



房價預測

以雙北地區住宅預測分析

BDSE31第三組

組長：曾子維

組員：陳郁文

簡宏晉

翁若芸

林文海

CONTENTS

1 專題簡介

團隊與專題

2 環境建置

Hadoop叢集架構

3 資料前處理

資料清洗

4 地域資料蒐集

經緯度及距離計算

5 模型建置

機器學習

6 後端框架

網站應用與開發

7 網頁視覺化

前端與操作

8 總結

成果分析

專題簡介

團隊與專題



曾子維

OUR TEAM



曾子維

組長

- 環境建置
- 模型建置



陳郁文

- 資料蒐集
- 資料清洗



簡宏晉

- 網頁視覺化
- 地域資料蒐集



OUR TEAM



翁若芸

- 資料蒐集
- 後端框架



林文海

- 後端框架
- 網頁視覺化



專題簡介

研究動機



台灣地狹人稠，近30年房價呈現長期上漲的趨勢,其中以雙北地區房價所得比創下歷年新高。



影響房價的因素眾多,除了房屋自身因素以外，外在條件以生活機能與交通便利為主要關鍵。

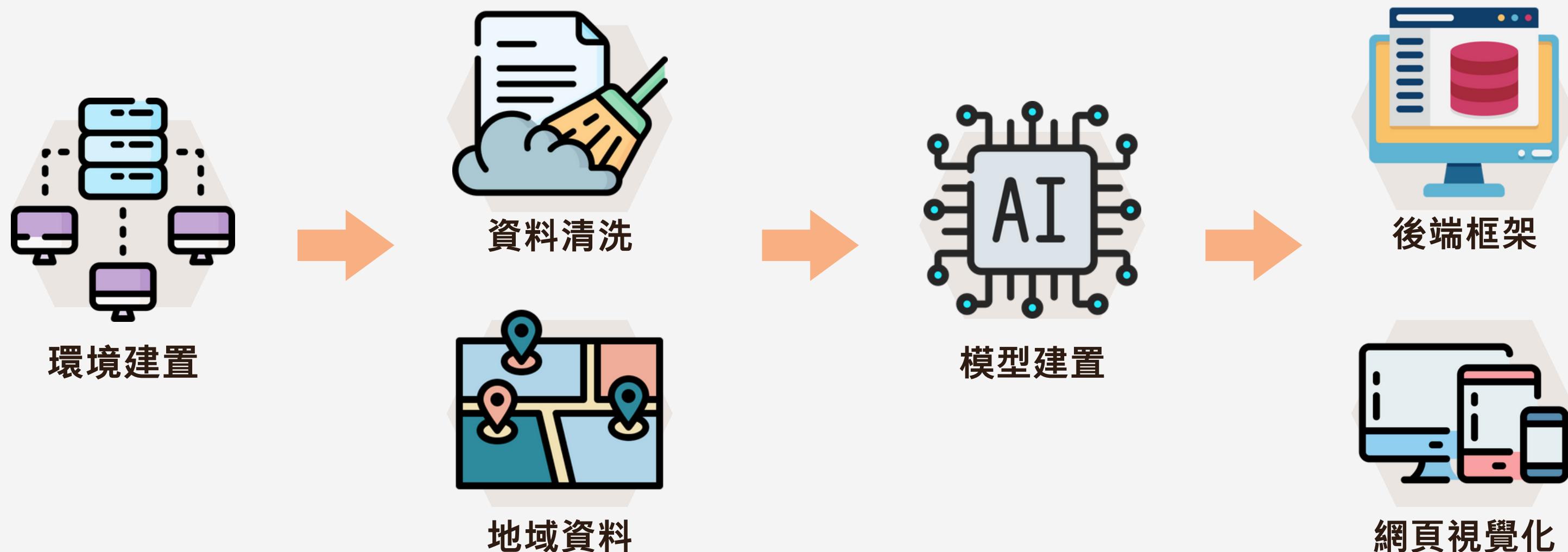


以雙北地區，交通、醫療、教育以及部分嫌惡設施(如殯葬、垃圾場)等因素，探討並預測房價。



專題簡介

研究流程與使用工具



專題簡介

研究流程與使用工具

資料蒐集



資料前處理



環境建置



ubuntu 22.04 LTS



模型建置



後端框架、視覺化



環境建置

Hadoop叢集架構



曾子維

環境建置

系統架構配置

| 系統架構配備 | |
|--------|--------------|
| 實體主機 | 6台 |
| 虛擬機 | 8台 |
| 記憶體 | 叢集96G(單機12G) |
| CPU | 叢集48核(單機6核) |



環境建置

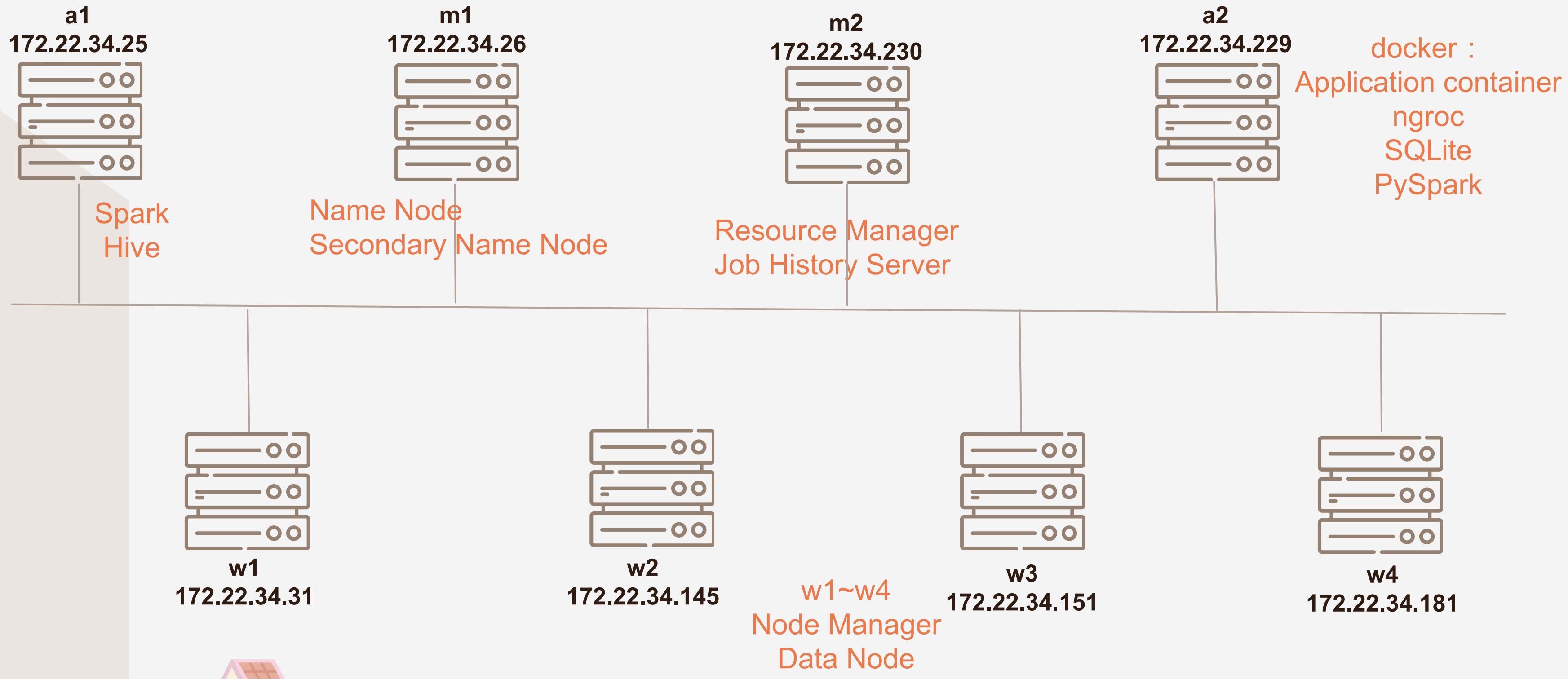
系統架構配置表

| 系統架構配置表 | |
|------------------|--|
| a1、a2 | client機 |
| m1 | Name Node Secondary Name Node |
| m2 | Resource Manager Job History Server |
| w1~w4 Worker機 | Data node |

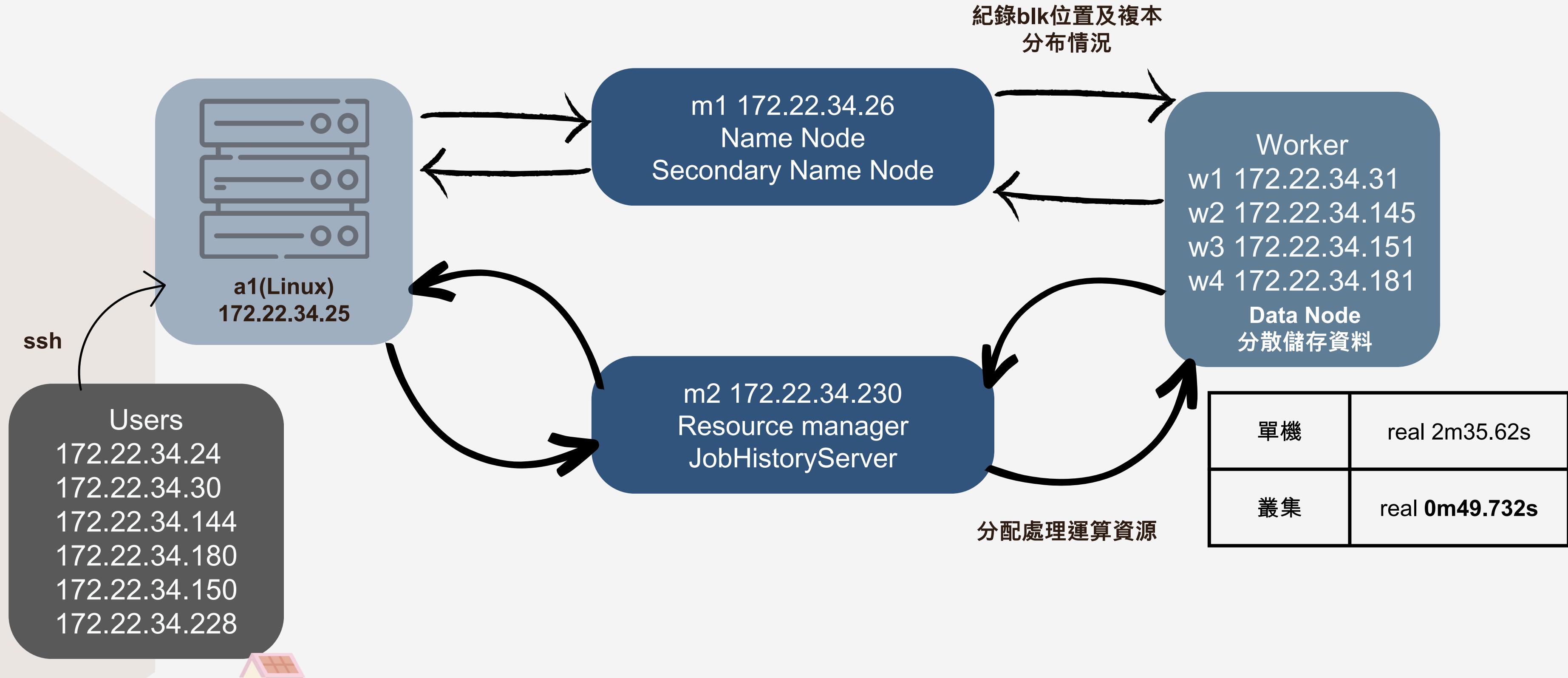


環境建置

Hadoop叢集生態



環境建置 Hadoop叢集生態運作圖



陳郁文

資料前處理

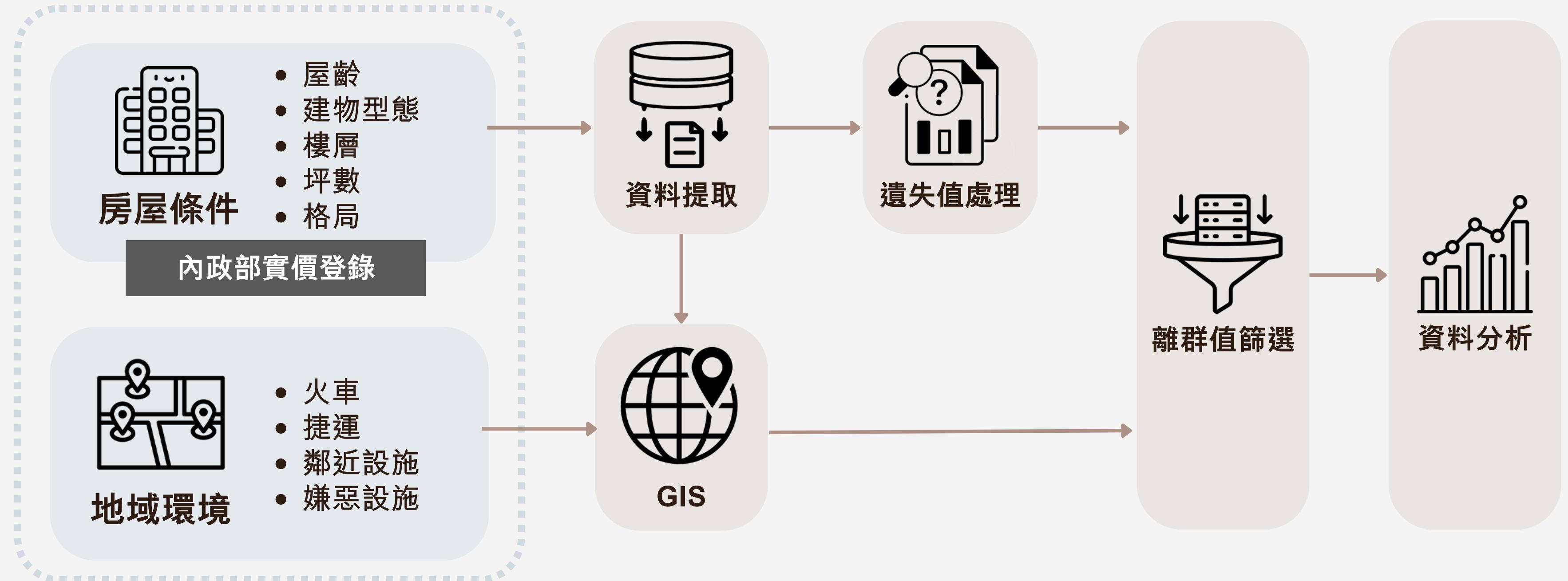
資料清洗



陳郁文

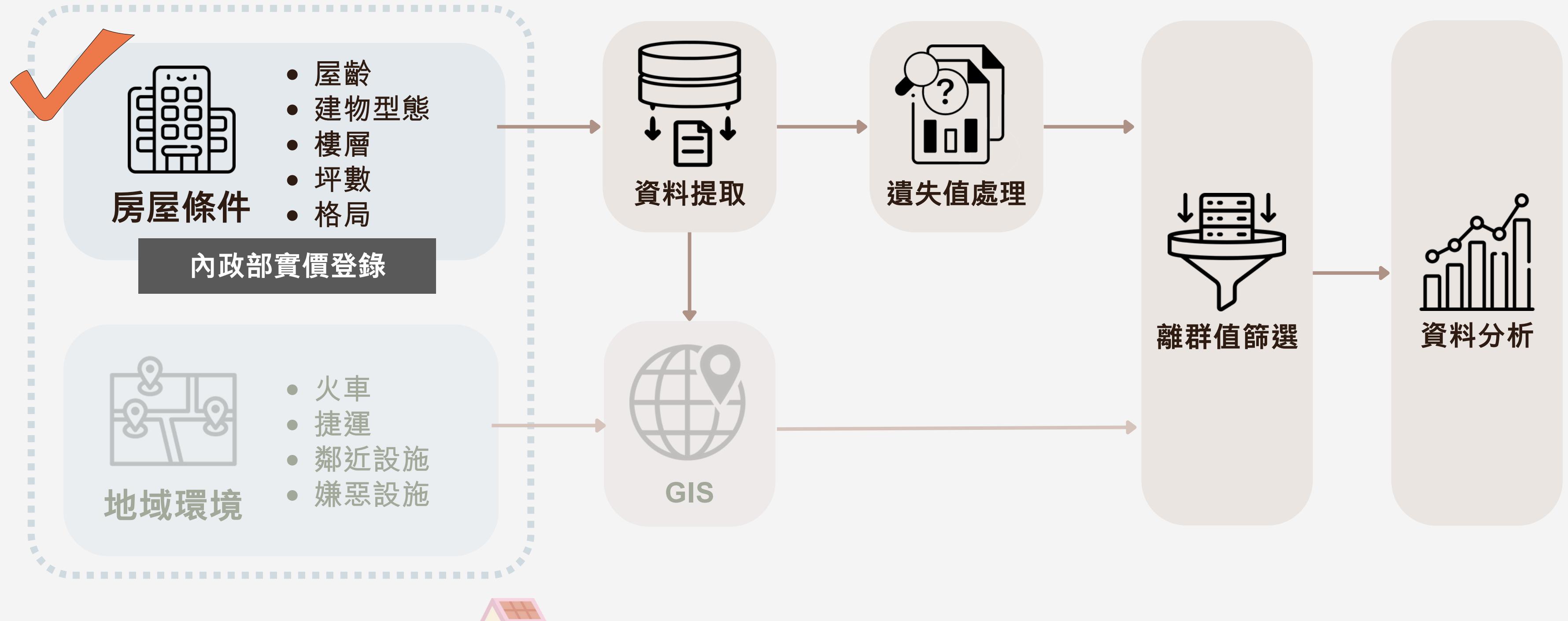
資料清理流程

影響房價因素



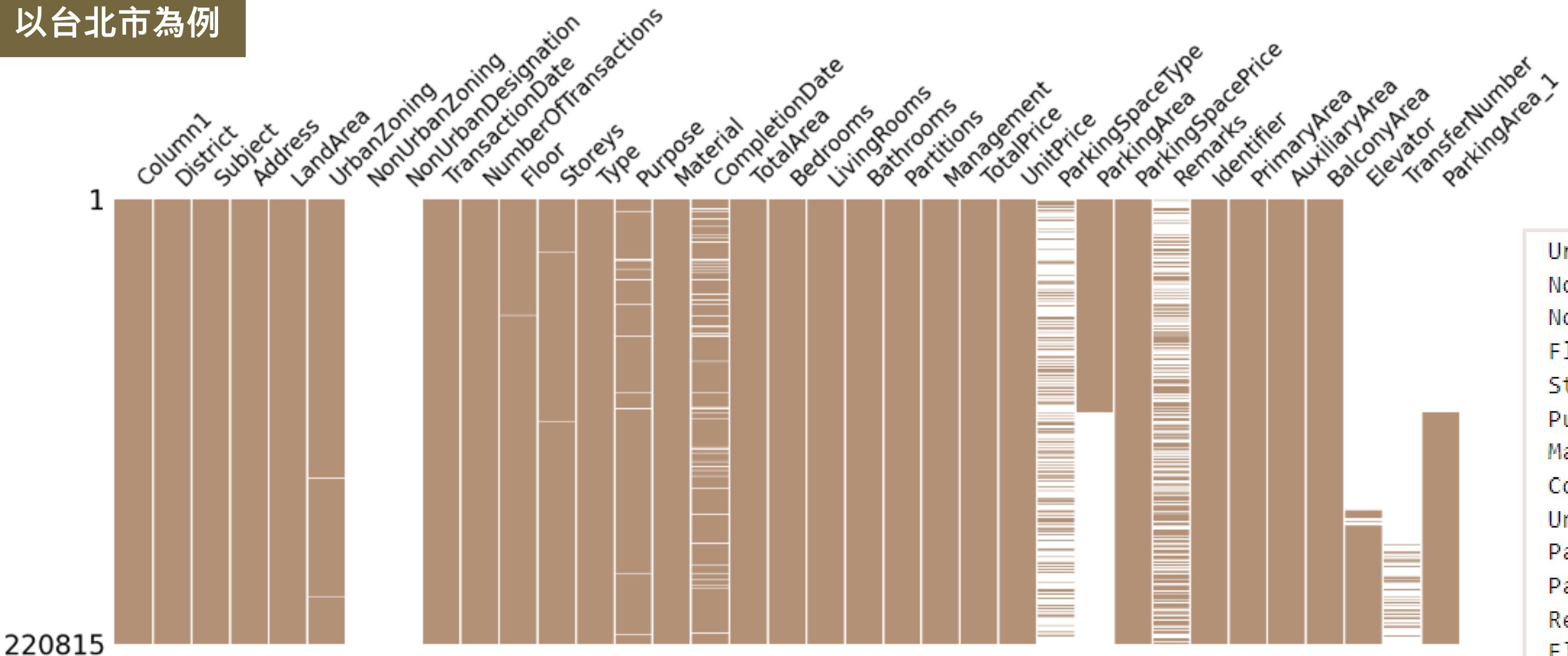
資料清理流程

影響房價因素



資料清理

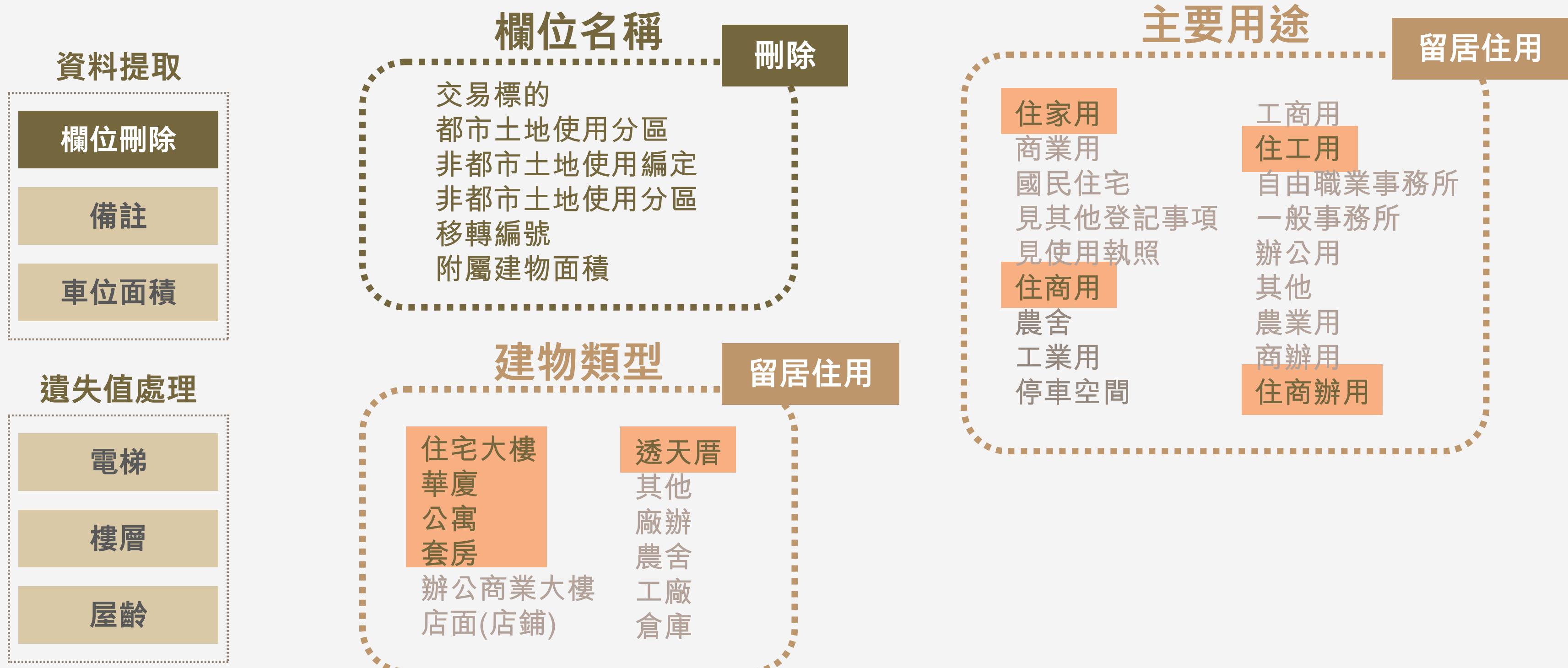
以台北市為例



Missing values (event plot)

| | |
|---------------------|--------|
| UrbanZoning | 1296 |
| NonUrbanZoning | 220786 |
| NonUrbanDesignation | 220815 |
| Floor | 244 |
| Storeys | 490 |
| Purpose | 4329 |
| Material | 52 |
| CompletionDate | 21565 |
| UnitPrice | 158 |
| ParkingSpaceType | 145382 |
| ParkingArea | 115106 |
| Remarks | 113582 |
| Identifier | 157673 |
| PrimaryArea | 206742 |
| AuxiliaryArea | 105709 |
| BalconyArea | |
| Elevator | |
| TransferNumber | |
| ParkingArea_1 | |
| dtype: int64 | |
| 220815 | |

資料提取



資料提取

資料提取

欄位刪除

備註

車位面積

| 備註 |
|---------|
| 伯姪買賣 |
| 預售屋買賣案件 |
| 二等親買賣 |

原備註文字



遺失值處理

電梯

樓層

屋齡

關鍵字

親父叔夫
伯兄姪妻
母婆朋友債務

刪除親友買賣

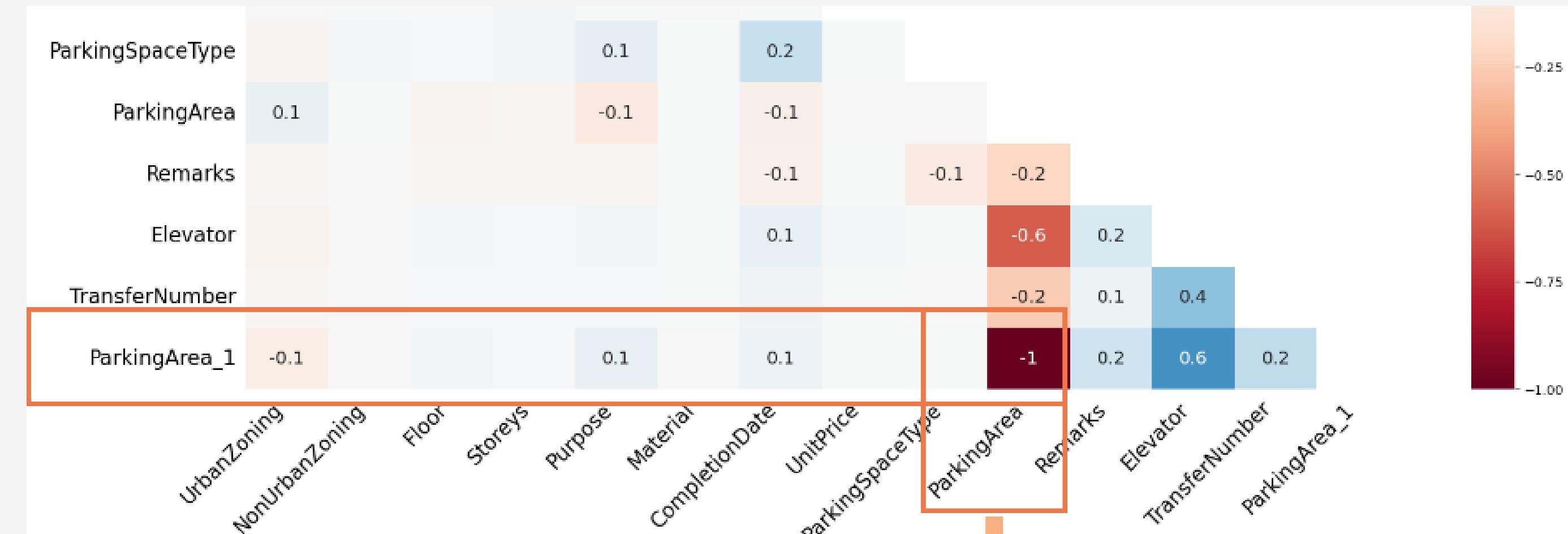


資料提取

資料提取

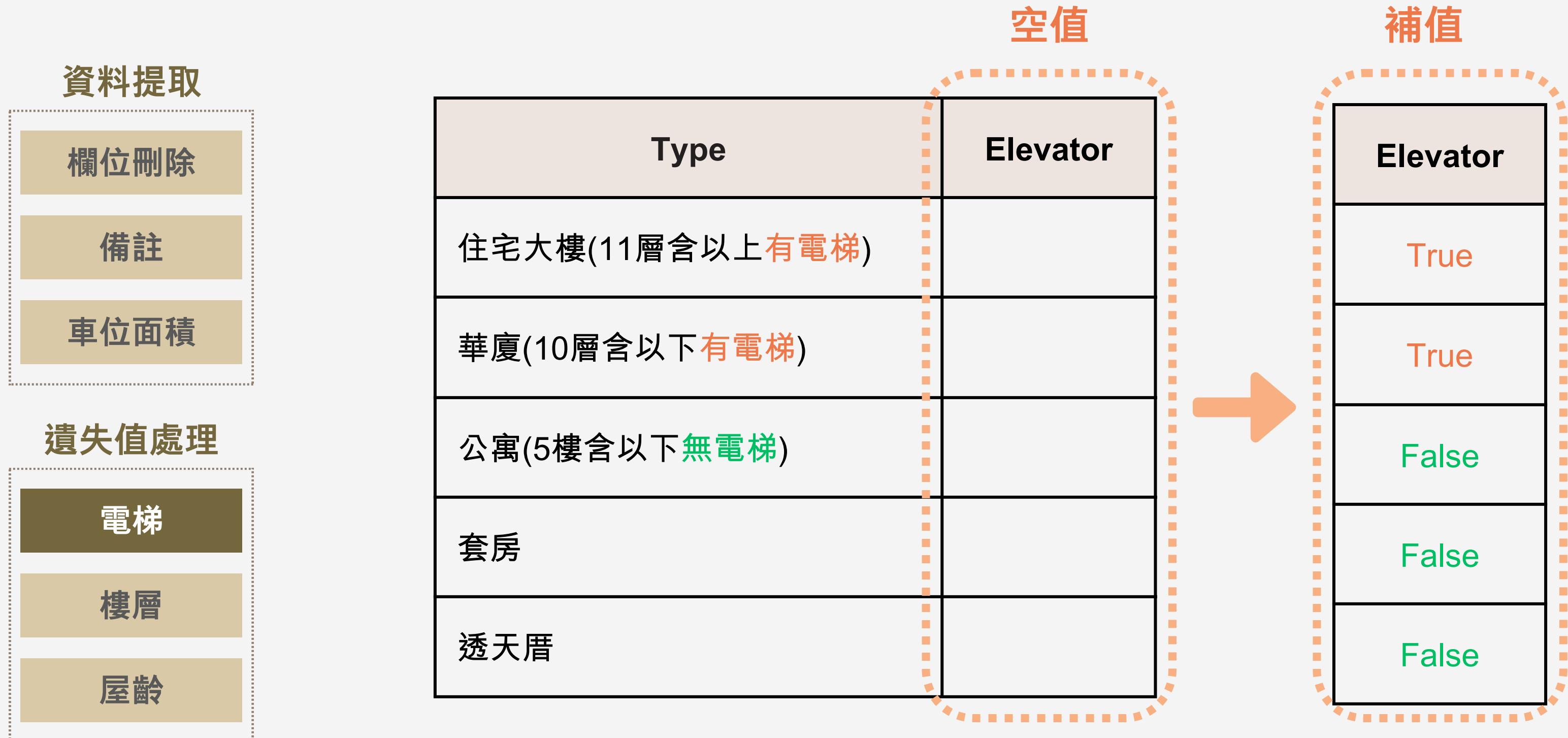


遺失值處理



合併欄位

遺失值處理



遺失值處理

資料提取

欄位刪除

備註

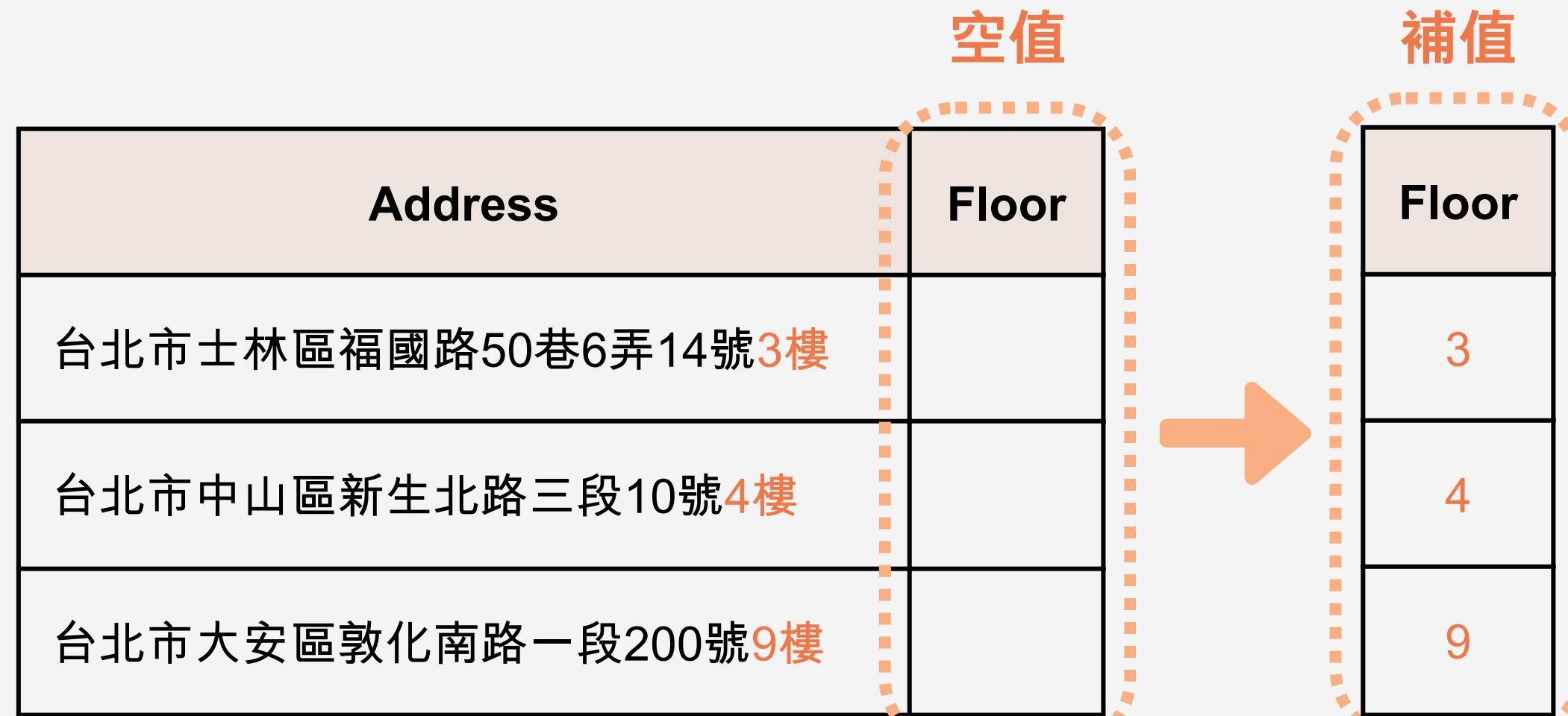
車位面積

遺失值處理

電梯

樓層

屋齡



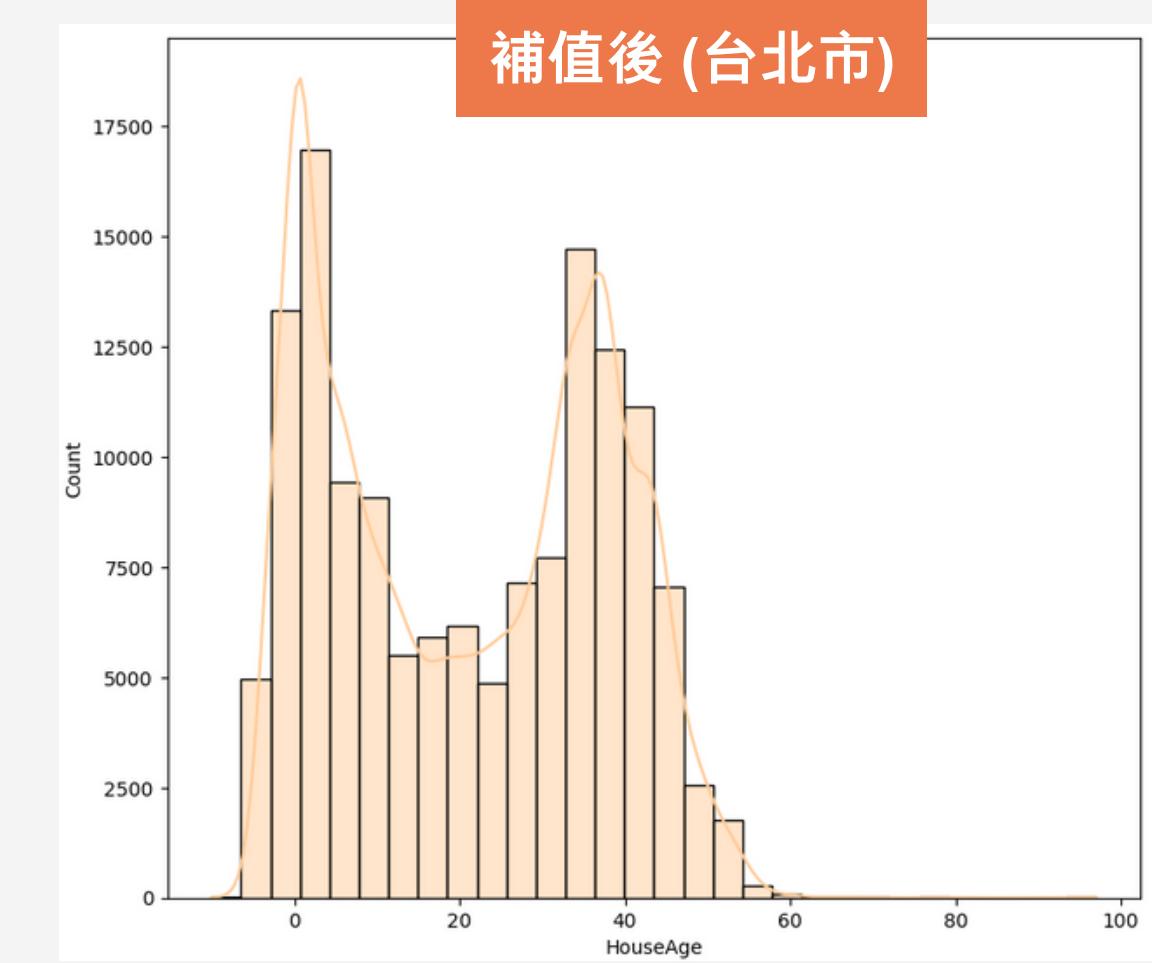
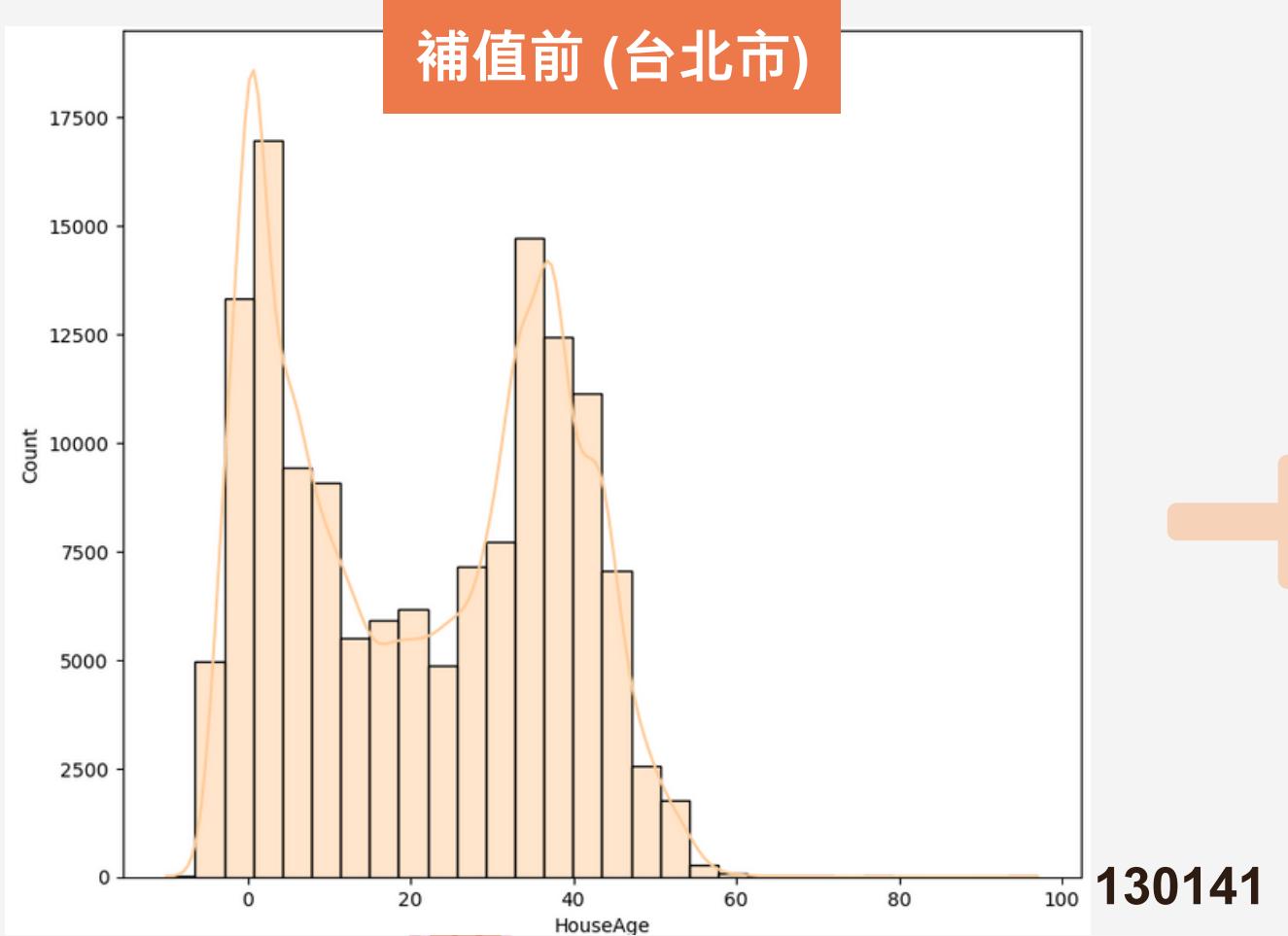
遺失值處理

資料提取



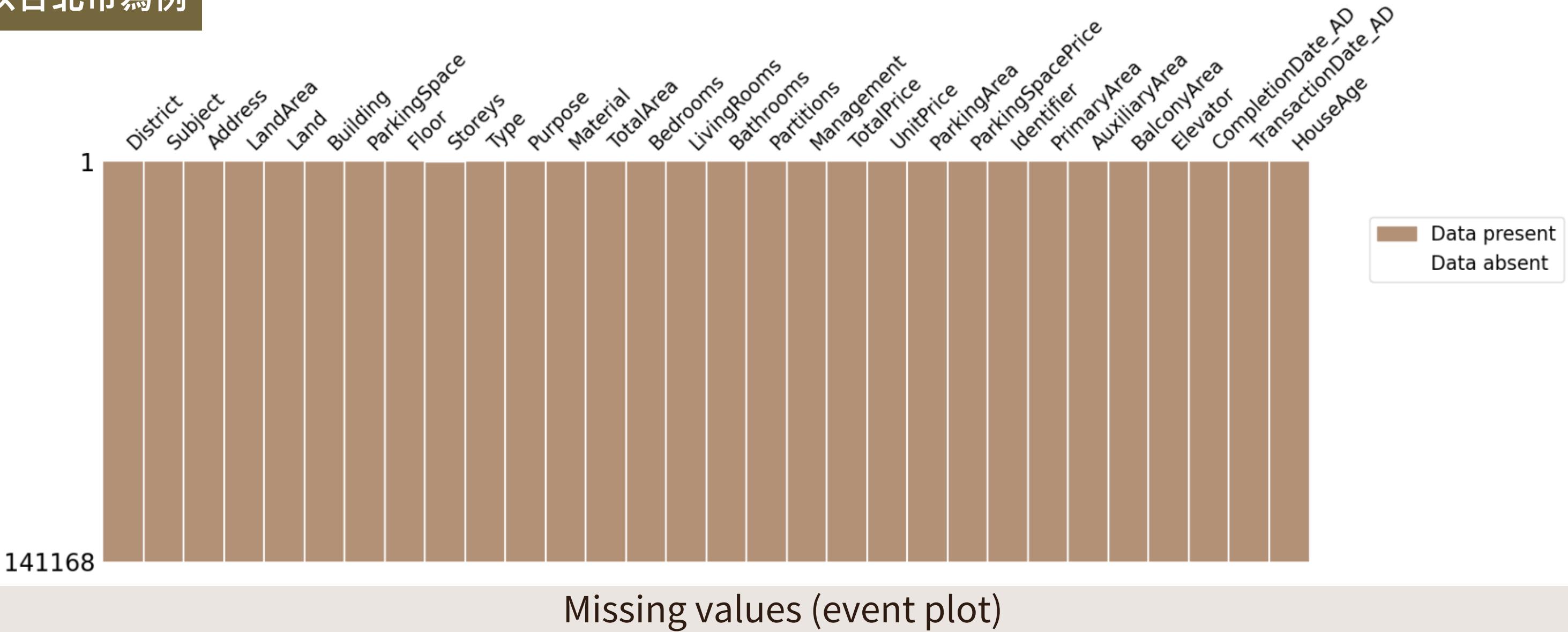
$$\text{屋齡 (年)} = \text{交易日期} - \text{完工日期}$$

- 「眾數」補值
- 台北市: 區域(12) * 房屋類型(5) = 60
- 新北市: 區域(29) * 房屋類型(5) = 145



資料清理

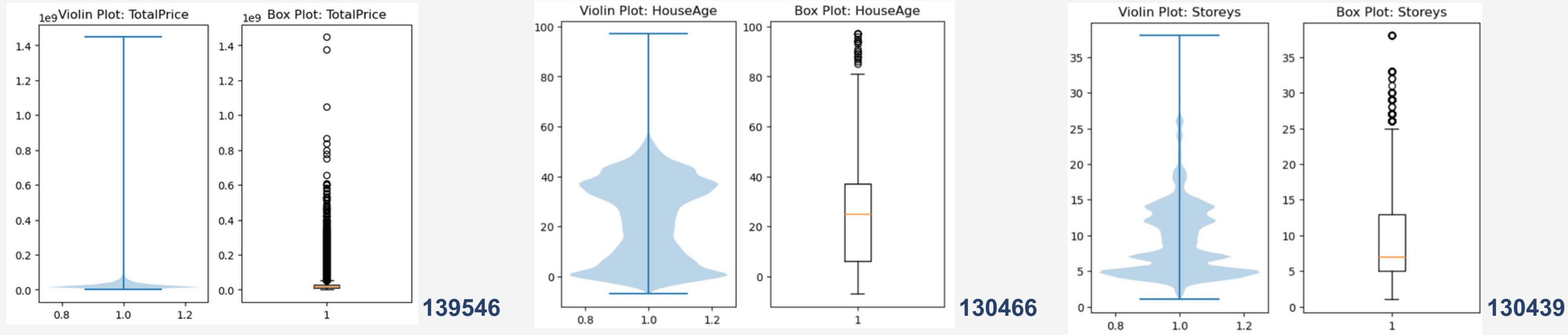
以台北市為例



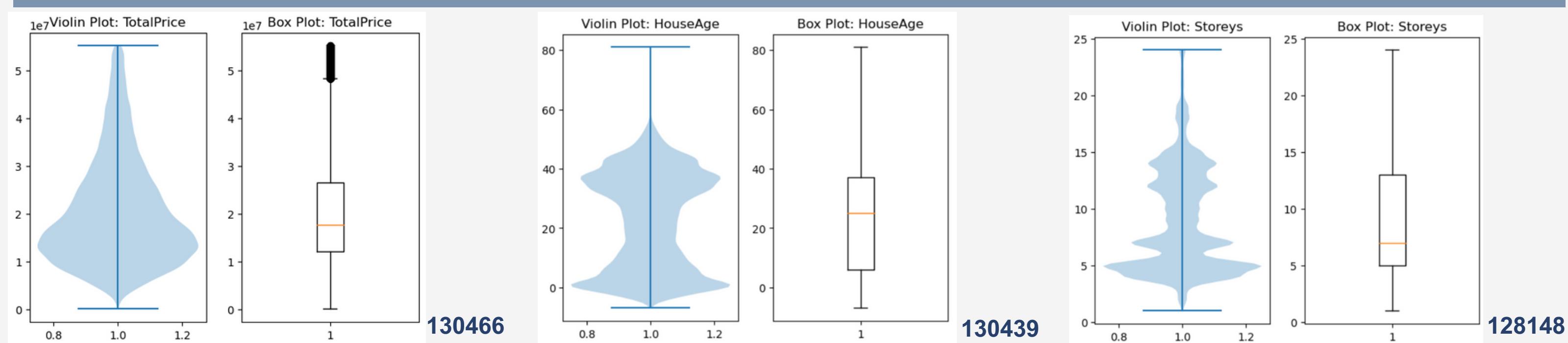
資料篩選

以台北市為例

資料分布 已刪除2012前交易



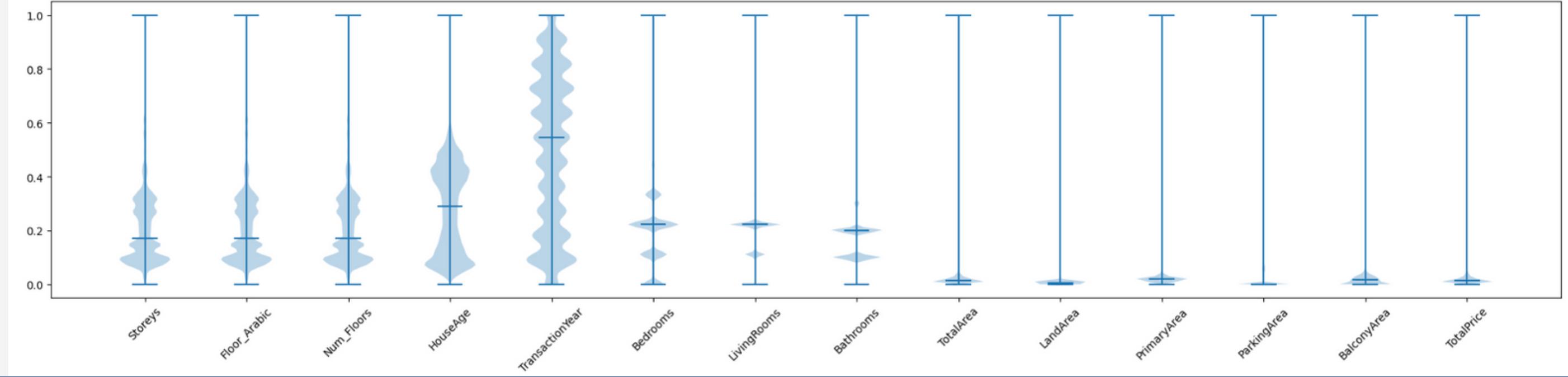
刪除離群值 (IQR的1.5倍)



資料篩選

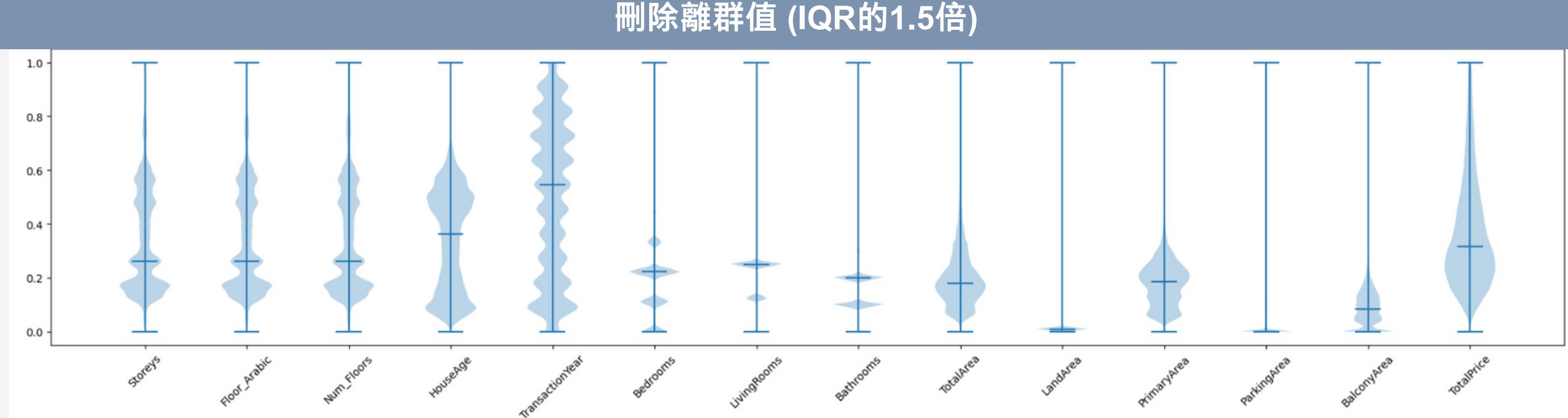
以台北市為例

資料分布



139546

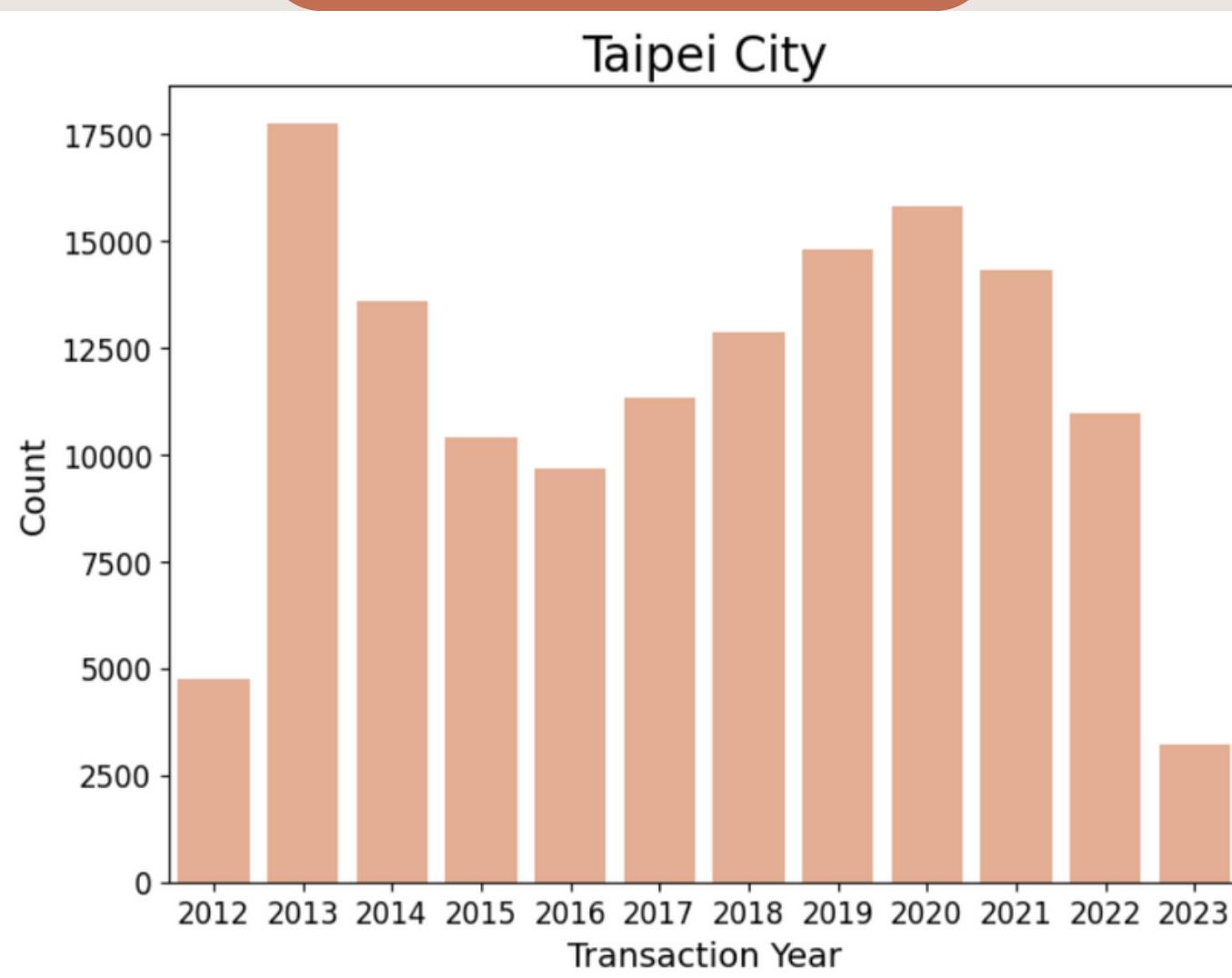
刪除離群值 (IQR的1.5倍)



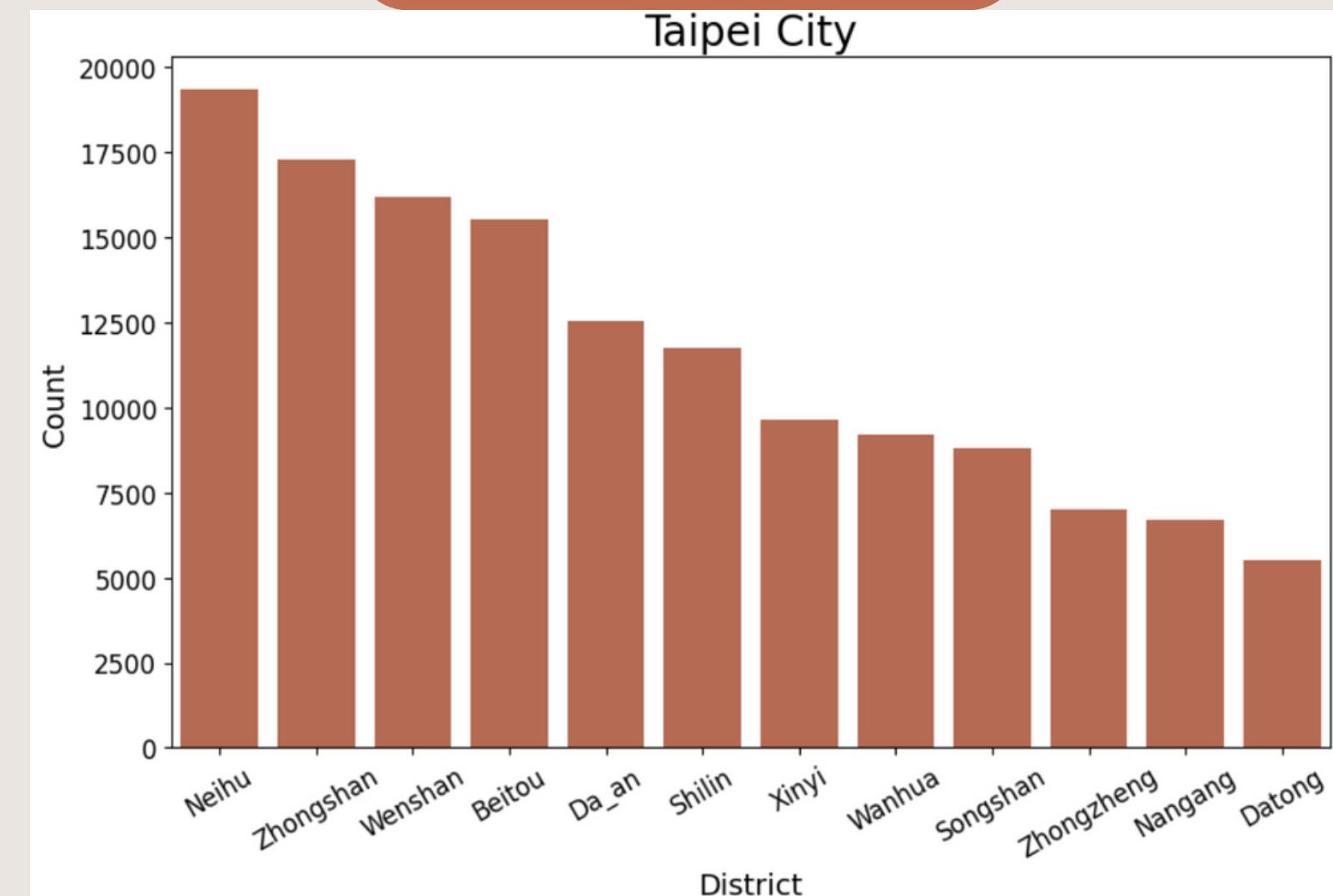
128148

資料分析-台北市

各年交易 (台北市)



各區域筆數 (台北市)

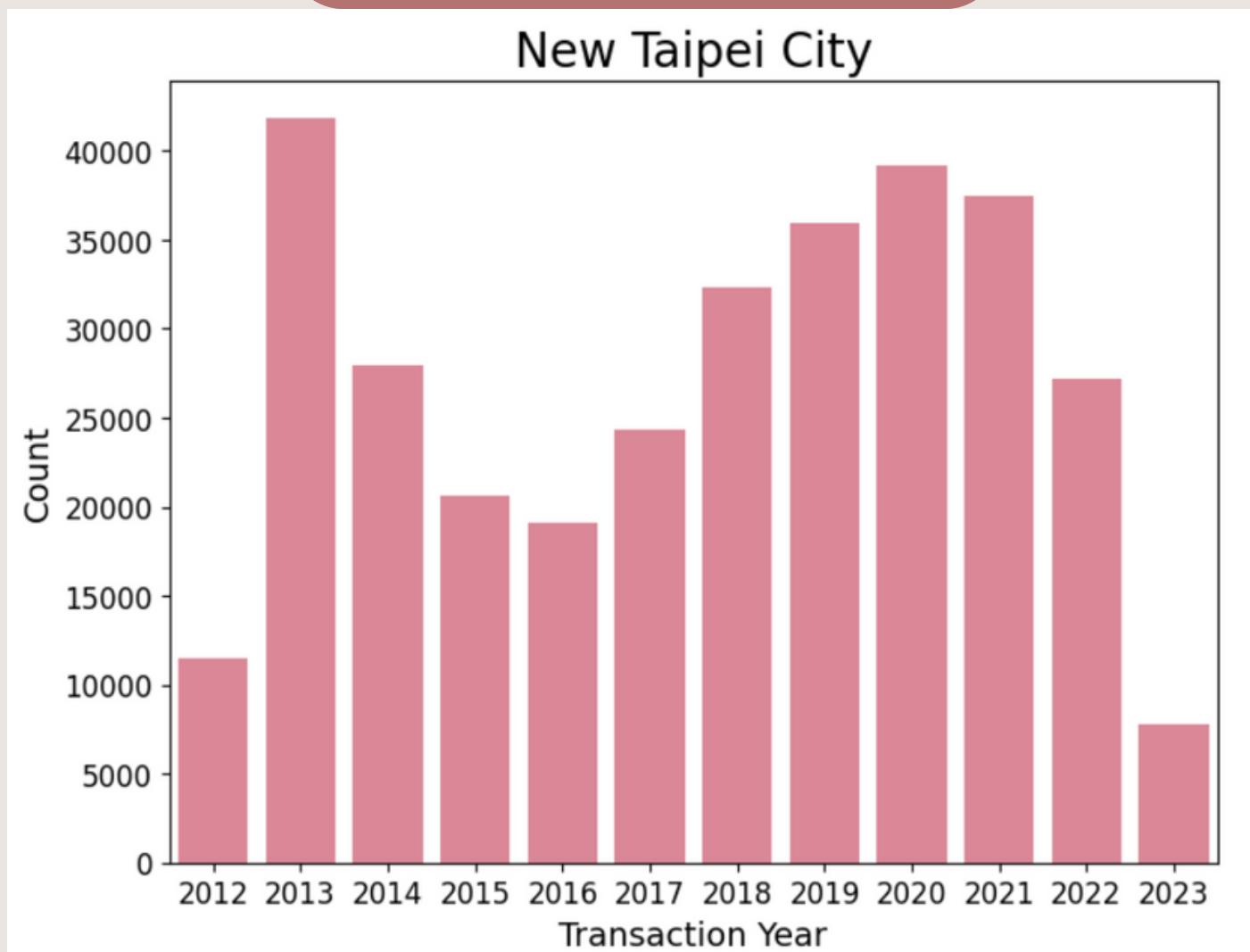


資料分析-新北市

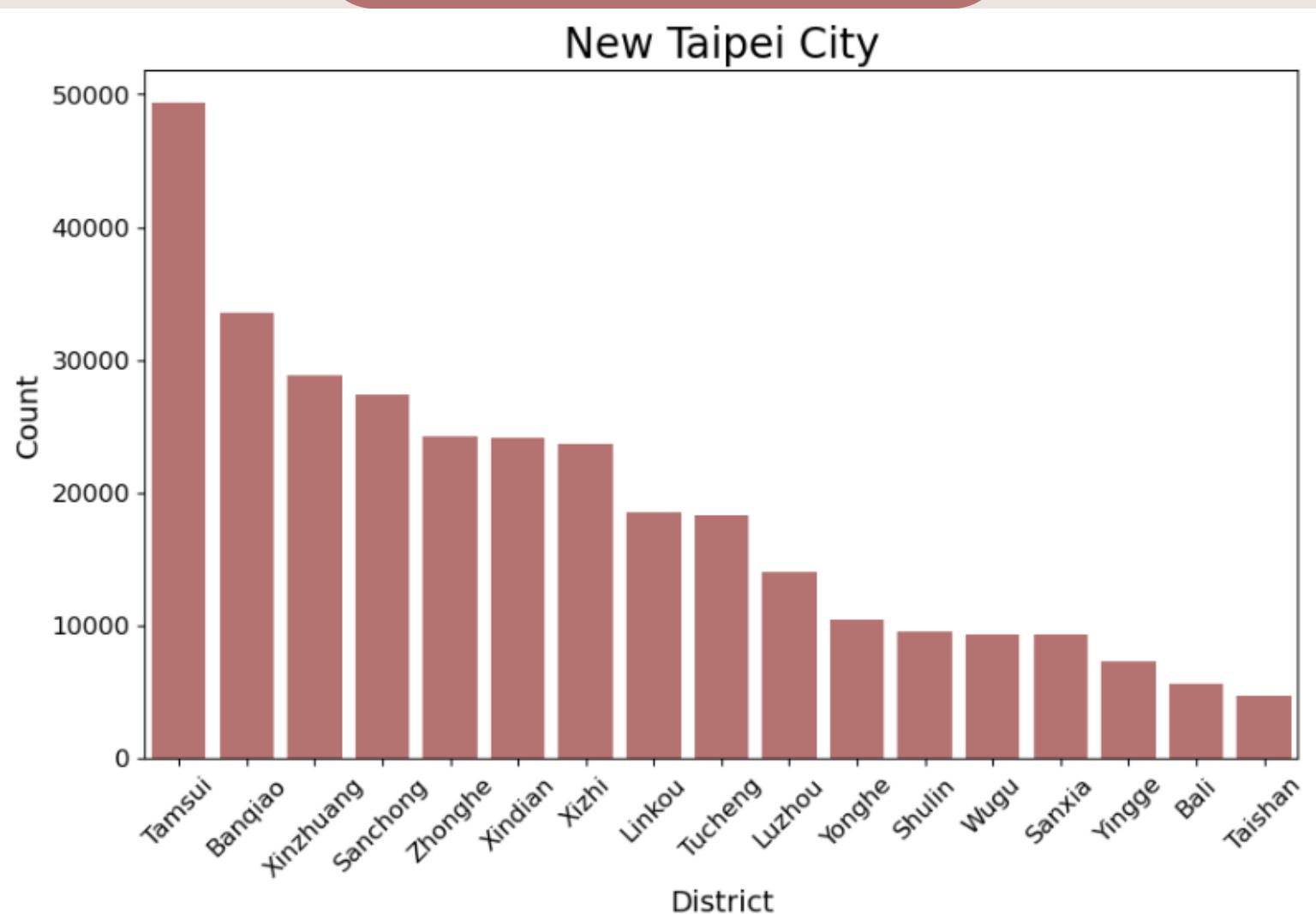
刪除筆數過少的區域

三芝、深坑、萬里、金山、瑞芳、貢寮、
石門、石碇、雙溪、烏來、坪林、平溪

各年交易 (新北市)

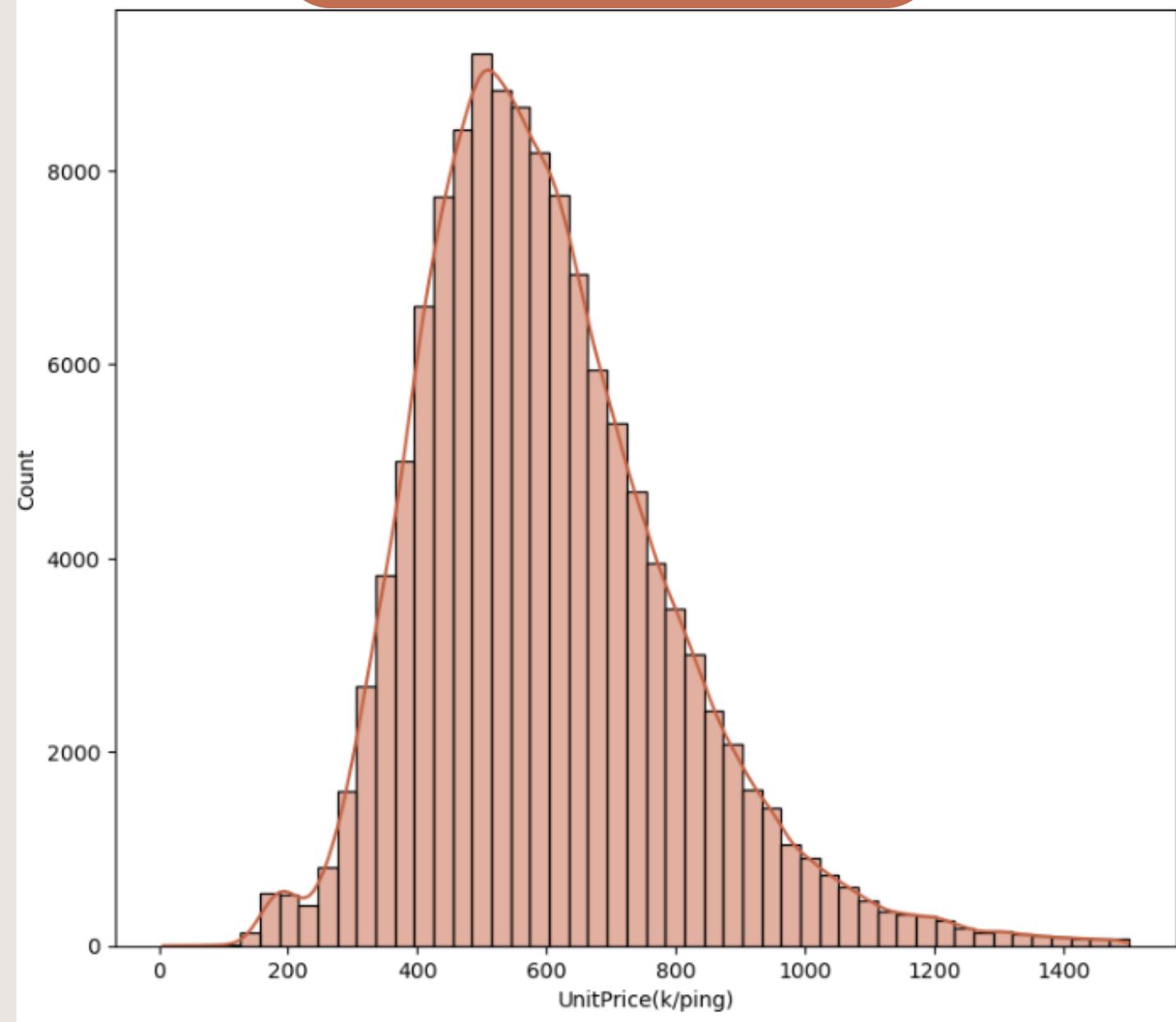


各區域筆數 (新北市)



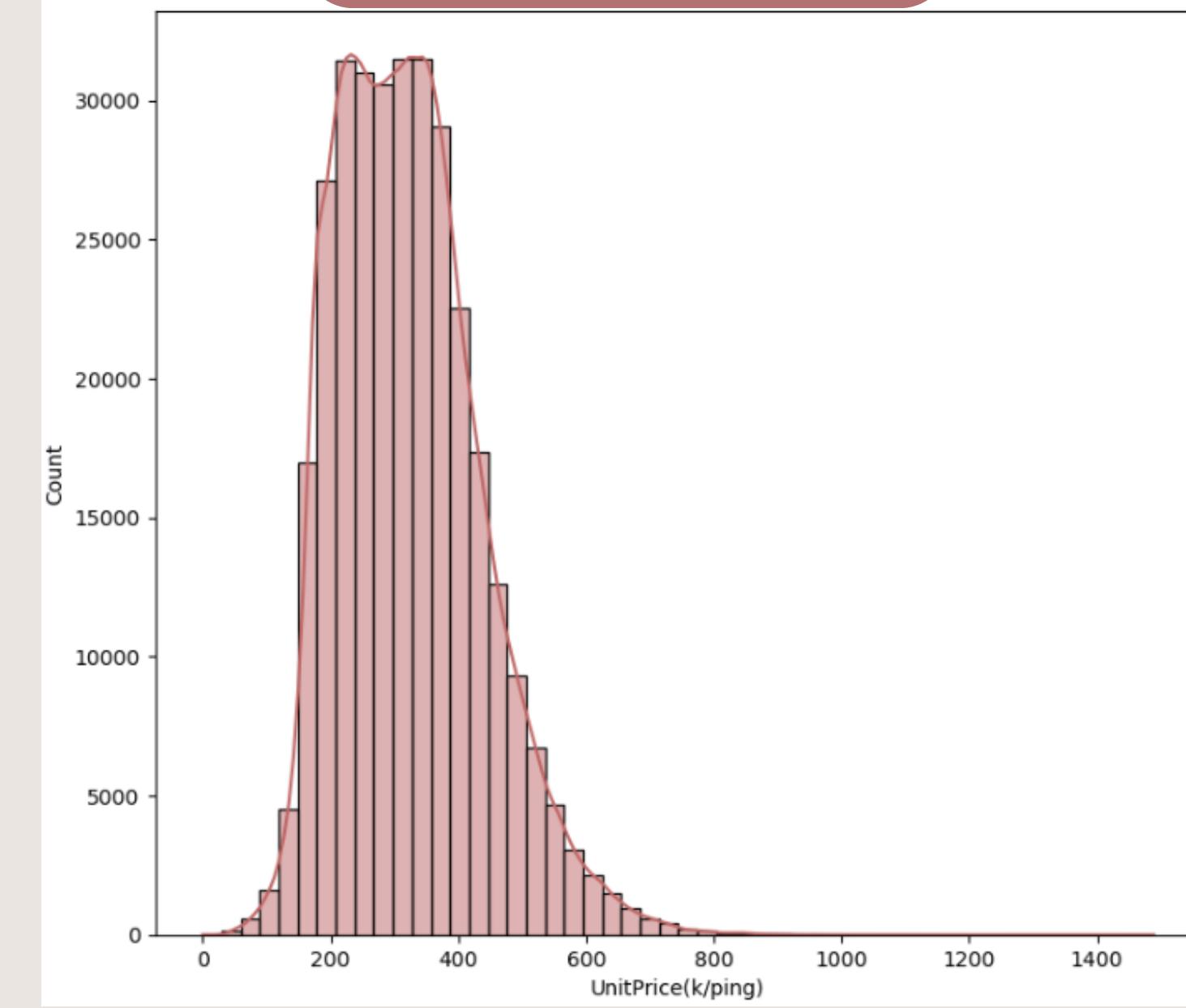
資料分析

台北市 (千/坪)



127886

新北市 (千/坪)

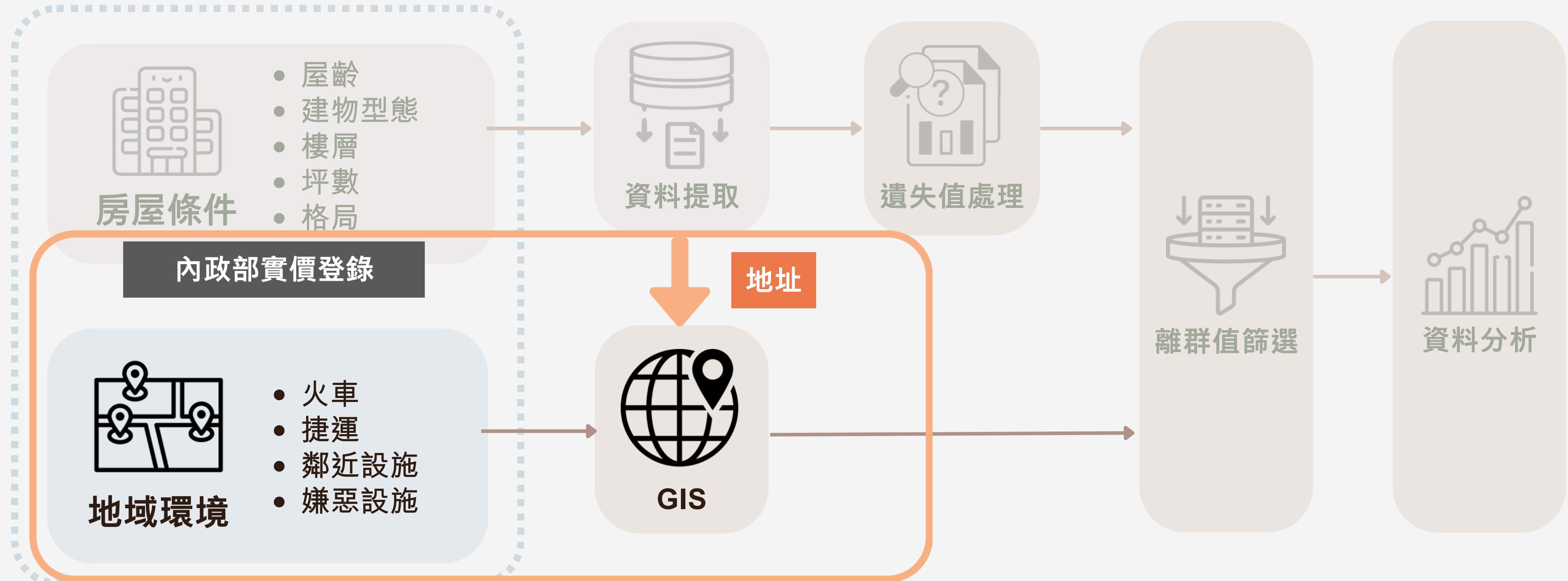


318335



資料清理流程

影響房價因素



資料清理-地址

原地址

新生北路三段 84 巷 40 號十樓 之 4, 之 5

僅示意，非實際地址

移除空格、標點符號

新生北路三段 84 巷 40 號十樓之 4 之 5

全形轉半形

新生北路三段 84 巷 40 號十樓之 4 之 5

數字轉換 (一/壹 -> 1)

新生北路三段 84 巷 40 號 10 樓之 4 之 5

補上「oo市oo區」

台北市中山區新生北路三段 84 巷 40 號 10 樓之 4 之 5

統一規格(號/樓)

台北市中山區新生北路三段 84 巷 40 號

修改後

台北市中山區新生北路三段 84 巷 40 號



地域資料蒐集

經緯度及距離計算



簡宏晉

地域資料蒐集



QGIS為一款免費且開源的
地理資訊系統

- 處理多種地理數據
- 地圖製作
- 空間分析
- 地圖投影和坐標系統支持
- 3D 地圖視覺化
- 插件和擴展性

→ [qgis2web](#)



地域資料蒐集

SELENIUM – 地址轉經緯度

雙北(約53萬筆)

- 台北
- 新北

雙北周遭物件(約800筆)

- 焚化廠
- 靈骨塔
- 醫院
- 國小
- 國中
- 捷運站
- 火車站

使用Selenium 將地址輸入googlemap查詢，
以Regular expression URL上面的經緯度



地域資料蒐集

SELENIUM – 地址轉經緯度

爬取成果

25000筆資料須5個小時

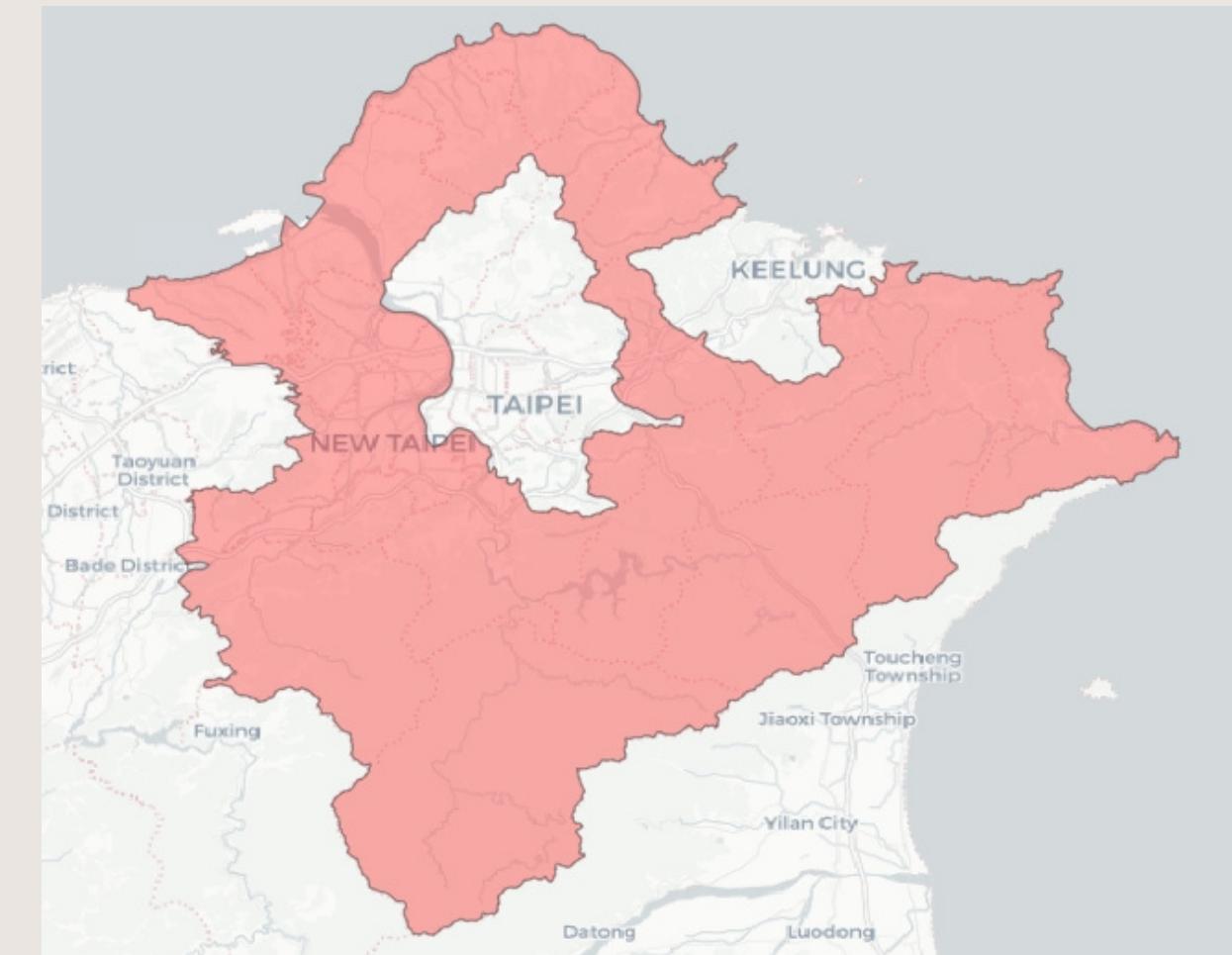
| All_tpe_AddressOnly_with_coords_cleaned.csv | X |
|---|--|
| 1 | Identifier, District, Address, latitude, longitude |
| 2 | RPQNMLMJJJIHFFAA98CA, 文山區, 台北市文山區興隆路四段145巷25號, 24.9853764, 121.5602683 |
| 3 | RPSNMLMJJJIHFFAA56CA, 文山區, 台北市文山區指南路三段22巷8弄5號, 24.9833289, 121.5765174 |
| 4 | RPQNMLNJJJIHFFAA09CA, 文山區, 台北市文山區文山區木新路3段50巷7弄11號, 24.9828688, 121.5611521 |
| 5 | RPXOMLQJJIHFFAA57CA, 文山區, 台北市文山區興隆路二段233巷5弄18號, 25.0022851, 121.5477942 |
| 6 | RPUOMLNLLJIHFFAA17CA, 文山區, 台北市文山區羅斯福路五段218巷38弄15號, 25.0026921, 121.5346292 |
| 7 | RPOOMLOKLHIFFAA76CA, 文山區, 台北市文山區萬盛街142號, 25.0054998, 121.5396215 |
| 8 | RPSNMLMJJJIHFFBA66CA, 大同區, 台北市大同區長安西路322號, 25.0519686, 121.5089523 |
| 9 | RPQNMLTJJJIHFFBA17CA, 萬華區, 台北市萬華區富民路155巷37號, 25.0184649, 121.4984255 |
| 10 | RPWPMLMJJJIHFFCA37CA, 內湖區, 台北市內湖區內湖路二段103巷90號, 25.0830397, 121.5778564 |
| 11 | RPTQMLMJJJIHFFCA17CA, 中山區, 台北市中山區建國北路一段150號, 25.0514707, 121.5338171 |



地域資料蒐集

SELENIUM – 地址轉經緯度

利用QGIS描繪雙北邊界



地域資料蒐集

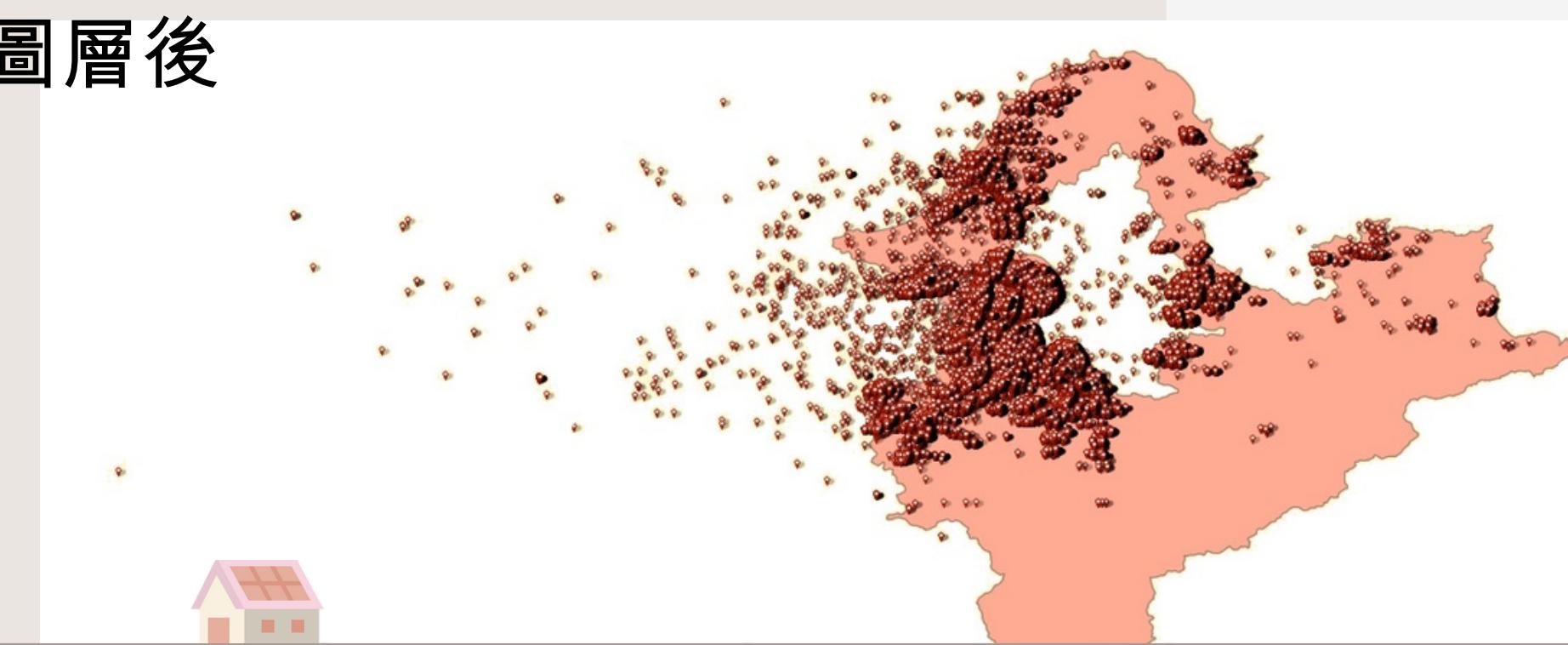
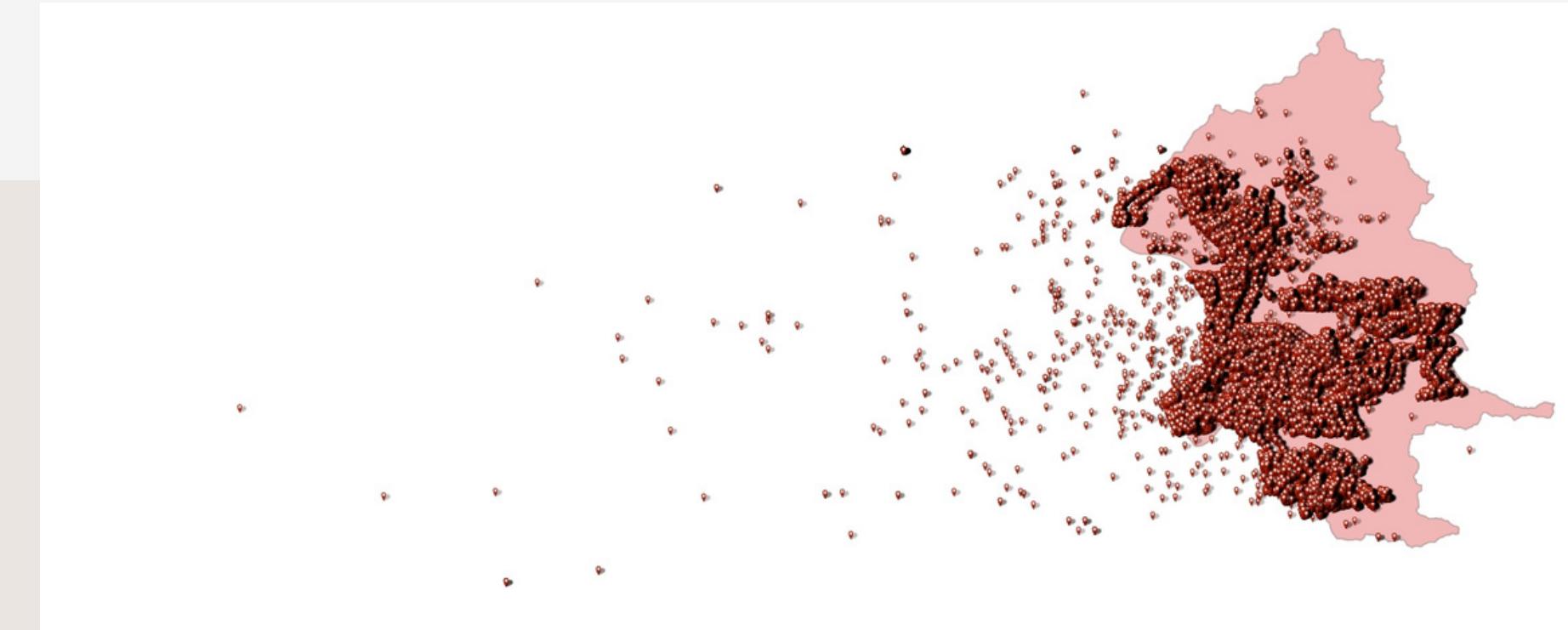
SELENIUM – 地址轉經緯度

疊圖成果

台北 12773筆在圖層外
新北 10420筆在圖層外

發現問題：

以 Selenium 查詢經緯度，套圖層後
許多地點位置嚴重偏移。



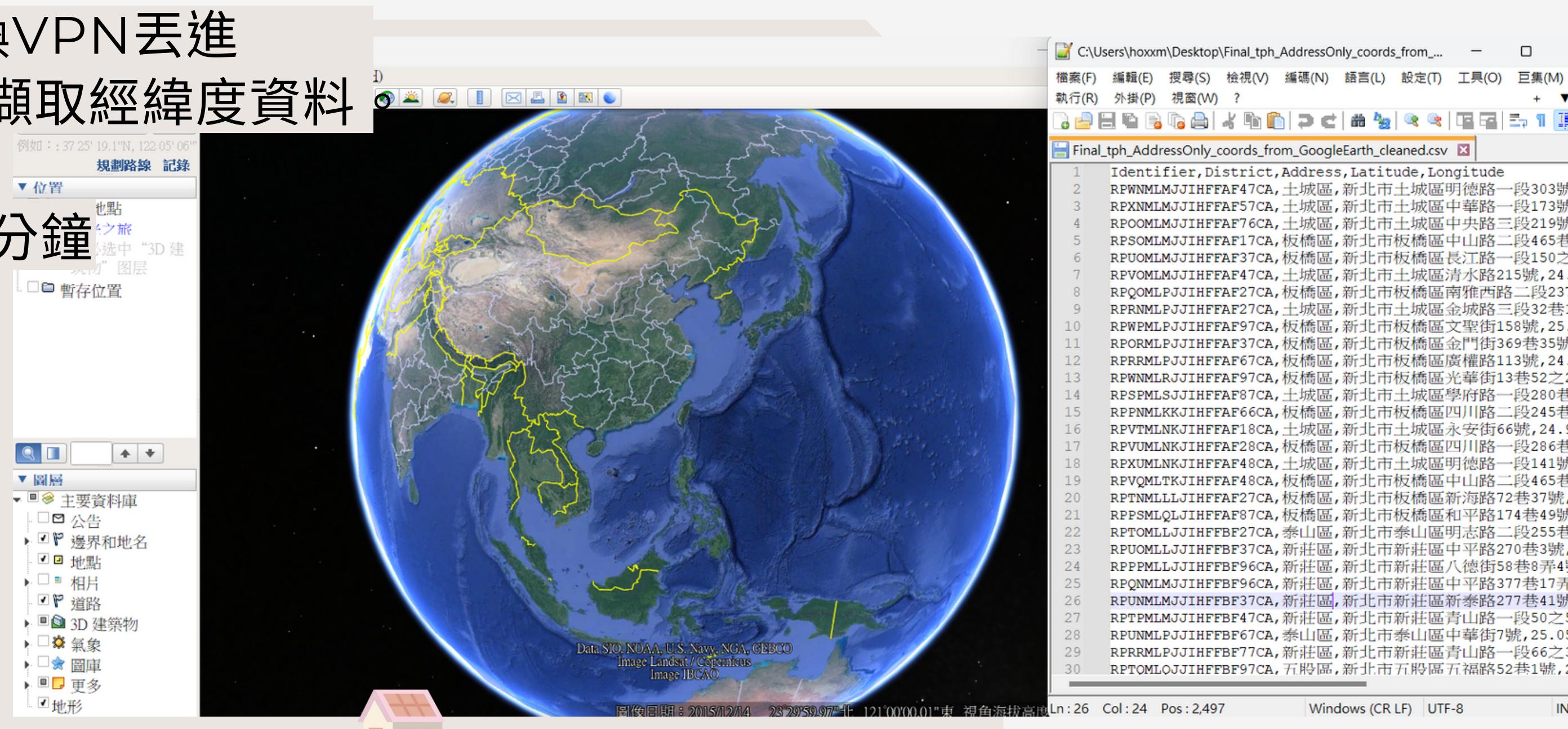
地域資料蒐集

GOOGLE EARTH – 地址轉經緯度

解決方法：

將原始檔案以每2500筆分割成多個
新檔案，不斷切換VPN丟進
GoogleEarth以擷取經緯度資料

2500筆資料約10分鐘



地域資料蒐集

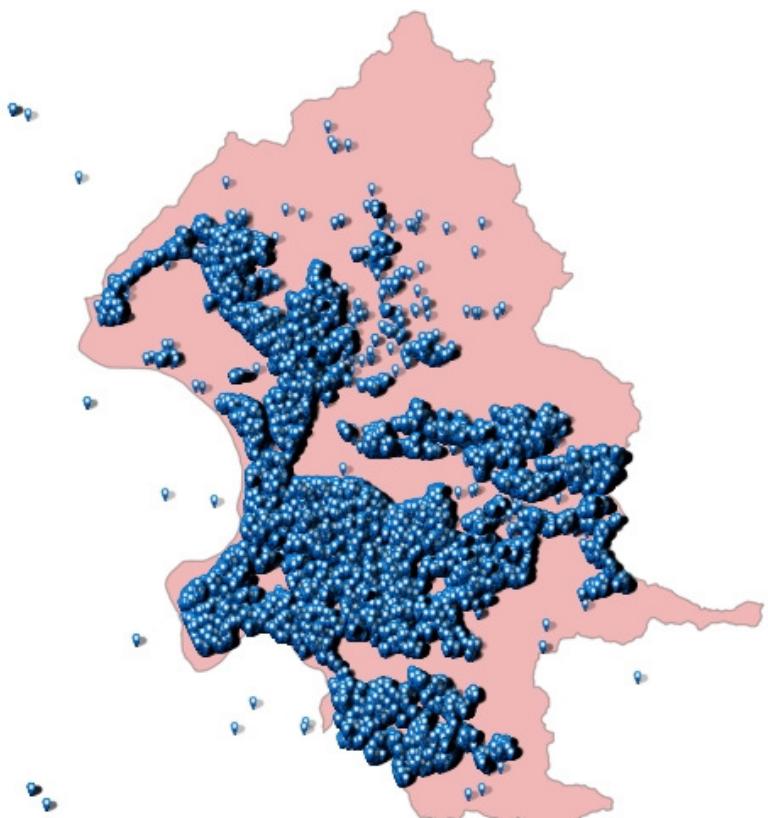
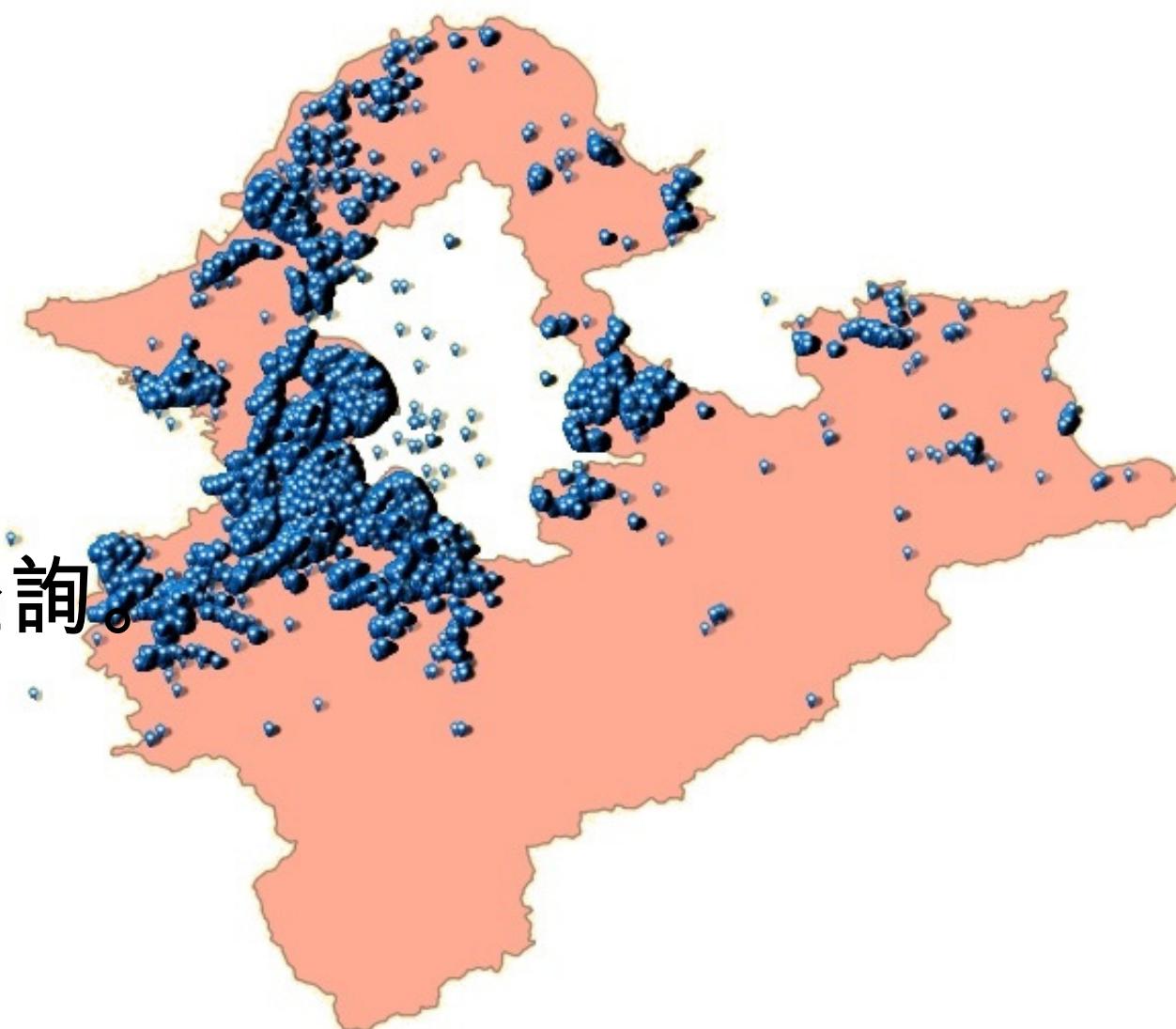
GOOGLE EARTH – 地址轉經緯度

疊圖成果

新北 約170筆在圖層外
台北 約30筆在圖層外

剩餘部分零星因地址格式，導致
經緯度讀取失敗的筆數，以手動查詢。

EX: 臺北市信義區忠孝東路5段



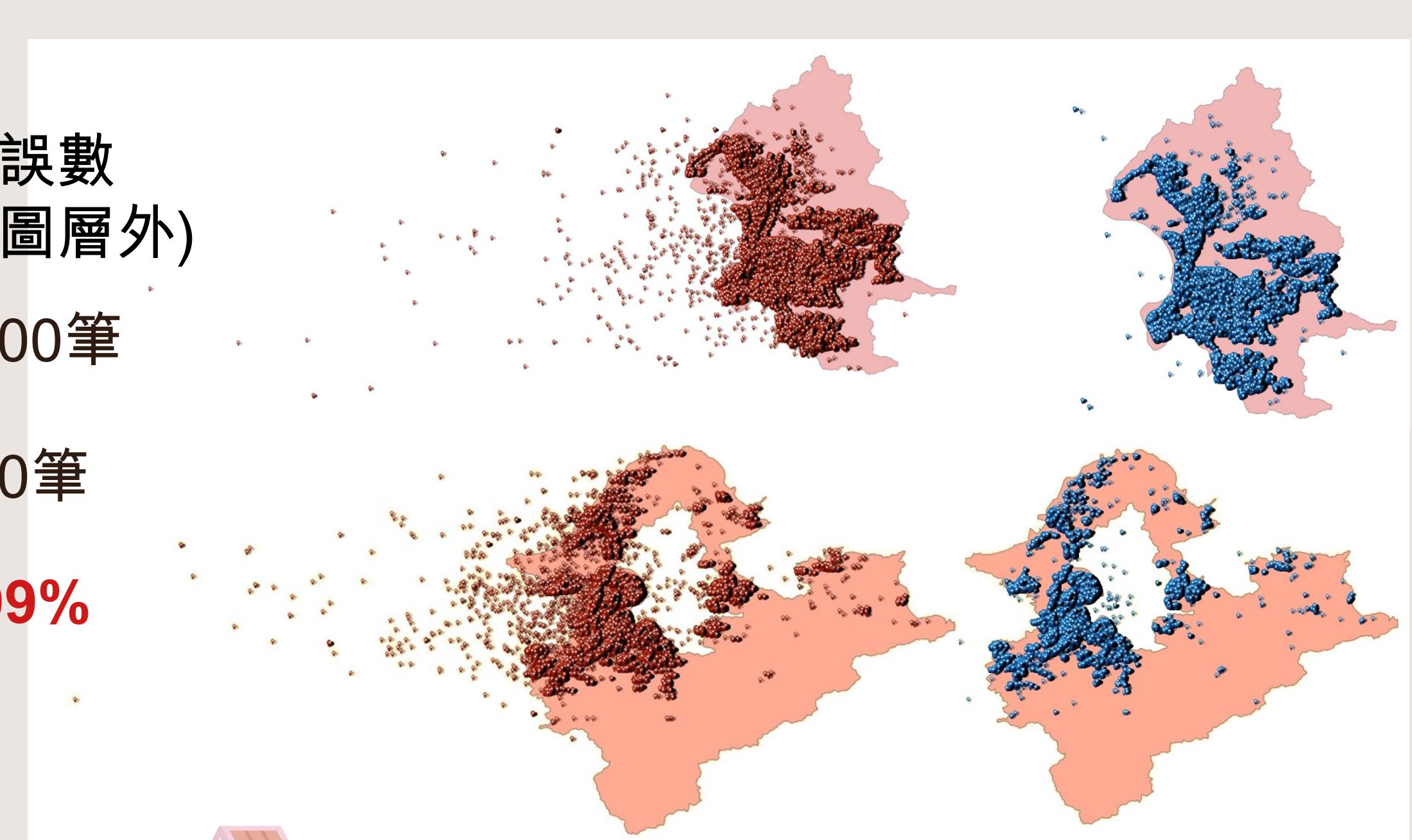
地域資料蒐集

成果比較 - 地址轉經緯度

BEFORE / AFTER

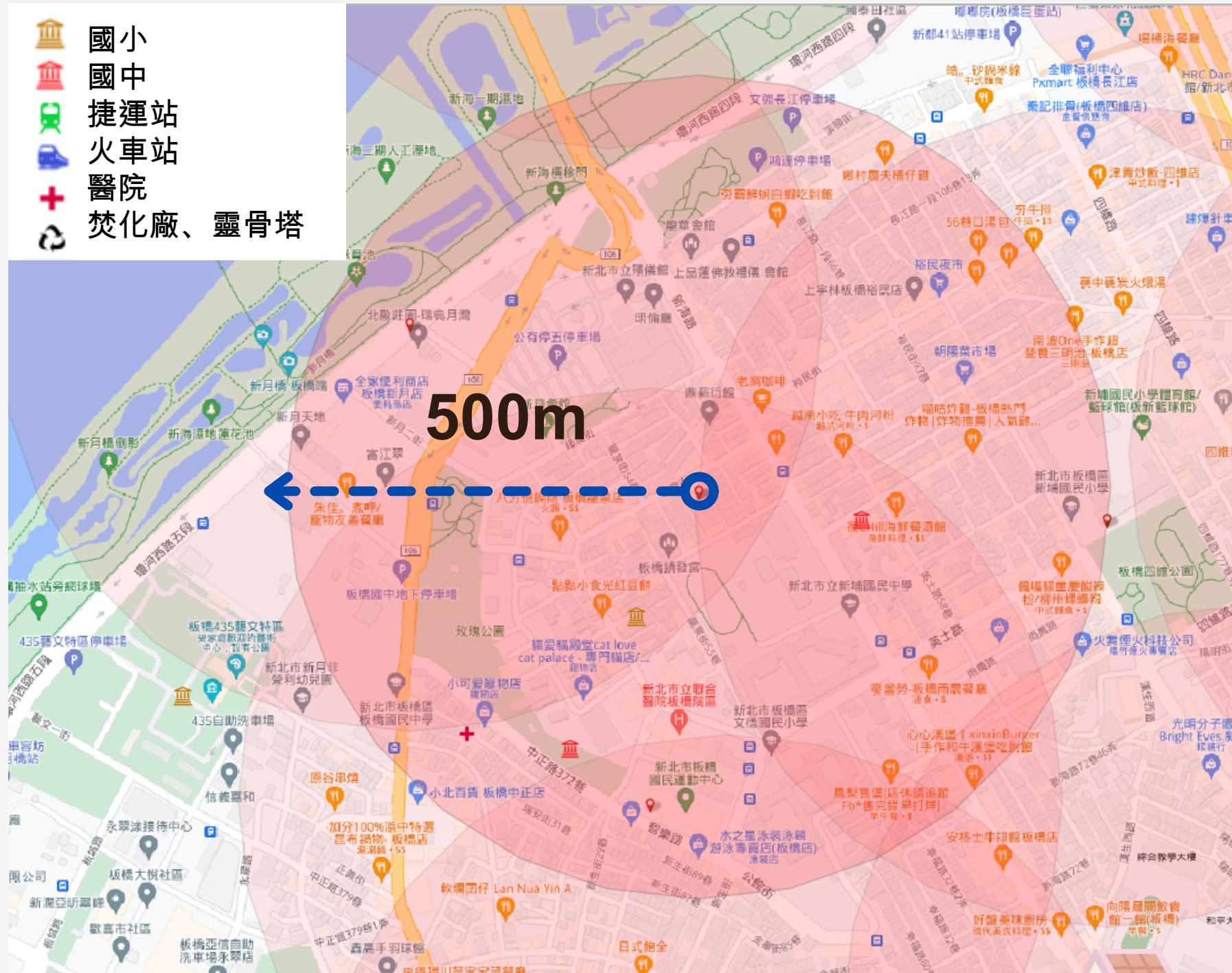
花費時間
(每25000筆資料) 錯誤數
(位於圖層外)

| | | |
|--------------|-------|--------|
| Selenium | 600分鐘 | 23000筆 |
| Google Earth | 50分鐘 | 200筆 |
| | ↓ 92% | ↓ 99% |



地域資料蒐集

500M內設施個數(步行10分鐘內可到達)



最近設施距離

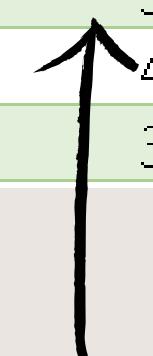


地域資料蒐集

500M內設施個數 / 最近設施距離

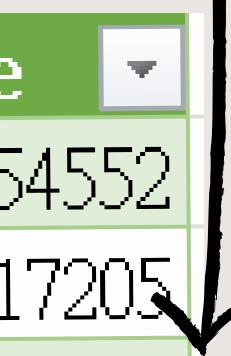
| Identifier | District | Address | latitude | longitude | NUMPOINTS |
|----------------------|----------|------------------------|------------|-------------|-----------|
| RPQNMLMJIHFFAA98CA | 文山區 | 台北市文山區興隆路四段145巷25號 | 24.9853764 | 121.5602683 | 0 |
| RPSNMLMJIHFFAA56CA | 文山區 | 台北市文山區指南路三段22巷8弄5號 | 24.9833289 | 121.5765174 | 0 |
| RPQNMLNJIHFFAA09CA | 文山區 | 台北市文山區文山區木新路3段50巷7弄11號 | 24.9828688 | 121.5611521 | 0 |
| RPXOMLQJJIHFFAA57CA | 文山區 | 台北市文山區興隆路二段233巷5弄18號 | 25.0022851 | 121.5477942 | 0 |
| RPUOMILNLJIHFFAA17CA | 文山區 | 台北市文山區羅斯福路五段218巷38弄15號 | 25.0026921 | 121.5346292 | 3 |
| RPOOMLOKLHIFFAA76CA | 文山區 | 台北市文山區萬盛街142號 | 25.0054998 | 121.5396215 | 4 |
| RPSNMLMJIHFFBA66CA | 大同區 | 台北市大同區長安西路322號 | 25.0519686 | 121.5089523 | 3 |

距離最近的捷運站為大坪林，1811.公尺



在500公尺內有3站捷運站

| InputID | TargetID | Distance |
|--------------------|----------|-------------|
| RPQNMLMJIHFFAA98CA | 萬芳醫院 | 1572.454552 |
| RPSNMLMJIHFFAA56CA | 動物園 | 1659.617205 |
| RPQNMLNJIHFFAA09CA | 大坪林 | 1811.609036 |



曾子維

模型建置

機器學習



曾子維

模型建置

模型選擇



dmlc
XGBoost

模型建置

模型選擇

台北市

| | Random Forest | XGBoost |
|-----------|---------------|---------|
| R-squared | 0.5564 | 0.7857 |
| MAE | 101.589 | 67.838 |
| 運行速度 | 58.7s | 37.9s |



模型建置

模型選擇

新北市

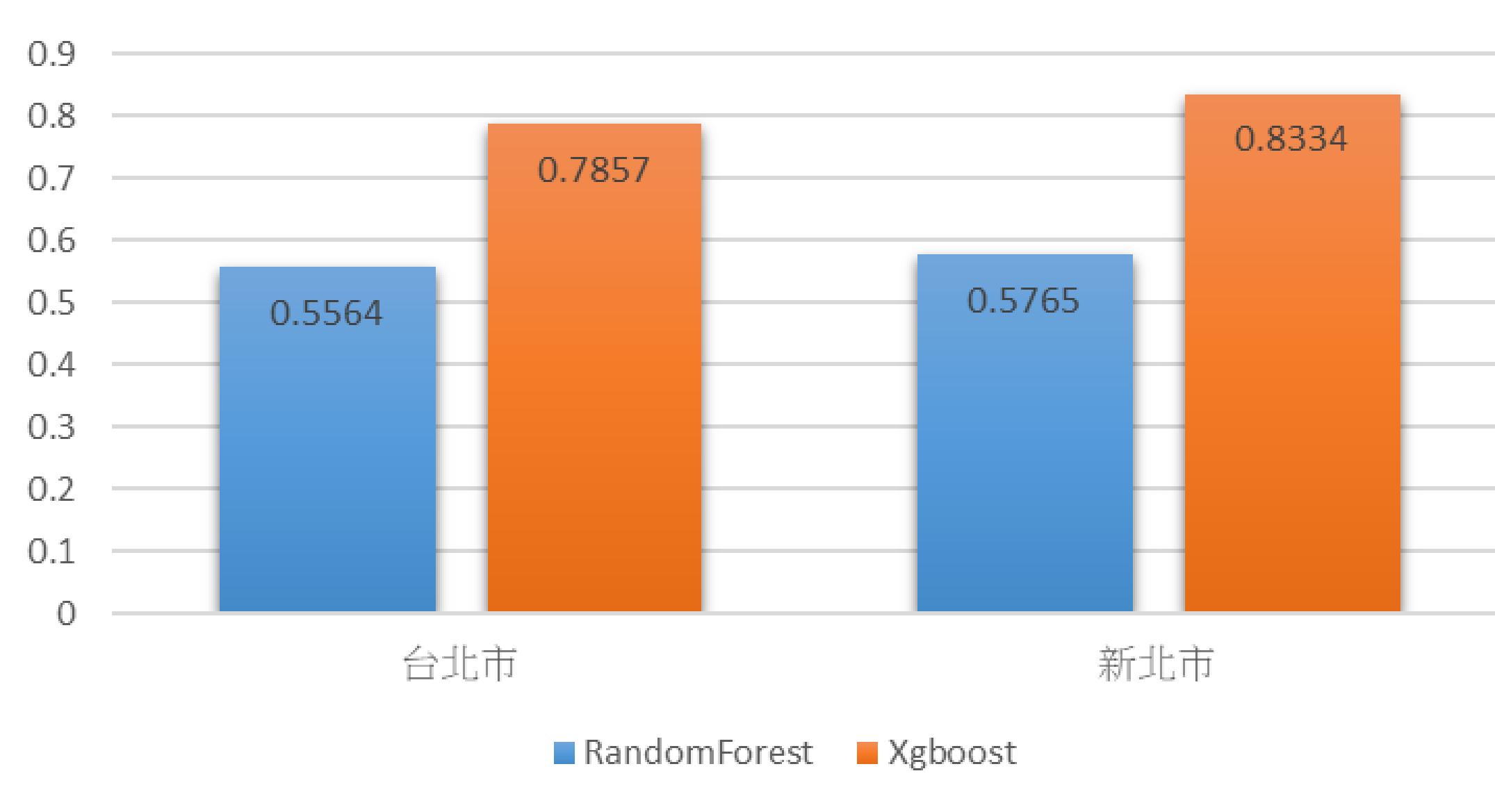
| | Random Forest | XGBoost |
|-----------|---------------|---------|
| R-squared | 0.5765 | 0.8334 |
| MAE | 55.92 | 33.22 |
| 運行速度 | 2m57.9s | 5.8 |



模型建置

模型選擇

R² Scores by Regressor



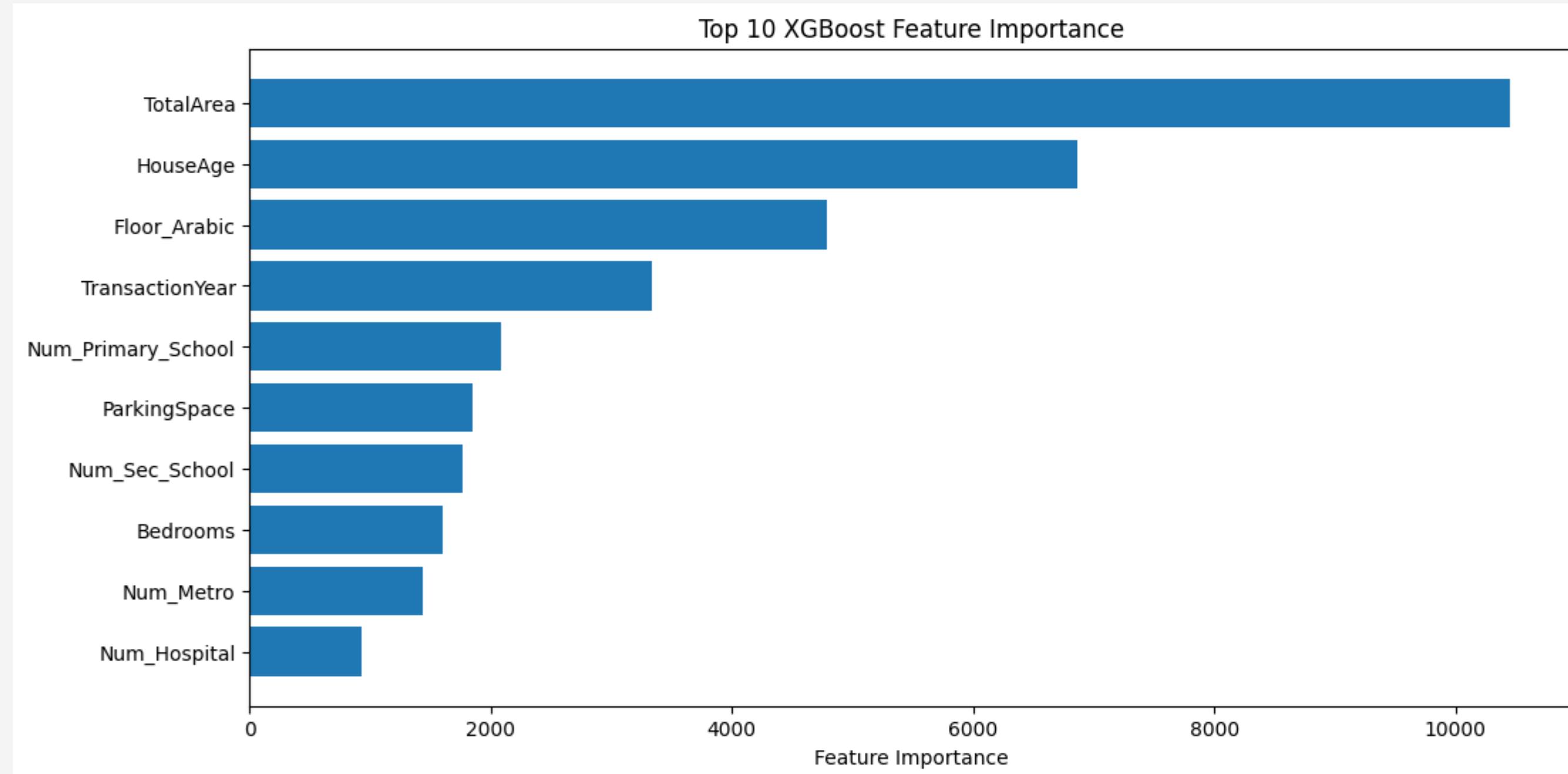
Xgboost優點：

- 速度快
- 靈活性高
- 防止過擬合

模型建置

模型解釋

台北市



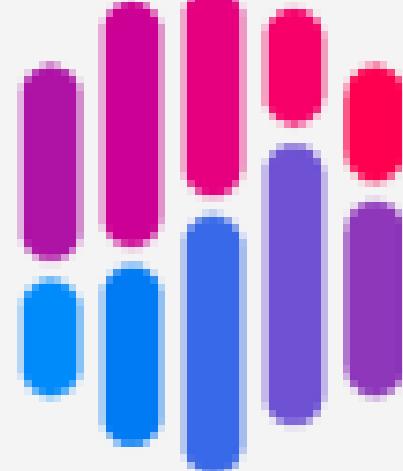
北市重要特徵：
 1. 總交易面積
 2. 屋齡
 3. 交易樓層

Feature Importance雖然可以顯示特徵對模型的重要程度,但沒有正負關係。

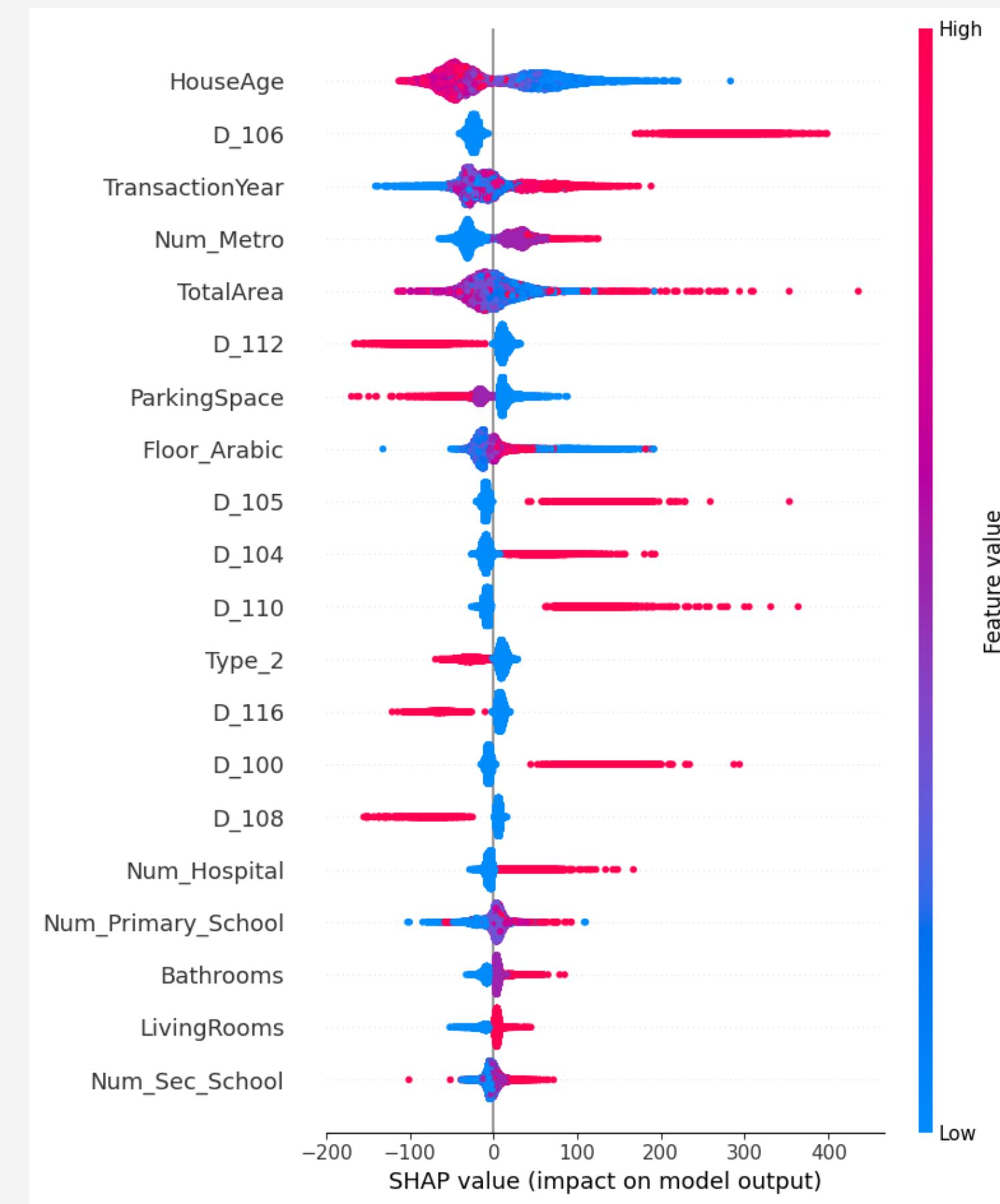
模型建置

台北市

模型解釋



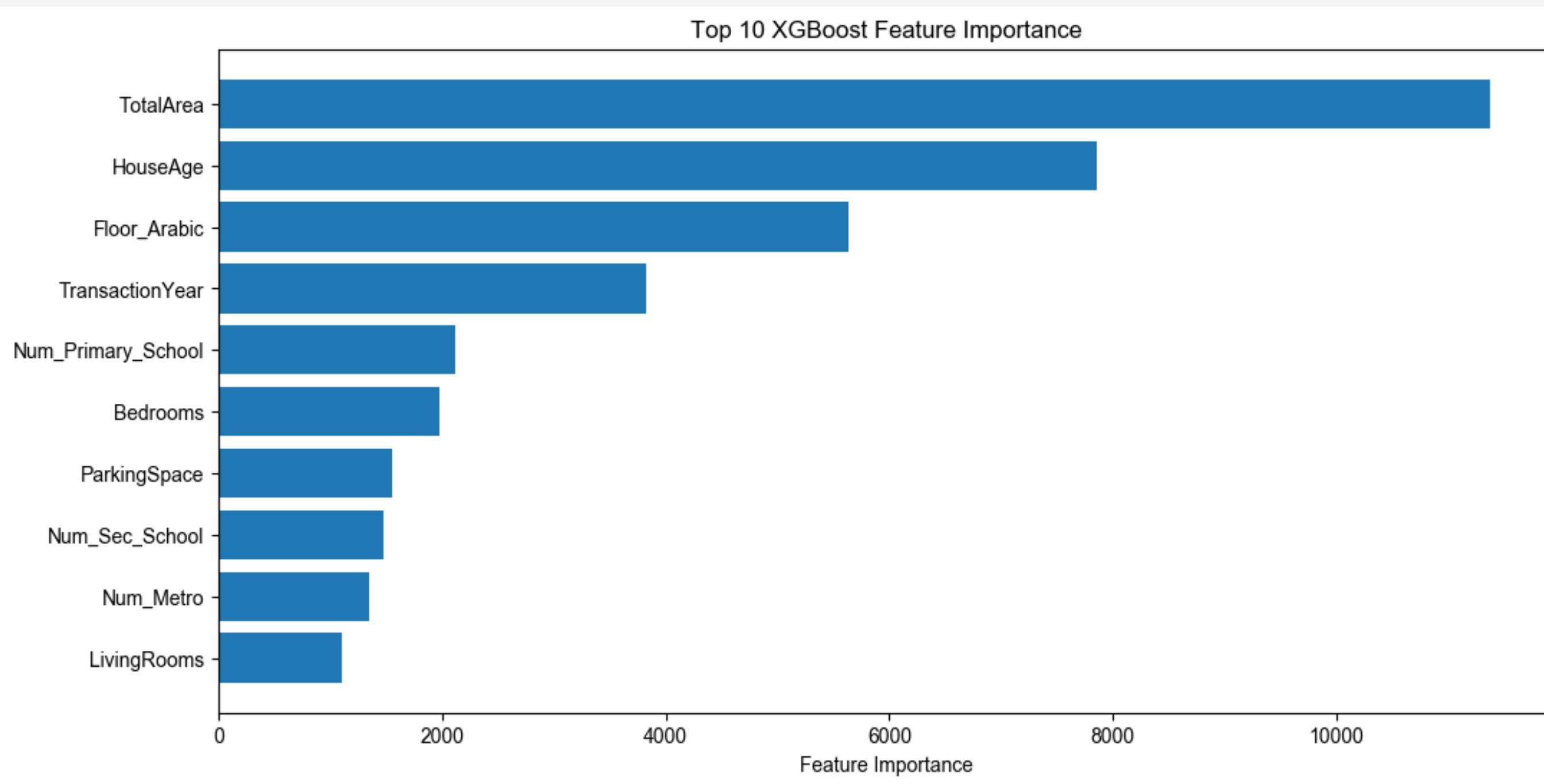
SHAP



模型建置

模型解釋

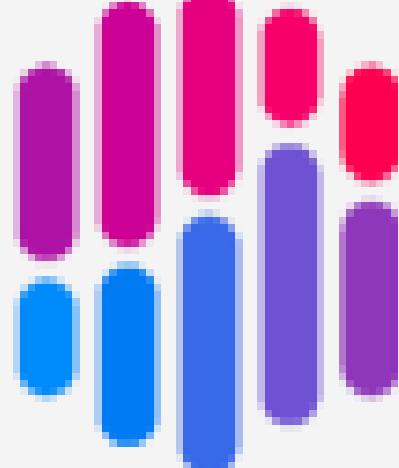
新北市



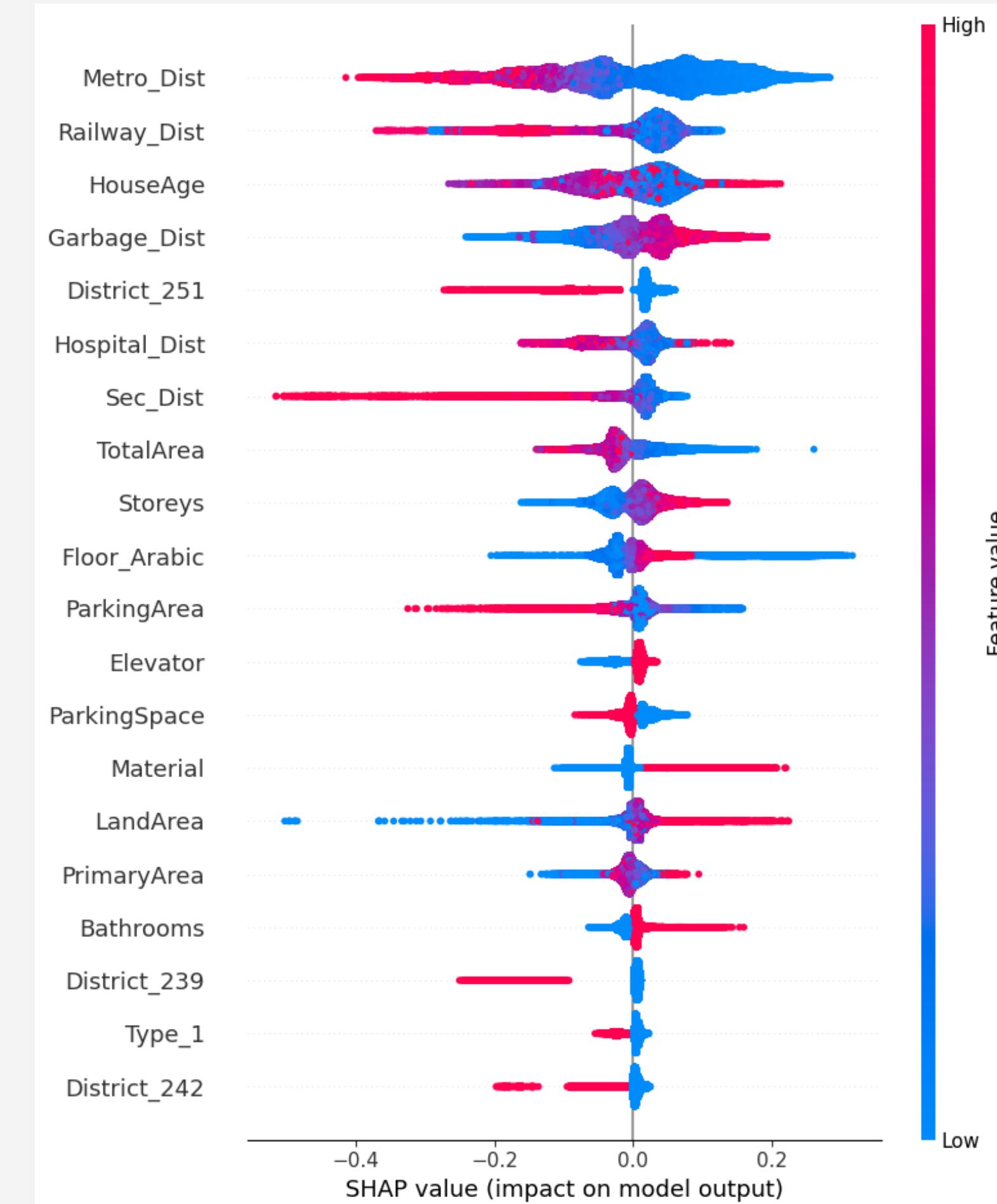
新北市重要特徵：
1. 總交易面積
2. 屋齡
3. 交易樓層

模型建置 新北市

模型解釋



SHAP

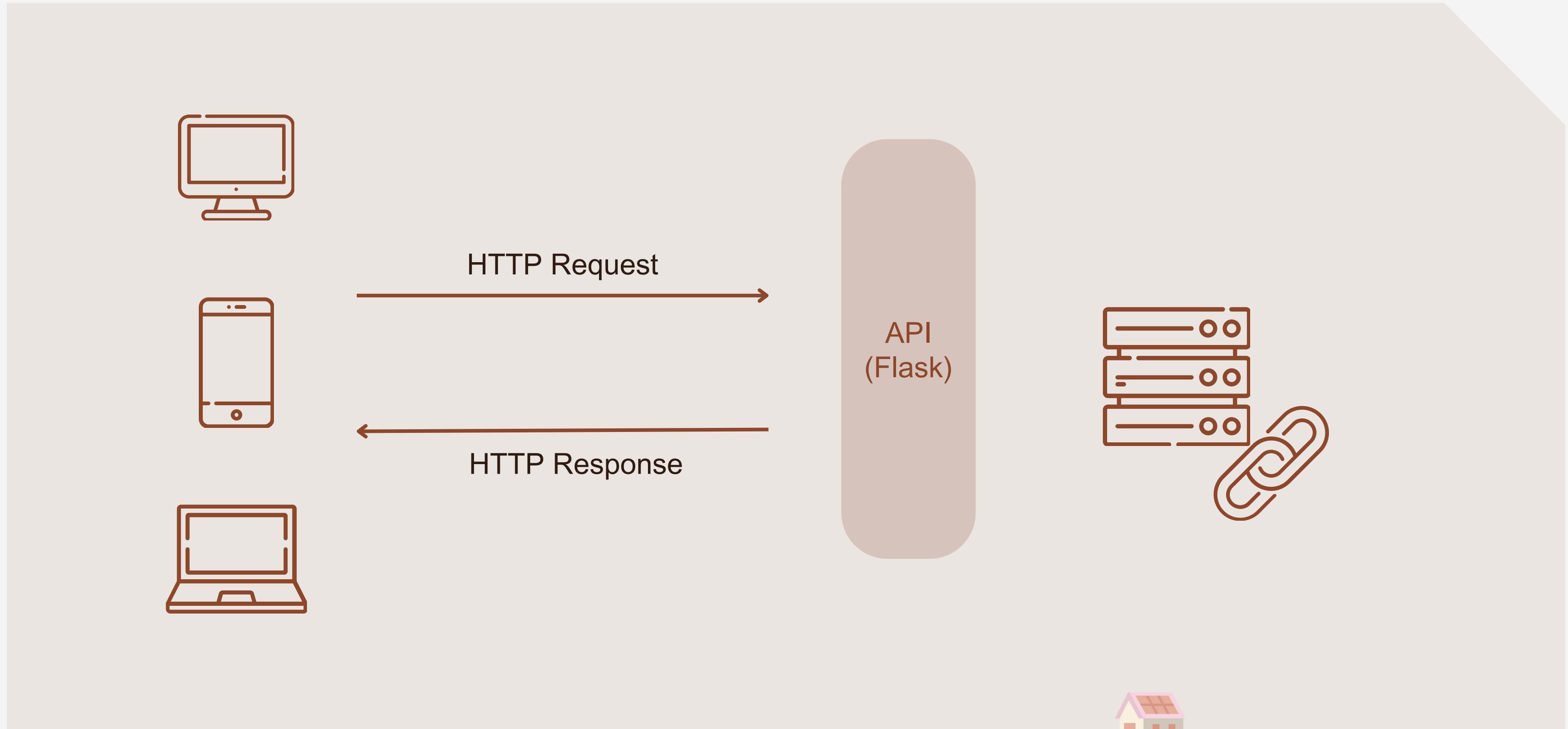


後端框架 網站應用與開發

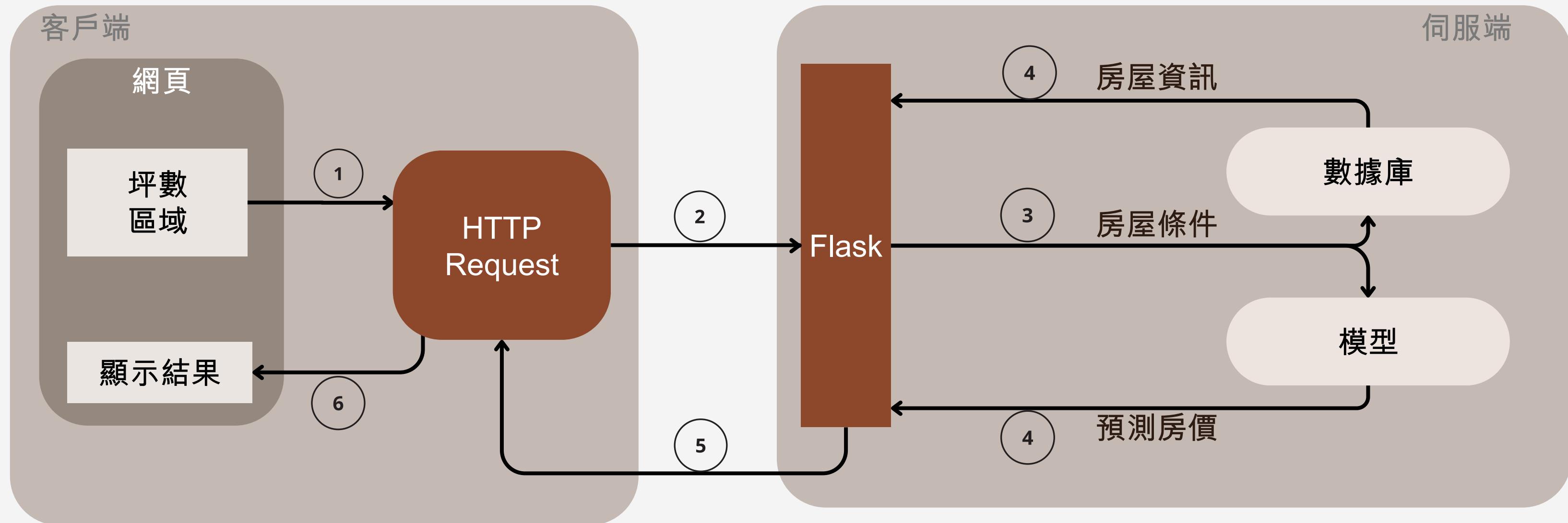


翁若芸

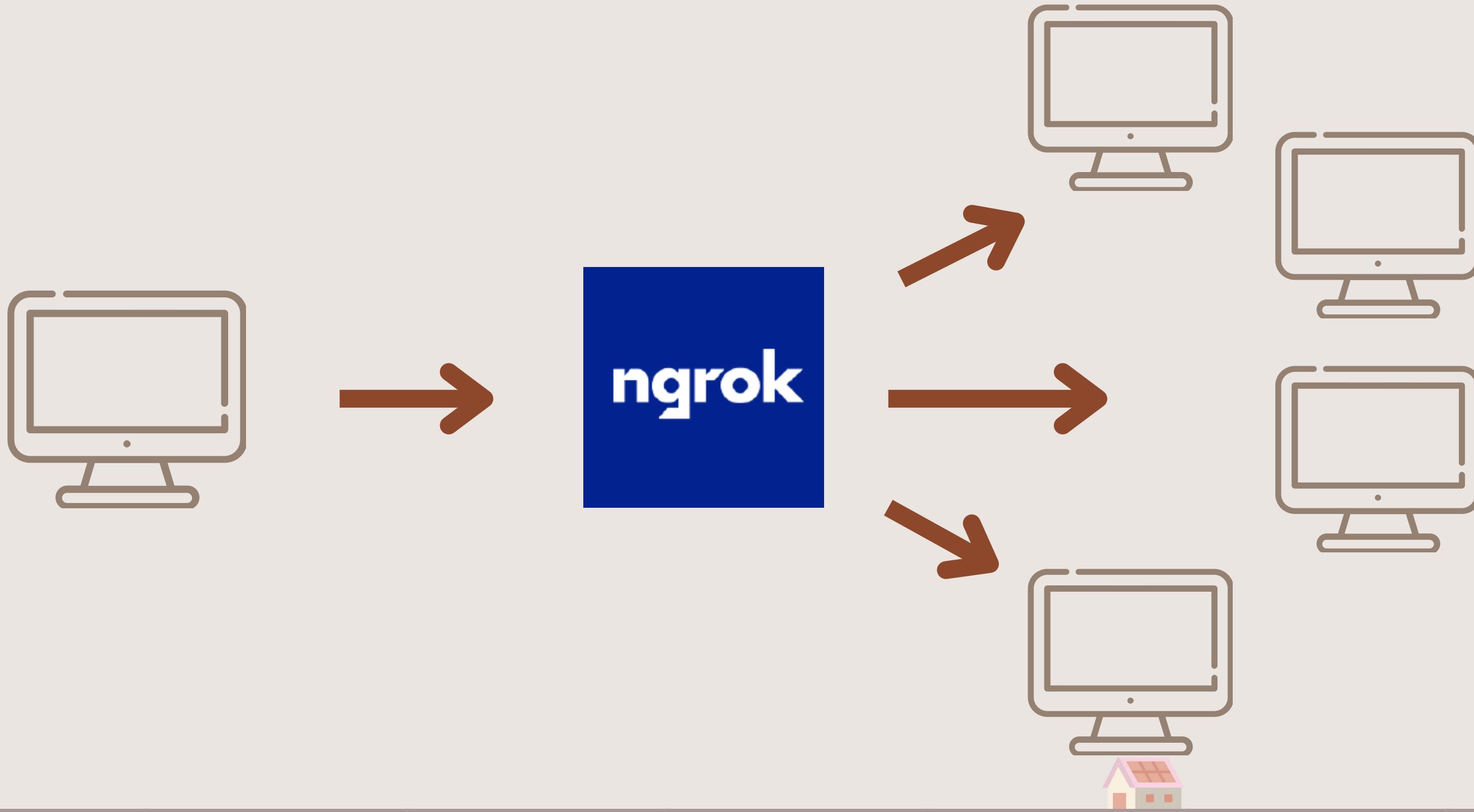
Flask



運作流程



網站架設



林文海



網頁視覺化



林文海



視覺化

雙北房價預測

臺北市

中正區

請選擇坪數

Please select an item in the list.

進階篩選

SUBMIT





視覺化

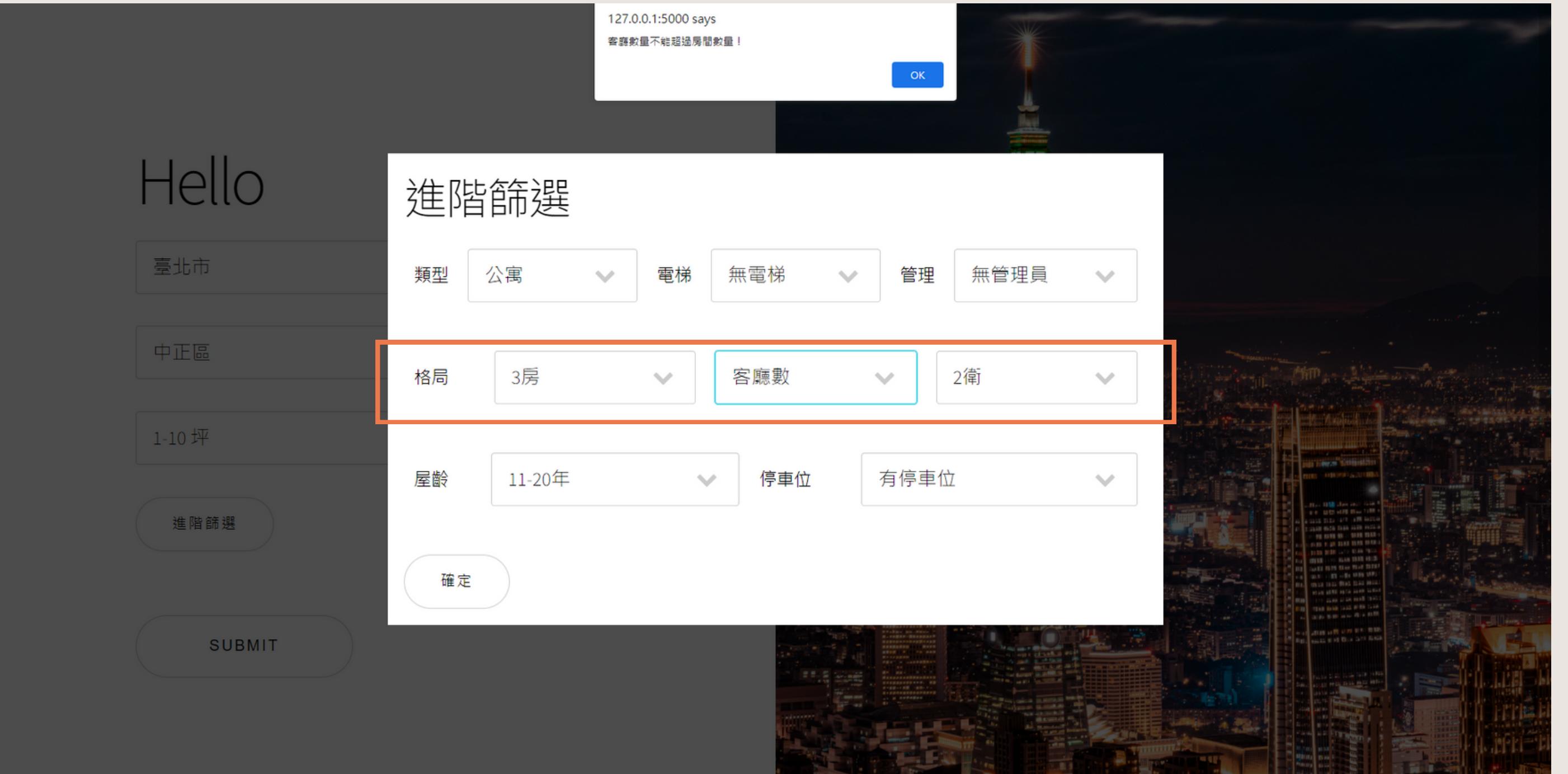
The screenshot shows a web application interface for searching real estate. In the background, there is a dark-themed search form with fields for location (臺北市), district (中正區), and area (1-10 坪). A prominent button labeled "進階篩選" (Advanced Filter) is visible. Overlaid on this is an "進階篩選" (Advanced Filter) dialog box. This dialog contains several dropdown menus and input fields:

- Top row: 電型 (Type) set to 公寓 (Apartment), 電梯 (Elevator) set to 無電梯 (No Elevator), and 管理 (Management) set to 無管理員 (No Manager).
- Middle row: 格局 (Layout) with options 7房 (7 Rooms), 5廳 (5 Halls), and 2衛 (2 Bathrooms).
- Bottom row: 屋齡 (Age) set to 11-20年 (11-20 years), and 停車位 (Parking) set to 有停車位 (With Parking).

A red rectangle highlights the top row of filters, specifically the "電梯" (Elevator) field which is set to "無電梯" (No Elevator). An error message box is displayed at the top of the screen, stating "127.0.0.1:5000 says 公寓通常無電梯，請重新選擇！" (127.0.0.1:5000 says Apartments usually have no elevators, please re-select!). An "OK" button is present in the error box.



視覺化





RWD響應式網頁設計 - HTML5UP

iphone 12 pro

Dimensions: iPhone 12 Pro ▾ 390 x 844 100% No throttling

Hello

臺北市

請選擇鄉鎮市區

請選擇坪數

進階篩選

SUBMIT

iPad Air

Dimensions: iPad Air ▾ 820 x 1180 70% No throttling

Hello

臺北市

請選擇鄉鎮市區

請選擇坪數

進階篩選

SUBMIT

Samsung Galaxy S20

Dimensions: Samsung Galaxy S20 Ultra ▾ 412 x 915 95% No throttling

Hello

臺北市

請選擇鄉鎮市區

請選擇坪數

進階篩選

SUBMIT

網頁導覽



- 1. 選擇縣市/區/坪數 (必填)**
- 1-1. 進階篩選(選填)**
- 2. Submit**
- 3. 重新查詢(回首頁)**



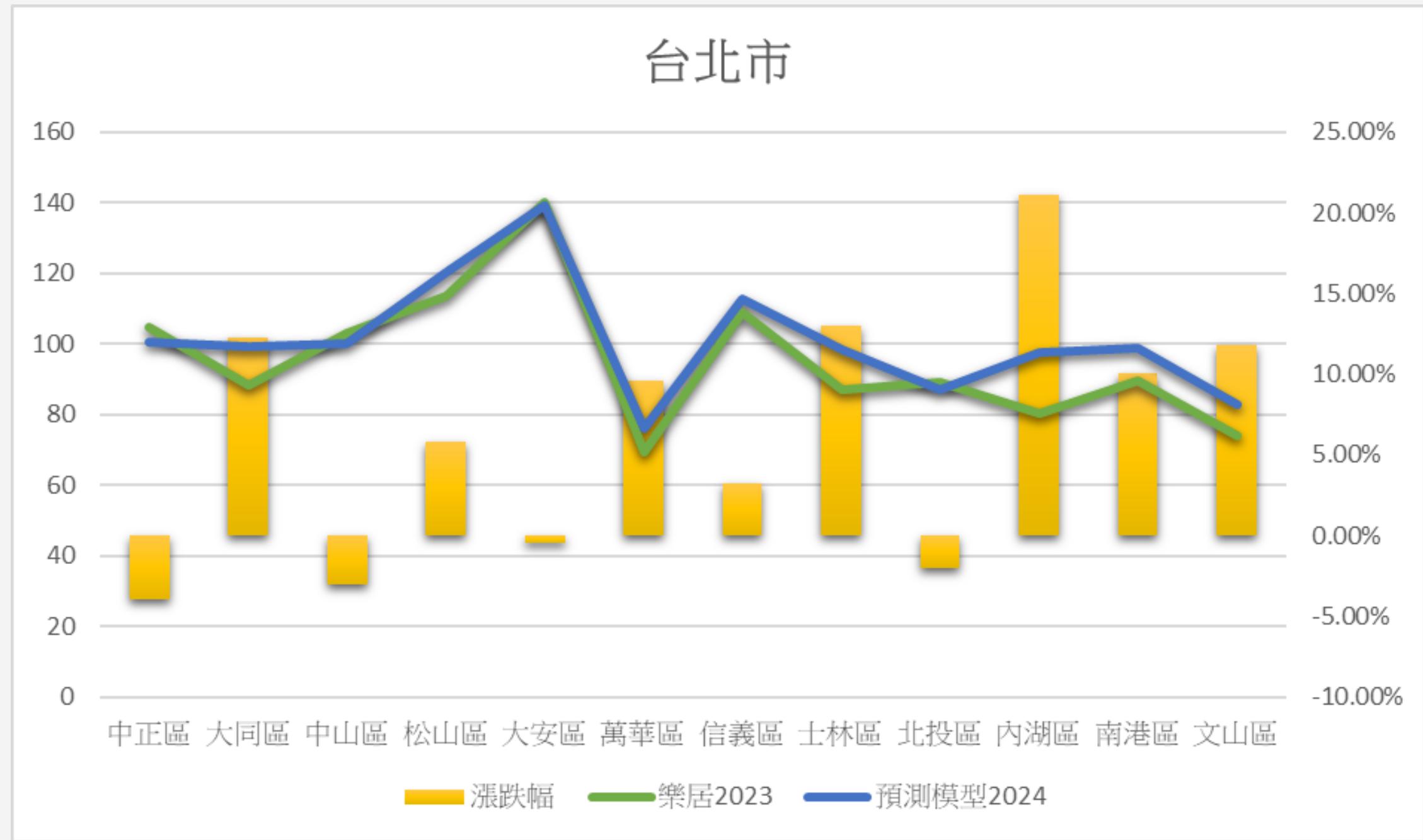
總結



曾子維

結論

成果分析

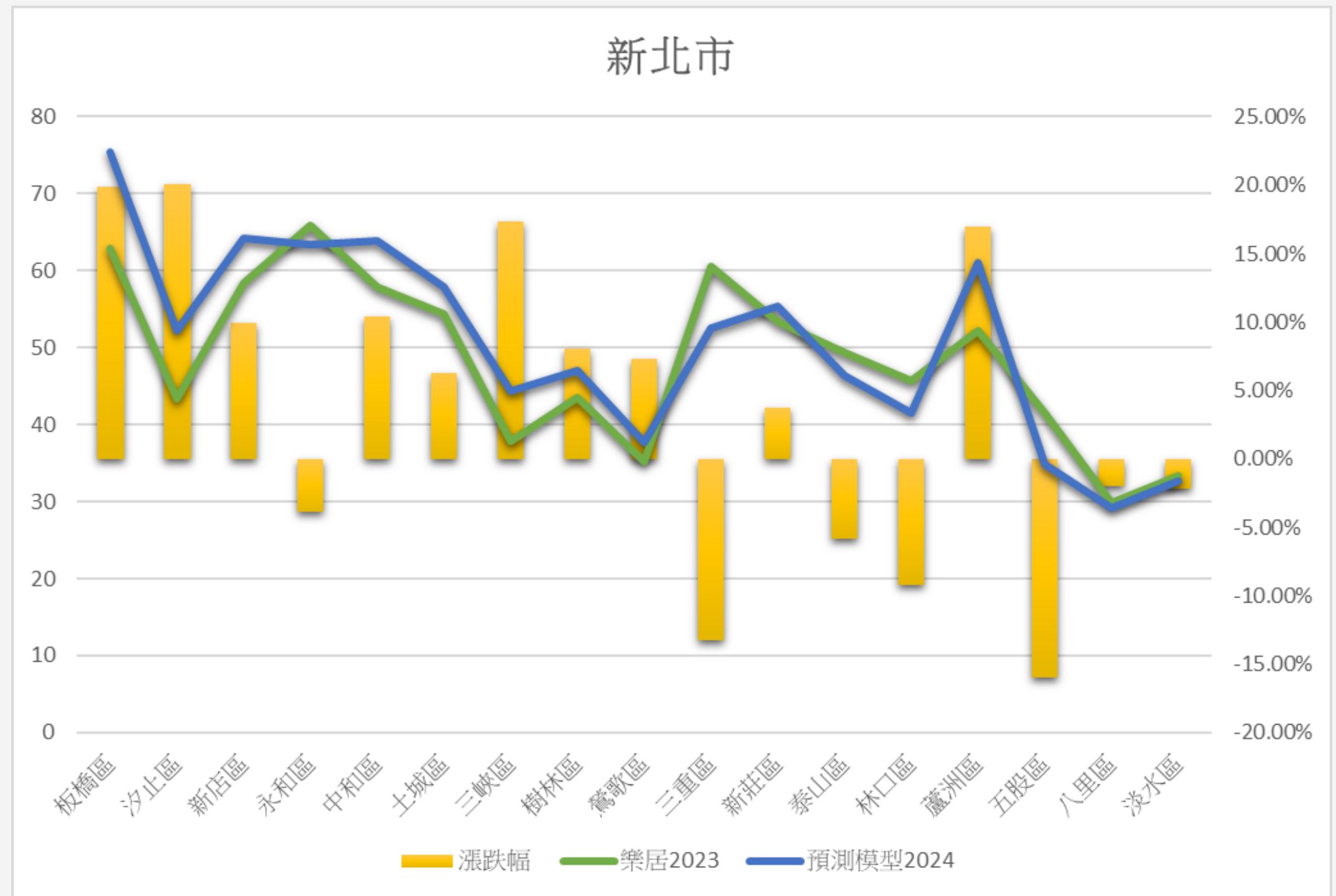


漲跌幅排行

| 區域 | 漲跌幅 |
|-----|---------|
| 內湖區 | +21.14% |
| 士林區 | +13.01% |
| 大同區 | +12.25% |
| 中正區 | -3.98% |
| 中山區 | -2.98% |
| 北投區 | -2.02% |

結論

成果分析



漲跌幅排行

| 區域 | 漲幅 |
|-----|---------|
| 汐止區 | +20.07% |
| 板橋區 | +19.86% |
| 三峽區 | +17.30% |

| 區域 | 漲跌幅 |
|-----|---------|
| 五股區 | -15.97% |
| 三重區 | -13.29% |
| 林口區 | -9.19% |

結論

目前現況：

- 目前模型僅適用雙北地區
- 地域資料目前僅納入交通、醫療、教育以及嫌惡設施等
- 前端網頁功能僅基礎功能

未來展望：

- 持續收集資料，優化模型
- 納入其他便利性設施資料，加以分析建模
- 改善前端其他功能性



Thank you!

