Milestone Report 1

1. The Problem and The Client

“Creating a Grocery Product Recommender for Instacart”

In the eCommerce shopping experience product recommendations come in many forms: they may be used to recommend other products on one product’s page (Amazon’s “Frequently bought together” feature for instance) or they may be used on the checkout page to show customers products they may be interested in based on their total order.

Instacart is an online grocery delivery service that allows users to place grocery orders through their website or app which are then fulfilled and delivered by a personal shopper- very similar to Uber Eats but for grocery stores. In 2017 they released a year of their data composed of about 3.3 million orders from about 200,000 customers.

Through the machine learning model and EDA, my best hope is to answer the questions like:

- What product will the user buy again?

Further, recommendations may be more helpful if they are targeted towards a specific segment of customers, rather than made uniformly. For instance, if one group of customers tends to buy a lot of non-dairy milk substitutes and another group tends to buy traditional milk, it may make sense to make different recommendations to go along with that box of Cheerios. To make tailored recommendations, Instacart users must be segmented based on their purchase history using K-Means clustering and then made recommenders based on the product association rules within those clusters.

2. The Data

The dataset is a relational set of files describing customers' orders over time. The goal 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, it is provided between 4 and 100 of their orders, with the sequence of products purchased in each order. We also provide the

week and hour of the day the order was placed and a relative measure of time between orders.

Dataset: https://www.kaggle.com/c/instacart-market-basket-analysis/data

Each entity (customer, product, order, aisle, etc.) has an associated unique id. Most of the files and variable names should be self-explanatory.

- aisles.csv
- departments.csv
- order_products__*.csv

  These files specify which products were purchased in each order. order_products__prior.csv contains previous order contents for all customers. 'reordered' indicates that the customer has a previous order that contains the product. Note that some orders will have no reordered items. You may predict an explicit 'None' value for orders with no reordered items.

- orders.csv

  This file tells to which set (prior, train, test) an order belongs. You are predicting reordered items only for the test set orders. 'order_dow' is the day of the week.

- products.csv
- sample_submission.csv

2.1 Data Wrangling

As it was mentioned above that the dataset is a relational set of files describing customer's orders over time. As expected, the dataset has no missing entries. Data wrangling on each of the files produced the following results:

- **Aisles**:

  There are 134 aisles. This table comprises 2 columns - unique aisle id and its corresponding aisle name.

- **Departments:**

  Every aisle is further grouped into 21 distinct departments. There is a department named "missing". Which is to be further examined later as to what products are in that department. There are departments such as frozen, bakery, dairy eggs, canned goods, personal care, produce, alcohol, etc.,.

- **Order_products__prior and order_products__train**

  Order_products__prior is the largest file among the existing files. It comprises 32434489 rows × 4 columns. And order_products__train contains 1384617 rows × 4 columns. These two files have the same column names.

  The difference between the two files being that the prior file comprises of the prior data of customer's orders whereas the order_products__train file corresponds to the current customer's purchase.

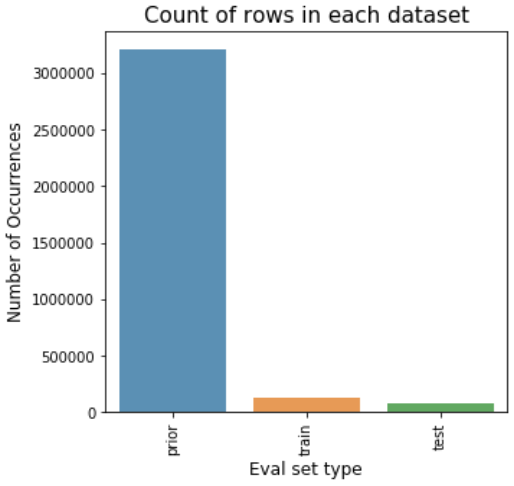
  The 4 columns for the order_products__* tables are:

  - Order_id
  - Product_id
  - Add_to_cart_order
  - Reordered (1 for yes and 0 for no)

- **Orders**

  Orders file is comprised of 3421083 rows × 7 columns. The 7 columns are:

  - Order_id
  - User_id
  - Eval_set
  - Order_no
  - Order_dow (day of the week)
  - Order_hour_of_day
  - Days_since_prior_order



The Eval_set has three outcomes - prior, train, or test. The prior and train discussed earlier. The test outcome corresponds to the data for which the recommendation has to be made. There are 206,209 customers in total. Out of which, the last purchase of 131,209 customers is given a train set and we need to predict for the rest 75,000 customers.

2.2 Data Visualization and EDA

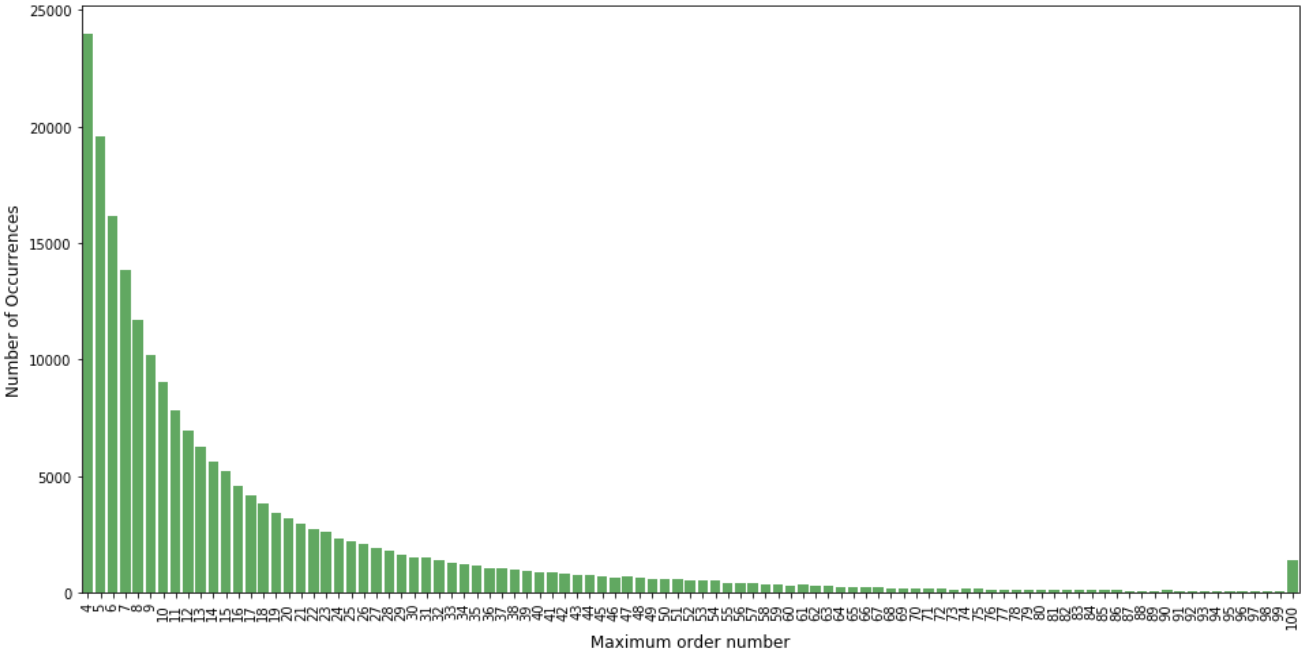
The main focus for Data Visualization and Exploratory Data Analysis is for the files

- Orders
- Order_products__*

After individually analyzing both the files, all the files are merged on order_id, user_id, product_id, and aisle_id for the sake of further analysis.

2.2.1 No. of orders by each customer:

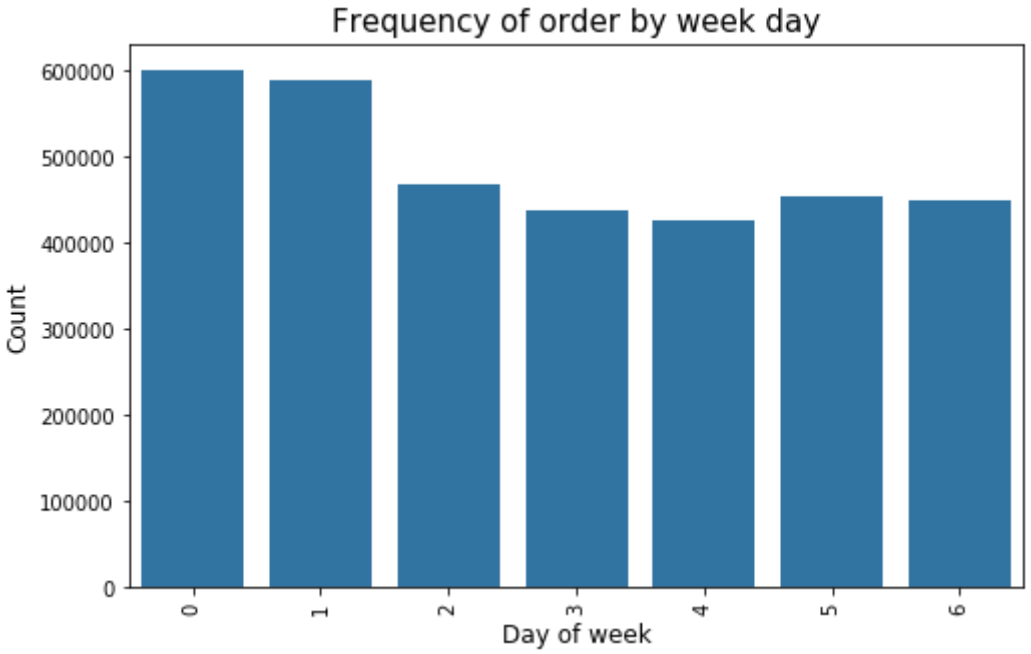
In the introduction of the dataset, it was mentioned earlier that order information of each customer was provided and that each customer made between 4 to 100 orders.



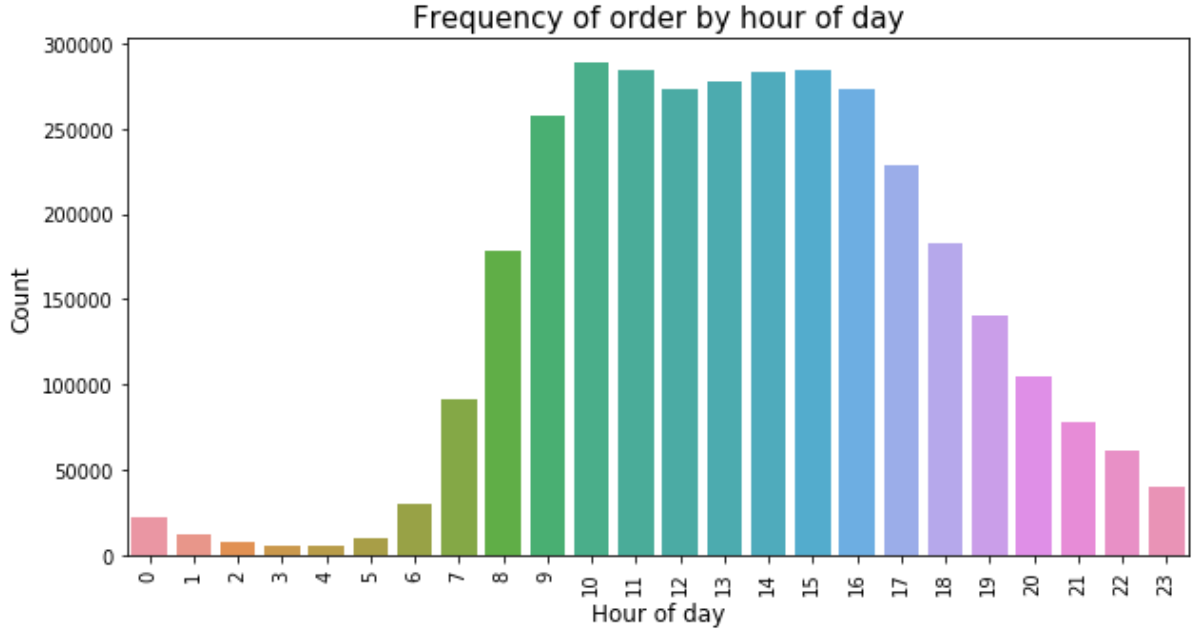
So there are no orders less than 4 and the maximum is capped at 100 as given in the introduction.

2.2.2 Ordering habit changes with day of the week:

The basic countplot of the orders vs their respective day of the week is given below. The orders are usually higher on Saturday and Sunday. And the lowest on Wednesday.

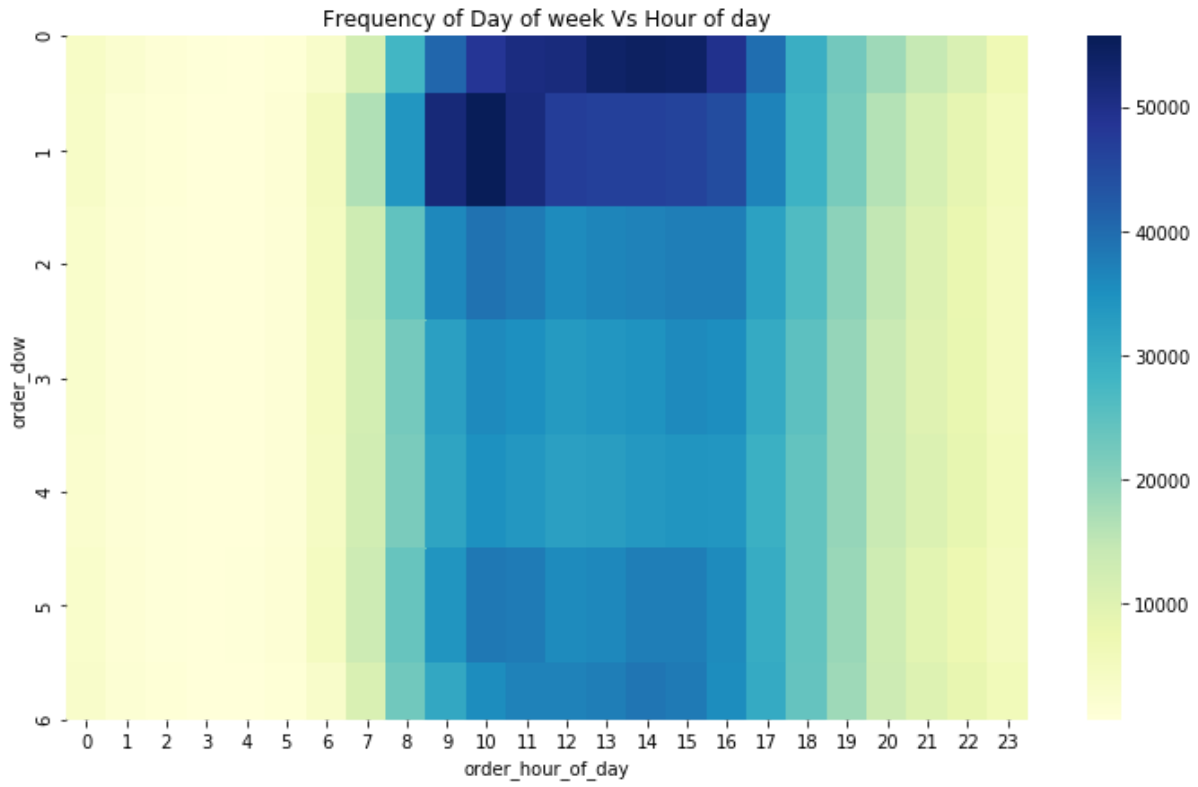


2.2.3 Order Distribution w.r.t time of day:



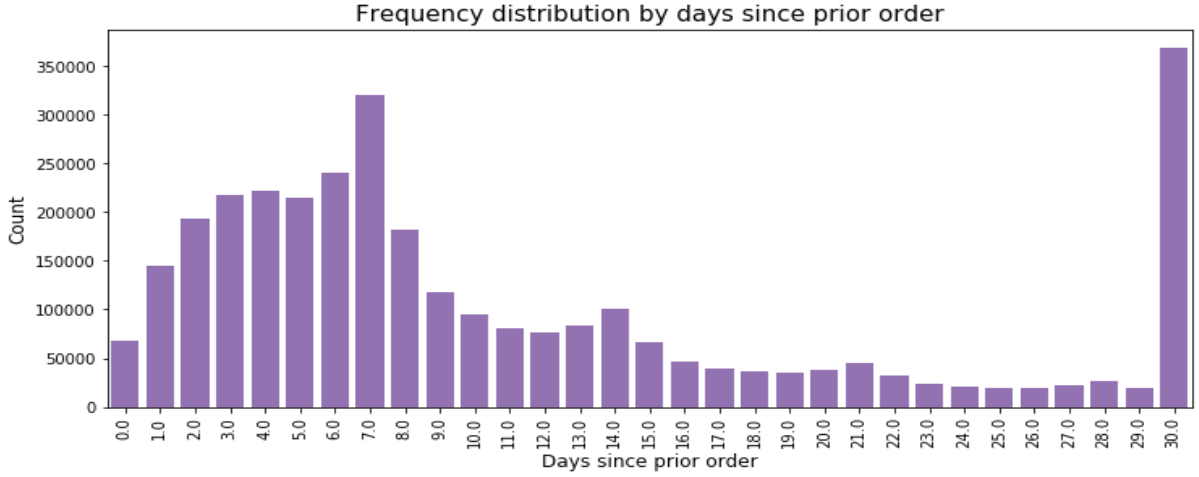
It can be inferred from the above barplot that most of the orders happen during the day time.

2.2.4 Order Distribution - Time of day vs. day of the week:



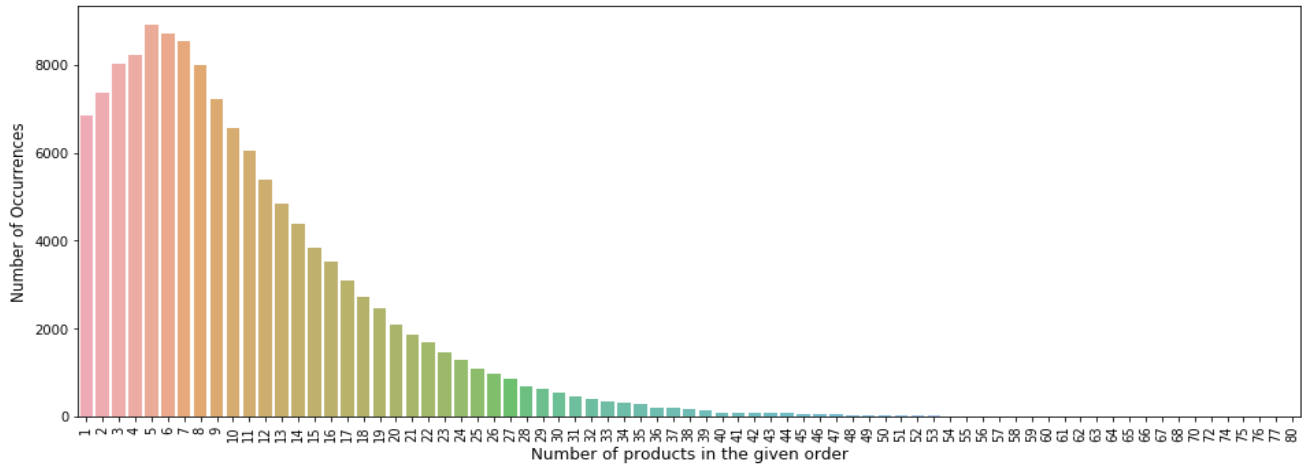
The inference we got from the two plots we plotted individually with Time of day and day of the week are combined to make a heatmap. From the heatmap, it can clearly be understood that Saturday evenings and Sunday mornings are prime time for orders.

2.2.5 The time interval between the orders:



Looks like customers order once every week (check the peak at 7 days) or once in a month (peak at 30 days). We could also see smaller peaks at 14, 21, and 28 days (weekly intervals).

2.2.6 Number of products bought in each order:



There is a peak of 5. But most customers ordered anywhere between 1 and 15 orders. It is interesting to note that the max number of products per order is 80. Although small, there is a sizable no of customers for whom the number of products per order is greater than 40.

2.2.7 Merging files:

The order_products__prior is merged with the files- products, aisles, and departments. This is done with the help of the left merge by aligning product_id, aisle_id, and department_id.

	order_id	product_id	add_to_cart_order	reordered	product_name	aisle_id	department_id	aisle	department
0	2	33120	1	1	Organic Egg Whites	86	16	eggs	dairy eggs
1	2	28985	2	1	Michigan Organic Kale	83	4	fresh vegetables	produce
2	2	9327	3	0	Garlic Powder	104	13	spices seasonings	pantry
3	2	45918	4	1	Coconut Butter	19	13	oils vinegars	pantry
4	2	30035	5	0	Natural Sweetener	17	13	baking ingredients	pantry

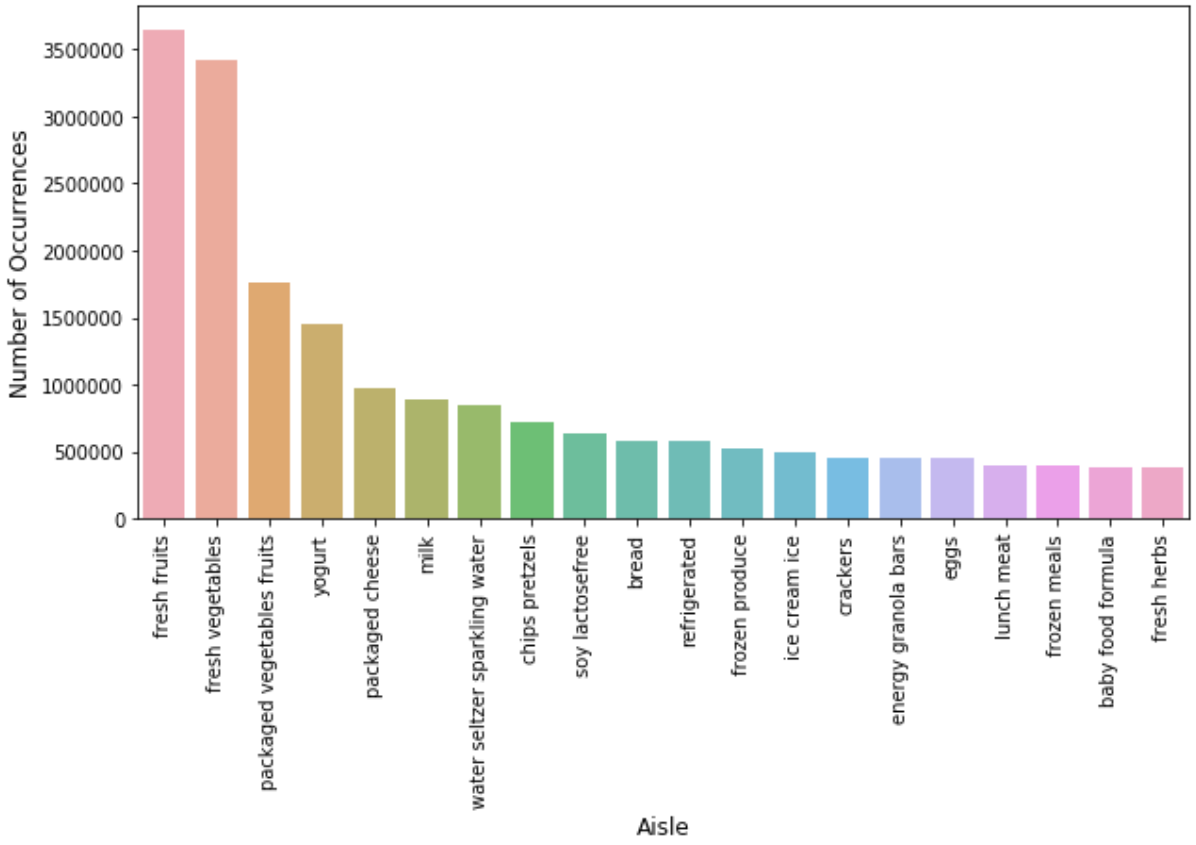
The top 20 products that were bought after repeatedly is then extracted from the merged file with the help of value_counts().

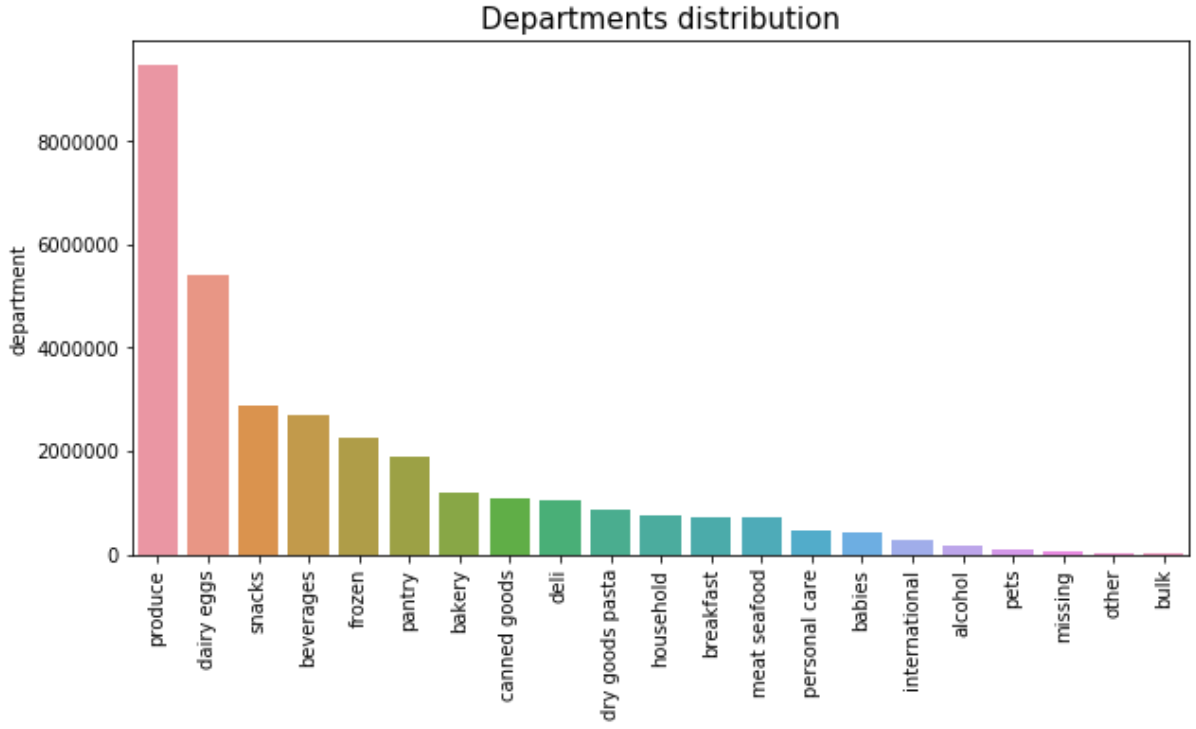
	product_name	frequency_count
0	Banana	472565
1	Bag of Organic Bananas	379450
2	Organic Strawberries	264683
3	Organic Baby Spinach	241921
4	Organic Hass Avocado	213584
5	Organic Avocado	176815
6	Large Lemon	152657
7	Strawberries	142951
8	Limes	140627
9	Organic Whole Milk	137905
10	Organic Raspberries	137057
11	Organic Yellow Onion	113426
12	Organic Garlic	109778
13	Organic Zucchini	104823
14	Organic Blueberries	100060
15	Cucumber Kirby	97315
16	Organic Fuji Apple	89632
17	Organic Lemon	87746
18	Apple Honeycrisp Organic	85020
19	Organic Grape Tomatoes	84255

It is interesting to note that most of the products are organic and the majority of them are fruits.

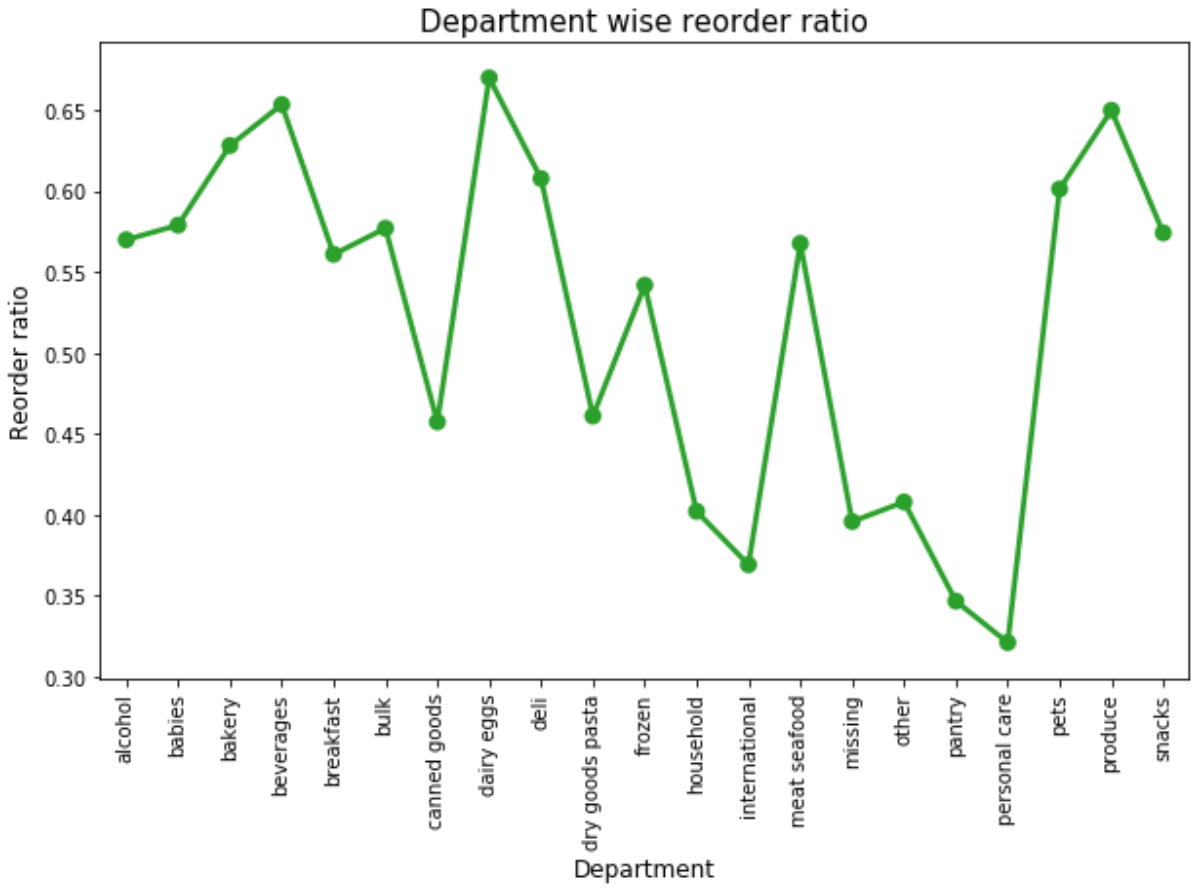
2.2.8 Important Aisles and Departments:

The most important aisles are the aisles that contain the produce (the top department)- fresh fruits, fresh vegetables, packaged vegetable fruits. And then comes dairy products. The plot describing the top 20 aisles and departments:





2.2.9 Re-order ratio of each department:



Personal care has the lowest reorder ratio and dairy eggs have the highest reorder ratio.

3. The Conclusion

The prime time for the orders is Saturday night and Sunday morning with the highest number of orders observed during Saturday night and the lowest is during the nights. The lowest number of orders during the day time is also observed on a Wednesday.

The customers usually order once a week. But it can also be noted that that there is a weekly(also bi-weekly, tri-weekly) and Monthly trend observed from the data.

The usual order for most customers is fruits and vegetables, dairy products. And the re-order ratio follows the same trend. Another important find is that customers prefer organic items.