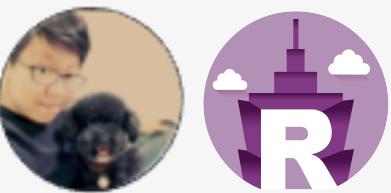




COSCUP 2017  
2017.0805



# 什麼是 kaggle ?





- Website :  
<https://www.kaggle.com>
- 資料科學和機器學習競賽平台
- 目前已累積超過50萬名、遍布超過194個國家的註冊用戶
- 涵蓋電腦科學、電腦視覺、生物、醫藥

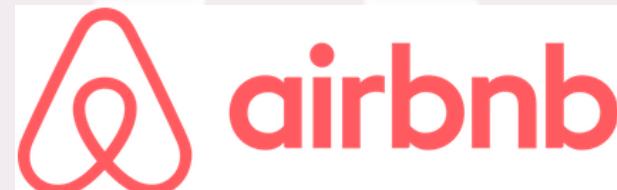
R-Ladies Taipei

# 所以 kaggle?



R-Ladies Taipei

# 誰在 kaggle™ 辦過比賽？



# 如何開始打 kaggle ?

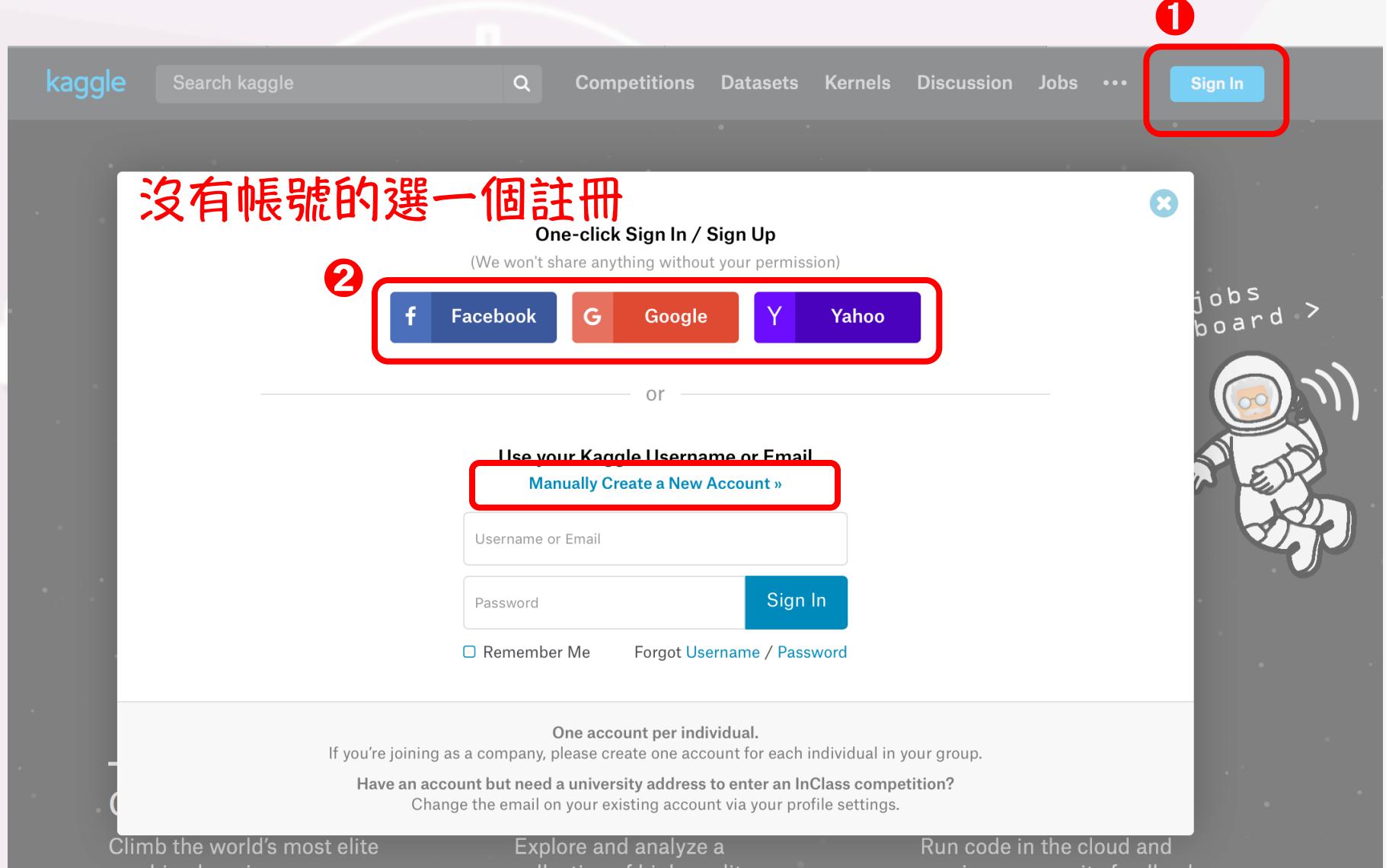


# 首先你需要...





- Account



The image shows the Kaggle sign-in page. At the top right, there is a red box labeled '1' around the 'Sign In' button. Below it, a large red box labeled '2' surrounds the 'One-click Sign In / Sign Up' section, which includes social media integration buttons for Facebook, Google, and Yahoo. A red box also surrounds the 'Manually Create a New Account' link. The page features a central text overlay in Chinese: '沒有帳號的選一個註冊' (For those without an account, choose one to register). There is also a small illustration of an astronaut on the right side.

Search kaggle

Competitions Datasets Kernels Discussion Jobs ...

Sign In

沒有帳號的選一個註冊

One-click Sign In / Sign Up  
(We won't share anything without your permission)

Facebook Google Yahoo

or

Use your Kaggle Username or Email

Manually Create a New Account »

Username or Email

Password

Sign In

Remember Me    [Forgot Username / Password](#)

One account per individual.  
If you're joining as a company, please create one account for each individual in your group.

Have an account but need a university address to enter an InClass competition?  
[Change the email on your existing account via your profile settings.](#)

Climb the world's most elite  
data science competition

Explore and analyze a  
library of high-quality datasets

Run code in the cloud and  
share your results with the world



- Account

**RLadies Taipei**

Taiwan  
Joined 7 months ago · last seen in the past day

Competitions Novice

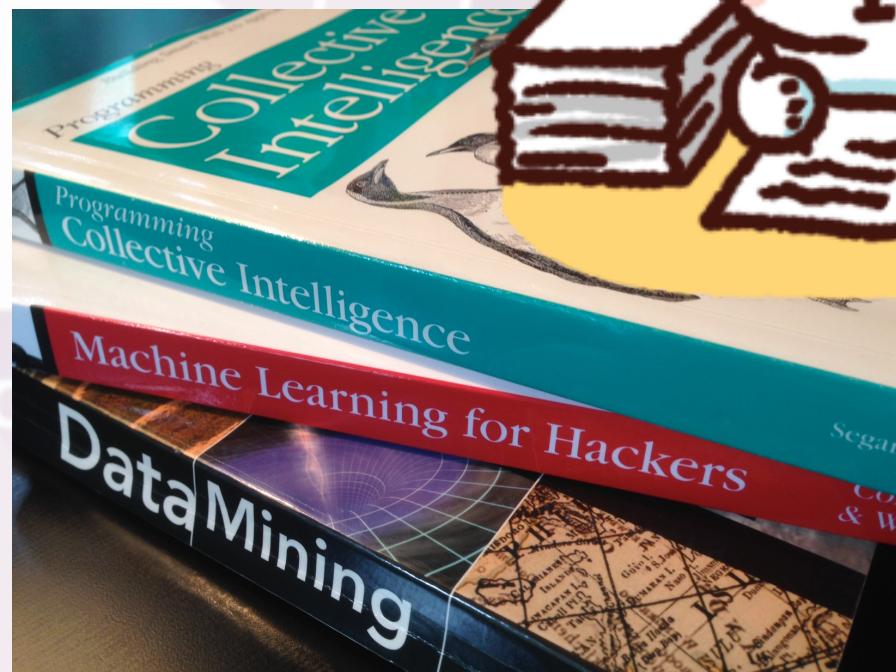
[Edit Profile](#)

[Home](#) Competitions (2) Kernels (1) Discussion (0) Datasets (0) ...

Competitions Novice	Kernels Novice	Discussion Novice
Unranked	Unranked	Unranked
0	0	0

接著你需要...

R-Lab

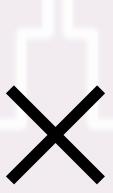


# 準備好了嗎！



R-Ladies Taipe





kaggle™

# R-LADIES TAIPEI 2017 KAGGLE 大賽

2017.0722 – 2017.0723

主辦單位



贊助單位



國泰慈善基金會  
Cathay Charity Foundation



Microsoft



In Bloom  
印花樂

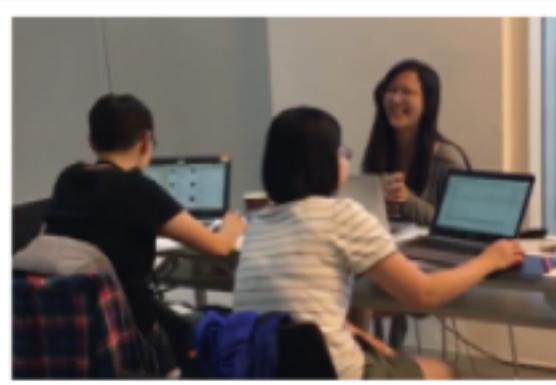
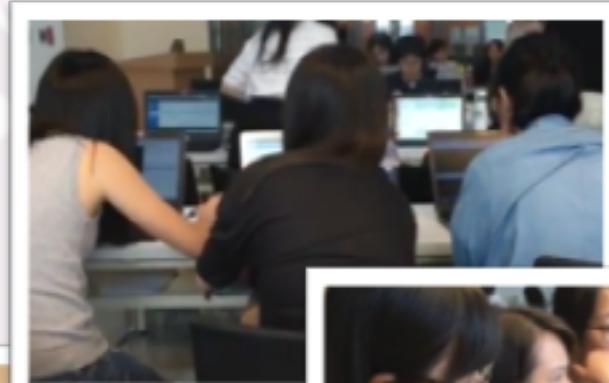


**7/22** **7/22**  
**11:00** **13:00**

**7/23** **7/23**  
**13:00** **14:30**



## Summit



# Competition

Featured Prediction Competition

## Instacart Market Basket Analysis

Which products will an Instacart consumer purchase again?

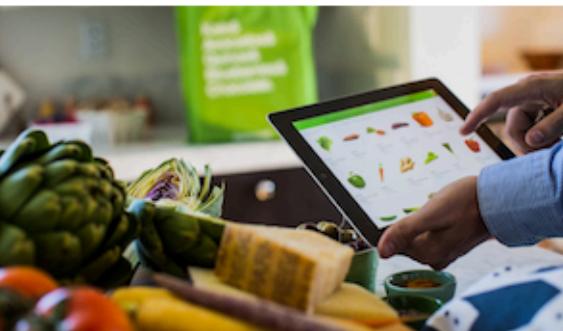
Instacart · 1,608 teams · a month to go

\$25,000 Prize Money

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#)

**Overview**

Description	Whether you shop from meticulously planned grocery lists or let whimsy guide your grazing, our unique food rituals define who we are. Instacart, a grocery ordering and delivery app, aims to make it easy to fill your refrigerator and pantry with your personal favorites and staples when you need them. After selecting products through the Instacart app, personal shoppers review your order and do the in-store shopping and delivery for you.
Evaluation	
Prizes	
Timeline	



# Competition

- Website

<https://www.kaggle.com/c/instacart-market-basket-analysis>

- Instacart

利用募集群眾外包配送的服務，主要為配送的物品以生活必需品與雜貨為主，商品大約為 30 多萬個品項。



詳細介紹：<https://www.inside.com.tw/2014/10/09/instacart>

# Competition

Featured Prediction Competition

## Instacart Market Basket Analysis

Which products will an Instacart consumer purchase again?

Instacart · 1,641 teams · a month to go

\$25,000 Prize Money

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Join Competition](#)

Overview

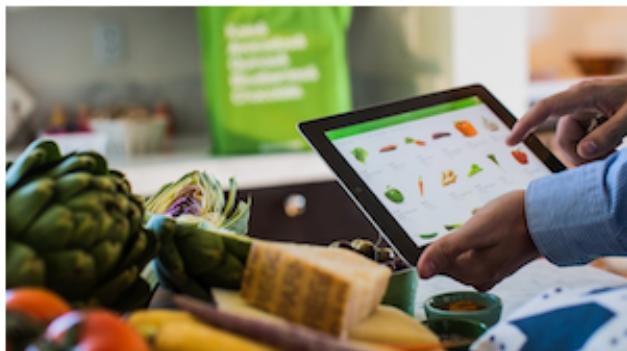
Description

Evaluation

Prizes

Timeline

Whether you shop from meticulously planned grocery lists or let whimsy guide your grazing, our unique food rituals define who we are. Instacart, a grocery ordering and delivery app, aims to make it easy to fill your refrigerator and pantry with your personal favorites and staples when you need them. After selecting products through the Instacart app, personal shoppers review your order and do the in-store shopping and delivery for you.



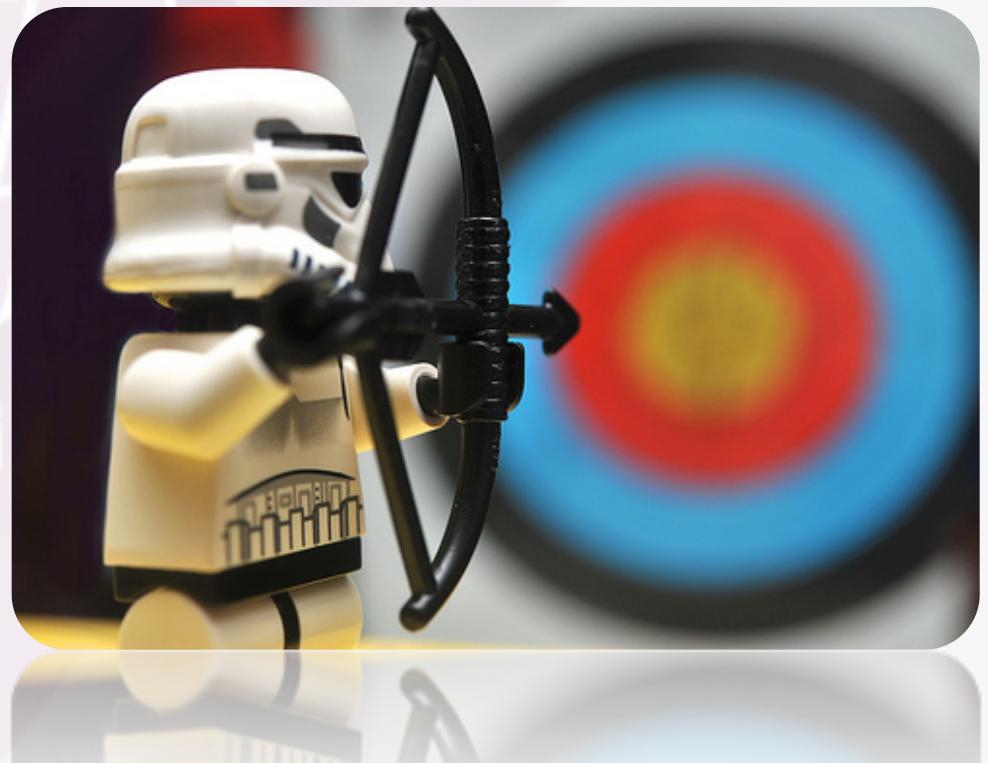
Join!!!!

Instacart's data science team plays a big part in providing this delightful shopping experience. Currently

# Competition

- Target

預測顧客下一筆訂單中會有哪些商品



# 讓我們介紹一下資料集

R-Ladie



# About the Data -- 所有檔案

## Data : 7 Files

aisles.csv	商品子類別
departments.csv	商品類別
products.csv	Instacart 販售的商品項目
orders.csv	各訂單描述
order_products_prior.csv	顧客最近一筆訂單的前一筆訂單
order_products_train.csv	顧客最近一筆訂單(Train)
sample_submission.csv	上傳格式範例

# About the Data -- 檔案關聯及資料屬性

商品子類別	AISLES.CSV
	+ aisle_id: integer in [1:134]
	+ aisle: string

商品資訊	PRODUCTS.CSV
	+ product_id: integer in [1:49688]
	+ product_name: string
	+ aisle_id: integer
	+ department_id: integer

顧客最近一筆的訂單(Train)

顧客以前的訂單	ORDER_PRODUCTS__PRIOR.CSV
	+ order_id: integer
	+ product_id: integer
	+ add_to_cart_order: integer
	+ reordered: boolean 0-1

各訂單描述	ORDER_PRODUCTS__TRAIN.CSV
	+ order_id: integer
	+ product_id: integer
	+ add_to_cart_order: integer
	+ reordered: boolean 0-1

上傳格式範例	SAMPLE_SUBMISSION.CSV
	+ order_id: integer
	+ product_id: integer

# About the Data -- 各資料集說明

**department.csv** (共分成21個商品類別)

department_id	商品類別編號
department	商品類別名稱



# About the Data -- 各資料集說明

## aisles.csv (共有134個商品子類別)

aisle_id	商品子類別編號
aisle	商品子類別名稱

Note.

商品子類別分類更細緻，有助於理解商品間的關係

Ref: <https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/34302>

R-Ladies Taipei

# About the Data -- 各資料集說明

Departments							
<b>Produce</b> Fresh Fruits Fresh Herbs Fresh Vegetables Packaged Vegetables &...	 <b>Meat &amp; Seafood</b> Packaged Poultry Hot Dogs, Bacon & Sau... Packaged Meat Meat Counter <a href="#">View more &gt;</a>	<b>Deli</b> Lunch Meat Specialty Cheeses Fresh Dips & Tapenades Prepared Meals Prepared Soups & Sala...	 <b>Bakery</b> Bread Tortillas & Flat Bread Buns & Rolls Breakfast Bakery Bakery Desserts				
<b>Dairy &amp; Eggs</b> Milk Cream Eggs Packaged Cheese <a href="#">View more &gt;</a>	 <b>Bulk</b> 	<b>Canned Goods</b> Canned & Jarred Veget... Canned Meals & Beans Soup, Broth & Bouillon Canned Meat & Seafood Canned Fruit & Apples...	 <b>Dry Goods &amp; Pasta</b> Dry Pasta Pasta Sauce Grains, Rice & Dried G... Fresh Pasta Instant Foods				
<b>Pantry</b> Condiments	 <b>International</b> Asian Foods	<b>Beverages</b> Tea	 <b>Breakfast</b> Cereal				

Ref : <https://www.instacart.com/store/browse>

# About the Data -- 各資料集說明

**products.csv** (Instacart 共販售 50K 個商品)

product_id	商品編號
product_name	商品名稱
aisle_id	商品子類別編號
department_id	商品類別編號

R-Ladies Taipei

# About the Data -- 各資料集說明

## order\_products\_prior.csv

order_id	訂單編號
product_id	商品編號
add_to_cart_order	商品放入購物車的順序
reordered	是否曾經購買過 (1 :表示有過購買記錄, 0 :表示第一次購買/沒有購買記錄)

R-Ladies Taipei

# About the Data -- 各資料集說明

## order\_products\_train.csv

order_id	訂單編號
product_id	商品編號
add_to_cart_order	商品放入購物車的順序
reordered	是否曾經購買過 (1 :表示有過購買記錄, 0 :表示第一次購買/沒有購買記錄)

R-Ladies Taipei

# About the Data -- 各資料集說明

**order.csv** (記錄各個顧客的所有訂單含 3.4M 筆訂單, 206K 個顧客)

order_id	訂單編號
user_id	顧客編號 3.2M筆 131K筆 75K筆
eval_set	屬於哪個資料集( prior / train / test)
order_number	該顧客訂單成立的順序
order_dow	訂單時間[星期] (0 = Saturday, 1 = Sunday,...)
order_hour_of_day	訂單時間[小時] (0 ~ 24)
day_since_prior_order	距離上一筆訂單的天數 (第一次購買沒有數據 : NAs 大於30天以30記)

# About the Data -- 上傳格式範例說明

sample\_submission.csv

order_id	訂單編號
product_id	商品編號



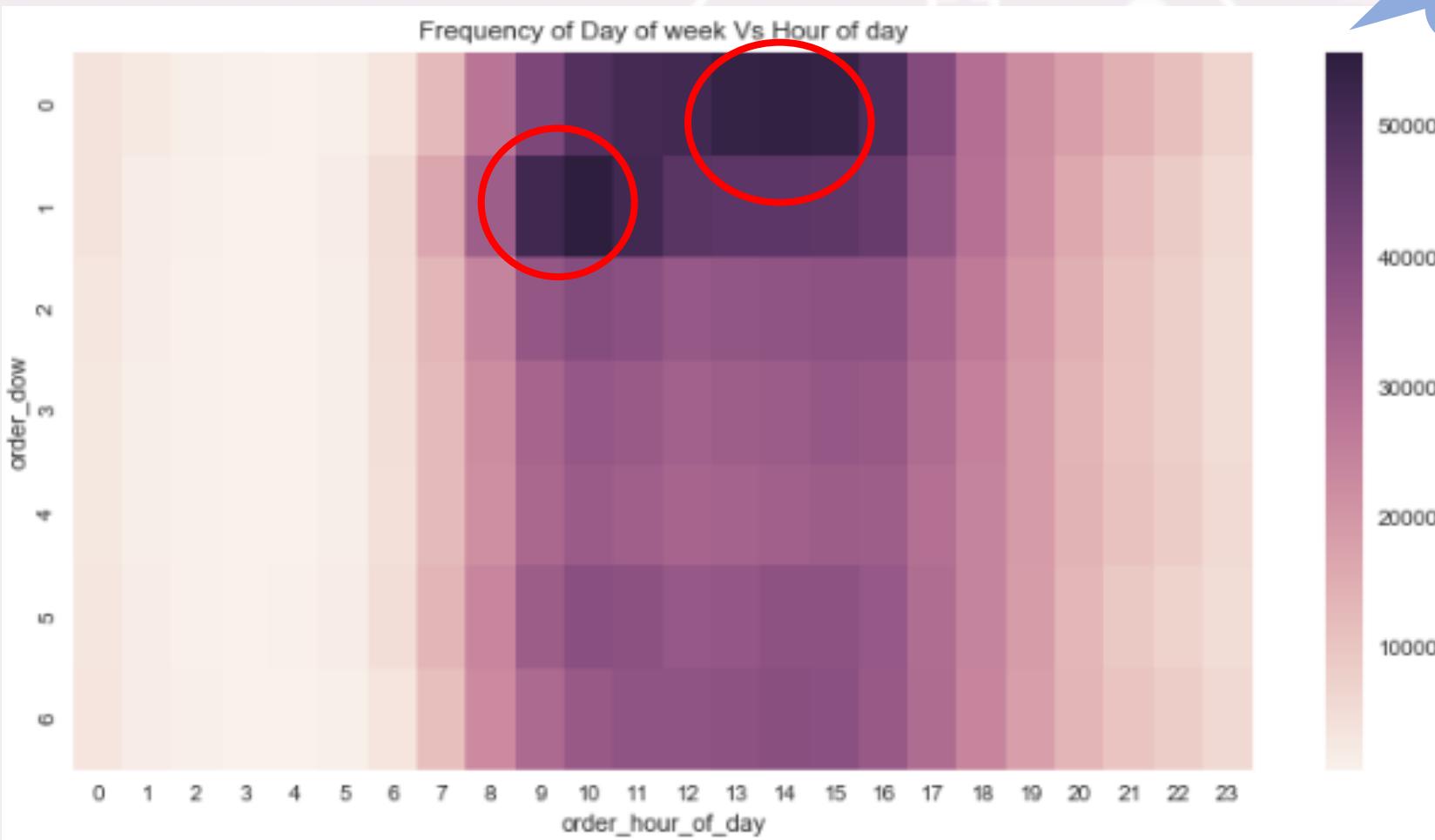
order_id	products
2774568	17668 21903 39190 47766 18599 43961 23650 24810
1528013	21903 38293
1376945	33572 28465 27959 44632 24799 34658 14947 30563 8309 13176
1356845	11520 14992 7076 28134 10863 13176
2161313	11266 196 10441 12427 37710 48142 14715 27839
1416320	5134 21903 21137 24852 17948 41950 24561

我們發現了...



# Exploratory Data Analysis

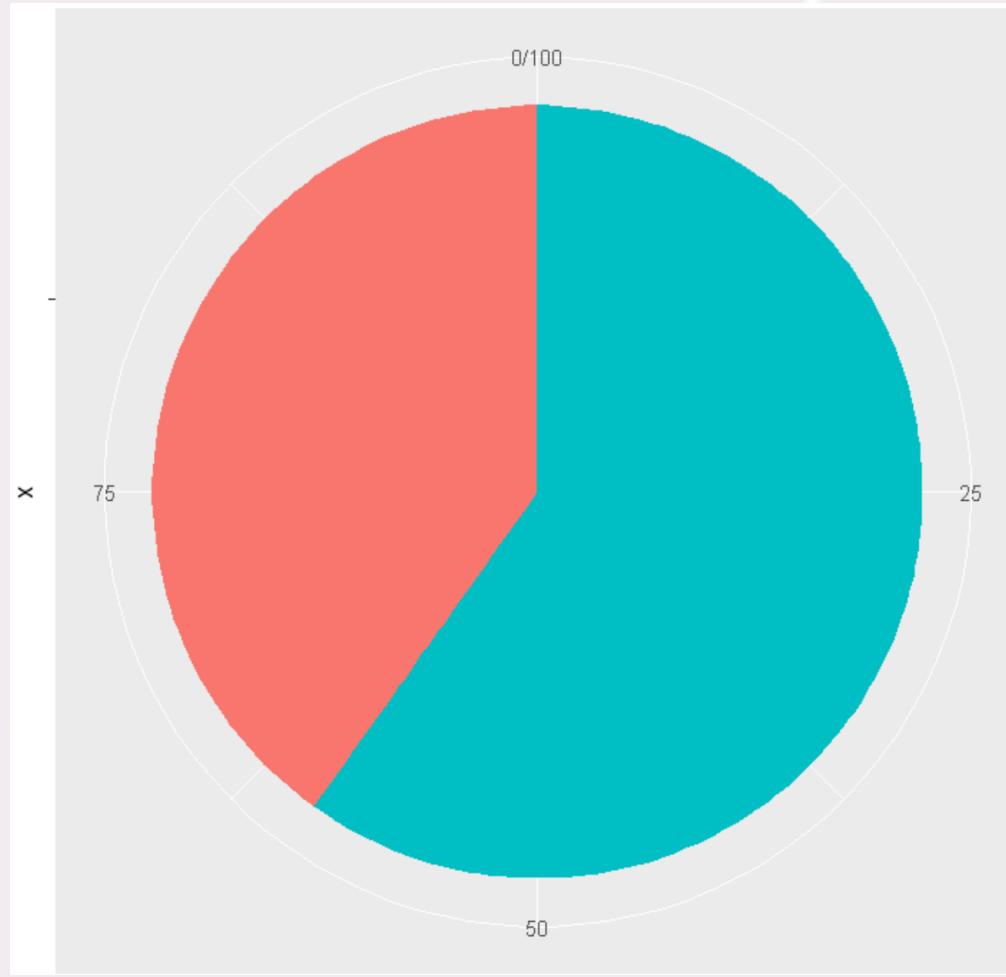
顧客大約都在何時訂購？



週六中午過後及週日早上為最密集的購買時間

# Exploratory Data Analysis

商品回購率？

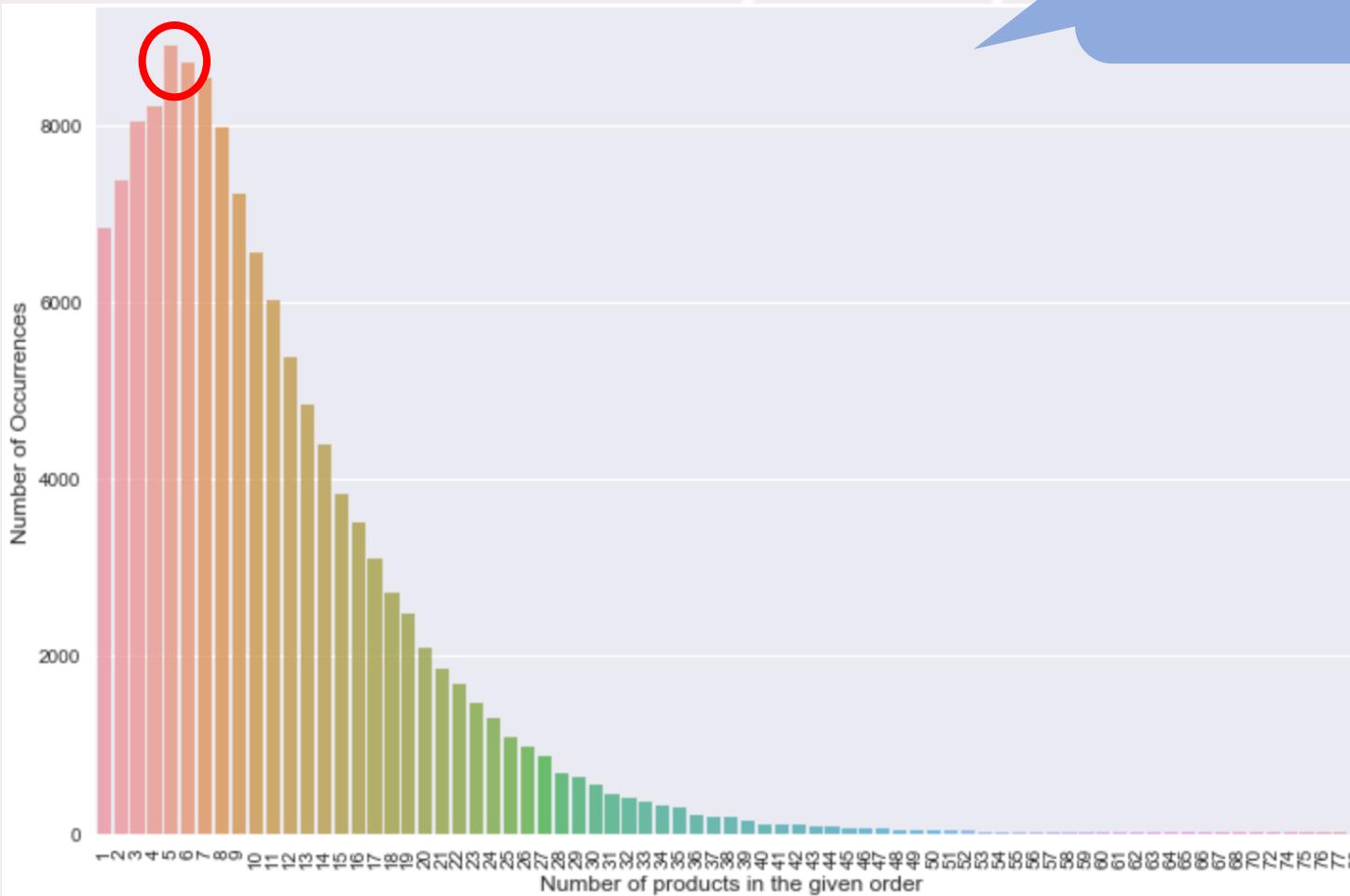


回購率大約為 59%

# Exploratory Data Analysis

一筆訂單中有多少件商品? (train)

大部分訂單一次購買 **5** 件商品



想看看更詳細的**EDA**整理結果？

寫 R 的朋友走這邊：<http://goo.gl/fThkMb>

寫 Python 的朋友來這裡：<http://goo.gl/H3ARDa>  
<http://goo.gl/ZEVjYK>

R-Ladies Taipei

# LET'S JOIN KAGGLE !!!

