

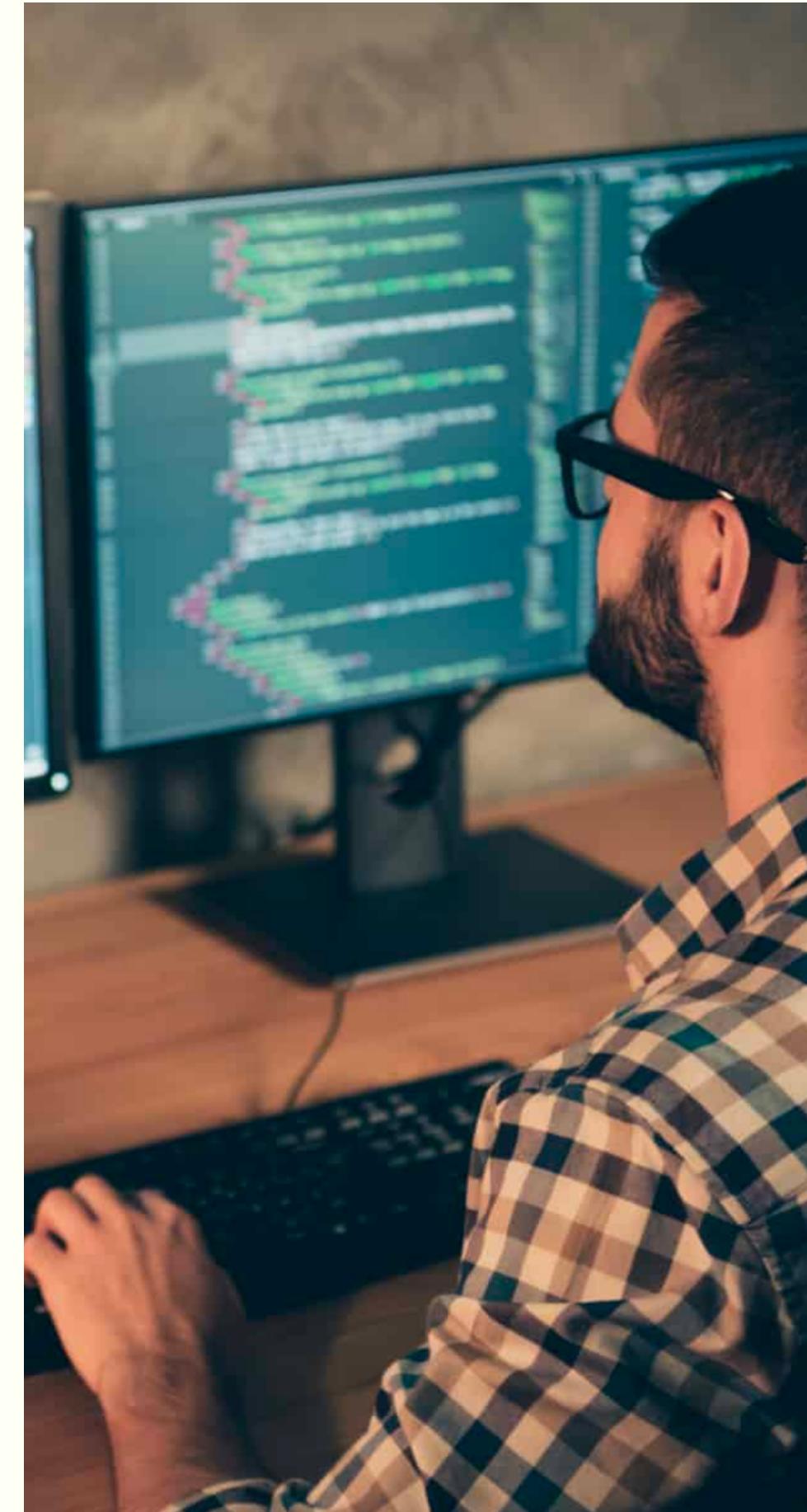
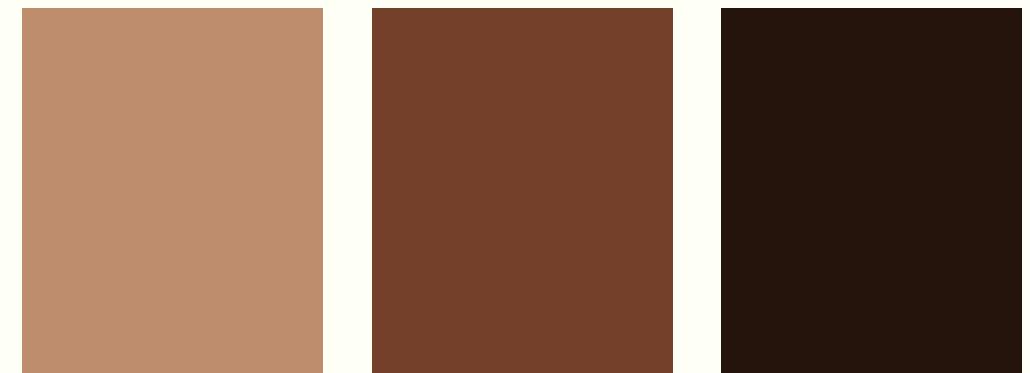
2024.01.10 WED

# TEXT MINING

NCCU X TEXT MINING

PRESENTED BY

蔡品洋、羅永富、葉瀚元



Hello world!

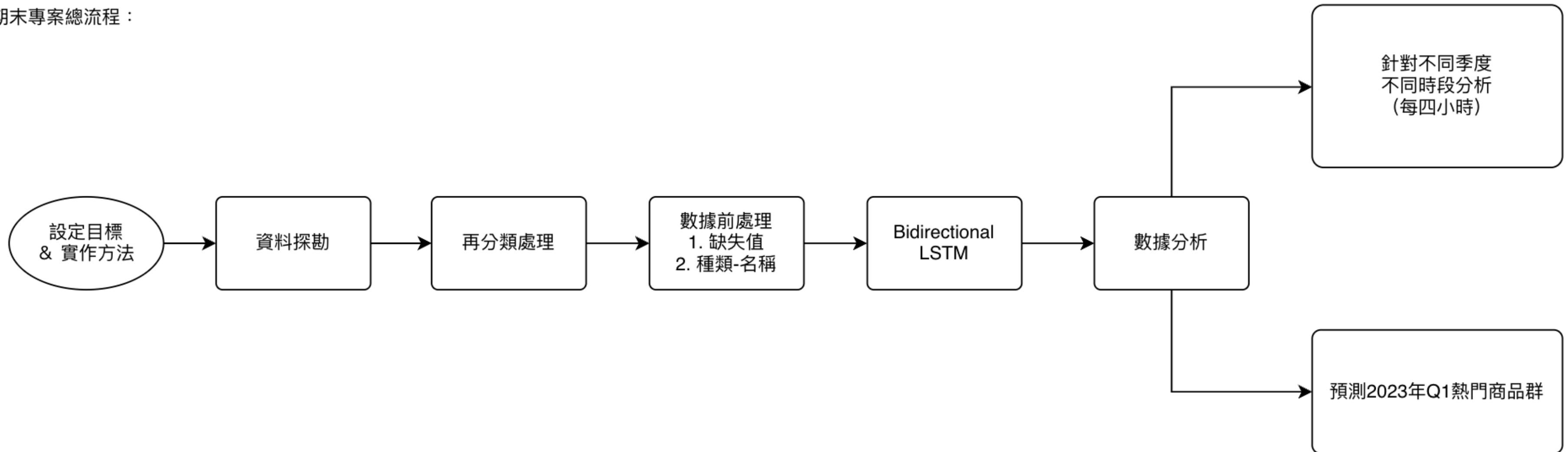
# 大綱

- 專案目標與架構
- 資料探勘與前處理
  - 資料總覽探索
  - 數據統計分析
  - 資料前處理與清洗
- 研究成果與結果分析
  - 時序分析的探勘與發現
  - 爆品預測結果與趨勢
- 數據洞見與建議

# 專案目標與架構

# 進行 2023 年 Q1~Q4 的爆品預測

期末專案總流程：

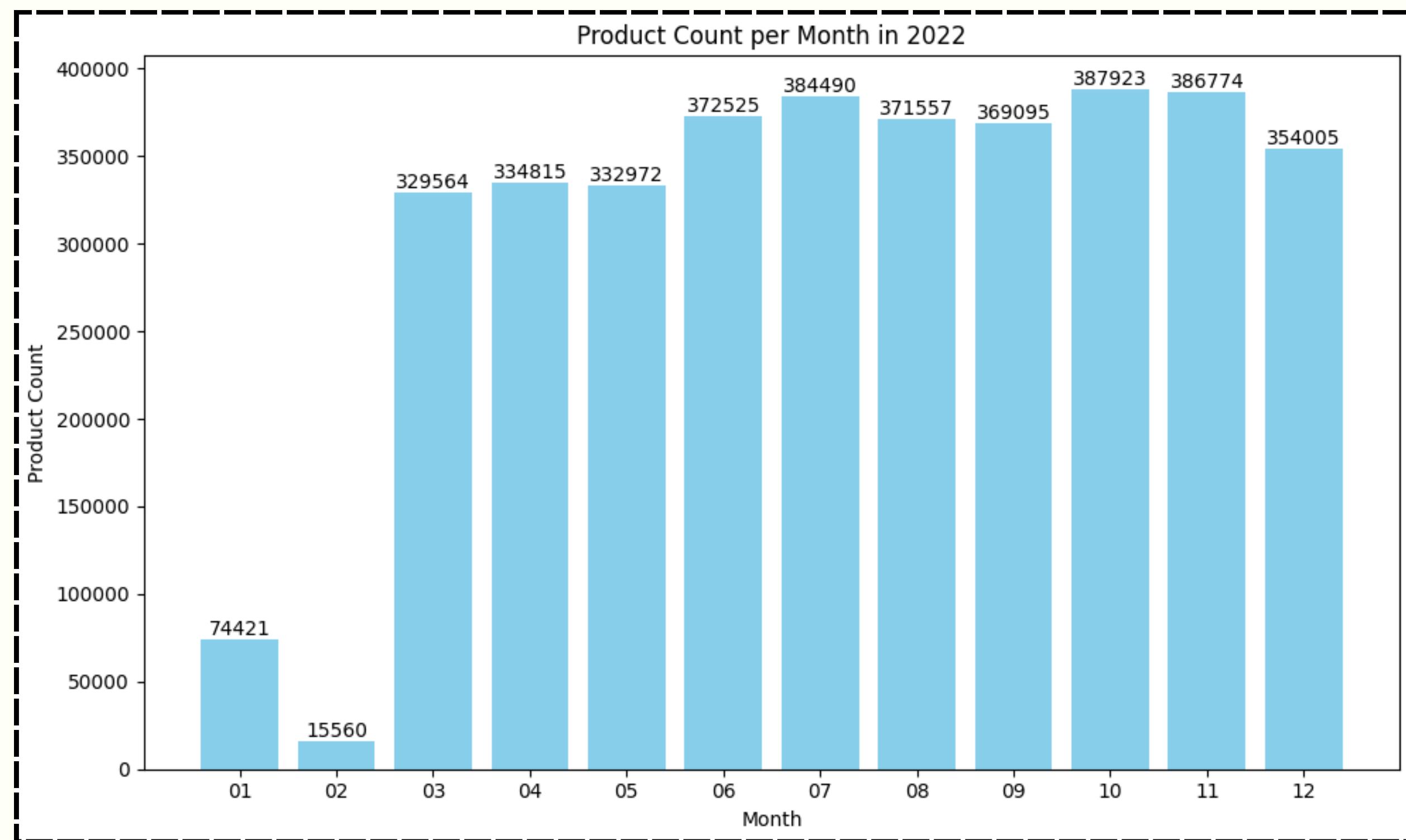


# 資料探勘與前處理

# 資料總覽

商品資料-2022\_XX  
XX = [01~12]

共有  
**3,713,701**  
商品資料



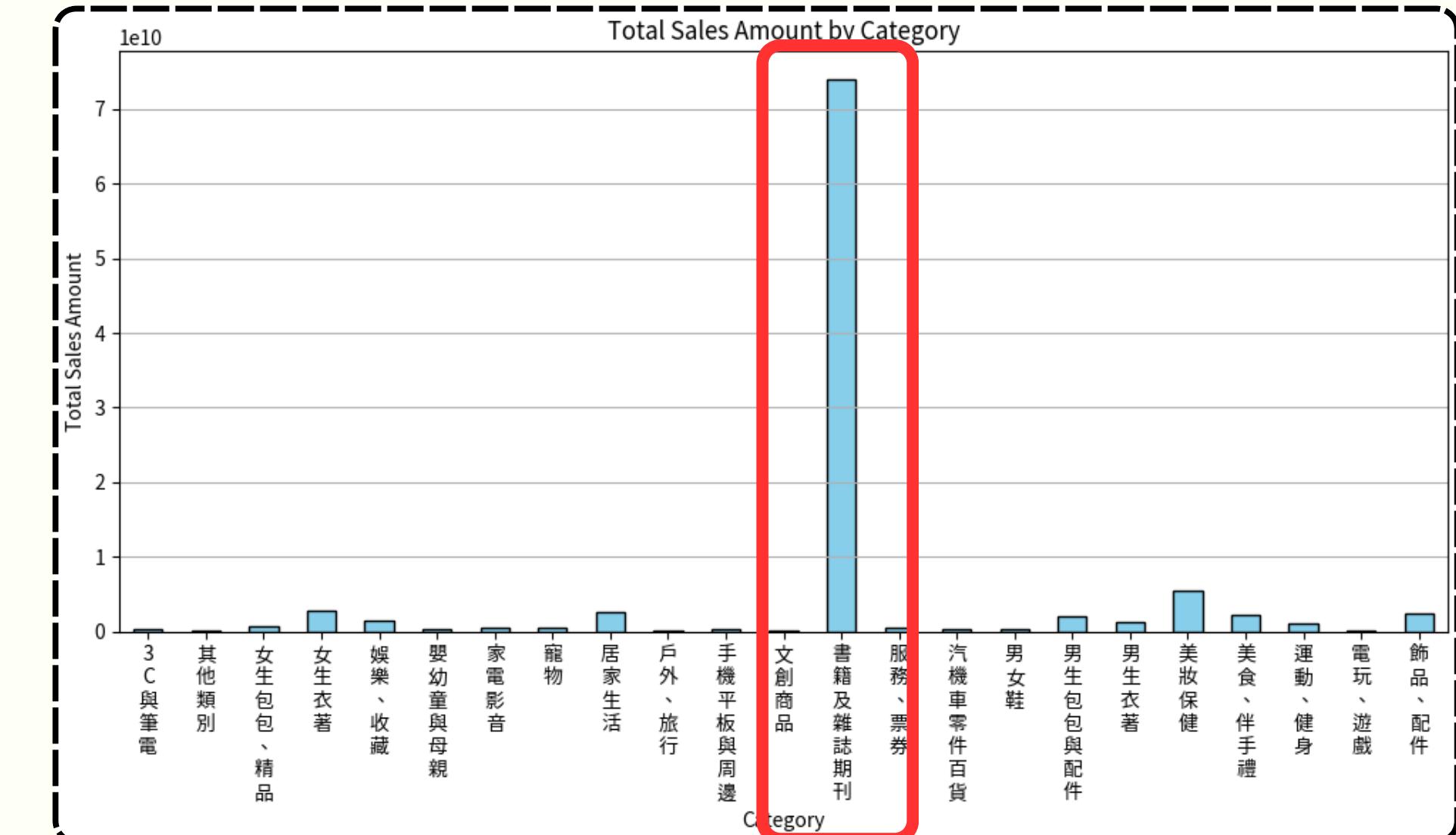
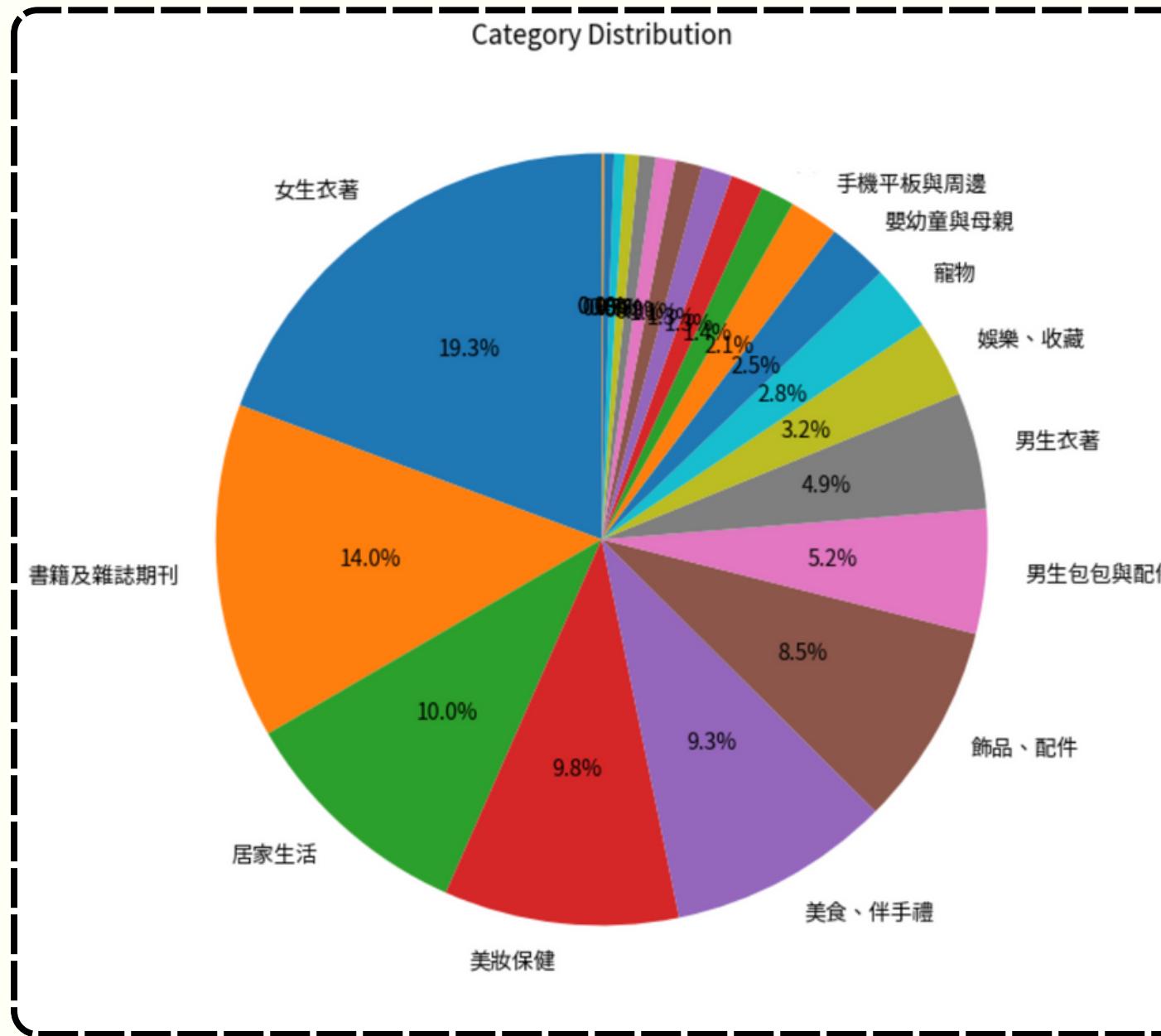
# 資料欄位介紹

欄位名稱	說明
id	交易的唯一識別編號。每筆交易都有一個獨一的 id。
time	交易發生的時間。格式為 “YYYY-MM-DD HH”
commodity_id	商品的唯一識別編號。每種商品都有一個獨特的 commodity_id。
category	商品的分類。格式為 “[‘類別名稱’]”

# 資料欄位介紹

欄位名稱	說明
name	商品的名稱。
price	商品的單價。
total_quantity	在這筆交易中購買的商品數量。
total_amount	交易的總金額。應等於 total_quantity 和 price 的乘積。

# 原始資料觀察



女生衣著      19.3%  
書籍及期刊雜誌      14.0%

# 重新分類書籍及雜誌期刊類

	<b>id</b>	<b>time</b>	<b>commodity_id</b>	<b>category</b>		<b>name</b>	<b>total_quantity</b>	<b>price</b>	<b>total_amount</b>
22	6693826	2022-03-01 08	5829911	書籍及雜誌期刊	第 008 標 (2/25)小誌-台式鹽酥雞 500g/包		12	90	1080
51	7807184	2022-03-24 09	5842286	書籍及雜誌期刊	第 048 標 90509BS 絶版黑鋒石爪鎖鎖式 (1/2)		1		
53	6920301	2022-03-06 11	5843348	書籍及雜誌期刊	022 新版探照燈		2		
157	7354539	2022-03-15 07	5876348	書籍及雜誌期刊	中藥包1包 \$150		3	150	450
178	6847611	2022-03-04 16	5883363	書籍及雜誌期刊	新油雞腿*1隻		10	262	2620
...	...	...	...	...	...		...	...	...
3484733	6657982	2022-02-28 13	6491492	書籍及雜誌期刊	32.WE83-DISNEY迪士尼超大環保購物袋(藍) 好看又可以一起愛地球只要199+1		12	199	2388
3484734	6658079	2022-02-28 13	6491542	書籍及雜誌期刊	33.WK99-拉斯維加斯限定超好看購物袋(按讚分享標) 關鍵字：2+1		2	0	0
3484768	6659677	2022-02-28 14	6492261	書籍及雜誌期刊	47.WJ83-GUESS春夏最新款滿版防刮黛妃包(粉) 0元起標一刀100 喜歡想要相信米...		2	2266	4532
3484832	6662251	2022-02-28 15	6493240	書籍及雜誌期刊	大師兄 0228 終極密碼-購物金1000元-一定要私訊		2	0	0
3484845	6663137	2022-02-28 15	6493648	書籍及雜誌期刊	2/28分享禮*1個(隨機出貨不可挑選)		19	0	0

「書籍及雜誌期刊」類別中有許多類別不合理的商品

022 新版探照燈  
中藥包1包 \$150  
新油雞腿 \*1隻

- 
- 
-

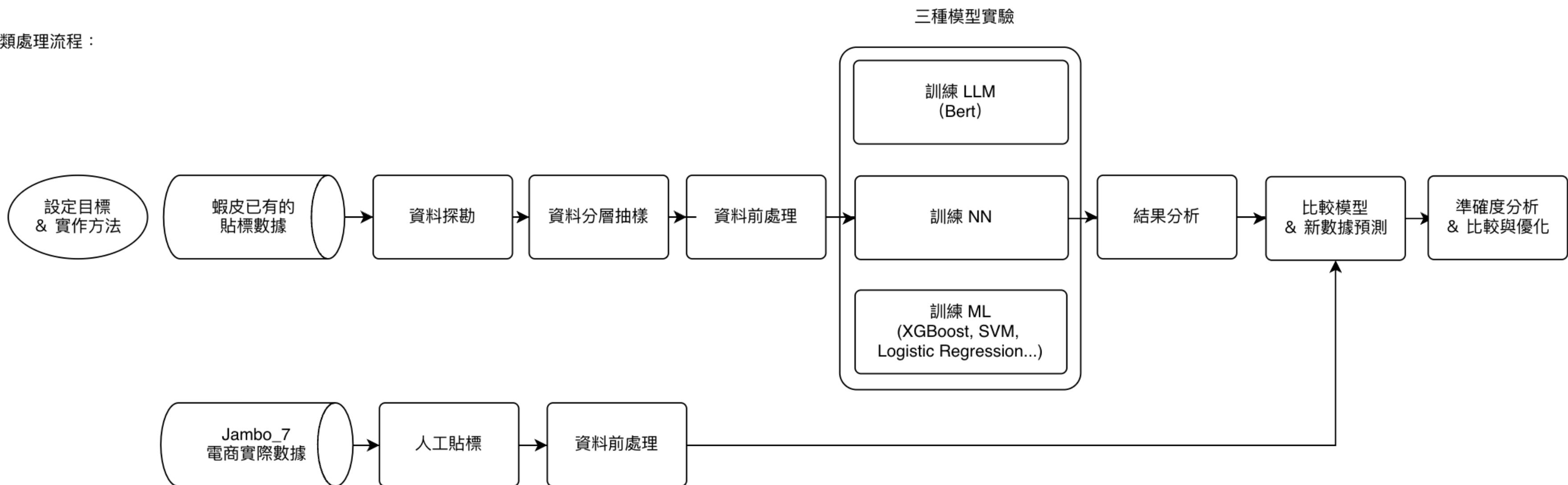
# 類別再分類

---

# 再分類 - 訓練流程介紹

小組使用包含25萬筆已貼上蝦皮標籤的訓練資料，訓練LLM進行商品分類。本次專案也利用該LLM，重新分類「書籍及雜誌期刊」，以獲得更精確的分析結果。

再分類處理流程：



# 再分類 - 訓練資料集介紹

## 一、shopee\_item\_category

資料數量為253335，並包含商品名稱、大類、小類。用於訓練模型。

	商品名稱	大類	小類	大小類合成
count	253335	253335	253335	253335
unique	253117	28	633	716
top	#ERROR! 書籍及雜誌期刊	其他	愛好與收藏品 - 公仔	
freq	5	23318	13822	5808

## 二、jambo\_7

資料數量為18752筆，並包含channel\_name、商品id、商品名稱、大類、小類。而大類與小類都是空的，需要進行預測填充。

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18752 entries, 0 to 18751
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   channel_name    18752 non-null   object 
 1   商品id          18752 non-null   int64  
 2   商品名稱        18752 non-null   object 
```

# 再分類 - 資料前處理

1. 缺失值檢查：最終檢查無NULL值
2. 中文的停用詞處理：去除掉文本中多餘的雜訊。
3. UNKNOWN 類別：於資料集中加入 UNKNOWN 類別，讓模型遇到不知所云的商品名稱也能進行分類。

# 再分類 - 資料前處理

根據「大類」和「小類」資料的比例分層採樣。

為讓訓練資料更符合真實分布，以及降低模型訓練時間，小組進行分層抽樣，並分為兩種訓練數據，分別為 **25K** 以及 **50K** 的兩份數據集。

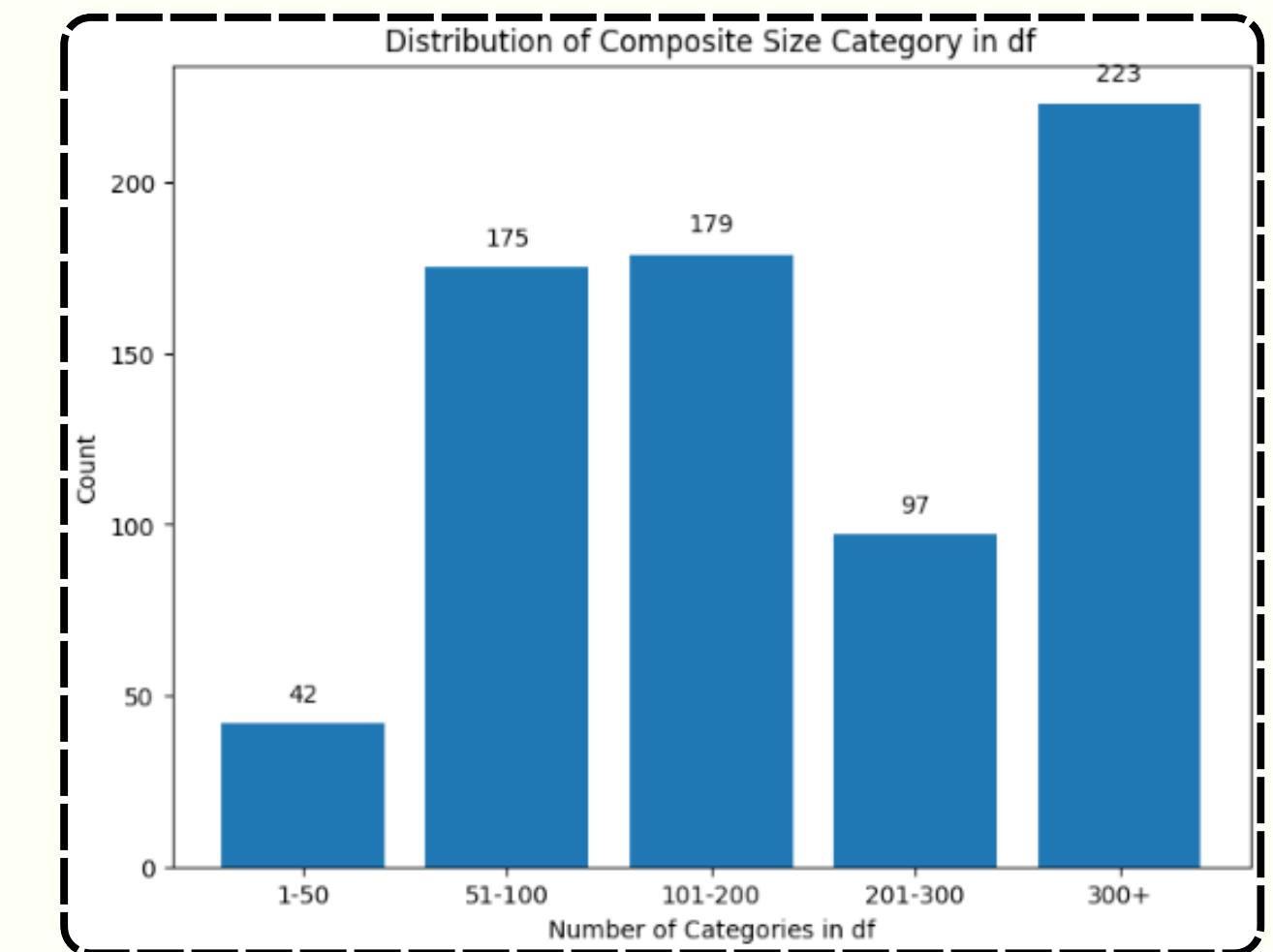
**DF\_25K**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25971 entries, 0 to 25970
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype  
--- 
 0   商品名稱    25971 non-null   object 
 1   大類        25971 non-null   object 
 2   小類        25971 non-null   object 
 3   大小類合成    25971 non-null   object 
dtypes: object(4)
memory usage: 811.7+ KB
None
```

**DF\_50K**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51244 entries, 0 to 51243
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype  
--- 
 0   商品名稱    51244 non-null   object 
 1   大類        51244 non-null   object 
 2   小類        51244 non-null   object 
 3   大小類合成    51244 non-null   object 
dtypes: object(4)
memory usage: 1.6+ MB
None
```

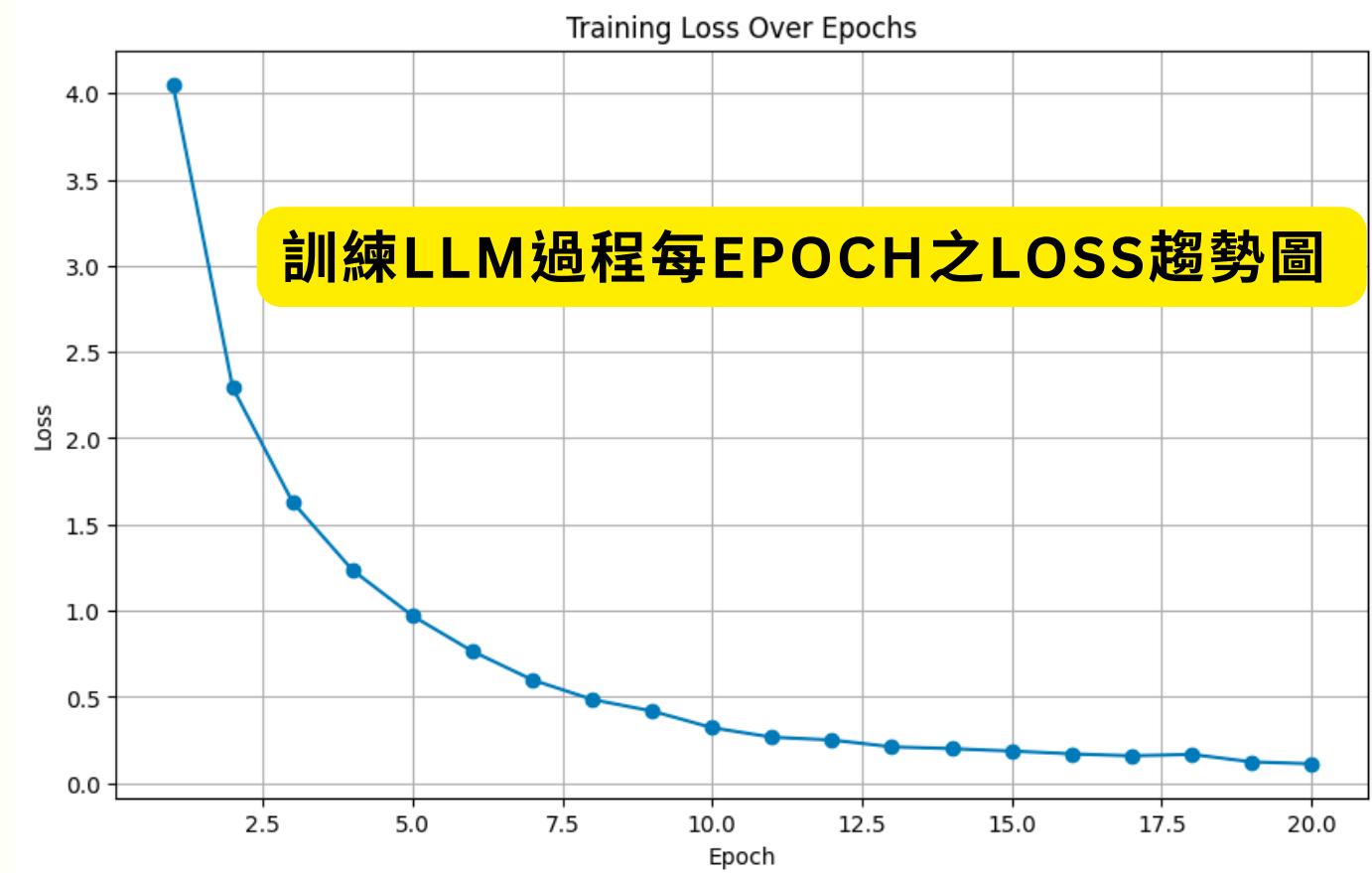
**原始資料類別數量分布**



# 再分類 - LLM 訓練成果

BERT為預訓練模型，進行文本分類準確度達**0.68**

這次訓練使用了**25K**和**50K**的資料。**25K**的資料量不僅確保了訓練的充足性，也避免了因為資料量龐大而增加訓練時間，最終的準確率為**0.68**。



訓練：

```
TOKENIZER = BERTTOKENIZER('BERT-BASE-CHINESE')
MODEL = BERTFORSEQUENCECLASSIFICATION('BERT-BASE-CHINESE')
OPTIMIZER = ADAMW(LR=5E-5)
LOSS_FN = CROSSENTROPYLOSS()
EPOCHS = 20
```

# 再分類 - 本次任務

將成果套用於「書籍及雜誌期刊」種類，進行再分類

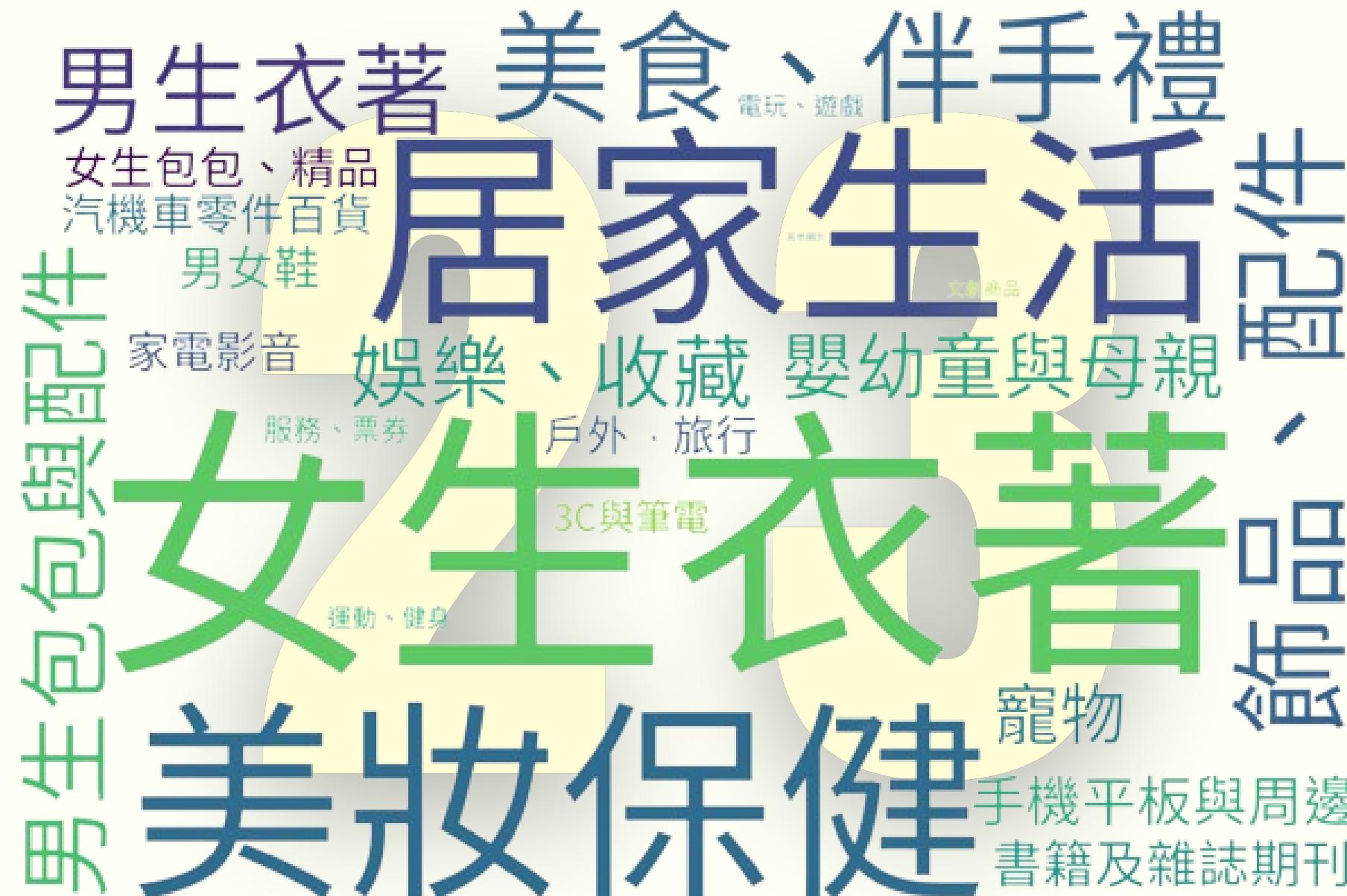
將原資料（種類為書籍及雜誌期刊）中的商品名稱欄位，僅保留中文名稱後，利用上次訓練完成的 **LLM** 模型進行分類，並取出「大類」作為新的種類。

舊種類 → LLM 預測出的新種類

C	D	E
commodity_id	category	name
5931051	書籍及雜誌期刊	胸愛心厚棉T
5936143	書籍及雜誌期刊	夢長野櫻紅
5936577	書籍及雜誌期刊	越南外海船凍土魠魚
5942050	書籍及雜誌期刊	Z2020 維尼小夜燈
5942319	書籍及雜誌期刊	1號秋刀魚(5入)
5955250	書籍及雜誌期刊	063 - 正韓挺版褲
5992117	書籍及雜誌期刊	K354 韓國身材管理錠 1罐 \$470 (1天 2餐飯前)
5999052	書籍及雜誌期刊	754 DavidsTEA good morning 早安茶包組合 20入 (有五種有機茶包)
5999054	書籍及雜誌期刊	755 DavidsTEA cold survival 感冒救助茶包組合 20入 (有五種有機茶包)

F	G
大類	小類
女生衣著	扭肩帶 / 繞頸背心
書籍及雜誌期刊	經典文學及古典/
美食、伴手禮	魚類
居家生活	登具
美食、伴手禮	海鮮零食
男生衣著	長褲
保健	幾能性食品
美食、伴手禮	茶葉、茶包
美食、伴手禮	茶葉、茶包

# 再分類 - 任務成果



商品總數 TOP 5

女生衣著	21.42%
居家生活	10.79%
美妆保健	10.76%
美食、伴手禮	10.02%
飾品、配件	9.90%

商品總數 BOTTOM 3

運動、健身	0.40%
文創商品	0.20%
其他類別	0.01%

# 資料前處理

---

合併十二份商品資料

對照再分類的  
書籍雜誌類

去除所有  
無意義資料

依照月份時段  
計算總銷售額

Name	Last modified	File size
商品資料-2022_01.xlsx	Dec 20, 2023 11:00	4.1 MB
商品資料-2022_02.xlsx	Dec 20, 2023 11:00	1.6 MB
商品資料-2022_03.xlsx	Dec 20, 2023 11:00	1.6 MB
商品資料-2022_04.xlsx	Dec 20, 2023 11:00	1.6 MB
商品資料-2022_05.xlsx	Dec 20, 2023 11:00	1.6 MB
商品資料-2022_06.xlsx	Dec 20, 2023 11:00	1.6 MB
商品資料-2022_07.xlsx	Dec 20, 2023 11:00	21.8 MB
商品資料-2022_08.xlsx	Dec 20, 2023 11:00	21.8 MB
商品資料-2022_09.xlsx	Dec 20, 2023 11:00	21.8 MB
商品資料-2022_10.xlsx	Dec 20, 2023 11:00	21.8 MB
商品資料-2022_11.xlsx	Dec 20, 2023 11:00	21.8 MB
商品資料-2022_12.xlsx	Dec 20, 2023 11:00	21.8 MB

商品資料-2022\_all.csv

ID	Name	Category	Price
00000001_2022-01-01-00	書籍-文學小說	小說	100
00000002_2022-01-01-00	書籍-歷史傳記	歷史	150
00000003_2022-01-01-00	書籍-哲學思想	哲學	200
00000004_2022-01-01-00	書籍-科學技術	科學	250
00000005_2022-01-01-00	書籍-外語	外語	300
00000006_2022-01-01-00	書籍-漫畫	漫畫	350
00000007_2022-01-01-00	書籍-經典名著	經典	400
00000008_2022-01-01-00	書籍-旅行指南	旅遊	450
00000009_2022-01-01-00	書籍-電影影評	影評	500
00000010_2022-01-01-00	書籍-生活哲理	生活	550
00000011_2022-01-01-00	書籍-政治軍事	軍事	600
00000012_2022-01-01-00	書籍-運動休閒	運動	650
00000013_2022-01-01-00	書籍-美術藝術	美術	700
00000014_2022-01-01-00	書籍-音樂	音樂	750
00000015_2022-01-01-00	書籍-電影影評	影評	800
00000016_2022-01-01-00	書籍-哲學思想	哲學	850
00000017_2022-01-01-00	書籍-文學小說	小說	900
00000018_2022-01-01-00	書籍-歷史傳記	歷史	950
00000019_2022-01-01-00	書籍-哲學思想	哲學	1000
00000020_2022-01-01-00	書籍-科學技術	科學	1050
00000021_2022-01-01-00	書籍-外語	外語	1100
00000022_2022-01-01-00	書籍-漫畫	漫畫	1150
00000023_2022-01-01-00	書籍-經典名著	經典	1200
00000024_2022-01-01-00	書籍-旅行指南	旅遊	1250
00000025_2022-01-01-00	書籍-政治軍事	軍事	1300
00000026_2022-01-01-00	書籍-運動休閒	運動	1350
00000027_2022-01-01-00	書籍-美術藝術	美術	1400
00000028_2022-01-01-00	書籍-音樂	音樂	1450
00000029_2022-01-01-00	書籍-電影影評	影評	1500
00000030_2022-01-01-00	書籍-哲學思想	哲學	1550
00000031_2022-01-01-00	書籍-文學小說	小說	1600
00000032_2022-01-01-00	書籍-歷史傳記	歷史	1650
00000033_2022-01-01-00	書籍-哲學思想	哲學	1700
00000034_2022-01-01-00	書籍-科學技術	科學	1750
00000035_2022-01-01-00	書籍-外語	外語	1800
00000036_2022-01-01-00	書籍-漫畫	漫畫	1850
00000037_2022-01-01-00	書籍-經典名著	經典	1900
00000038_2022-01-01-00	書籍-旅行指南	旅遊	1950
00000039_2022-01-01-00	書籍-政治軍事	軍事	2000
00000040_2022-01-01-00	書籍-運動休閒	運動	2050
00000041_2022-01-01-00	書籍-美術藝術	美術	2100
00000042_2022-01-01-00	書籍-音樂	音樂	2150
00000043_2022-01-01-00	書籍-電影影評	影評	2200
00000044_2022-01-01-00	書籍-哲學思想	哲學	2250
00000045_2022-01-01-00	書籍-文學小說	小說	2300
00000046_2022-01-01-00	書籍-歷史傳記	歷史	2350
00000047_2022-01-01-00	書籍-哲學思想	哲學	2400
00000048_2022-01-01-00	書籍-科學技術	科學	2450
00000049_2022-01-01-00	書籍-外語	外語	2500
00000050_2022-01-01-00	書籍-漫畫	漫畫	2550
00000051_2022-01-01-00	書籍-經典名著	經典	2600
00000052_2022-01-01-00	書籍-旅行指南	旅遊	2650
00000053_2022-01-01-00	書籍-政治軍事	軍事	2700
00000054_2022-01-01-00	書籍-運動休閒	運動	2750
00000055_2022-01-01-00	書籍-美術藝術	美術	2800
00000056_2022-01-01-00	書籍-音樂	音樂	2850
00000057_2022-01-01-00	書籍-電影影評	影評	2900
00000058_2022-01-01-00	書籍-哲學思想	哲學	2950
00000059_2022-01-01-00	書籍-文學小說	小說	3000
00000060_2022-01-01-00	書籍-歷史傳記	歷史	3050
00000061_2022-01-01-00	書籍-哲學思想	哲學	3100
00000062_2022-01-01-00	書籍-科學技術	科學	3150
00000063_2022-01-01-00	書籍-外語	外語	3200
00000064_2022-01-01-00	書籍-漫畫	漫畫	3250
00000065_2022-01-01-00	書籍-經典名著	經典	3300
00000066_2022-01-01-00	書籍-旅行指南	旅遊	3350
00000067_2022-01-01-00	書籍-政治軍事	軍事	3400
00000068_2022-01-01-00	書籍-運動休閒	運動	3450
00000069_2022-01-01-00	書籍-美術藝術	美術	3500
00000070_2022-01-01-00	書籍-音樂	音樂	3550
00000071_2022-01-01-00	書籍-電影影評	影評	3600
00000072_2022-01-01-00	書籍-哲學思想	哲學	3650
00000073_2022-01-01-00	書籍-文學小說	小說	3700
00000074_2022-01-01-00	書籍-歷史傳記	歷史	3750
00000075_2022-01-01-00	書籍-哲學思想	哲學	3800
00000076_2022-01-01-00	書籍-科學技術	科學	3850
00000077_2022-01-01-00	書籍-外語	外語	3900
00000078_2022-01-01-00	書籍-漫畫	漫畫	3950
00000079_2022-01-01-00	書籍-經典名著	經典	4000
00000080_2022-01-01-00	書籍-旅行指南	旅遊	4050
00000081_2022-01-01-00	書籍-政治軍事	軍事	4100
00000082_2022-01-01-00	書籍-運動休閒	運動	4150
00000083_2022-01-01-00	書籍-美術藝術	美術	4200
00000084_2022-01-01-00	書籍-音樂	音樂	4250
00000085_2022-01-01-00	書籍-電影影評	影評	4300
00000086_2022-01-01-00	書籍-哲學思想	哲學	4350
00000087_2022-01-01-00	書籍-文學小說	小說	4400
00000088_2022-01-01-00	書籍-歷史傳記	歷史	4450
00000089_2022-01-01-00	書籍-哲學思想	哲學	4500
00000090_2022-01-01-00	書籍-科學技術	科學	4550
00000091_2022-01-01-00	書籍-外語	外語	4600
00000092_2022-01-01-00	書籍-漫畫	漫畫	4650
00000093_2022-01-01-00	書籍-經典名著	經典	4700
00000094_2022-01-01-00	書籍-旅行指南	旅遊	4750
00000095_2022-01-01-00	書籍-政治軍事	軍事	4800
00000096_2022-01-01-00	書籍-運動休閒	運動	4850
00000097_2022-01-01-00	書籍-美術藝術	美術	4900
00000098_2022-01-01-00	書籍-音樂	音樂	4950
00000099_2022-01-01-00	書籍-電影影評	影評	5000
00000100_2022-01-01-00	書籍-哲學思想	哲學	5050
00000101_2022-01-01-00	書籍-文學小說	小說	5100
00000102_2022-01-01-00	書籍-歷史傳記	歷史	5150
00000103_2022-01-01-00	書籍-哲學思想	哲學	5200
00000104_2022-01-01-00	書籍-科學技術	科學	5250
00000105_2022-01-01-00	書籍-外語	外語	5300
00000106_2022-01-01-00	書籍-漫畫	漫畫	5350
00000107_2022-01-01-00	書籍-經典名著	經典	5400
00000108_2022-01-01-00	書籍-旅行指南	旅遊	5450
00000109_2022-01-01-00	書籍-政治軍事	軍事	5500
00000110_2022-01-01-00	書籍-運動休閒	運動	5550
00000111_2022-01-01-00	書籍-美術藝術	美術	5600
00000112_2022-01-01-00	書籍-音樂	音樂	5650
00000113_2022-01-01-00	書籍-電影影評	影評	5700
00000114_2022-01-01-00	書籍-哲學思想	哲學	5750
00000115_2022-01-01-00	書籍-文學小說	小說	5800
00000116_2022-01-01-00	書籍-歷史傳記	歷史	5850
00000117_2			

# 合併十二份 商品資料

## 對照再分類的書籍雜誌類

# 去除所有 無意義資料

# 依 照 月 份 時 段 計 算 總 銷 售 額

22	6693826	2022-03-01 08	5829911	書籍及雜誌期 刊	第 008 標 (2/25)小誌-台式鹽酥雞 500g/包			12	90	1080	手錶	男錶	
51	7807184	2022-03-24 09	5842286	書籍及雜誌期 刊	第 048 標 90509BS 絶版黑鋸石爪鑲鎖式(1/2)			1	150	150	寵物	牽繩項圈類	
53	6920301	2022-03-06 11	5843348	書籍及雜誌期 刊	022 新版探照燈			2	299	598	居家生活	燈具	
157	7354539	2022-03-15 07	5876348	書籍及雜誌期 刊	中藥包1包 \$150			3	150	450	旅行相關用品/ 行李箱	旅行收納包	
178	6847611	2022-03-04 16	5883363	書籍及雜誌期 刊	新油雞腿*1隻			10	262	2620	美食、伴手禮	雞肉	
...	...	...	...	...	...			...	...	...	...	...	
3484733	6657982	2022-02-28 13	6491492	書籍及雜誌期 刊	32.WE83-DISNEY迪士尼超大 物袋(藍)好看又可以 一起愛地球只要199+1			12	199	2388	居家生活	垃圾袋	
3484734	6658079	2022-02-28 13	6491542	書籍及雜誌期 刊	33.WK99-拉斯維加斯限定超好 袋(按讚分享標)關鍵 字 : 2+1			2	0	0	女生包包/精品	手提包	
3484768	6659677	2022-02-28 14	6492261	書籍及雜誌期 刊	47.WJ83-GUESS春夏最新款滿版防刮黛妃包(粉) 0元起標 一刀100 喜歡想要相信米...			2	2266	4532	書籍及雜誌期 刊	生活風格書籍	
3484832	6662251	2022-02-28 15	6493240	書籍及雜誌期 刊	大師兄 0228 終極密碼-購物金1000元-一定要私訊			2	0	0	書籍及雜誌期 刊	經典文學及古 典小說	
3484845	6663137	2022-02-28 15	6493648	書籍及雜誌期 刊	2/28分享禮*1個(隨機出貨不可挑選)			19	0	0	美食、伴手禮	糕餅點心	

合併十二份  
商品資料

對照再分類的  
書籍雜誌類

去除所有無意義資料

依照月份時段  
計算總銷售額

LDWAFFLE SU DJ4877300  
1992-2202 SW 5414410 S  
6 CD CEWE  
77601 (4/29)  
8813 QUẦN NAM  
第 012 標 (6/29) DO TUOI DE KHI  
LDWAFFLE SU DJ4877300  
1703-2202 RB1972-7504  
NIKE WAFFLE ONE-DA7995 100  
第 076 標 (4/29) #Y1426  
M16.BODY ĐÈM3  
35  
K144  
1723 DÉP X TRẮNG  
SC SHUI TONG BAO ĐEN  
116 1229 (4/28)

不合理價格

**86,755**

無意義名稱

**498,400**

剩餘資料

**3,215,301**

合併十二份  
商品資料

對照再分類的  
書籍雜誌類

去除所有  
無意義資料

依照月份時段計算總銷售額

## 資料集欄位 TIME

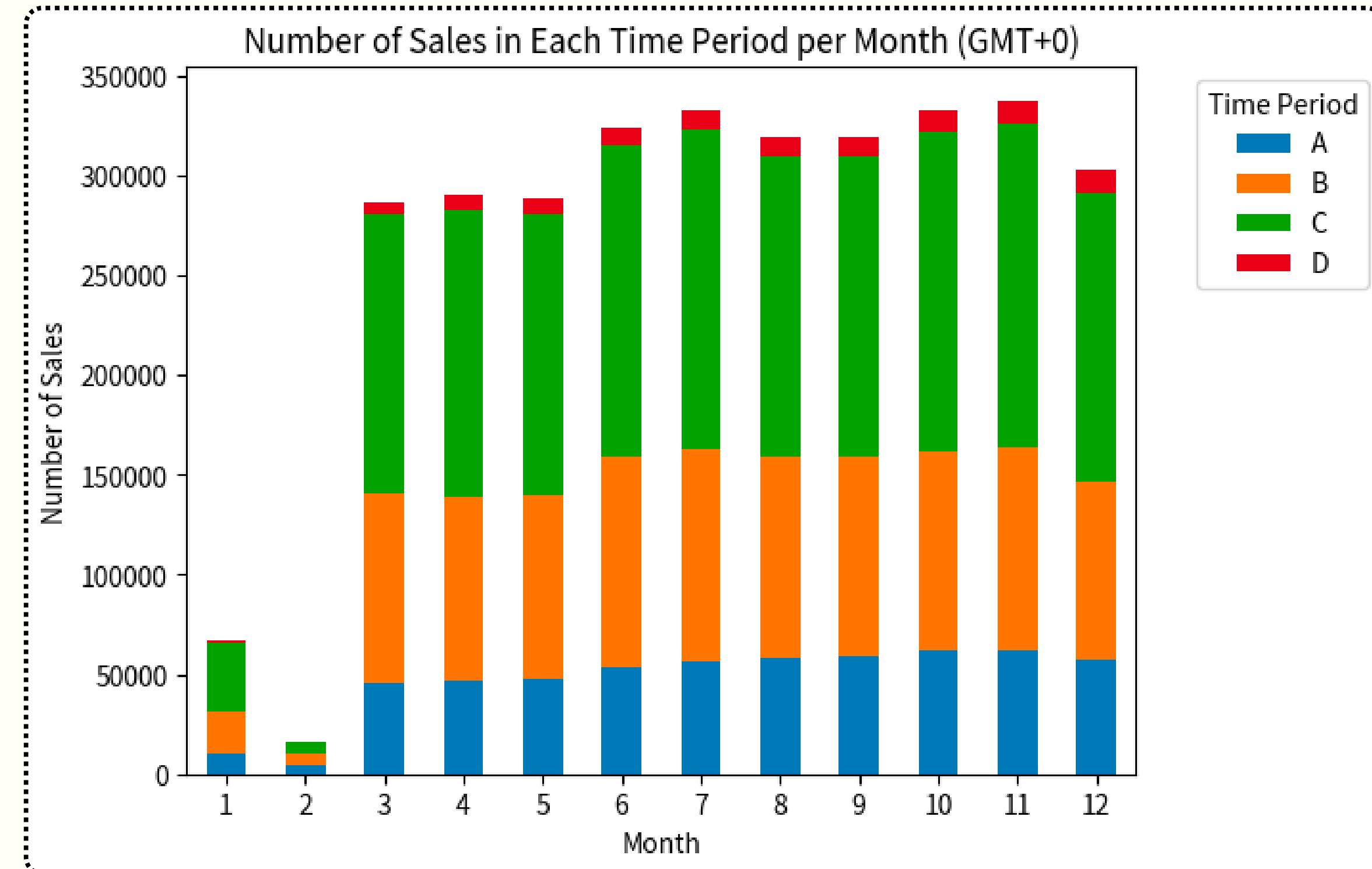
(推測時區為 GMT + 0 )

2022-01-05 21  
2022-02-25 15  
2022-03-18 04

.

.

01~06 → A 時段  
07~12 → B 時段  
13~18 → C 時段  
19~00 → D 時段



合併十二份  
商品資料

對照再分類的  
書籍雜誌類

去除所有  
無意義資料

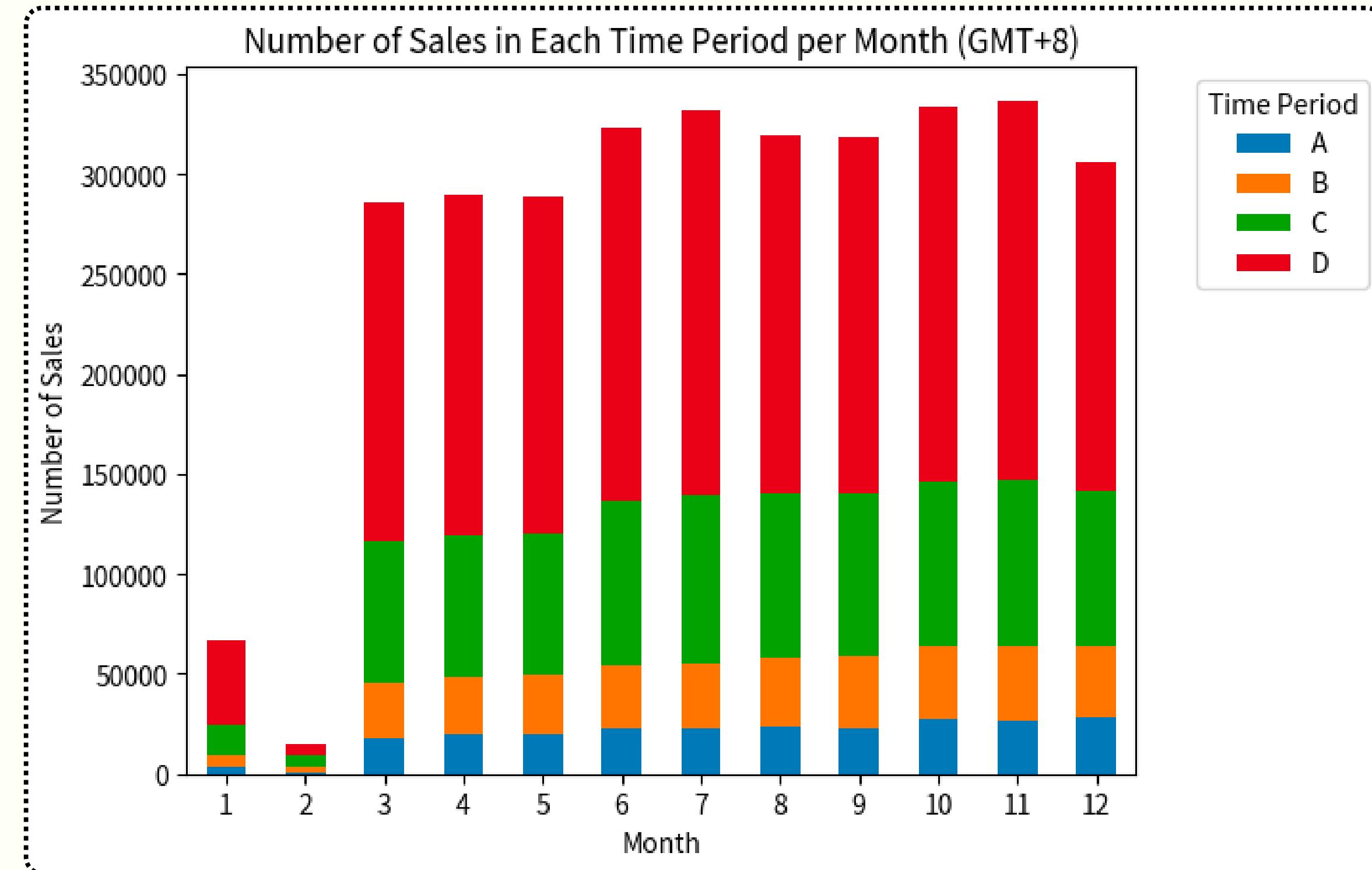
依照月份時段計算總銷售額

## 資料集欄位 TIME

(推測時區為 GMT+0)

2022-01-05	21
2022-02-25	15
2022-03-18	04

01~06 → A 時段  
07~12 → B 時段  
13~18 → C 時段  
19~00 → D 時段



合併十二份  
商品資料

對照再分類的  
書籍雜誌類

去除所有  
無意義資料

依照月份時段計算總銷售額

12個月份 X 4個時段



23  
個  
類  
別

	1A	1B	1C	1D	2A	2B	2C	2D	3A	3B	...	10C
運動、健身	815949	693000	762490	203741	462289	260599	288283	28083	1273165	2036963	...	6111212
戶外、旅行	767450	816696	1559003	17470	215687	269762	298520	850	1442851	3902786	...	6630783
美妝保健	23336299	43191066	52460807	230680	24425848	28808480	20155361	175748	68377019	85922506	...	381431715
男女鞋	1586898	1939246	2178948	247030	1438827	1139314	673865	38212	4498180	5740257	...	24763823
書籍及雜誌期刊	907884	1216705	2599271	35224	1907156	2619327	2460945	63738	2902745	5052841	...	12900592

# 時序分析預測模型 所需的前處理

# 行列轉換

為符合時序分析模型需求，我們將行列轉換，並把48個時段換成1~48的數字表示

time_points		書籍及雜誌期刊	美食、伴手禮	嬰幼童與母親	運動、健身	家電影音	美妝保健	其他類別	飾品、配件	手機平板與周邊	...	男生衣著	3C與筆電	服務、票券	戶外、旅行
0	1	153260	4003524	195038	40629	2697205	4714161	10782	2716677	321704	...	826949	1337692	10292	54889
1	2	254754	1550611	2040072	899457	1539139	14323238	0	6530340	706898	...	7526794	408633	0	573046
2	3	1033464	11988327	2140208	285787	2267156	31478911	4589	5727800	1721200	...	8586338	3248127	112667	451371
3	4	3311299	60221655	8228245	1249307	27221491	68702559	1227	17280946	5182561	...	19315684	4781110	743985	2081513
4	5	294581	504455	129608	5451	735388	2059604	0	183421	92319	...	58222	1280	3861	1280
5	6	1611421	2378925	756868	431521	1134416	14904762	0	1213122	615194	...	6632074	215161	4295	155280
6	7	1923956	4102485	1066292	238093	1867116	24942072	1797	2402272	1546676	...	6394996	386916	0	164996
7	8	3217424	15233428	1939201	313460	5235235	30693588	399	2363461	1194085	...	8237930	674475	4300	430948
8	9	828160	6373156	2017139	139402	5252615	27911115	2550	10102944	2672908	...	10401229	1270881	245512	317628
9	10	1406219	8290295	2444072	733536	6028297	39430692	3040	13507333	3923872	...	10812497	1883365	1476622	870102
10	11	3380705	20343582	7155449	1538069	11535591	77848639	16030	21843028	4556652	...	17351684	2913070	3035840	1599583
11	12	11505405	143374583	18179053	5396254	45531484	251910543	6967	114797948	19355139	...	68200748	13154185	15571671	8191526

# 切分資料集

將 48 筆資料以保有時序的方式進行切分，訓練集與測試集比為 8 : 2

TRAINING

TESTING



time\_points : 1 ~ 38

time\_points : 39 ~ 48

# 銷售額標準化

我們使用MinMaxScaler，將銷售額轉換成0~1之間的數值。

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{\min}}{x_{\max} - x_{\min}}$$

```
1 # 將categories進行標準化
2 scaler = MinMaxScaler(feature_range=(0, 1))
3 train_data_categories_scaled = scaler.fit_transform(train_data_categories)
4 test_data_categories_scaled = scaler.transform(test_data_categories)
```

# **BIDIRECTIONAL LSTM MODEL**

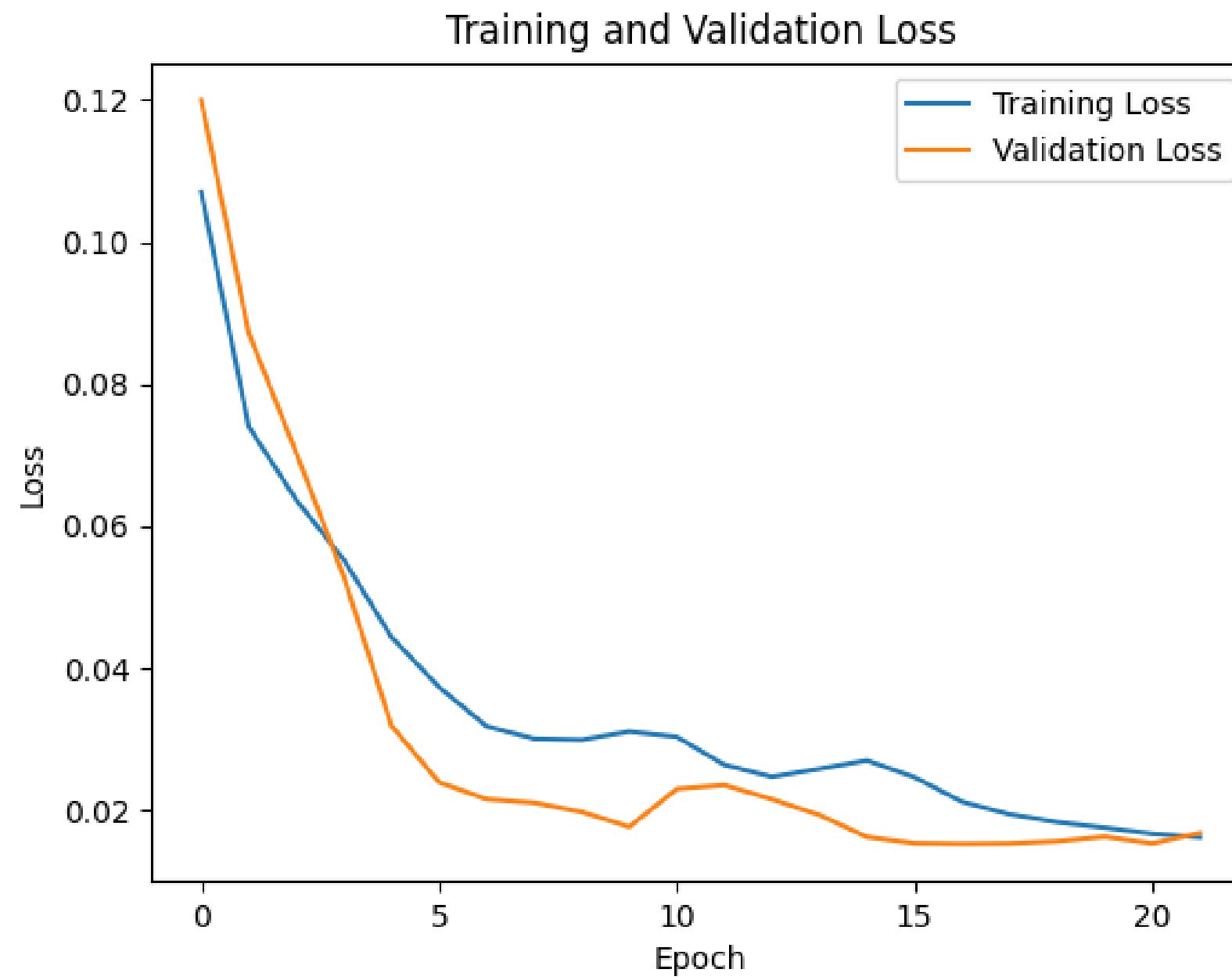
# 模型架構與超參數設定

```
1 # 設定time_steps以建立sequences  
2 time_steps = 4
```

```
1 # 建立模型  
2 model = Sequential()  
3 model.add(Bidirectional(LSTM(units=64, activation='relu', return_sequences=True), input_shape=(time_steps, train_data_categories.shape[1])))  
4 model.add(Bidirectional(LSTM(units=64, activation='relu', return_sequences=False)))  
5 model.add(Dense(train_data_categories.shape[1], activation='linear'))  
6 model.compile(optimizer='adam', loss='mse')
```

```
1 # 設定early_stopping機制  
2 early_stopping = EarlyStopping(monitor='val_loss', patience=5, restore_best_weights=True)  
3  
4 # 訓練  
5 history = model.fit(X_train, y_train, epochs=50, batch_size=1, validation_split=0.2, shuffle=False, callbacks=[early_stopping])
```

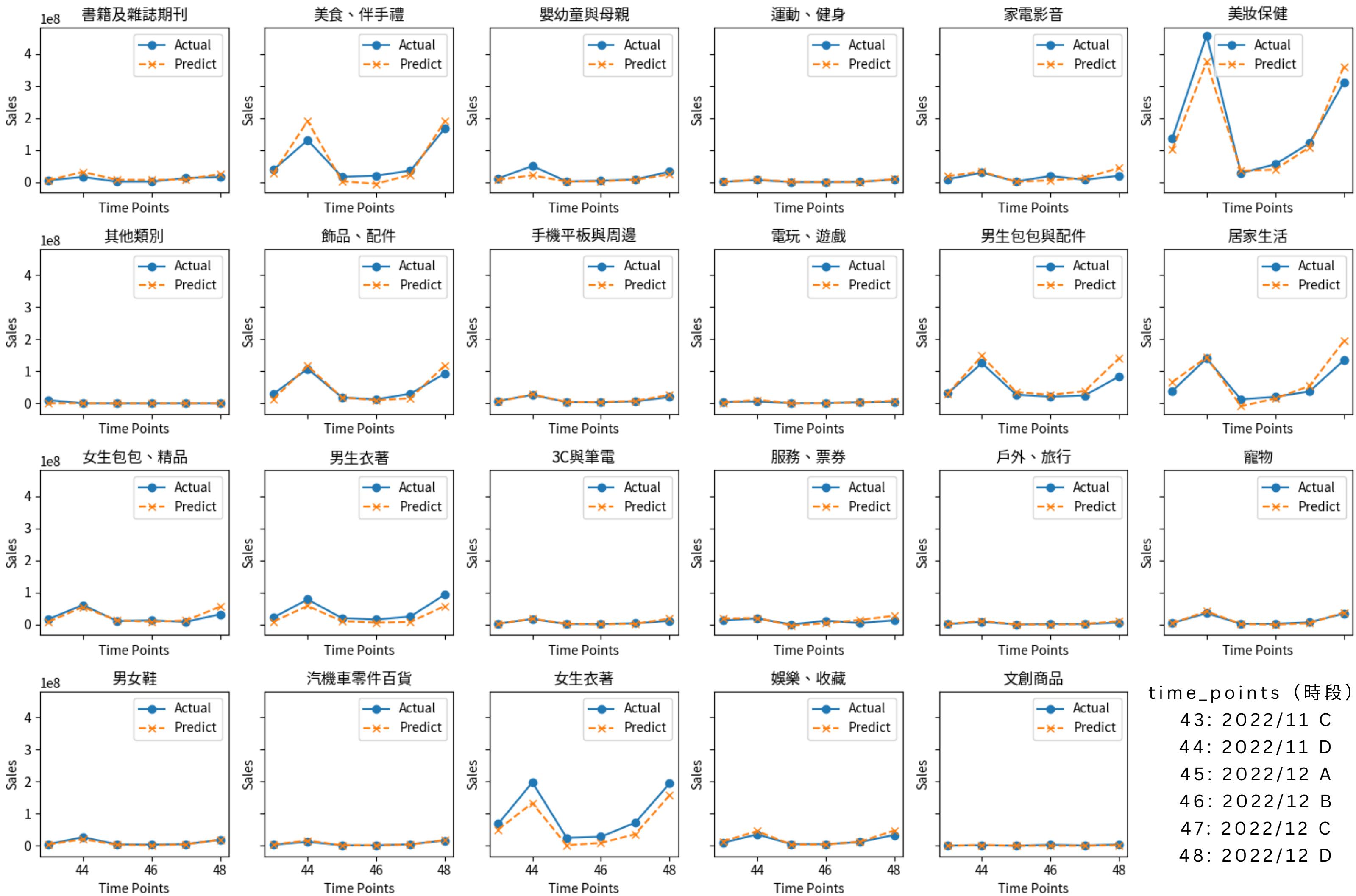
# 模型訓練過程



Epoch : 22/50

Training Loss : 0.0162

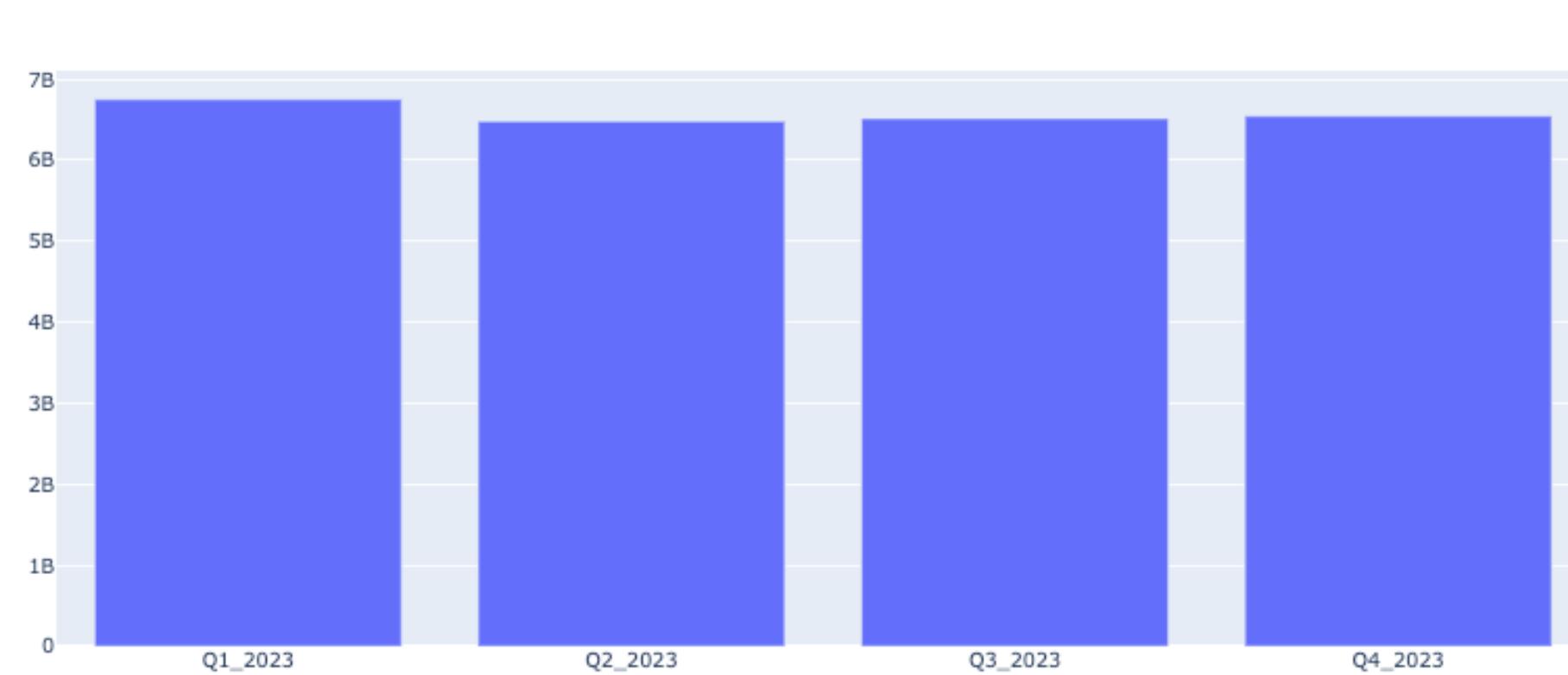
Validation Loss : 0.0168



# 預測結果分析

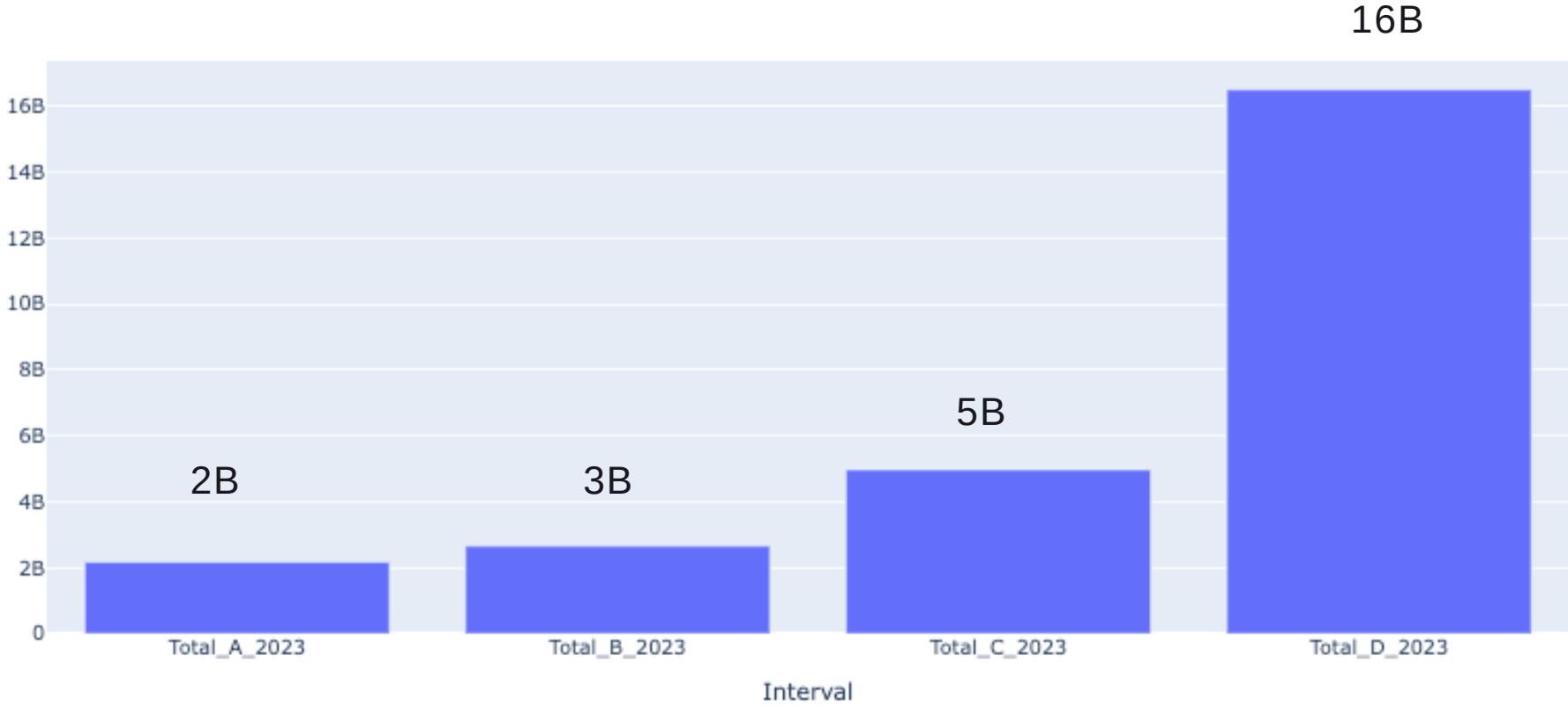
預測2023年不同季度與四大時段總銷售額比較

2023 每個季度的銷售總和



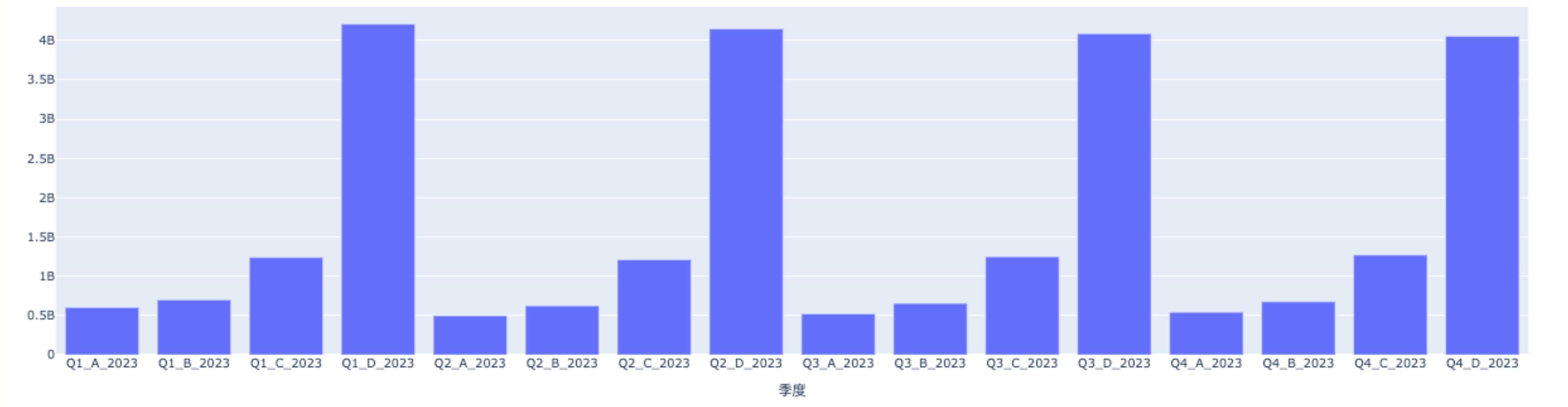
不同季度之預測總銷售額較為平均，2023 年無較強勢季度存在。

2023 每天四個時段的銷售對比



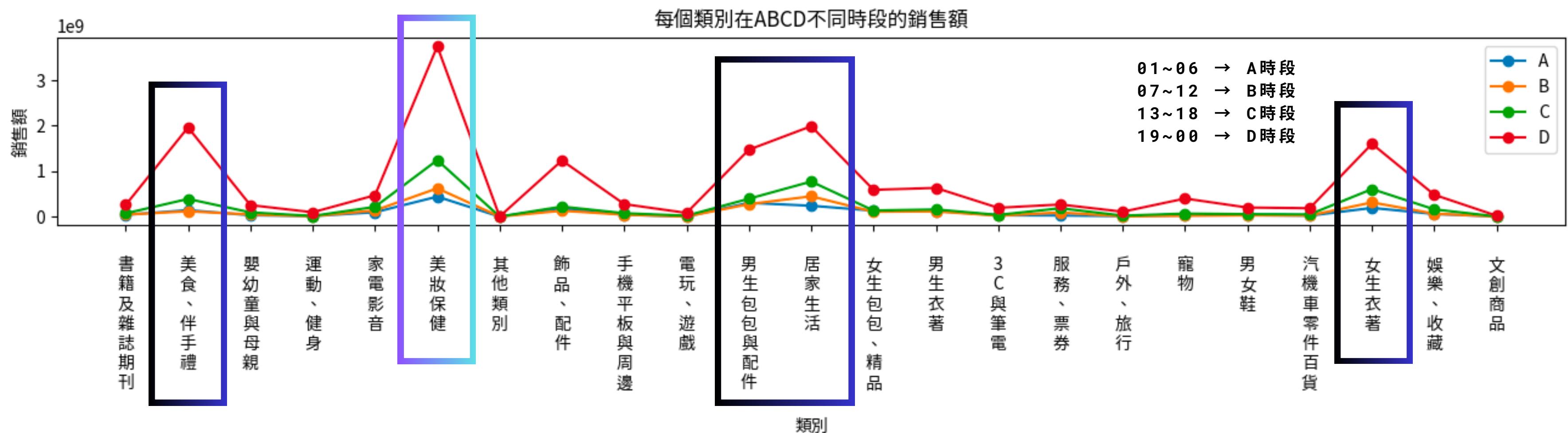
預測結果顯示D時段（19~00）為平台最熱門銷售時段，並且比起第二名時段銷售總額差超過三倍以上，建議平台能多加強該黃金時段，或是確保該時段系統會湧入高流量的預期。

# 預測結果分析 2023年季度與時段的預測銷售額綜合分析



1. D 時段於每個季度皆為平台最熱門時段
2. 同一時段於不同季度的銷量大略持平，無較特殊的變化
3. 銷售排行： $D > C > B > A$  時段

# 預測結果分析 預測時段銷售排行，美妝保健為全時段銷售王

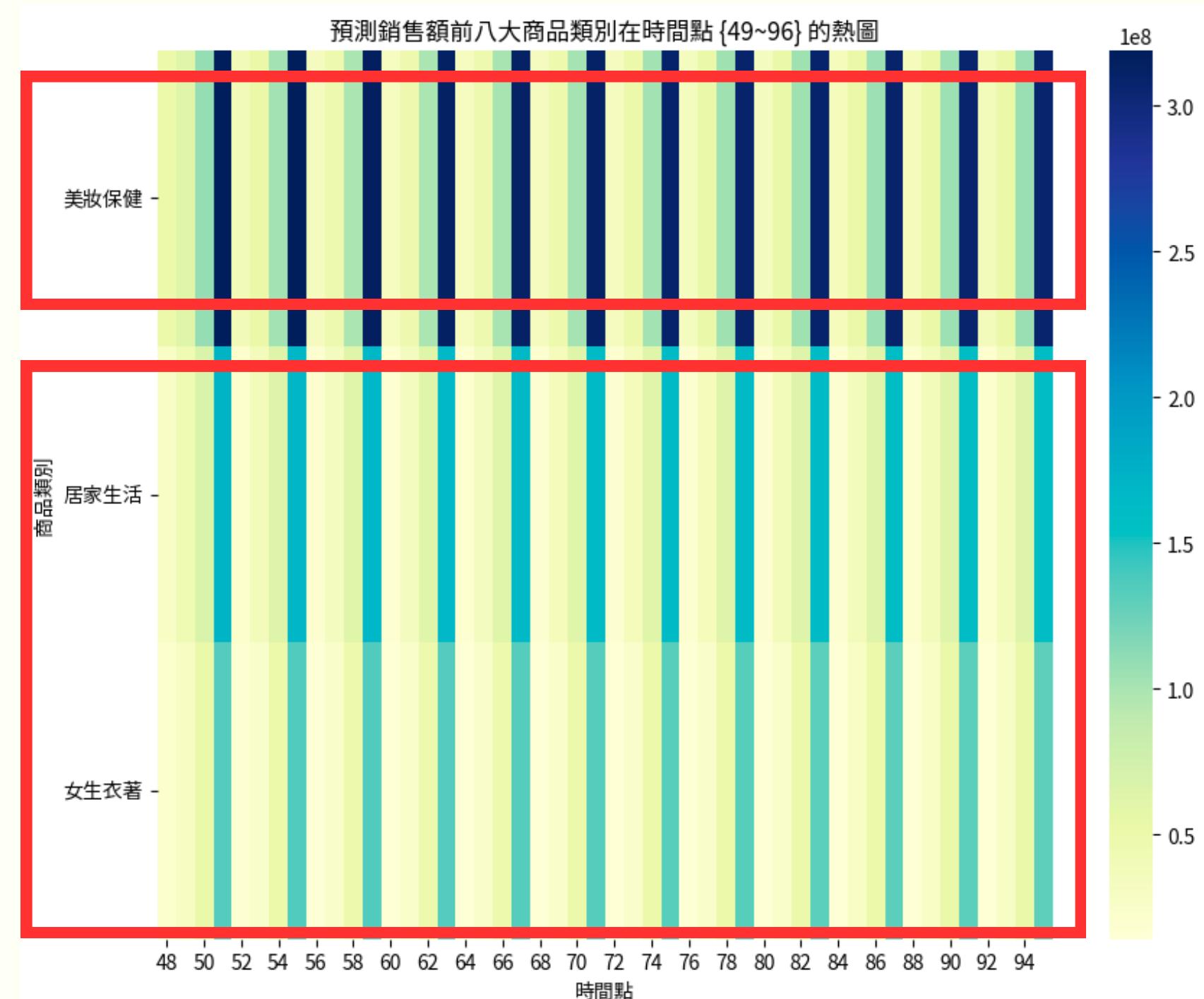


美妝保健於所有時段皆為銷售冠軍。

家電影音、男生包包與配件、居家生活、女生衣著類別中銷售額於不同時段銷售表現有較顯著的差異，尤其於D時段最為熱賣。

# 預測結果分析 2023年每個時間點銷售最高皆為「美妝保健」

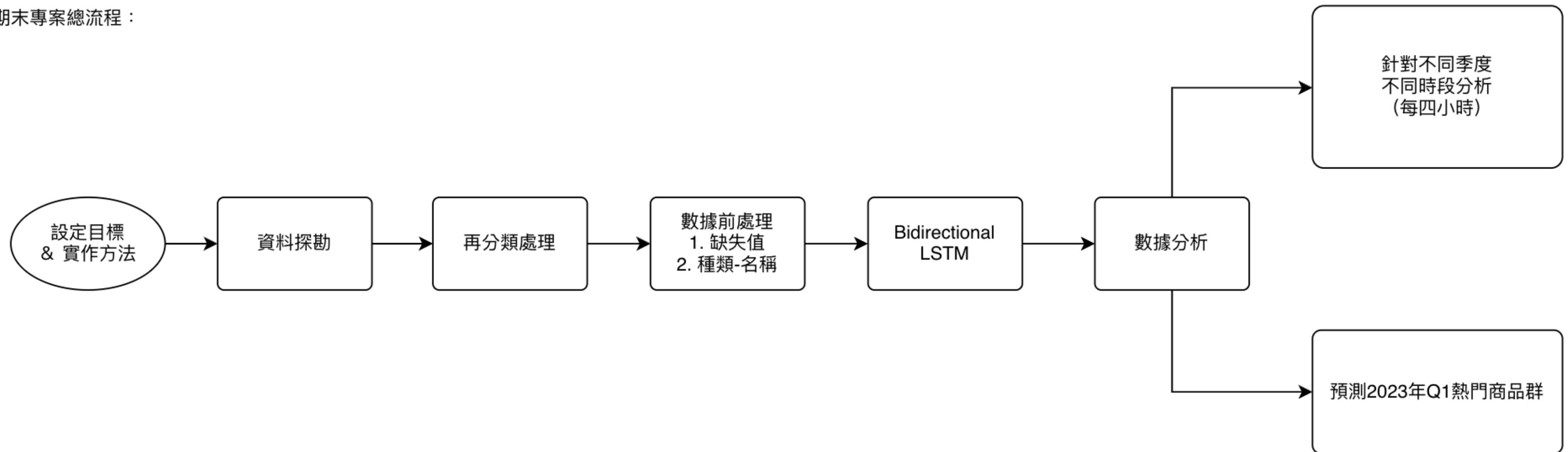
- 美妝保健於不只於Q1為最佳銷售產品，2023年整個年度從預測來看都擁有最高銷售額。
- 除「美妝保健」外，不容忽視的產品群為「居家生活」、「女生衣著」同樣於2023年可能會有不錯的表現。



# 數據洞見與建議

# RECAP：實作再分類與雙向LSTM模型，進行 2023 年 Q1~Q4 的爆品預測

期末專案總流程：



# DATA INSIGHTS

1. 2023年最熱門時段仍為D時段（19-00）
2. 最熱門商品為「美妝保健」類產品。「居家生活」、「女生衣著」則可能會有不錯的表現。

# FUTURE WORKS

1. 取得「顧客類」與「店家類」資料，能夠結合商品名稱進行更多元的購物籃分析，或是相關統計。
2. 硬體運算設備升級，加速訓練工作：希望能買COLAB PRO

# SUGGESTION

1. 聚焦預測出的熱門時段與產品，進行相關商業策略與系統策略制定。
2. 升級產品分類器以提高精確度和擴充類別，便於未來數據分析。
3. 資料紀錄準確性問題：例如商品名稱、價格和數量等欄位經常出現異常資料。建議未來提醒直播主確實輸入商品名稱，或擴充資料庫欄位以存儲直播主所需的更多資料，例如商品編號等。

# 分工

---

# 團隊分工

蔡品洋：資料探勘分析、資料前處理、簡報

羅永富：時序分析預測模型所需的前處理與模型建置、時序分析預測、簡報

葉瀚元：資料前處理、再分類模型、預測結果分析、簡報