Data Modeling & Structured Analysis: Final Project

Instacart analysis

Mimi Liu

2024

Contents

[Introduction 2](#_Toc161097162)

[Executive Summary 2](#_Toc161097163)

[Section One: Descriptive Statistics 3](#_Toc161097164)

[5](#_Toc161097165)

[Section Two: User Behavior 5](#_Toc161097166)

[Section Three: Product Information 7](#_Toc161097167)

[Section Four: Market Basket Analysis 8](#_Toc161097168)

[Conclusions & Recommendations 9](#_Toc161097169)

[Appendix: SQL Statement per Table 10](#_Toc161097170)

# Introduction

Instacart is known nationally as a grocery delivery and pickup service, operating mainly throughout the United States. It is an online-based system that allows users to select products from many different grocery stores and have their items delivered by another person, their “shopper”, to the door. The company’s aim is to make it easy for the users to order and receive their groceries when they need and without having to leave the comfort of their homes. This report is to analyze their consumer purchase data and gain insight from the data to help the company make better decisions so they can keep their competitive advantage. The objective of this purchase database is to predict which products will be in a user’s next order. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, Instacart provided between 4 and 100 of their orders, along with the sequence in which products were placed in the cart.

# Executive Summary

Throughout this report, it will be analyzing many different areas of Instacart’s purchase database. The database consists of f tables: In\_order\_products, In\_products, In\_aisles, In\_orders, and In\_departments. In the first section of this report, it will be diving further into the tables with explanations of the attributes and giving insight into the descriptive statistics, such as the number of orders, number of different customers, count of products and department, frequency of orders based on time and day of the week, etc. For example, based on a quick query, it is noted that the most orders are placed during work hours of 10 am to 6 pm. This hour period may likely be due to the fact that it is also when most groceries stores are opening. High order periods are from 9 am to 6 pm, which are work hours for most professionals and before dinnertime.

The next section is analyzation of their users’ behavior. The section will answer questions on types of users, their preferences, average spending, and much more. This is highly important since current users, aka Instacart’s customers, can show the company what many of the non-users would want, but have just not yet joined. The strategy focus will the social aspect of society as in what are the popular products are usually ordered. Looking at the In\_order table, it can be seen that most users, about 10,430 of them, place orders every 4 days. There could be many reasons for that, and it can all depend on age, job, and household size of the users, for example, a single young professional could be ordering on Sunday for meal prepping for the week or a family with kids could go through perishables quickly and would need to restock heading into the weekend.

The third section and last sections will be all about the products and the market base analysis of them. As Instacart’s platform is mainly about which groceries and other service stores it should be offering their users and at what prices, this section is just as important as the customer one since they go hand in hand. The report will compare the relationships between the top five types of users and products ordered. This will be a great summary of where, what, and who Instacart should focus their target market on. So therefore, using this database, this report will explain the insights gained from each entity and their relationships with each other. There will be summed up in a few recommendations for how Instacart can make better business decisions and optimize their predictions feature for current and future orders for their users.

# Section One: Descriptive Statistics

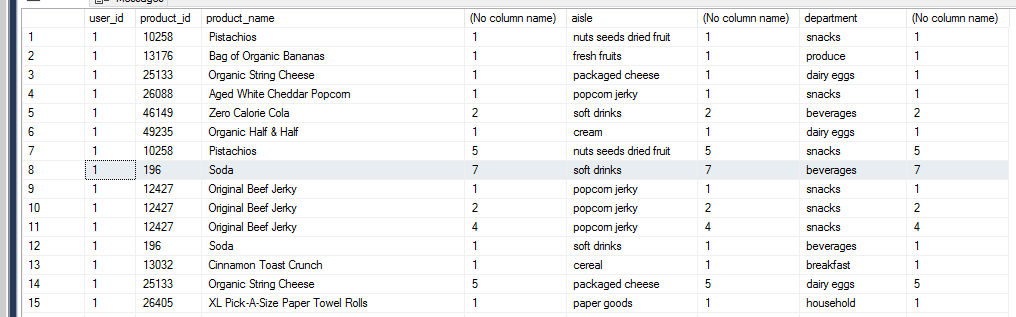
Out of 10,000 users (based on distinct user\_id in the in\_order table), on average they would order on a Tuesday at 1 pm and places an order every 11 days. Statistically speaking, this can mean that many users are working professionals who would do their shopping at their lunchtime towards the beginning of the week.

*A screenshot of a computer

Description automatically generated1.1 Averages of order for day and hour of the orders and days between each order*

Let’s also look at the reminder of the tables for statistics on the products and their identifiers of 49,688 different products, 134 aisles, and 21 departments. Based on the database for reordered products, some examples to be seen in the below query, user 1 likes to snack as their highest orders are snacks and beverages (1.2), user 2 maybe part of a family as their order have many products and is spread out in number of times ordered between one and three (1.3). Or user 1001 could likely be a young single person as their list of products ordered is small compared to the rest and likes to eat breakfast at home (1.4).

*1.2 Data of Product for User 1’s Order*



*1.3 Data of Product for User 2’s Order (Only half is shown)*



*1.4* *Data of Product for User 1001’s Order*

# 

As there is now a basic understanding of the purchase dataset, the report can go into details of user behavior, purchased product movement, and the market-based analysis and what each of the findings can mean.

# Section Two: User Behavior

On a user level analysis, most of the users only made few orders ranging from one to three. The number of orders a user made decreased significantly along the order numbers. The maximum orders any users have made is 99.

*A screenshot of a computer

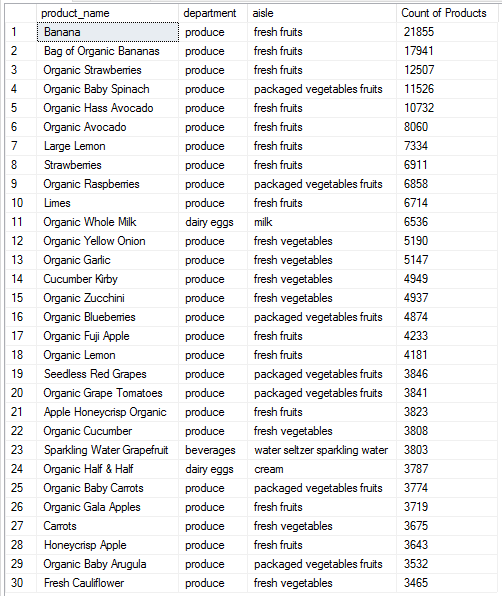
Description automatically generated2.1 Table of Count of Users by their Maximum Number of Orders*

It is also obvious that most users made their orders either weekly (every 7 days) or monthly (every 30 days) (2.2) and as mentioned in the previous section, most users typically like to shop on a Tuesday around lunchtime as it could mean it is around the time frames where errands are usually run (1.1). Another quick insight on user behavior is that out of 136,824 different orders, approximately 9,853 users, or 7.2%, would go about reordering the same products. Some broad implications, as deeper insight is beyond the scope of this section, on this would be that these 7% users are reordering the everyday essentials, such as produces, snacks, or household goods.

A screenshot of a computer

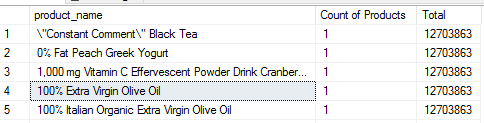
Description automatically generated*2.2 See the peak of day is 7 and day 30.*

One last set of analysis investigated would be what were the most popular products, say top 30, purchased. As seen in the chart below bananas, strawberries, and spinach were among the top of the tier (2.3). As reported in the previous section, these are grouped into the produce department in the fresh fruits and vegetables aisle (see 1.2 to 1.4)

*2.3 Top 30 List of Products in Users’ Orders*

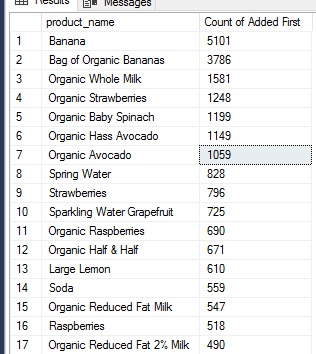
# Section Three: Product Information

As this section ties into both section one and two, many of the mentioned insights and analysis will likely have been briefly mentioned before. To start, let’s investigate how often a product is purchased. As seen in the table below (3.1) and like table 2.3, the top 5 most purchased items are bananas, strawberries, spinach, and avocados, purchased 74,561 times out of 12,703,863 purchased items, or about 0.6%; whereas the bottom 5 least purchased items are black tea, Greek yogurt, vitamin c powdered drink, and olive oil, each purchased only 1 time (3.2). The most likely reason for this is that the popular products are produced that a big population of users can eat at any time of the day and can be used in many ways for cooking and usually are brought in small quantities. Therefore, the opposite can be said for the least purchased: brought in bulk or not frequently used or even just a special one-time buy.

A screenshot of a computer

Description automatically generated*3.1 Tables of Most (left) and 3.2 Least Purchased Products (right)*

Next analysis will be looking into the orders, such as how many orders (in percentages) would contain the top 5 items of banana, strawberries, spinach, and avocados and how often a product is the first to be added to cart. Based on the SQL query, the number of orders that contains the top 5 items is 49,618, which is about 0.4% of the total. This is important as this means many of the future orders would likely have these produces added to their cart and could be useful to have this listed on a most frequently purchased predicative list. Lastly, as can be predicted, the table below shows that the previously mentioned top 5 most popular products are also ones that are most frequently to be added first to the cart. Top sold items for most shoppers are known to be the first to be on the list.

*3.3 Frequency of Products Added First to the Cart*

# Section Four: Market Basket Analysis

In this last section, market basket analysis was performed to propose recommendations by showing the support, confidence, and lift among the top 5 most ordered products mentioned throughout this report. The table below shows that in the Support column there is a 0.19 support ratio that when a user orders 13176, which is Organic Banana, then he/she will also order 47209, which is Avocados. The next highest support ratio is 0.16 for ordering 24852, a regular banana, also ordering 21137, which are strawberries.

Next, for the confidence rule of the above-mentioned supports. From the table below in the Confidence column, there are about 17 times where the situation could happen with the user ordering Organic Bananas and Avocados. Or 11 times where the user orders regular bananas and strawberries. So far, these rules are quite positive and signifies a high chance of the event occurring and recommendation-wise, Instacart should always feature these produces in their predictive list for all the grocery stores that sells these items.

Last, but certainly not least, the lift rule or the ratio of support to confidence. In the same table below, in the lift column, the top 2 correlations are, of course, Organic Bananas to Avocados, but unpredicted, the regular bananas to strawberries does not have as strong of a correlation comparatively to the rest in the top 5, is still quite strong if among the rest of the dataset.

*A table of numbers and text

Description automatically generated with medium confidence4.1 Table Showing Support, Confidence, Lift Ratios of Top 5 Ordered Products*

# Conclusions & Recommendations

Recommendations were made throughout this report, so here are some summary points to wrap up all four sections as a conclusion to our analysis:

* Out of 10,000 users, most orders are placed on a Tuesday at around 1 pm.
* Based on various users’ ordering habits, there can be seen so far of three different one:
  + One who likes to snack, a parent, and a single, young professional, who eats breakfast at home, but out for dinner.
* There is a maximum of 99 orders by a single user and the most frequent ordering times between each new order is either 7 or 30 days.
  + Most have only ordered 1 to 3 times so far.
* Produces have a high chance of getting added to cart first and getting reordered.

So, the overall recommendation would be to have a profile setting of the category of the user, so in their predictive items feature for their next order, it could be the top 5 produces of bananas, strawberries, spinach, and/or avocados or drinks such as sodas and nuts. In addition, there could also be a reminder made out according to the user’s ordering habits either every 7 days or 30 day at 1 pm on a Tuesday, for example.

# Appendix: SQL Statement per Table

use A\_Instacart\_project;

select \*

from in\_order\_products;

select \*

from in\_orders;

select \*

from in\_products;

select \*

from in\_aisles;

select \*

from in\_departments;

--Query for Table 1.1

select AVG(order\_dow) as [Avg. day of the week], AVG(order\_hour\_of\_day) as [Avg. Order Time], AVG(days\_since\_prior\_order) as [Avg. Days bt Orders]

from in\_orders;

--Query for Section 2 data (No table in report)

select days\_since\_prior\_order, COUNT(days\_since\_prior\_order) as [Count of Prior Order Days]

from in\_orders

group by days\_since\_prior\_order

order by [Count of Prior Order Days] desc;

--Query for Tables 1.2 to 1.4

select distinct o.user\_id, op.product\_id, p.product\_name, COUNT(p.product\_name), a.aisle, COUNT(a.aisle), d.department, COUNT(d.department)

from in\_orders o join in\_order\_products op on o.order\_id=op.order\_id

join in\_products p on op.product\_id=p.product\_id

join in\_aisles a on p.aisle\_id=a.aisle\_id

join in\_departments d on p.department\_id=d.department\_id

group by o.user\_id, op.product\_id, op.add\_to\_cart\_order, op.reordered, p.product\_name, a.aisle, d.department

having reordered=1

order by user\_id;

--Query for Table 2.1

select count(distinct user\_id) as [Count of Users], max(order\_number) as [Highest No. of Orders per User]

from in\_orders

group by order\_number

order by [Highest No. of Orders per User];

--Query for Table 2.2

select count(order\_number) as [Count of Orders], MAX(days\_since\_prior\_order) as [Days Since Prior for Each Order]

from in\_orders

where days\_since\_prior\_order is not null or order\_number>1

group by days\_since\_prior\_order

order by [Count of Orders] desc;

--Query for Table 2.3

select top 30 p.product\_name, d.department, a.aisle, COUNT(op.add\_to\_cart\_order) as [Count of Products]

from in\_orders o join in\_order\_products op on o.order\_id=op.order\_id

join in\_products p on op.product\_id=p.product\_id

join in\_aisles a on p.aisle\_id=a.aisle\_id

join in\_departments d on p.department\_id=d.department\_id

group by product\_name, aisle, department

order by [Count of Products] desc;

--Queries for Section 3 (no table in report)

select count(distinct o.user\_id), count(distinct op.order\_id)

from in\_order\_products op left join in\_orders o on op.order\_id=o.order\_id

where reordered = 1;

select count(op.order\_id) as [No. of Orders], (select sum(add\_to\_cart\_order) from in\_order\_products) as Total

from in\_order\_products op left join in\_products p on op.product\_id=p.product\_id

where product\_name like '% banana' or product\_name like '% strawberries' or product\_name like '% bananas' or

product\_name like '% spinach' or product\_name like '% avacados';

--Query for Tables 3.1 & 3.2

select p.product\_name, COUNT(op.add\_to\_cart\_order) as [Count of Products], (select sum(add\_to\_cart\_order) from in\_order\_products) as Total

from in\_order\_products op left join in\_products p on op.product\_id=p.product\_id

group by product\_name

order by [Count of Products];

--Query for Table 3.3

select p.product\_name, COUNT(op.add\_to\_cart\_order) as [Count of Added First]

from in\_order\_products op left join in\_products p on op.product\_id=p.product\_id

where add\_to\_cart\_order=1

group by product\_name

order by [Count of Added First] desc;

--Queries for Table 4.1

select \*

from in\_products

where product\_id in (13176,21137,21903,47209);

with support as (

select op1.product\_id as [Product 1], op2.product\_id as [Product 2],

COUNT(1) as Frequency,

(select count(order\_id) from in\_order\_products) as [Total Orders],

(select count(order\_id) from in\_order\_products op where op1.product\_id=op.product\_id) as [Frequency LHS],

(select count(order\_id) from in\_order\_products op where op2.product\_id=op.product\_id) as [Frequency RHS]

from in\_order\_products op1

join in\_order\_products op2

on op1.order\_id=op2.order\_id

where (op1.product\_id > op2.product\_id) and (op1.product\_id in (24852,13176) and op2.product\_id=21137) or (op1.product\_id in (24852,13176) and op2.product\_id=21903)

or (op1.product\_id in (24852,13176) and op2.product\_id=47209) or (op1.product\_id=21137 and op2.product\_id=21903) or (op1.product\_id=21137 and op2.product\_id=47209)

or (op1.product\_id=21903 and op2.product\_id=47209)

group by op1.product\_id, op2.product\_id

)

select [Product 1], [Product 2], Frequency, FORMAT(Frequency\*100.00/[Total Orders], '0.##') as Support, FORMAT(Frequency\*100.00/[Frequency LHS], '0.##') as Confidence,

Format(Frequency\*100.00/[Total Orders]/([Frequency LHS]\*100.00/[Total Orders])\*([Frequency RHS]\*100.00/[Total Orders]), '0.##') as Lift

from support

order by Frequency desc;