

Appendix

Data Analysis

The Dataset

Orders from Instacart are available in four .csv files: "orders.csv", "order_products__train.csv", "order_products__prior.csv" and "sample_submission.csv". The key to understand the dataset and the train / test split is the orders table ("orders.csv").

Take for example User 1[Fig 1] , who happens to be a train user. User 1 has 10 prior orders, and 1 train order whose details are provided in "order_products__prior.csv" and in "order_products__train.csv" respectively.

Similarly, User 4 is a test user. He has 5 prior orders, and his 6th is a test order. Their details are available in "order_products__prior.csv" and "sample_submission.csv" respectively.

Figure 2 is a glimpse at "order_products__prior.csv" when merged with three other .csv files that represent products, aisles and departments. The format of "sample_submission.csv" and "order_products__train.csv" is exactly the same.

Exploration and Analysis

How many orders have users placed?

The below histogram validates the claim that 4 to 100 orders of a customer are given.

| 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 |
| 2254736 | 1 | prior | 4 | 4 | 07 | 29.0 |
| 431534 | 1 | prior | 5 | 4 | 15 | 28.0 |
| 3367565 | 1 | prior | 6 | 2 | 07 | 19.0 |
| 550135 | 1 | prior | 7 | 1 | 09 | 20.0 |
| 3108588 | 1 | prior | 8 | 1 | 14 | 14.0 |
| 2295261 | 1 | prior | 9 | 1 | 16 | 0.0 |
| 2550362 | 1 | prior | 10 | 4 | 08 | 30.0 |
| 1187899 | 1 | train | 11 | 4 | 08 | 14.0 |

(a) User 1

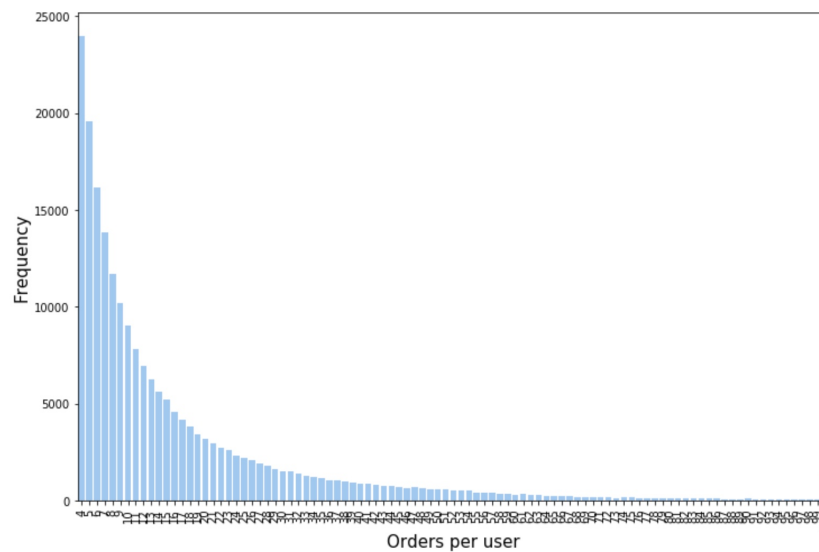
| | | | | | | |
|---------|---|-------|---|---|----|------|
| 3343014 | 4 | prior | 1 | 6 | 11 | |
| 2030307 | 4 | prior | 2 | 4 | 11 | 19.0 |
| 691089 | 4 | prior | 3 | 4 | 15 | 21.0 |
| 94891 | 4 | prior | 4 | 5 | 13 | 15.0 |
| 2557754 | 4 | prior | 5 | 5 | 13 | 0.0 |
| 329954 | 4 | test | 6 | 3 | 12 | 30.0 |

(b) User 4

Figure 1: Train/Test Split

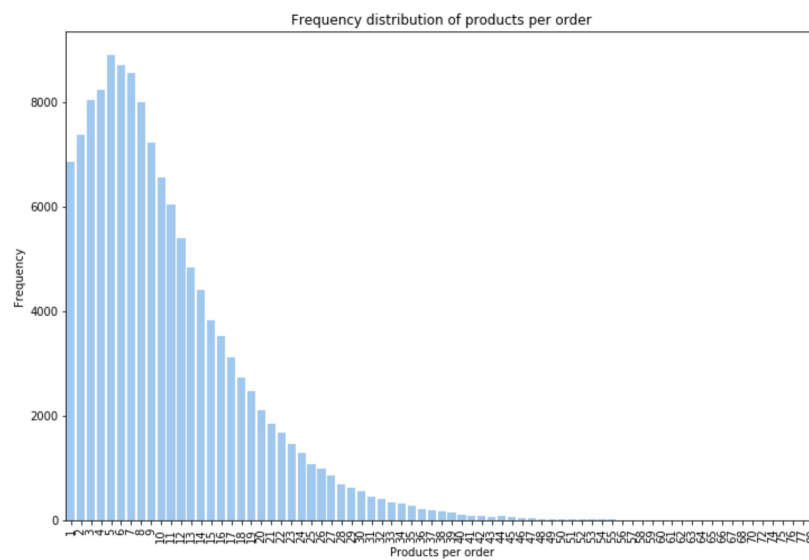
| order_id | product_id | add_to_cart_order | reordered | product_name | aisle_id | department_id | aisle | department |
|----------|------------|-------------------|-----------|--------------|-----------------------|---------------|-------|---------------------------|
| 0 | 2 | 33120 | 1 | 1 | Organic 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 |

Figure 2: Merged Prior Orders



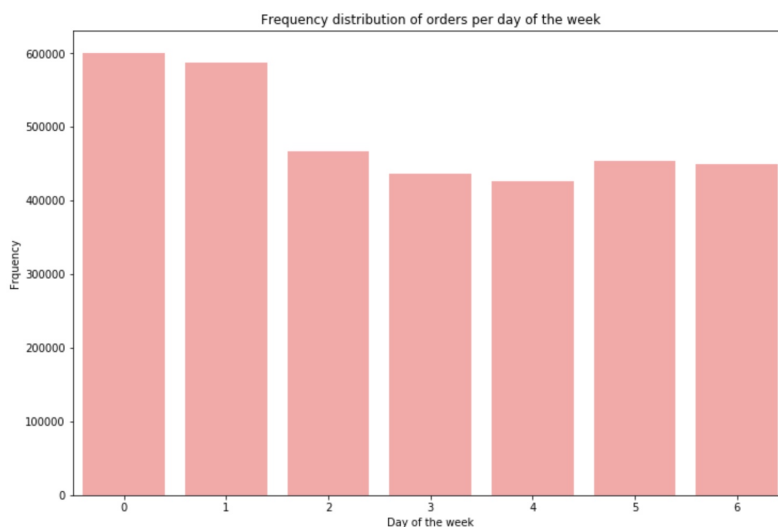
How many products does an order have??

The "long tail" phenomenon is clearly visible here.



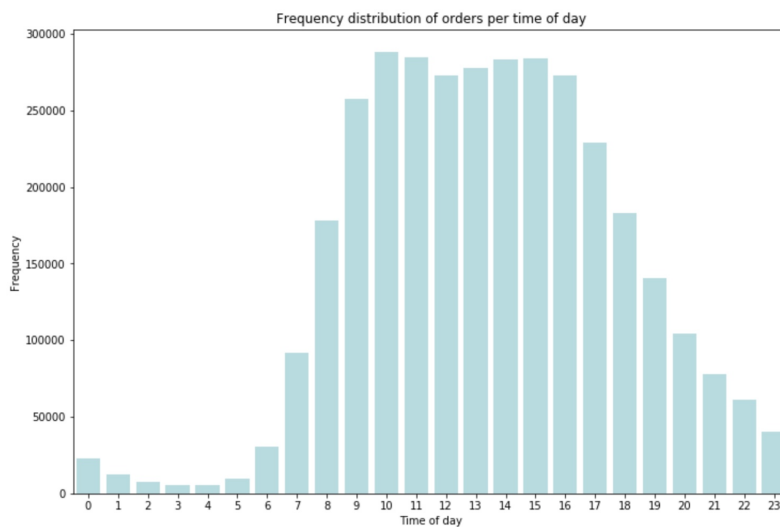
Does day of the week influence user order habits?

There is a clear effect of day of the week. Most orders are on days 0 and 1. However, there is no information about which values represent which day, but, it's reasonable to assume they are weekends.



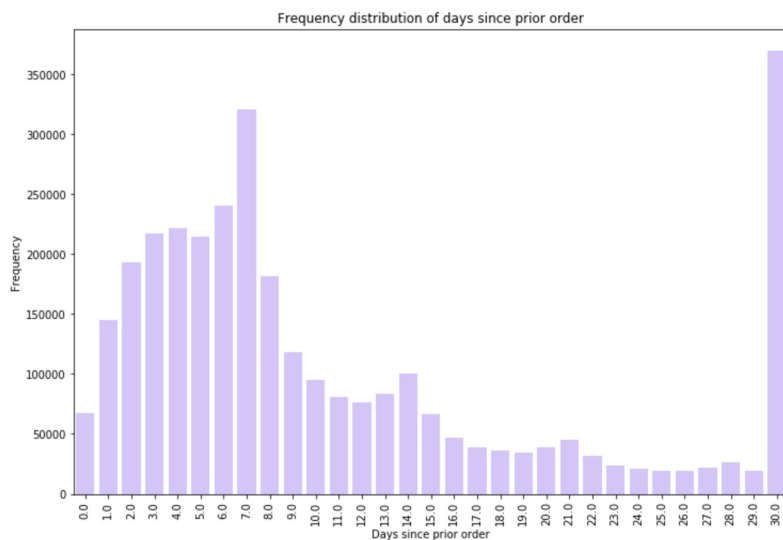
Does time of day influence user order habits?

There is a clear effect. Most of the orders are placed between 7.00 am and 10.00pm.



When do users reorder?

Users seem to order on a weekly and monthly basis.



What are the top 20 Products?

