



# Python 一條龍專案

待售物件分析

Branden Darren Howard Jeff

Get Started

## 主要目標

- 整合開放資料和房仲網
- 幫助消費者在購屋時，找出議價空間

# 爬蟲動機

## Crawler Motive

購買房子對大多數人而言，可能是一生中，最大筆的支出交易。如何選擇價格合理且符合需求的房子，是一個重要課題。以往購買房屋的資訊來源，價格資訊單方面掌握在房仲業者手上。

# 時程分工

MAY

				題目 更改		
9			12			



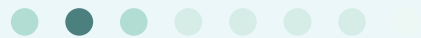
# 時程分工

JUNE

建立 資料庫 6	上傳 資料庫 7	8			10	11 Git hub
13 視覺化				17 API		
Tableau API 20	簡報 製作 21	簡報 製作 22		Big Day 24		

# 大綱

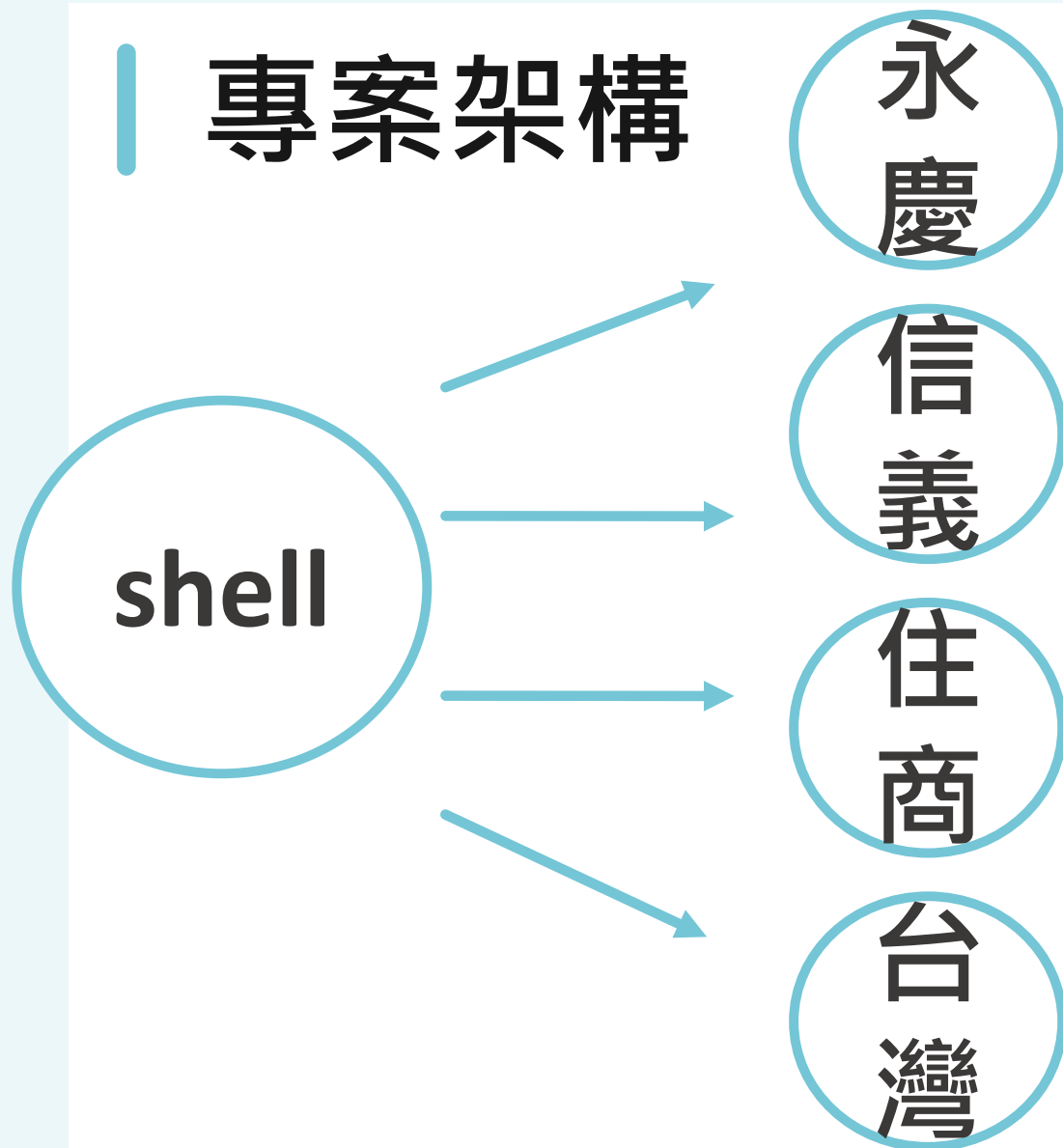




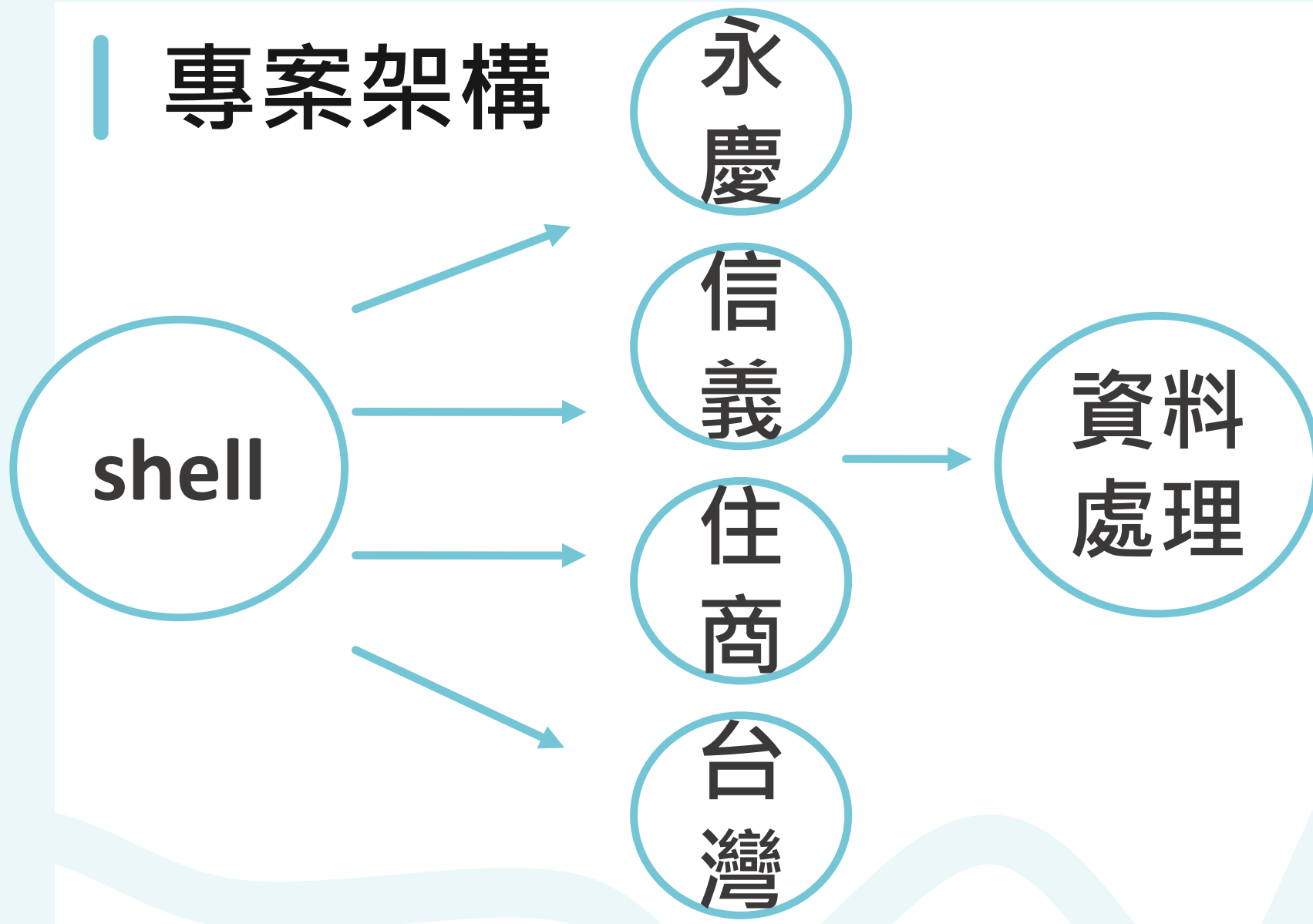
# | 專案架構



# 專案架構

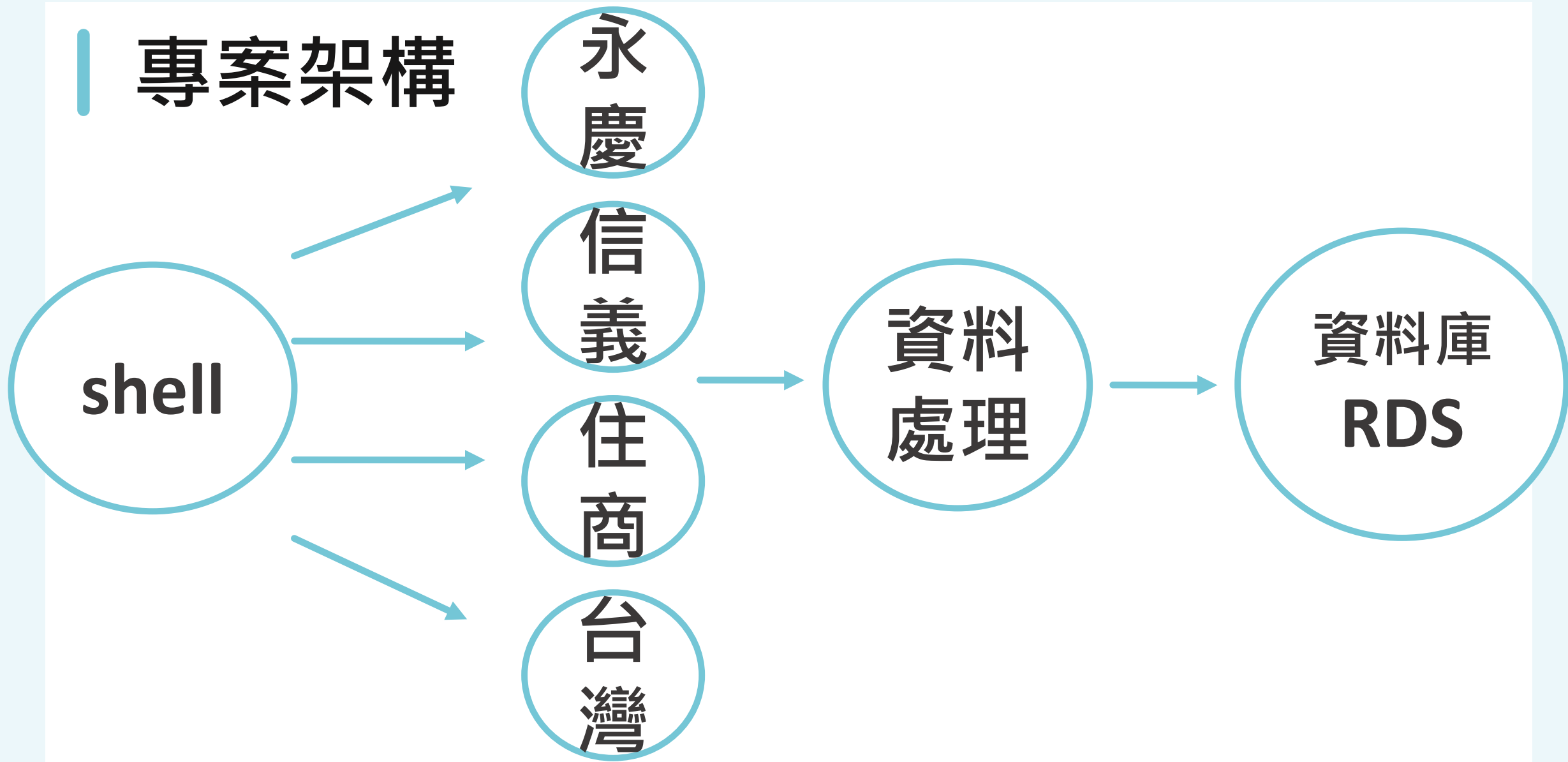


# 專案架構





# 專案架構



# 資料蒐集

物件名

地址

房屋型態

屋齡

坪數

售價

每坪單價

房

廳

衛

樓層

總樓層數

# 資料格式

Home Page - Select or create 05.23住商不動產信義 - Jupyter | 住商不動產信義.csv - Jupyter Te | +

localhost:8888/notebooks/05.23住商不動產信義.ipynb

jupyter 05.23住商不動產信義 Last Checkpoint: 2022年6月7日 (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [23]: `df = pd.DataFrame(alldata)`

In [24]: `df`

Out[24]:

	title	address	house_type	age	area	price	perprice	room	hall	bathroom	floor	total_floor
0	信義區公寓2樓美3房	台北市信義區虎林街	公寓	52.2	20.45	1800	88.02	3	2	1	2	4
1	近象山站公寓2樓	台北市信義區虎林街	公寓	52.2	20.45	1800	88.02	3	2	1	2	
2	永春雙捷運黃感二樓	台北市信義區虎林街	公寓	52.2	20.45	1800	88.02	3	2	1	2	4
3	松仁新廈2房附車位	台北市信義區松仁路	大廈	8.2	17.38	1838	105.75	2	2	2	4	9
4	永春捷運精品1樓	台北市信義區忠孝東路五段	公寓	22.7	10.17	1880	184.86	1	1	1	1	5
5	北醫看101頂家	台北市信義區吳興街	公寓	43.9	26.05	1880	72.17	5	4	2	5	5
6	博愛設籍電梯美屋	台北市信義區松山路	大廈	24.1	16.94	1888	111.45	2	2	1	2	9
7	永春2房花園一樓\$	台北市信義區松山路	大廈	24.1	16.94	1888	111.45	2	1	1	1	9
8	文盲世紀高樓辦公	台北市信義區基隆路二段	大廈	25.3	17.31	1888	109.07	1	室	1	13	21
9	永春捷運危老收租屋	台北市信義區松山路	公寓	45.7	19.53	1888	96.67	2	2	1	3	3

# API

## Jousesell Project <sup>v1</sup>

/swagger/

### Price

GET

/perprice/{location}/{data\_number}

請輸入地區關鍵字，查詢坪數、每坪單價(萬)，以及希望呈現的資料筆數

Parameters

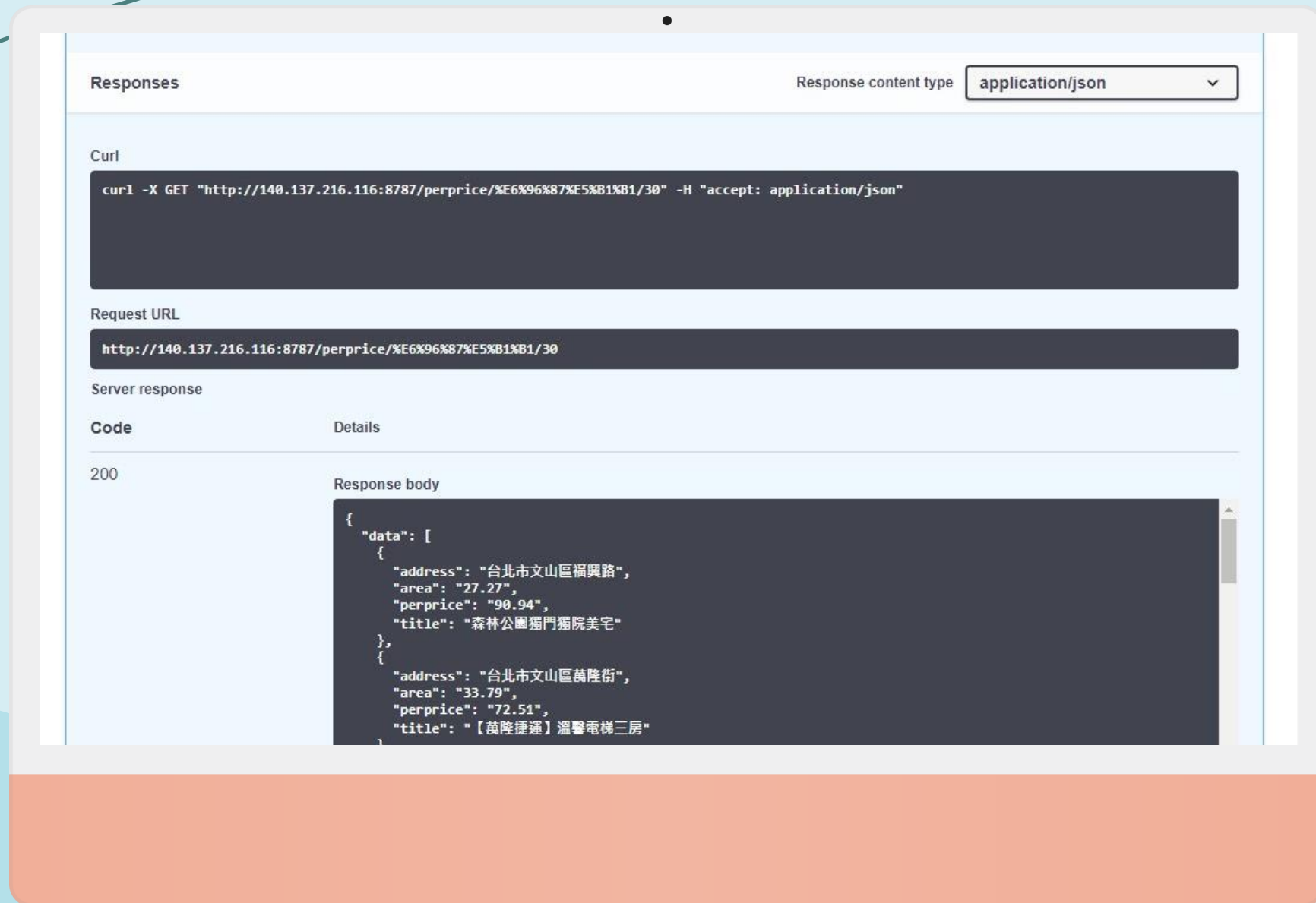
Cancel

Name	Description
<b>data_number</b> * required string (path)	<input type="text" value="30"/>
<b>location</b> * required string (path)	<input type="text" value="文山"/>

Execute

Clear

# API



分析

# Analysis



Xinyi +

Daan +

Wenshan +

# 視覺化分析



# 各房型

下降

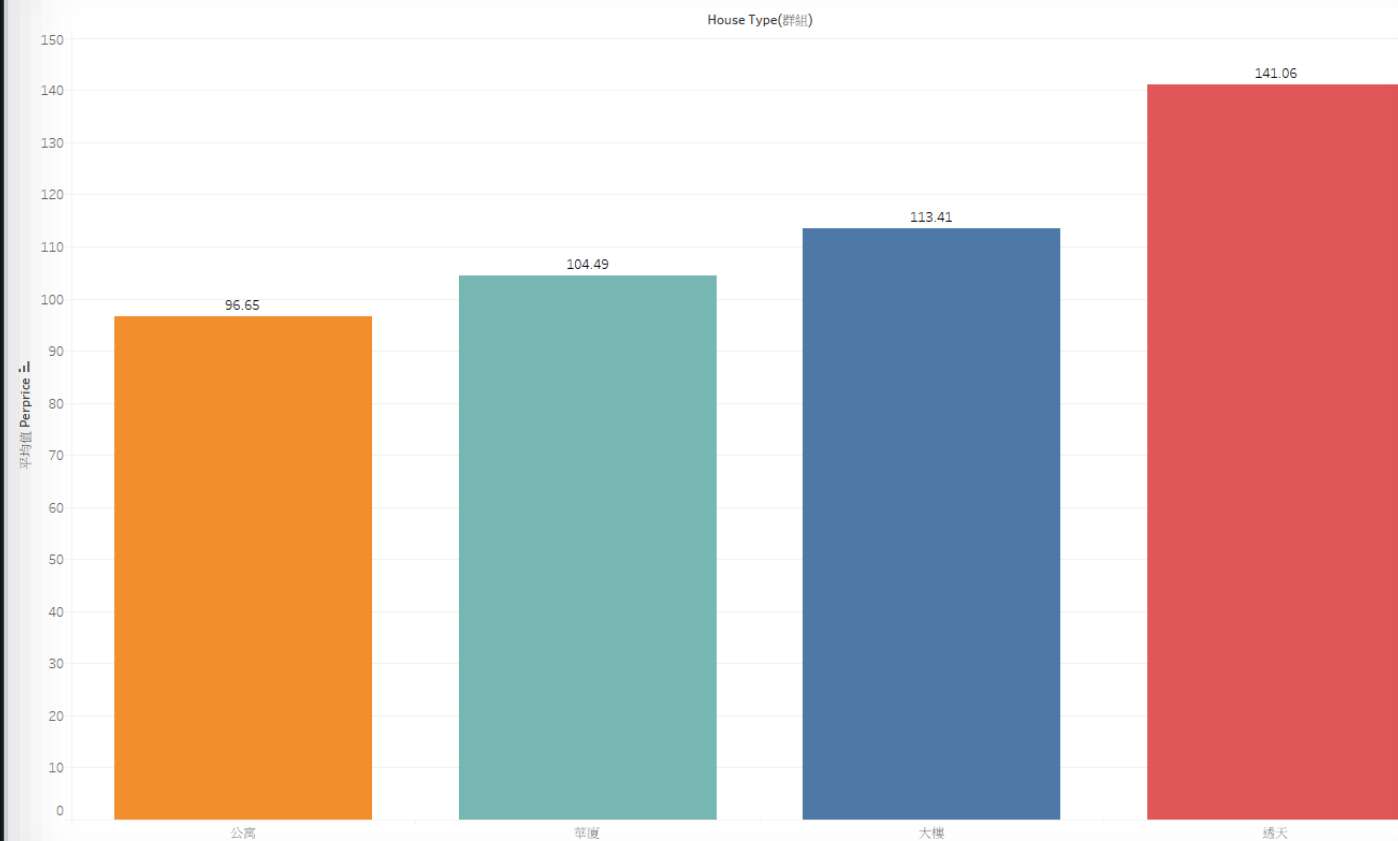
透天是  
大樓  
公寓的  
華廈  
1.46X  
公寓  
5.9%

7.4%

8.5%

1.46X  
5.9%

不同房型每坪定價(萬)



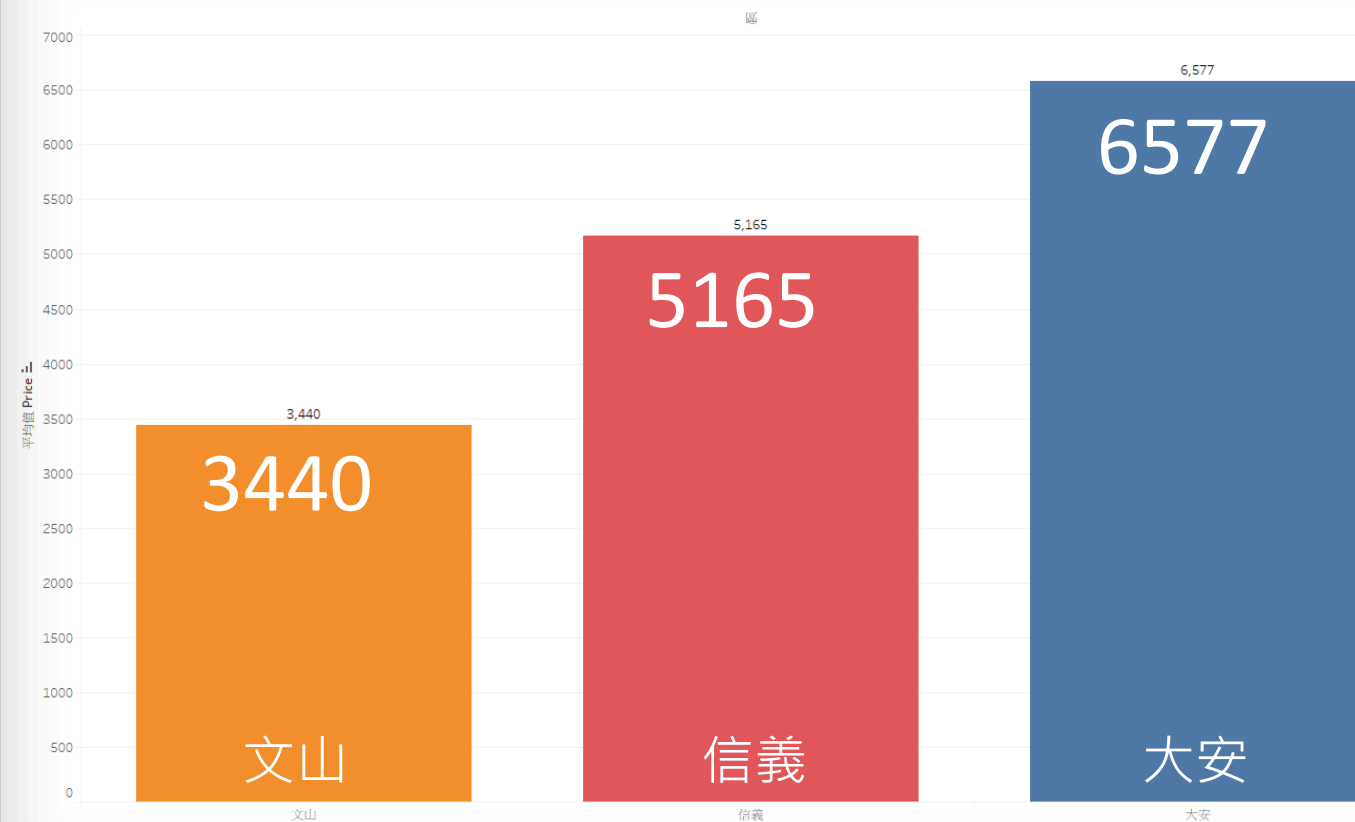
每個 House Type(群組) 的 Perprice 平均值。色彩顯示有關 House Type(群組) 的詳細資訊。標記按 Perprice 平均值 進行標記。檢視按 House Type(群組) 進行篩選。這會保留 公寓,大樓,華廈與透天。

# 視覺化分析



# 各區/總房價

各區平均總房價



每個區的 Price 平均值。色彩顯示有關區的詳細資訊。標記按 Price 平均值進行標記。

大安是  
文山的

1.91 X

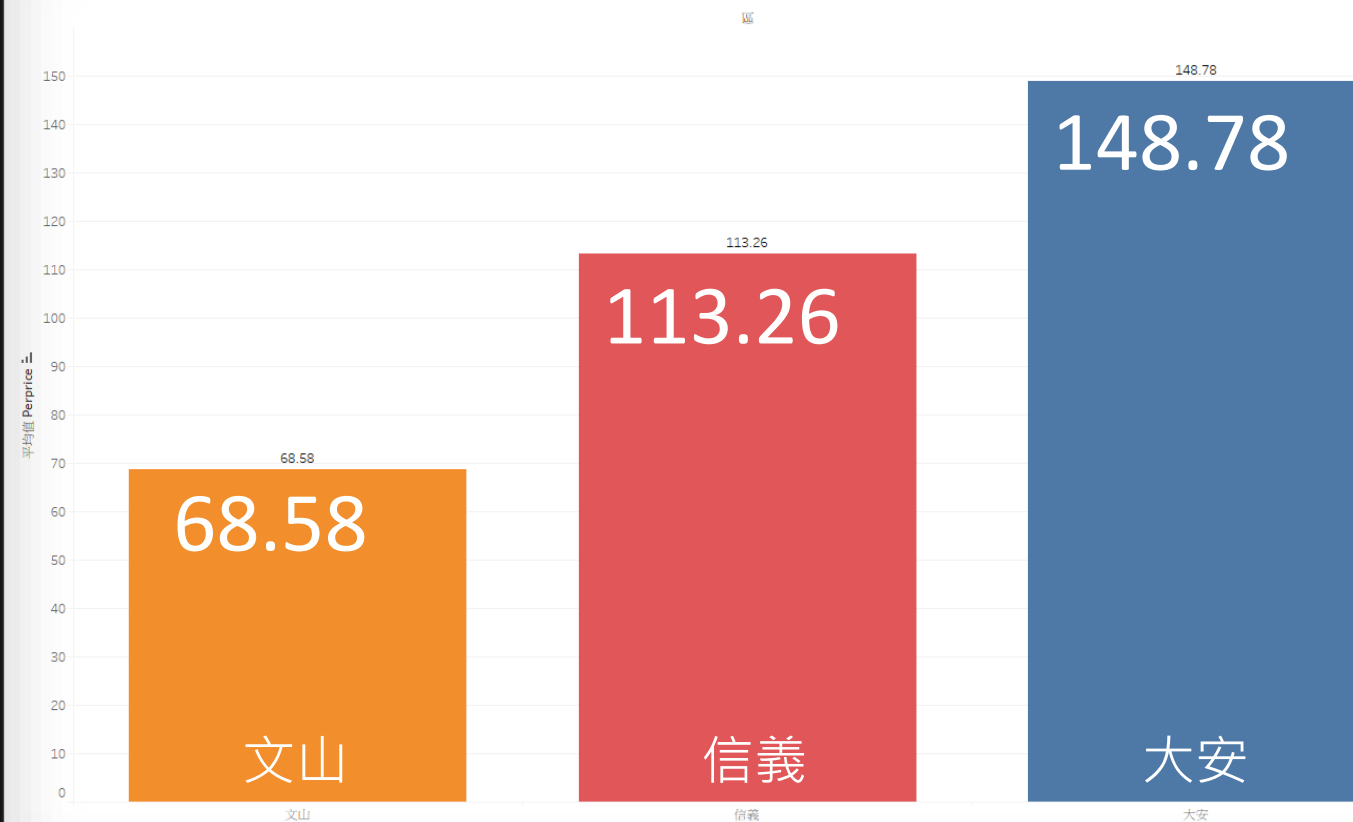


# 視覺化分析

×

# 各區/每坪

各區每坪定價(萬)



每個區的 Perprice 平均值。色彩顯示有關區的詳細資訊。標記按 Perprice 平均值進行標記。

大安是  
文山的

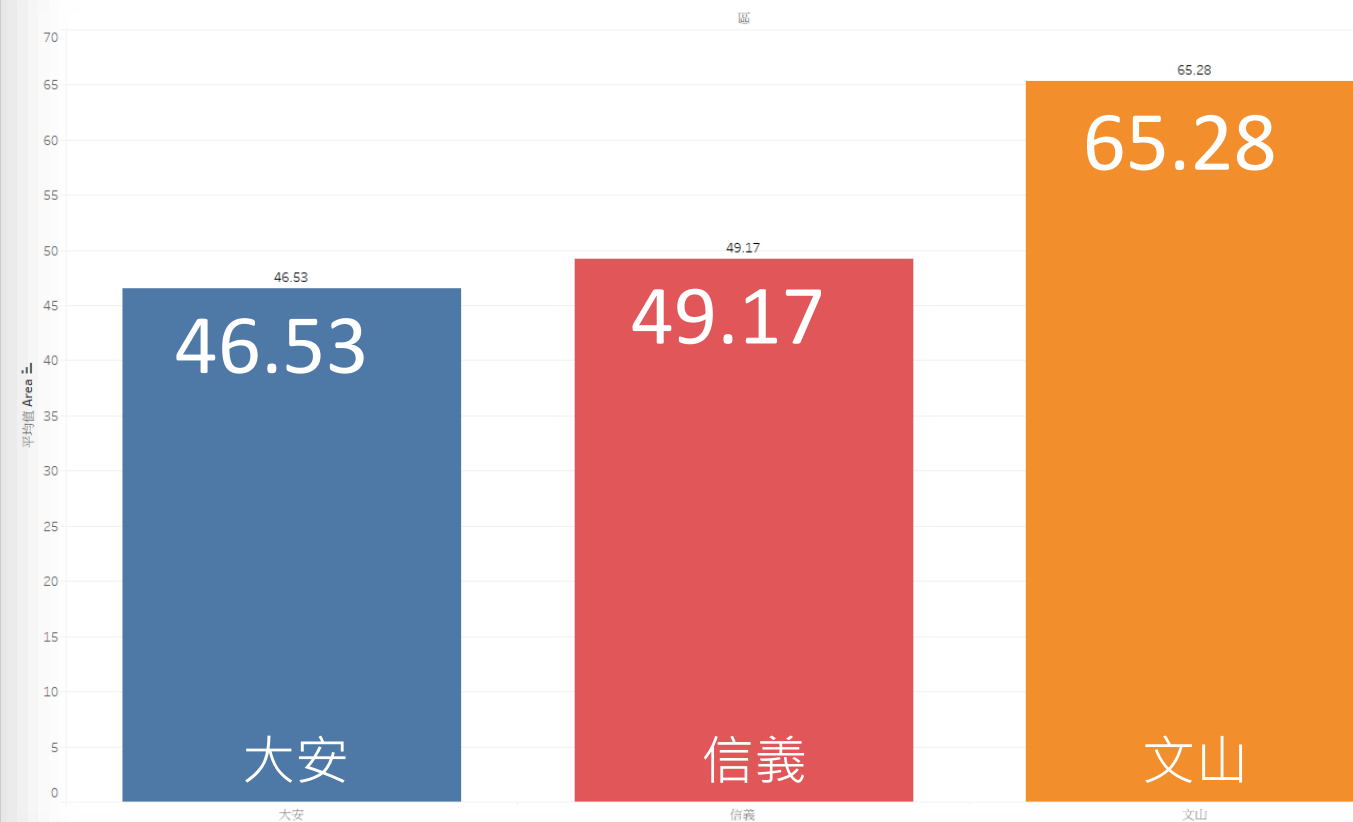
2.16 X

# 視覺化分析



# 各區/坪數

各區每房總坪數(平均)



每個區的 Area 平均值。色彩顯示有關區的詳細資訊。標記按 Area 平均值進行標記。

文山是  
大山的

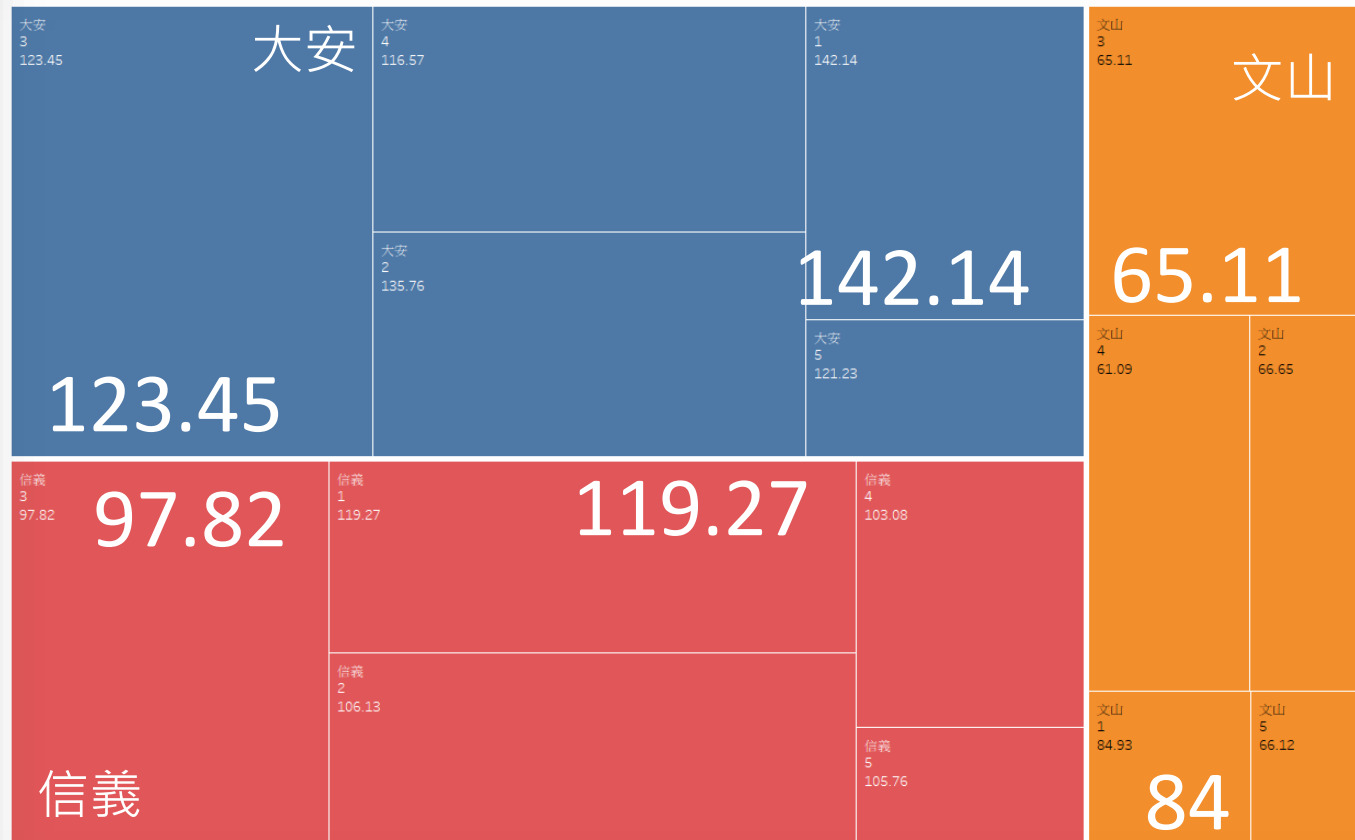
1.4 X

# 視覺化分析



# 各區/房數/單價

各區針對房數比較/每坪價格(萬)



區,Room與Perprice 平均值, 色彩顯示有關於 區 的詳細資訊, 尺寸顯示 Room 計數, 標記按 區,Room與Perprice 平均值 進行標記, 檢視按 Room 進行篩選, 這會保留 1,2,3,4與5。

1 房

最貴

3房

CP王

# 視覺化分析



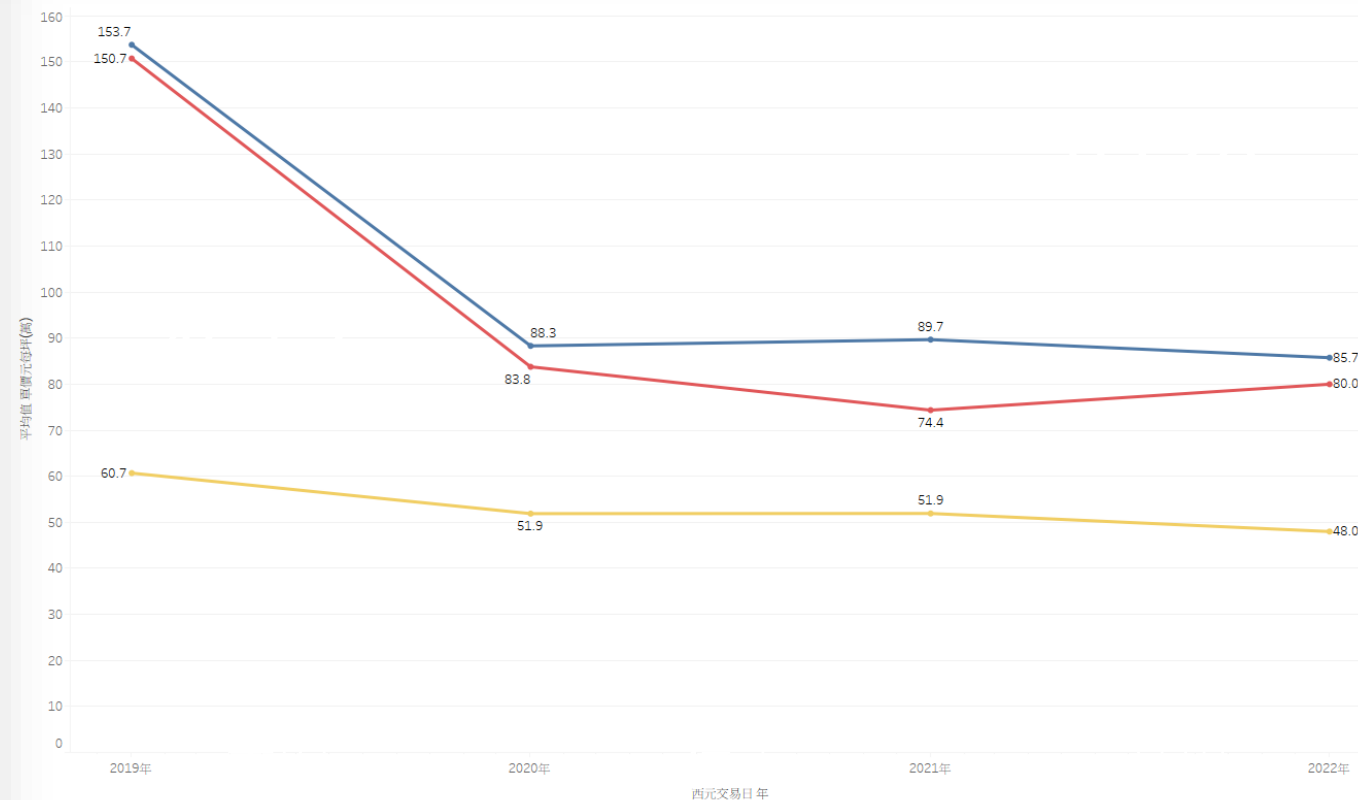
# 各區/單價

- 大安
- 信義
- 文山

信義/大安  
降幅

達 44 %

各區整年度每坪房價

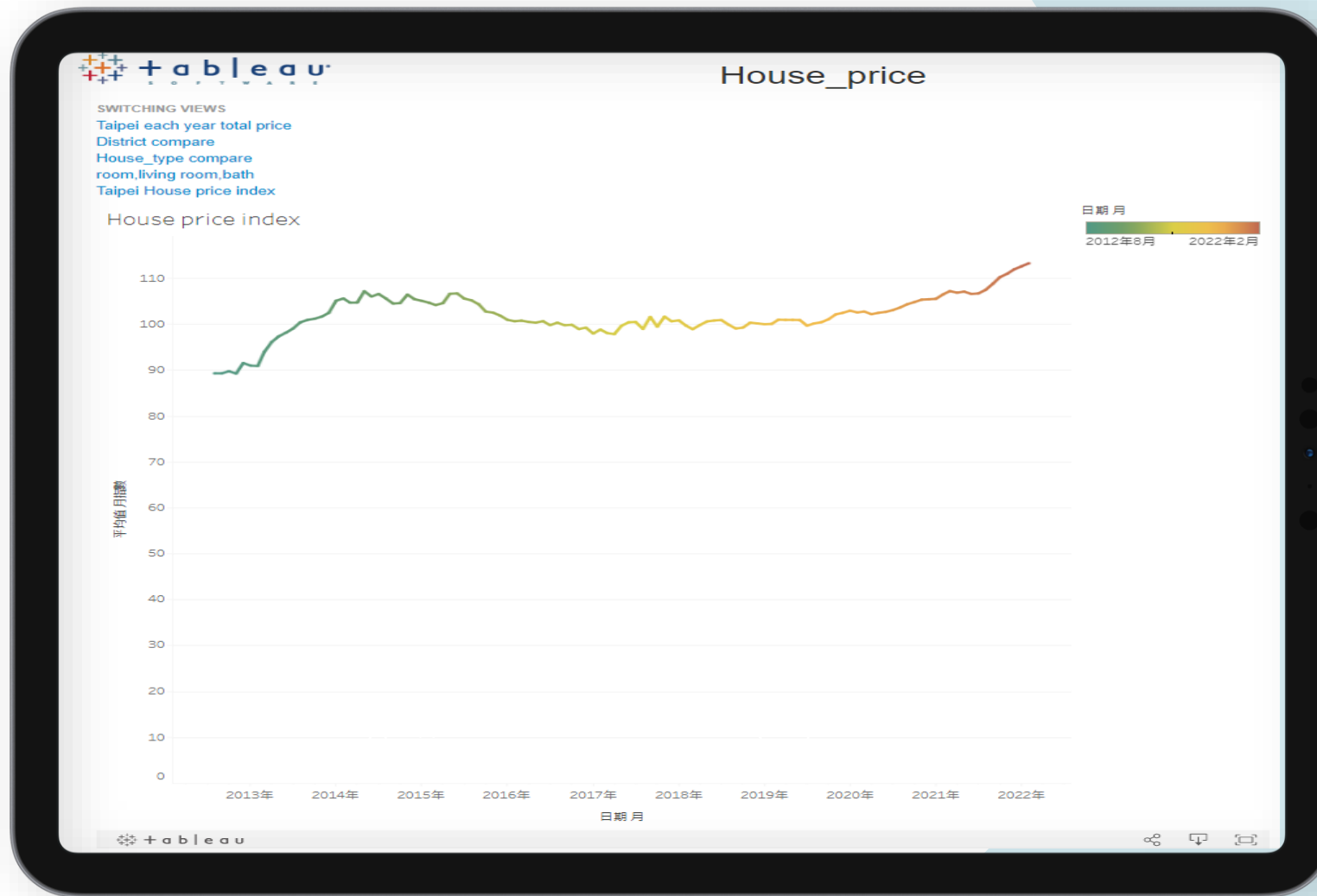


西元交易日 年 的 單價元每坪(萬) 平均值的趨勢。色彩顯示有關 鄉鎮市區 的詳細資訊。標記按 單價元每坪(萬) 平均值 進行標記。資料按 西元交易日 年 進行篩選，這會保留 2019,2020,2021與2022。檢視按 鄉鎮市區 進行篩選，這會保留 信義區,大安區與文山區。

# 視覺化分析



# 台北住房指數



困

# Dilemma



境

Difficult +  
Solution ?

How ?

# 06.24

## 困境

## 回顧

# 57 DAY

# Branden

## 爬蟲

- 樓層欄位的定義以及房/廳/衛的資料的格式切分較為複雜，並且包含許多不需要的雜項字元。
- 因為缺少每坪價位的資訊，因此需要做Dataframe上的計算，過程中產生inf的無窮值。

## API

- 撰寫API時在JSON格式上有錯誤。

## 解決方法

- Dataframe的調整，包含計算上，除了做到numeric的轉型之外，在資料清理時也要檢查錯誤值做取代。
- 撰寫JSON時可以習慣做JSON Formatter的檢查

# 解决方案

檔案(F) 編輯(E) 選取項目(S) 檢視(V) 移至(G) 執行(R) 終端機(T) 說明(H)

app1.py - PJ - Visual Studio Code

檔案總管

▼ PJ

- 永慶房屋.py 8
- 住商不動產.py
- app1.py 6
- cra\_api\_route.py 1
- cra\_api.py 5
- drive-download-20220623T014336Z...
- util.py

app1.py > ...

```
1 from flask import Flask
2 from flask_restful import Api
3 from cra_api import *
4 from apispec import APISpec
5 from apispec.ext.marshmallow import MarshmallowPlugin
6 from flask_apispec.extension import FlaskApiSpec
7 from flask_jwt_extended import JWTManager
8
9
10 app = Flask(__name__)
11 api = Api(app)
12
13 app.config["DEBUG"] = True # Able to reload flask without exit the process
14 app.config["JWT_SECRET_KEY"] = "secret_key" #JWT token setting
15
16 # Swagger setting
17 app.config.update({
18     'APISPEC_SPEC': APISpec(
19         title='Jousesell Project',
20         version='v1',
21         plugins=[MarshmallowPlugin()],
22         openapi_version='2.0.0'
23     ),
24     'APISPEC_SWAGGER_URL': '/swagger/', # URI to access API Doc JSON
25     'APISPEC_SWAGGER_UI_URL': '/swagger-ui/' # URI to access UI of API Doc
26 })
27 docs = FlaskApiSpec(app)
28
29 api.add_resource(Search, '/perprice/<string:location>/<string:data_number>')
30 docs.register(Search)
31
32
33 if __name__ == '__main__':
34     JWTManager().init_app(app)
35     app.run(host='0.0.0.0', port=8787, debug=True)
```



# 06.24

## 困境

## 回顧

# 57 DAY

## Darren

### 爬蟲

- 價格的區塊有多種顯示方式，無法正確定位到總價
- 房廳衛的格式有多種可能性，資料切割處理繁雜

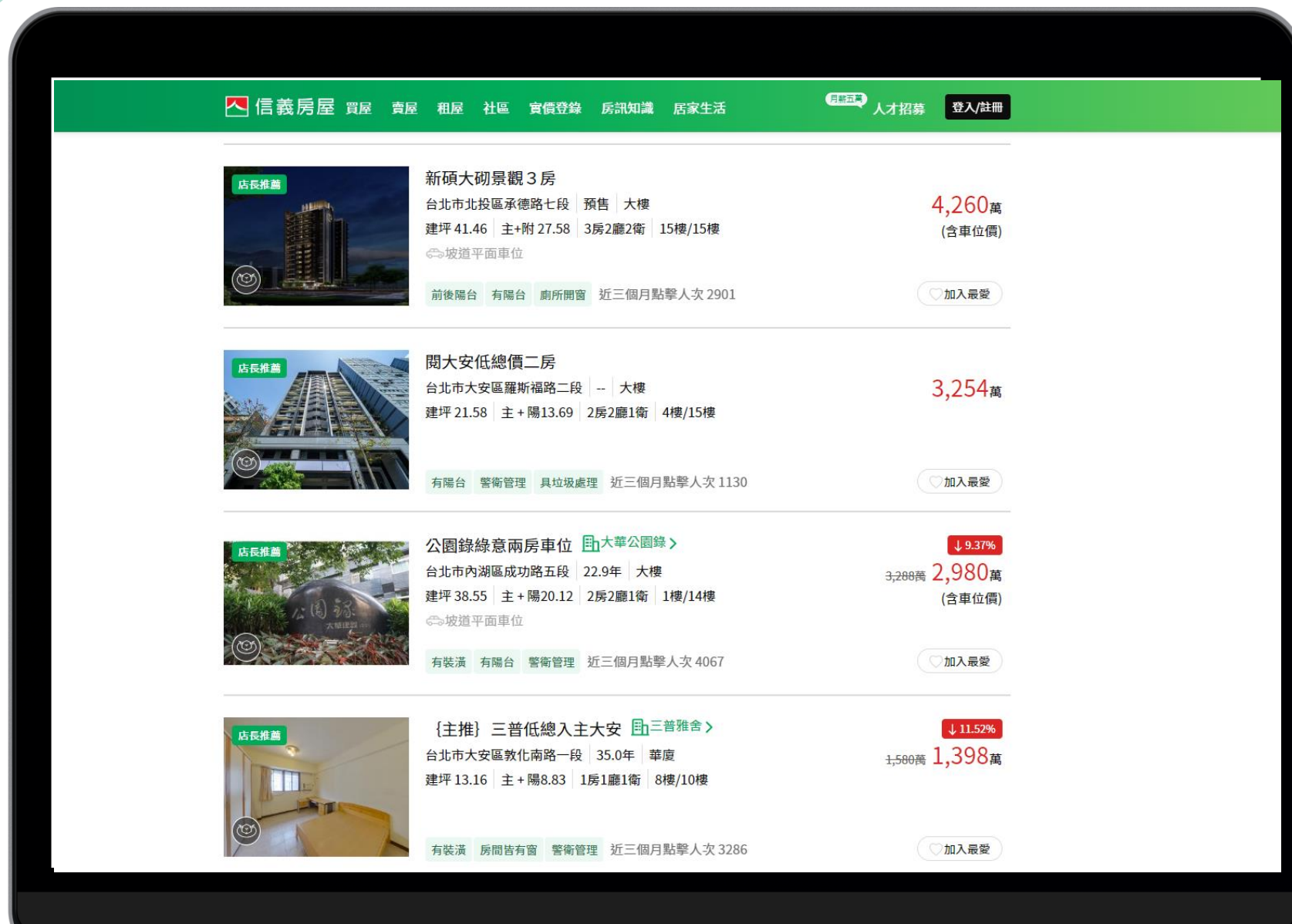
### 資料庫

- 透過AWS的RDS功能架設，主要的問題只有一開始忘了開RDS的3306 port的權限，導致無法連線登入

### 解決方法

- 發現總價的字體色彩與其他字體皆不同，因此透過find\_all定位屬性的方法去抓取
- 列舉8項可能性 使用if elif else的判斷式去篩選

# 問題



# 刀解 决

檔案(F) 編輯(E) 選取項目(S) 檢視(V) 移至(G) 執行(R) 終端機(T) 說明(H)

信義房屋.py - PJ - Visual Studio Code

檔案總管

▼ PJ

- 永豐房屋.py
- 住商不動產.py
- 信義房屋.py 3
- app1.py
- cra\_api\_route.py
- cra\_api.py
- drive-download-20220623T014336Z...
- util.py

> 大綱  
> 時間表

信義房屋.py 3

信義房屋.py > getdata

```
17 for i in range(20):
18     table = soup.select('.row')[i]
19     title = table.find_all('div', 'LongInfoCard_Type_Name')[1].text
20     address = table.find_all('div', 'LongInfoCard_Type_Address')[0].find_all('span')[0].text
21     house_type = table.find_all('div', 'LongInfoCard_Type_Address')[0].find_all('span')[2].text
22     age = table.find_all('div', 'LongInfoCard_Type_Address')[0].find_all('span')[1].text.replace('--', '0').replace('年', '')
23     structure = table.find_all('div', 'LongInfoCard_Type_HouseInfo')[0].find_all('span')
24     area = structure[0].text.replace('建坪', '').replace('地坪', '')
25     price = soup.find_all('span', style='font-size:1.75em;font-weight:500;color:#dd2525')[1].text.replace(',', '')
26
27     s = structure[2].text
28     if ('房' in s) & ('廳' in s) & ('衛' in s):
29         room=s.split('房')[0]
30         hall=s.split('房')[1].split('廳')[0]
31         bathroom=s.split('房')[1].split('廳')[1].split('衛')[0]
32     elif ('房' in s) & ('廳' in s):
33         room=s.split('房')[0]
34         hall=s.split('房')[1].split('廳')[0]
35         bathroom='0'
36     elif ('廳' in s) & ('衛' in s):
37         room='0'
38         hall=s.split('廳')[0]
39         bathroom=s.split('廳')[1].split('衛')[0]
40     elif ('房' in s) & ('衛' in s):
41         room=s.split('房')[0]
42         hall='0'
43         bathroom=s.split('房')[1].split('衛')[0]
44     elif '房' in s:
45         room=s.split('房')[0]
46         hall='0'
47         bathroom='0'
48     elif '廳' in s:
49         room='0'
50         hall=s.split('廳')[0]
51         bathroom='0'
52     elif '衛' in s:
53         room='0'
54         hall='0'
55         bathroom=s.split('衛')[0]
56     else:
57         print(s)
58
59     if "-" in structure[3].text.split('/')[0]:
60         floor=structure[3].text.split('/')[0].split('-')[1].replace('樓', '')
61     else:
62         floor=structure[3].text.split('/')[0].replace('樓', '')
63
64     total_floor=structure[3].text.split('/')[1].replace('樓', '')
65     perprice = round(int(price) / float(area), 2)
```

# 06.24

## 困境

## 回顧

# 57 DAY

# Howard

## 爬蟲

- 爬蟲時遇到網頁有透過ajax技術，因此沒辦法用 requests html 來爬到需要的資料。

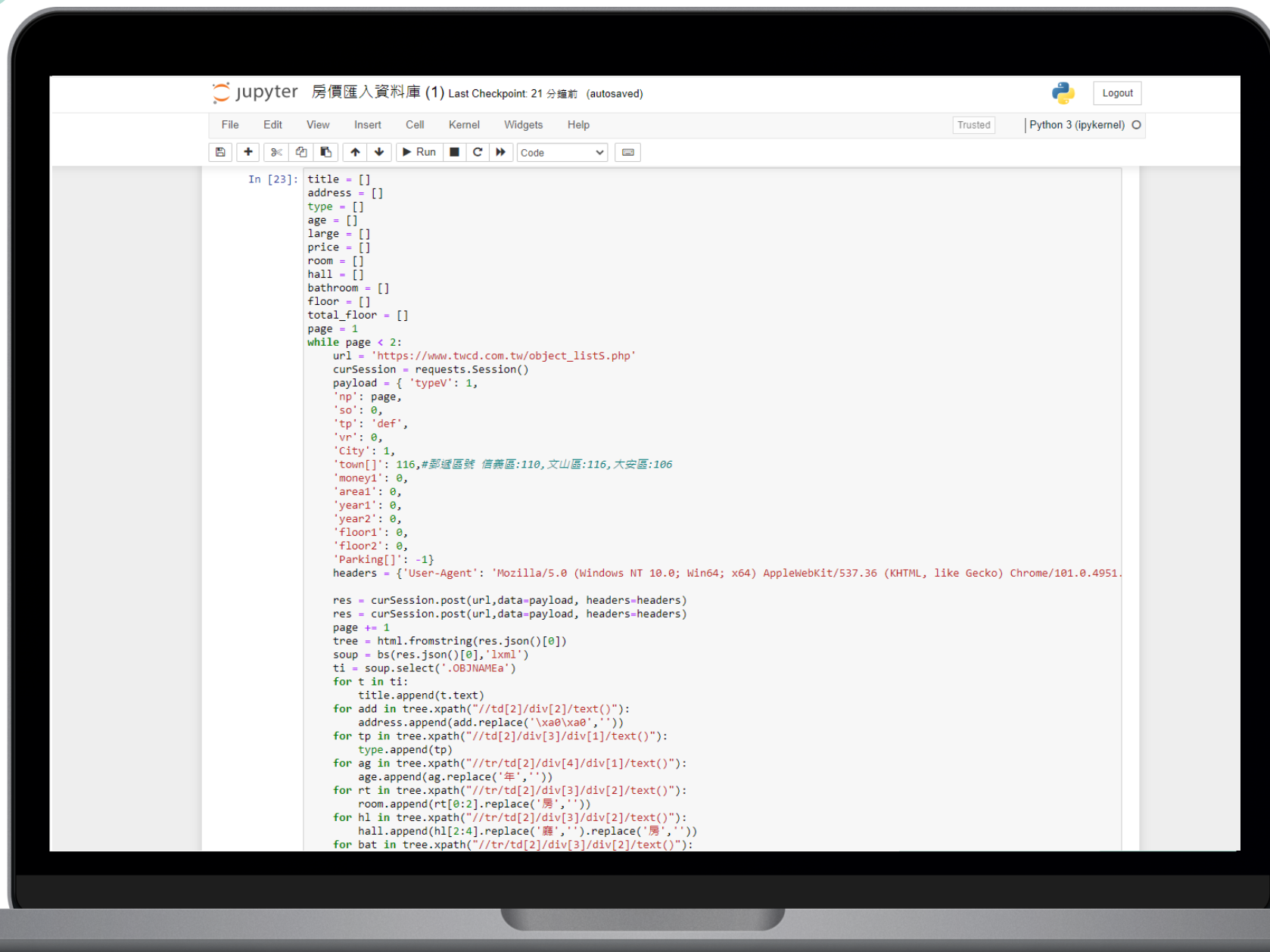
## 資料整理

- 由於爬下來的資料很雜，有出現空值或是格式不符等錯誤，影響視覺化的呈現。

## 解決方法

- 在Network 可以找到資料需要的Url，透過post或 get 請求，來取得資料。
- 空值的部分填入整欄的平均值，民國的日期改以西元方式呈現。

# 刀解 牛決



# 06.24

困境

回顧

## 57 DAY

# Jeff

## 爬蟲

- 無法使用while迴圈把每一頁資料Append到同列表
- 無法使用Request.get的方式來取得資料

## 解決方法

- 使用雙For迴圈把資料取出來，就可以Append到列表。
- 從Data-Service中的Payload來去GET

# 問題

```
[3]: response = []

i = 2
while i < 3:
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/101.0.4951.54 Safari/537.36',
        'Referer': 'https://www.hbhousing.com.tw/BuyHouse/%E5%8F%B0%E5%8C%97%E5%B8%82/116',
        'params': {'job': 'search', 'path': 'house'},
    }

    payload = {
        'q': '^1^3^116^^1_9^2_3_4_5^^^^^^^^^^^^^0^^{}^0".format(i),
        'rlg': '0'
    }

    i += 1
    response2 = requests.post('https://www.hbhousing.com.tw/ajax/dataService.aspx', params=params, headers=headers, data=payload)
    soup=bs(response2.text, 'lxml')
    x=json.loads(response2.text)
    response.append(x)

data2 = response.json()

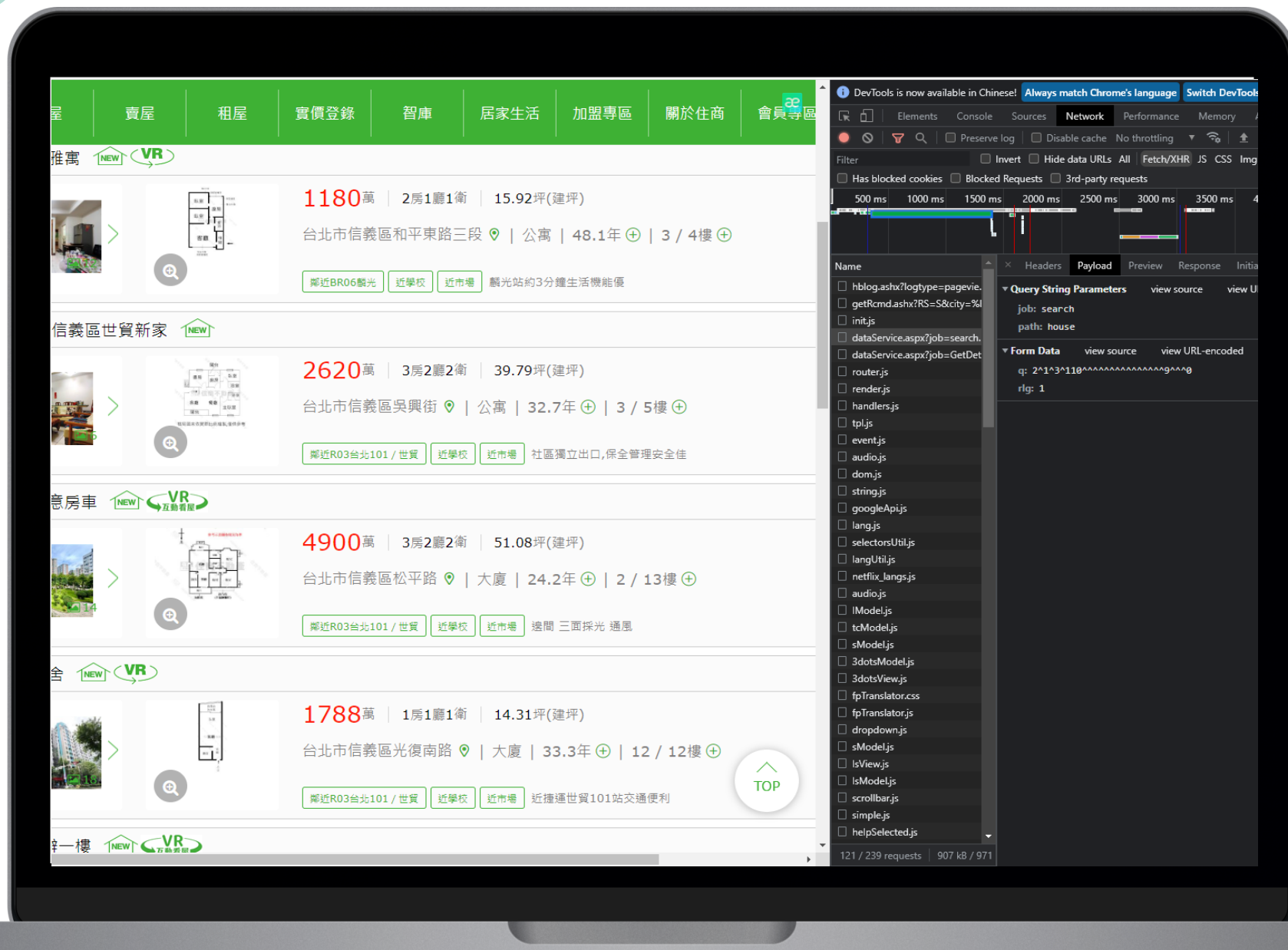
# 'n': '景美捷運美套房' 'x': '台北市文山區景興路' 't': '大廈' 'k': 28.2 'a': 8.45 'np': '630'

title = []
address = []
house_type = []
age = []
area = []
price = []

for t in data2['data']:
    title.append(t["n"])
    address.append(t["x"])
    house_type.append(t["t"])
    age.append(t["k"])
    area.append(t["a"])
    price.append(t["np"])

ave = [int(a) / float(b) for a, b in zip(price, area)]
ave = [round(i,2) for i in ave]
```

# 問題





# 解 決

```
In [29]: response = []

i = 2
while i < 22:
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/101.0.4951.54 Safari/537.36',
        'Referer': 'https://www.hbhousing.com.tw/BuyHouse/%E5%8F%B0%E5%8C%97%E5%B8%82/116',}
    params = {'job': 'search', 'path': 'house',}

    payload = {
        'q': '^1^3^116^^1_9^2_3_4_5^^^^^^^^^0^^{}^0".format(i),
        'rlg': '0'
    }

    i += 1
    response2 = requests.post('https://www.hbhousing.com.tw/ajax/dataService.aspx', params=params, headers=headers, data=payload)
    x=response2.json()
    response.append(x)

title = []
address = []
house_type = []
age = []
area = []
price = []
room = []
floor = []
total_floor = []

dd = {i:v for i,v in enumerate(response)}
for z in dd.values():
    for t in z['data']:
        title.append(t["n"])
        address.append(t["x"])
        house_type.append(t["t"])
        age.append(t["k"])
        area.append(t["a"])
        price.append(t["np"])
        room.append(t["p"])
        floor.append(t["w"])
        total_floor.append(t["z"])

ave = [int(a) / float(b) for a, b in zip(price, area)]
ave = [round(i,2) for i in ave]
```

未來

# Future

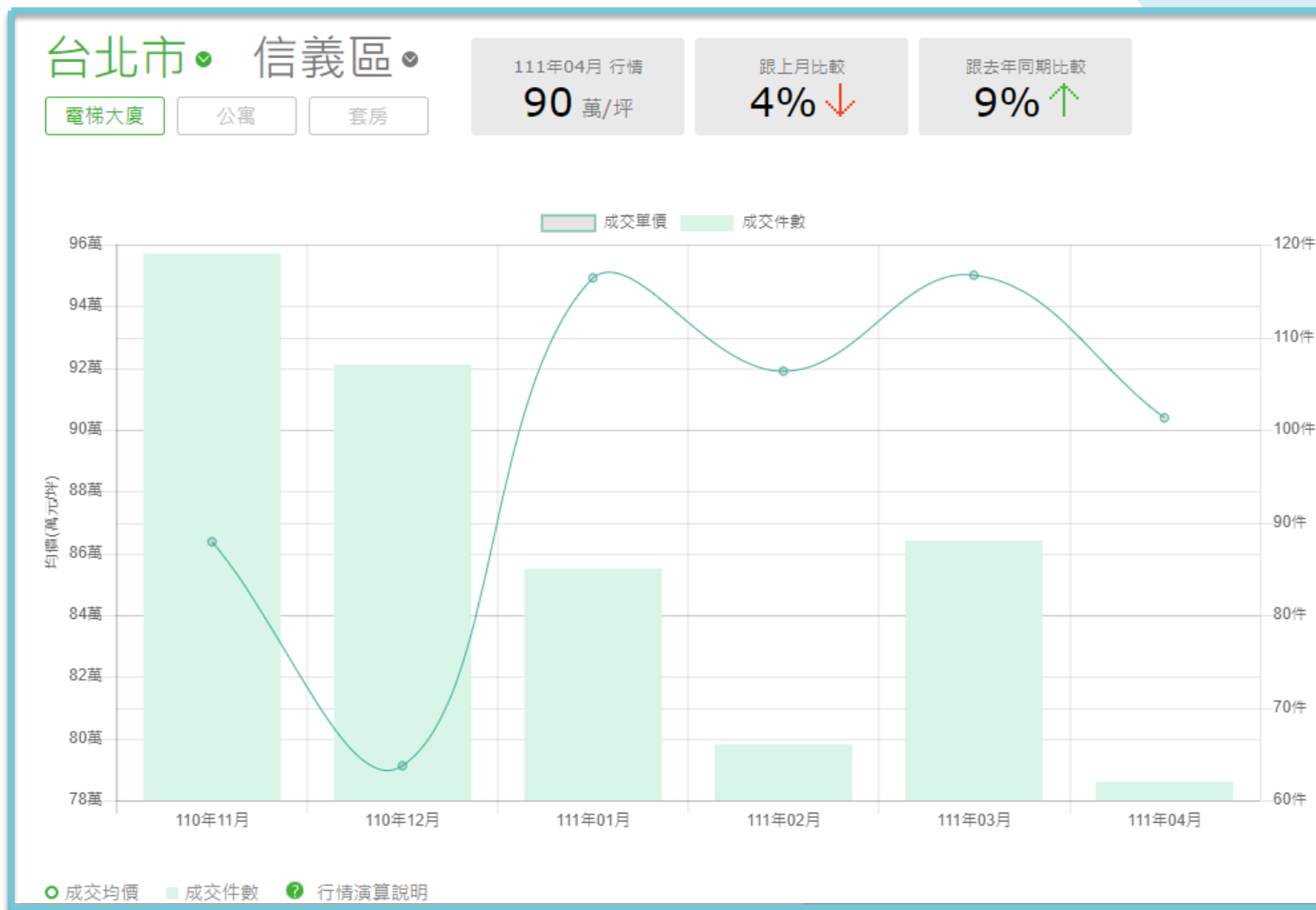


MORE ?

Enrich ?

Vision ?

# 未來工作



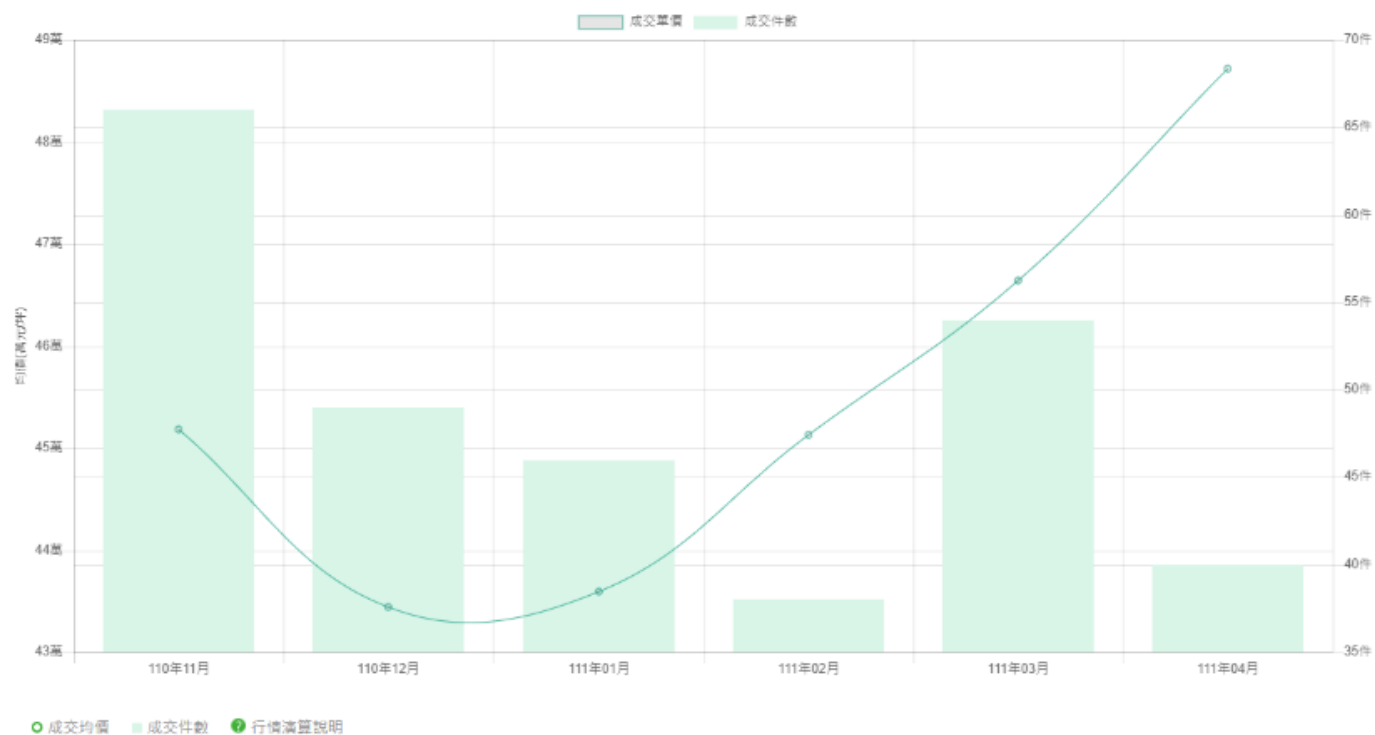
# 未來工作



# 未來工作

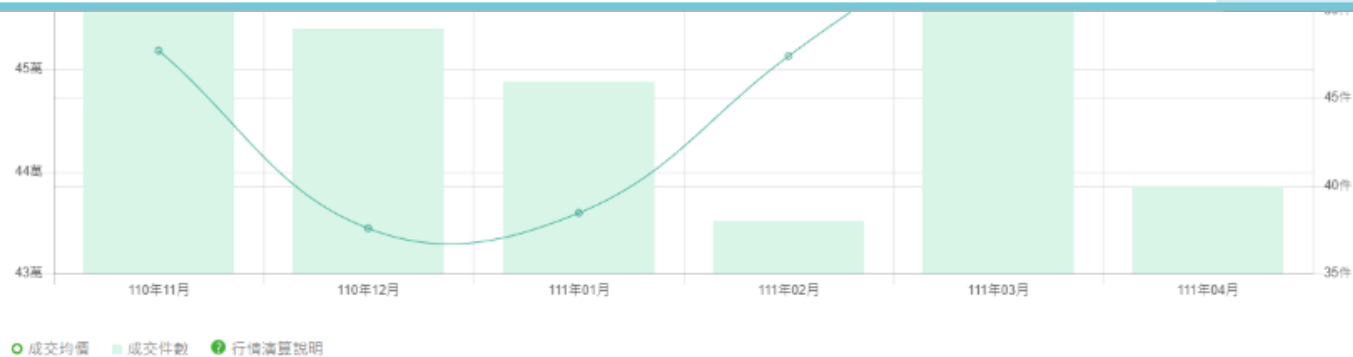
台北市 文山區 實價登錄走勢

電梯大廈 公寓 套房 其他



台北市 前10大成交排行

# 未來工作



## 台北市 前10大成交排行

台北市4月 成交單價排行

1. 大安區	95.2萬
2. 信義區	82.6萬
3. 中正區	79.4萬
4. 中山區	77.4萬
5. 松山區	77.4萬
6. 士林區	68.3萬
7. 大同區	65.3萬
8. 南港區	63.9萬
9. 內湖區	61.5萬
10. 萬華區	56.2萬

台北市4月 成交漲跌排行

1. 士林區	9.3 %
2. 萬華區	6.1 %
3. 文山區	5.4 %
4. 中山區	4.1 %
5. 北投區	3.4 %
6. 內湖區	0 %
7. 大安區	-0.5 %
8. 松山區	-2.1 %
9. 大同區	-2.8 %
10. 信義區	-3.5 %

台北市4月 成交量排行

1. 中山區	160件
2. 內湖區	113件
3. 士林區	102件
4. 文山區	94件
5. 信義區	90件
6. 萬華區	77件
7. 大安區	77件
8. 北投區	76件
9. 松山區	67件
10. 中正區	48件

Darren X aws

The Amazon logo arrow, a thick orange curved line, is positioned below the text "aws".

# 未來工作

## Amazon API Gateway

建立、維護和保護任何規模的 API

建立 AWS 帳戶

Amazon API Gateway 是一種全受管的服務，可讓開發人員輕鬆地建立、發佈、維護、監控和保護任何規模的 API。API 可作為應用程式的「前門」，以便從後端服務存取資料、商業邏輯或功能。使用 API Gateway 時，您可以建立 RESTful API 和 WebSocket API，以啟用即時雙向通訊應用程式。API Gateway 支援容器化、無伺服器工作負載和 Web 應用程式。

API Gateway 負責處理有關接受和處理多達數十萬個並行 API 呼叫的所有工作，包括流量管理、CORS 支援、授權和存取控制、調節、監控和 API 版本管理。API Gateway 沒有最低費用或啟動成本。您要為收到的 API 呼叫和資料傳輸量支付費用，而使用 API Gateway 分級定價模型，可在 API 用量擴展時減少成本。

**100 萬個 API 呼叫接收免費**

(使用 AWS 免費方案 12 個月的每個月)

[免費試用 »](#)



# 未來工作

## API Gateway 運作方式



# 未來工作

## 優勢

### 有效率的 API 開發

使用 API Gateway 同時執行相同 API 的多個版本，可讓您快速重複執行、測試和發行新的版本。您要為 API 呼叫和資料傳輸付費；既沒有最低費用，也沒有前期承諾。

### 輕鬆監控

從 API Gateway 儀表板監控效能指標，以及 API 呼叫、資料延遲和錯誤率的相關資訊，這可讓您使用 [Amazon CloudWatch](#)，以視覺化方式監控對您服務的呼叫。

### 任何規模的效能

使用 Amazon CloudFront 來妥善利用節點的全球網路，為最終使用者提供 API 請求和回應的最低可能延遲。調節流量和授權 API 呼叫，以確保後端操作可以承受流量高峰，而且不需要呼叫後端系統。

### 靈活的安全性控制

利用 AWS Identity and Access Management (IAM) 和 Amazon Cognito，授與對 API 的存取權。如果您使用 OAuth 字符，則 API Gateway 會提供原生 OIDC 和 OAuth2 支援。為了支援自訂的授權要求，您可以從 [AWS Lambda](#) 執行 Lambda 授權方。

### 大規模節省成本

API Gateway 提供 API 請求的分級定價模式。在最高級之中，每百萬請求的 API 請求價格低至 0.90 USD，這是因為在各個 AWS 帳戶中，每個區域的 API 用量都有所增加，所以您的成本會下降。

### RESTful API 選項

使用 HTTP API 或 REST API 建立 RESTful API。對大多數使用案例而言，HTTP API 是建置 HTTP API 的最佳方法，最高可比 REST API 便宜 71%。如果您的使用案例需要單一解決方案中有 API 代理功能和管理功能，您就可以使用 REST API。

The End