網路爬梳在消費者物價調查的應用與限制

報告人：周治良
行政院主計總處
綜合統計處
110年3月17日

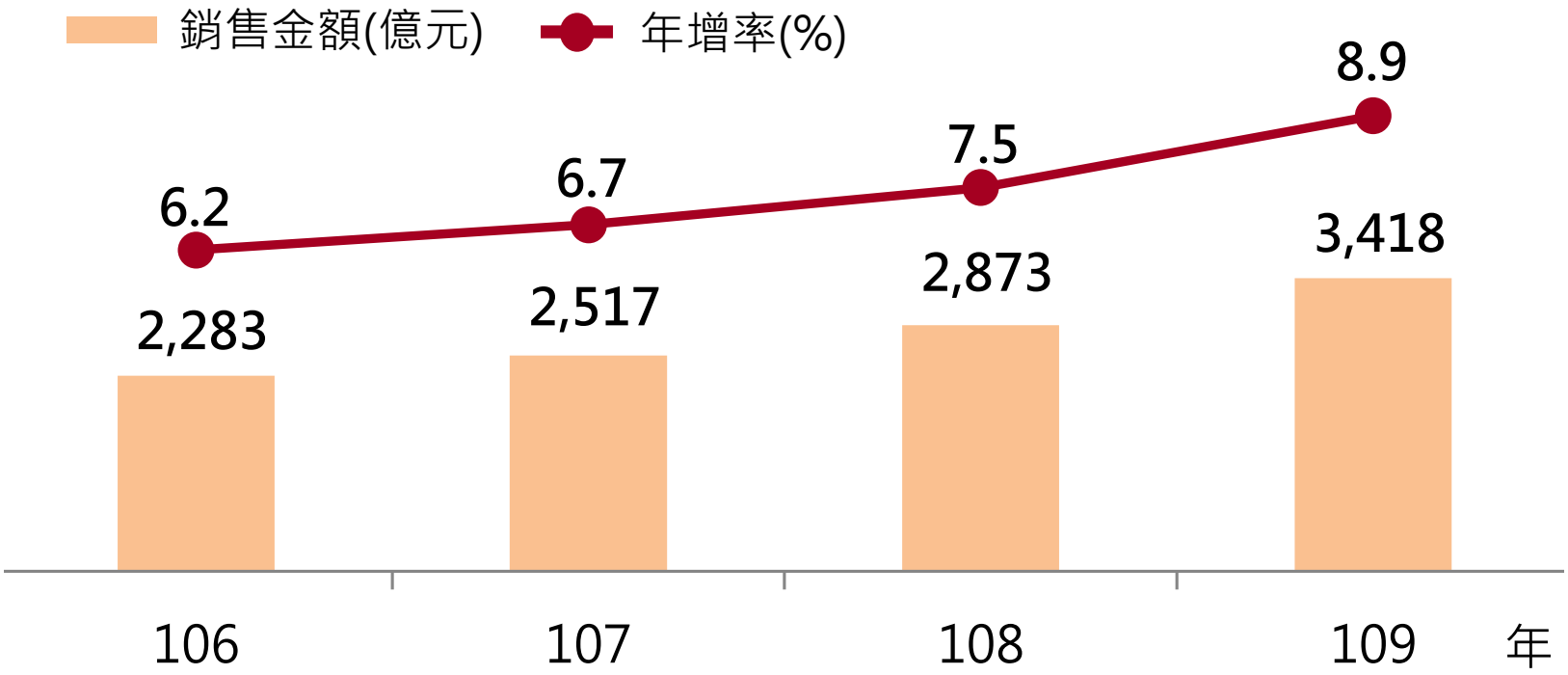
大綱

- 為什麼需要網路爬梳
- 如何進行網路爬梳
- 網路爬梳的限制
- 結語

- 為什麼需要網路爬梳
- 如何進行網路爬梳
- 網路爬梳的限制
- 結語

為什麼需要網路爬梳(1/6)

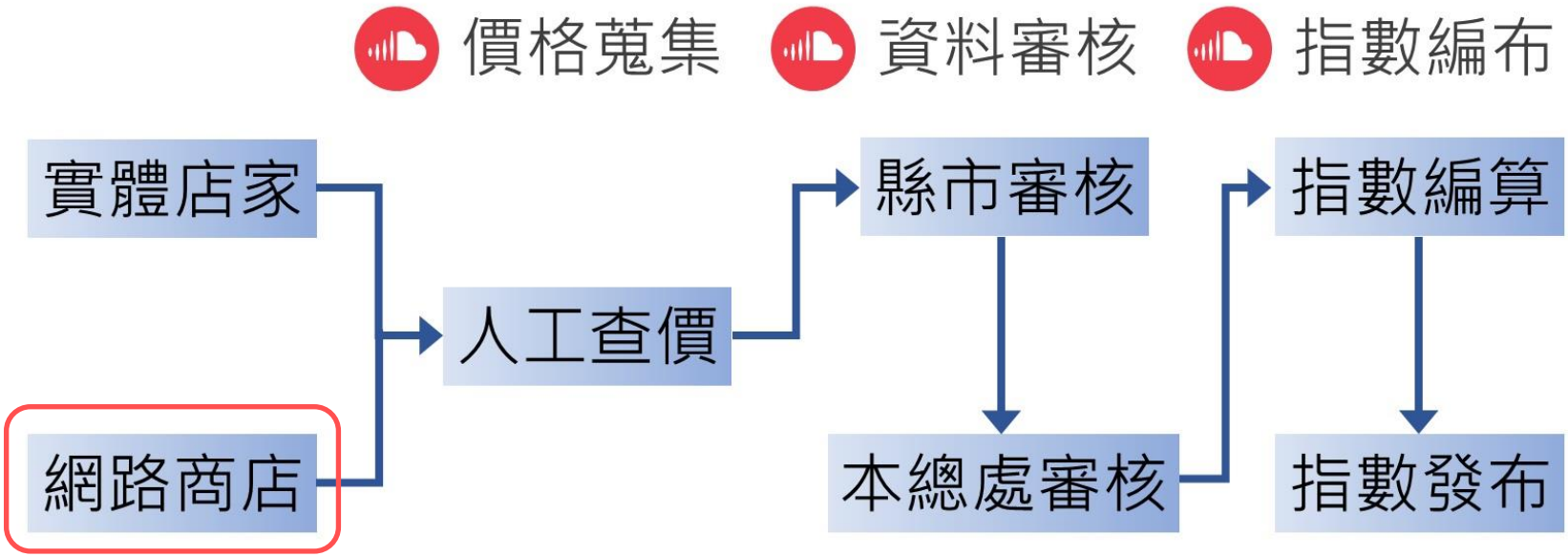
近年我國零售業網路銷售概況



資料來源：經濟部統計處

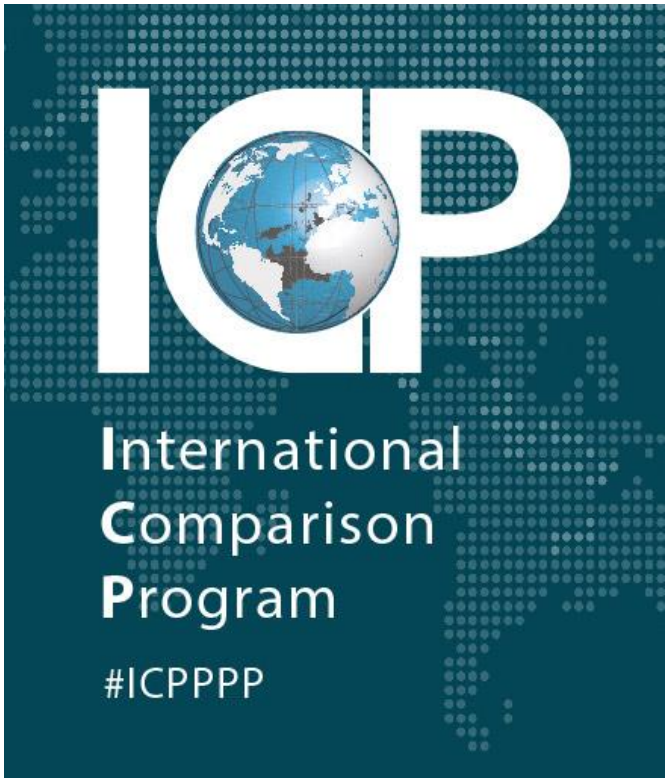
為什麼需要網路爬梳(2/6)

➢CPI編製流程



為什麼需要網路爬梳(3/6)

- 購買力平價(PPP)國際比較計畫(ICP)
  (https://www.worldbank.org/en/programs/icp)



- 主要目的
  - 物價水準
  - 實質GDP規模
- 週期
  - 109年起每3年 1 次

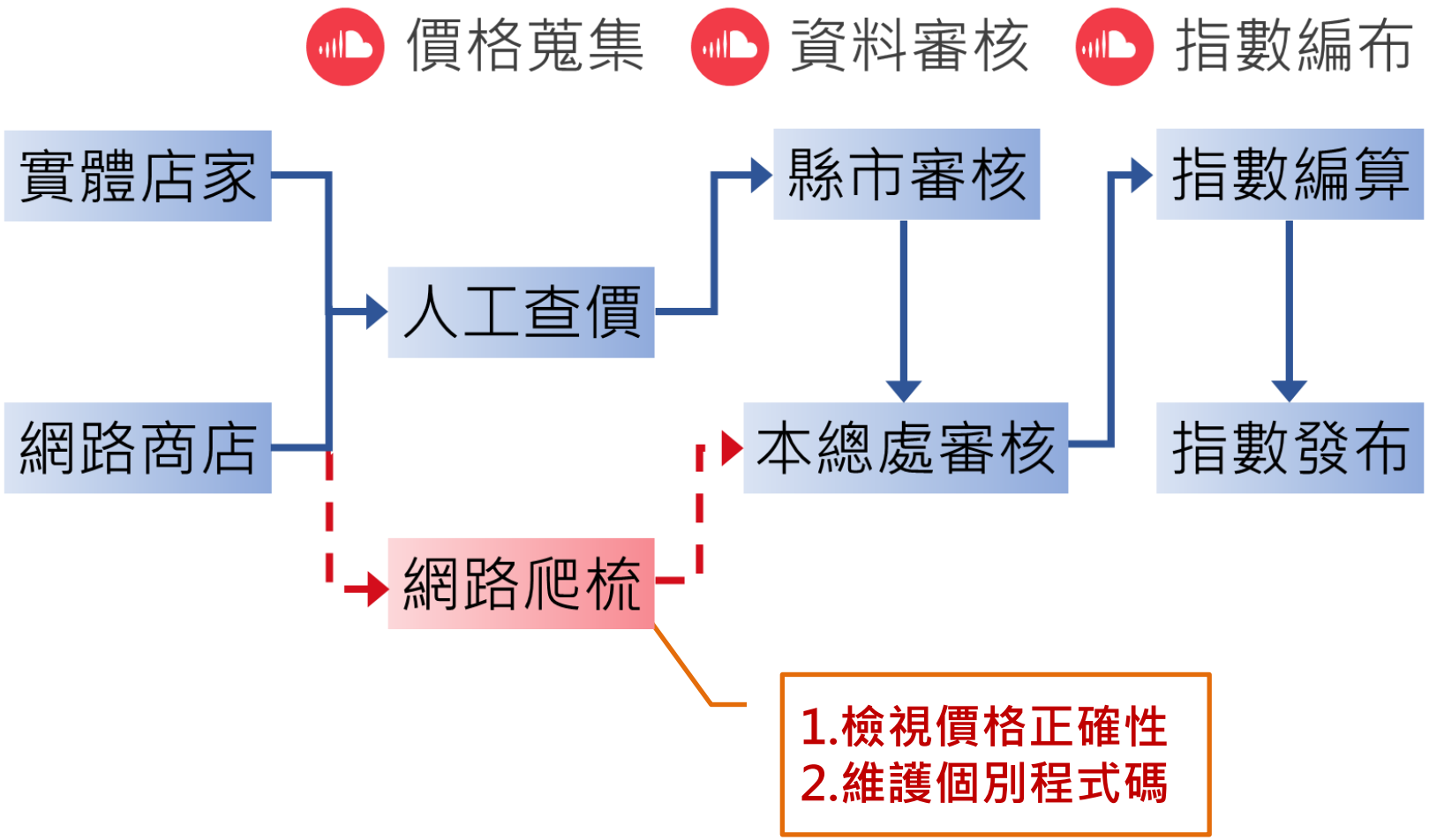
說明：全球近200個國家參與。

為什麼需要網路爬梳(4/6)

➢比價網可蒐得物價所需資料

為什麼需要網路爬梳(5/6)

➢CPI編製流程(108年7月起納入網路爬梳)



為什麼需要網路爬梳(6/6)

➢多數國家亦已投入研究

國家	CPI預計(已)採用項目
英國	預計2023年起逐年採用：資訊產品、影音光碟、書籍、旅遊團費、衣服。
日本	預計2021年7月採用：國外旅行團費、機票、旅館住宿費。
荷蘭	已採用：衣服。
挪威	已採用：機票、牙醫醫療費。
義大利	已採用：電子產品、火車票。

資料來源：自行整理

- 為什麼需要網路爬梳
- **如何進行網路爬梳**
- 網路爬梳的限制
- 結語

如何進行網路爬梳(1/13)

- 程式語言
  - PHP
  - Java
  - C/C++
  - VBA
  - Python

1. 語法簡潔
2. 數據分析功能完善
3. 使用的人多

- 資料取得方式
  - 循網址
  - 關鍵字
  - 循目錄

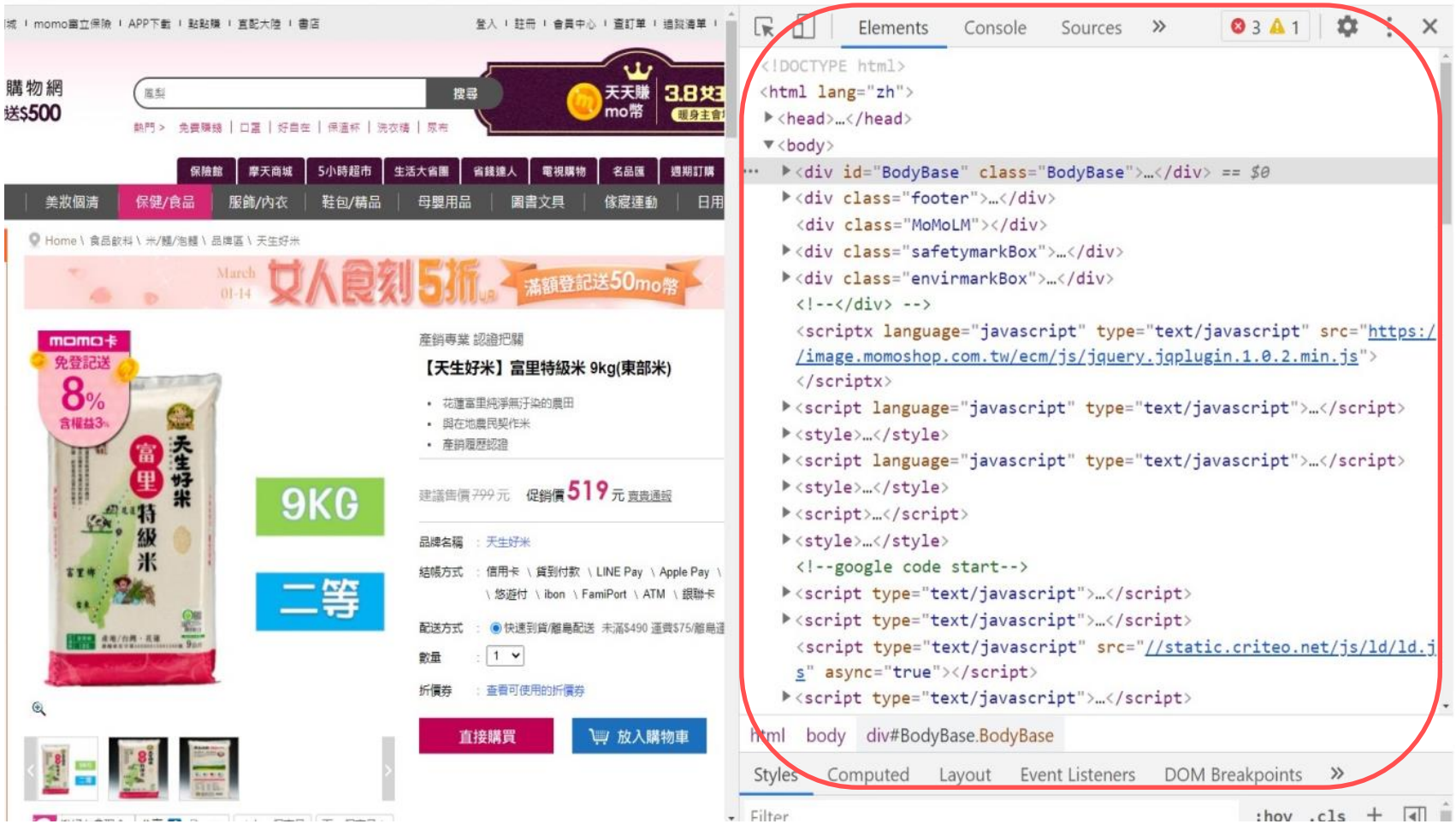
如何進行網路爬梳_循網址(2/13)

➢ 爬梳前：人工點選網址並登錄資料

	B	C	D	E	G	H	AC
1	項號	查價項目	網路廠商	花色規格	是否無	開網頁	109/01
2	1	米	momo商城	【天生好米】富里特級米 9kg(東部米)		https://www.momosho	650
3	1	米	樂天市場	三好米履歷台梗九號2.2KG【愛買】		https://www.rakuten.c	179
4	1	米	udn	【池上農會】台灣好米月光米(1.8kg /		https://shopping.udn.c	705
5	1	米	PChome24H購物	袋裝；<中興米>經典米(CNS一等)；1包2		https://24h.pchome.co	109
6	1	米	PChome24H購物	《三好米》三好長鮮米(12kg/包)		https://24h.pchome.co	479
7	2	糯米	奇摩購物	溪州尚水米 紫米800gx10包(黑糯米)；賣		https://tw.buy.yahoo.c	2099
8	2	糯米	PChome24H購物	【米屋】嚴選長糯米(600g)		https://24h.pchome.co	68
9	2	糯米	奇摩購物中心	【光豐農會】太巴塱有機黑糯		https://tw.buy.yahoo.c	654
10	2	糯米	momo商城	【米屋】嚴選黑糯米(600g/包x1)		https://www.momosho	140
11	2	糯米	momo商城	【米樂銀川】銀川有機長糯白米(600g)		https://www.momosho	53
12	3	米粉、粿條及其	博客來網路書局	老鍋米粉-純米台灣米粉炒(4人)		https://www.books.co	225
13	3	米粉、粿條及其	奇摩購物	佛祖牌 新竹米粉(200g)；賣場編號：718		https://tw.buy.yahoo.c	58
14	3	米粉、粿條及其	udn	【米樂】調合米粉-米含量90%, 200g/包(		https://shopping.udn.c	100
15	3	米粉、粿條及其	momo商城	【永盛米粉-聖光牌】100%純米米V		https://www.momosho	88
16	3	米粉、粿條及其	PChome24H購物	虎牌米粉(純米製造) 260g		https://24h.pchome.co	89

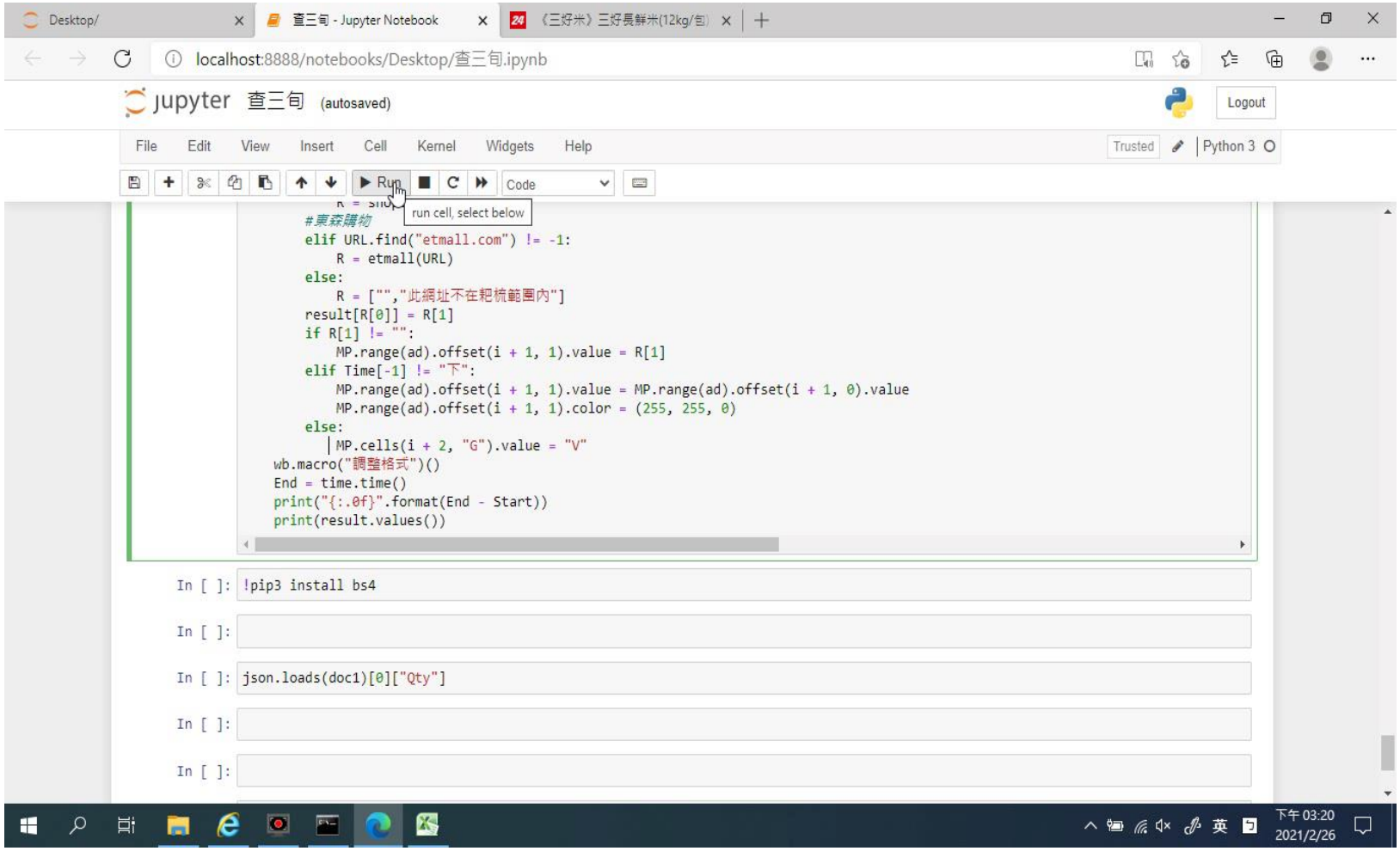
如何進行網路爬梳_循網址(3/13)

- 人工查價的網頁、爬梳程式搜尋的網頁程式



如何進行網路爬梳_循網址(4/13)

➢ 循網址爬梳實際過程



如何進行網路爬梳_循網址(5/13)

- 目前進度
  - 針對不同網站撰寫程式碼
    - ✓ 直接搜尋主程式語法
    - ✓ 透過API取得價格 (Application Programming Interface)
  - 完成10幾個主要網路電商平台
  - 蒐集近千筆不同規格樣本
    - ✓ 分散在CPI近4成(139個)查價項目群

如何進行網路爬梳_關鍵字(6/13)

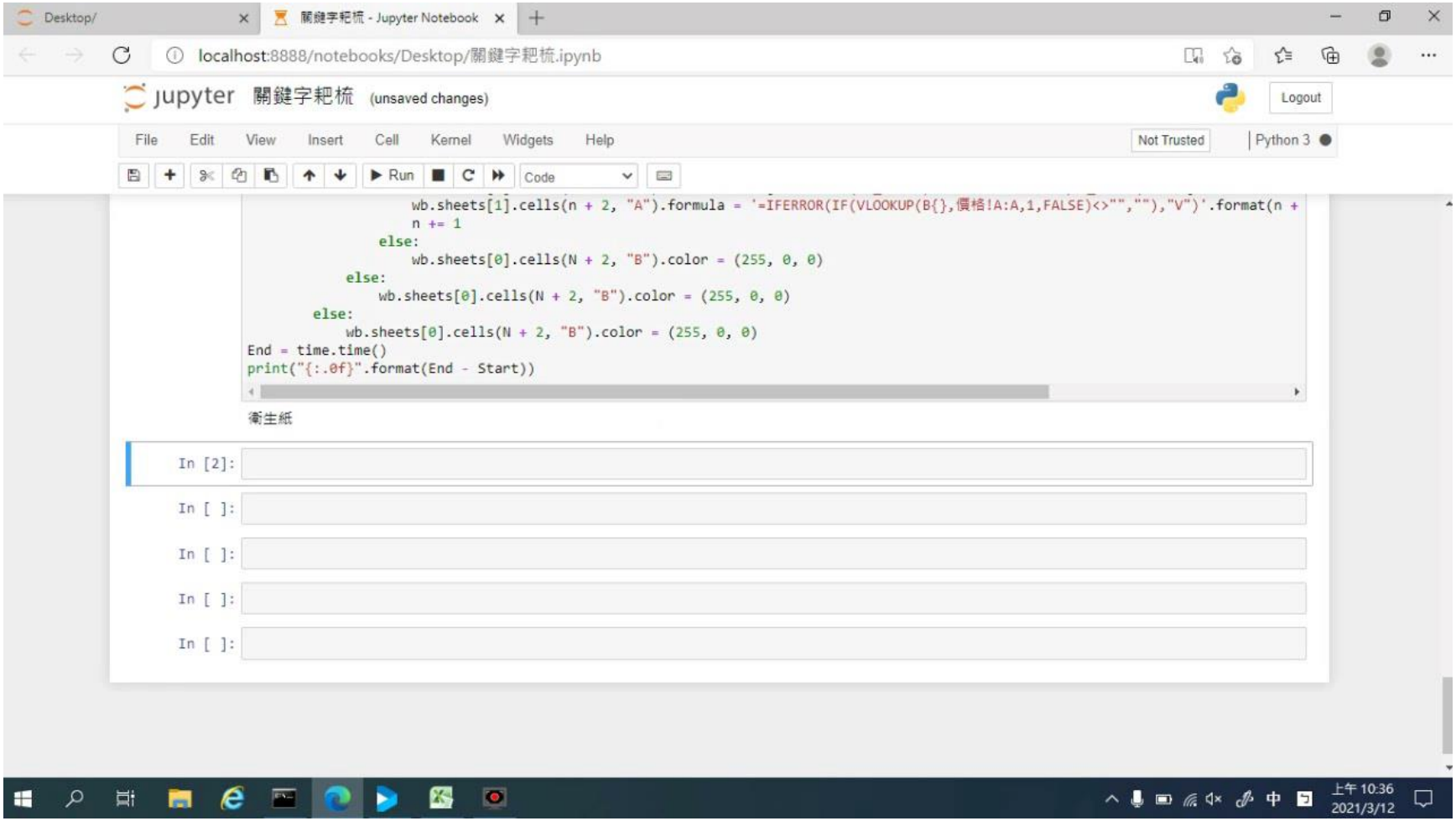
- 針對各界關注的CPI單一項目群
- 大量增加樣本、提升代表性
- 不同電商平台搜尋資料筆數差異大

如何進行網路爬梳_關鍵字(7/13)

- 目前進度
  - 建立資料清洗機制
    - ✓ 須有：衛生紙、家用紙
    - ✓ 剔除：~~衛生紙~~架、~~衛生紙~~盒、廚房~~紙巾~~
  - 每月觀察新增爬梳品項
    - ✓ 是否符合目前「資料清洗機制」
  - 撰寫價格分析程式
  - 預計今年底前完成納入CPI編算

如何進行網路爬梳_關鍵字(8/13)

➢ 關鍵字爬梳實際過程_爬梳過程



如何進行網路爬梳_關鍵字(9/13)

➢ 關鍵字爬梳實際過程_新增品項判斷

	A	B	C	D	E
1	新增	商品碼	網址	花色	價格
5	V	6781146	https://www	【倍潔雅】柔軟舒適抽取式衛生紙150抽x60包/箱	669
9	V	7252730	https://www	【倍潔雅】花漾柔感抽取式衛生紙200抽x48包/箱	618
11	V	6955871	https://www	【倍潔雅】花漾柔感抽取式衛生紙(150抽70包/箱)	760
13	V	3631231	https://www	【得意-重裝版】抽取式花紋衛生紙(150抽*12入*7袋)	929
14	V	238168	https://www	【興亞】大捲筒衛生紙架(圓形透明黑)	389
17	V	3379026	https://www	【五月花】極柔頂級抽取式衛生紙(110抽*12包*6袋)	899
19	V	7344913	https://www	【MAY FLOWER 五月花】花園版抽取式衛生紙(100抽*8包*8袋)	679
20	V	6780113	https://www	【唯潔雅】唯潔雅潔淨柔感抽取式衛生紙(100抽12包6袋/箱)	499
22	V	7493278	https://www	【春風】抽取式衛生紙-輕柔膚觸-100抽*24包*3串	699

如何進行網路爬梳_循目錄(10/13)

- 複合品項電商：分類難以對應至CPI項目

如何進行網路爬梳_循目錄(11/13)

➢ 專門品項電商：資料較容易(無須)清洗

飯店資料：　池上 日暉國際渡假村

房　型：　雙人房

專案名稱：

○【出發吧！找茶趣】-翠綠山嵐雙人房〔１大床或２小床〕

○【✿FUN寒假 兒童免費】-翠綠山嵐雙人房〔１大床或２小床〕

○【出發吧！找茶趣】-伯爵別墅villa〔１大床〕

◉【Airstream露營車體驗-自在假期】-露營車〔依飯店安排為主〕

<<上二個月　　2021年3月

日	一	二	三	四	五	六
	01	02	03	04	05	06
07	08 7,590元 額滿	09 7,590元 額滿	10 7,590元 額滿	11 7,590元 額滿	12 8,140元 額滿	13 9,790元 額滿
14 8,140元 餘1間	15 7,590元 餘1間	16 7,590元 餘1間	17 7,590元 餘1間	18 7,590元 餘1間	19 8,140元 餘1間	20 9,790元 餘1間
21 8,140元 餘1間	22 7,590元 餘1間	23 7,590元 餘1間	24 7,590元 餘1間	25 7,590元 餘1間	26 8,140元 餘1間	27 9,790元 餘1間
28 8,140元 餘1間	29 7,590元 餘1間	30 7,590元 餘1間	31 7,590元 餘1間			

2021年4月　　下二個月>>

日	一	二	三	四	五	六
				01 7,590元 餘1間	02 9,790元 額滿	03 9,790元 餘1間
04 9,790元 餘1間	05 8,140元 餘1間	06 7,590元 餘1間	07 7,590元 餘1間	08 7,590元 餘1間	09 8,140元 餘1間	10 9,790元 餘1間
11 8,140元 餘1間	12 7,590元 餘1間	13 7,590元 餘1間	14 7,590元 餘1間	15 7,590元 餘1間	16 8,140元 餘1間	17 9,790元 餘1間
18 8,140元 餘1間	19 7,590元 餘1間	20 7,590元 餘1間	21 7,590元 餘1間	22 7,590元 餘1間	23 8,140元 餘1間	24 9,790元 餘1間
25 8,140元 餘1間	26 7,590元 餘1間	27 7,590元 餘1間	28 7,590元 餘1間	29 7,590元 餘1間	30 8,140元 餘1間	

如何進行網路爬梳_循目錄(12/13)

➢ 蒐集每日價格：資料量龐大但爬梳費時

	A	B	C	D	E	F	G	H	I	J
1	飯店	方案	房型	1	2	3	4	5	6	7
2	北投 加賀屋溫泉飯店	2021一泊一食-溫馨標準套房〔和洋或和式房〕	雙人房	12100	12100	12100	12100	14300	14300	12100
3	北投 加賀屋溫泉飯店	2021一泊一食-園景和洋標準套房〔１大床〕	雙人房	13200	13200	13200	13200	15400	15400	13200
4	北投 加賀屋溫泉飯店	2021一泊二食-溫馨標準套房〔和洋或和式房〕	雙人房	16720	16720	16720	16720	18920	18920	16720
5	北投 加賀屋溫泉飯店	2021一泊二食-園景和洋標準套房〔１大床〕	雙人房	17820	17820	17820	17820	20020	20020	17820
6	北投 加賀屋溫泉飯店	2021一泊一食-園景和式標準套房〔２鋪床〕	雙人房	14300	14300	14300	14300	16500	16500	14300
7	北投 加賀屋溫泉飯店	2021一泊一食-園景和洋雅緻套房〔２小床〕	雙人房	14300	14300	14300	14300	16500	16500	14300
8	北投 加賀屋溫泉飯店	2021一泊一食-園景露天風呂套房〔１大床或２小床〕	雙人房	16500	16500	16500	16500	18700	18700	16500
9	北投 加賀屋溫泉飯店	2021一泊二食-園景和式標準套房〔２鋪床〕	雙人房	18920	18920	18920	18920	21120	21120	18920
10	北投 加賀屋溫泉飯店	2021一泊二食-園景和洋雅緻套房〔２小床〕	雙人房	18920	18920	18920	18920	21120	21120	18920
11	北投 加賀屋溫泉飯店	2021一泊二食-園景露天風呂套房〔１大床或２小床〕	雙人房	20020	20020	20020	20020	22220	22220	20020
12	台北 凱撒大飯店	指定日住房專案-精緻客房〔１大床或２小床〕	雙人房	1700	1700	1700	1700	1800	2100	1700
13	台北 凱撒大飯店	可樂訂房專案~不含早-精緻客房〔１大床或２小床〕	雙人房	3800	3800	3800	3800	4400	4400	3800
14	台北 凱撒大飯店	可樂訂房專案~不含早-悠遊客房〔１大床〕	雙人房	4400	4400	4400	4400	5000	5000	4400
15	台北 凱撒大飯店	可樂訂房專案~不含早-綺幻客房〔２小床〕	雙人房	4600	4600	4600	4600	5200	5200	4600
16	台北 凱撒大飯店	可樂訂房專案~不含早-站前套房〔１大床或２小床〕	雙人房	5800	5800	5800	5800	6400	6400	5800
17	台北 凱撒大飯店	春暖花開遊凱撒-精緻客房〔１大床或２小床〕	雙人房	1999	1999	1999	1999	1999	2599	1999
18	台北 凱撒大飯店	春暖花開遊凱撒-悠遊客房〔１大床〕	雙人房	2299	2299	2299	2299	2299	2899	2299
19	台北 凱撒大飯店	春暖花開遊凱撒-綺幻客房〔２小床〕	雙人房	2399	2399	2399	2399	2399	2999	2399
20	台北 凱撒大飯店	2021 台北凱撒 城市慢遊 連住專案-精緻客房〔１大床或	雙人房	4900	4900	4900	4900	4900	4900	4900
21	台北 凱撒大飯店	2021 台北凱撒 城市慢遊 連住專案-悠遊客房〔１大床〕	雙人房	5700	5700	5700	5700	5700	5700	5700
22	台北 凱撒大飯店	2021 台北凱撒 城市慢遊 連住專案-綺幻客房〔２小床〕	雙人房	5900	5900	5900	5900	5900	5900	5900
23	台北 凱撒大飯店	可樂訂房專案~不含早-精緻四人房〔２大床〕	四人房	5500	5500	5500	5500	6100	6100	5500
24	台北 凱撒大飯店	春暖花開遊凱撒-精緻四人房〔２大床〕	四人房	3899	3899	3899	3899	3899	4499	3899
25	台北 凱撒大飯店	2021 台北凱撒 城市慢遊 連住專案-精緻四人房〔２大床	四人房	7900	7900	7900	7900	7900	7900	7900

如何進行網路爬梳_循目錄(13/13)

- 目前進度
  - 旅館住宿、國內旅遊
    - ✓ 500個飯店、3,000種方案及房型、8萬筆價格
    - ✓ 300個行程、1,000種方案、3萬筆價格
  - 撰寫價格分析程式
  - 預計今年底前完成納入CPI編算

- 為什麼需要網路爬梳
- 如何進行網路爬梳
- **網路爬梳的限制**
- 結語

網路爬梳的限制(1/3)

- 國際勞工組織(ILO)、國際貨幣基金(IMF)於2020年版CPI編製手冊中提及
  Prices can also be collected centrally online or by using web scraping or scanner data.
  - 提高資料代表性、節省調查負擔
  - 程式需要不斷維護
  - 商品筆數多時可能需要大量時間
  - 若網站會依IP實體位置定價，不容易爬梳到各地區價格
  - 只能針對有網站的零售廠商

網路爬梳的限制(2/3)

- 實務上執行時
  - 程式需要不斷維護
    - ✓ 試行至今遇到4至5次網頁改版
  - 商品筆數多時可能需要大量時間
    - ✓ 循網址：近千樣本，約40分鐘
    - ✓ 循目錄：8萬筆價格，約2小時

網路爬梳的限制(3/3)

- 實務上執行時
  - 若網站會依IP實體位置定價，不容易爬梳到各地區價格
    - ✓ 我國各地區價格一致
  - 只能針對有網站的零售廠商
    - ✓ 爬梳僅是部分CPI查價樣本
    - ✓ 9成以上零售消費皆在實體商店

- 為什麼需要網路爬梳
- 如何進行網路爬梳
- 網路爬梳的限制
- **結語**

結語

- 適當調配調查人力
  - 用於CPI及ICP的實體商家
- 爬梳僅是工具，須考量工作負荷
  - 物價科審核人力
  - 程式維護成本
- 建立完善且可持續的審核機制，提升CPI統計品質及代表性

報告完畢
敬請指教