Table of Contents

1. Problem Statement

Instacart is a grocery ordering and delivery app that aims to make it easy to fill your refrigerator and pantry with your personal favorites when you need them. The Instacart app allows you to browse thousands of products from your favorite stores, from groceries and alcohol to home essentials and more. After selecting the products you want, personal shoppers will pick up those products for you as well as deliver them to you. Kaggle has challenged the data science community to use anonymized data on customer orders over time to predict which previously purchased products will be in a user's next order.

The goal of this project is to develop a model capable of predicting whether an order is reordered or not. The obvious client here is Instacart, as it will be the one profiting from the predictive model. The users of Instacart will also benefit from the project, as the app would be more skilled at recommending products to users. Therefore, this is a classification problem, but it is also a recommendation system.

2. Data Wrangling

The data for this project was supplied by Kaggle. The file containing the aisle names and id's is called aisles.csv. Similarly, the file containing the department names and id's is called departments.csv, and the file containing the product names and id's is called products.csv. Another csv file, order_products__prior.csv contains previous order contents for all customers. Yet another csv file, orders.csv, tells which set (prior, train, test) an order belongs. As there were many csv files to work with, many merges were done to better view the data. Fortunately, there were no missing values to deal with.

3. Exploratory Data Analysis

a. Order Analysis

First, an analysis of the orders.csv file was completed. First, I plotted which days of the week most orders were made. This is seen in Figure 1.

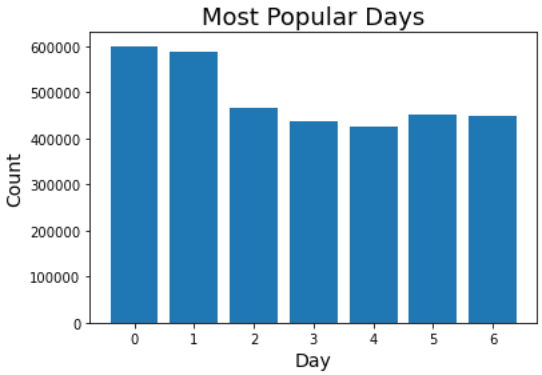

Figure 1: Orders by day of week

As you can see in the figure, days 0 and 1 were the most popular days for orders. It is reasonable to assume that days 0 and 1 correspond to Saturday and Sunday, respectively. Therefore, it is clear that more orders were made on the weekend than on the weekdays. Next, I plotted which hours of the day most orders were made. This can be seen in Figure 2.

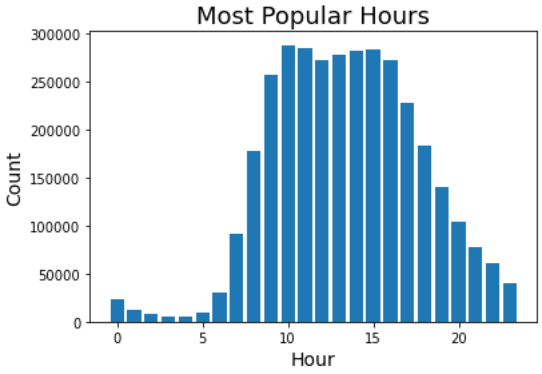

Figure 2: Orders by hour of day

Most orders were made between the hours of 10am and 3pm. Understandably, there were very few orders in the middle of the night (12am to 6am). I also plotted a heat map of the orders for day of week vs. hour of day. This is seen in Figure 3.

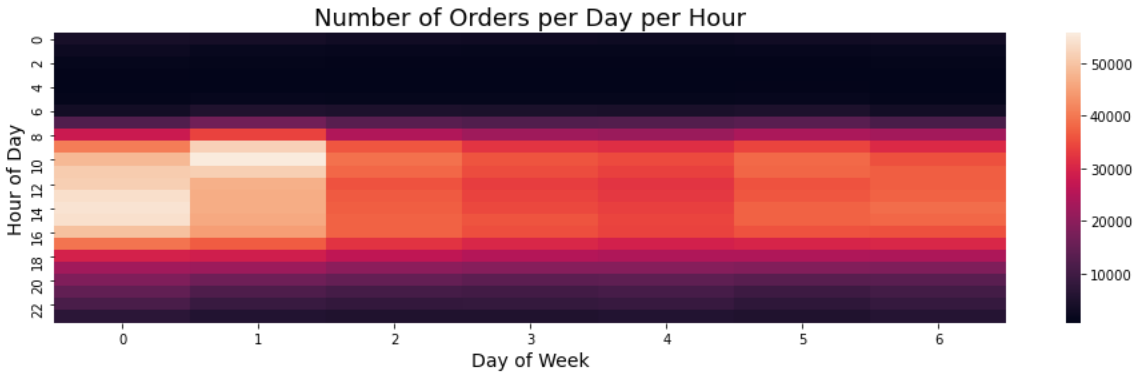

Figure 3: Orders day of week vs. orders hour of day

This heat map shows more clearly that there are very few orders at the beginning of the day (12am to 6am). Then, orders pick up until the end of the night. It is also clear that days 0 and 1 (Saturday and Sunday, respectively) have the most orders. Finally, I plotted the days since prior order, which is seen in Figure 4. The majority of the days since prior order is 30. This is most likely due to any number of days over 30 being assigned to the 30 category. There also appears to be many customers who order weekly, as the bars for 7, 14, and 21 days are local maxima.

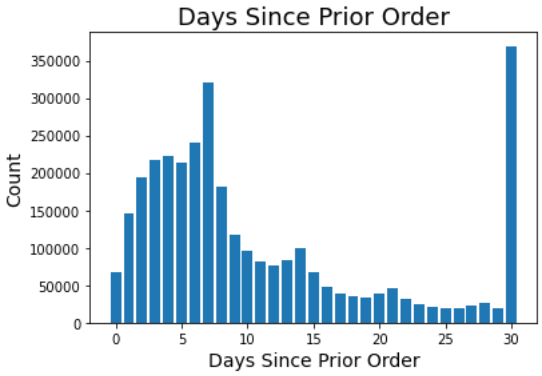

Figure 4: Days since prior order

b. Product and Order Analysis

Next, an analysis of the products and orders combined was conducted. I was able to do a simple inventory check to determine which aisles have the most products and which departments have the most products. These are seen in Figure 5 and 6, respectively.

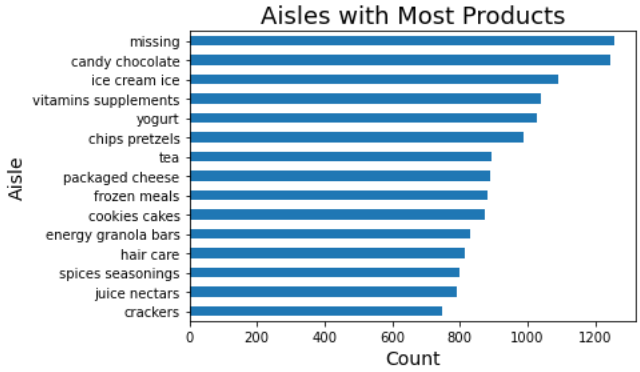

Figure 5: Aisles with most products

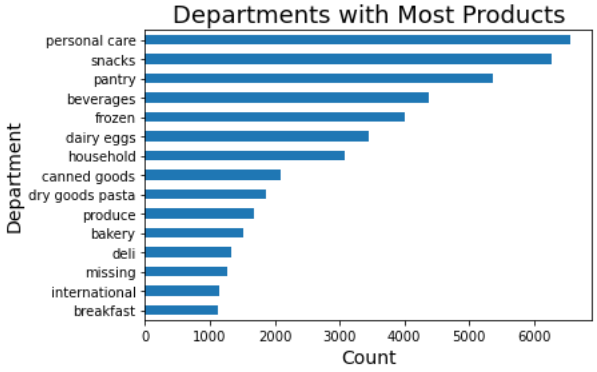

Figure 6: Departments with most products

I also determined the top products ordered, top aisles ordered from, and top departments ordered from. These are seen in Figures 7, 8, and 9, respectively.

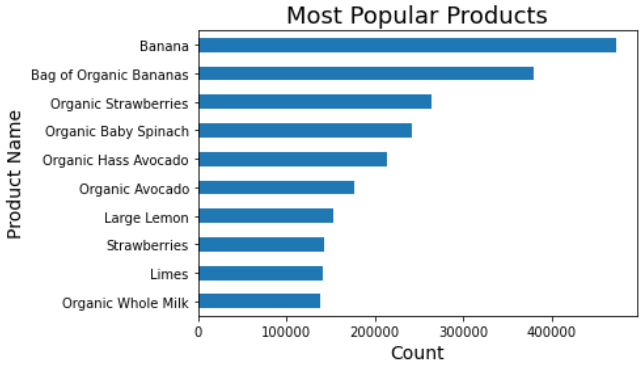

Figure 7: Top products ordered

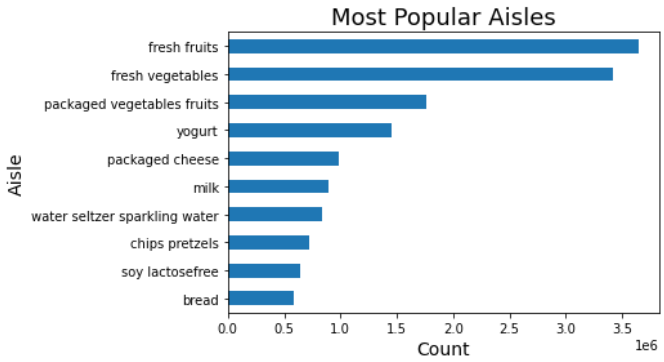

Figure 8: Top aisles ordered from

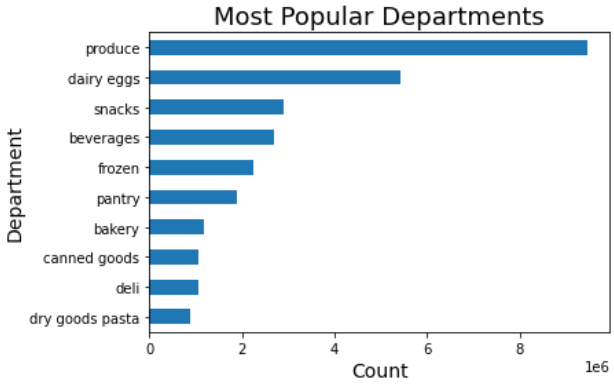

Figure 9: Top departments ordered from

You can see from Figure 7 that most of the top products ordered were actually organic products. The most popular product, however, was regular bananas. The most popular aisles were fresh fruits and vegetables, which makes sense given that most of the top

products were fruits and vegetables. Similarly, the most popular department was produce. Next, I analyzed the most popular reordered products, which can be seen in Figure 10.

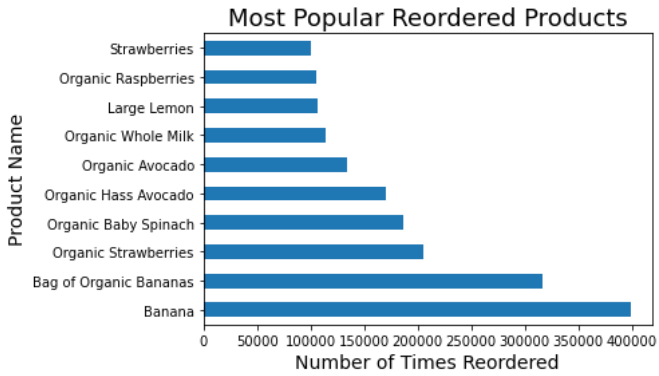

Figure 10: Most popular reordered products

I also calculated that about 59% of the products in the orders were reordered products, while about 12% of the orders had no reordered items. Figure 11 displays a histogram of the number of products in a given order. It has a normal distribution with a long right tail. The count was highest for 5-6 items in an order.

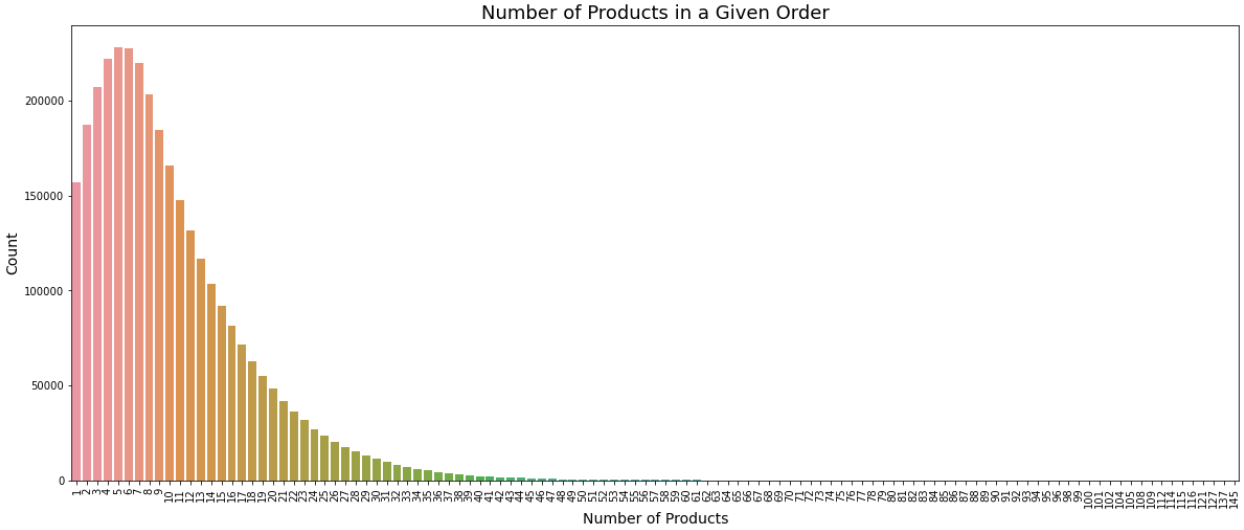

Figure 11: Number of products in a given order

Figure 12 displays a histogram of the number of reordered products in a given order. This is similar to the right side of a normal distribution with a long right tail.

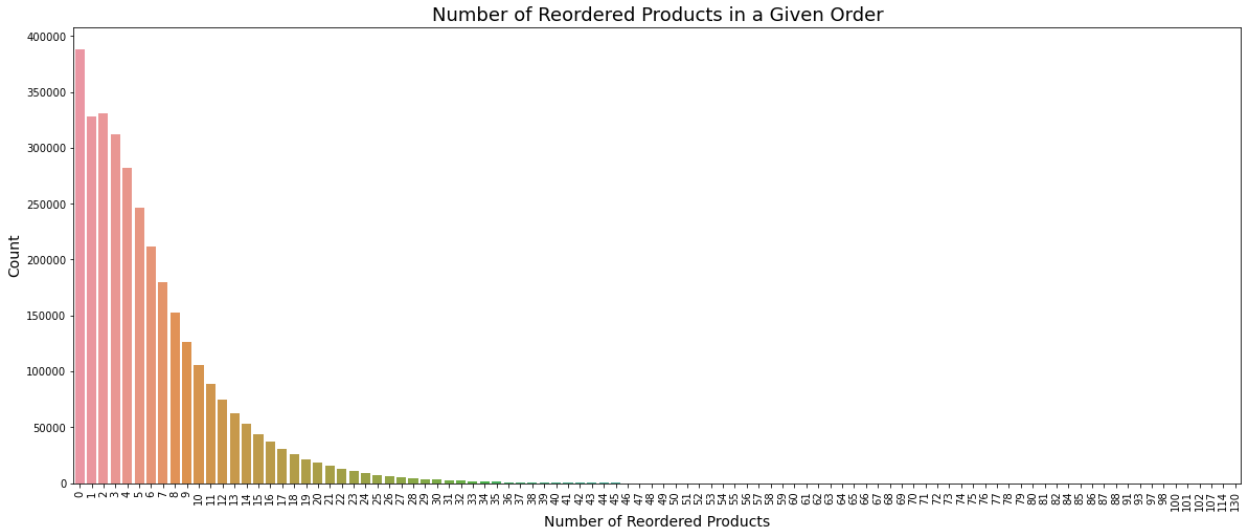

Figure 12: Number of reordered products in a given order

Finally, I completed a user analysis. The average number of prior orders by a customer is 15.59 orders. The minimum is 3 orders, and the maximum is 99 orders. The average number of days between orders for all customers is 15.47, with a minimum of 0 and a maximum of 30. The average basket size for all customers is 9.95 items, with a minimum of 1 and a maximum of 70.25**.**

4. Feature Engineering

a. User Features

The features calculated for each user include the reordered ratio of each user, the total number of orders of each user, the total items purchased by each user, the average days since prior order for each user, and the average basket size of each user.

b. Product Features

The features calculated for each product include the reordered ratio of each product and the number of purchases of each product.

c. Order Features

The features for each order include the order number, aisle ID, department ID, day of week, hour of day, and days since prior order.