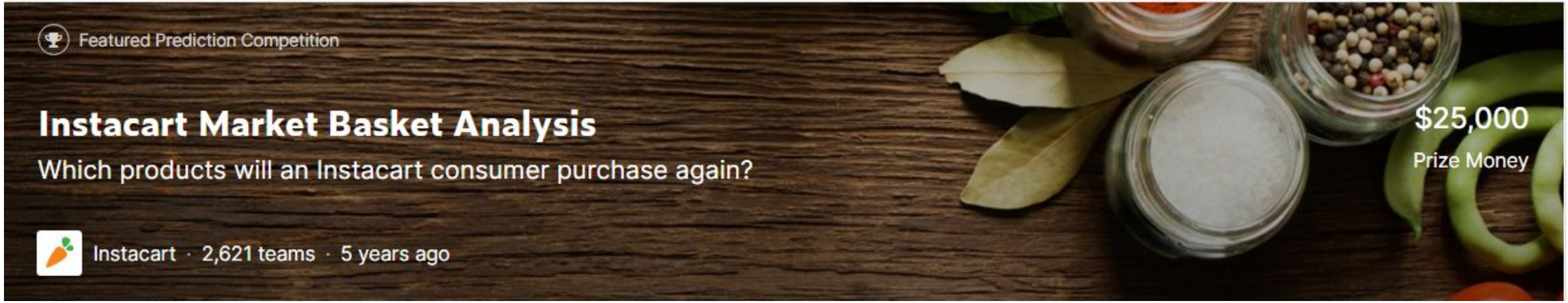
Instacart market basket analysis

Karen 2022/07/11

Outline

- Background
- Exploratory Data Analysis (EDA)
- Modeling (Xgboost)
- Apyori association analysis
- Summary

Background



Featured Prediction Competition

Instacart Market Basket Analysis

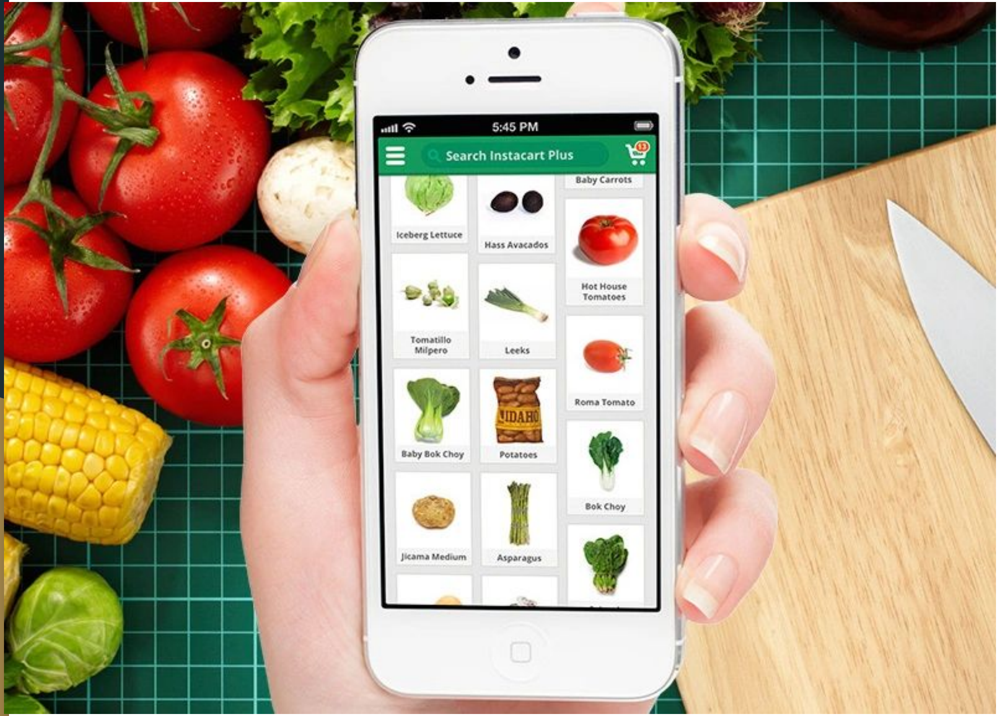
Which products will an Instacart consumer purchase again?

Instacart · 2,621 teams · 5 years ago

$25,000

Prize Money

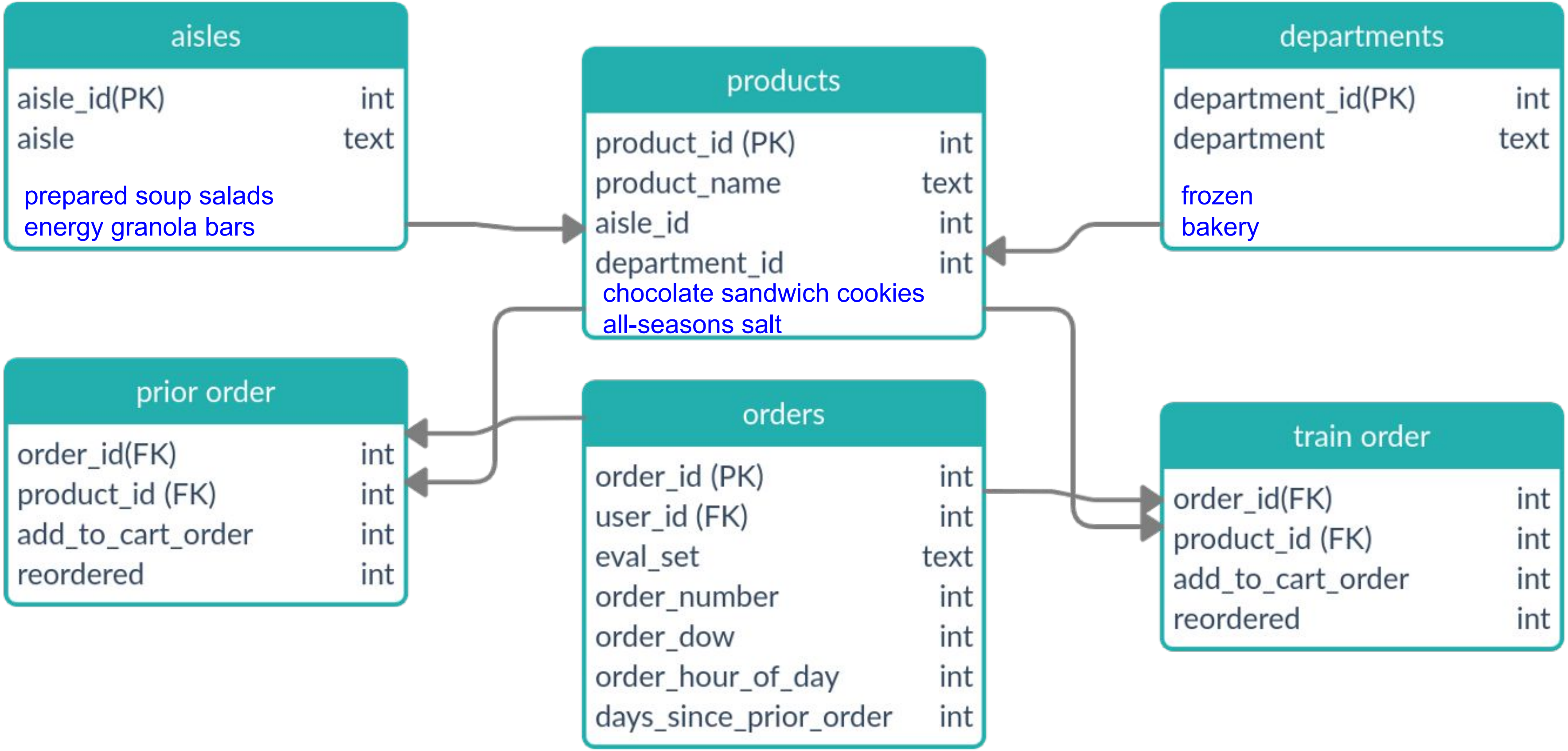
- 公開資料集 :Instacart Online Grocery Shopping Dataset 2017
- 根據之前的購買資料計算客戶再度訪問平台時可能再次購買的商品



- 成立於 2012 年
- 提供 O2O 生鮮雜貨代買代送服務
    - O2O (Online to Offline)
    - 團購服務
- 提供使用者比價資訊
- 提供購物專家購買的最佳路徑, 依照天氣、交通調整運費

EDA - Briefly review

- 來自約 20 萬名 Instacart 用戶
- 約 340 萬的訂單數量
- 將近 5 萬件商品項目
- 這些商品的類別, 分佈 21 種
- 商品擺放的位置, 約有 134 個商品陳列走道位置
- 對於每個用戶提供 4 ~ 100 個訂單資料



aisles	
aisle_id(PK)	int
aisle	text

prepared soup salads
energy granola bars

products	
product_id (PK)	int
product_name	text
aisle_id	int
department_id	int

chocolate sandwich cookies
all-seasons salt

departments	
department_id(PK)	int
department	text

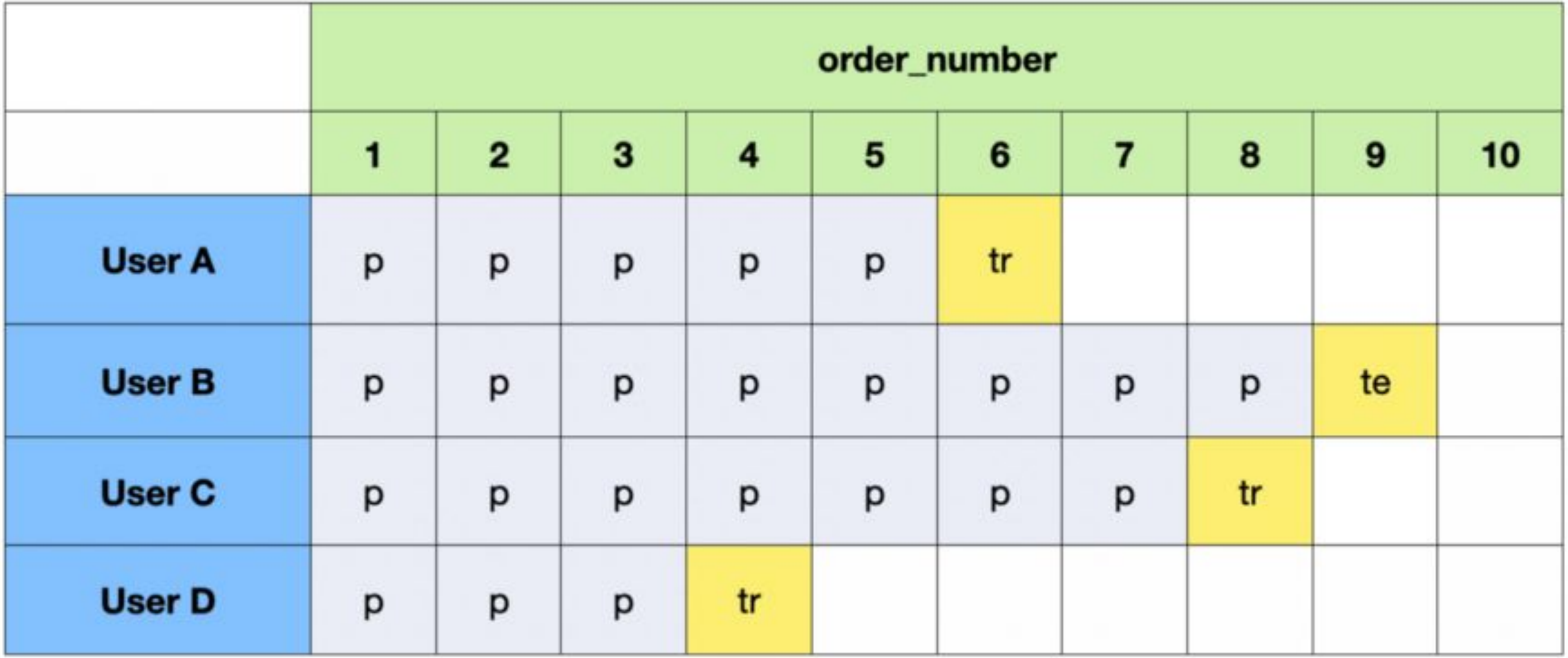
frozen
bakery

prior order	
order_id(FK)	int
product_id (FK)	int
add_to_cart_order	int
reordered	int

orders	
order_id (PK)	int
user_id (FK)	int
eval_set	text
order_number	int
order_dow	int
order_hour_of_day	int
days_since_prior_order	int

train order	
order_id(FK)	int
product_id (FK)	int
add_to_cart_order	int
reordered	int

ref:https://parmarjigar4.medium.com/kaggle-instacart-market-basket-analysis-9148608cdf18



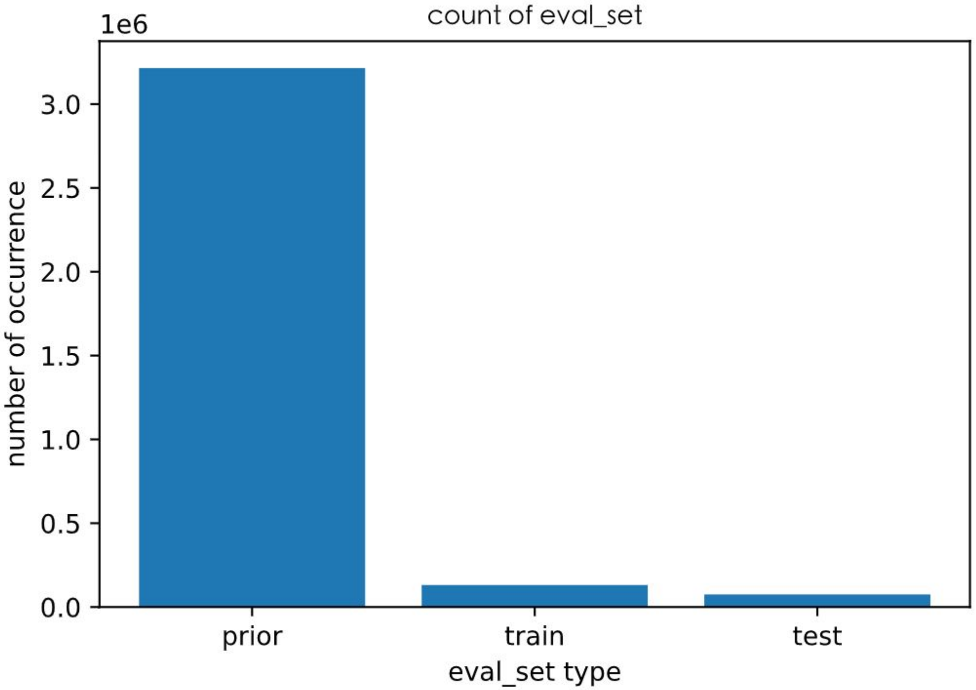
	order_number									
	1	2	3	4	5	6	7	8	9	10
User A	p	p	p	p	p	tr				
User B	p	p	p	p	p	p	p	p	te	
User C	p	p	p	p	p	p	p	tr		
User D	p	p	p	tr						

Prior (p)
Train (tr)
Test (te)

ref:https://medium.com/@PTLin0519/kaggle%E7%AB%B6%E8%B3%BD-instacart-market-basket-analysis-%E4%B8%80-%E7%AB%B6%E8%B3%BD%E7%B0%A1%E4%BB%8B%E8%88%87%E6%8E%A2%E7%B4%A2%E6%80%A7%E6%95%B8%E6%93%9A%E5%88%86%E6%9E%90-972183f2a19b

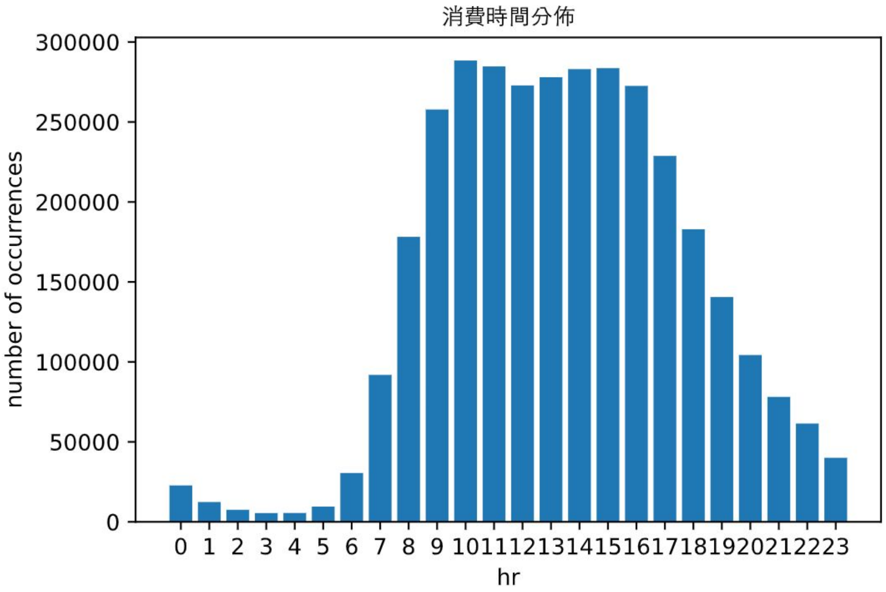
EDA- Dataset review

- 拆分三個資料集
  - Prior dataset
  - Train dataset
  - Test dataset



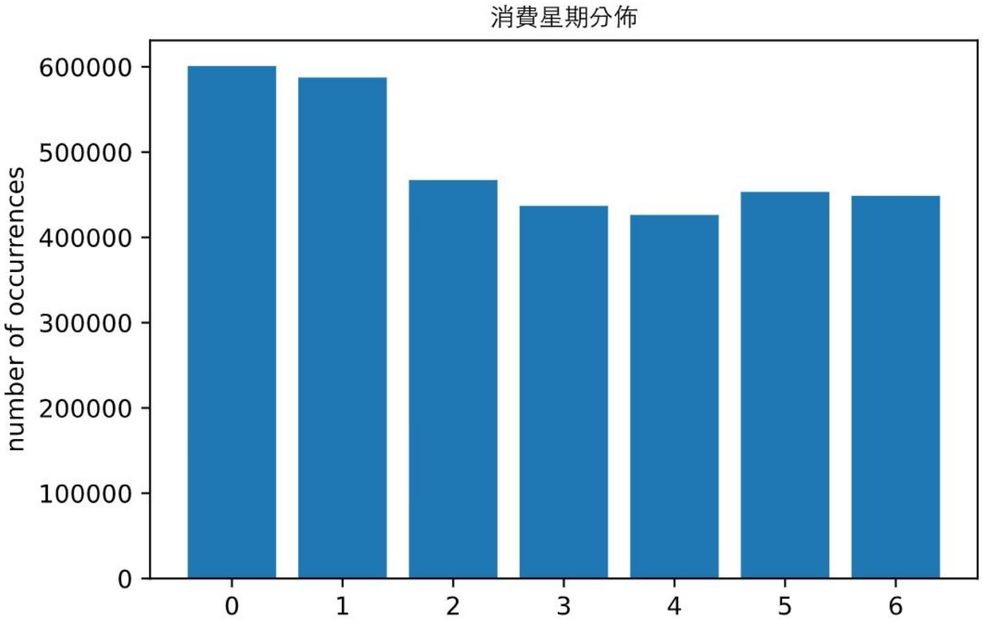
EDA- Distribution of purchase time

- 消費集中在10am~4pm

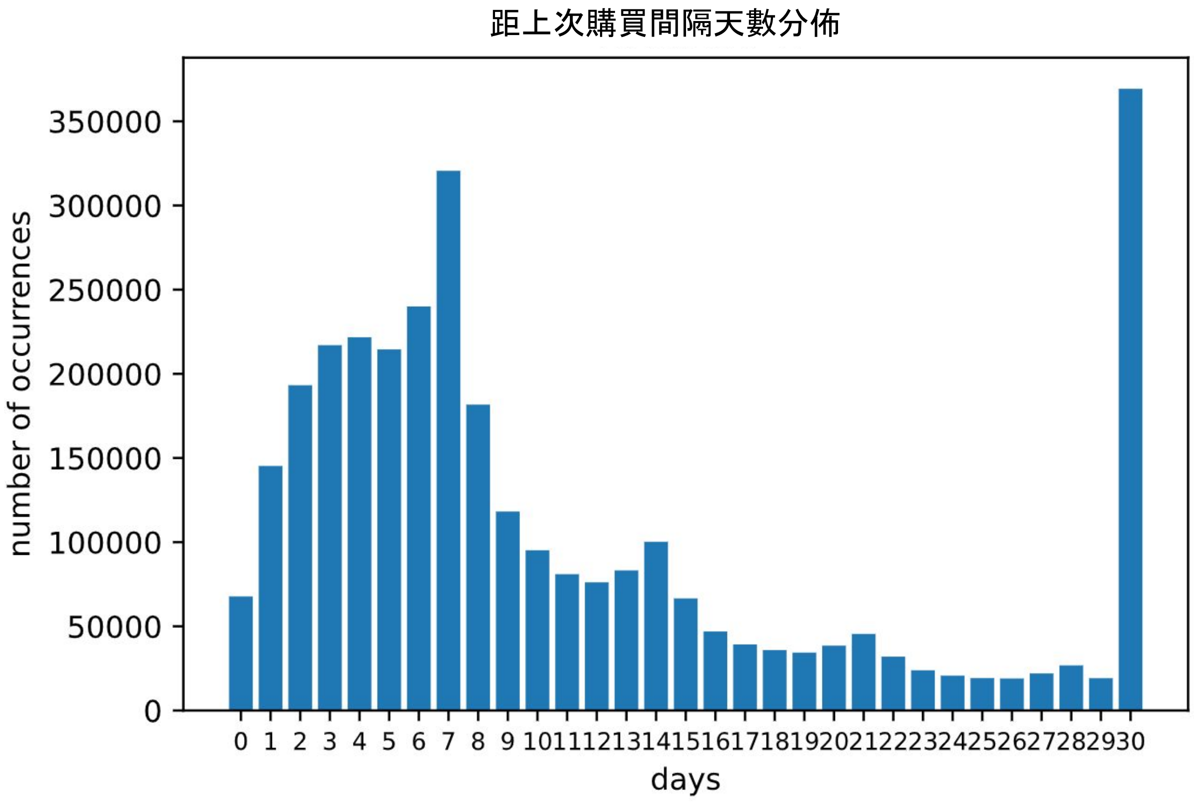


EDA- Distribution of purchase weekday

- 0為週六、1為週日以此類推
- 週六最愛買！
- 週三小週末反而消費少

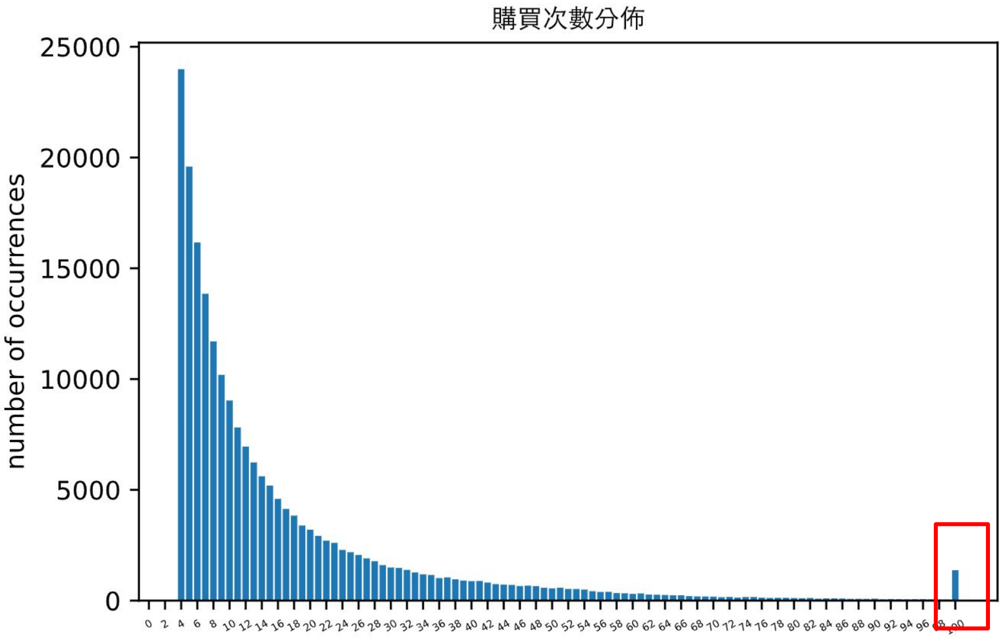


EDA- Distribution of order interval

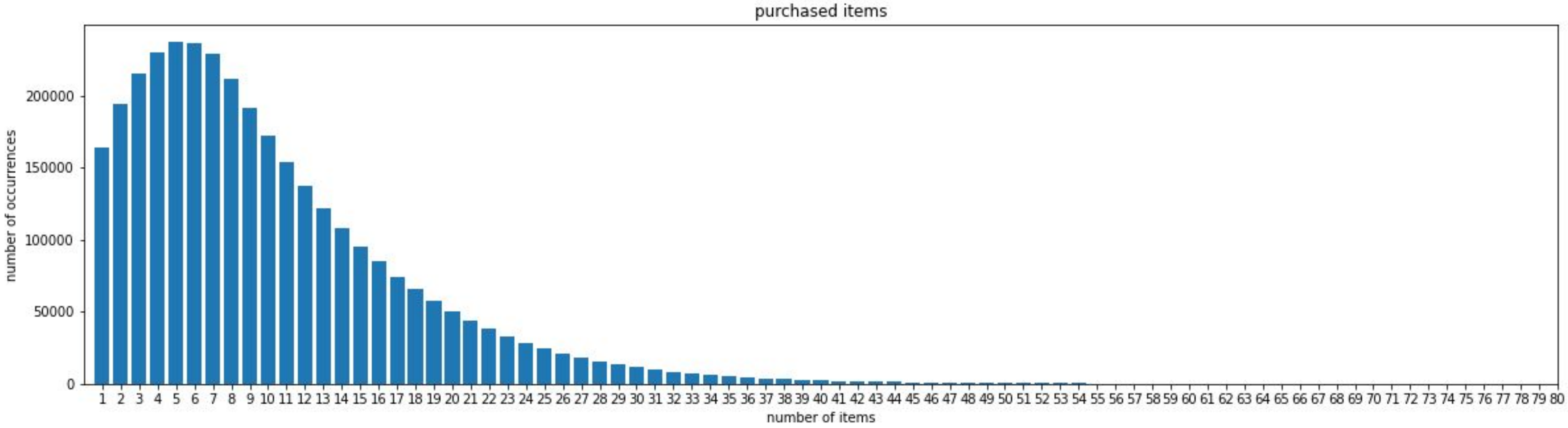


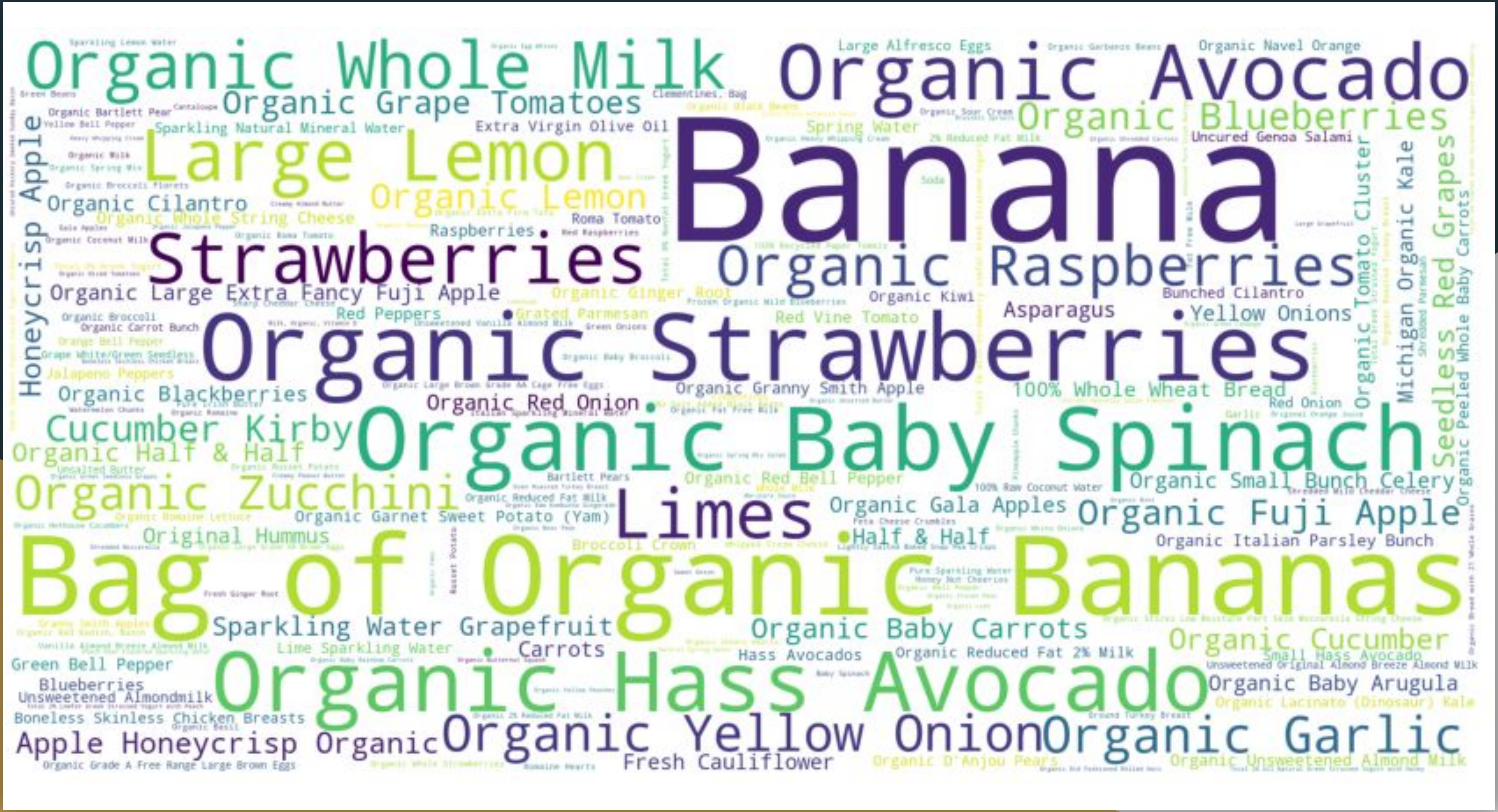
EDA- Purchase frequency

- 最低購買次數為4次
- 最高為100次



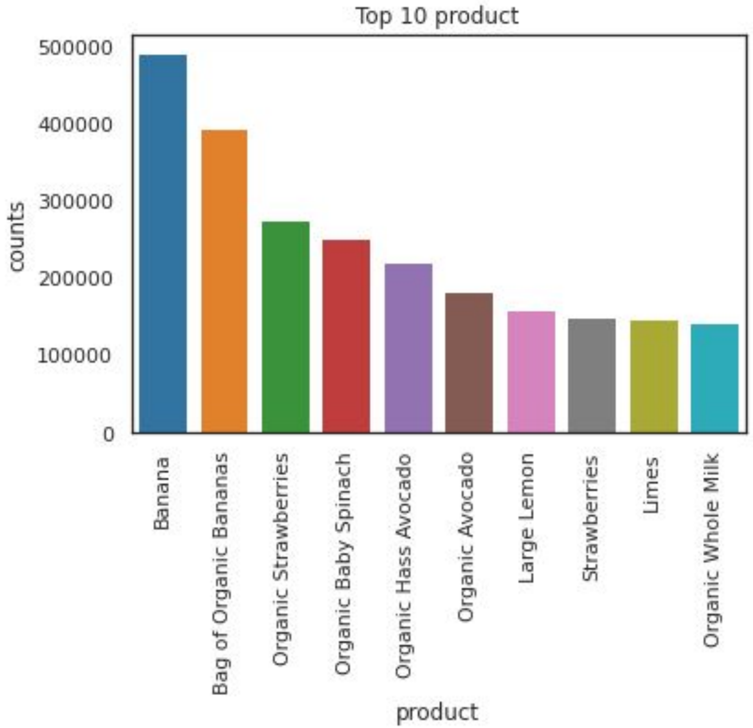
EDA- Distribution of numbers of items



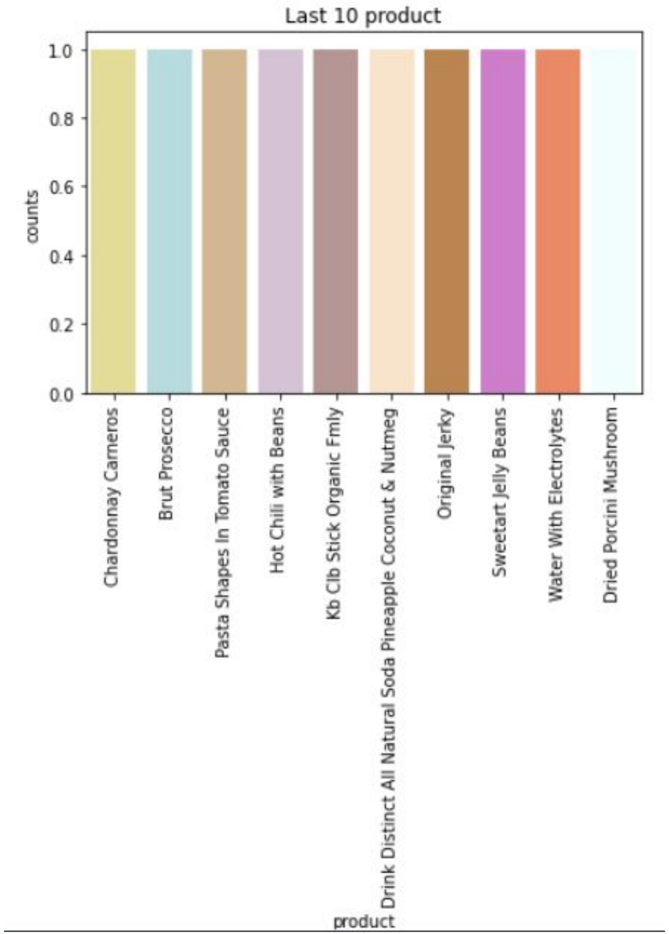

Banana
Bag of Organic Bananas
Organic Strawberries
Organic Baby Spinach
Organic Hass Avocado
Organic Avocado
Organic Whole Milk
Large Lemon
Strawberries
Organic Raspberries
Limes
Organic Zucchini
Organic Yellow Onion
Organic Garlic
Organic Blueberries
Organic Grape Tomatoes
Organic Lemon
Cucumber Kirby
Organic Fuji Apple
Organic Baby Carrots
Organic Cucumber
Sparkling Water Grapefruit
Apple Honeycrisp Organic
Honeycrisp Apple
Seedless Red Grapes
Michigan Organic Kale
Organic Tomato Cluster
Organic Cilantro
Organic Half & Half
Organic Large Extra Fancy Fuji Apple
Organic Blackberries
Organic Gala Apples
Organic Small Bunch Celery
Organic Italian Parsley Bunch
Organic Baby Arugula
Asparagus
Yellow Onions
Bunched Cilantro
100% Whole Wheat Bread
Red Onion
Organic Granny Smith Apple
Organic Red Onion
Organic Red Bell Pepper
Organic Garnet Sweet Potato (Yam)
Original Hummus
Half & Half
Broccoli Crown
Fresh Cauliflower
Carrots
Hass Avocados
Organic Reduced Fat 2% Milk
Lime Sparkling Water
Green Bell Pepper
Blueberries
Unsweetened Almondmilk
Boneless Skinless Chicken Breasts
Organic Grade A Free Range Large Brown Eggs
Organic D'Anjou Pears
Organic Unsweetened Almond Milk
Organic Lacinato (Dinosaur) Kale
Unsweetened Original Almond Breeze Almond Milk
Small Hass Avocado
Organic Peeled Whole Baby Carrots
Large Alfresco Eggs
Organic Navel Orange
Clementines, Bag
Extra Virgin Olive Oil
Sparkling Natural Mineral Water
Spring Water
Uncured Genoa Salami
2% Reduced Fat Milk
Roma Tomato
Raspberries
Organic Whole String Cheese
Organic Ginger Root
Red Peppers
Grated Parmesan
Red Vine Tomato
Organic Kiwi
Organic Broccoli
Organic Carrot Bunch
Orange Bell Pepper
Grape White/Green Seedless
Jalapeno Peppers
Organic Bartlett Pear
Yellow Bell Pepper
Organic Reduced Fat Milk
Bartlett Pears
100% Raw Coconut Water
Organic Romaine Lettuce
Unsalted Butter
Sparkling Lemon Water
Green Beans
Cantaloupe
Organic Sour Cream
Frozen Organic Wild Blueberries
Unsweetened Vanilla Almond Milk
Green Onions
Organic Baby Broccoli
Organic Black Beans
Gala Apples
Organic Coconut Milk
Organic Roma Tomato
Red Raspberries
Organic Broccoli Florets
Organic Spring Mix
Organic Milk
Shredded Parmesan
Fresh Ginger Root
Granny Smith Apples
Russet Potato
Pure Sparkling Water
Honey Nut Cheerios
Feta Cheese Crumbles
Garlic
Soda
Watermelon Chunks
Organic Romaine
Pure Irish Butter
Italian Sparkling Mineral Water
Organic Fat Free Milk
Ground Turkey Breast
Romaine Hearts
Organic Whole Strawberries
Baby Spinach
Organic Basil
Organic Baby Rainbow Carrots
Organic Butternut Squash

EDA- Purchase frequency by product

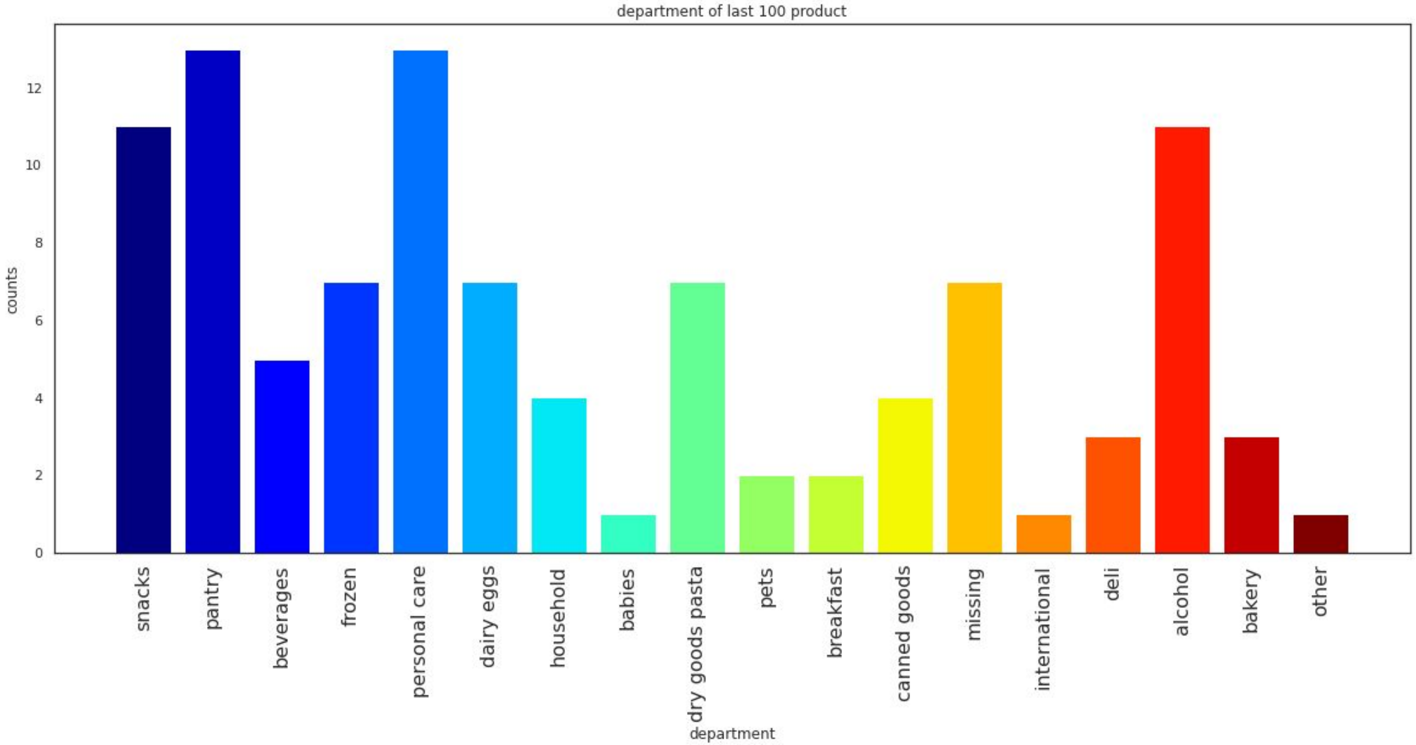
- 分析過往訂單記錄, 找出 TOP 10 明星商品
- Banana, the king of the fruit?



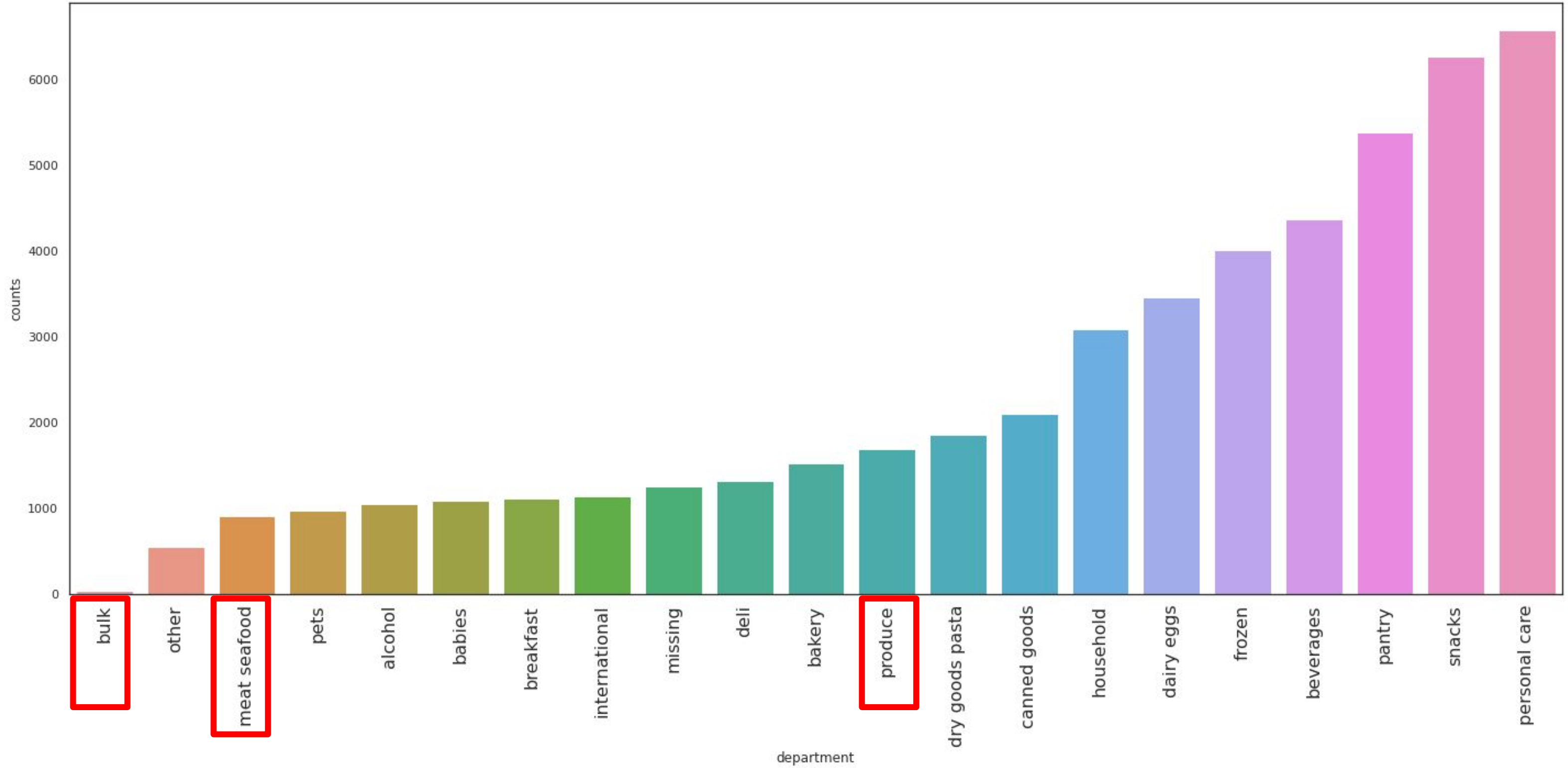
- 分析過往訂單記錄，找出 Last 10 不受歡迎商品
- 特殊食品、酒？！



EDA- Distribution of the department which product is only purchased one



EDA- Distribution of product department



Define reorder

- reordered: 1 if this product has been ordered by this user in the past, 0 otherwise

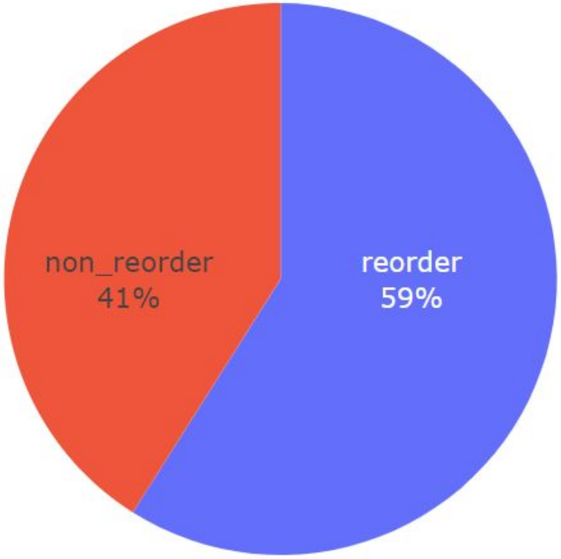
first purchase	
product_name	reordered
A	0
B	0

second purchase	
product_name	reordered
A	1
C	0
D	0

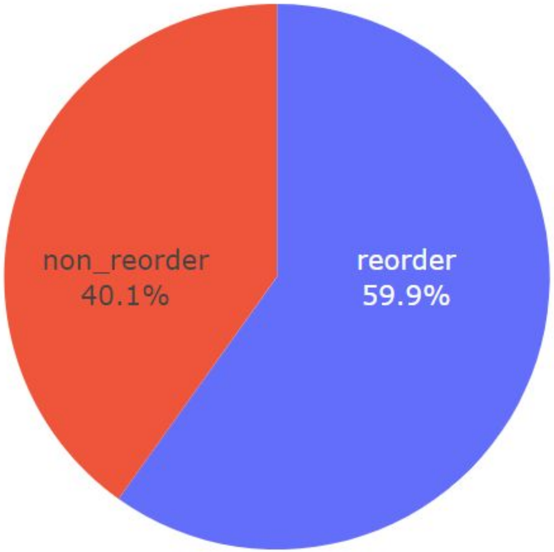
third purchase	
product_name	reordered
A	1
B	1
E	0

EDA - Reorder ratio

prior

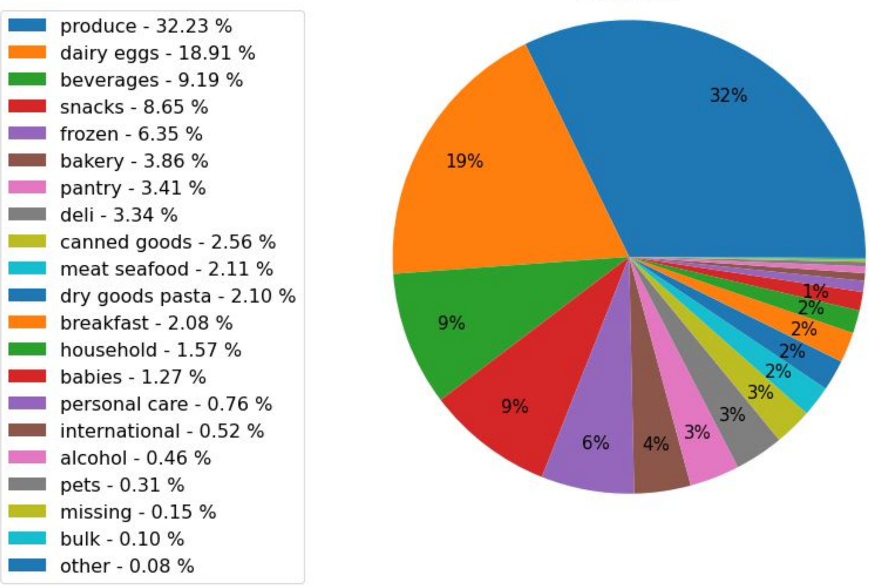


train

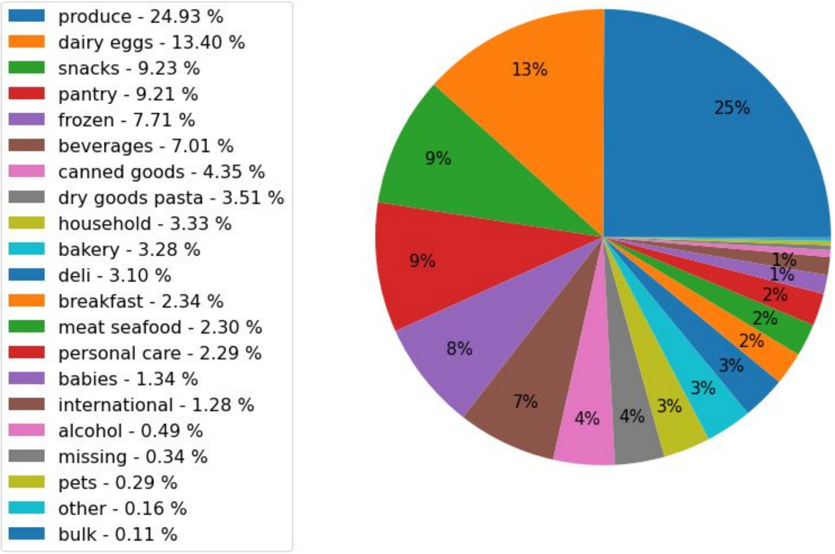


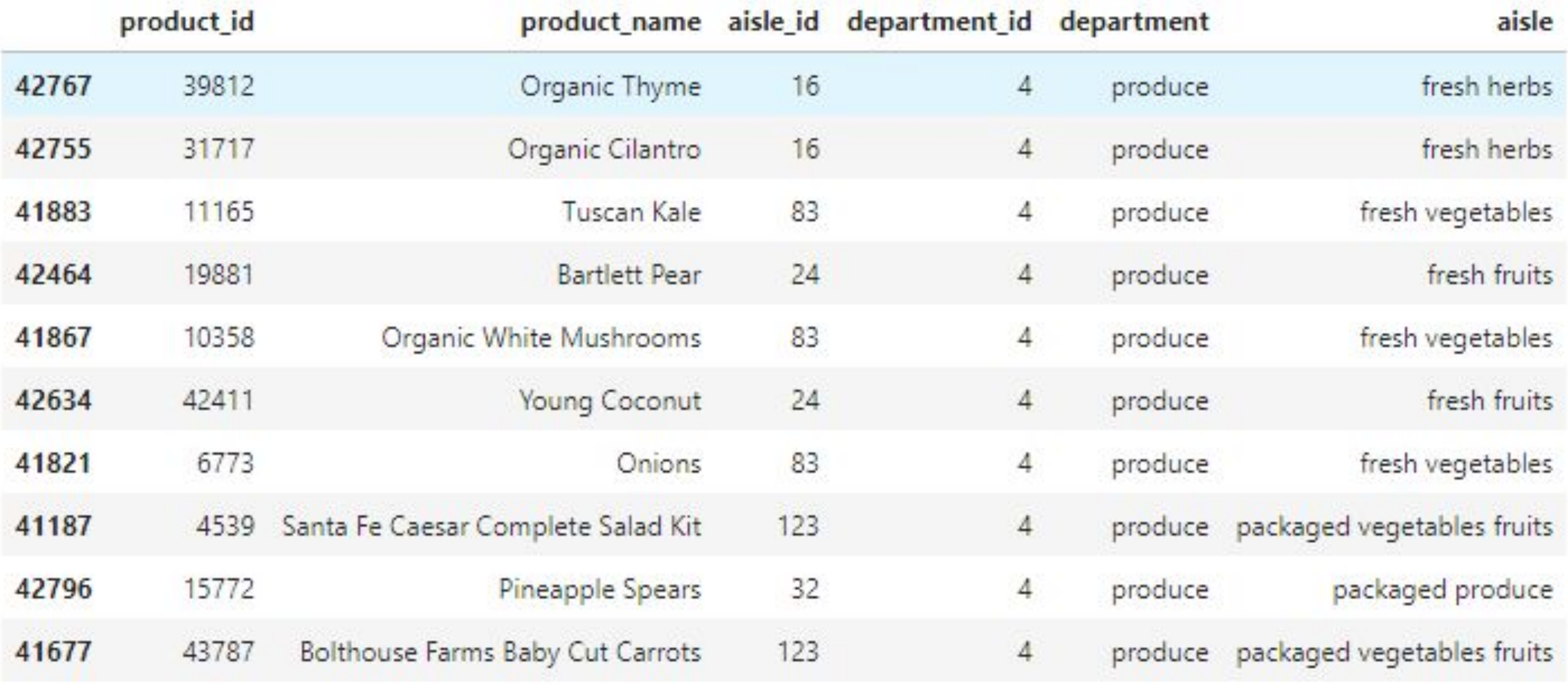
EDA - Distibution of department by reorder

reordered



non_reordered

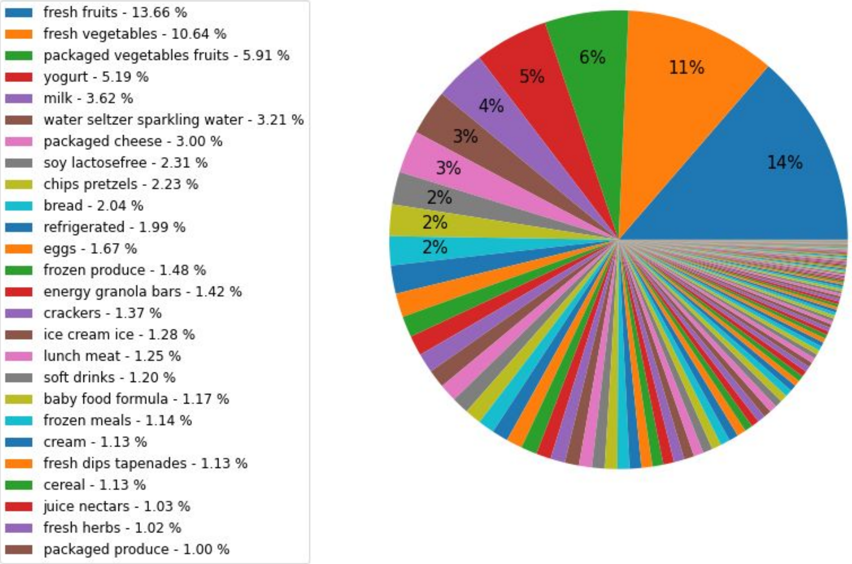




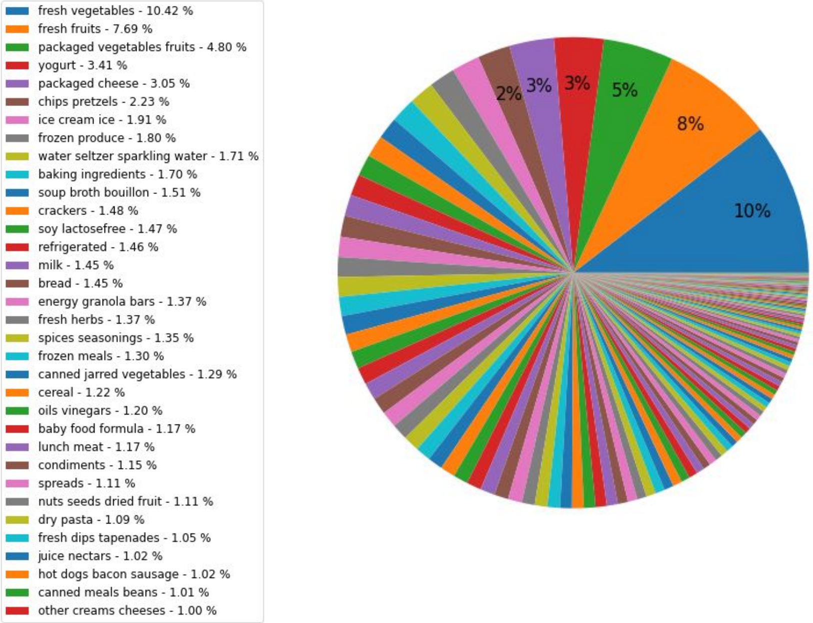
	product_id	product_name	aisle_id	department_id	department	aisle
42767	39812	Organic Thyme	16	4	produce	fresh herbs
42755	31717	Organic Cilantro	16	4	produce	fresh herbs
41883	11165	Tuscan Kale	83	4	produce	fresh vegetables
42464	19881	Bartlett Pear	24	4	produce	fresh fruits
41867	10358	Organic White Mushrooms	83	4	produce	fresh vegetables
42634	42411	Young Coconut	24	4	produce	fresh fruits
41821	6773	Onions	83	4	produce	fresh vegetables
41187	4539	Santa Fe Caesar Complete Salad Kit	123	4	produce	packaged vegetables fruits
42796	15772	Pineapple Spears	32	4	produce	packaged produce
41677	43787	Bolthouse Farms Baby Cut Carrots	123	4	produce	packaged vegetables fruits

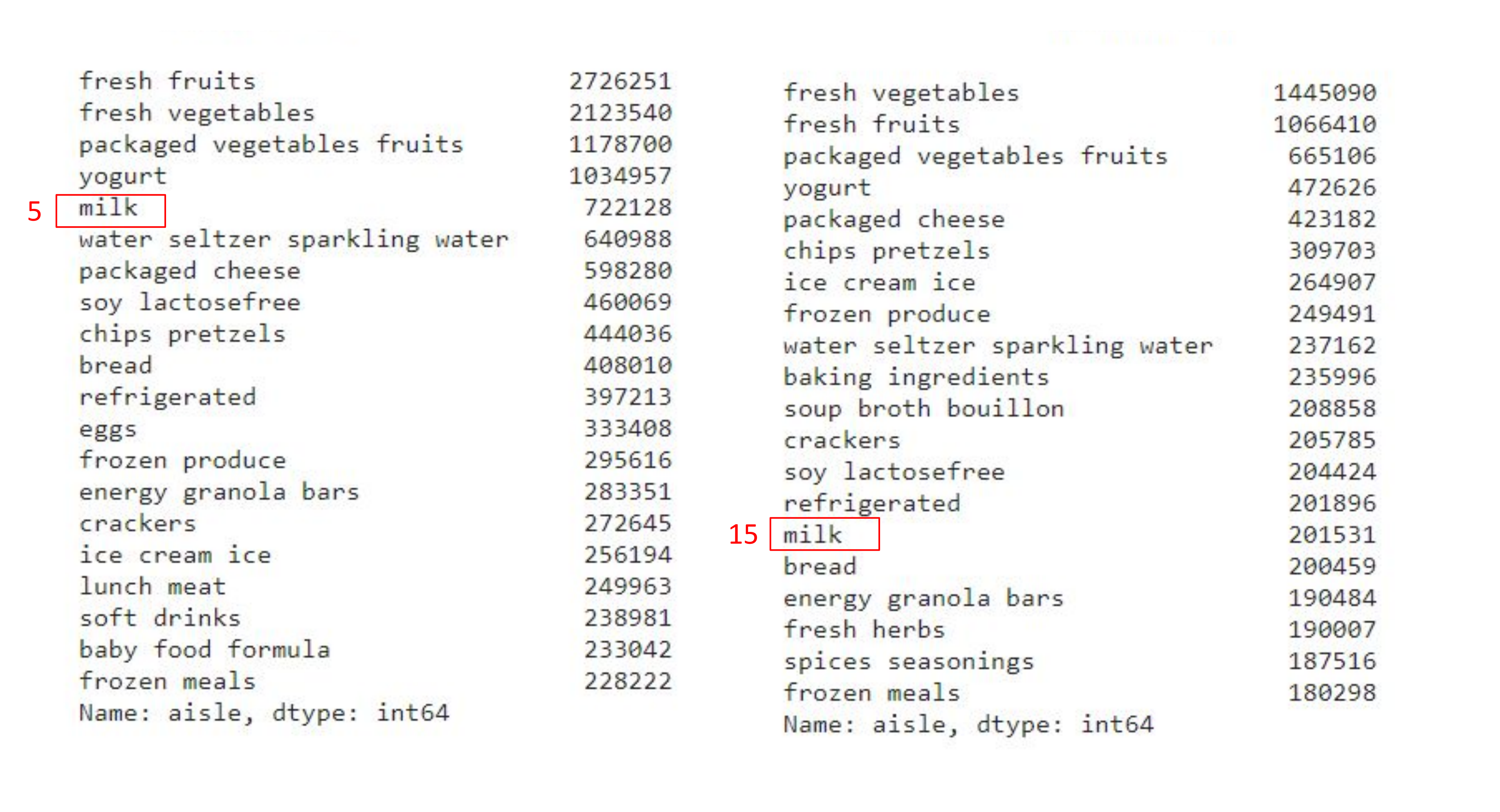
EDA - Distibution of aisles by reorder

reordered



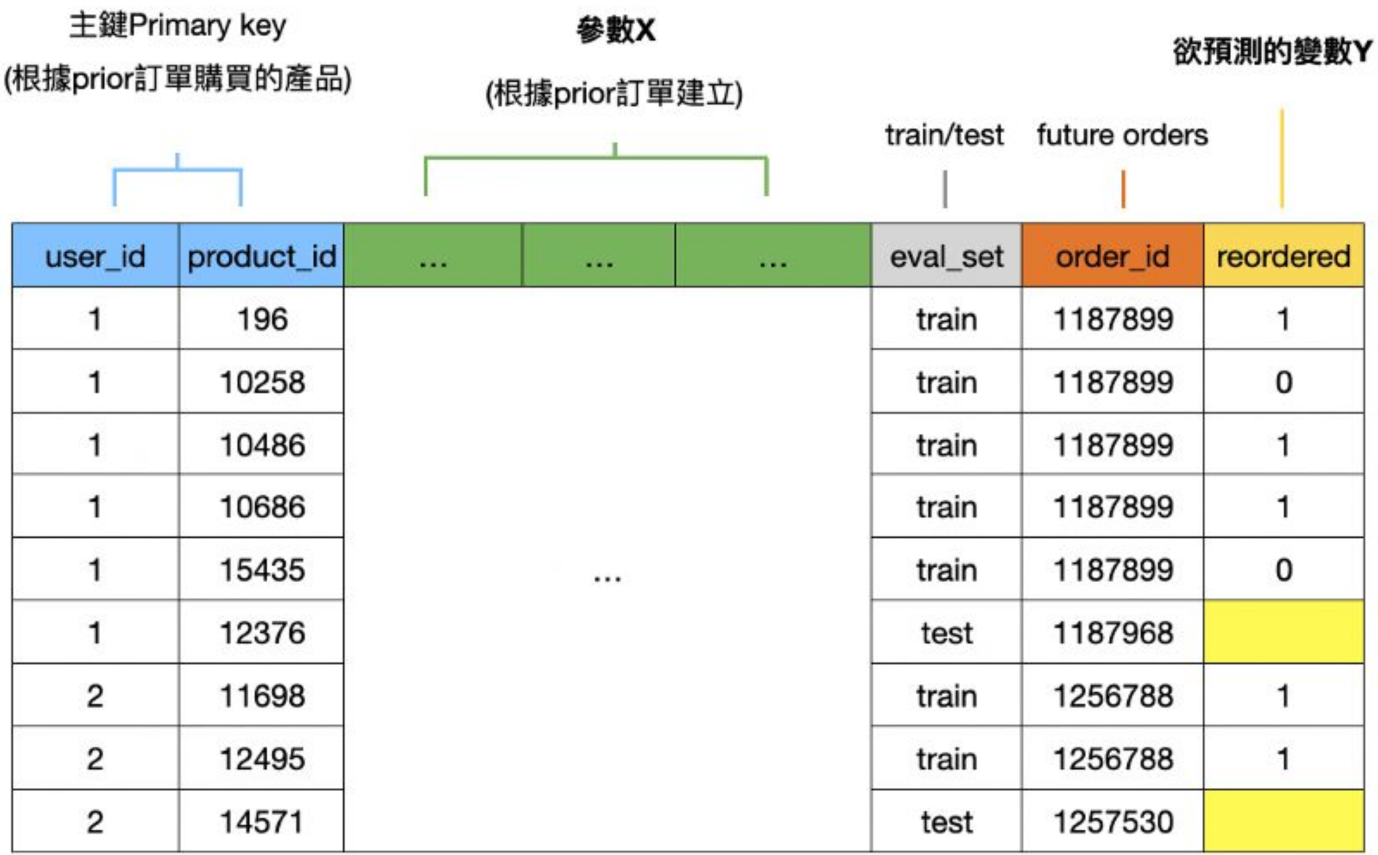
non_reordered





```
  fresh fruits                     2726251
  fresh vegetables                 2123540
  packaged vegetables fruits       1178700
  yogurt                           1034957
5 milk                              722128
  water seltzer sparkling water     640988
  packaged cheese                   598280
  soy lactosefree                   460069
  chips pretzels                    444036
  bread                             408010
  refrigerated                      397213
  eggs                              333408
  frozen produce                    295616
  energy granola bars               283351
  crackers                          272645
  ice cream ice                     256194
  lunch meat                        249963
  soft drinks                       238981
  baby food formula                 233042
  frozen meals                      228222
  Name: aisle, dtype: int64
```

```
   fresh vegetables                 1445090
   fresh fruits                     1066410
   packaged vegetables fruits        665106
   yogurt                            472626
   packaged cheese                   423182
   chips pretzels                    309703
   ice cream ice                     264907
   frozen produce                    249491
   water seltzer sparkling water     237162
   baking ingredients                235996
   soup broth bouillon               208858
   crackers                          205785
   soy lactosefree                   204424
   refrigerated                      201896
15 milk                              201531
   bread                             200459
   energy granola bars               190484
   fresh herbs                       190007
   spices seasonings                 187516
   frozen meals                      180298
   Name: aisle, dtype: int64
```



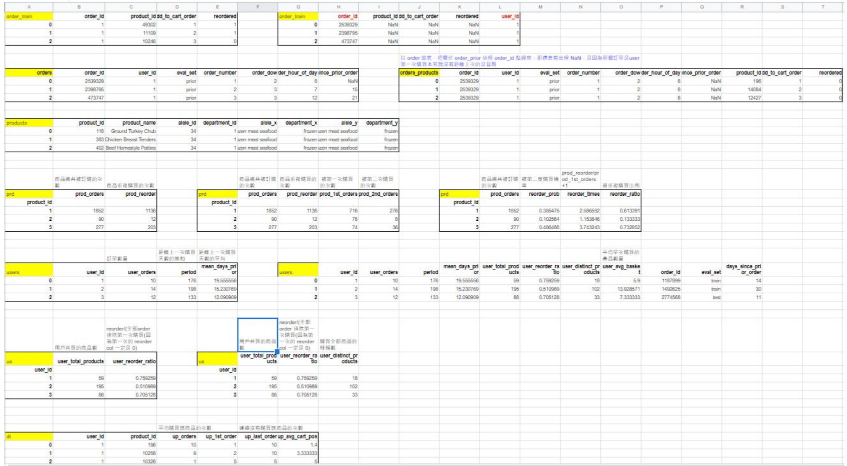
主鍵Primary key
(根據prior訂單購買的產品)
參數X
(根據prior訂單建立)
train/test
future orders
欲預測的變數Y

user_id	product_id	...	...	...	eval_set	order_id	reordered
1	196				train	1187899	1
1	10258				train	1187899	0
1	10486				train	1187899	1
1	10686				train	1187899	1
1	15435		...		train	1187899	0
1	12376				test	1187968	
2	11698				train	1256788	1
2	12495				train	1256788	1
2	14571				test	1257530	

ref:https://medium.com/@PTLin0519/kaggle%E7%AB%B6%E8%B3%BD-instacart-market-basket-analysis-%E4%B8%80-%E7%AB%B6%E8%B3%BD%E7%B0%A1%E4%BB%8B%E8%88%87%E6%8E%A2%E7%B4%A2%E6%80%A7%E6%95%B8%E6%93%9A%E5%88%86%E6%9E%90-972183f2a19b

Feature Engineering

- 商品面
    - 被購買的次數
    - 被重複購買比例
    - 被第一次購買的次數
    - 被第二次購買的次數
- 客戶面
    - 距離上一次購買天數的總和和平均
    - 平均單次購買的產品數量
    - 平均將該商品放入購物車的順序
    - 平均購買該商品的次數
    - 連續沒有購買該商品的次數



order_train	order_id	product_id	dd_to_cart_order	reordered
0	1	49302	1	1
1	1	11109	2	1
2	1	10246	3	0

order_train	order_id	product_id	dd_to_cart_order	reordered	user_id
0	2539329	NaN	NaN	NaN	1
1	2398795	NaN	NaN	NaN	1
2	473747	NaN	NaN	NaN	1

orders	order_id	user_id	eval_set	order_number	order_dow	der_hour_of_day	ince_prior_order
0	2539329	1	prior	1	2	8	NaN
1	2398795	1	prior	2	3	7	15
2	473747	1	prior	3	3	12	21

以 order 當底，把關於 order_prior 依照 order_id 黏過來，那邊會有出現 NaN，是因為那個訂單是user 第一次購買本來就沒有距離上次的日期

orders_products	order_id	user_id	eval_set	order_number	order_dow	der_hour_of_day	ince_prior_order	product_id	dd_to_cart_order	reordered
0	2539329	1	prior	1	2	8	NaN	196	1	0
1	2539329	1	prior	1	2	8	NaN	14084	2	0
2	2539329	1	prior	1	2	8	NaN	12427	3	0

products	product_id	product_name	aisle_id	department_id	aisle_x	department_x	aisle_y	department_y
0	118	Ground Turkey Chub	34	1	uzen meat seafood	frozen	uzen meat seafood	frozen
1	383	Chicken Breast Tenders	34	1	uzen meat seafood	frozen	uzen meat seafood	frozen
2	402	Beef Homestyle Patties	34	1	uzen meat seafood	frozen	uzen meat seafood	frozen

	商品總共被訂購的次數	商品重複購買的次數
prd	prod_orders	prod_reorder
product_id		
1	1852	1136
2	90	12
3	277	203

	商品總共被訂購的次數	商品重複購買的次數	被第一次購買的次數	被第二次購買的次數
prd	prod_orders	prod_reorder	prod_1st_orders	prod_2nd_orders
product_id				
1	1852	1136	716	276
2	90	12	78	8
3	277	203	74	36

	商品總共被訂購的次數	被第二度購買機率	prod_reorder/prod_1st_orders +1	被重複購買比例
prd	prod_orders	reorder_prob	reorder_times	reorder_ratio
product_id				
1	1852	0.385475	2.586592	0.613391
2	90	0.102564	1.153846	0.133333
3	277	0.486486	3.743243	0.732852

		訂單數量	距離上一次購買天數的總和	距離上一次購買天數的平均
users	user_id	user_orders	period	mean_days_prior
0	1	10	176	19.555556
1	2	14	196	15.230769
2	3	12	133	12.090909

								平均單次購買的產品數量			
users	user_id	user_orders	period	mean_days_prior	user_total_products	user_reorder_ratio	user_distinct_products	user_avg_basket	order_id	eval_set	days_since_prior_order
0	1	10	176	19.555556	59	0.759259	18	5.9	1187899	train	14
1	2	14	196	15.230769	195	0.510989	102	13.928571	1492625	train	30
2	3	12	133	12.090909	88	0.705128	33	7.333333	2774568	test	11

	用戶共買的商品數	reorder/(全部order 排除第一次購買(因為第一次的 reorder col 一定是 0)
us	user_total_products	user_reorder_ratio
user_id		
1	59	0.759259
2	195	0.510989
3	88	0.705128

	用戶共買的商品數	reorder/(全部order 排除第一次購買(因為第一次的 reorder col 一定是 0)	購買全部商品的種類數
us	user_total_products	user_reorder_ratio	user_distinct_products
user_id			
1	59	0.759259	18
2	195	0.510989	102
3	88	0.705128	33

			平均購買該商品的次數		連續沒有購買該商品的次數	
dt	user_id	product_id	up_orders	up_1st_order	up_last_order	up_avg_cart_pos
0	1	196	10	1	10	1.4
1	1	10258	9	2	10	3.333333
2	1	10326	1	5	5	5

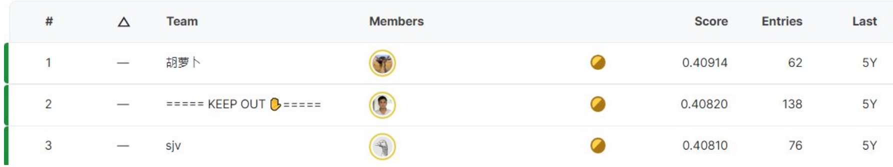
Modeling

XGBoost

- 高準確度
- 適合作為 baseline model
- 準確度較不受無效特徵干擾

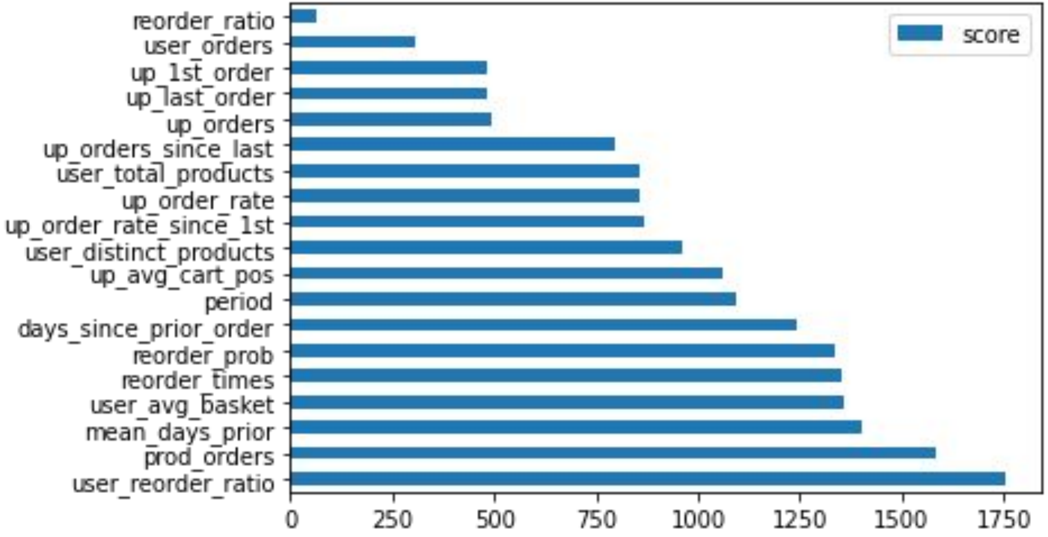
$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

- Baseline model score : 0.37625
- Grid Search score : 0.37968 (38.3%)



#	△	Team	Members		Score	Entries	Last
1	—	胡萝卜		●	0.40914	62	5Y
2	—	===== KEEP OUT ✋=====		●	0.40820	138	5Y
3	—	sjv		●	0.40810	76	5Y

Features importance analysis



- 重複購買率 (user_reorder_ratio)
- 被購買的次數(prod_order)
- 距離上一次購買天數的平均 (mean_days_prior)
- 平均單次購買的產品數量 (user_avg_basket)

Apyori association analysis

- 關聯分析: 找尋資料彼此之間的關聯, 透過兩種主要的方式來進行分析
  - 頻繁項集: 經常一起出現的物品集合
  - 關聯規則: 表達數據之間的可能存在很強關聯性

- 支持度(Support) : 表示為 item-set 在整個 AllSamples 中出現的頻率

Support(X) = number(X) / number(AllSamples)

- 信心度(Confidence) : 表示當事件X發生的情況下, 同時會發生Y的可能性

Confidence(X→Y) = P(Y|X) , = P(X∩Y) / P(X)

交易編號	商品
0	豆漿、萵苣
1	萵苣、尿布、葡萄酒、甜菜
2	豆漿、尿布、葡萄酒、橙汁
3	萵苣、豆漿、尿布、葡萄酒
4	萵苣、豆漿、尿布、橙汁

Apyori Association Analysis Case

prior

test

Organic Strawberries (275577)

Organic Hass Avocado (42333)

15%!

Support: 0.028
Confidence: 0.45

Boneless Skinless Chicken Breasts (52369)

Organic Baby Spinach (8823)

16%!

Support: 0.006
Confidence: 0.278

Summary

- 藉由過去的訂單數據，利用 model 能預測下次顧客是否再度購買商品
- 觀察 organic 在資料集中，可能成為重要的特徵
- 通過關聯分析發現訂單中的商品有交互關係，例如購買A產品經常購買B

Future works

- data augmentation: 加入用戶最近的 3 - 5 個訂單提高訓練數據量
- product feature engineering: organic feature, alternative item
- none prediction model: 有可能用戶下一次訂單中不回購任何商品
- other models: RNN, CNN, LGBMClassifier
- design new training flow: 由於產品間有交互作用，可以設計新的訓練流程，不只是將每一個產品當成獨立的分類問題

Thank you