Appier Interview Report

廖偉丞 Vincent Liao

2023.02.21

自我介紹

廖偉丞
Vincent Liao

學歷	● 台大國企系畢業（2017.09 ~ 2022.06） ● 政大資管所碩一就讀中（2022.09 ~
工作/實習經驗	● 後端工程師實習生 | 國泰金控數數發中心（2022.02 ~ 2022.08） ● JAVA 程式設計助教 | 政大資管系（2022.09 ~ ）
課外活動	● NTU DAC 台大資料分析社（2021.09 ~ 2022.06） ○ 企業合作專案 feat.烘焙找材料（2021.10 ~ 2021.12） ○ 企業合作專案 feat.PChome（2022.03 ~ 2022.06）
程式技能	Python, Java, Javascript, React.js, MySQL, PostgreSQL, Docker, Kubernetes
興趣	踢足球、下圍棋

Agenda

I. Data Overview

II. 資料清理與缺失值處理

III. 找到目標客群

IV. Create User Persona

IV. Advanced Questions

我將簡單分享 H&M 資料集的分析流程，並提供分析結果以及目標客群之user persona，最後會分享對於這份資料集的一些其他發現與後續之可研究方向。

Key Takeaways

1. **資料清理**：針對 customers.csv 中的資料做缺失 值處理
2. **目標客群**：基於 RFM 分析之結果選擇 Champions 作為目標客群
3. **User Persona**：| Jennifer | 25y | 化妝品產業 PM | 未婚 | 中產階級 |
4. **其他發現**：從 2020-05-12 開始，每日訂單數量以及訂單金額有非常大幅度的增長

Agenda

此份資料為 H&M 所提供，一共三個 csv 檔以及一個圖片資料夾。

images	• 裡面的每張服飾圖片分別對應了unique *article_id*
articles.csv	• 存有每一個 *article_id* 的 metadata，可以清楚了解每個訂單的購買商品資訊 • 總共 105542 種商品資料（包含不同色號）只有*detail_desc* 的部分有缺失值
customers.csv	• 存有每位顧客的metadata，每位顧客都對應了unique *customer_id* • 總共 1371980 位顧客資料，包含了*Active, club_member_status, fashion_news_frequency, age* 等等屬性資料
transactions_train.csv	• 有從 2018.09.20 - 2020.09.22 的訂單資料，內容包含 *t_dat, article_id, customer_id*，以及其他屬性資料 • 總共 31788324 筆交易資料，一筆資料代表一樣被購買之商品，若有N 筆完全重複之資料，則為購買了N 次

Appier

articles

```
RangeIndex: 105542 entries, 0 to 105541
Data columns (total 25 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   article_id                    105542 non-null  int64
 1   product_code                  105542 non-null  int64
 2   prod_name                     105542 non-null  object
 3   product_type_no               105542 non-null  int64
 4   product_type_name             105542 non-null  object
 5   product_group_name            105542 non-null  object
 6   graphical_appearance_no       105542 non-null  int64
 7   graphical_appearance_name     105542 non-null  object
 8   colour_group_code             105542 non-null  int64
 9   colour_group_name             105542 non-null  object
 10  perceived_colour_value_id     105542 non-null  int64
 11  perceived_colour_value_name   105542 non-null  object
 12  perceived_colour_master_id    105542 non-null  int64
 13  perceived_colour_master_name  105542 non-null  object
 14  department_no                 105542 non-null  int64
 15  department_name               105542 non-null  object
 16  index_code                    105542 non-null  object
 17  index_name                    105542 non-null  object
 18  index_group_no                105542 non-null  int64
 19  index_group_name              105542 non-null  object
 20  section_no                    105542 non-null  int64
 21  section_name                  105542 non-null  object
 22  garment_group_no              105542 non-null  int64
 23  garment_group_name            105542 non-null  object
 24  detail_desc                   105126 non-null  object
dtypes: int64(11), object(14)
memory usage: 20.1+ MB
```

customers

```
RangeIndex: 1371980 entries, 0 to 1371979
Data columns (total 7 columns):
 #   Column                  Non-Null Count    Dtype
---  ------                  --------------    -----
 0   customer_id             1371980 non-null  object
 1   FN                      476930 non-null   float64
 2   Active                  464404 non-null   float64
 3   club_member_status      1365918 non-null  object
 4   fashion_news_frequency  1355971 non-null  object
 5   age                     1356119 non-null  float64
 6   postal_code             1371980 non-null  object
dtypes: float64(3), object(4)
memory usage: 73.3+ MB
```

transactions

```
RangeIndex: 31788324 entries, 0 to 31788323
Data columns (total 5 columns):
 #   Column            Dtype
---  ------            -----
 0   t_dat             object
 1   customer_id       object
 2   article_id        int64
 3   price             float64
 4   sales_channel_id  int64
dtypes: float64(1), int64(2), object(2)
memory usage: 1.2+ GB
```

Agenda

主要使用 Python 中 的 Pandas 套件做資料分析處理。在看過資料後我判斷只有customers.csv 中的資料需要去做缺失值處理。

缺失值處理

1. 將 *FN* 屬性中的 nan 值改為 *0*, 以符合 binary value 的模式
2. 將 *Active* 屬性中的 nan 值改為 *0*, 以符合 binary value 的模式
3. 將 *club_member_status* 屬性中的 nan 值改為 *NON-ACTIVE*
4. 將 *fashion_news_frequency* 屬性中的 *None* 值改為 *NONE*

```
customers.FN.loc[customers.FN.isna() == True] = 0.0
customers.Active.loc[customers.Active.isna() == True] = 0.0
customers.club_member_status.loc[customers.club_member_status.isna() == True] = 'NON-ACTIVE'
customers.fashion_news_frequency.loc[customers.fashion_news_frequency == 'None'] = 'NONE'
```

Agenda

Appier

利用 RFM analysis 將所有顧客區分成多種類型，並從中挑出我們的目標客群。

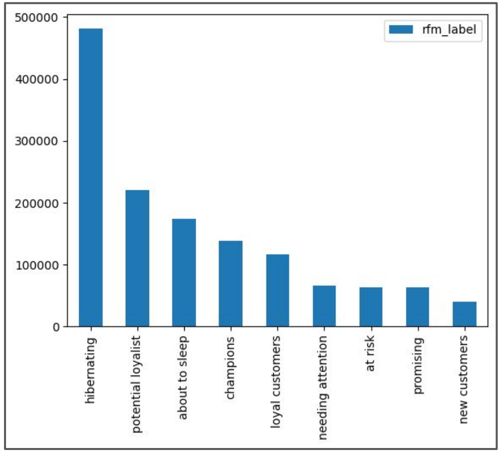
合併資料	將 *articles.csv* 以及 *transactions.csv* 合併以便後續的分析
計算 RFM & 建立 RFM score	1. 計算出每位顧客的 Recency, Frequency, Monetary 指標 2. 根據計算出的結果給予 1~5 分的分級，5 分代表最高，反之 1 分代表最低。
顧客分群	根據計算出來的分數將顧客分群，一共十種類型的顧客

顧客類型介紹

代號	Description
Champions	Bought recently, buy often and spend the most
Loyal Customers	Spend good money and often, responsive to promotions
Potential Loyalist	Recent customers, but spent a good amount and bought more than once
New Customers	Bought most recently, but not often
Promising	Recent shoppers, but haven’t spent much
Needing Attention	Above average recency, frequency and monetary values; may not have bought very recently though
About to Sleep	Below average recency, frequency and monetary values; will lose them if not reactivated
At Risk	Spent big money and purchased often but long time ago; need to bring them back
Can’t Loose Them	Made biggest purchases, and often but haven’t returned for a long time
Hibernating	Last purchase was long back, low spenders and low number of orders

Dataset Overview

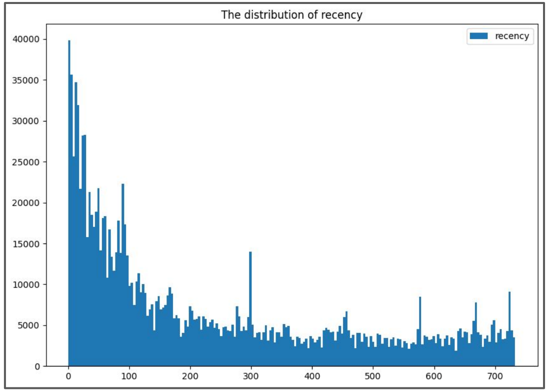
	customer_id	recency	monetary	frequency	r_score	f_score	m_score	rfm_sum	rfm_label
0	00000dbacae5abe5e23885899a1fa44253a17956c6d1c3...	17	0.6490	10	5	3	4	12	potential loyalist
1	0000423b00ade91418cceaf3b26c6af3dd342b51fd051e...	76	2.6019	23	4	4	5	13	loyal customers
2	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	7	0.7048	7	5	3	4	12	potential loyalist
3	00005ca1c9ed5f5146b52ac8639a40ca9d57aeff4d1bd2...	471	0.0610	1	1	1	1	3	hibernating
4	00006413d8573cd20ed7128e53b7b13819fe5cfc2d801f...	41	0.4697	6	4	3	4	11	potential loyalist
5	000064249685c11552da43ef22a5030f35a147f723d5b0...	356	0.1016	1	2	1	2	5	hibernating
6	0000757967448a6cb83efb3ea7a3fb9d418ac7adf2379d...	8	0.1660	3	5	2	3	10	potential loyalist
7	00007d2de826758b65a93dd24ce629ed66842531df6699...	132	3.8236	16	3	4	5	12	loyal customers
8	00007e8d4e54114b5b2a9b51586325a8d0fa74ea23ef77...	261	0.0534	1	2	1	1	4	hibernating
9	00008469a21b50b3d147c97135e25b4201a8c58997f787...	680	0.0781	1	1	1	2	4	hibernating

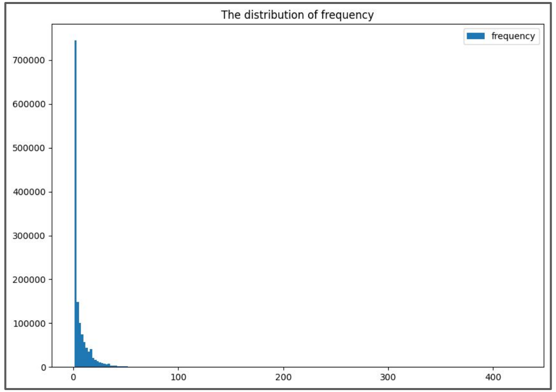


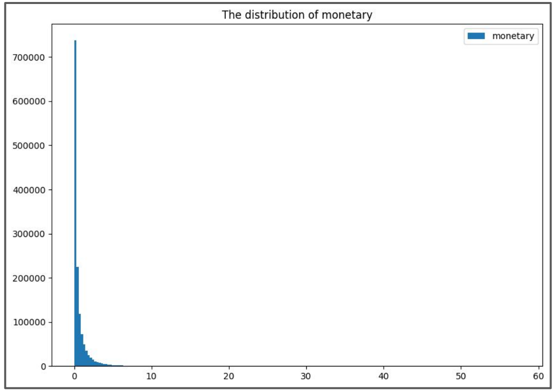
Appier

透過圖表檢視 Recency, Frequency, Monetary 指標的分佈情況。

Recency	Frequency	Monetary







Appier

尋找對品牌最有價值的顧客群。

指標 -> 平均每人貢獻之金額比例(monetary / number_of_people)

結果 -> 以 champions 族群作為我們的分析目標客群

	rfm_label	number_of_people	frequency	monetary	per_buy	per_person
0	about to sleep	173716	336880	34047.5503	0.1011	0.1960
1	at risk	63721	503749	47830.5122	0.0949	0.7506
2	champions	138804	3451812	341548.4287	0.0989	2.4607
3	hibernating	481013	762828	72659.5944	0.0953	0.1511
4	loyal customers	116095	2089726	202154.3818	0.0967	1.7413
5	needing attention	65524	455429	46999.4234	0.1032	0.7173
6	new customers	40195	56348	5497.7601	0.0976	0.1368
7	potential loyalist	219848	1333623	125991.5242	0.0945	0.5731
8	promising	63365	89784	7916.7991	0.0882	0.1249

Agenda

I. Data Overview

II. 資料清理與缺失值處理

III. 找到目標客群

IV. Create User Persona

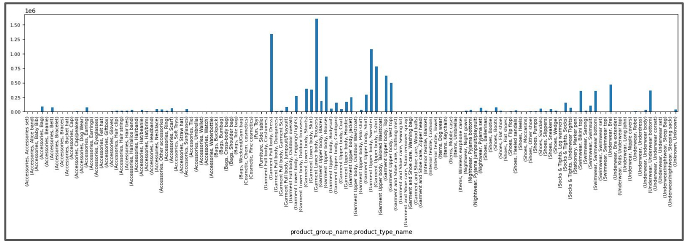
IV. Advanced Questions

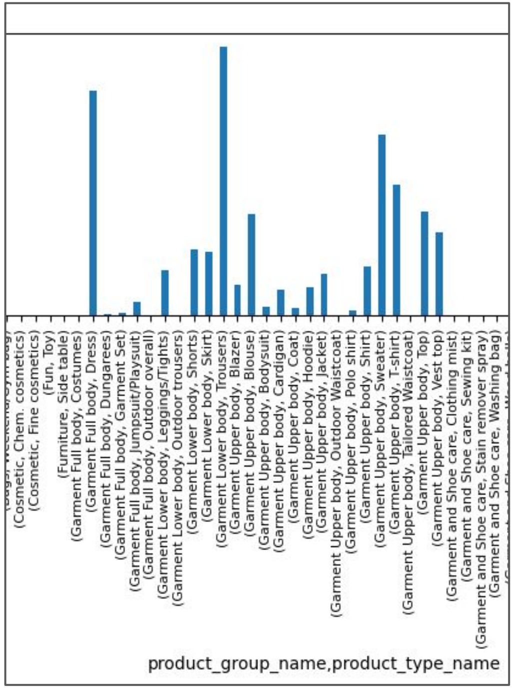
Appier

選定目標客群後，開始從各個資料維度去尋找資料集中性，作為user persona 之輪廓。
首先必須定義集中性，網路上並沒有資料集中性的明確定義，因此我自行訂定了幾個規則

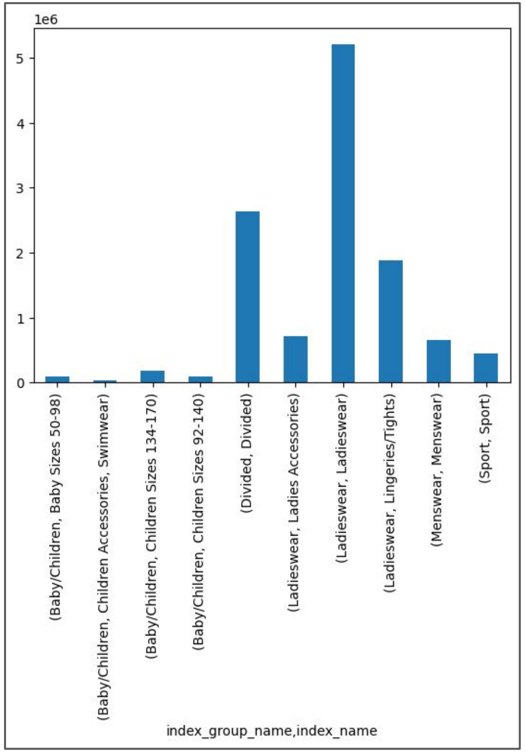
Rule	Description
Rule1	該屬性指標資料量必須超過總資料量的 10%
Rule2	若為連續性資料，則看分佈趨勢是否大致符合 Rule1

Transactions -> *Product_Group & Product_Type* -> (Garment Lower body, Trousers), (Garment Full body, Dress)

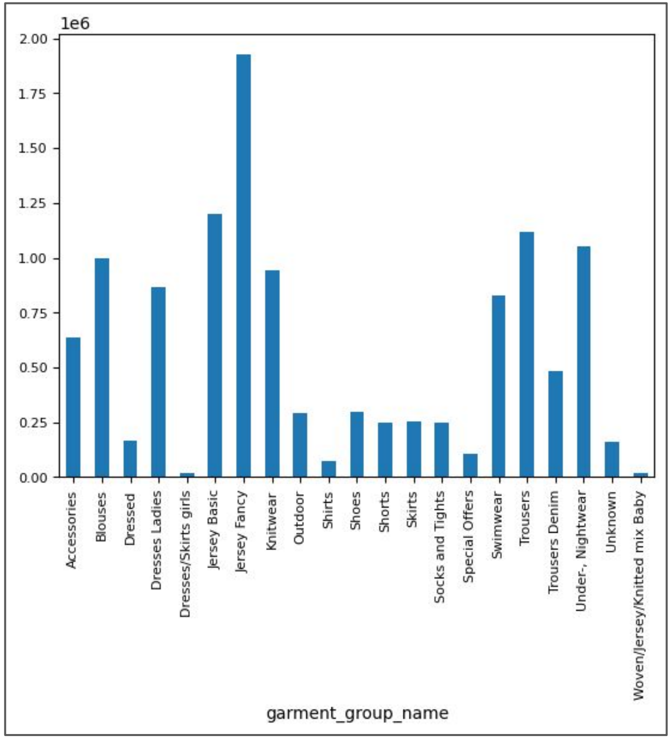




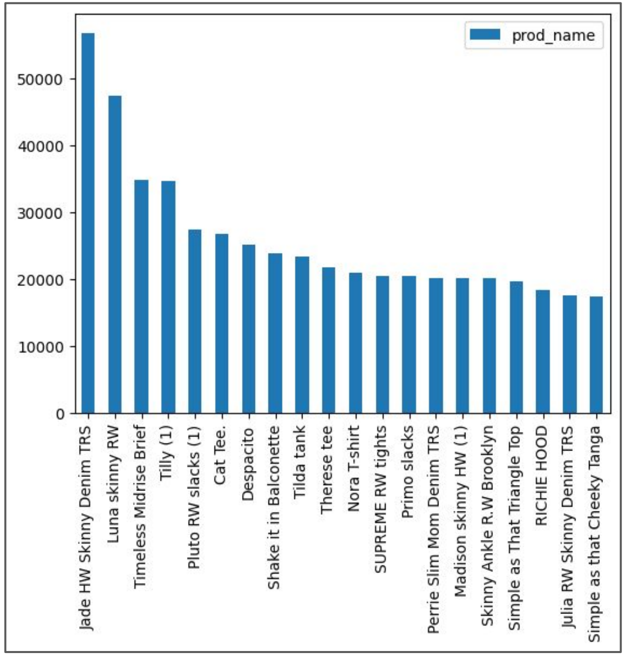
Transactions -> *Index_Group_Name & Index_Name* -> (Ladieswear, Ladieawear)

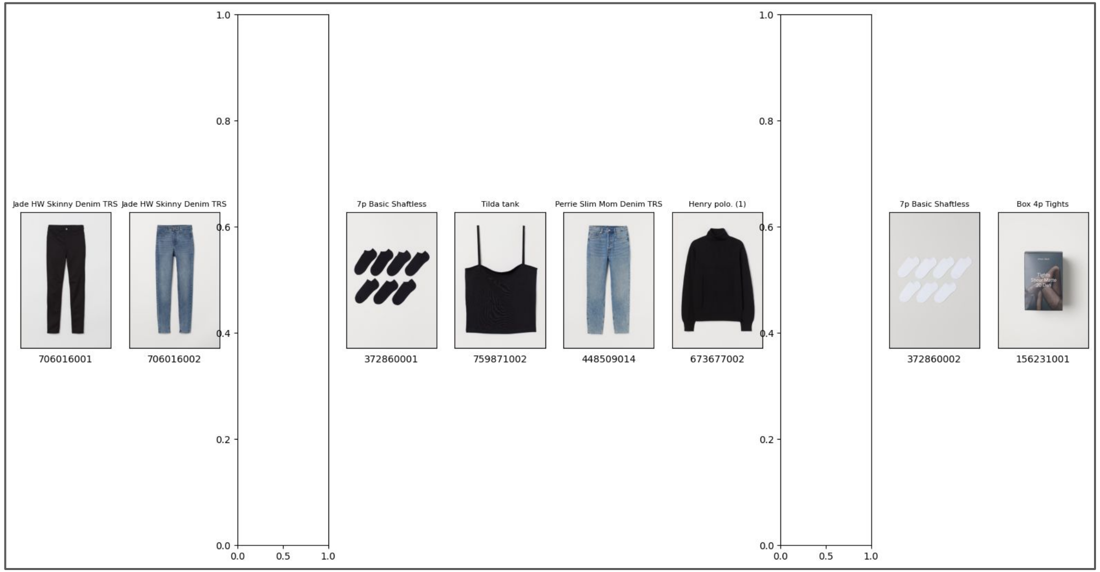


Transactions -> *Garment_Group_Name* -> Jersey_Fancy



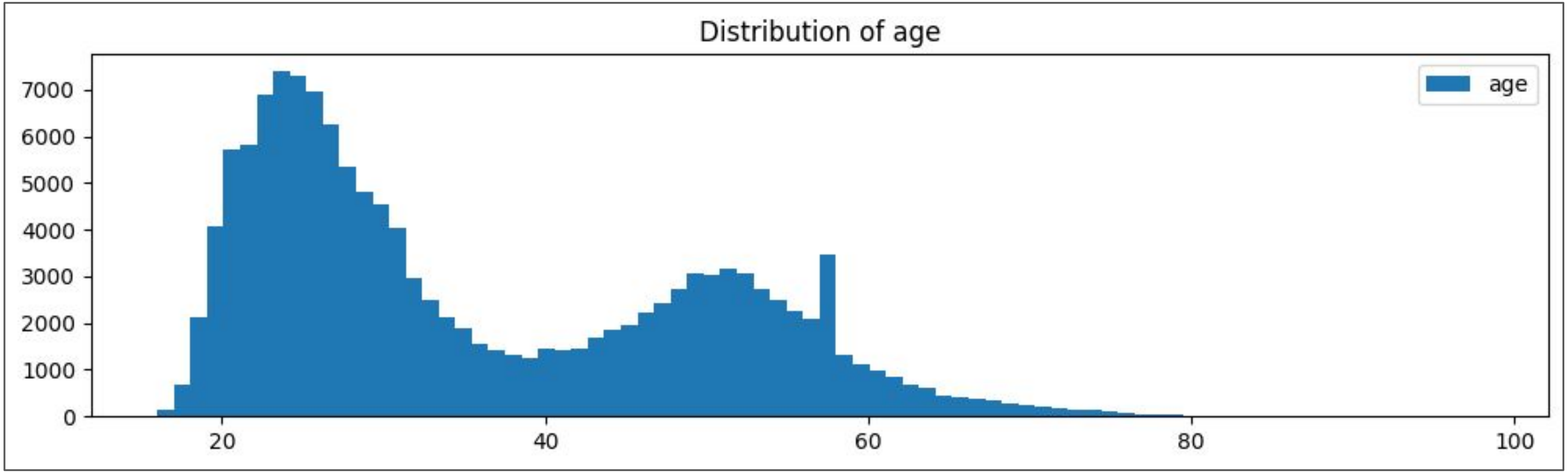
Transactions -> *Prod_Name* -> 集中性較不明顯



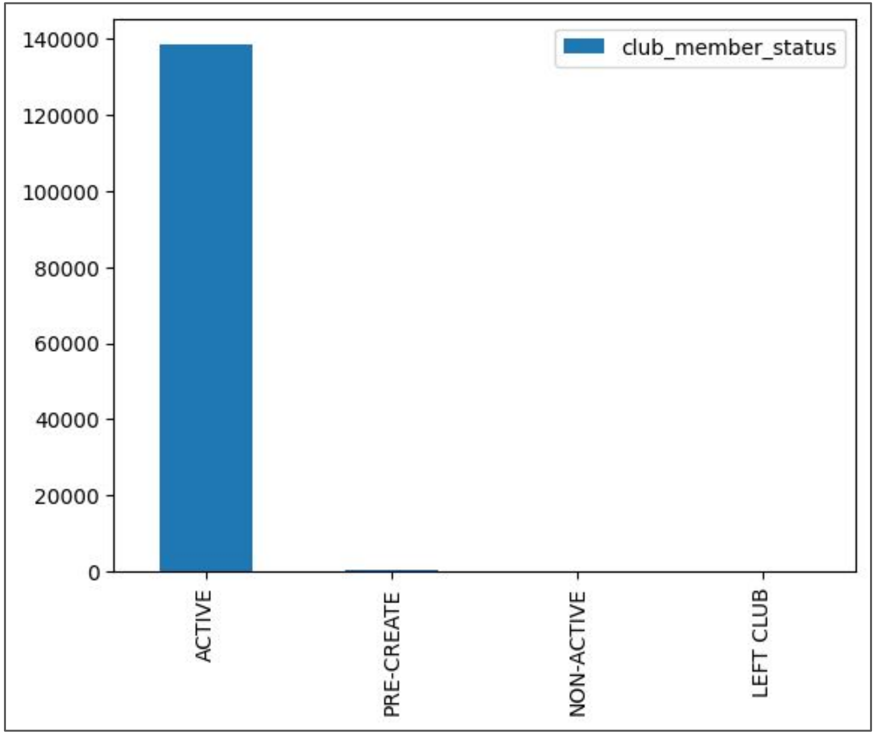


備註：空白為找不到圖片

Customers -> *Age* -> 23~26



Customers -> *Club_Member_Status* -> ACTIVE



根據上述各維度之集中性描繪出Champions 客群之 User Persona

基本資料	Description
姓名	Jennifer
身份	上班族
職業	化妝品產業 PM
性別	女
年齡	25
婚姻狀態	未婚
薪資水準	middle-income level
Club_Member_Status	ACTIVE

Appier

Jennifer’s challenge

Personal Challenge	三年前 Jennifer 作為初出社會茅廬的小資女孩，每天都非常認真努力的上班，身處於化妝品產業的她，身邊的同事們個個都很會打扮，在同儕壓力下她也開始研究起穿搭。在可支配金錢較少的限制之下，她成為了快時尚品牌 H&M 的忠實顧客之一，其多樣且高 CP值的服飾風格滿足了 Jennifer 的需求。然而隨著年紀漸長，Jennifer 在手頭逐漸寬裕的情況下逐漸有嘗試其他服飾品牌的念頭，她現在最大的問題是她不知道 H&M 是不是還適合現在的她 ...

Agenda

I. Data Overview

II. 資料清理與缺失值處理

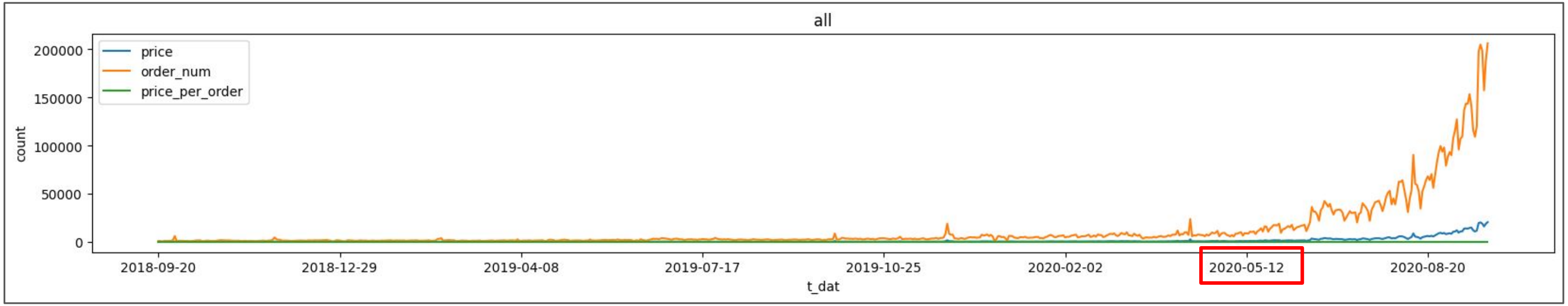
III. 找到目標客群

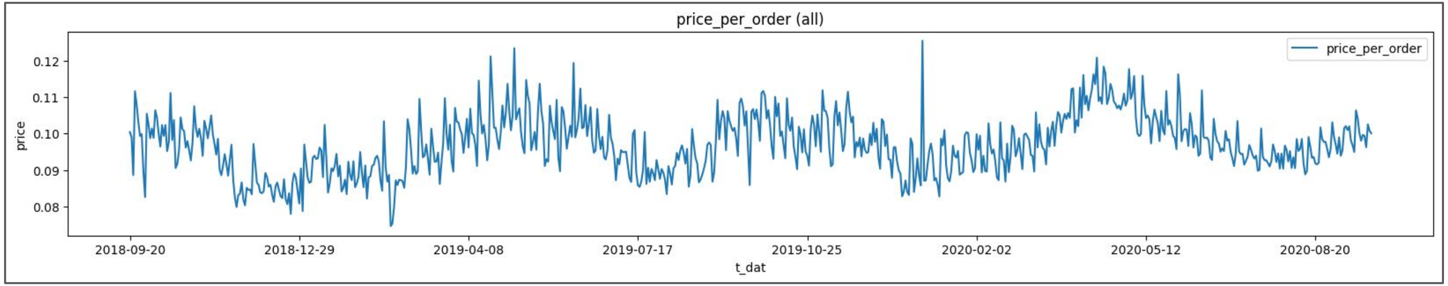
IV. Create User Persona

IV. Advanced Questions

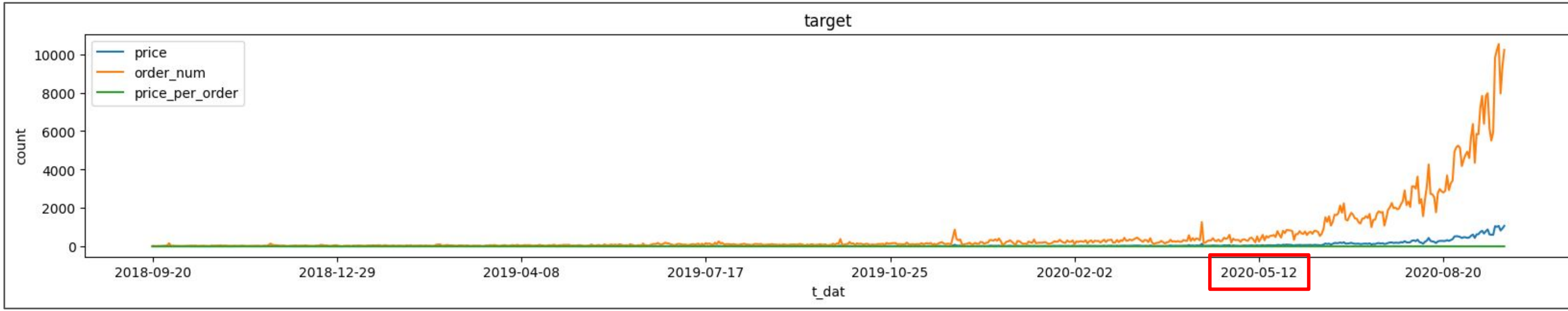
在分析資料過程中發現有趣的現象以及值得探討的問題

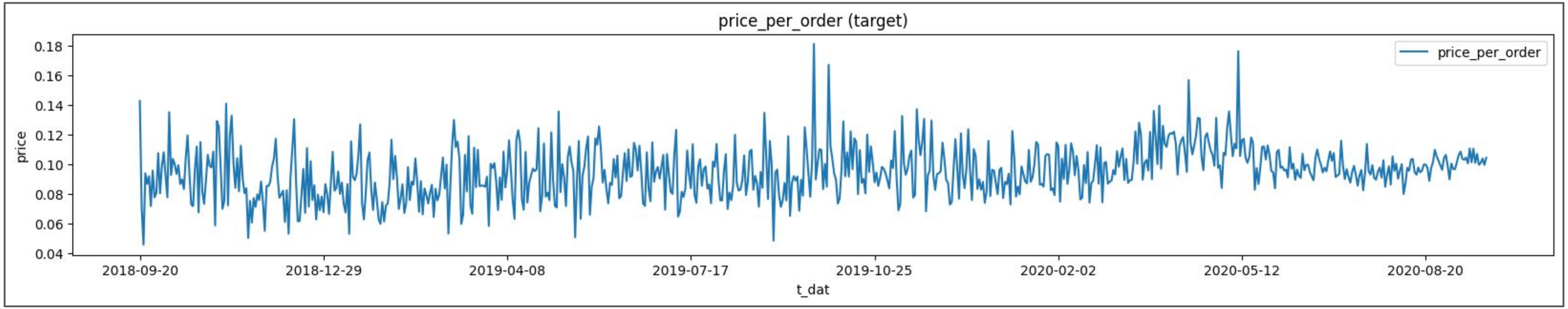
現象	從 2020-05-12 開始，每日訂單數量以及訂單金額有非常大幅度的增長



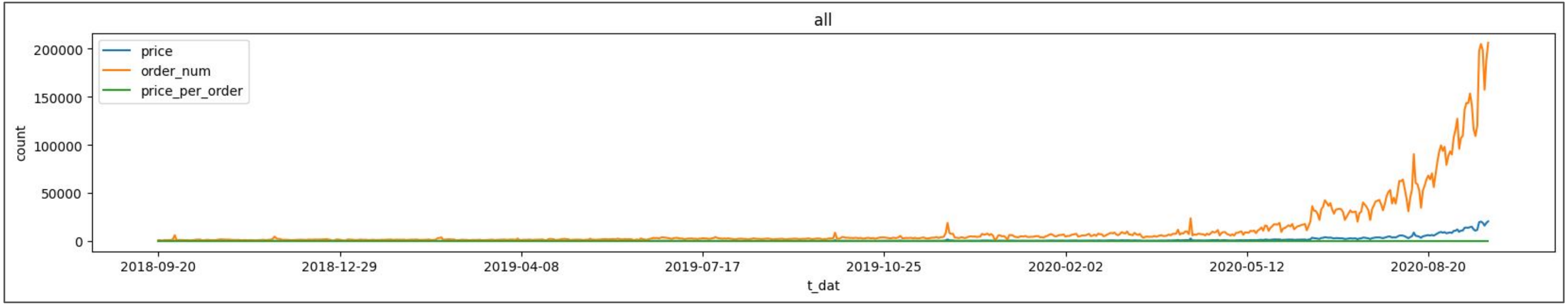


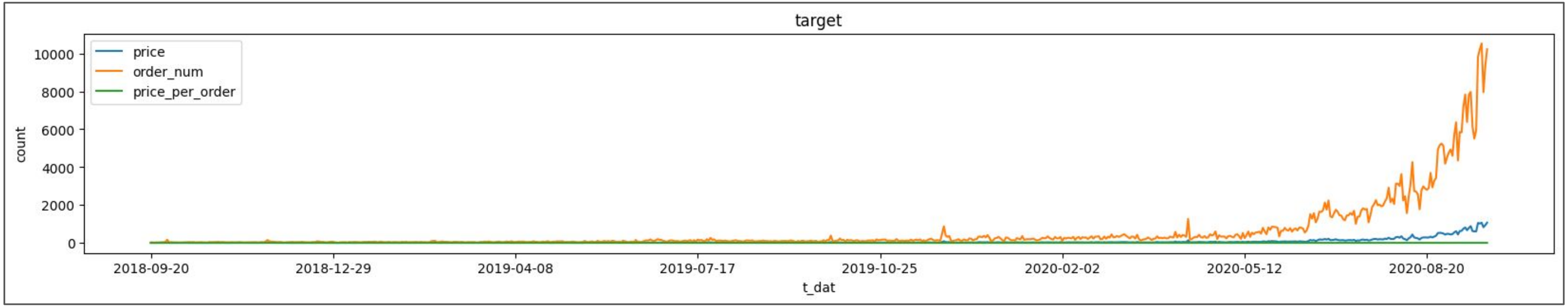
基於這個現象去分析了目標客群的相關數據，發現也是相同的趨勢。



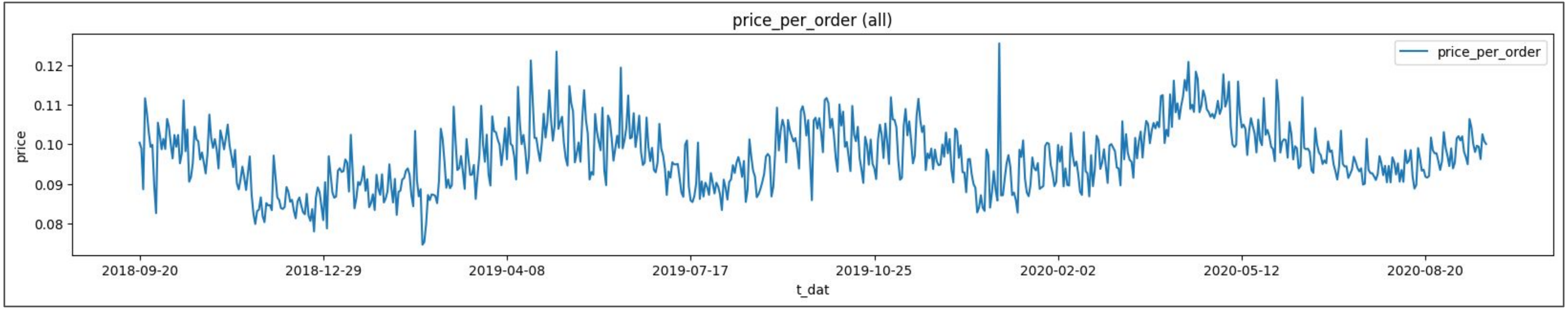


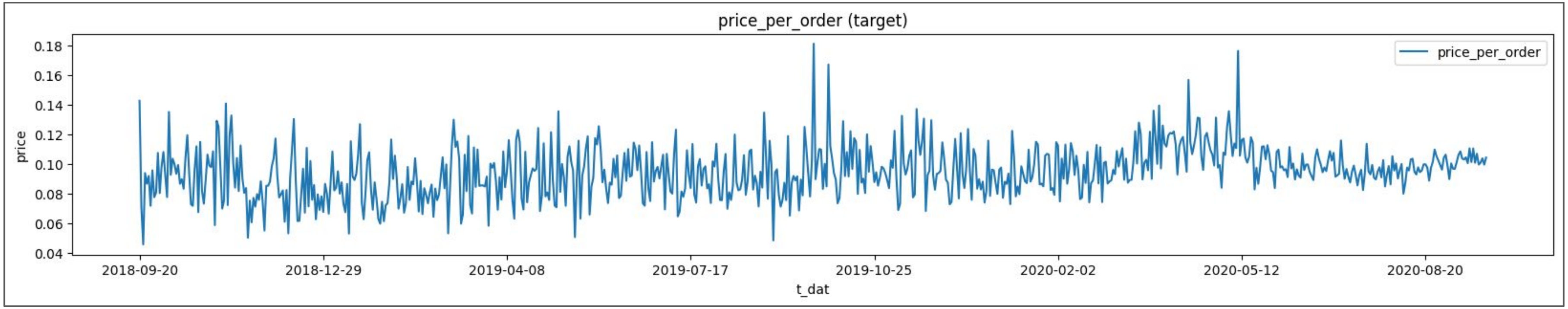
All vs Target





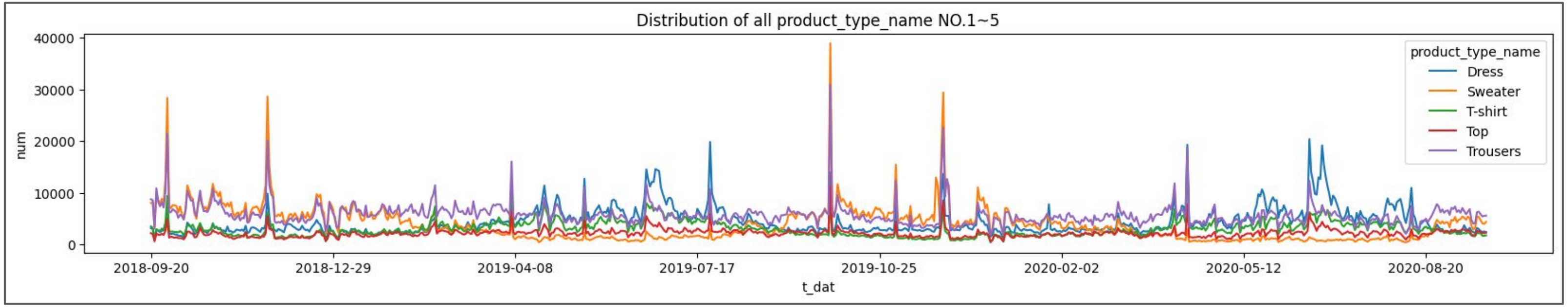
All vs Target

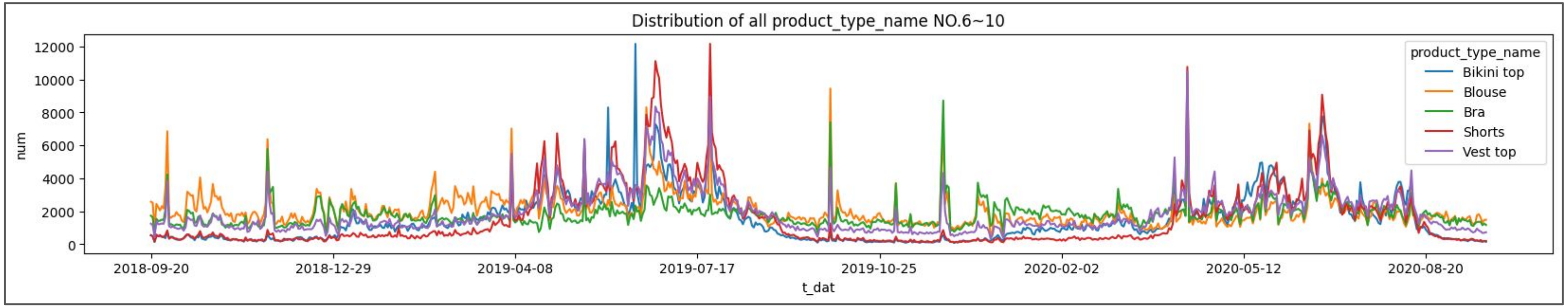




更進一步分析在暴增的訂單量中，又是哪些*product_type_name* 更常被購買。這邊一樣分為全體和 Target 來分析，分別找出這次交易資料區間內最常被購買的 product_type_name 前 20 名來分析。

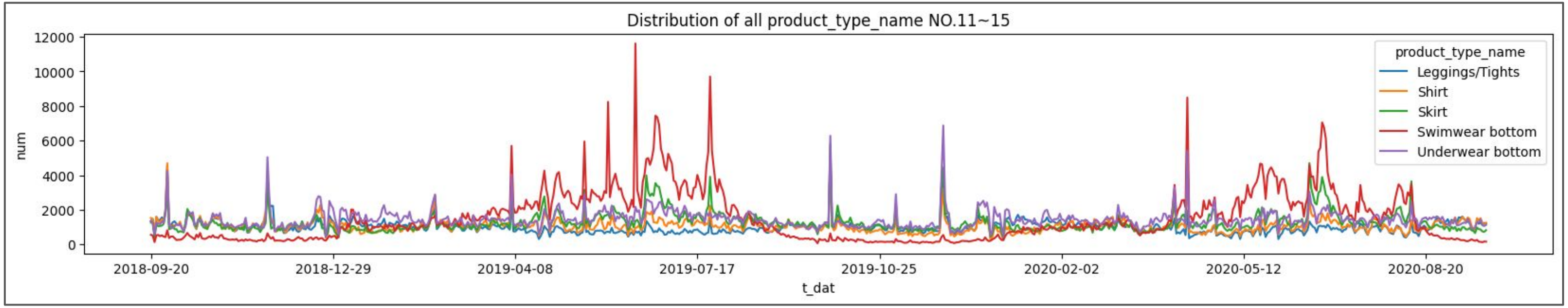
ALL

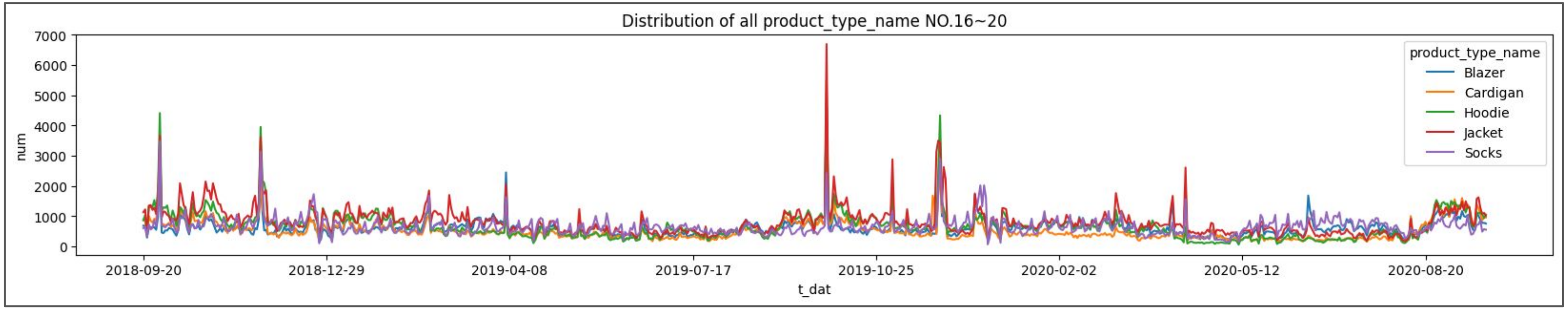




更進一步分析在暴增的訂單量中，又是哪些*product_type_name* 更常被購買。這邊一樣分為全體和 Target 來分析，分別找出這次交易資料區間內最常被購買的 product_type_name 前 20 名來分析。

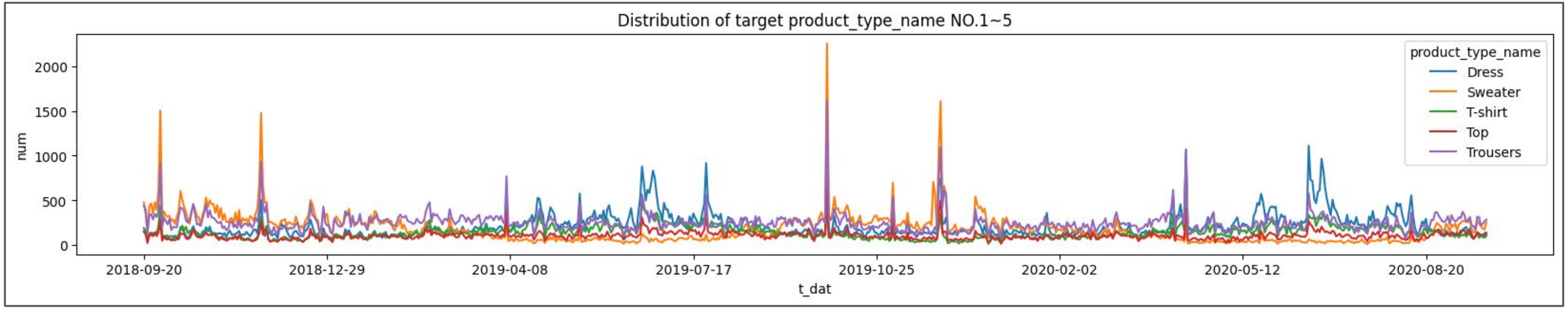
ALL

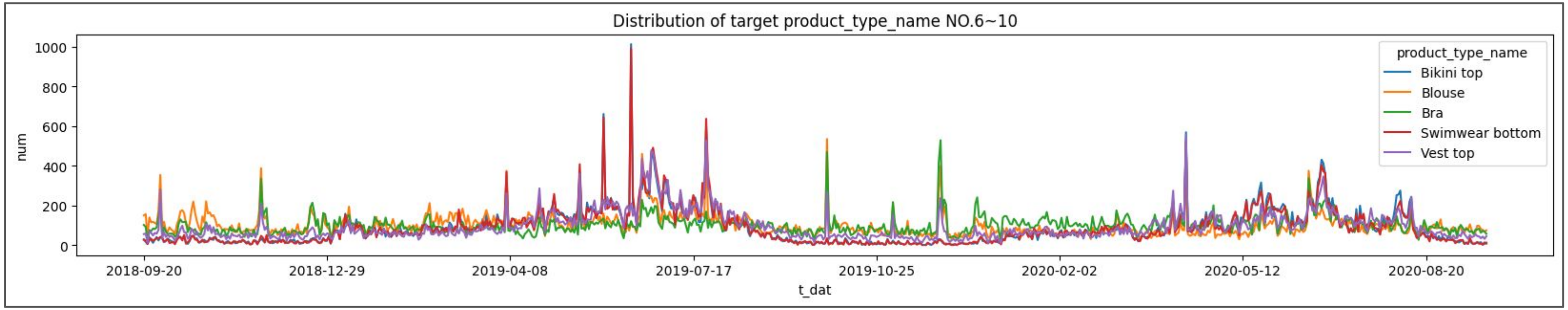




更進一步分析在暴增的訂單量中，又是哪些*product_type_name* 更常被購買。這邊一樣分為全體和 Target 來分析，分別找出這次交易資料區間內最常被購買的 product_type_name 前 20 名來分析。

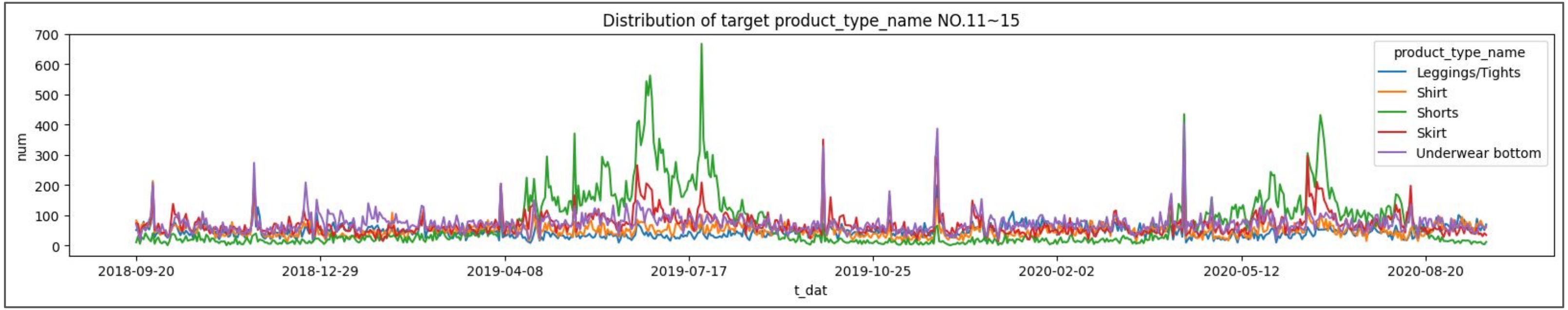
TARGET

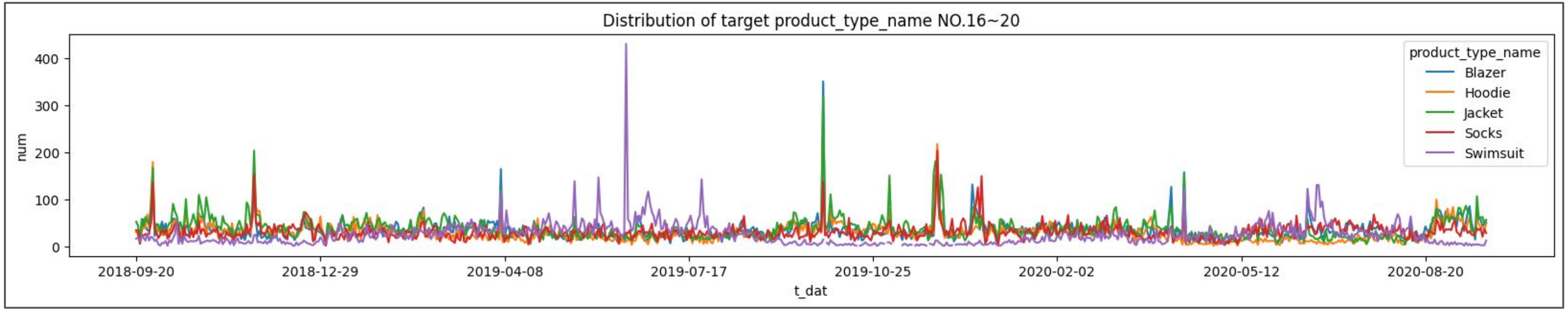




更進一步分析在暴增的訂單量中，又是哪些*product_type_name* 更常被購買。這邊一樣分為全體和 Target 來分析，分別找出這次交易資料區間內最常被購買的 product_type_name 前 20 名來分析。

TARGET





Appier

綜合以上結果，歸納出幾個未來可以繼續研究的方向。

研究方向

1. 若能夠獲得更多**消費者旅程**的相關數據，就可以更加清楚在各階段的各項進階數據。舉例來說：
   - 消費者旅程：搜尋 -> 瀏覽 -> 加入購物車 -> 購買
   - 數據：點擊、商品頁代碼、停留時間、結帳與否
2. 透過品牌的提供的**行銷活動資料**，就可以了解哪些活動以及商品會受到 TA 的喜愛，能夠刺激購買率。舉例來說：
   - 行銷活動資料：活動碼、日期、活動商品
3. 可以比較 TA 以及所有顧客的輪廓，善用 user persona 將**顧客形象具體化**並剖析其中的差異，以將所有消費者都轉化為 Champions 類型顧客作為核心目標。

End