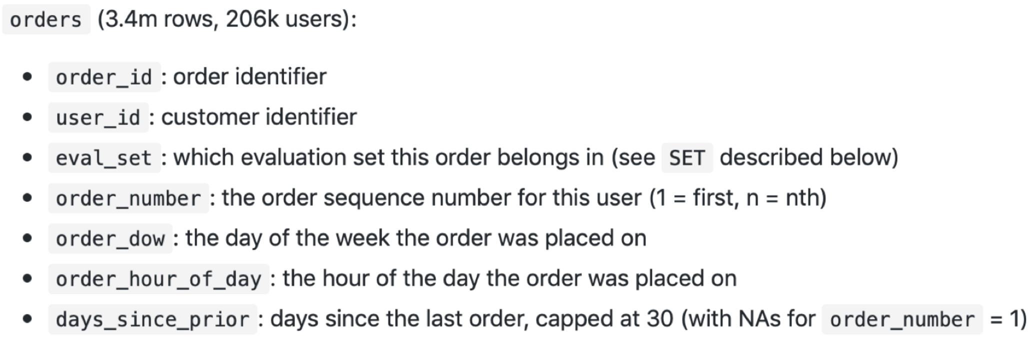
Instacart Basket Analysis

Amal Radhakrishnan

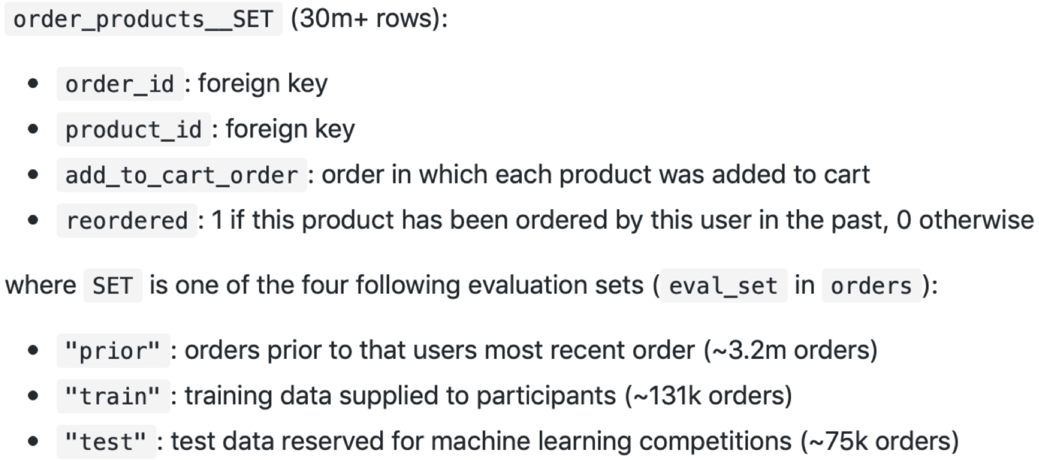
The Data

- The dataset is a relational set of files describing customers’ orders over time. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users.



`orders` (3.4m rows, 206k users):

- `order_id` : order identifier
- `user_id` : customer identifier
- `eval_set` : which evaluation set this order belongs in (see `SET` described below)
- `order_number` : the order sequence number for this user (1 = first, n = nth)
- `order_dow` : the day of the week the order was placed on
- `order_hour_of_day` : the hour of the day the order was placed on
- `days_since_prior` : days since the last order, capped at 30 (with NAs for `order_number` = 1)

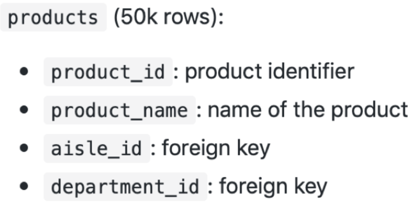


`order_products__SET` (30m+ rows):

- `order_id` : foreign key
- `product_id` : foreign key
- `add_to_cart_order` : order in which each product was added to cart
- `reordered` : 1 if this product has been ordered by this user in the past, 0 otherwise

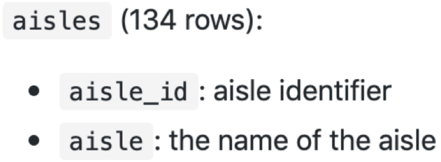
where `SET` is one of the four following evaluation sets (`eval_set` in `orders`):

- `"prior"` : orders prior to that users most recent order (~3.2m orders)
- `"train"` : training data supplied to participants (~131k orders)
- `"test"` : test data reserved for machine learning competitions (~75k orders)



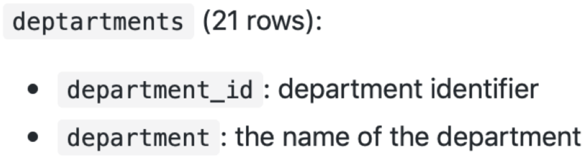
`products` (50k rows):

- `product_id` : product identifier
- `product_name` : name of the product
- `aisle_id` : foreign key
- `department_id` : foreign key



`aisles` (134 rows):

- `aisle_id` : aisle identifier
- `aisle` : the name of the aisle

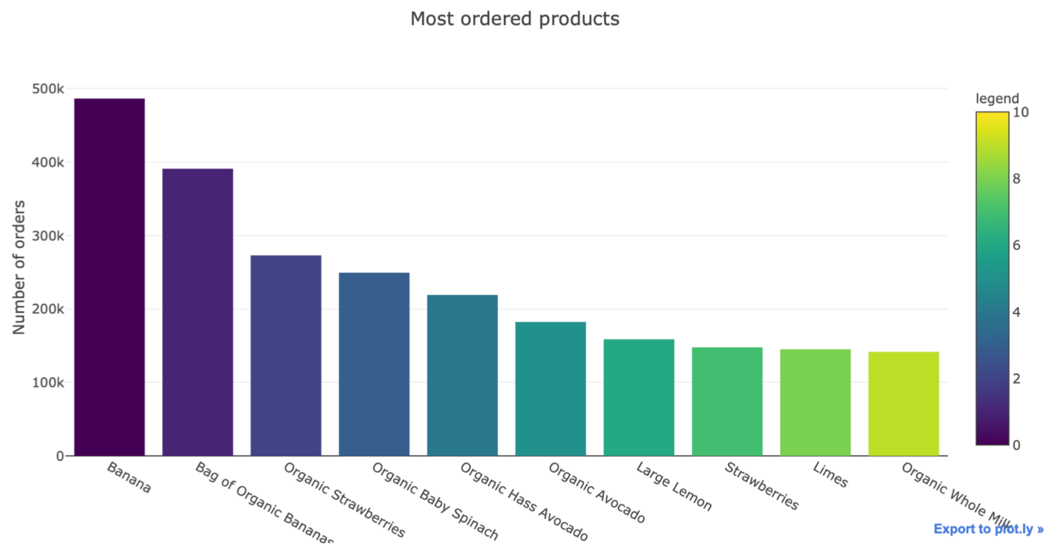


`deptartments` (21 rows):

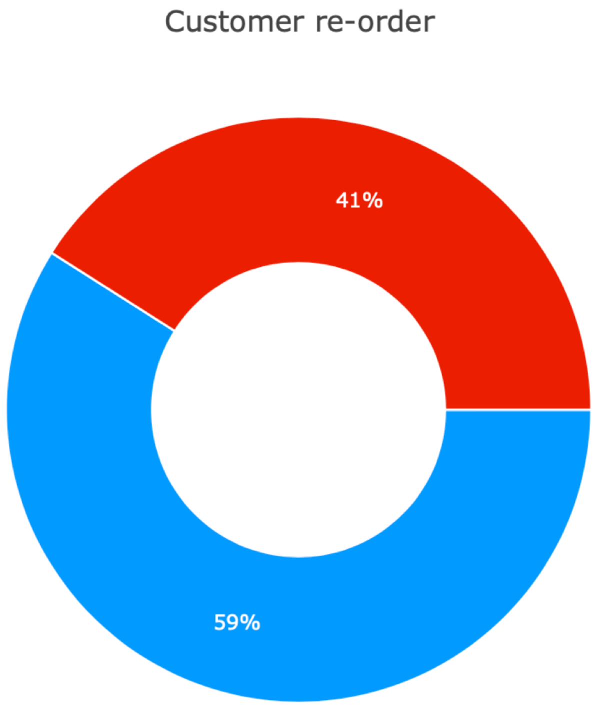
- `department_id` : department identifier
- `department` : the name of the department

Exploratory Data Analysis

The most ordered products were banana, strawberries, avocado, limes, milk etc.



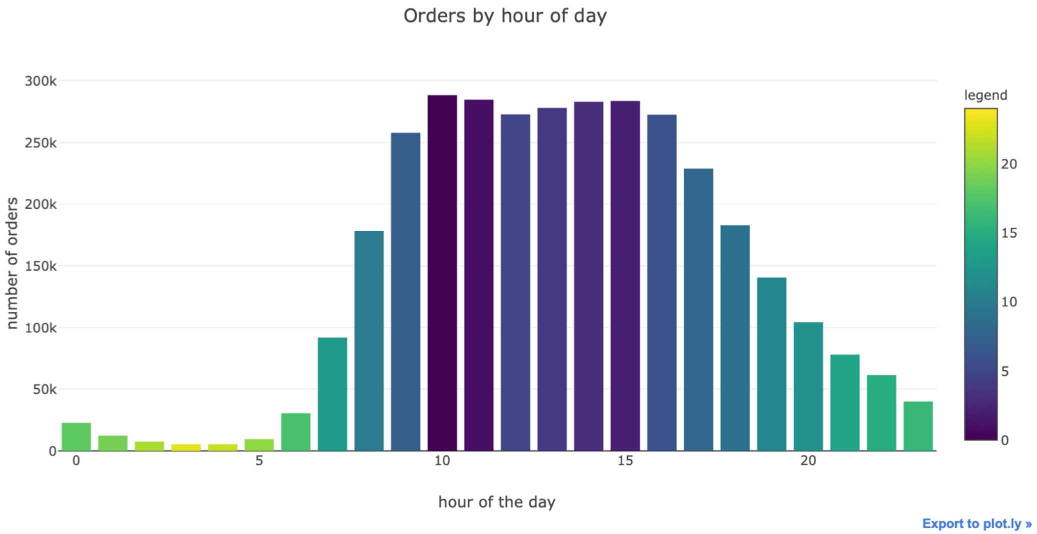
Around 59% percentage of all products in orders are reordered at some point.

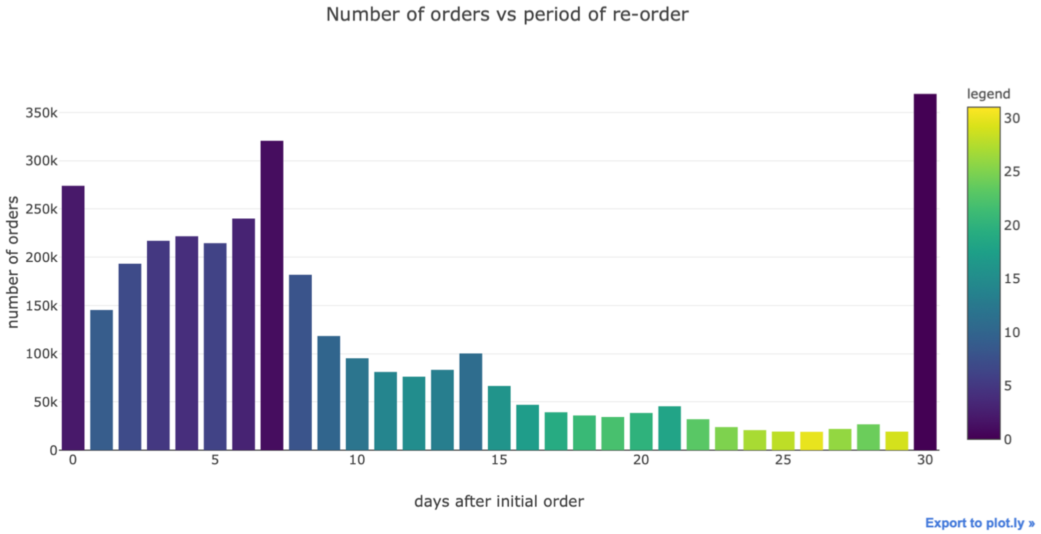


Exploratory Data Analysis

Peak hours for online shopping tends to be around 10 am to 3 pm.

Customers also tend to keep recurring reorders either at an interval of 7 days or 30 days.

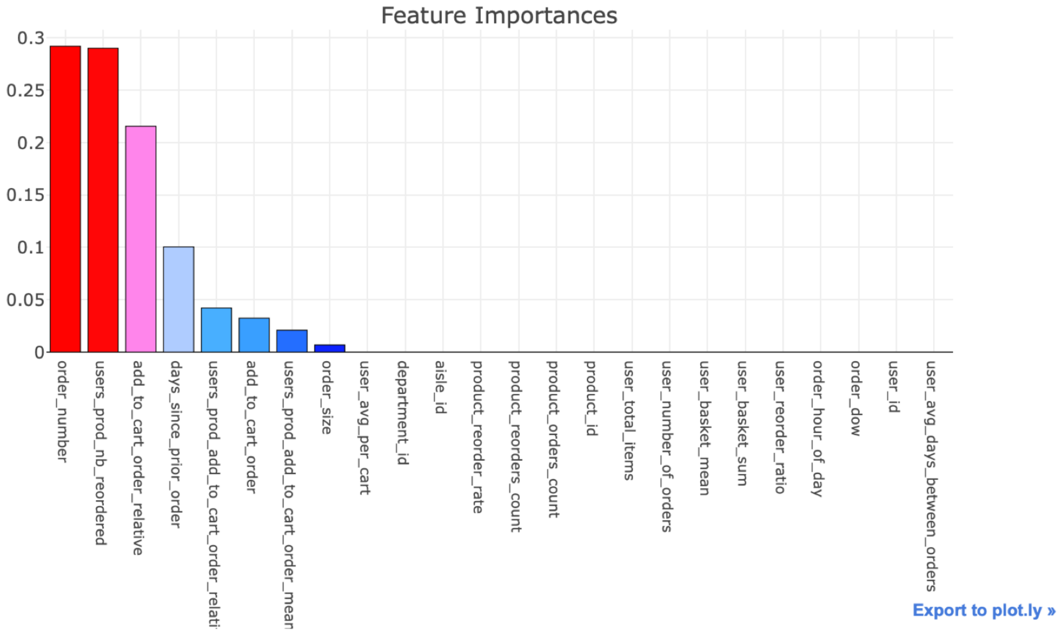




Exploratory Data Analysis

1. About 6.4% of the columns in the prior orders contained null values.
2. Chocolate, Milk, Energy drink, Banana etc. had the highest reorder probability (with a minimum threshold of 100 reorders)
3. Most popular (number of orders) department is Produce and Personal Care has the greatest number of distinct products.
4. I observed a right skewed distribution for number of orders vs number of products. Most orders have 4-8 different products in the cart.

Important features



These are the identified important features for predicting reorder behavior of a customer.

- order_number
- user same product reorder number
- add to cart order relative
- days since prior order
- user product add to cart order relative
- add to cart order
- size of the order

The Recommendation Engine

The Engine was built using spark FP-growth. FP-growth is a program for frequent item set mining, a data mining method that was originally developed for market basket analysis.

Example recommendations -

Order Number 2115

Last order - [**'Organic** Mixed Vegetables', **'Organic** Broccoli Florets', 'Cheese Pizza Snacks', **'Organic** Spring Mix Salad']

Recommendation - ['Bag of **Organic** Bananas', 'Banana', **'Organic** Strawberries', 'Large Lemon', **'Organic** Raspberries', **'Organic** Baby Spinach', **'Organic** Hass Avocado’]

Order Number 904

Last Order - ['Cup Noodles Chicken Flavor', **'Zero Calorie Cola**']

Recommendation - [**'Soda**']

Future Work

- Improving feature engineering for analyzing reorder patterns – Exploring ability to process more features using word2vec from product, department texts. Having access to more hardware resources will allow generating more user-product features.
- Using more powerful Neural Networks.
- Using feature interactions for descriptive analysis and reorder prediction.