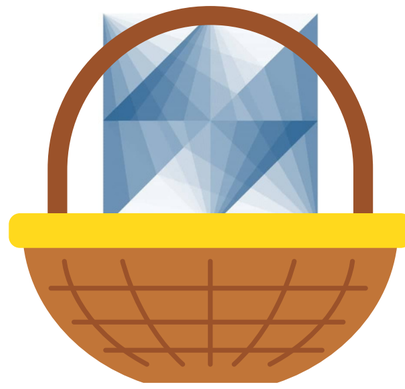
Recommendation and Market Basket Analysis

Supervised By: Dr. Doaa Mahmoud

Who are we?

Noran Hany

Toka Khaled



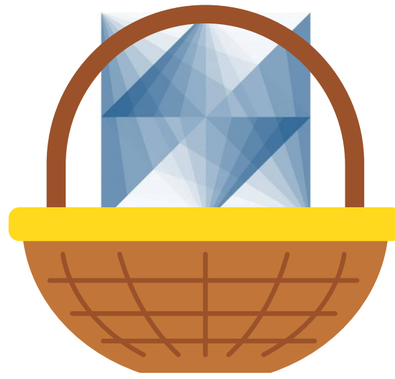
Motivation

Ecommerce is a very rich problem. A lot of rich insights, and ∞ ideas can be extracted. Also, new contributions can be added.

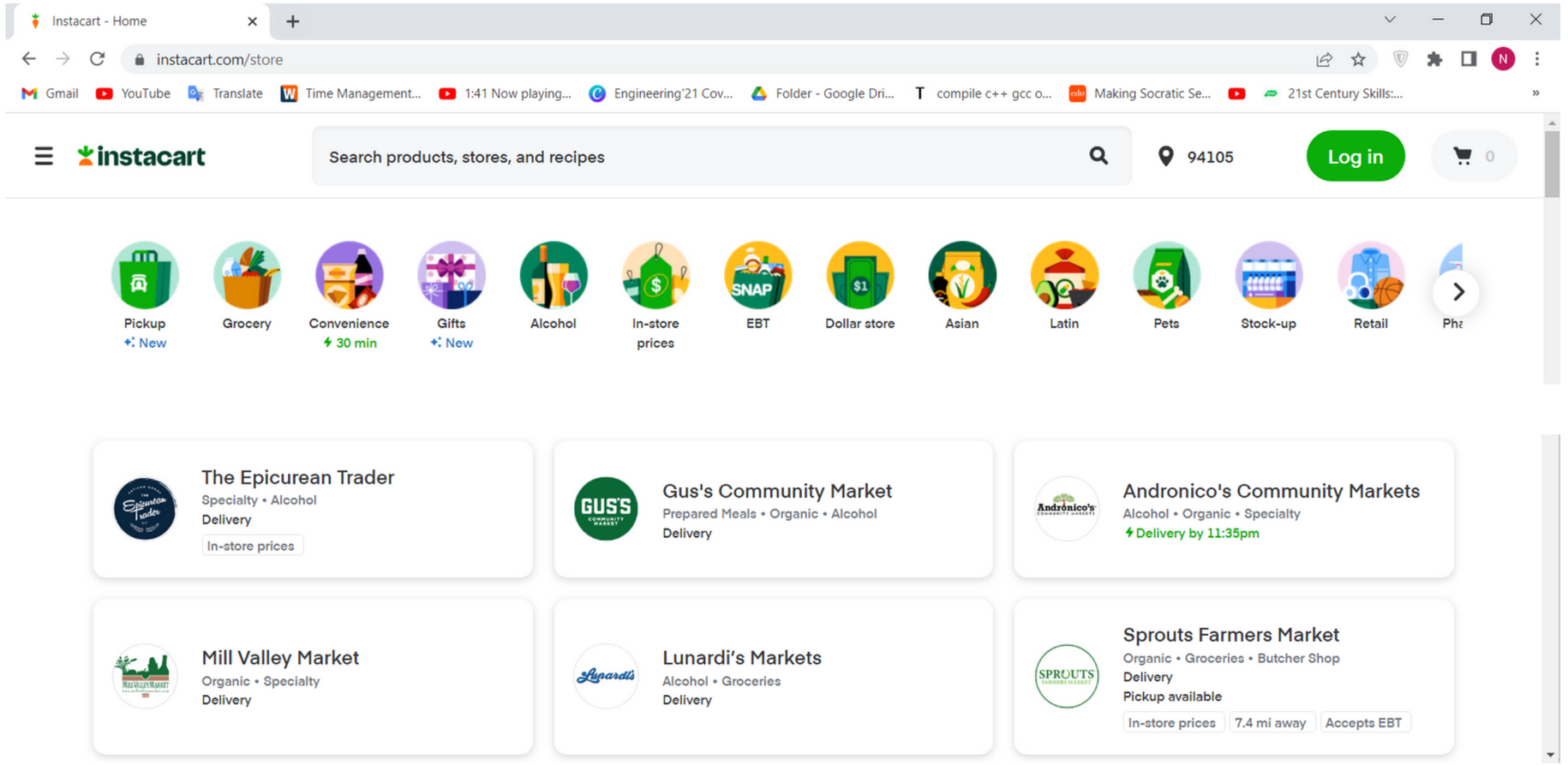
Problem Definition

What's instacart?

Instacart, a grocery ordering and delivery app. After selecting products through the Instacart app, personal shoppers review your order and do the in-store shopping and delivery for you.



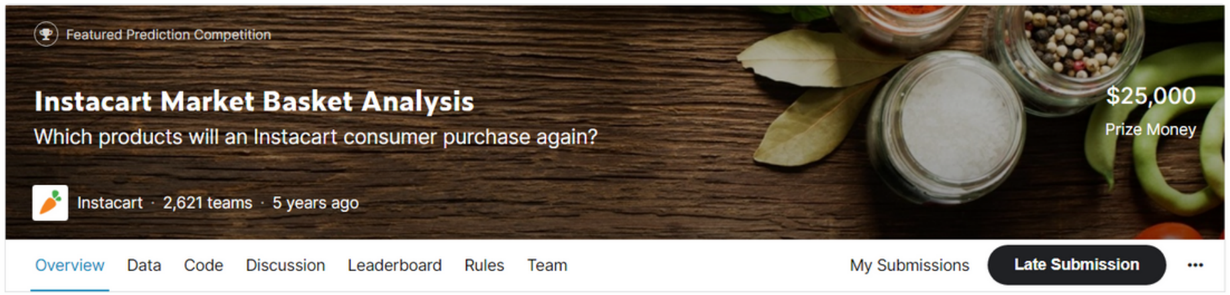
Problem Definition |



Problem Definition

At 2017, they Open Sourced for the first time 3 Millions instacart orders

They hosted a competition to use this anonymized data on customer orders over time to predict: **which previously purchased products will be in a user's next order?**



Literature Review

We searched and investigated a paper and many kaggle notebooks. Contributors mainly worked on:

Customer Segmentation

Recommender Model for next order

Association Rules between products

Paper Link:
SESUG Paper 252-2019 | Market Basket Analysis on Instacart
Aravind Dhanabal,
Oklahoma State University

Literature Gap

Contributors didn't make good use of the time-dependent data provided.

Contributors we just using associations rules, for the sake of getting the rules. They didn't introduced how to make use of these rules from business point of view.

Data Description

Data is divided to 3 sets:

- Prior data : Orders' history of every user.
  - Nearly 3–100 past orders per user
- Train data : Future order data of every user.
- Test data : Future order data of every user.

Data Description

Aisles [~134 sub-department]

	aisle_id	aisle
0	1	prepared soups salads
1	2	specialty cheeses
2	3	energy granola bars

Departments [~21department]

	department_id	department
0	1	frozen
1	2	other
2	3	bakery
3	4	produce

Products [~48K products]

	product_id	product_name	aisle_id	department_id
0	1	Chocolate Sandwich Cookies	61	19
1	2	All-Seasons Salt	104	13
2	3	Robust Golden Unsweetened Oolong Tea	94	7

Data Description

Data is divided across 6 files:

order_products [2 files]

- One for the basket of previously orders. [~33M rows]
- Another for the basket of the future orders.

	order_id	product_id	add_to_cart_order	reordered
0	1	49302	1	1
1	1	11109	2	1
2	1	10246	3	0

Data Description

orders [~3M orders]

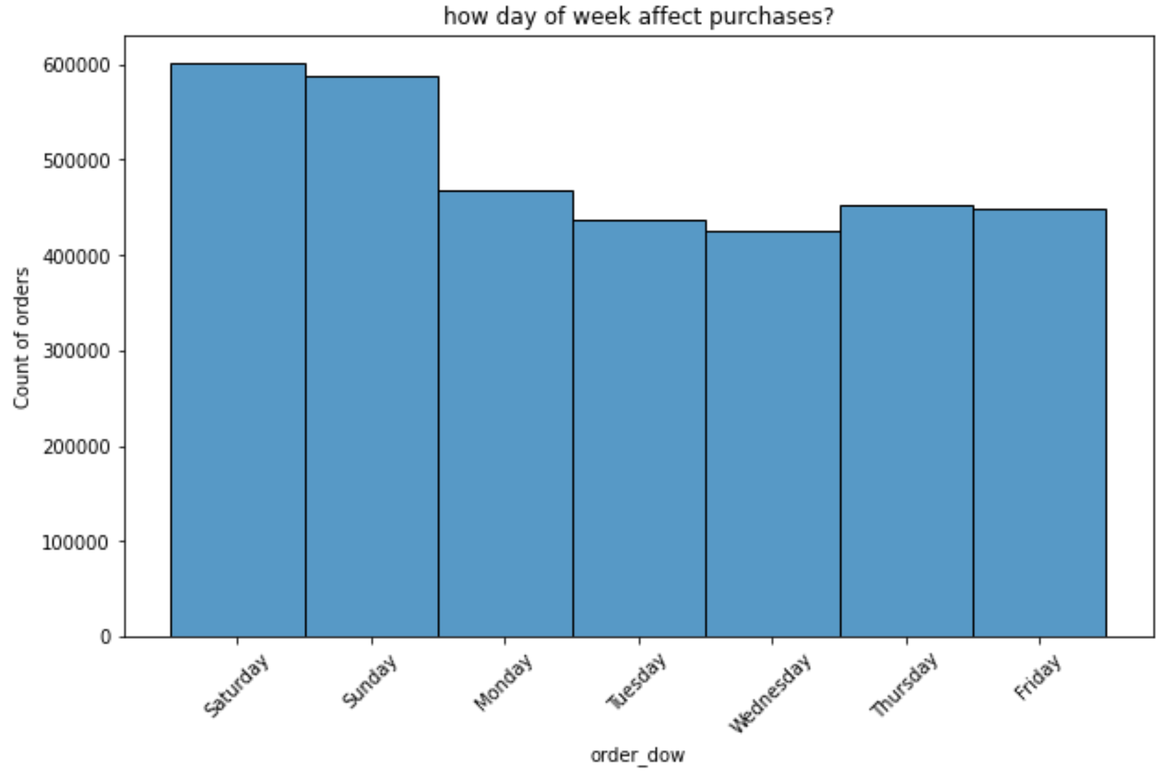
order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
2539329	1	prior	1	2	8	NA
2398795	1	prior	2	3	7	15
473747	1	prior	3	3	12	21

EDA

Days of week column only have numbers from 0-6

It didn't hold, which number correspond to which day of week

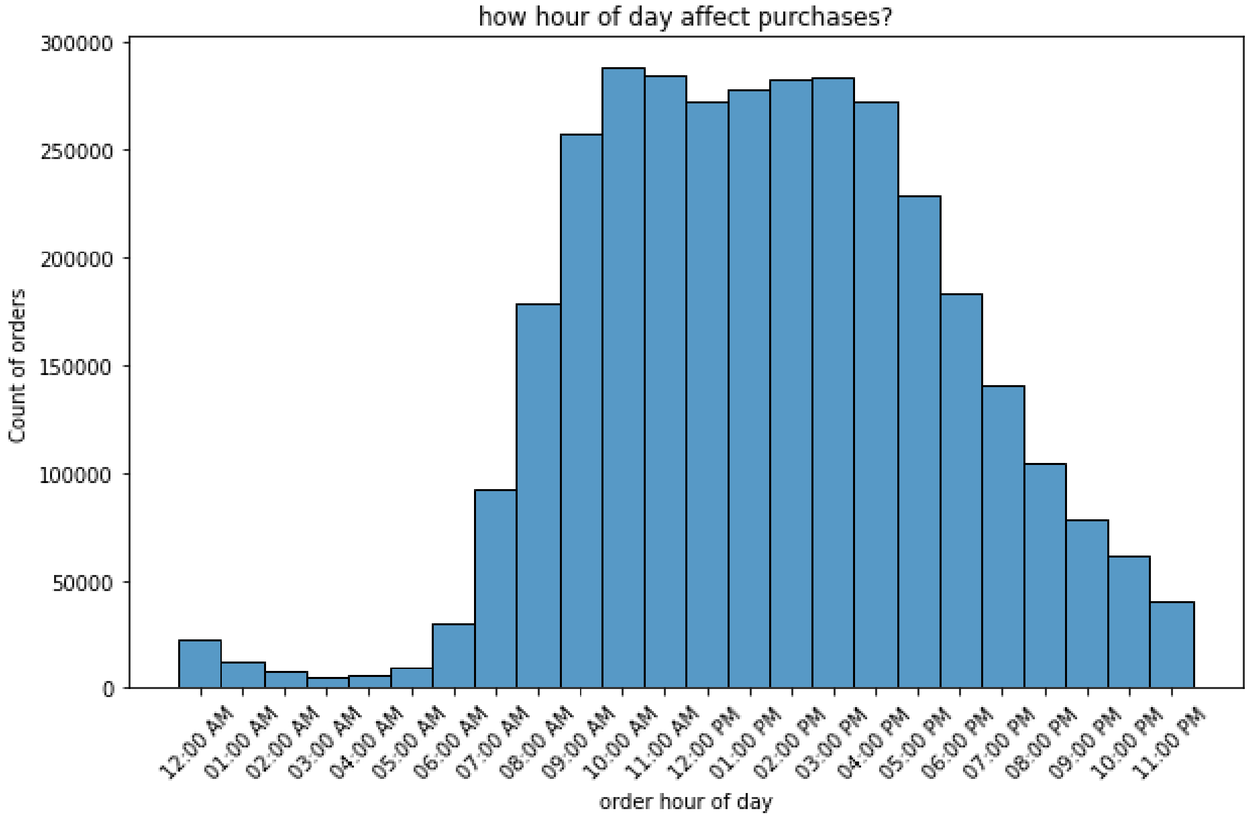
From the days of week histogram, Most orders were purchased on Day 0 and Day 1.



Thus we inferred that these 2 days seems to be the weekend.

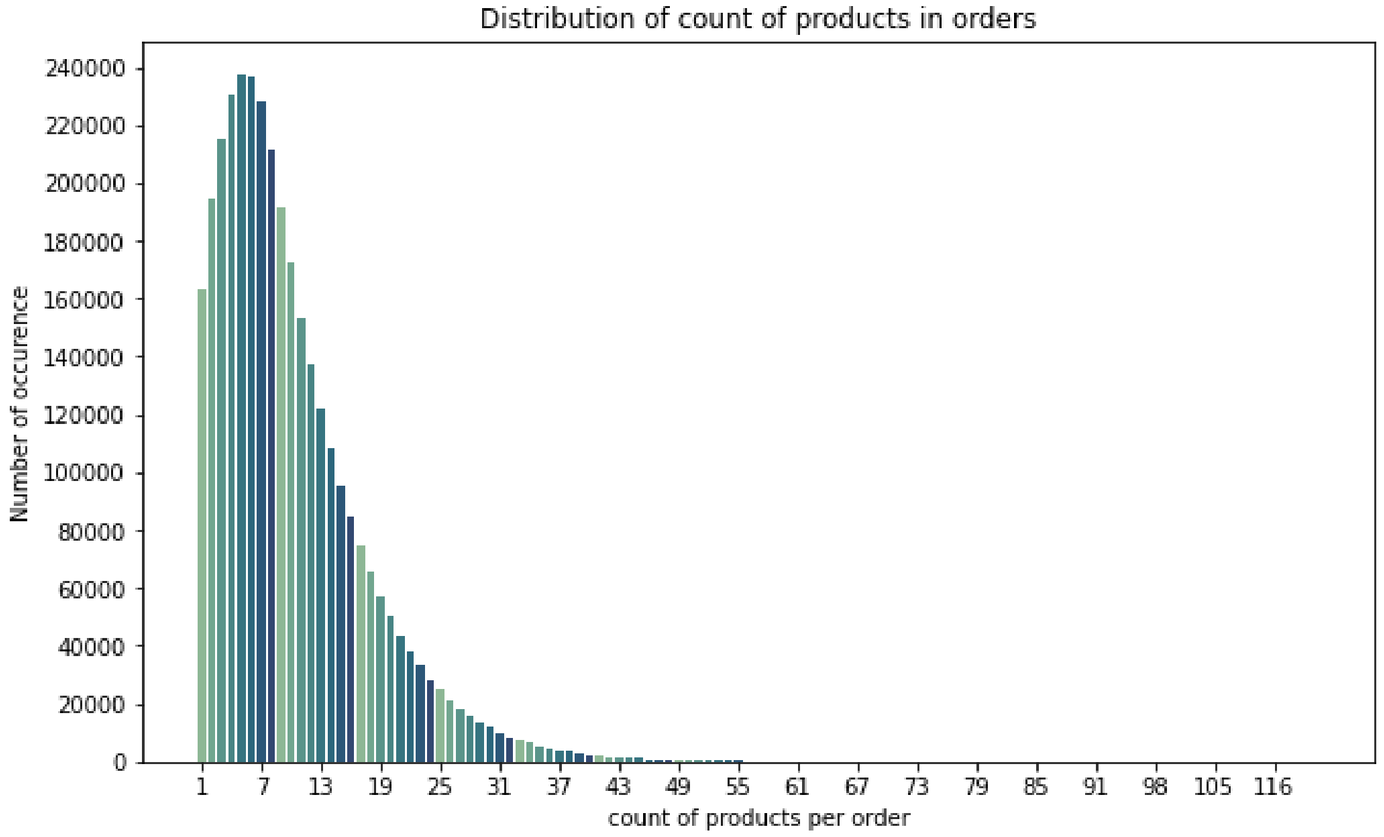
EDA

Intuitively, orders are mostly ordered during day, from 9:00 AM to 4:00 PM.



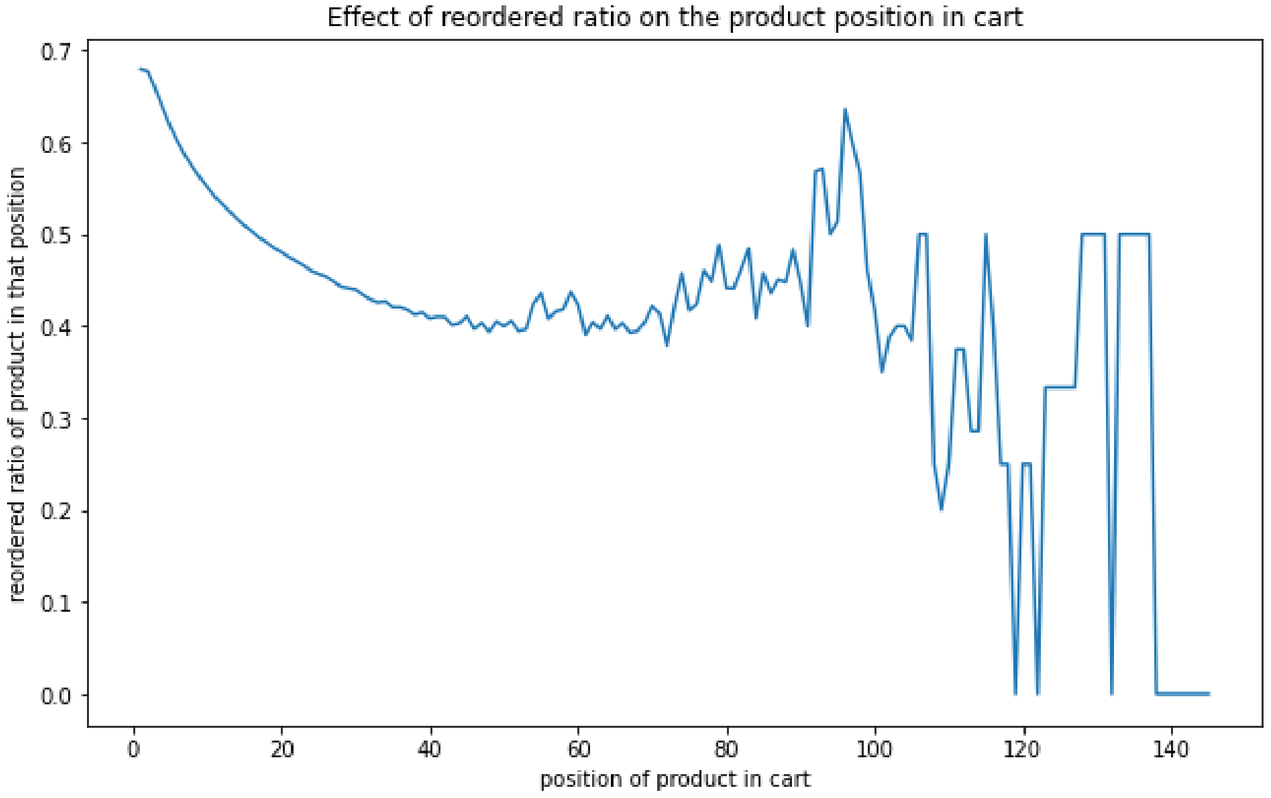
EDA

Most baskets contain from 5–8 products.



EDA

Intuitively, products placed first in cart are the products mostly reordered.
People tend to put first the products they already know.



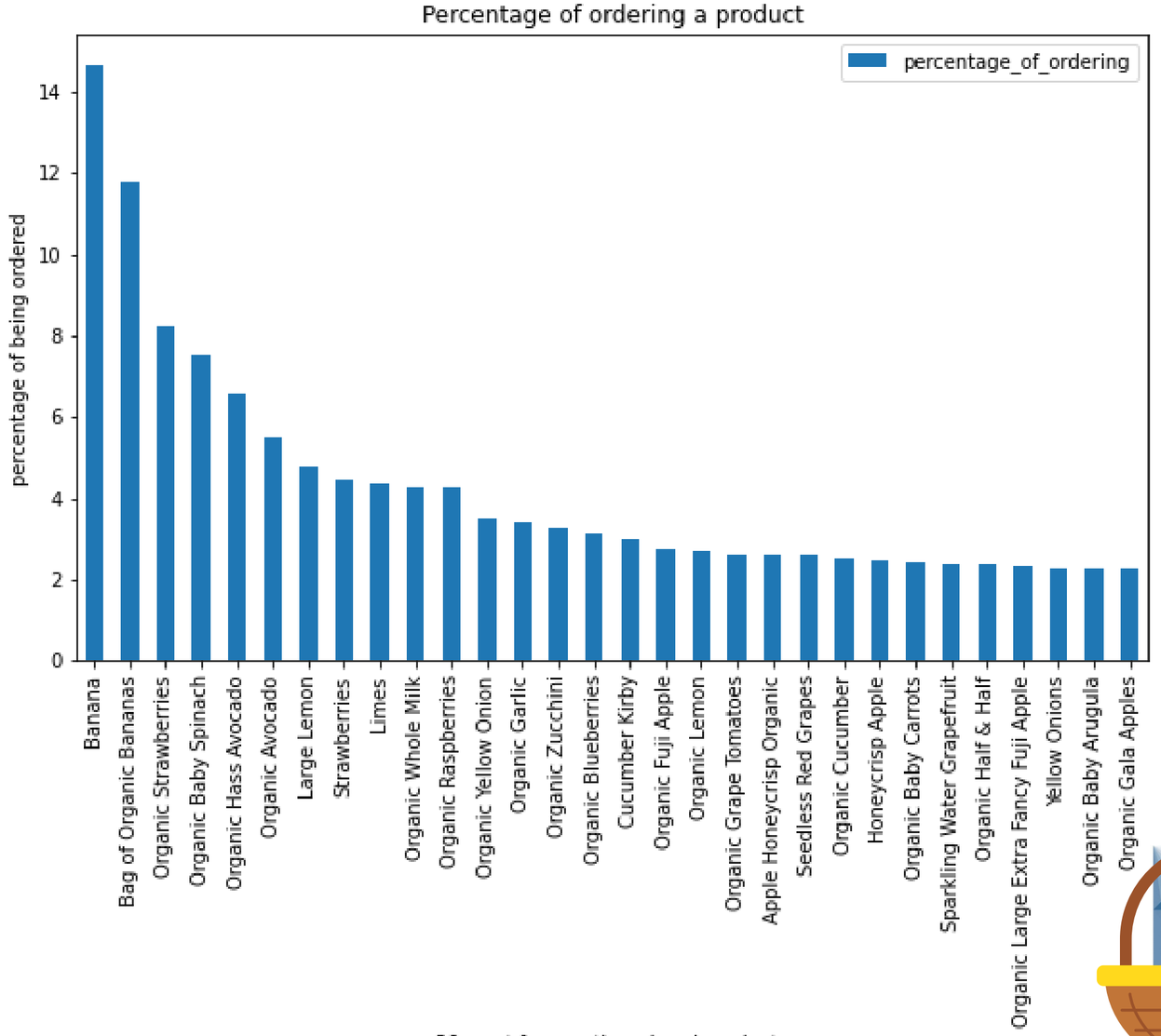
EDA

Analyzing products

5 Most Ordered Products

- Banana
- Bag of Organic Bananas
- Organic Strawberries
- Organic Baby Spinach
- Organic Hass Avocado

- 14% of all purchases are bananas.
- Organic products are frequently ordered.



EDA

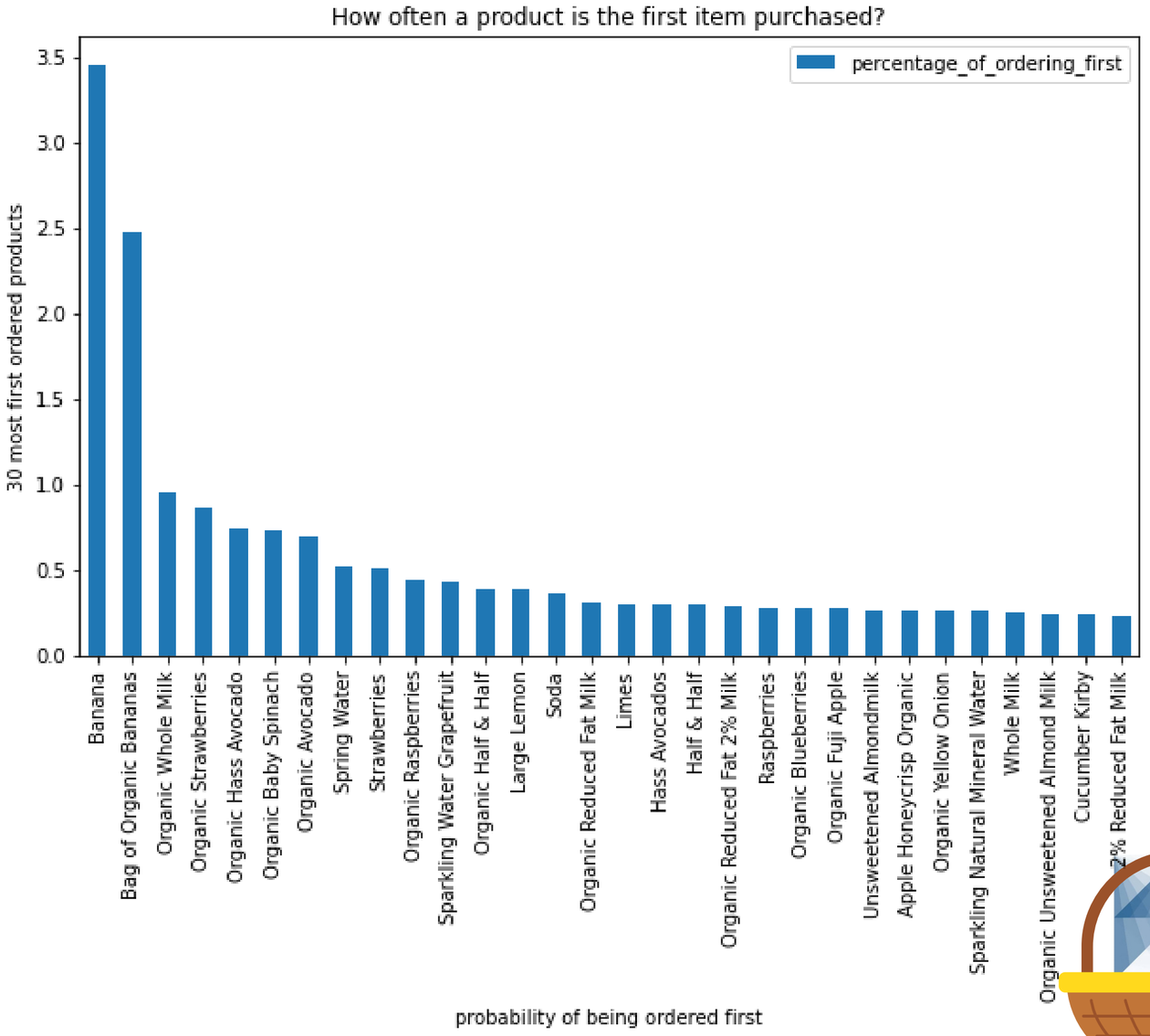
Analyzing products

From all purchases, how often a product is the first item purchased?

- 3.4% of the orders, Banana is the first product added to cart.

Intuitive, since banana is frequently bought.

Let's find another metric to measure how often a product is first placed!

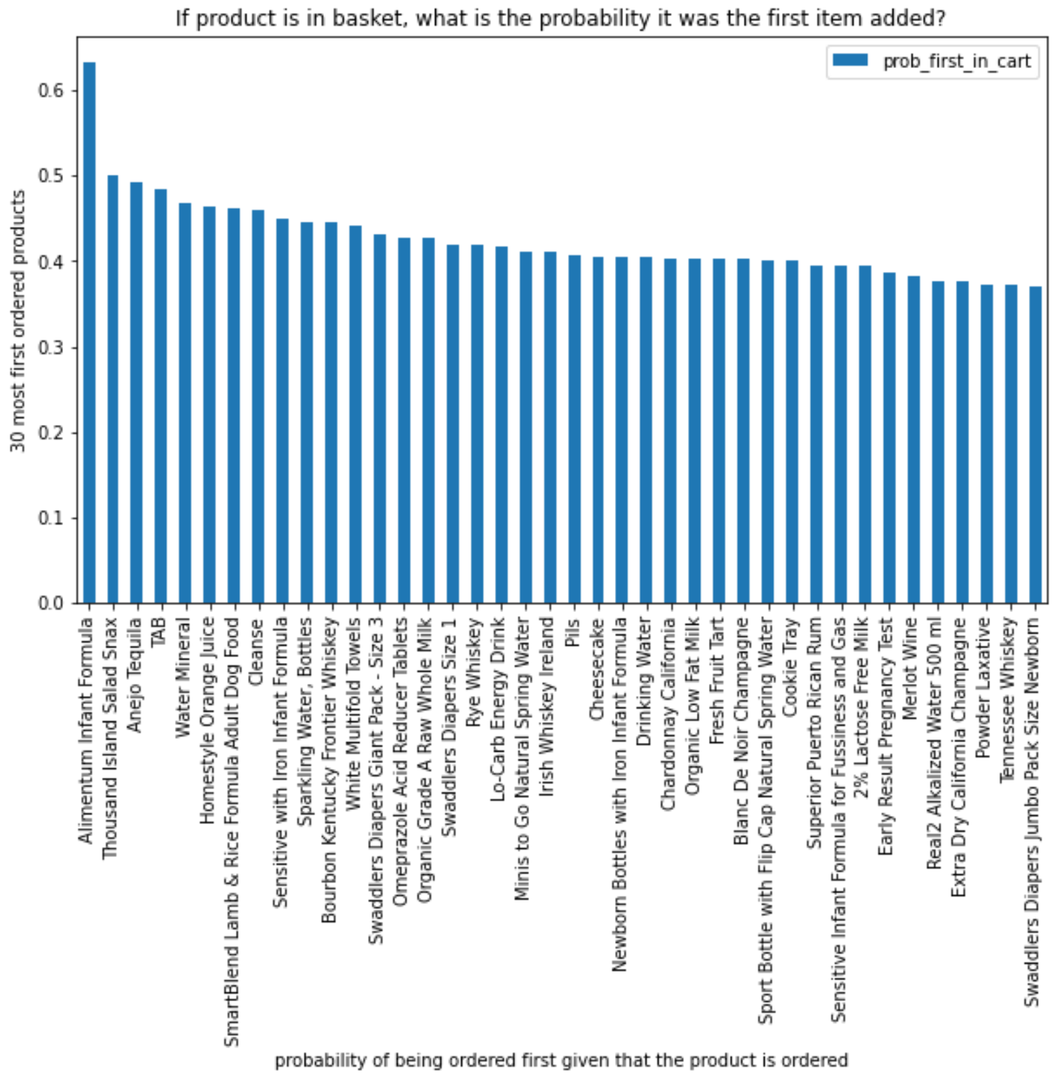


EDA

Analyzing products

Products which have been bought more than 100 times. From the orders the product only exist in, how often people put it first?

Didn't find a pattern or product if exists in basket, it will always be the first placed.



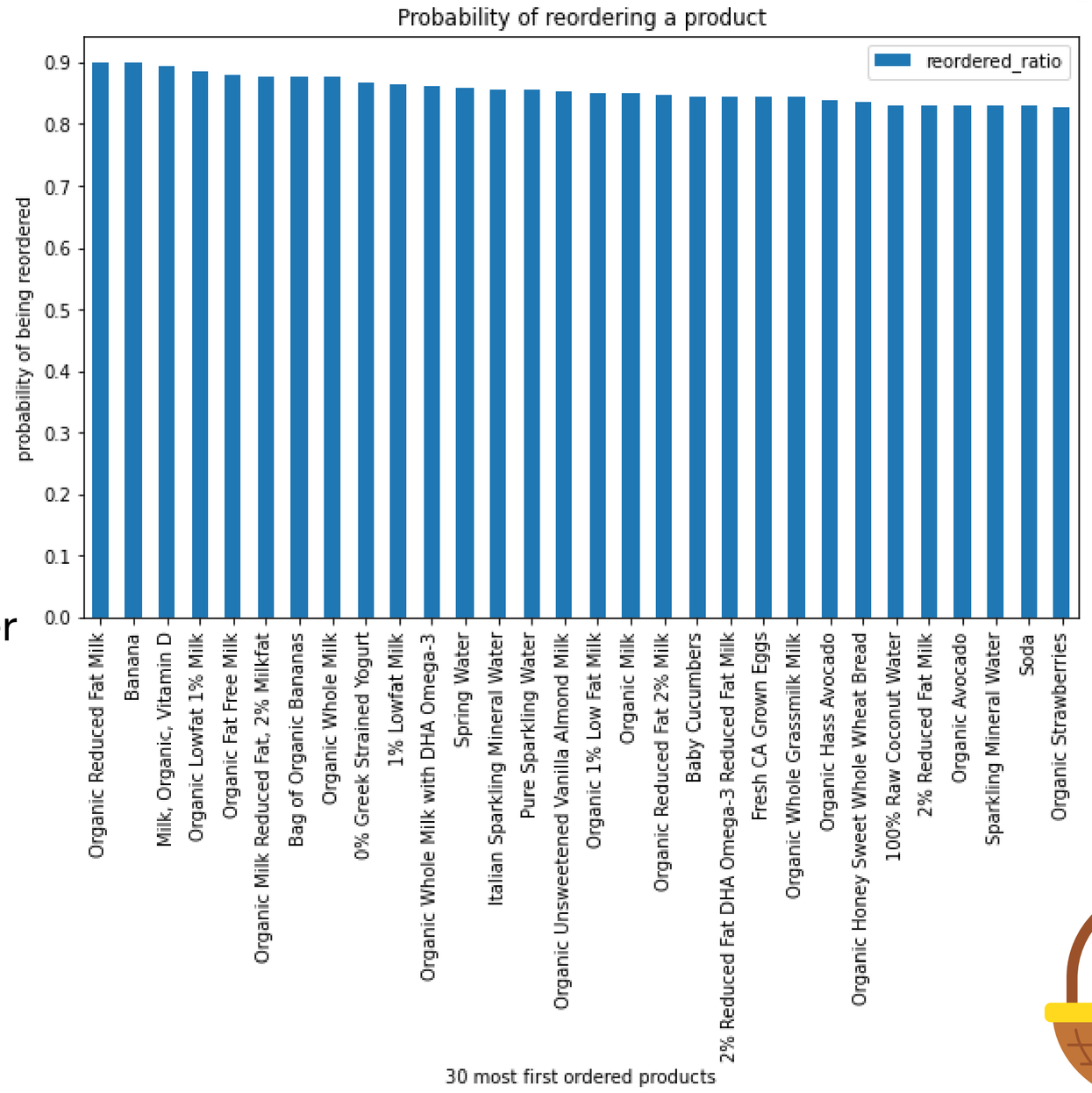
EDA

Analyzing products

For products which has over 10k purchases, we found products that most their purchases are from users who bought them before.

These products that when a customer buys, he/she will most probably buy again.

E.g. Count of total Organic Reduced Fat Milk purchases: ~36k
Count of reordered Organic Reduced Fat Milk purchases: ~31k



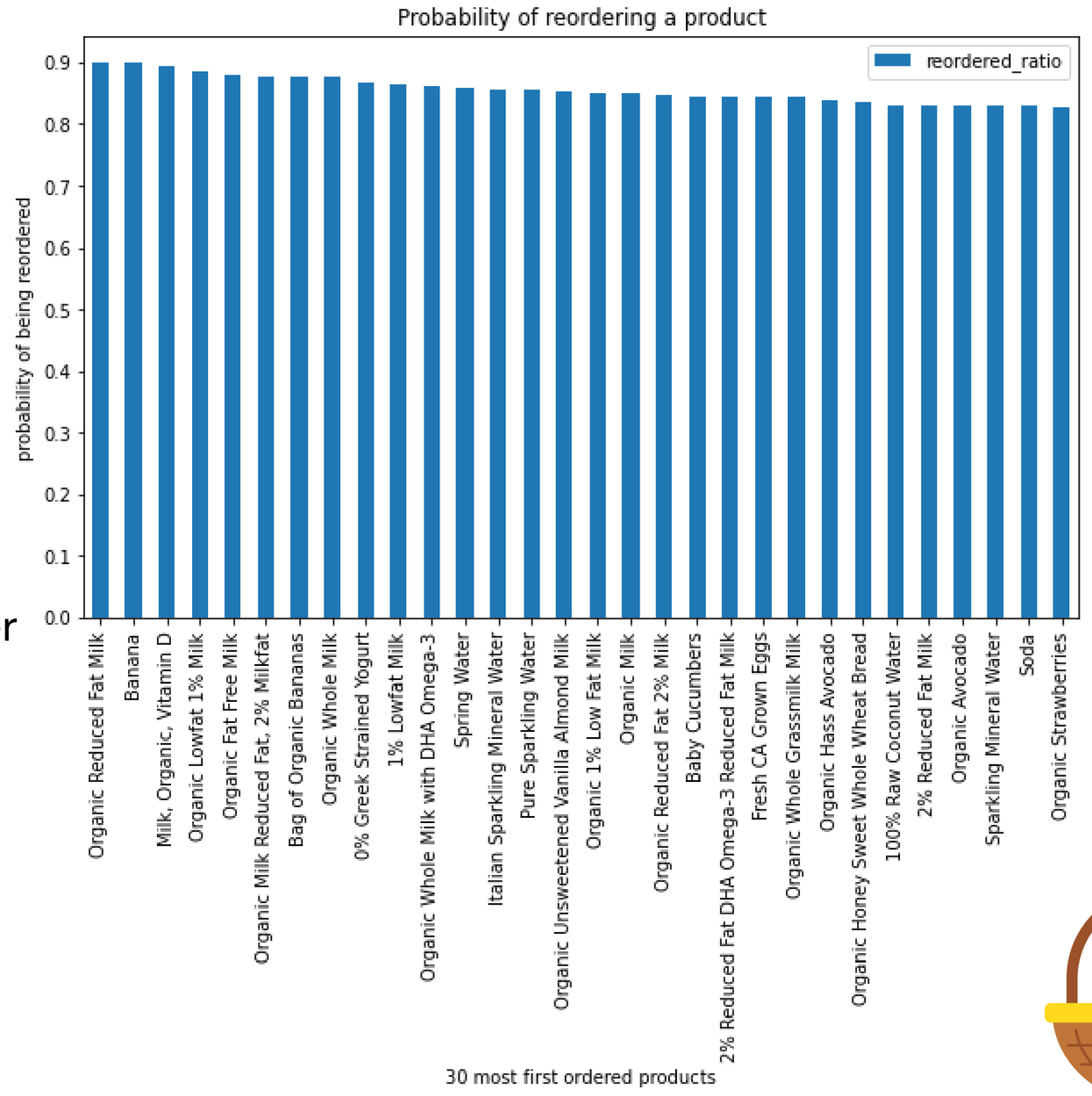
EDA

Analyzing products

For products which has over 10k purchases, we found products that most their purchases are from users who bought them before.

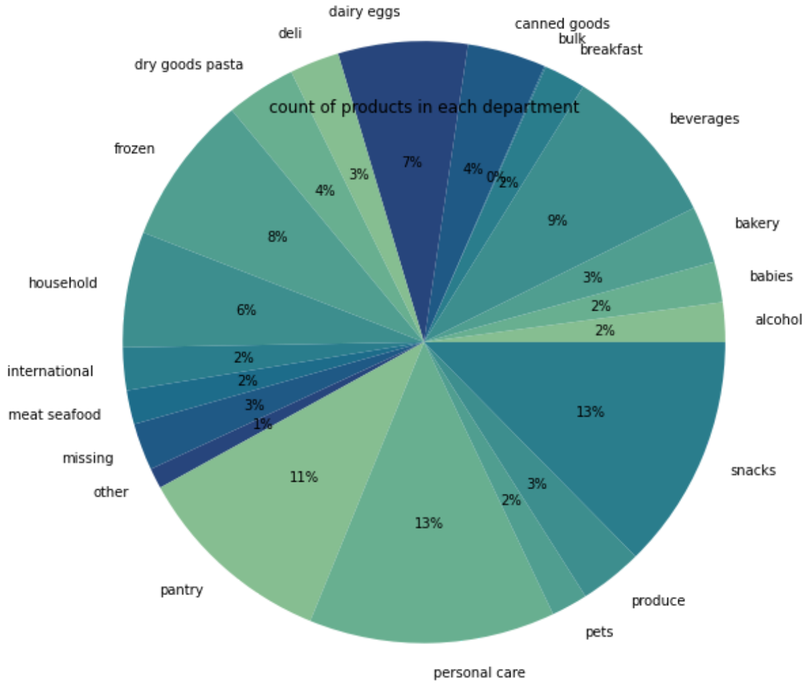
These products that when a customer buys, he/she will most probably buy again.

E.g. Count of total Organic Reduced Fat Milk purchases: ~36k
Count of reordered Organic Reduced Fat Milk purchases: ~31k

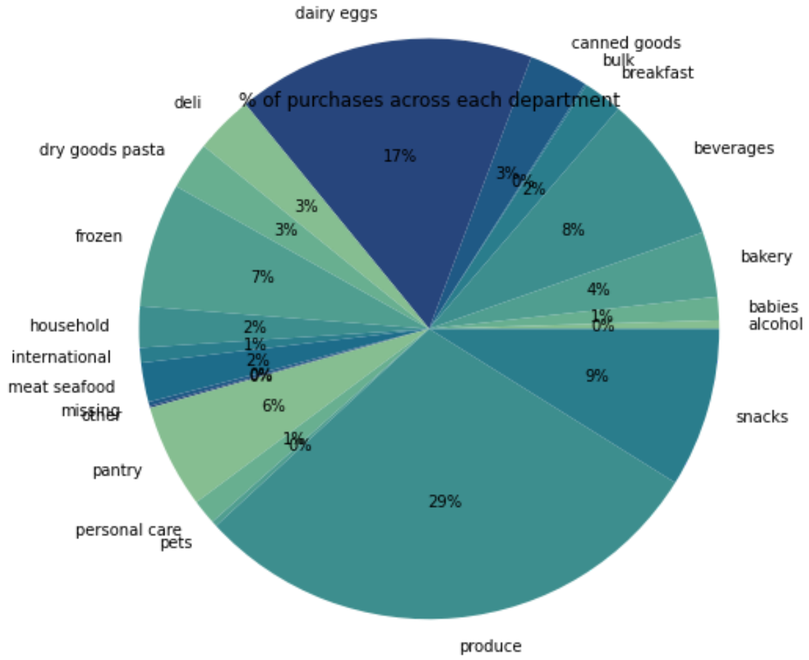


EDA

Analyzing departments



Distribution pf products across departments



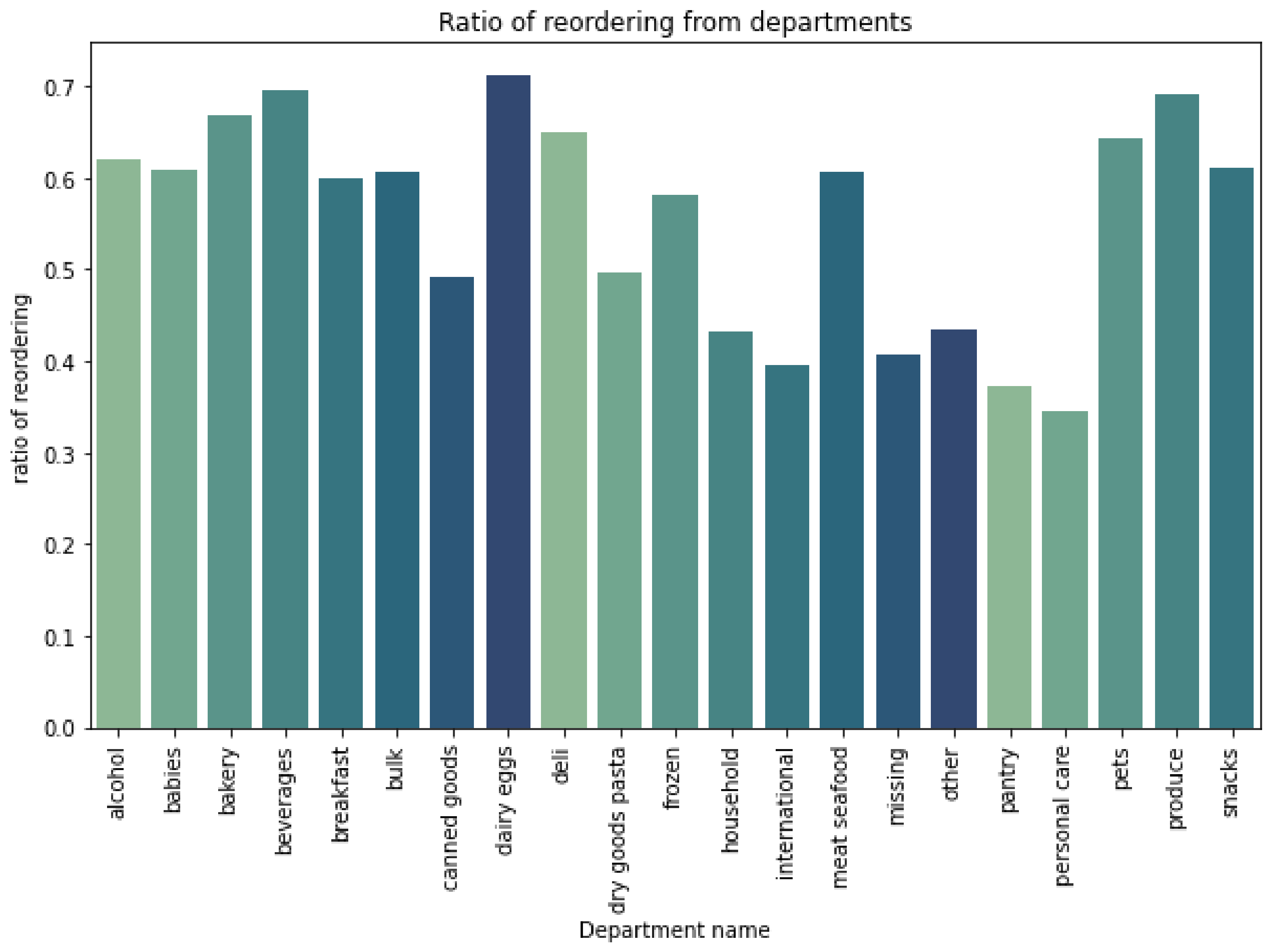
Distribution pf purchases across departments

It seems that produce contains very few products (2% of the products), but 29% of total purchases.

EDA

Analyzing departments

People how buy products from **dairy eggs and produce** will most probably buy them again.



Business Questions

How to make customers never forget instacart?

Which aisle or department to consider adding or offering more products to it?

When to avoid recommending new products?

When it's best to recommend a customer new products or a less frequently bought product?

When to recommend customized products to a user?

When is it most beneficial to both customer and business to make free coupons and offers?

Business Questions

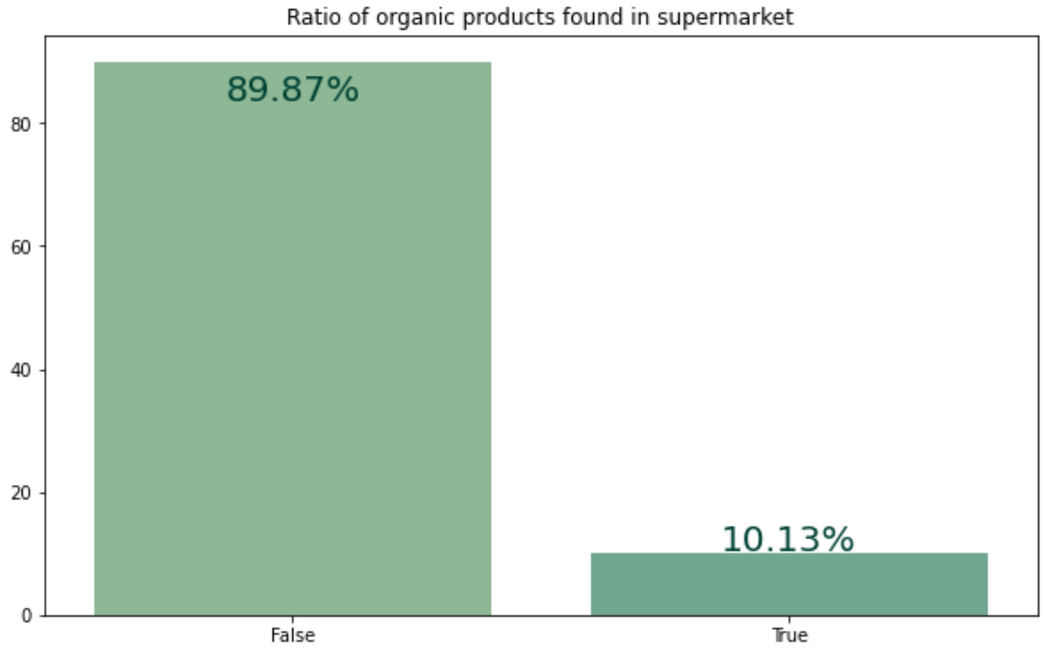
Which aisle or department to consider adding or offering more products to it?

We found that some product names holds "organic" in them! So we decided to analyze them.

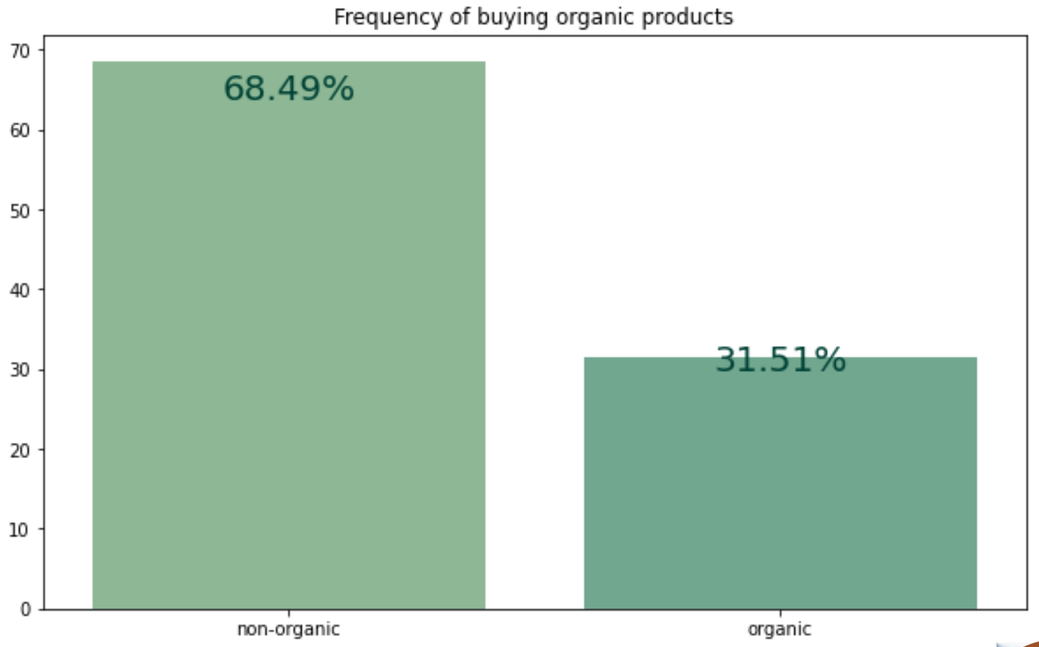
	product_id	product_name
20	21	Small & Medium Dental Dog Treats
21	22	Fresh Breath Oral Rinse Mild Mint
22	23	Organic Turkey Burgers

Business Questions

Analyzing organic products



They are few!



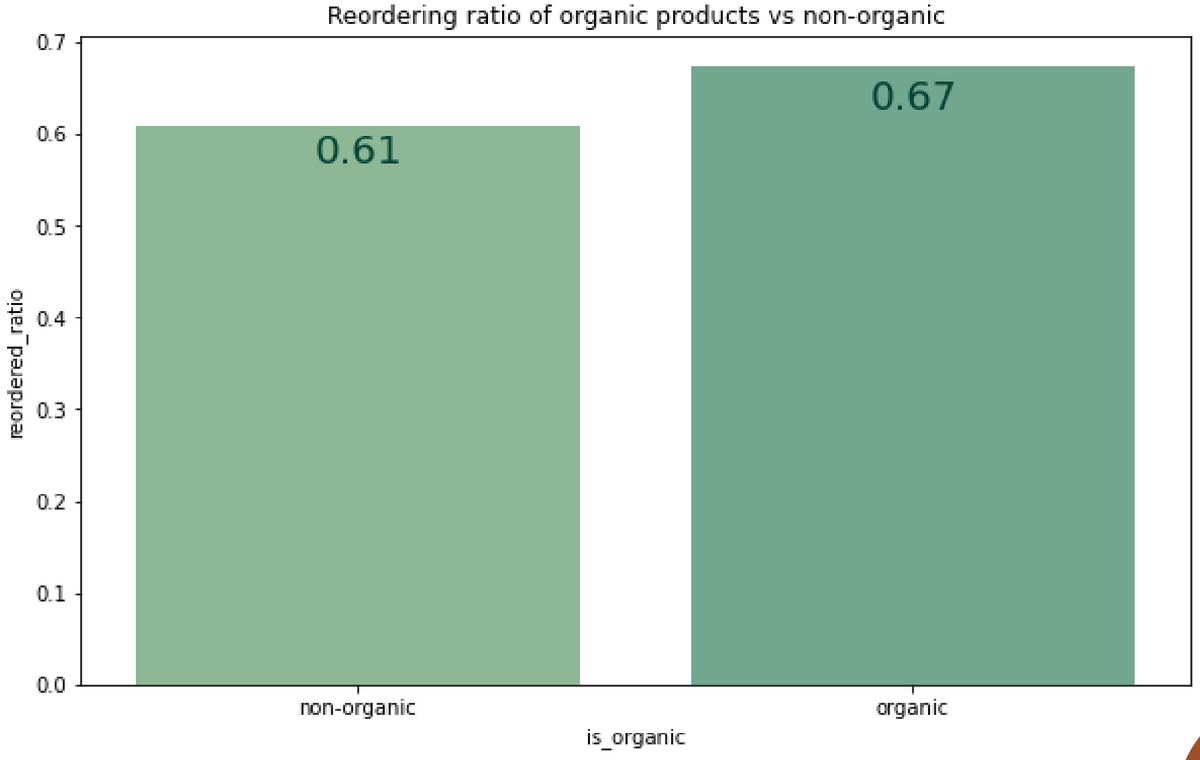
They are 31% of purchases!

Business Questions

Analyzing organic products

67% of organic purchases are users who previously experienced and bought these products.

People are loving them!



Business Questions

Which aisle or department to consider adding or offering more products to it?

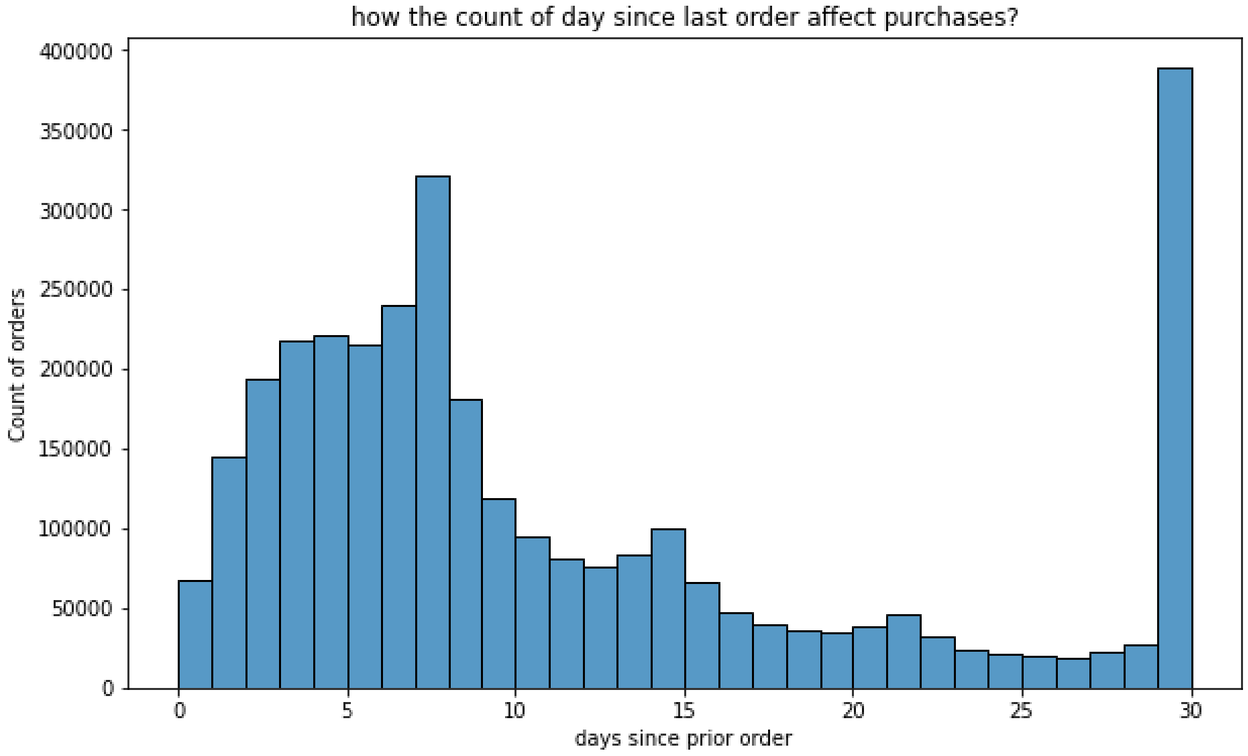
- Introducing new organic products to instacart’s retailers portfolio.
- Consider introducing new products in the departments that offer few product choices, but these products are highly being purchased.

Business Questions

How to make customers never forget instacart?

Most users make orders after a week from their last order.

Send reminders to users who haven’t ordered since 7 days from their last order.



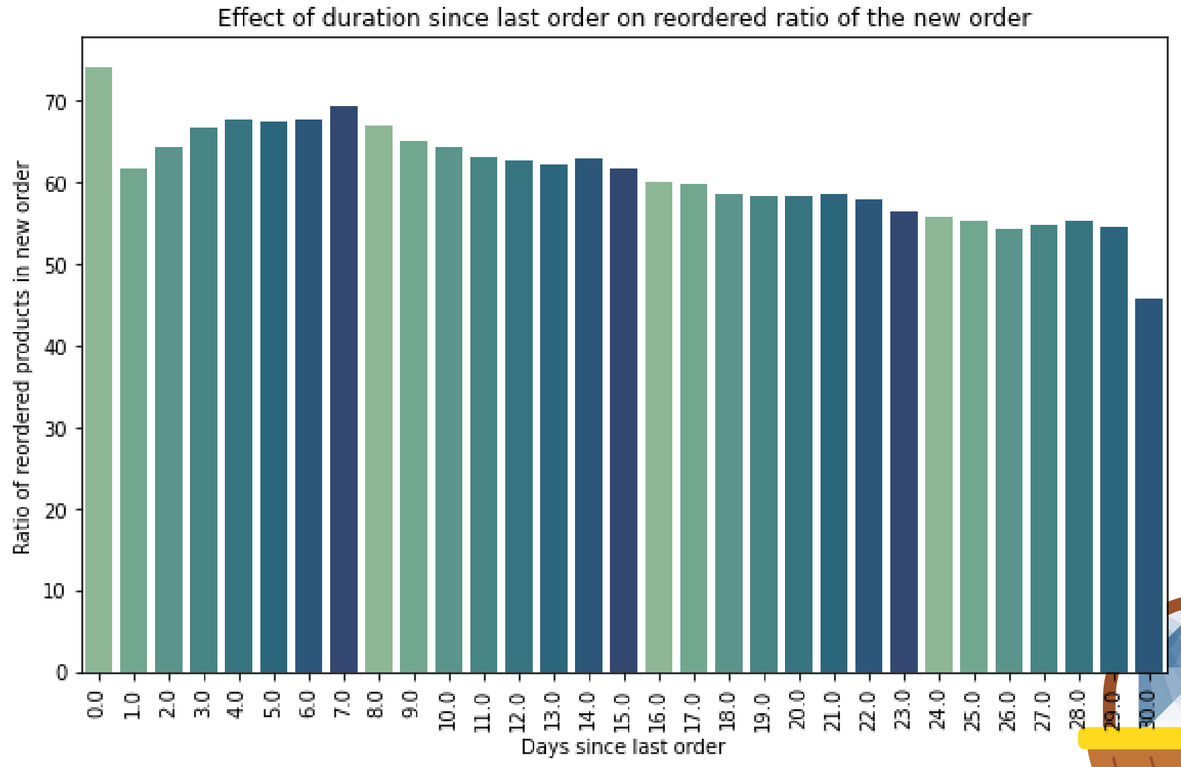
Business Questions

When to avoid recommending new products?

When it's most safe to recommend the user products they already know?

- 74 % of products bought at the same day of prev order, are reorders.
- 69% of products bought after one week from the previous order are reorders.

These are good timings to recommend products that have highly reordering ratio

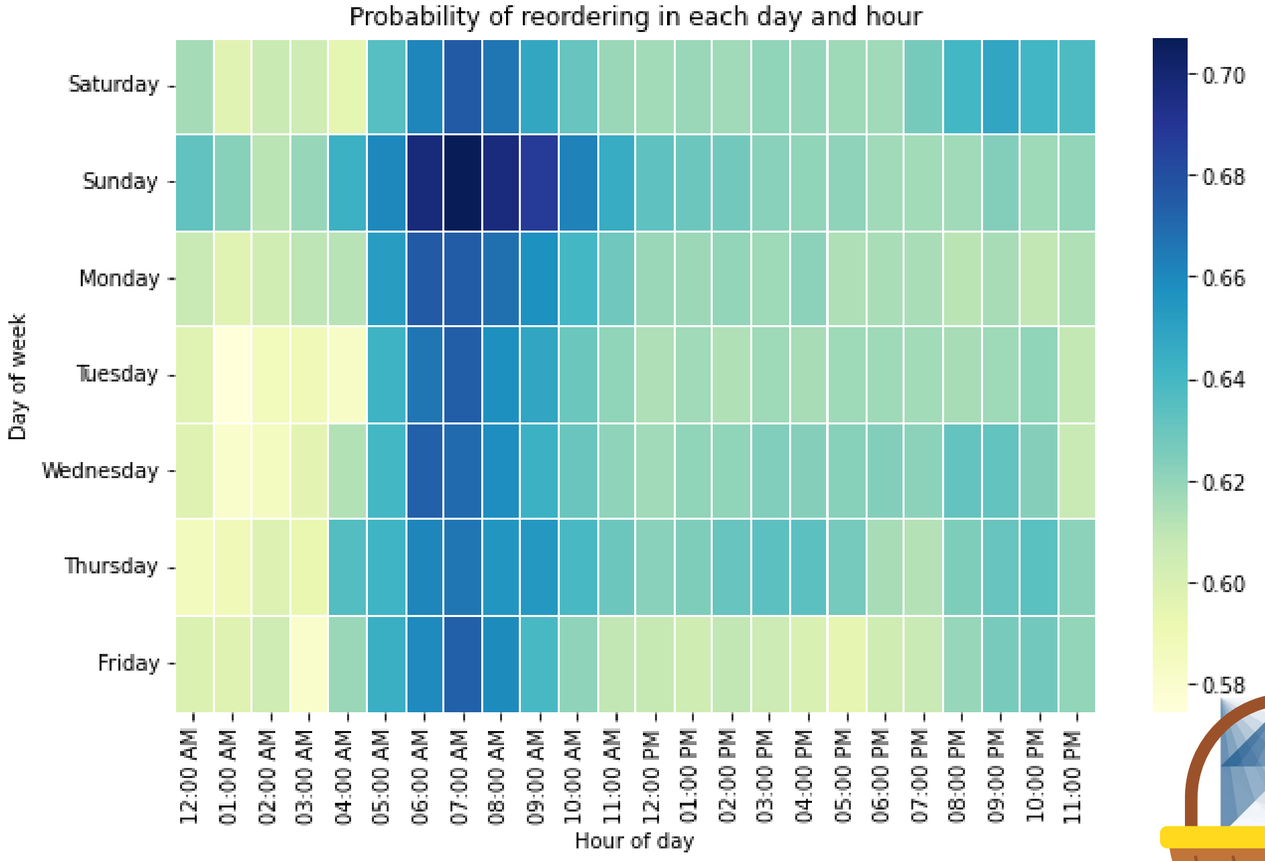


Business Questions

When to avoid recommending new products?

By more than 65%, People usually buy previously ordered products from 6:00AM to 8:00AM

Recommend previously ordered products at these hours, while avoiding recommending new products at these hours.



Business Questions

Who should we avoid recommending new products?

We found that 685 users always buy previously ordered products. starting from their 2nd order they never buy something new.

Users with **strong behavior!**

E.g. user_id 197064

Purchases of user_id 197064

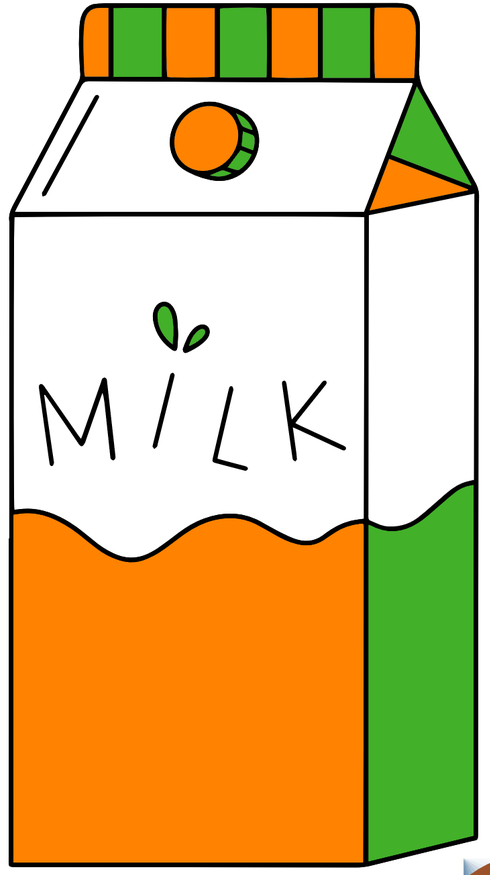

```
Orders of user 197064:
Order number2981954:
['Organic White Onions', 'Natural Spring Water', '100% Natural Spring Water', 'Beef Short Ribs']
--------------
Order number1089895:
['Organic White Onions', 'Beef Short Ribs']
--------------
Order number1892685:
['Organic White Onions']
--------------
Order number1374661:
['Organic White Onions']
--------------
Order number1234679:
['Organic White Onions']
--------------
Order number849677:
['Organic White Onions']
--------------
Order number698794:
['Organic White Onions']
--------------
Order number66061:
['Organic White Onions']
--------------
Order number1923289:
['Organic White Onions']
--------------
Order number2685353:
['Organic White Onions']
--------------
Order number2401566:
['Organic White Onions', 'Beef Short Ribs']
--------------
Order number2139480:
['Organic White Onions']
--------------
Order number858962:
['Organic White Onions']
```

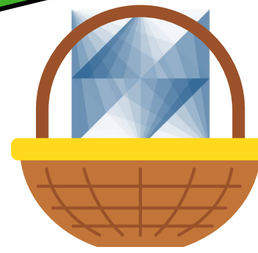
Purchases of user_id 99753

100 Orders, all contian MILK?


```
Orders of user 99753:
Count of his orders: 100
Order number2646617:
['Organic Reduced Fat Milk', 'Organic Whole Milk']
--------------
Order number208307:
['Organic Whole Milk', 'Organic Reduced Fat Milk']
--------------
Order number1849591:
['Organic Reduced Fat Milk', 'Organic Whole Milk']
--------------
Order number653264:
['Organic Reduced Fat Milk', 'Organic Whole Milk']
--------------
Order number260804:
['Organic Whole Milk', 'Organic Reduced Fat Milk']
--------------
Order number2587421:
['Organic Whole Milk', 'Organic Reduced Fat Milk']
--------------
Order number2483168:
['Organic Whole Milk', 'Organic Reduced Fat Milk']
--------------
Order number2850443:
['Organic Reduced Fat Milk', 'Organic Whole Milk']
--------------
Order number530304:
['Organic Reduced Fat Milk', 'Organic Whole Milk']
--------------
Order number3359243:
['Organic Whole Milk', 'Organic Reduced Fat Milk']
--------------
Order number2106073:
['Organic Reduced Fat Milk', 'Organic Whole Milk']
--------------
Order number2811161:
['Organic Whole Milk', 'Organic Reduced Fat Milk']
--------------
Order number798001:
['Organic Whole Milk', 'Organic Reduced Fat Milk']
```

```
Order number2337263:
['Organic Reduced Fat Milk', 'Organic Whole Milk']
--------------
Order number1458323:
['Organic Reduced Fat Milk', 'Organic Whole Milk']
--------------
Order number2578828:
['Organic Whole Milk', 'Organic Reduced Fat Milk']
--------------
Order number1627754:
['Organic Reduced Fat Milk', 'Organic Whole Milk']
--------------
Order number3087896:
['Organic Whole Milk', 'Organic Reduced Fat Milk']
--------------
Order number1293833:
['Organic Whole Milk', 'Organic Reduced Fat Milk']
--------------
Order number1917685:
['Organic Reduced Fat Milk']
--------------
Order number2950519:
['Organic Reduced Fat Milk', 'Organic Whole Milk']
--------------
Order number2924794:
['Organic Whole Milk', 'Organic Reduced Fat Milk']
--------------
Order number1469126:
['Organic Reduced Fat Milk', 'Organic Whole Milk']
--------------
Order number3018636:
['Organic Whole Milk', 'Organic Reduced Fat Milk']
--------------
Order number530239:
['Organic Reduced Fat Milk', 'Organic Whole Milk']
--------------
Order number3309651:
['Organic Reduced Fat Milk', 'Organic Whole Milk']
```





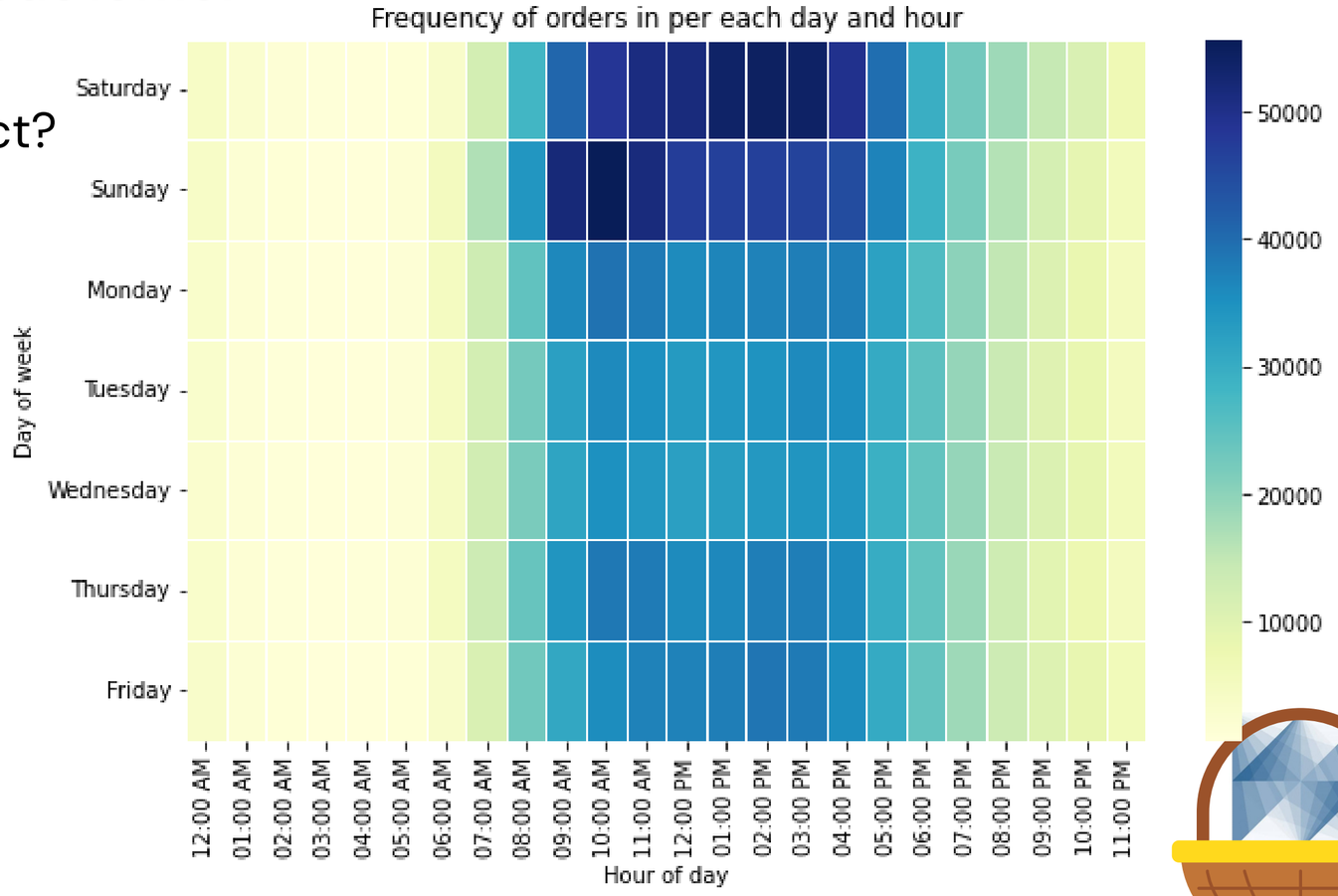
Business Questions

When it's best to recommend a customer

- new products he/she never tried?
- or a less frequently bought product?

Probability of a user buying during the afternoon of the weekends is high.

Thus can target the weekends to recommend users to try new products they haven’t bought before.



Predictive Analysis Model

Problem formulation

What do we want? We want to predict the products that will be in user's next future order.

Forming the data: We take each user with his/her previously ordered products, and form each record to be a user-product pair.

Then predict a boolean, wether this user will order or not this product in his/her future order.

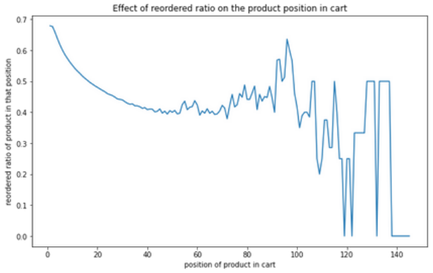
Let's extract features that relate to this user-product pair.

Feature Engineering

We wanted to extract features that strongly describe the relation between user and product

User-product features

Remember from analysis, product placed first, has high probability of being reordered.
up_average_cart_position: The average position in a user's cart of a product



up_order_rate_since_first_time: measures the degree a user like a product. It's the ratio by which a user will buy a product from the first moment he/she knew about it.

up_orders_since_last_order: measures how long the user hasn't considered buying a specific product.

Feature Engineering

User-product features

up_first_order: What was the first time a user purchased a product
up_order_rate: Percentage of user's orders that include a specific product

User features

user_orders: count of user's orders. We can rely more on data obtained from users who purchased many orders.
user_reorder_ratio: measures how this user is likely to buy something new!

user_average_basket: how many products user put on average.
user_period: how many days since the user starts shopping at instacart.
user_mean_days_since_prior: how many days passed since last order.

Feature Engineering |

Products features

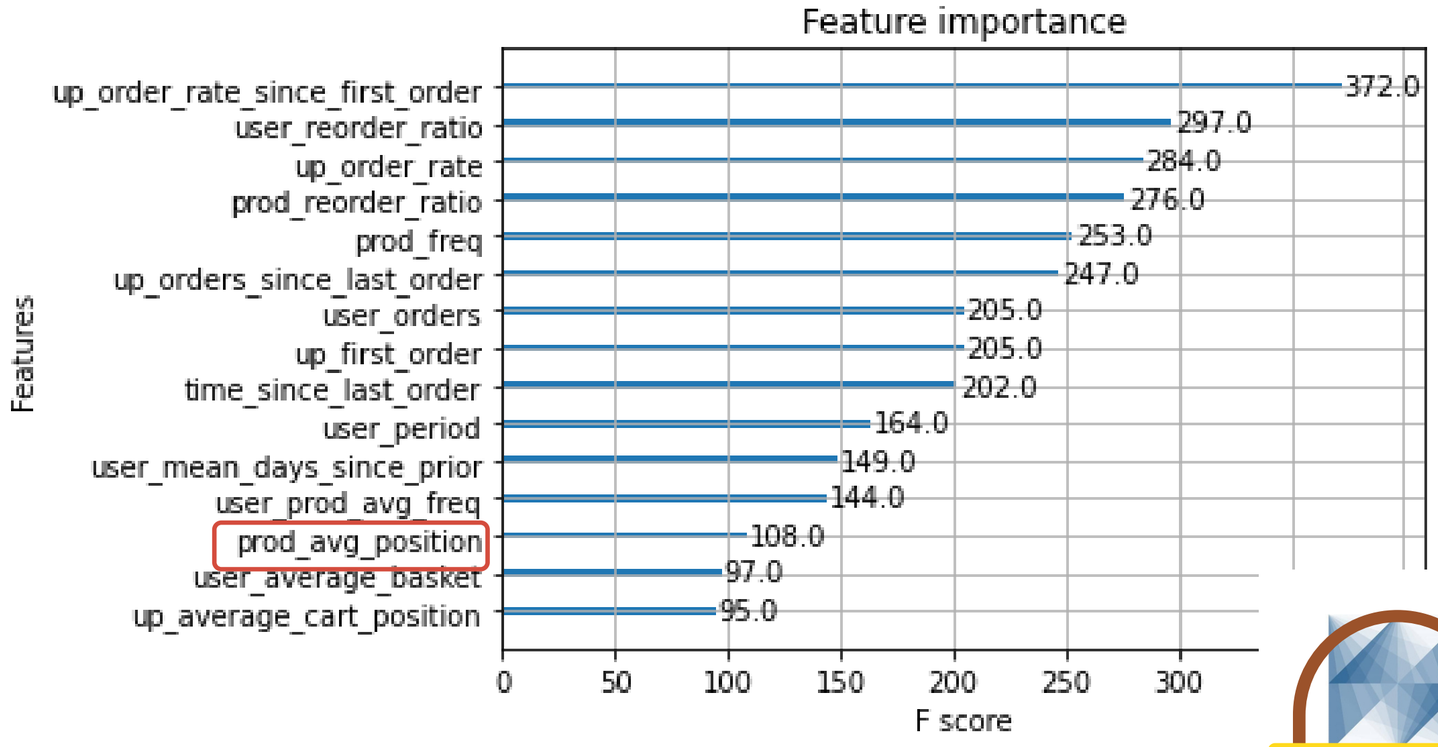
prod_freq: Total number of orders per product.
prod_reorder_ratio: Ratio that this product is being reordered from all purchases.
user_prod_avg_freq: In average how many times a product has been purchased by the users who purchased it at least once.

Feature Engineering

Product Features

prod_avg_position: The position in cart mostly repeated for the product

From analysis, we already knew that prod_avg_position is not an indicator or good feature. Since we found no product that is always placed first in cart.



Predictive Analysis | Model

Time-dependent features

Difference between

- On avg the hour the user buys this product at - The future order hour.
- On avg which day of week the user buys this product at - The future order's day of week.
- On avg the hour the product is most bought at - The future order hour.
- On avg which day of week the product is most bought at - The future order's day of week

Predictive Analysis | Model

Time-dependent features

We faced a problem calculating difference and averages between times.

Saturday takes 0
and Friday takes 6
However, they are not 6 days apart! They're 1 day apart.

How we implemented this?

Predictive Analysis | Model

Time-dependent features

Changed numbers to angles, since angles have a cyclic property.
Then averaged the angles using the following formula:
Then changed the average angle to number.

$$\bar{\alpha} = \operatorname{atan2}\left(\frac{1}{n} \cdot \sum_{j=1}^{n} \sin \alpha_j, \frac{1}{n} \cdot \sum_{j=1}^{n} \cos \alpha_j\right)$$

Predictive Analysis | Model

Classes are imbalanced

Data is very skewed to the negative class. Class distribution: 10 negative points to 1 positive point.

Train Set positive class count: 579915.0
Train Set negative class count: 5352347.0

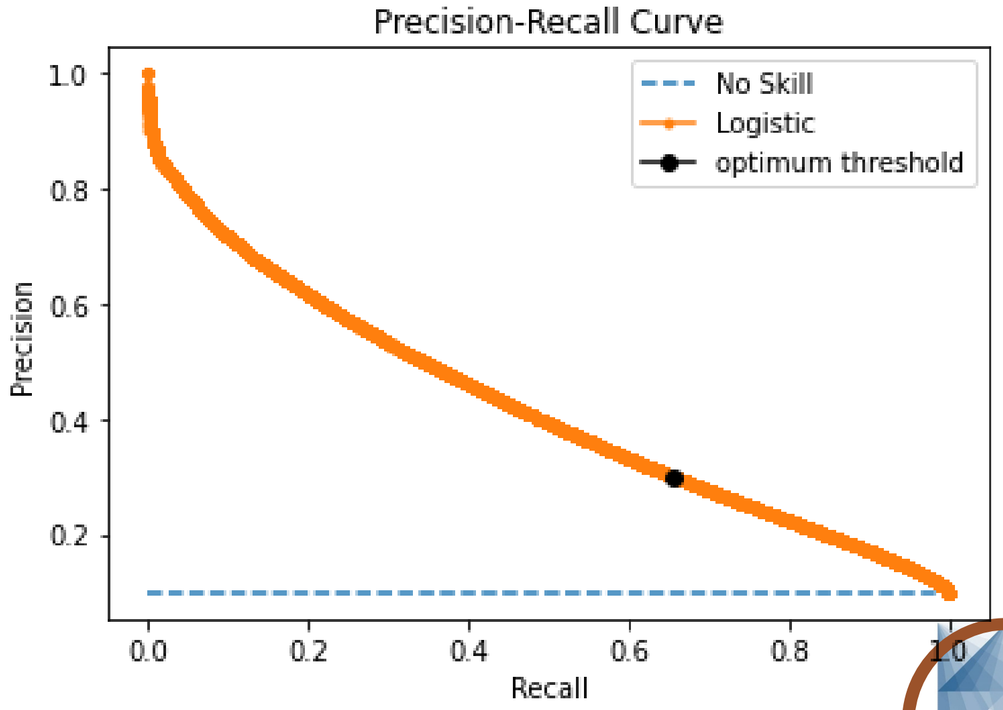
Predictive Analysis | Model

First, we've found that there's a lot of **false negatives**

So we wanted to reduce the number of products the model say user won't predict in the future while he/she will actually does.

On the other side, it's okay to allow some false positives, when the model recommends a products the user will less likely buy in his/her next order.

We changed the threshold to maximize the recall, while keeping the precision above a certain threshold 0.3



Predictive Analysis | Model

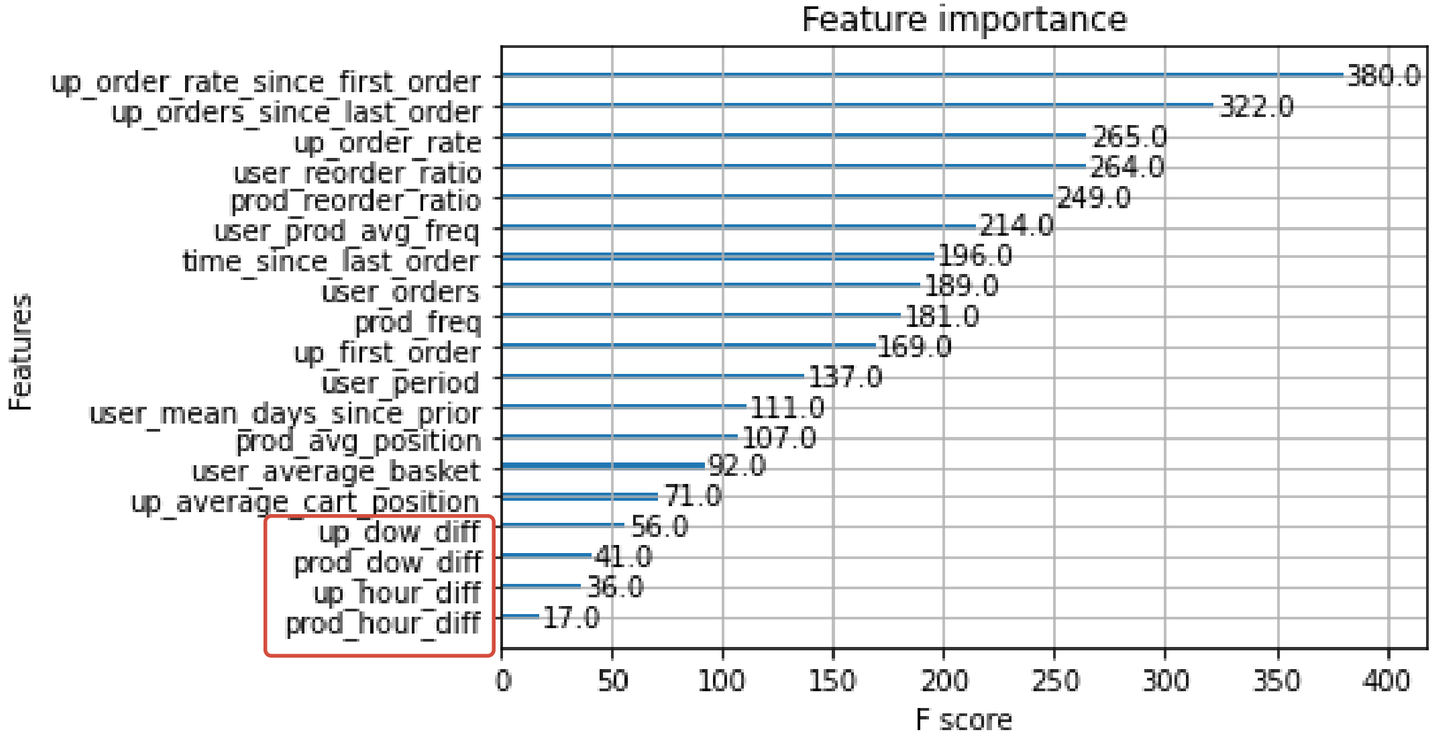
Performance

Classification report

```
-----------------CLASSIFICATION REPORT--------------------
Train positive class count:  579915.0
Train negative class count:  5352347.0
Train Set tn, fp, fn, tp: [4463290  889057  199430  380485]
Train Set report:               precision    recall  f1-score   support

         0.0       0.96      0.83      0.89   5352347
         1.0       0.30      0.66      0.41    579915

    accuracy                           0.82   5932262
   macro avg       0.63      0.74      0.65   5932262
weighted avg       0.89      0.82      0.84   5932262

Validation positive class count:  248909.0
Validation negative class count:  2293490.0
Validation Set tn, fp, fn, tp: [1912045  381445   85432  163477]
Validation Set report:               precision    recall  f1-score   support

         0.0       0.96      0.83      0.89   2293490
         1.0       0.30      0.66      0.41    248909

    accuracy                           0.82   2542399
   macro avg       0.63      0.75      0.65   2542399
weighted avg       0.89      0.82      0.84   2542399
```

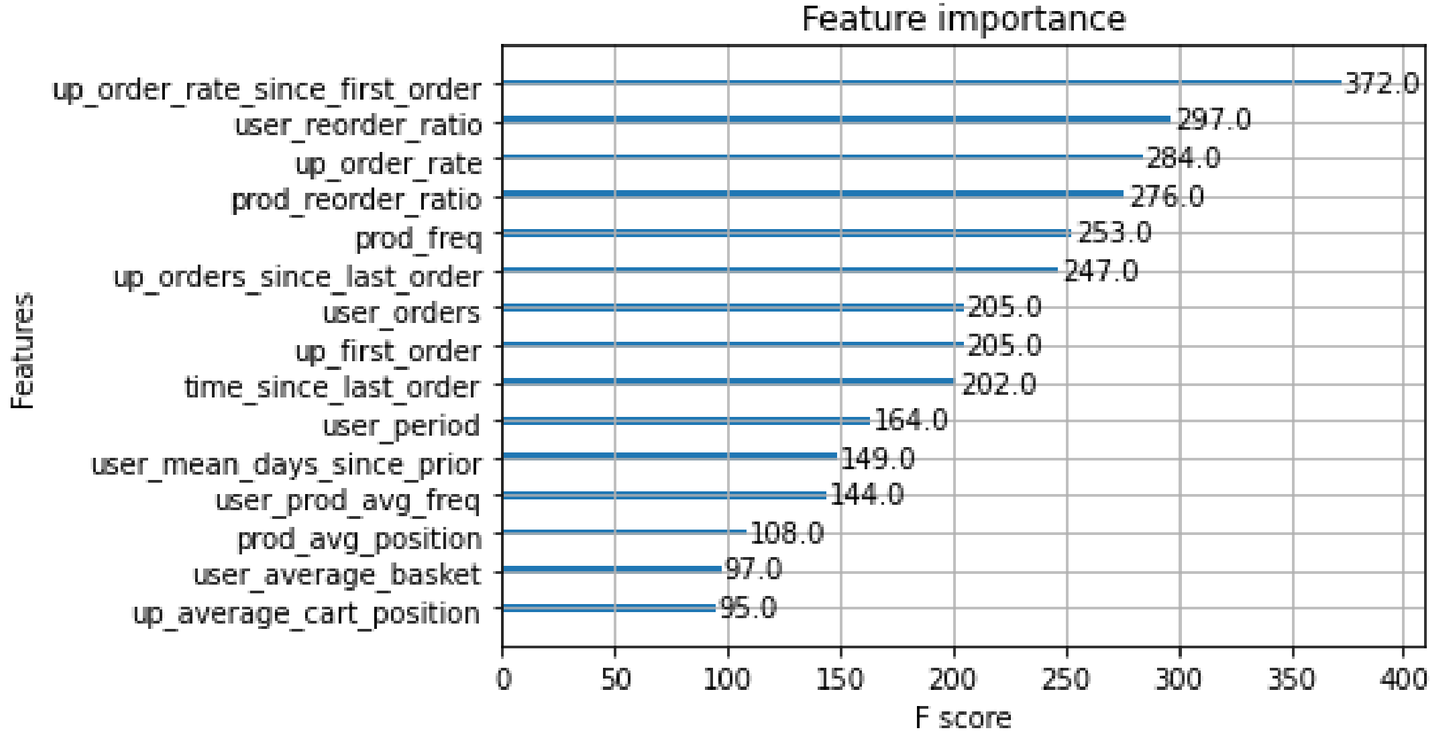
Predictive Analysis | Model

Time-dependent didn't perform well



Predictive Analysis | Model

Feature importance without time features



Another Question

When is it most beneficial to both customer and business to make free coupons and offers?

Future work

Citation

"The Instacart Online Grocery Shopping Dataset 2017", Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on 2022, May