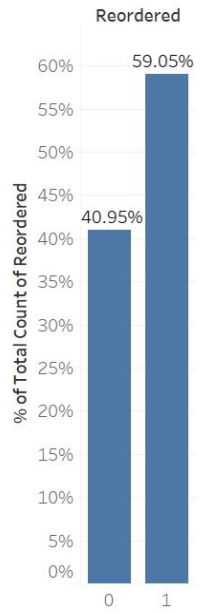


Which products are sold the most?

What were the top 10 reordered products?

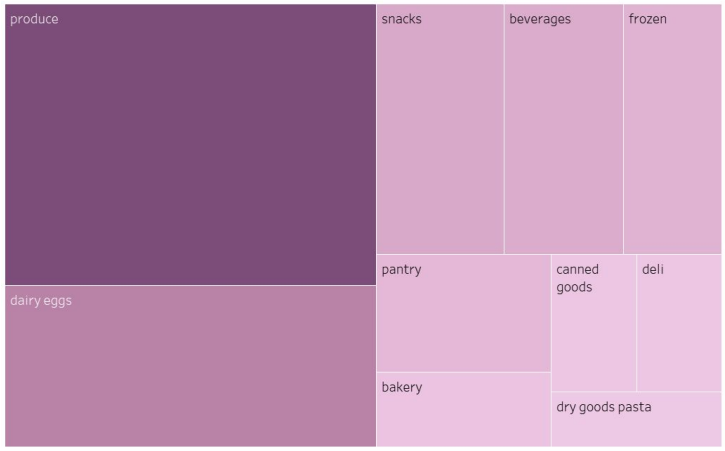


Proportion of reorders and not reordered



Top 10 and bottom 10 departments departments

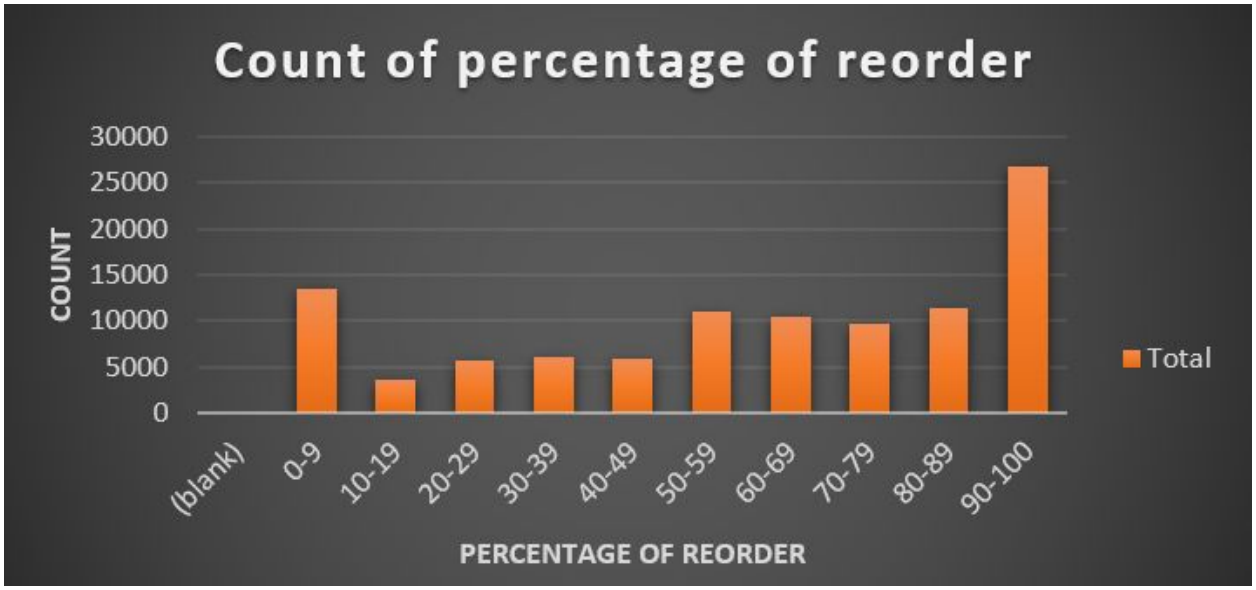
Top 10 department



Bottom 10 department

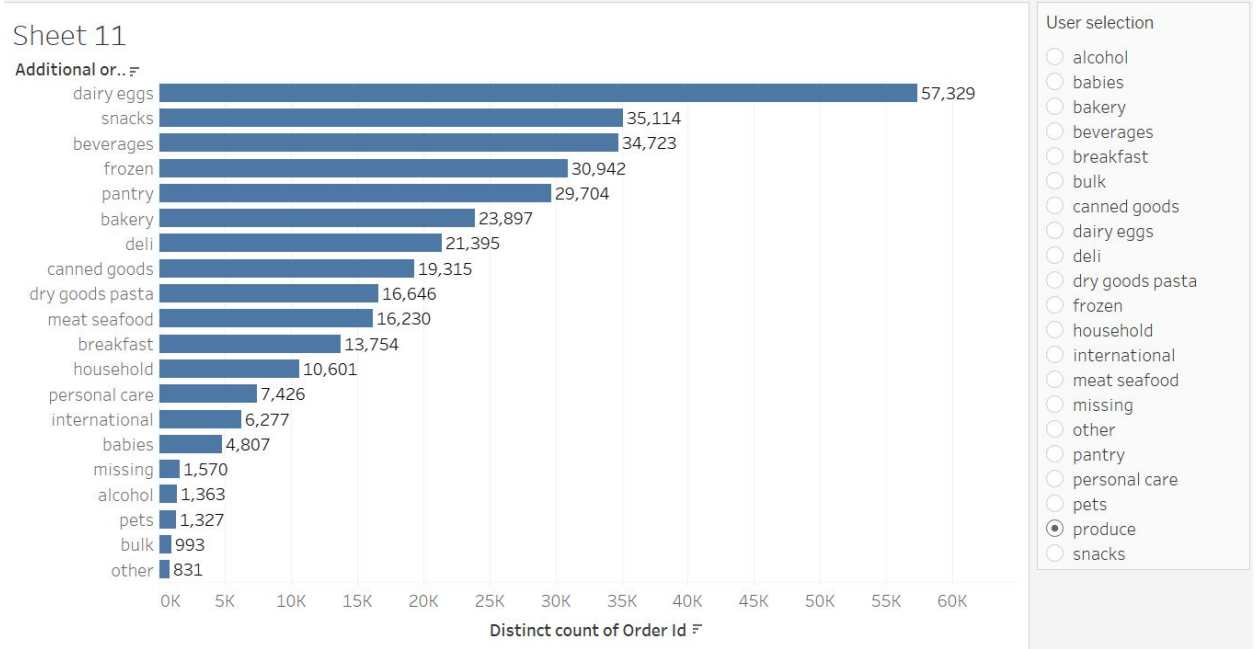
Department	
bulk	1,133
other	1,180
missing	2,239
pets	3,306
alcohol	4,856
international	8,778
babies	13,775
personal care	14,577
meat seafood	22,958
breakfast	23,007

What percentage is usually recorded?
A lot of orders have 90-100% reorder products.

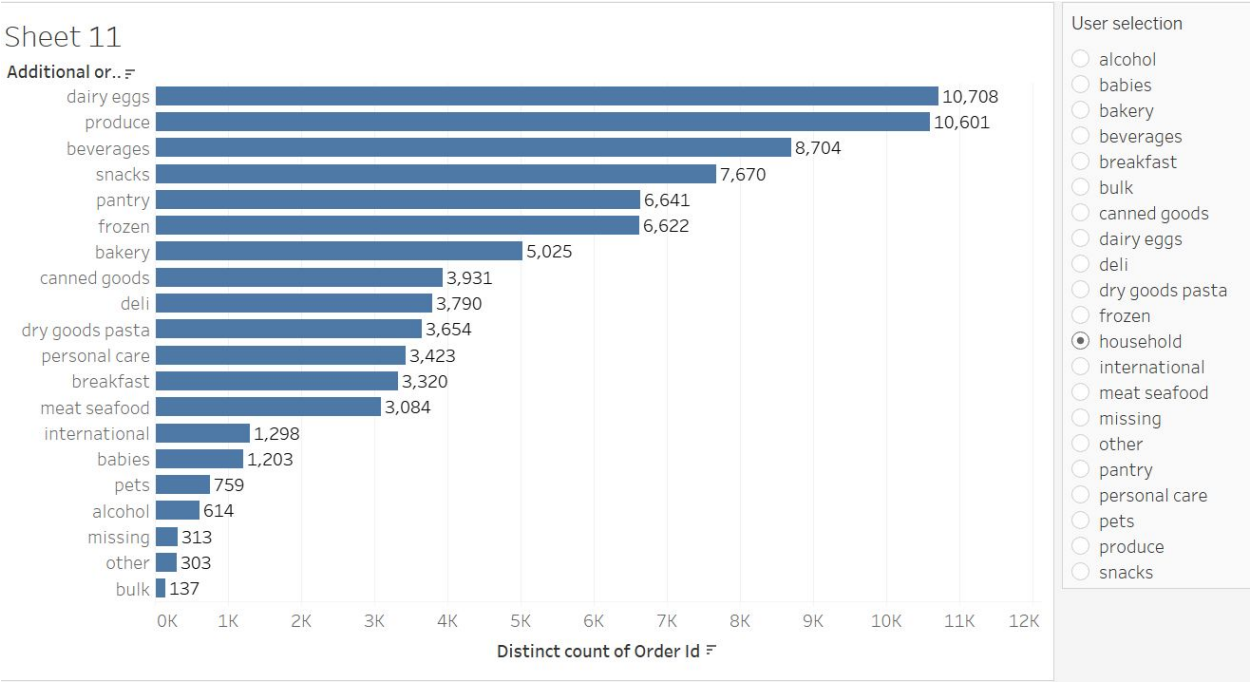


Market Basket analysis

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



Similarly households will mostly be bought along with eggs,veggies,beverages.

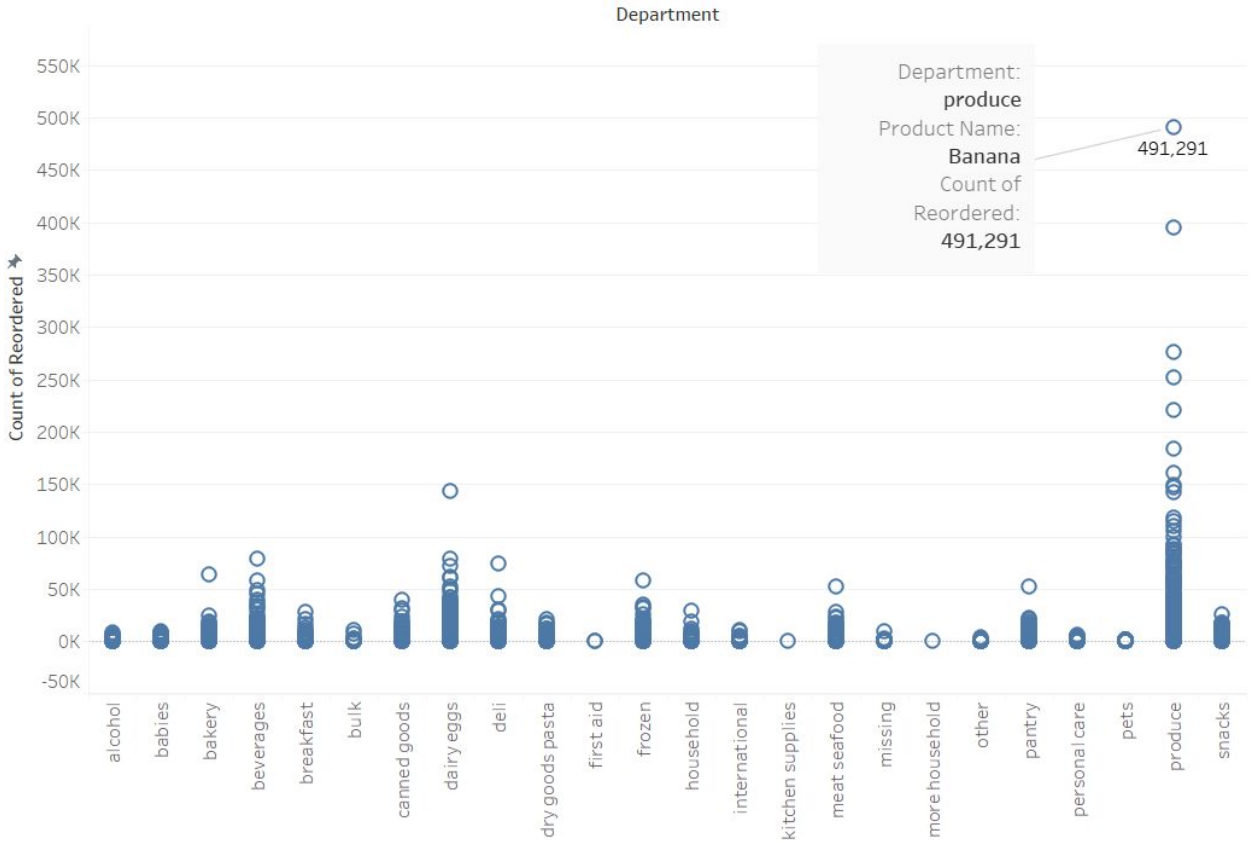


Tests for Question 3:

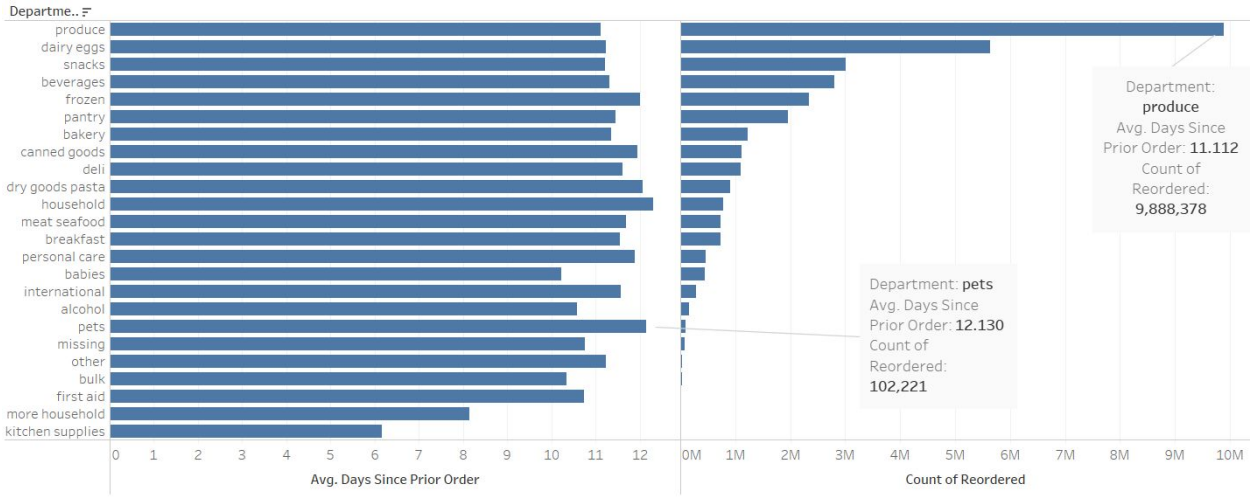
To understand what products were reordered the most and from which departments and time of the day, I joined all the secondary tables to form a primary table for the analysis.

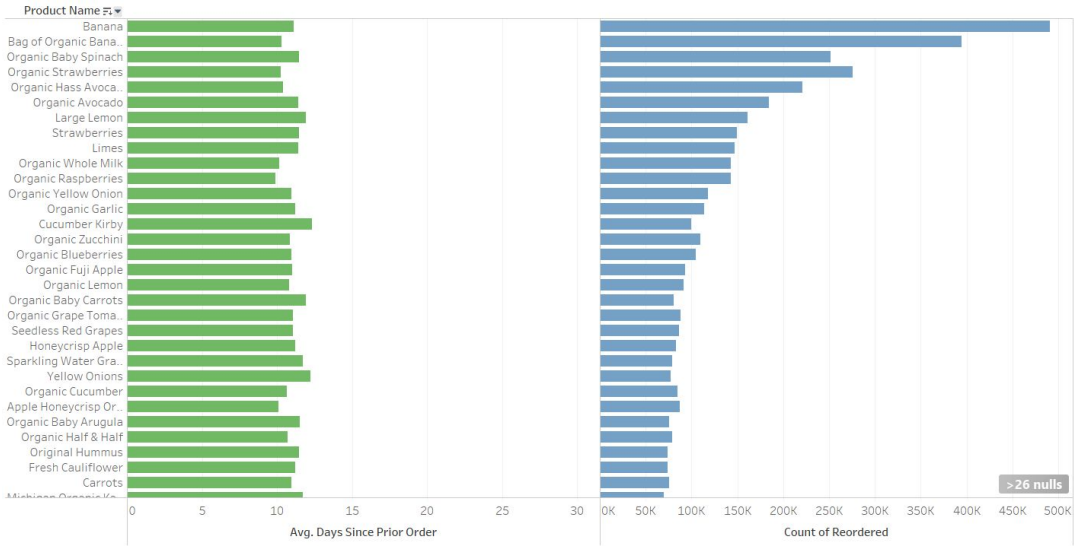
Q1)Which products and in which department make more sales on reorders?

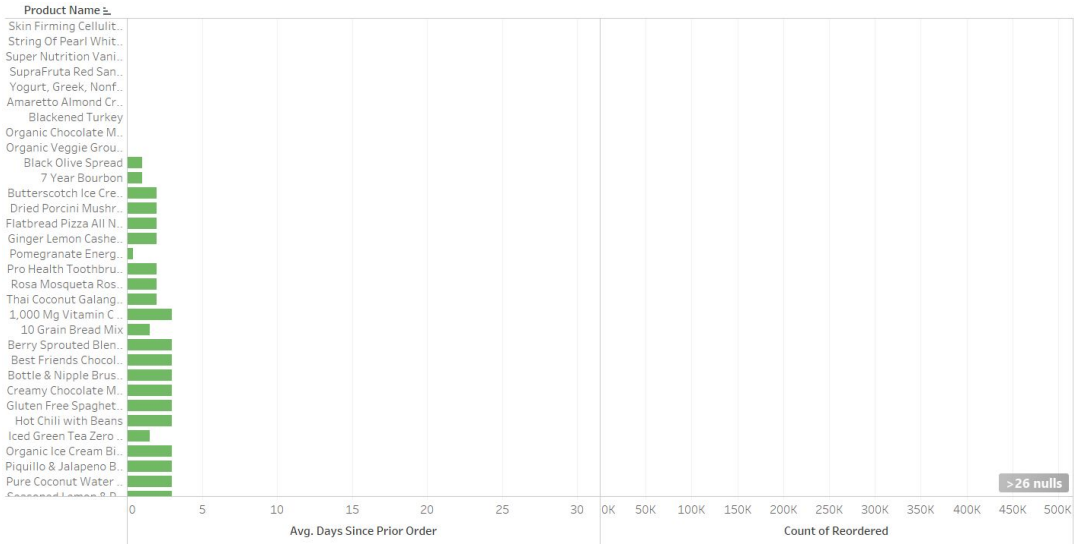
In order to answer this question, I used products, orders, order_products__train, order_products__prior tables to do a group by and measure. We can see that the banana in produce is sold the most - 491,291.



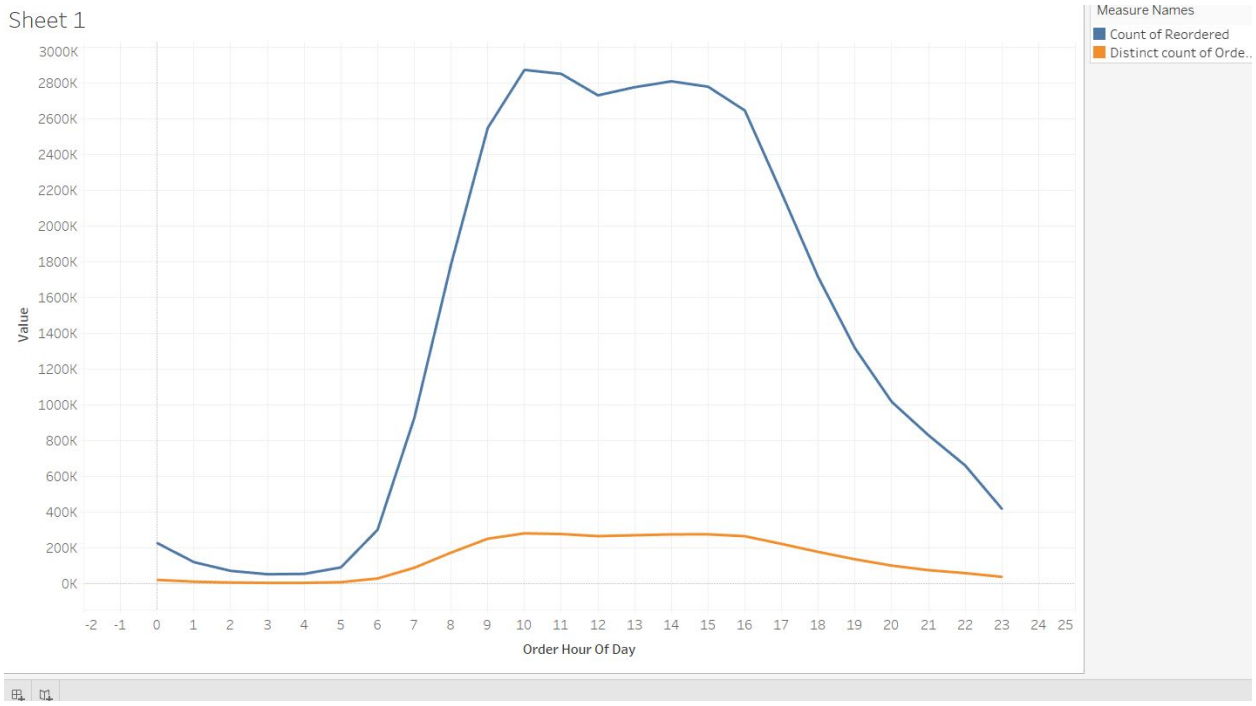
Q2) What could be the reason behind the produce department having more reorders? Which other departments have low reorders and what additional attribute supports this answer?
If plotting the graph of departments and reorders, we can clearly see that the produce department generally has a greater reorder 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.


Departme..
produce
dairy eggs
snacks
beverages
frozen
pantry
bakery
canned goods
deli
dry goods pasta
household
meat seafood
breakfast
personal care
babies
international
alcohol
pets
missing
other
bulk
first aid
more household
kitchen supplies
0 1 2 3 4 5 6 7 8 9 10 11 12
Avg. Days Since Prior Order
0M 1M 2M 3M 4M 5M 6M 7M 8M 9M 10M
Count of Reordered
Department: produce
Avg. Days Since Prior Order: 11.112
Count of Reordered: 9,888,378
Department: pets
Avg. Days Since Prior Order: 12.130
Count of Reordered: 102,221


Product Name
Banana
Bag of Organic Bana..
Organic Baby Spinach
Organic Strawberries
Organic Hass Avoca..
Organic Avocado
Large Lemon
Strawberries
Limes
Organic Whole Milk
Organic Raspberries
Organic Yellow Onion
Organic Garlic
Cucumber Kirby
Organic Zucchini
Organic Blueberries
Organic Fuji Apple
Organic Lemon
Organic Baby Carrots
Organic Grape Toma..
Seedless Red Grapes
Honeycrisp Apple
Sparkling Water Gra..
Yellow Onions
Organic Cucumber
Apple Honeycrisp Or..
Organic Baby Arugula
Organic Half & Half
Original Hummus
Fresh Cauliflower
Carrots
0 5 10 15 20 25 30
Avg. Days Since Prior Order
0K 50K 100K 150K 200K 250K 300K 350K 400K 450K 500K
Count of Reordered
>26 nulls


Product Name
Skin Firming Cellulit..
String Of Pearl Whit..
Super Nutrition Vani..
SupraFruta Red San..
Yogurt, Greek, Nonf..
Amaretto Almond Cr..
Blackened Turkey
Organic Chocolate M..
Organic Veggie Grou..
Black Olive Spread
7 Year Bourbon
Butterscotch Ice Cre..
Dried Porcini Mushr..
Flatbread Pizza All N..
Ginger Lemon Cashe..
Pomegranate Energ..
Pro Health Toothbru..
Rosa Mosqueta Ros..
Thai Coconut Galang..
1,000 Mg Vitamin C ..
10 Grain Bread Mix
Berry Sprouted Blen..
Best Friends Chocol..
Bottle & Nipple Brus..
Creamy Chocolate M..
Gluten Free Spaghet..
Hot Chili with Beans
Iced Green Tea Zero ..
Organic Ice Cream Bi..
Piquillo & Jalapeno B..
Pure Coconut Water ..
0 5 10 15 20 25 30
Avg. Days Since Prior Order
0K 50K 100K 150K 200K 250K 300K 350K 400K 450K 500K
Count of Reordered
>26 nulls

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.



Test for question 4
Q1: Which products are sold the most?

I am using tableau for this task**.** 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, Bag of Organic bananas, organic Strawberries are very popular among other products.

Q2: Among all products, Which products are more likely to be in added_to_cart_order from which Aisle, which department? Thoughts on organic products.

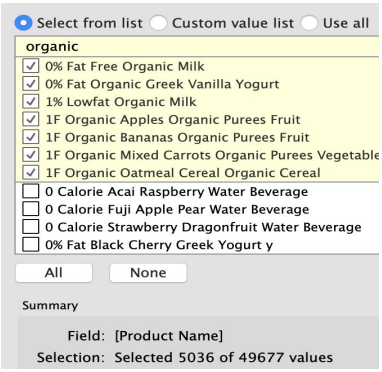
Sheet 1

Product Name	Aisle	Department	
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

0K 200K 400K 600K 800K 1000K 1200K 1400K 1600K 1800K 2000K 2200K 2400K 2600K

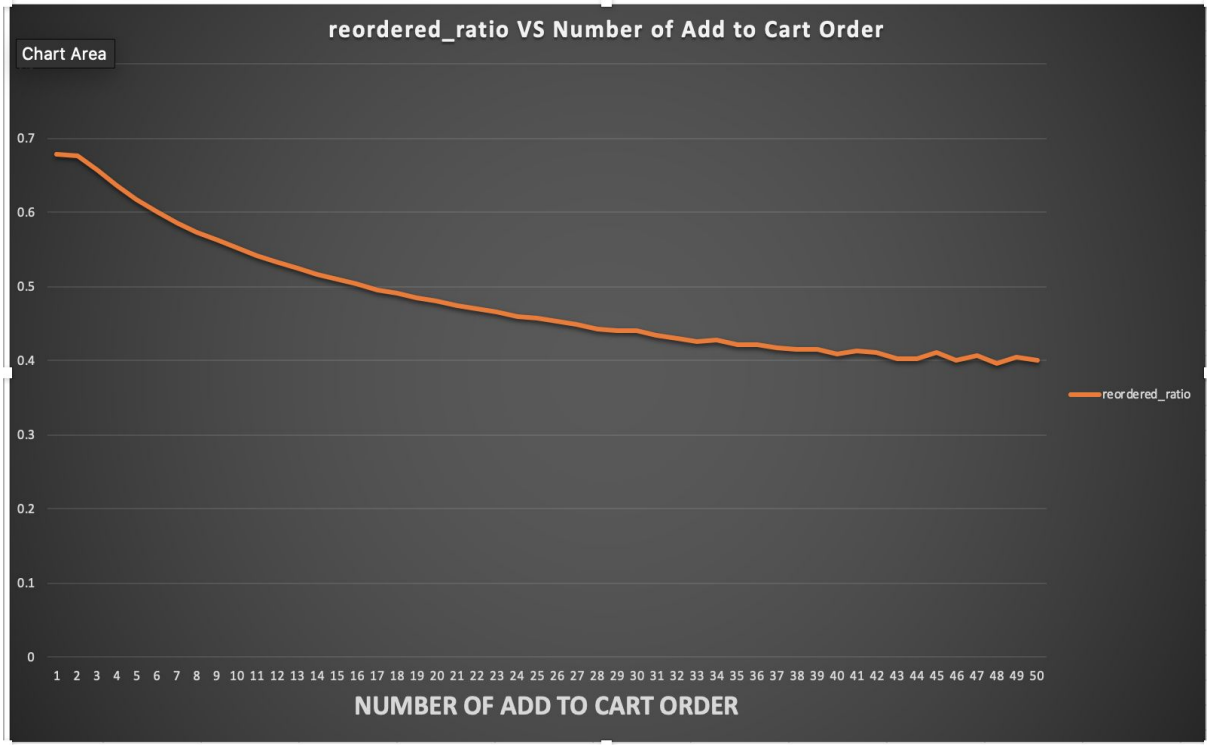
Add To Cart Order

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 49667 types of different product names. 5036 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 184864556. Number of organic_add to cart orders is 85998311. Next, plot the pie chart to show the proportion organic and non-organic related add to cart order.



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.

Q3: What's the relationship between add_to_cart order and reorder order? Any specific pattern?If there is, please explain.



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.

From this plot, we observe that add to cart orders products are more likely to be reordered VS products that are added later. I would suggest Instacart to remind their customer in the early period of time.

5. Conclusion/Recommendations

Question2 Conclusions/Recommendations

Daily used products like vegetables, fruits, eggs, snacks, and beverages should be kept close to each other and close to the entrance and less frequently used products like alcohol can be kept at the end.

Also reordered products are also mostly fruits & vegetables.

Most of the orders are placed on Sunday and between 7am to 5pm.

Most of the orders have 90-100% reorders.

Most of the reorders are placed after 7 days on average.

Question3 Conclusions/Recommendations

1. We can see that the banana in produce is sold the most - 491,291. Produce based products have greater reorder rate

2. 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
3. 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.

Question 4 Conclusions/Recommendations

1. Send out notification to customers to remind them that they still have product in their shopping cart.
2. Organic products are really popular, I would suggest Instacart add more organic products to their inventory.
3. we observe that add to cart orders products are more likely to be reordered compared with products that are added later. I would suggest Instacart to remind their customer in the early period of time and build a recommendation system based on the customer's most recent order.
4. I would suggest Instacart to open more promotion events during the weekend. Because most of their customer will be reorder during the weekend.

Link-- https://www.kaggle.com/c/instacart-market-basket-analysis/overview/evaluation