# Trends and Patterns in Instacart Data

José Gutiérrez
Computer Science
University of Colorado, Boulder
Boulder, Colorado, USA
jogu0215@colorado.edu

Anna Malawista
Computer Science
University of Colorado, Boulder
Boulder, Colorado, USA
anma7757@colorado.edu

Michael Whitlock
Computer Science
University of Colorado, Boulder
Boulder, Colorado, USA
miwh5549@colorado.edu

Keaton Hoeger
Computer Science
University of Colorado, Boulder
Boulder, Colorado, USA
keho3676@colorado.edu

Aidan O'Connor
Computer Science
University of Colorado, Boulder
Boulder, Colorado, USA
aioc7481@colorado.edu

## 1 Problem Statement/Motivation

Instacart is an online grocery delivery service. Our group's motivation is to analyze patterns from Instacart orders. We plan to determine if there are trends in the types or rate of products ordered depending on the time of day, week, and frequency of repurchases. Potential questions include: "Which items are most likely to be added order, based on the items already in the cart?"; "Based on purchase history, can we predict which new items a customer is most likely to purchase?"; and "Can we predict the length of time between orders?"

Ideally, our research could be used to help companies determine how best to target and advertise to customers, what products to stock in the highest quantity and when, and possibly area where the market appears to be untapped. One very interesting area for possible future research would be to compare these finding to similar measures for standard grocery stores (purchased from in person) to determine what, if any, differences are seen when grocery purchases are provided through a digital, delivery service. The potential applications for this research is quite broad, and trends that are found could have implications or not only Instacart but also any other grocery and food delivery service, as well as any companies selling products through these services.

## 2 Literature Survey

Instacart posted the dataset on their website and a brief description on Medium [1, 2]. We have found that most of the prior work on this dataset is exploratory in nature. Instacart ran a competition on Kaggle in early 2016. Various analyses of this dataset were conducted as a result of this competition. One of the primary areas of research done for this competition looked at the trends related to repurchasing [3, 4].

Instacart's Medium post provides two short examples of patterns derived from the dataset; one on repurchasing and item popularity, and the other on the difference in "healthiness" of orders depending on the time of day [2]. This data (and other similar datasets) have also been used in a number of studies looking at buyer recommendation and customer targeting strategies [5, 6, 7].

## 3 Proposed Work

Since the dataset appears to be relatively clean and does not contain many missing values, we will begin by using exploratory data analysis techniques to examine the data. This analysis will use charts, graphs, and other visualization methods to give us an initial idea of the relationship between attributes. The Instacart dataset is comprised of 6 tables that are linked by ID attributes for orders, departments, etc. In order to complete our project, we will need to join the tables by these attributes, so we can fully analyze all the order information available. We have 3.2 million data points in total for analysis. There are pre-designated subsets of data for training and testing machine learning models.

Ideally, our group will implement some simple machine learning and data mining techniques to predict future order activity.

## 4 Data Set

Our dataset is from Instacart, a company which created an app that provides a grocery ordering and delivery service. This dataset is made up of 6 different CVS files [1]. The dataset is very clean and simple to follow; there are no null or missing values. According to Instacart, the data is anonymized and is a sample of over 3 million orders from more than 200,000 unique Instacart users [1]. There are prelabeled training and test sets. The data set provides information on what was ordered (including department and aisle location), the day of week and time of day that the order was placed, as well as the amount of time between orders. However, no information is provided on what time of year items were purchased, so no information can be drawn on the fluctuations in ordering across seasons. A "data dictionary" was provided by Instacart and is available on GitHub [8].

There are a few issues we have noted about this dataset. The purchase day of the week is provided as an integer from 0-6; however, thus far we have been unable to find any information on what day of the week each integer corresponds to. Others have documented this same issue, and most have concluded that based on the rate of ordering 0 represents the beginning of the weekend (Saturday) and that the days of the week are in order from that point [4,8]. This conclusion was drawn because there is a large spike in orders on day 0 and day 1, which one could attribute to the amount of free time and social gatherings people tend to have on the weekend. However, in the process of our initial research, we have discovered multiple pieces of data that have actually lead us to the conclusion that day 0 is in fact Monday. These findings are explained in the "Orders across Time" subsection of our results, which can be found on page 5 of this text.

One other issue we will face is that all data in the "days_since_prior" order attribute is capped at 30 and the number of orders for any user is capped at 99. Rather than truncating the data at these points, we have concluded that all data greater than these limits were simply assigned that max value (censored). This conclusion is based on a large spike in the data at 30 for "days_since_prior" and 99 for the number of orders per user. In most of our data visualization, we will be able to simply label this part of the data as >=30 or >=99. However, we are still assessing if we should try to account for the possibility of this censored data skewing our results through preprocessing, and if so, what the best method would be. Overall, this data set is large, clean, and fairly comprehensive; and it should provide ample opportunity for analysis and prediction.

## 5 Evaluation Method

We want to try using as many of the following analysis methods as can be reasonably applied: data visualizations (box plots, scatter plots and trend lines, histograms, etc.), data analysis (correlation coefficient, chi-square test, and other correlation statistics, association rules),

classification and prediction with machine learning. We intend on using the approximately 3.2 million order information for the analysis methods discussed above. Then for the machine learning models, there are additional orders for both training and testing our machine learning models of size 131,000 and 75,000 respectively. We intend to test multiple models in order to find the one that performs best and then optimize that model through hyperparameter tuning. While accuracy is likely a useful measurement for this dataset other measurements such as precision, recall, etc. will be considered for the final model. We will continue to modify and add to our intended evaluation methods as we learn more throughout the course.

## 6 Tools

We will mainly be using Python, as the language provides a lot of libraries that are helpful in both data analysis and any machine learning we want to attempt on the datasets. NumPy is one library we plan to use, which grants access to useful functions and features. We plan to use Pandas as the main Python library for data analysis, it will give us the ability to load and interpret the CSV files which the dataset is made up of. Having them loaded, allows us to combine and perform the necessary data analysis on the Instacart dataset. MySQL is another tool that could be valuable for the manipulation of the dataset and could be useful in other exploratory analysis of the dataset. Finally, Scikit-learn is a Python library that we can use for any machine learning we want to try and do on the data set.

Other tools include Google Docs for creating our milestones, allowing for easy editing by our whole group. Similarly, our group is using Google Colaboratory to collectively work on the same Python Notebook file. GroupMe and Zoom will serve as our communication platforms. And Github for communication and version control [9].

## 7 Milestones

To date, our group members have each downloaded and saved the dataset on their local machines. We've also completed an initial project description that outlines our goals and proposed work. We have also performed an initial survey of the data and previous research that has been performed with this dataset. Looking ahead, these are our groups planned milestones.

**Milestones Completed**:

Data preprocessing

Joining data tables

Initial exploratory data analysis

**Milestones Todo:**

End of March:

Classification and clustering

End of April:

Narrow down machine learning models

Optimize selected model's hyperparameters

Visualizations for final model's performance

Visualizations for data analysis

May 4th:

Final paper

## 8 Results So Far

<u>Data Cleaning</u>

As mentioned in the dataset section, our dataset is very clean with very few null or missing values. However, the days since prior order and number of orders from a single individual are "censored" at 30 and 99 respectively. After doing research on censored data, we have decided to maintain the data set as is and simply note that these data points refer to any data that is greater than or equal to that value. We will also aim to predominantly use our many uncensored data attributes whenever possible.

## Size of Orders

In our exploratory data analysis, we looked at rates at which different types of products are ordered, trends across time, and the amount that products are reordered. Our exploratory data analysis has shown that over 90.56% of the orders contain 20 items or less, and between 5 and 7 items is the most common size for an Instacart order (Figure 1).

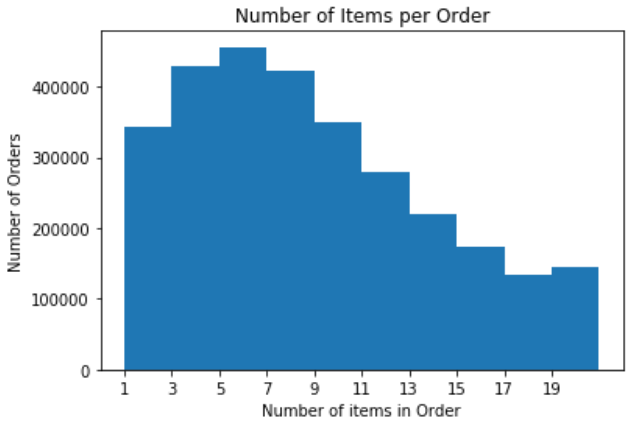


Figure 1: Histogram of number of items per order for all orders with 20 items or less

## Product Reorders by Department

We found that the most commonly ordered department by a large margin is produce making up 29.23% of the items ordered. Dairy/Eggs were second at 16.29%, with all other departments making up less than 10% of the items ordered (Figure 2). Dairy/Eggs and produce being the two highest ordered items makes sense, as they last a short time before spoiling and must be ordered frequently. The departments, bulk, other, missing, pets, alcohol, and international; none of which are known for having very perishable items, each make up less than 1% of the total items ordered.

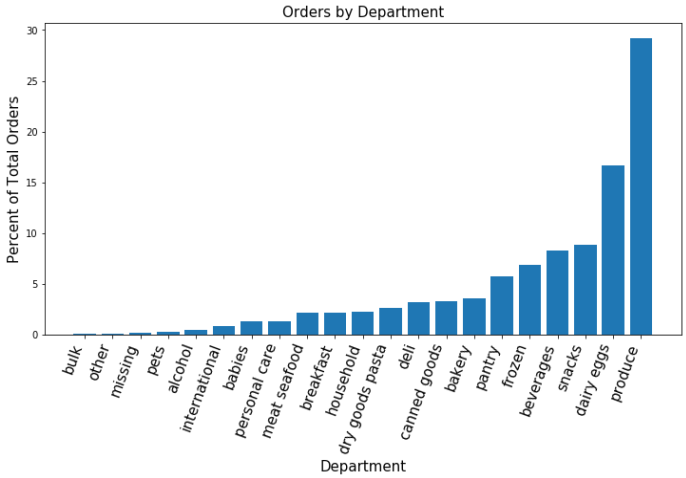


Figure 2: Bar graph of percent of total items ordered from each department

We also checked the percent of products re-ordered by department. Dairy/Eggs was the department that items were most commonly reordered from, followed by beverages and then produce (Figure 3).

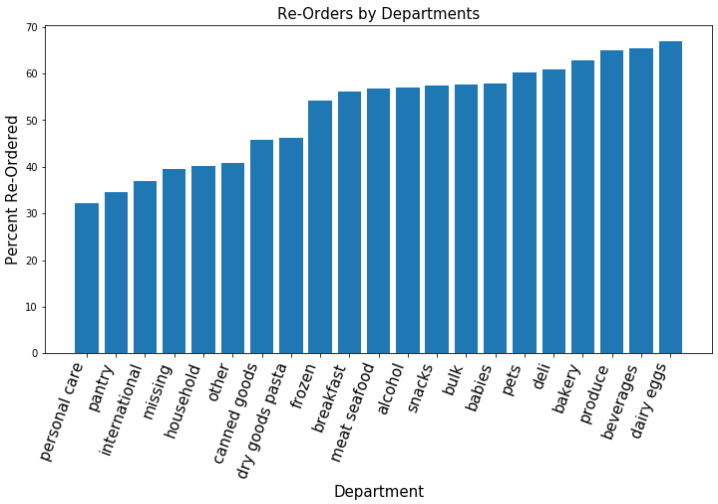


Figure 3: Bar graph of percent of items re-ordered from each department

The discrepancy between produce being the most commonly ordered from department by a large margin and being third for items being repurchased, indicates that shoppers tend to vary what produce they purchase between orders more than in the case of dairy/eggs or beverages. This makes sense given the great variety of produce and that people preferences for produce tend to change by season, as compared to dairy/eggs or beverages which stay more consistent and are may be more likely to foster brand loyalty.

## Product Reorders by Aisle

We found the aisles most commonly ordered from were fresh fruits and fresh vegetables at 11.23% and 10.54% of items ordered respectively, with all other aisles making up less

than 10% of the items ordered (Figure 4). This is consistent with produce being the most commonly ordered from department.

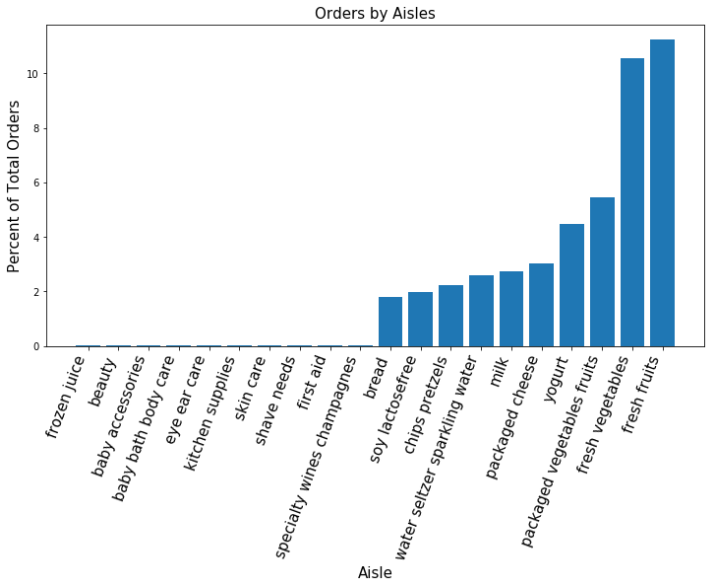


Figure 4: Bar graph of percent of total items ordered from each of the top and bottom 10 most ordered from aisles

The re-orders by aisle have milk as the most re-ordered aisle, followed by the water then fresh fruits aisle. Similar to the departments, milk and water being the most reordered from aisles, when they each make up less than 10% of the total items ordered, indicates that customers are more consistent and loyal with these purchases as compared to produce from which customers more often vary their purchases. The least commonly reordered from aisles are spices, baking supplies, and first aid; all of which tend to sell non-perishable or long shelf long items, which could last customers for extended periods of time.

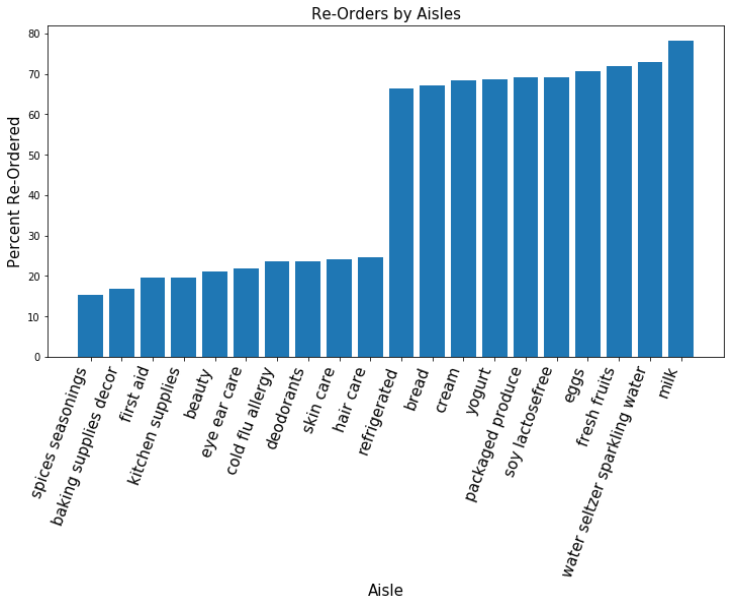


Figure 5: Bar graph of percent of items re-ordered from each of the top and bottom 10 most re-ordered from aisles

## Orders across Time

As mentioned in the dataset section, others have concluded that based on large spike in orders on day 0 and day 1, 0 represents the beginning of the weekend (Saturday) and that the days of the week are in order from that point [4,8]. However, based on our initial research, we have hypothesised that day 0 is in fact Monday. The evidence that lead us to this conclusion is presented within this section.

We began looking at how instacart orders varied across time by analyzing and plotting the frequency of total orders across day of the week and through time of the day. We looked at the volume of ordering throughout the week in two different ways, the first being the total number of items ordered in a day and the second being the number of different orders placed regardless of the number of items in each of those orders. Day 0 is the day on which a majority of items are ordered (19.15%), with day 1 having the second highest percentage of items ordered (17.47%) (following days: day 2 = 13.00%, day 3 = 11.85%, day 4 = 11.68%, day 5 = 12.98%, day 6 = 13.88%).

Likewise day 0 & 1 are the days on which the greatest number of orders were placed (regardless of the number of items); however, the percentage of orders placed on these two days is actually nearly equivalent (17.35% and 17.32% respectively) (following days: day 2 = 13.74%, day 3 = 12.83%, day 4 = 12.48%, day 5 = 13.25%, day 6 = 13.03%).

This difference between the number of items ordered and the number of orders placed led us to conclude that the size of orders placed on day 0 is larger than the size of orders placed on day 1. We then calculated the mean and median order size for both day 0 and day 1 to confirm this finding, and we found that on day 0 the mean order size is 11.13 items and the median is 9 items, whereas, on day 1 the mean order size is 10.18 items and the median is 8 items. All the other days of the week have a mean in the range of 9.32-9.88 items with a median of 8 items, except day 6, which has a mean of 10.74 items and a median of 9 items.

The most common hours of the day for total items ordered and number of orders placed were between 10 am and 2 pm, with approximately 8% of items being ordered each hour, and the least common hours of the day were between 12 and 6 am, with less than 1% of items being ordered each hour.

The visualization of this data was difficult to create because of the size of the dataset, so we used a sample of 10,000 items ordered to make the KDE plot. The visualization displays the distribution of the number of items ordered by day of week and hour of the day. This graph shows that on day 0 most items are ordered in the afternoon, while on day 1 the bulk of items are around 10:00 am (Figure 6).

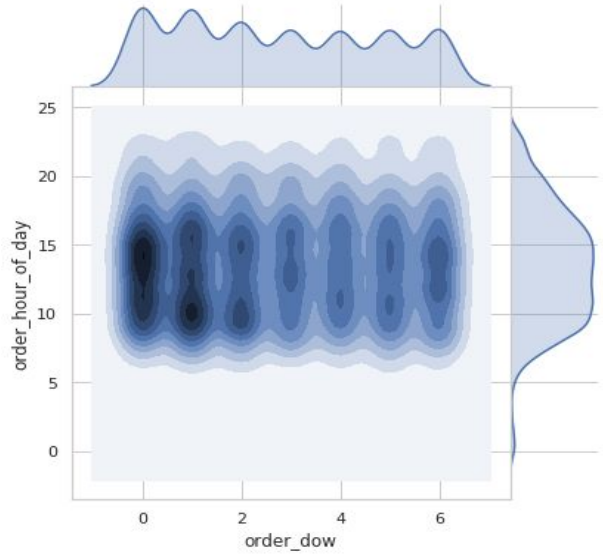


Figure 6: KDE plot of the number of items ordered by day of week and hour of the day

We then looked at how orders varied throughout the week separated by department. Every department except alcohol followed the pattern of total items ordered, with day 0 or day 1 being the day that the most items from that department were ordered. The alcohol department, on the other hand, had the most items sold on day 4 & 5 (17.01% and 17.85% of alcohol items sold respectively), and on day 0 alcohol was actually sold at the lowest rate (11.35% of alcohol items sold). So, despite the spike in orders placed on day 0 & 1 (which led others to conclude that this was the weekend), alcohol, a product that is often served at parties and other social gatherings or consumed when individuals do not need to work the next day, is still sold at a much higher rate on day 4 & 5. This was our first piece of evidence that day 0 may in fact be monday, making day 4 friday and day 5 saturday.

We thought this conclusion made sense intuitively because we typically encounter individuals beginning their count of weekdays on monday or sunday, to fit with common calendars; however, we wanted to further test this hypothesis. We next compared the most commonly products ordered from departments that would likely contain “junkfood” on day 0, which we hypothesised to be monday, to the most common products ordered on day 5, which we hypothesised to be saturday. We assert that if food we deemed to be unhealthy or “junkfood” appears to be ordered more frequently on day 5 than day 0 then this would support or hypothesis that day 0 is monday as “junkfood”, along with alcohol, is often consumed at parties and other social gatherings that tend to be more common on the weekend.

We looked at two departments snacks and frozen. We chose these departments because the snacks department includes items like chips and cookies and the frozen department includes items like ice cream, all of which are considered common types of “junkfood”. For the snacks department we found that the top 5 most

commonly purchased items on day 0 and day 5 were as follows:

| Day 0 | Day 5 |
|---|---|
| Lightly Salted Baked Snap Pea Crisps | Lightly Salted Baked Snap Pea Crisps |
| Trail Mix | Sea Salt Pita Chips |
| Original Veggie Straws | Organic Tortilla Chips |
| Pretzel Crisps Original Deli Style Pretzel Crackers | Sea Salt & Vinegar Potato Chips |
| Sea Salt Pita Chips | Chocolate Chip Cookies |

In the case of the frozen department, we found that the top 9 most ordered items on Day 0 and Day 5 are the same (with the only the order swapped for a couple items) and primarily consist of frozen produce. However, the following 5 items in the frozen aisle do differ (items 10-14), and those are as follows:

| Day 0 | Day 5 |
|---|---|
| Gluten Free Whole Grain Bread | Chocolate Ice Cream |
| Berry Medley | Gluten Free Whole Grain Bread |
| Organic Brown Rice | Frozen Broccoli Florets |
| Frozen Broccoli Florets | Vanilla Ice Cream |
| Organic Cheese Frozen Pizza | Organic Cheese Frozen Pizza |

As can be seen from these lists, potato chips, cookies, and ice cream are all ordered ranked higher in the top ordered products on day 5 than day 0, and none of these items actually appear at all in the day 0 top 5 differing products for these two departments of interest. Thus, we concluded that day 0 likely represents monday, and that it we will operate under this assumption throughout our final analysis.

## 8 References


[1] Instacart. 2017. The Instacart Online Grocery Shopping Dataset 2017. Retrieved March 10, 2019 from https://www.instacart.com/datasets/grocery-shopping-2017

[2] Stanley, Jeremy. 2017. 3 Million Instacart Orders, Open Sourced. (May 2017) Retrieved March 10, 2019 from https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2

[3] Annie George. 2017. Instacart Market Basket Analysis - Reorder Analysis. (Sept. 2017) Retrieved March 10, 2019 from https://nycdatascience.com/blog/student-works/capstone/instacart/

[4] Philipp Spachtholz. 2017. Exploratory Analysis - Instacart. Retrieved March 10, 2019 from https://www.kaggle.com/philippsp/exploratory-analysis-instacart

[5] Wan, Mengting, Di Wang, Jie Liu, Paul Bennett, and Julian McAuley. 2018. Representing and Recommending Shopping Baskets with Complementarity, Compatibility and Loyalty. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, (Oct. 2018), 1133-1142. DOI: https://doi.org/10.1145/3269206.3271786

[6] Bhade, Kalyani, Vedanti Gulalkari, Nidhi Harwani, and Sudhir N. Dhage. 2018. A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), (Jul. 2018) 1-6. DOI: https://doi.org/10.1109/ICCCNT.2018.8494019

[7] Asha, K. N., and R. Rajkumar. 2017. Pre-processing of user behaviour for e-commerce. In 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), (Aug, 2017) 715-719. DOI:

https://doi.org/10.1109/ICCCNT.2018.8494019

[8] Stanley, Jeremy. 2017. data_description. (May 2017) Retrieved March 10, 2019 from https://gist.github.com/jeremystan/c3b39d947d9b88b3ccff3147dbcf6c6b

[9] Our Group's GitHub repository: https://github.com/MichaelWhitlock/CSCI4502Group1.git