Capstone Project 1: Exploratory Data Analysis

Ravi Maranganti

Springboard

**Abstract**

Instacart is a grocery ordering and delivery app through which *users* can select and place products in their cart and *shoppers* review the orders, do in-store shopping and deliver the items to the users. In 2017, Instacart released an anonymized dataset containing over 3 million grocery orders belonging to more than 200,000 Instacart users as part of a Kaggle competition. The objective of the competition was to use the provided dataset to predict which of a customer’s previously purchased products would be in his/her latest order. In this report, we will present results obtained by performing an exploratory data analysis (EDA) on the dataset and discuss some interesting correlation that was found between discrete features present in the dataset.

**Github Code:** <https://github.com/ravimaranganti/Springboard_Capstone_1/blob/master/reorder_eda.ipynb>

**Capstone Project 1: Exploratory Data Analysis**

In 2017, Instacart released a dataset consisting of over 3 million orders belonging to more than 200,000 customers for the first time to the public. Concurrently, a Kaggle competition was also announced where Instacart challenged the machine learning community to use the dataset to develop a model to predict which among of a customer’s previously purchased products would be in that customer’s next order. The problem at the very outset is akin to developing a recommender system but with an added temporal element to it. This is because while a typical recommender system (say for instance Netflix’s movie recommendation) assumes that user preferences and tastes do not change over relatively short periods of time, an Instacart user might be likely to re-order milk on a weekly basis but might re-order a spice such as cinnamon only on a bimonthly basis.

# Data

The data is provided in the form of 6 comma separated (.csv) files. As mentioned earlier, the data is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, between 4 and 100 of their orders are provided, with the sequence of products purchased in each order. In addition, the week and hour of day the order was placed, and a relative measure of time between orders is also given. Each order is also classified using the categorical variables of ‘prior’, ‘train’ and ‘test’. Apart from the latest orders of customers, all other orders are categorized under “prior”. The latest orders are further subdivided into “train” and “test”. It should be noted that it is for the orders categorized under “test” that we have to predict products which will be reordered. It should be also noted that while all customers have orders which are categorized under “prior”, some customers have their last orders categorized under “train” while the rest of the customers have their last orders categorized under “test”.

Let us examine the files provided in some detail.

1. **aisles.csv**

The aisles.csv dataset contains 134 rows, with each row containing an aisle\_id (integer form) and the corresponding aisle name (string). There are no null entries.

aisle\_id,aisle

1,prepared soups salads

2,specialty cheeses

3,energy granola bars

...

1. **departments.csv**

The departments dataframe consists of 21 rows , each row containing a department id (integer form) and the corresponding department name (string). There are no null entries.

department\_id,department

1,frozen

2,other

3,bakery

...

### order\_products\_\_prior.csv

order\_id,product\_id,add\_to\_cart\_order,reordered

1,49302,1,1

1,11109,2,1

1,10246,3,0

...

### order\_products\_\_train.csv

Consists of orders classified under “train”

### orders.csv

### This file tells to which set (prior, train, test) an order belongs. An order is identified by its unique order\_id. This file also contains the user\_id associated with that order, the order number of the user ,the day of the week and the hour of the day this order was placed ( 'order\_dow' , ‘order\_hour\_of\_day’) and lastly the number of days since the user’s previous order. This file contains 3421083 rows which is equal to the total number of orders contained in this dataset.

order\_id,user\_id,eval\_set,order\_number,order\_dow,order\_hour\_of\_day,days\_since\_prior\_order

2539329,1,prior,1,2,08,

2398795,1,prior,2,3,07,15.0

473747,1,prior,3,3,12,21.0

...

### products.csv

product\_id,product\_name,aisle\_id,department\_id

1,Chocolate Sandwich Cookies,61,19

2,All-Seasons Salt,104,13

3,Robust Golden Unsweetened Oolong Tea,94,7

...

## Exploratory Data Analysis

Visual Exploratory Data Analysis (EDA) is a good way to better understand our data and get a sense of which features might matter more than others as predictors of the target variable (“reorder” in our case). Visual EDA can also serve to provide sanity checks and identify any inconsistencies ,missing values and outliers in the dataset provided. While an exhaustive visual EDA of the data can be found in the corresponding ipython notebook, some representative graphs and charts will be shown below.

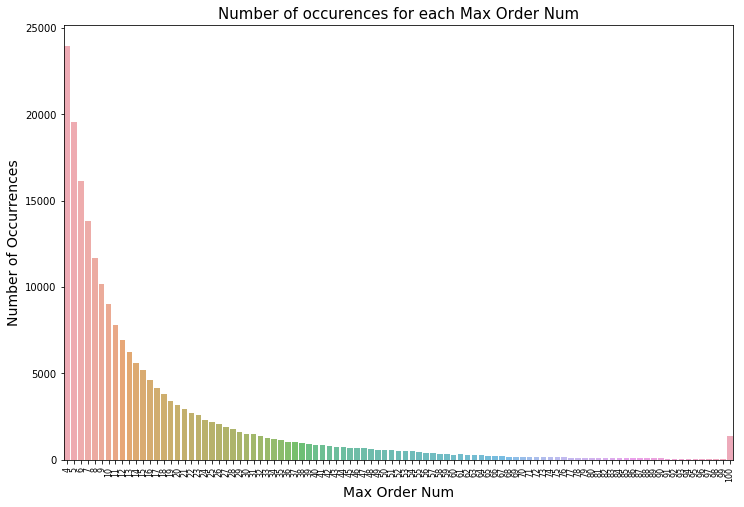
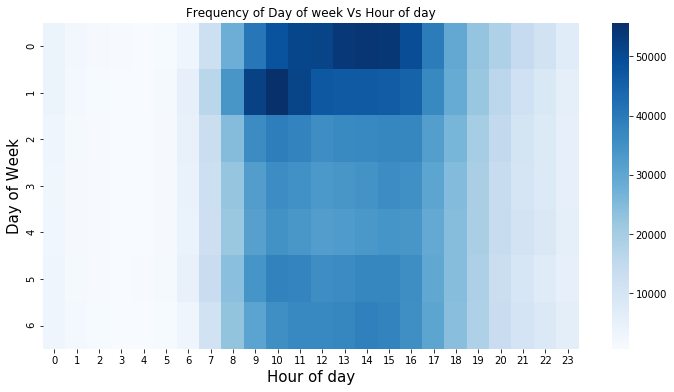


Figure 1: Plot of maximum number of orders for customer versus frequency

We can see that the number of occurrences for a given max order number decreases in an exponential manner with increasing max order number. This seems reasonable because most customers may be using Instacart sparingly while fewer customers may be using Instacart on a regular basis. There is a spike at Max Order Num=100 because the max order num is capped at that number.

Figure 2: Heatmap showing order volume for different days of week and hours of day.

We can see that Saturday (day 0) and Sunday (day 1) have the highest order volume compared to other days of the week. Further, Sunday mornings and Saturday afternoon and early evening hours have higher order volumes than other day-hour combinations. We can also see that the number of orders placed during early morning hours (before 6 am) and late evening and night hours (after 8 pm) are low.

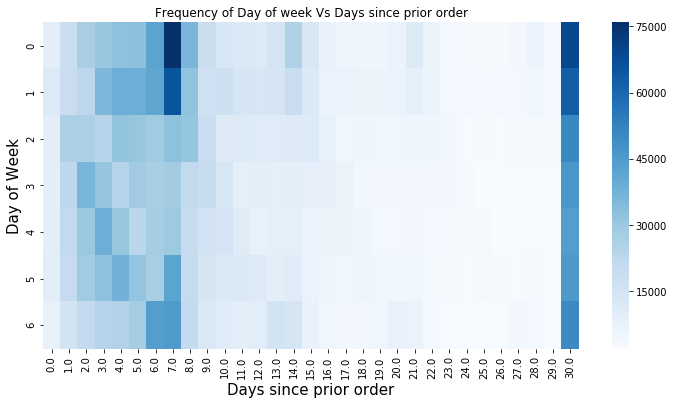
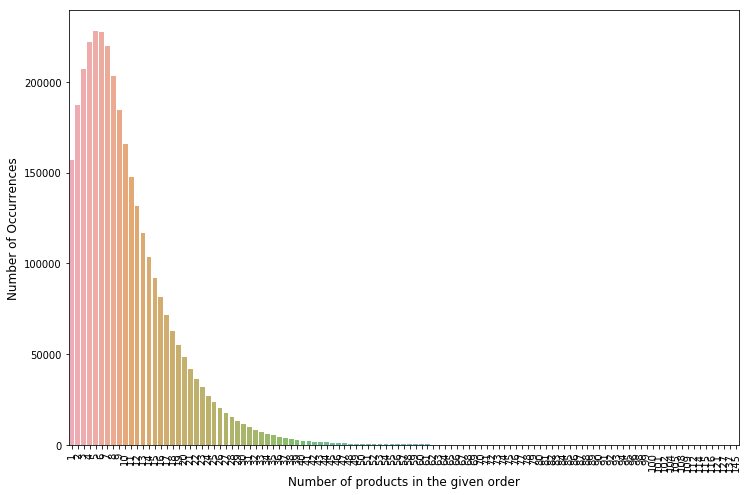


Figure 3: Heatmap showing order volume for different days of week and hours of day.

While we saw from Figure 2 that Saturday and Sunday (Day 0 and Day 1 respectively) accounted for more orders than other days of the week, From the above heatmap, we can see that that many customers who buy on Saturday and Sunday also buy on a weekly basis (7 days). So for quite of few customers ordering weekly on either Saturday night or Sunday morning in preparation for upcoming weeks is routine. Perhaps these customers will also be more predictable?

Now let us look at the cart size of different orders in the dataset.

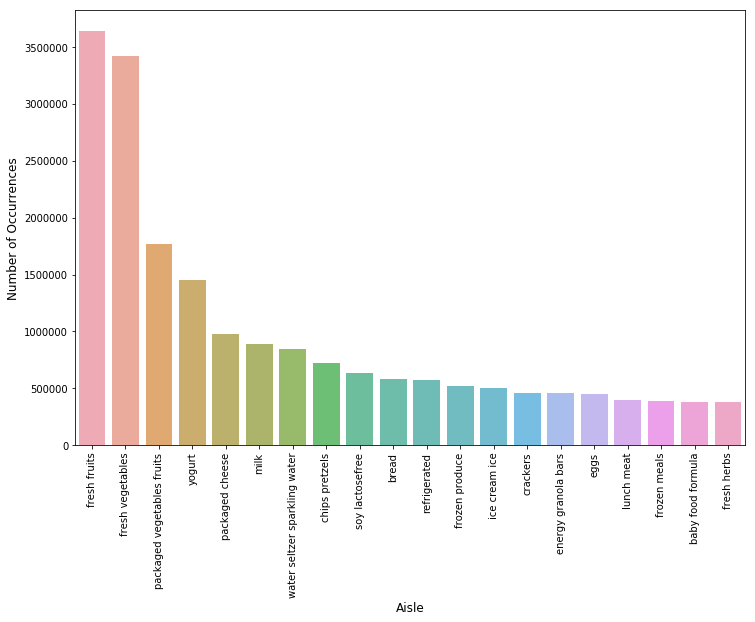


The above chart suggests that carts with 5 or 6 items are the most frequently occurring. While the largest cart size is 145 items, the graph is heavily skewed to the left.

Now let us look at the most frequently ordered items. The below table shows the top 15 most ordered items and we can clearly see that fruits and vegetables dominate with the only exception being organic milk coming in at number 9.

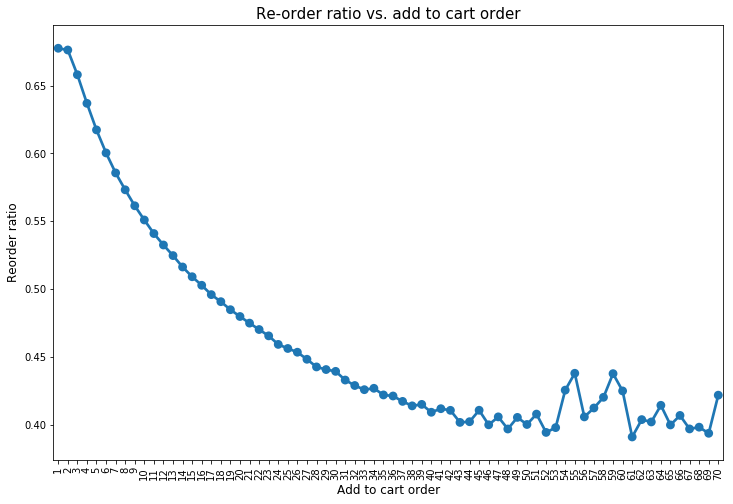
|  | **product\_name** | **count** |
| --- | --- | --- |
| **0** | Banana | 472565 |
| **1** | Bag of Organic Bananas | 379450 |
| **2** | Organic Strawberries | 264683 |
| **3** | Organic Baby Spinach | 241921 |
| **4** | Organic Hass Avocado | 213584 |
| **5** | Organic Avocado | 176815 |
| **6** | Large Lemon | 152657 |
| **7** | Strawberries | 142951 |
| **8** | Limes | 140627 |
| **9** | Organic Whole Milk | 137905 |
| **10** | Organic Raspberries | 137057 |
| **11** | Organic Yellow Onion | 113426 |
| **12** | Organic Garlic | 109778 |
| **13** | Organic Zucchini | 104823 |
| **14** | Organic Blueberries | 100060 |

A look at the most popular aisles also reveals interesting information about the buying habits of customers.



The above picture shows that fruits and vegetables are the most popular aisles.

Yet another interesting trend is revealed upon examining the graph of the re-order ratio (defined as the proportion of reorders versus total orders for a specific product) plotted against the add-to-cart order.

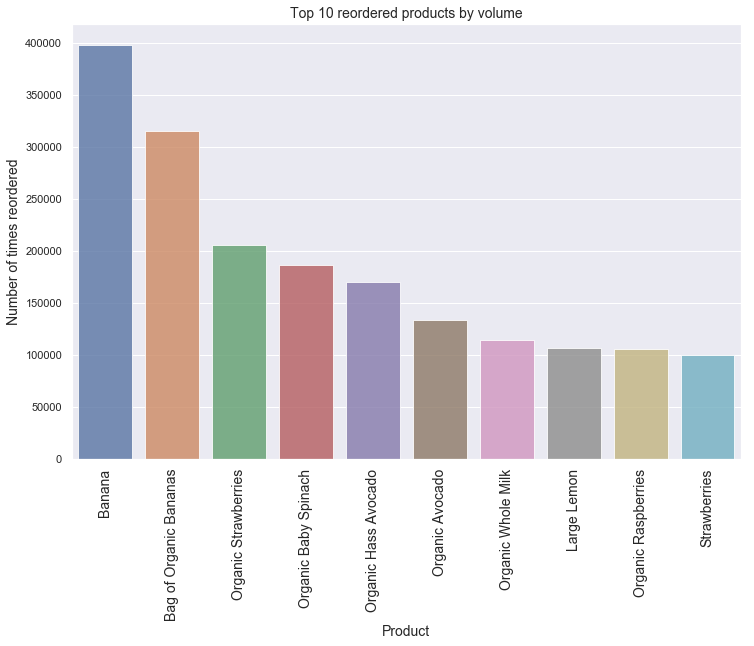


This graph shows that items which are added first to the cart are also the ones with higher re-order ratios.

## Reorder Variable (EDA)

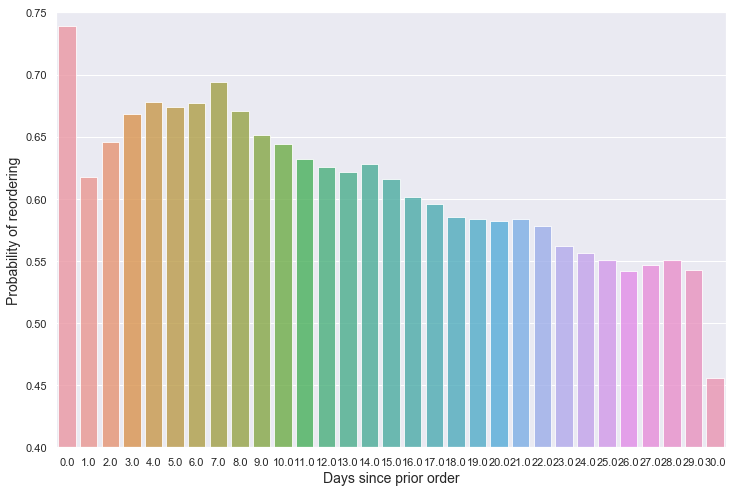
Having conducted this general investigation of the data, let us now focus on the variable of interest (which we have to predict): the reorder variable.

Let us look at which are the top re-ordered items.

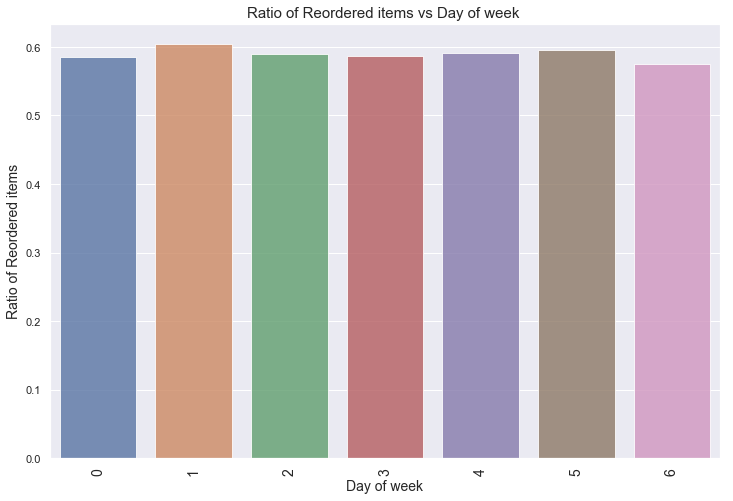


The above plot shows that the most popular items are also the ones with high re-order ratios which makes intuitive sense.

Now let us look at a bar plot of the probability of re-ordering versus days since prior order.

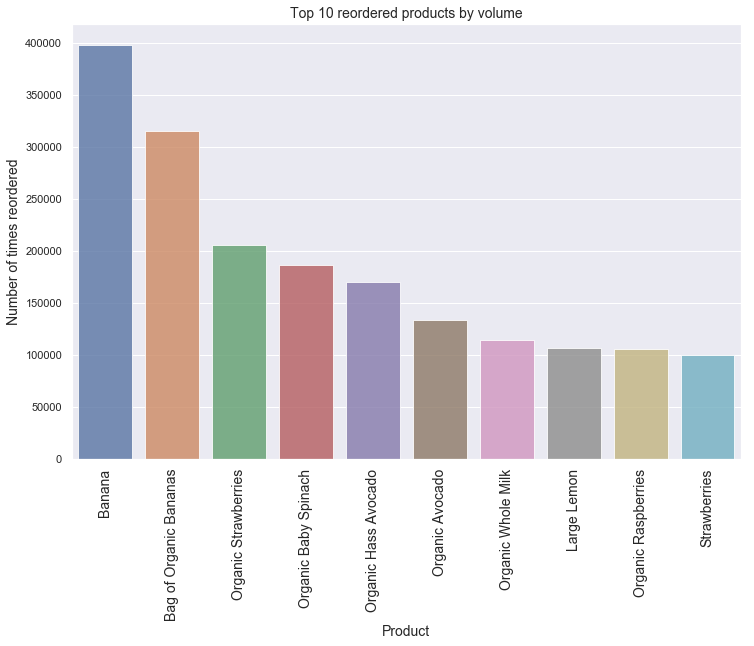


Let us look if the ratio of reordered items and total ordered items varies according to day of the week.



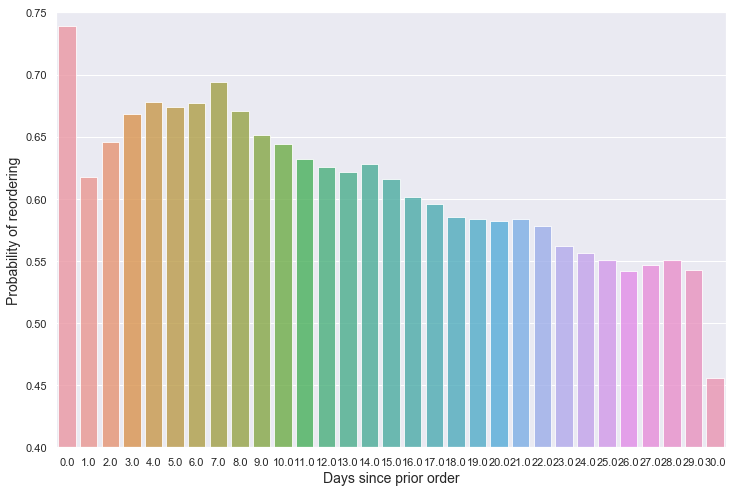
The bar plot above shows that there is no real variation in the re-order ratio with the day of the week. This suggests that the same items are re-ordered by customers (albeit in different quantities) on different days of the week.

Now let us examine the top re-ordered products by volume.



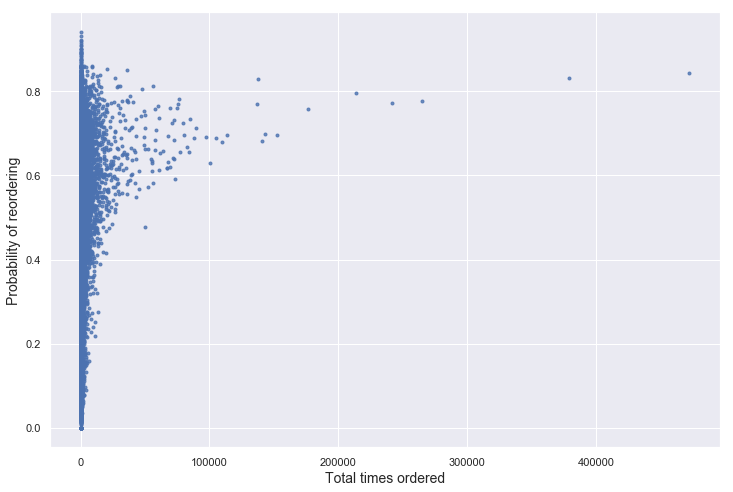
Not surprisingly, the top re-ordered products by volume are the same as the top 10 most popular products which we had plotted earlier.

Now let us look at the probability of reordering plotted against days since prior order.



It seems that if somebody orders again on the same day, it is more likely than not that they are reordering the same product. Maybe because they ordered less? It will be interesting to see what kind of products are reordered on the same day. Again, there are smaller peaks at 7, 14 , 21 and 28 days. At 30 there is a reasonably big drop which indicates that items which are ordered less frequently (more than or equal to 30 days between orders) have less probability of being reordered because customers may want some of these products rarely.

Finally, let us look at the probability of a product being re-ordered versus the total times a product is purchased.



We can see that items which have been ordered lower number of times have probability of reordering all across the spectrum. On the other hand, items which are ordered a lot have higher probability of being reordered (which makes sense).

## Investigating Correlations between Features Present in the Dataset

The “add\_to\_cart\_order” is an interesting feature of the dataset provided to us. The “add\_to\_cart\_order” of a product for a particular order is the position at which that product is added to the cart. It is noteworthy that unlike a physical store where a grocery shopper is likely to add produce or fruit to the cart earlier than say dairy (because of the physical location of these items in a grocery store), when it comes to an Instacart customer adding items to the cart, the physical location of items in the store does not matter. It may be assumed that popular products which are ordered regularly also get added to the cart earlier than items which are not very popular and are added to carts as an after-thought. On the other hand, there may be products which are not ordered regularly and are not uniformly popular products but are ordered by a niche set of customers and are inherently more “urgent” in nature. Products such as baby diapers or formula come to mind which may only be ordered by new parents and may not be ordered regularly but when they are ordered, they are added to the cart early enough because these products are on the customers’ top of mind.

In order to capture this feature, I created a feature which scales directly with the total times a product has been purchased across all customers but which scales inversely with the number of unique customers. This ensures that niche products which are popular with those set of customers are weighed more. I also scaled this feature further by multiplying it with the average days since prior order of this product. Therefore products which are ordered more infrequently have higher weight assigned to them. The formula for this so-called niche-factor is shown below.

order\_avg\_aco\_buys["niche\_factor"]=(order\_avg\_aco\_buys["user\_prod\_purchases"]/order\_avg\_aco\_buys["unique\_users"])\*order\_avg\_aco\_buys["days\_since\_prior\_order"]

Upon taking the Pearson correlation coefficient of the above product based feature with the average add to cart order of a product, we obtain a correlation of ~ -0.67 which hints at a moderately strong negative correlation. This means that the higher this “niche\_factor” the lower the average add to cart order for the product (or that the sooner the item is added to the cart). It also turns out that the item with the highest “niche\_factor” (low fat organic milk) is also the item with the third lowest “average add to cart order”. While this item is neither the most popular (8438 total purchases by only 1398 unique users) and is not so frequently ordered (with an average frequency of around 14 days), it has the highest niche-factor and also a very low average add to cart order.

The ipython notebook with the above analysis can be found below

<https://github.com/ravimaranganti/Springboard_Capstone_1/blob/master/reorder_eda.ipynb>