

# BeautifulSoup

HTML網頁爬蟲

孫善堂

台灣跨領域人才培訓發展協會  
理事長 2023.07.27



# 孫善堂 理事長

## 現職

台灣跨領域人才培訓發展協會 理事長  
勞動部勞動力發展署 關鍵就業力 KC講師

## 學歷

國立台灣大學 森林系 學士  
元智大學 工業工程與管理 碩士班  
輔仁大學 經濟學系 研究所

## 經歷

台灣人工智能產業協會 AI課程講師  
社團法人中華勞動力職能發展協會  
大數據 / 物聯網 / AI課程講師  
華梵大學、佛光大學、致理科技大學 講師  
力新創意有限公司 行銷副理  
數十家企業內訓講師及專案輔導顧問





# 1.HTML語法簡介

(1)標籤組成

(2)標籤屬性

(3)網頁架構

## 2.BeautifulSoup模組應用

## 3.104職缺爬蟲專案

# 最基礎的標籤

標籤開頭

`<a>`

一行文字

標籤結尾

`</a>`

標籤內容(字串)

# 最基礎的標籤

`<a>` 一行文字 `</a>`

標籤名稱：a



The diagram illustrates the structure of a basic HTML tag. It shows the opening tag `<a>` and the closing tag `</a>` separated by the text "一行文字" (a line of text). The 'a' in both tags is highlighted with a red square. Two yellow arrows point from these red squares to the text "標籤名稱：a" (Tag name: a) below.



# 有屬性的標籤

標籤屬性(Attributes)置於「標籤開頭」之內，以”名/值對”表現

`<h1 class="abc" href="https://www.google.com.tw/">Google</h1>`

第一對屬性

第二對屬性

# 有屬性的標籤

屬性名稱為HTML預設，屬性值則以字串形式呈現



# 清楚了解標籤結構了嗎？

標籤名稱？      標籤開頭？      標籤結尾？

```
<h1 class="abc" href="https://www.google.com.tw/">Google</h1>
```

屬性名稱？      屬性值？      標籤內容？





# 多層次網頁架構

標籤內容可以是"字串"，或者"其他標籤"

```
<body>
```

```
  <h1 class="abc" href="123">abc</h1>
```

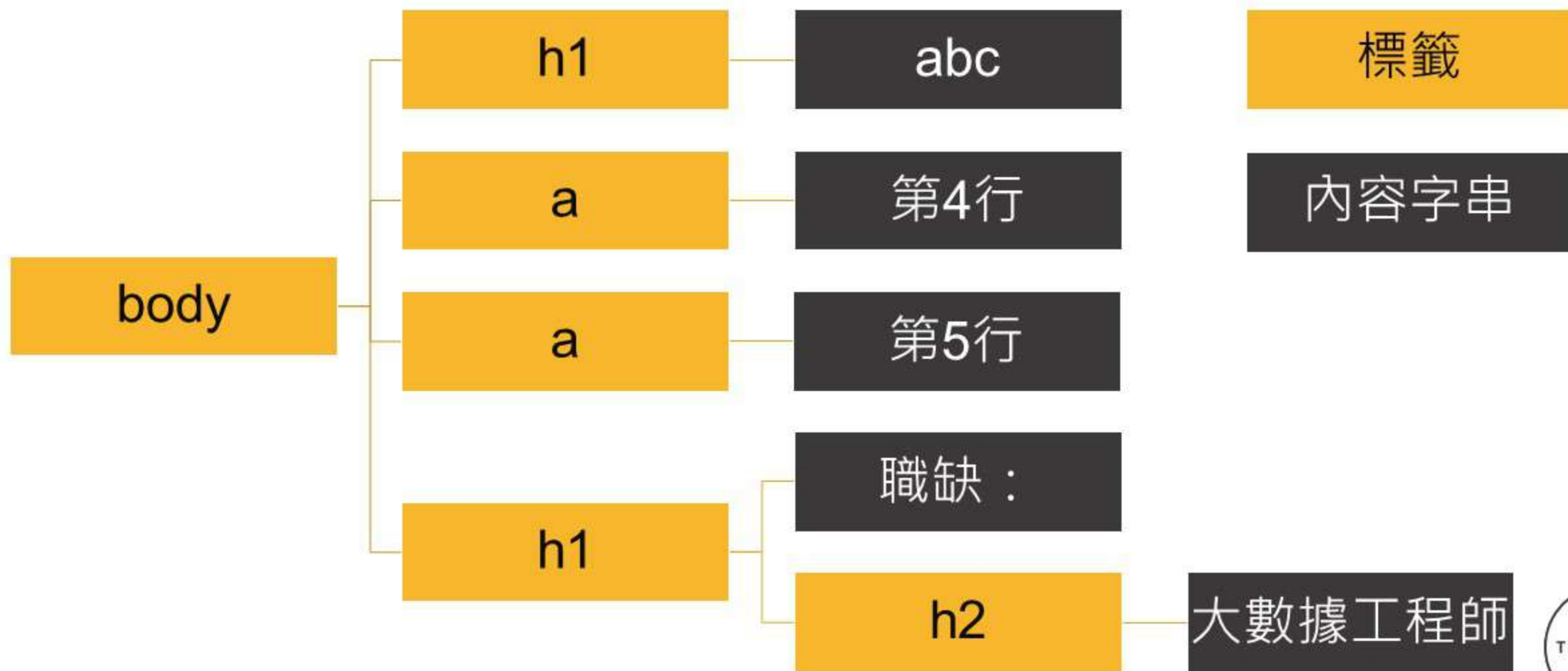
```
  <a> 第4行 </a>
```

```
  <a> 第5行 </a>
```

```
  <h1 class="job">職缺：<h2>大數據工程師</h2></h1>
```

```
</body>
```

# 多層次網頁架構





1.HTML語法簡介

2.BeautifulSoup模組應用

3.104職缺爬蟲專案

# BeautifulSoup常用功能

| 常用指令、函數                   | 功能                             |
|---------------------------|--------------------------------|
| .標籤名稱                     | 跨階層搜索符合標籤名稱的”第一個”標籤            |
| .select(標籤名稱)             | 跨階層搜索符合標籤名稱的所有標籤，並以list回傳      |
| .find(標籤名稱, 標籤屬性名/值對)     | 跨階層搜索符合標籤名稱及標籤屬性的”第一個”標籤       |
| .find_all(標籤名稱, 標籤屬性名/值對) | 跨階層搜索符合標籤名稱及標籤屬性的所有標籤，並以list回傳 |
| .string                   | 回傳純字串標籤的字串                     |
| .text                     | 跨階層回傳標籤底下的所有字串                 |
| 標籤名稱[“屬性名稱”]              | 回傳標籤內，特定屬性名稱的屬性值               |



1.HTML語法簡介

2.BeautifulSoup模組應用

3.104職缺爬蟲專案



# 104職缺爬蟲專案流程

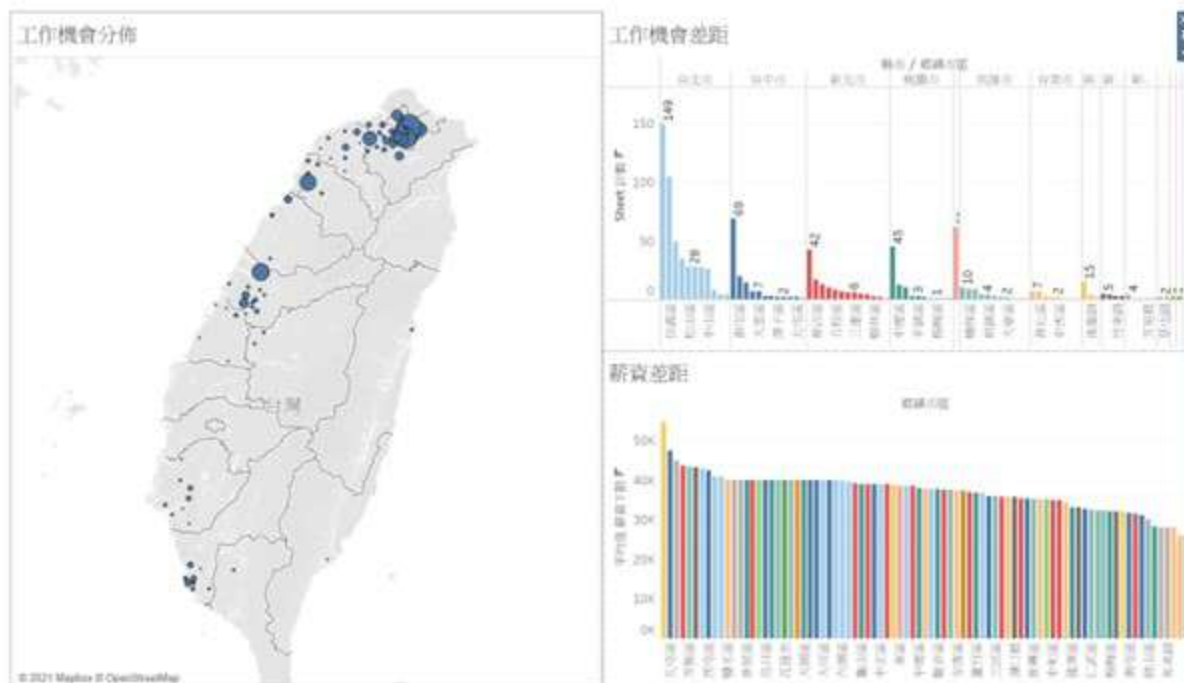
104 人力銀行  
Resource



Data



+tableau



Data Visualization



## 3.104職缺爬蟲專案

(1)資料抓取

(2)導入excel

(3)資料清洗

(4)資料視覺化

# 爬蟲的第一步→資料擷取

In [1]:

```
1 import requests
2 from bs4 import BeautifulSoup
3
4 res=requests.get('https://www.104.com.tw/jobs/search/?keyword=%E5%
5 soup=BeautifulSoup(res.text)
6
7 print(soup)
```

BeautifulSoup物件轉換

```
<!DOCTYPE html>
<html lang="zh-hant"><head>
<meta charset="utf-8"/>
<meta content="IE=edge" http-equiv="X-UA-Compatible"/>
<title>「大數據」找工作職缺 - 2023年5月 | 104人力銀行</title>
<meta content="2023/5/16 - 2236 個工作機會 | 大數據分析【杏豐實業股份有限公司
```



# 抓取最小單位資料(單一職缺)

104人力銀行 My104 履歷診療室 學習精靈 職場力 BeAGiver 快速測驗 更多 帳號 註冊 登入

6/2 新增條件: 新增文章時請注意

大數據 地區 職務類別

更新日期 出對制度 薪資待遇 經歷要求 技能相關 更多條件 排除條件

article.b-block--top-bord.job-list-item.b-clearfix.js-job-item 800 x 162.33

5/08 大數據分析

杏豐實業股份有限公司 藥品 / 化妝品及清潔用品零售業

新北市三重區 1年以上 大學

1. 結構及非結構化數據彙整、整理 2. 大數據分析、報告撰寫與簡報

待遇面議 員工240人 距捷運先靈宮站約300公尺

5/03 AI 大數據科學家

昱峰智能大數據科技股份有限公司 | 電腦軟體服務業

台北市信義區 經歷不拘 專科

1. 運用統計及機器學習等技術, 針對客戶需求, 研發與實作演算法程式碼。 2. 演算法設計、開發與驗證。 3. 協助專案研發與新技術的導入。 4. 數據分析、建模等相關工作, 並進行模型評估。 5. 資料視覺化設計。 6. coding skill in python, pyspark, mongodb

待遇面議 員工10人

## 元素比對

```
"1011004001" data-indcat-desc="不動產經營業" data-is-save="0" data-is-apply="0" data-jobsources="hot job_chr" data-qa-id="jobSeachResult">...</article>
<article class="b-block--top-bord job-list-item b-clearfix js-job-item" data-job-no="13005458" data-job-name="大數據分析" data-job-ro="1" data-cust-no="16998578000" data-cust-name="杏豐實業股份有限公司" data-indcat="1003002005" data-indcat-desc="藥品 / 化妝品及清潔用品零售業" data-is-save="0" data-is-apply="0" data-jobsources="jolist_a_relevance" data-qa-id="jobSeachResult"> == $0
<div class="b-block__left">
  <h2 class="b-tit">
    <span class="b-tit__date">5/08 </span>
    <a href="//www.104.com.tw/job/7qr2q?jobsources=jolist_a_relevance" data-qa-id="jobSeachResultTitle" class="is-job-link" target="_blank">
```

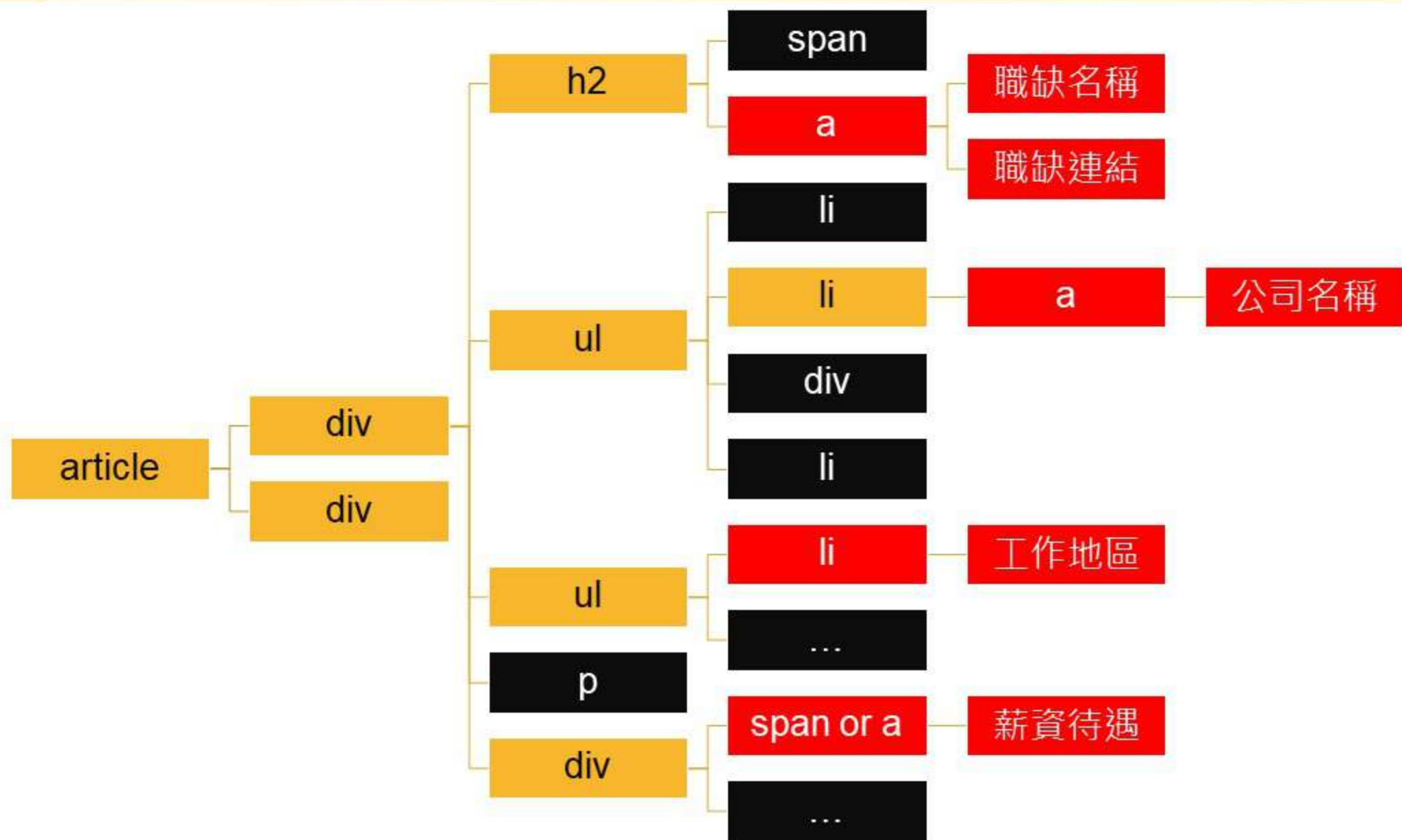
# 以find\_all抓取職缺標籤(article)

```
In [3]: 1 import requests
2 from bs4 import BeautifulSoup
3
4 res=requests.get('https://www.104.com.tw/jobs/search/?keyword=%E5%A4%A7%E6%95%B8%E6%93%9A&order=1&')
5 soup=BeautifulSoup(res.text)
6
7 print(soup.find_all('article',class_="b-block--top-bord job-list-item b-clearfix js-job-item")[0])
```

```
<article class="b-block--top-bord job-list-item b-clearfix js-job-item" data-cust-name="杏豐實業股份有
限公司" data-cust-no="16998578000" data-indcat="1003002005" data-indcat-desc="藥品 / 化妝品及清潔用品零售
業" data-is-apply="0" data-is-save="0" data-job-name="大數據分析" data-job-no="13005458" data-job-ro
="1" data-jobsorce="jolist_d_relevance" data-qa-id="jobSeachResult">
<div class="b-block__left">
<h2 class="b-tit">
<span class="b-tit__date">5/08 </span>
<a class="js-job-link" data-qa-id="jobSeachResultTitle" href="//www.104.com.tw/job/7qr2q?jobsorce=jo
list_d_relevance" target="blank">com class="b-txt highlight">大數據分析</a>
```



# 解析最小單位資料



# 抓取目標資料

職缺名稱

職缺連結

公司名稱

工作地區

```
6 fix js-job-item")[0].a.text)
7
8
9 n b-clearfix js-job-item")[0].a['href'])
10
11 fix js-job-item")[0].find('ul',class_="b-list-inline b-clearfix").a.text.strip())
12
13 fix js-job-item")[0].find('ul',class_="b-list-inline b-clearfix job-list-intro b-content").li.text)
14
15 js-job-item")[0].find('div',class_="job-list-tag b-content").select('span')==[:
16 learfix js-job-item")[0].find('div',class_="job-list-tag b-content").a.text)
17
18 learfix js-job-item")[0].find('div',class_="job-list-tag b-content").span.text)
```

# 薪資待遇不同標籤名稱處理

```
if job.find('div',class_="job-list-tag b-content").select('span')!=[] and job.find('div',class_="job-list-tag b-content").select('span')[0].text=="待遇面議":  
    e=job.find('div',class_="job-list-tag b-content").span.text  
else:  
    e=job.find('div',class_="job-list-tag b-content").a.text
```

如果搜尋article底下的所有span標籤  
回傳不是[]空list 且 第一個span.text是“待遇面議”

則印出span.text → 待遇面議

否則印出a.text → 薪資待遇數字範圍



# 以for迴圈印出整頁20個職缺

```
8 for job in soup.find_all('article',class_="b-block--top-bord job-list-item b-clearfix js-job-item")
9     print(job.a.text)
10
11     print('https:'+job.a['href'])
12
13     print(job.find('ul',class_="b-list-inline b-clearfix").a.text.strip())
14
15     print(job.find('ul',class_="b-list-inline b-clearfix job-list-intro b-content").li.text)
16
17     if job.find('div',class_="job-list-tag b-content").select('span')==[]:
18         print(job.find('div',class_="job-list-tag b-content").a.text)
19     else:
20         print(job.find('div',class_="job-list-tag b-content").span.text)
21
22     print('-----')
```

將find\_all回傳的list中的所有"article標籤" 統稱為job

# 以while迴圈印出所有頁數

```
page=1
while soup.find_all('article',class_="b-block--top-bord job-list-item b-clearfix js-job-item")!=[]:
    print("-----")
    print('正在讀取第',page,'頁...')
    print("=====")
```

Article list沒東西則跳出迴圈

```
for job in soup.find_all('article',class_="b-block--top-bord job-list-item b-clearfix js-job-item"):
    print(job.a.text)

    print('https: '+job.a['href'])

    print(job.find('ul',class_="b-list-inline b-clearfix").a.text.strip())

    print(job.find('ul',class_="b-list-inline b-clearfix job-list-intro b-content").li.text)

    if job.find('div',class_="job-list-tag b-content").select('span')==[]:
        print(job.find('div',class_="job-list-tag b-content").a.text)
    else:
        print(job.find('div',class_="job-list-tag b-content").span.text)

    print('-----')
page+=1
res=requests.get('https://www.104/?keyword=%E5%A4%A7%E6%95%B8%E6%93%9A&order=1&jobsources=2018indexpoc&ro=0&page='+str(page))
soup=BeautifulSoup(res.text)
```



# 以while迴圈印出所有頁數

page=1

```
while soup.find_all('article',class_="b-block--top-bord job-list-item b-clearfix js-job-item")!=[]:  
    print("=====  
    print('正在讀取第',page,'頁...')  
    print("=====
```

```
for job in soup.find_all('article',class_="b-block--top-bord job-list-item b-clearfix js-job-item"):  
    print(job.a.text)  
  
    print('https: '+job.a['href'])  
  
    print(job.find('ul',class_="b-list-inline b-clearfix").a.text.strip())  
  
    print(job.find('ul',class_="b-list-inline b-clearfix job-list-intro b-content").li.text)  
  
    if job.find('div',class_="job-list-tag b-content").select('span')==[]:  
        print(job.find('div',class_="job-list-tag b-content").a.text)  
    else:  
        print(job.find('div',class_="job-list-tag b-content").span.text)
```

page+=1

```
res=soup.get('https://www.104/?keyword=%E5%A4%A7%E6%95%B8%E6%93%9A&order=1&jobsource=2018indexpoc&ro=0&page='+str(page))  
soup=BeautifulSoup(res.text)
```

每一個while迴圈  
都把頁數+1  
以query修改網址

# 跨頁數資料爬取成功

=====  
正在讀取第 112 頁...

=====  
設計師 Art

[https://www.104.com.tw/job/4yuvn?jobsource=jolist\\_a\\_relevance](https://www.104.com.tw/job/4yuvn?jobsource=jolist_a_relevance)

亞瑞特數位社群行銷有限公司

台北市松山區

月薪28,000~38,000元

-----  
資深資料庫工程師-M115

[https://www.104.com.tw/job/49m93?jobsource=jolist\\_a\\_relevance](https://www.104.com.tw/job/49m93?jobsource=jolist_a_relevance)

精誠資訊股份有限公司

台北市中正區

待遇面議  
-----



## 3.104職缺爬蟲專案

(1)資料抓取

(2)導入excel

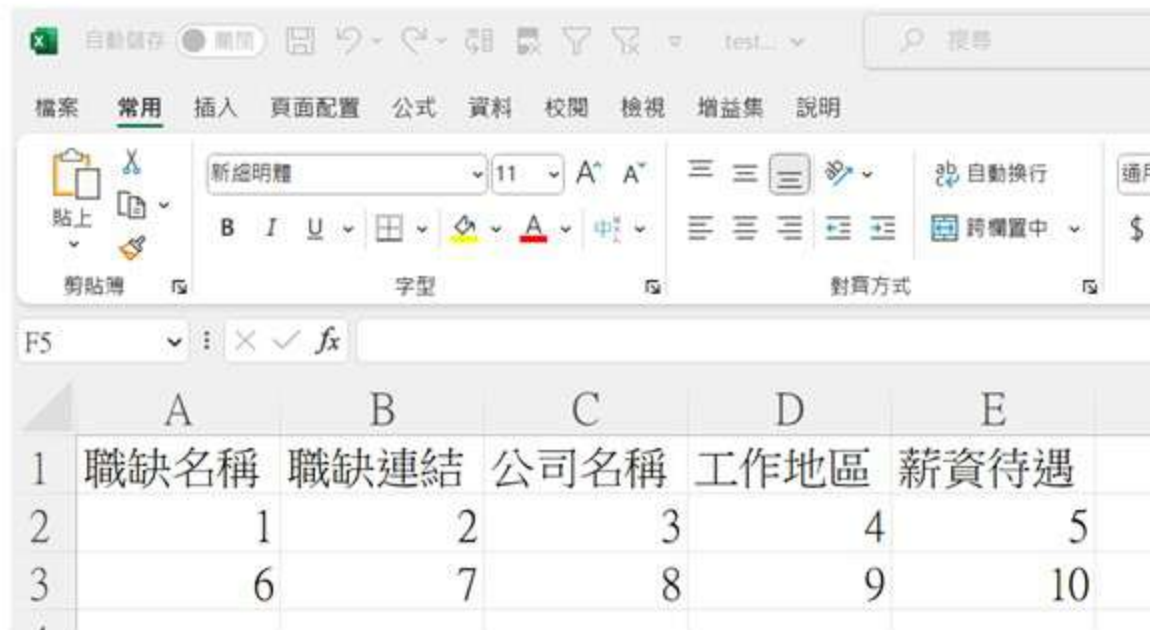
(3)資料清洗

(4)資料視覺化



# 用Python製作Excel資料表→openpyxl

```
In [ ]: 1 import openpyxl
        2
        3 #建立新工作簿workbook=xlsx檔案
        4 wb = openpyxl.Workbook()
        5
        6 #在工作簿中啟用一個新工作表worksheet
        7 ws = wb.active
        8
        9 #在第一列定義好欄位名稱
       10 ws['A1'] = '職缺名稱'
       11 ws['B1'] = '職缺連結'
       12 ws['C1'] = '公司名稱'
       13 ws['D1'] = '工作地區'
       14 ws['E1'] = '薪資待遇'
       15
       16 #用append一次匯入一列資料
       17 ws.append([1,2,3,4,5])
       18 ws.append([6,7,8,9,10])
       19
       20 #儲存工作簿
       21 wb.save('C:\\test\\test0703.xlsx')
```



|   | A    | B    | C    | D    | E    |
|---|------|------|------|------|------|
| 1 | 職缺名稱 | 職缺連結 | 公司名稱 | 工作地區 | 薪資待遇 |
| 2 | 1    | 2    | 3    | 4    | 5    |
| 3 | 6    | 7    | 8    | 9    | 10   |

# 在開始爬蟲之前完成excel設定

In [\*]:

```
1 import requests
2 from bs4 import BeautifulSoup
3 import openpyxl
4
5 #建立新工作簿workbook=xlsx檔案
6 wb = openpyxl.Workbook()
7
8 #在工作簿中啟用一個新工作表worksheet
9 ws = wb.active
10
11 #在第一列定義好欄位名稱
12 ws['A1'] = '職缺名稱'
13 ws['B1'] = '職缺連結'
14 ws['C1'] = '公司名稱'
15 ws['D1'] = '工作地區'
16 ws['E1'] = '薪資待遇'
17
18 res=requests.get('https://www.104.com.tw/jobs/search/?key
19 soup=BeautifulSoup(res.text)
20
```

爬蟲中置入  
openpyxl功能  
與excel設定



# 在爬蟲迴圈中直接匯入Excel檔

```
while soup.find_all('article',class_="b-block--top-bord job-list-item b-clearfix js-job-item")
    print("=====")
    print('正在讀取第',page,'頁...')
    print("=====")
    for job in soup.find_all('article',class_="b-block--top-bord job-list-item b-clearfix js-j
        a=job.a.text

        b='https:'+job.a['href']

        c=job.find('ul',class_="b-list-inline b-clearfix").a.text.strip()
        d=job.find('ul',class_="b-list-inline b-clearfix job-list-intro b-content").li.text

        if job.find('div',class_="job-list-tag b-content").select('span')==[]:
            e=job.find('div',class_="job-list-tag b-content").a.text
        else:
            e=job.find('div',class_="job-list-tag b-content").a.text

        ws.append([a,b,c,d,e])

    page+=1
res=requests.get('https://www.104.com.tw/jobs/search/?keyword=%E5%A4%A7%E6%95%B8%E6%93%9A&
soup=BeautifulSoup(res.text)
```

將之前print的五種資料  
定義成物件(a,b,c,d,e)

並用append([])匯入

#儲存工作簿

wb.save('職缺清單.xlsx')

完成記得存檔!

# 資料匯入成功

| G2242 |                                    |   |                       |        |                       |   |
|-------|------------------------------------|---|-----------------------|--------|-----------------------|---|
|       | A                                  | B   | C                     | D      | E                     | F |
| 1     | 職缺名稱                               | 職缺連結  | 公司名稱                  | 工作地區   | 薪資待遇                  |   |
| 2226  | 無經驗保障六個月業務補貼3萬5                    | <a href="https://www.104.com.tw/job/7brke?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/7brke?jobsourc=jolist_b_relevance</a> | 蔡鼎地產股份有限公司            | 台中市北屯區 | 月薪30,000~100,000元     |   |
| 2227  | JAVA後端工程師 I F000002402             | <a href="https://www.104.com.tw/job/7eni9?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/7eni9?jobsourc=jolist_b_relevance</a> | 104獵才顧問 一零四資訊科技股份有限公司 | 台北市松山區 | 待遇面議                  |   |
| 2228  | 無經驗保障六個月業務補貼                       | <a href="https://www.104.com.tw/job/7cdk0?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/7cdk0?jobsourc=jolist_b_relevance</a> | 蔡鼎地產股份有限公司水滄蔡鴻分公司     | 台中市北屯區 | 月薪30,000~100,000元     |   |
| 2229  | 【我在找你，你知道嗎？】                       | <a href="https://www.104.com.tw/job/7jp7e?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/7jp7e?jobsourc=jolist_b_relevance</a> | 蔡鼎地產股份有限公司水滄蔡鴻分公司     | 台中市北屯區 | 月薪35,000元以上           |   |
| 2230  | 永慶不動產儲備店長                          | <a href="https://www.104.com.tw/job/7sjkp?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/7sjkp?jobsourc=jolist_b_relevance</a> | 永慶不動產清華加盟店 捷報不動產仲     | 新竹市    | 論件計酬60,000~2,000,000元 |   |
| 2231  | iOS程式設計師                           | <a href="https://www.104.com.tw/job/62ln2?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/62ln2?jobsourc=jolist_b_relevance</a> | 移動商務股份有限公司            | 台北市南港區 | 待遇面議                  |   |
| 2232  | 不動產高專經紀人((台北市))                    | <a href="https://www.104.com.tw/job/42xny?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/42xny?jobsourc=jolist_b_relevance</a> | 群義房屋 群義不動產經紀股份有限公     | 台北市松山區 | 月薪26,400~100,000元     |   |
| 2233  | Android開發工程師(媒體營運部)                | <a href="https://www.104.com.tw/job/3jql4?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/3jql4?jobsourc=jolist_b_relevance</a> | 艾堡媒體有限公司              | 台北市中山區 | 待遇面議                  |   |
| 2234  | 保險業務員/保險業務儲備幹部(特                   | <a href="https://www.104.com.tw/job/71aiy?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/71aiy?jobsourc=jolist_b_relevance</a> | 新光人壽保險股份有限公司 仰德通訊     | 台北市中山區 | 月薪26,400~40,000元      |   |
| 2235  | RD20165 Senior Data Infrastructure | <a href="https://www.104.com.tw/job/7bra8?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/7bra8?jobsourc=jolist_b_relevance</a> | 華碩電腦股份有限公司            | 台北市北投區 | 待遇面議                  |   |
| 2236  | 【HealthToday】網站工程師 Full            | <a href="https://www.104.com.tw/job/7vj9n?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/7vj9n?jobsourc=jolist_b_relevance</a> | 大源國際實業有限公司            | 台北市內湖區 | 月薪45,000~70,000元      |   |
| 2237  | 研發後端工程師                            | <a href="https://www.104.com.tw/job/7zvj?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/7zvj?jobsourc=jolist_b_relevance</a>   | 成強科技股份有限公司            | 新北市林口區 | 待遇面議                  |   |
| 2238  | Data Architect 數據架構師               | <a href="https://www.104.com.tw/job/6jhb6?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/6jhb6?jobsourc=jolist_b_relevance</a> | 德義資訊股份有限公司            | 台北市大同區 | 待遇面議                  |   |
| 2239  | 5萬每月最高保障不動產仲介儲備                    | <a href="https://www.104.com.tw/job/7wch8?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/7wch8?jobsourc=jolist_b_relevance</a> | 中信房屋 昱信不動產仲介經紀有限公     | 基隆市安樂區 | 月薪35,000~50,000元      |   |
| 2240  | Android / iOS APP 軟體開發工程師          | <a href="https://www.104.com.tw/job/7n8f5?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/7n8f5?jobsourc=jolist_b_relevance</a> | 亞瑞特數位社群行銷有限公司         | 台北市松山區 | 月薪28,000~40,000元      |   |
| 2241  | 【不動產經紀人員】(台北市全區                    | <a href="https://www.104.com.tw/job/6bo5s?jobsourc=jolist_b_relevance">https://www.104.com.tw/job/6bo5s?jobsourc=jolist_b_relevance</a> | 信義房屋股份有限公司            | 台北市中山區 | 月薪50,000元             |   |
| 2242  |                                    |   |                       |        |                       |   |



# 補充：爬蟲速度控管→time.sleep()

```
1 import requests
2 from bs4 import BeautifulSoup
3 import openpyxl
4 from time import sleep
5
6 #建立新工作簿workbook=xlsx檔案
7 wb = openpyxl.Workbook()
8
9 #在工作簿中啟用一個新工作表worksheet
10 ws = wb.active
11
```

```
40
41 ws.append([a,b,c,d,e])
42 page+=1
43 res=requests.get('https://www.104.com.tw/jobs/sear
44 soup=BeautifulSoup(res.text)
45
46 #儲存工作簿
47 wb.save('職缺清單.xlsx')
48 sleep(2)
```

每頁爬取完成休息2秒  
以避免被鎖ip反爬蟲



# 補充：自動帶入系統日期於檔名

In [1]:

```
1 import requests
2 from bs4 import BeautifulSoup
3 import openpyxl
4 from time import sleep
5 import datetime
6
7 #建立新工作簿workbook=xlsx檔案
8 wb = openpyxl.Workbook()
```



職缺清單2022-08-11.xlsx

```
42     ws.append([a,b,c,d,e])
43     page+=1
44     res=requests.get('https://www.104.com.tw/jobs/search/?key
45     soup=BeautifulSoup(res.text)
46
47     #儲存工作簿
48     wb.save('職缺清單'+str(datetime.date.today())+'.xlsx')
49     sleep(2)
```





## 3.104職缺爬蟲專案

(1)資料抓取

(2)導入excel

(3)資料清洗

(4)資料視覺化

# 新增薪資細節欄位

```
7 /  
8 #開啟並設定好excel檔及欄位名稱作為資料容器  
9 wb=openpyxl.Workbook()  
10 ws=wb.active  
11 ws['A1']='職缺名稱'  
12 ws['B1']='職缺連結'  
13 ws['C1']='公司名稱'  
14 ws['D1']='工作地區'  
15 ws['E1']='薪資待遇'  
16 ws['F1']='給薪方式'  
17 ws['G1']='薪資下限'  
18 ws['H1']='薪資上限'  
19
```

# 提取給薪方式

注意資料清洗階層→僅有數字  
字的薪資待遇做處理

```
#針對搜尋結果第一層，把所有找到的資料，把每層資料都存到job
for job in soup.find_all('article',class_="b-block--top-bord job-list-item b-clearfix js-job-item"):
    #以下為薪資待遇做處理
    a=job.h2.a.text
    b='https:'+job.h2.a['href']
    c=job.ul.a.text.strip()
    d=job.find('ul',class_="b-list-inline b-clearfix job-list-intro b-content").li.text

    if job.find('div',class_="job-list-tag b-content").select('span')!=[] and job.find('div',class_="
        e=job.find('div',class_="job-list-tag b-content").span.text
    else:
        e=job.find('div',class_="job-list-tag b-content").a.text

    #量化薪資資料清洗
    #取前兩字作為給薪方式f
    f=e[:2]
```

取a.text前兩個字作為  
給薪方式類別

# 清洗多餘字元

```
#清洗多餘字元：中文、，  
salary=''   
for char in e:  
    if char.isdigit() or char=='~':  
        salary+=char
```



## 區分範圍(“~”)，提取薪資上下限

```
#區分有無"~"，並取上下限
if '~' in salary:
    g=salary[:salary.find('~')]
    h=salary[salary.find('~')+1:]
else:
    g=salary
    h=salary
```

## 薪資轉換成數值型態，以利Excel計算平均

#將薪資從字串轉換成int

g=int(g)

h=int(h)

# 匯入清洗完成資料

```
if job.find('div',class_="job-list-tag b-content").select('span')  
e=job.find('div',class_="job-list-tag b-content").span.text  
f=''  
g=''  
h=''
```

「待遇面議」則後三格給空字串

記得將後三格加入ws.append以匯入Excel

```
ws.append([a,b,c,d,e,f,g,h])
```



# 薪資待遇資料清洗完成

|    | A       | B   | C    | D    | E                | F    | G     | H     |
|----|---------|---|------|------|------------------|------|-------|-------|
| 1  | 職缺名稱    | 職缺連結  | 公司名稱 | 工作地區 | 薪資待遇             | 給薪方式 | 薪資下限  | 薪資上限  |
| 2  | AI 大數據  | <a href="https://www.https://www">https://www</a> | 昱峰智能 | 台北市信 | 待遇面議             |      |       |       |
| 3  | 大數據實    | <a href="https://www.https://www">https://www</a> | 台北神策 | 台北市中 | 時薪180~200元       | 時薪   | 180   | 200   |
| 4  | 大數據分    | <a href="https://www.https://www">https://www</a> | 杏豐實業 | 新北市三 | 待遇面議             |      |       |       |
| 5  | [DI] 大數 | <a href="https://www.https://www">https://www</a> | 動力安全 | 台北市內 | 月薪30,000~40,000元 | 月薪   | 30000 | 40000 |
| 6  | 大數據產    | <a href="https://www.https://www">https://www</a> | 台北神策 | 台北市中 | 時薪180~200元       | 時薪   | 180   | 200   |
| 7  | 大數據部    | <a href="https://www.https://www">https://www</a> | 典通股份 | 台北市中 | 月薪40,000~45,000元 | 月薪   | 40000 | 45000 |
| 8  | 產品業務    | <a href="https://www.https://www">https://www</a> | 大數據股 | 台北市中 | 待遇面議             |      |       |       |
| 9  | 大數據部    | <a href="https://www.https://www">https://www</a> | 典通股份 | 台北市中 | 月薪40,000~45,000元 | 月薪   | 40000 | 45000 |
| 10 | 【智慧製    | <a href="https://www.https://www">https://www</a> | 台灣恩悌 | 高雄市鼓 | 待遇面議             |      |       |       |
| 11 | 語意大數    | <a href="https://www.https://www">https://www</a> | 亞洲指標 | 台北市松 | 月薪35,000~60,000元 | 月薪   | 35000 | 60000 |
| 12 | 大數據部    | <a href="https://www.https://www">https://www</a> | 典通股份 | 台北市中 | 月薪43,000~50,000元 | 月薪   | 43000 | 50000 |
| 13 | 【資訊】    | <a href="https://www.https://www">https://www</a> | 群益金鼎 | 台北市松 | 待遇面議             |      |       |       |
| 14 | 健康大數    | <a href="https://www.https://www">https://www</a> | 典通股份 | 台北市中 | 月薪40,000~45,000元 | 月薪   | 40000 | 45000 |
| 15 | 大數據分    | <a href="https://www.https://www">https://www</a> | 博揚機械 | 新北市橫 | 月薪38,000~42,000元 | 月薪   | 38000 | 42000 |
| 16 | 大數據產    | <a href="https://www.https://www">https://www</a> | 台北神策 | 台北市中 | 待遇面議             |      |       |       |
| 17 | 【研究發    | <a href="https://www.https://www">https://www</a> | 現觀科技 | 台北市中 | 月薪35,000~55,000元 | 月薪   | 35000 | 55000 |
| 18 | 大數據分    | <a href="https://www.https://www">https://www</a> | 群健科技 | 台北市中 | 時薪200~250元       | 時薪   | 200   | 250   |
| 19 | 大數據分    | <a href="https://www.https://www">https://www</a> | 中嘉數位 | 台北市內 | 待遇面議             |      |       |       |



# 將縣市及鄉鎮市區層級切分

|   | D      | E           |      |
|---|--------|-------------|------|
|   | 工作地區   | 縣市          | 鄉鎮市區 |
| 支 | 台北市信義區 | =LEFT(D2,3) | 信    |
| 部 | 台北市中山區 |             |      |

|   | D      | E   | F                   |      |
|---|--------|-----|---------------------|------|
|   | 工作地區   | 縣市  | 鄉鎮市區                | 薪資類別 |
| 支 | 台北市信義區 | 台北市 | =REPLACE(D2,1,3,"") | 待命   |
| 部 | 台北市中山區 | 台北市 |                     | 時薪   |
| 注 | 新北市三重區 | 新北市 |                     | 待命   |

# 計算薪資平均

|    | G     | H     | I               |
|----|-------|-------|-----------------|
| 方▼ | 薪資下▼  | 薪資上▼  | 薪資平均            |
|    | 180   | 200   | =AVERAGE(G5:H5) |
|    | 35000 | 60000 |                 |



## 3.104職缺爬蟲專案

(1)資料抓取

(2)導入excel

(3)資料清洗

(4)資料視覺化

# Tableau Public



定價 登入 

為什麼選擇 Tableau ▾ 產品 ▾ 解決方案 ▾ 資源 ▾ 合作夥伴 ▾

立即試用

立即購買

## 幾分鐘內就能開始探索

在幾分鐘內建立互動式圖形、絕佳的地圖和即時儀表板。將您的視覺化項儲存到 Tableau Public 設定檔，並在網路上的任何位置分享。任何人都可以做到，就是這麼簡單，而且免費。

下載 TABLEAU PUBLIC

2023.1.2 適用於 WINDOWS 和D MAC | [系統需求](#)

<https://www.tableau.com/zh-tw/products/public/download>





# 匯入Excel資料表

Tableau Public - 工作簿 1

檔案(F) 資料(D) 說明(H)

## 連線

到本地

- Microsoft Excel
- 文字檔
- JSON 檔案
- Microsoft Access
- PDF 檔案
- 空間檔案
- 統計檔案

到伺服器

- OData
- 更多...

儲存到本機，使用大資料，連接到更多資料來源。

立即升級

## 開啟

組合管理 新增資料夾

常用

孫善堂 - 個人

桌面

下載

文件

圖片

Google 雲端

20221102探訪

檔案名稱(N): 大數據職缺清單2023-05-23.xlsx

Excel 工作簿(\*.xls \*.xlsx \*.xlsm)

開啟(O) 取消

## 探索

教學影片

概述

介面基本介紹

圖表類型

更多教學影片...

本日精選視覺化作品

探索本日精選視覺化作品

部落格 - 閱讀最新文章

範例資料組

目前狀態

# 地理角色資料格式：縣市→州/省



# 地理角色資料格式：鄉鎮市區→郡/縣



# 篩選雜亂資料(國外、無分區)

青單2023-05-23)

篩選條件

0 新增

編輯資料來源篩選條件

篩選條件 詳細資訊

新增... 編輯... 移除

確定 取消

新增篩選條件

選擇欄位:

搜尋

公司名稱  
工作地區  
給薪方式  
縣市  
職缺名稱  
職缺連結  
薪資上限  
薪資下限  
薪資平均  
薪資待遇  
鄉鎮市區

確定 取消

剔除國外資料

剔除無分區資料

個欄位 2250 個資料行

100

列

設定

下拉





# 建立階層資料

搜尋

表

- Abc 公司名稱
- Abc 工作地區
- Abc 給薪方式
- 縣市**
- Abc 職缺名稱
- Abc 職缺連結
- Abc 薪資待遇
- 鄉鎮市區**
- Abc 度量名稱
- # 薪資上限
- # 薪資下限
- # 薪資平均
- # Sheet (計數)
- 經度(產生)
- 緯度(產生)
- # 度量值

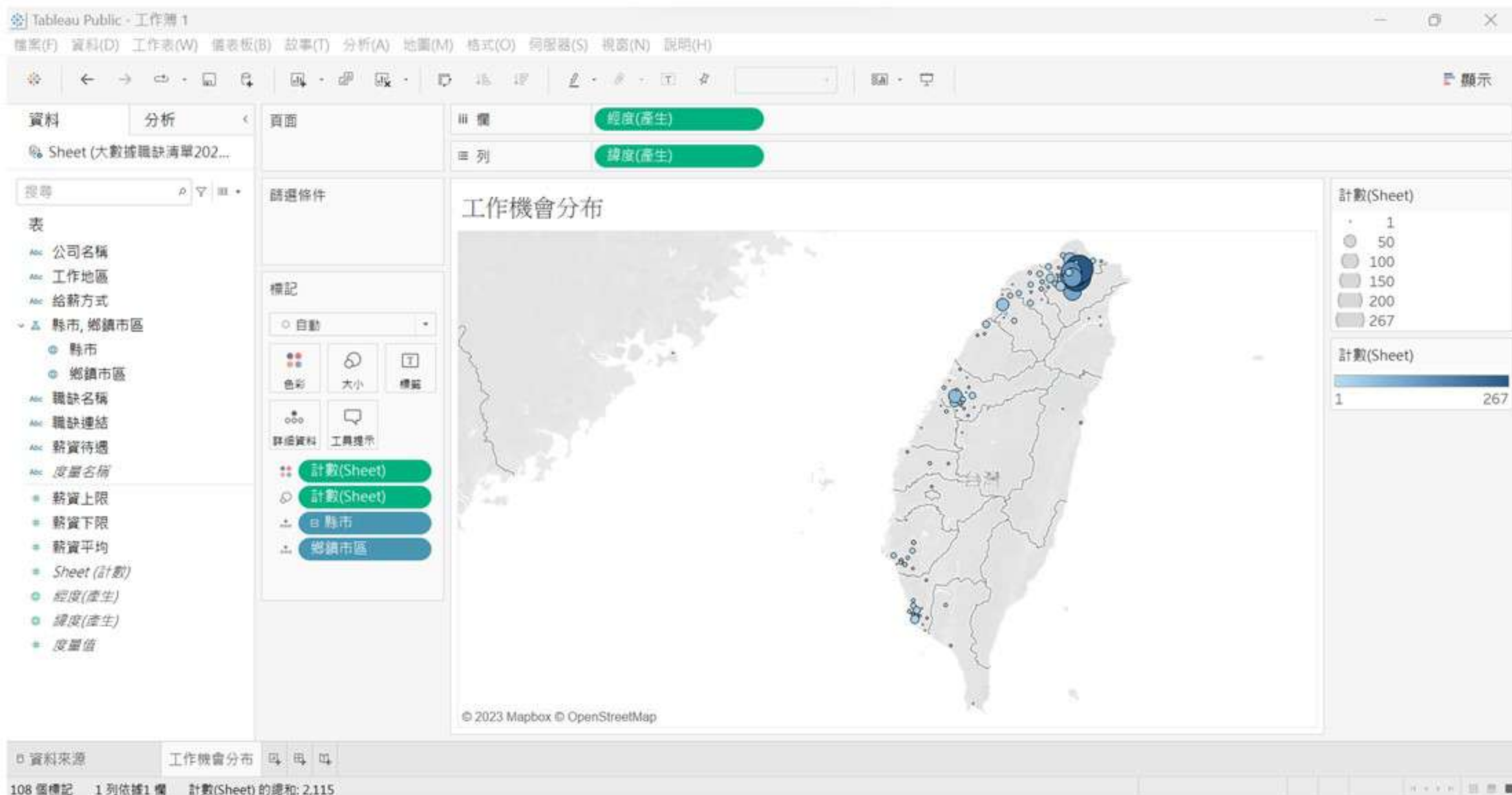


搜尋

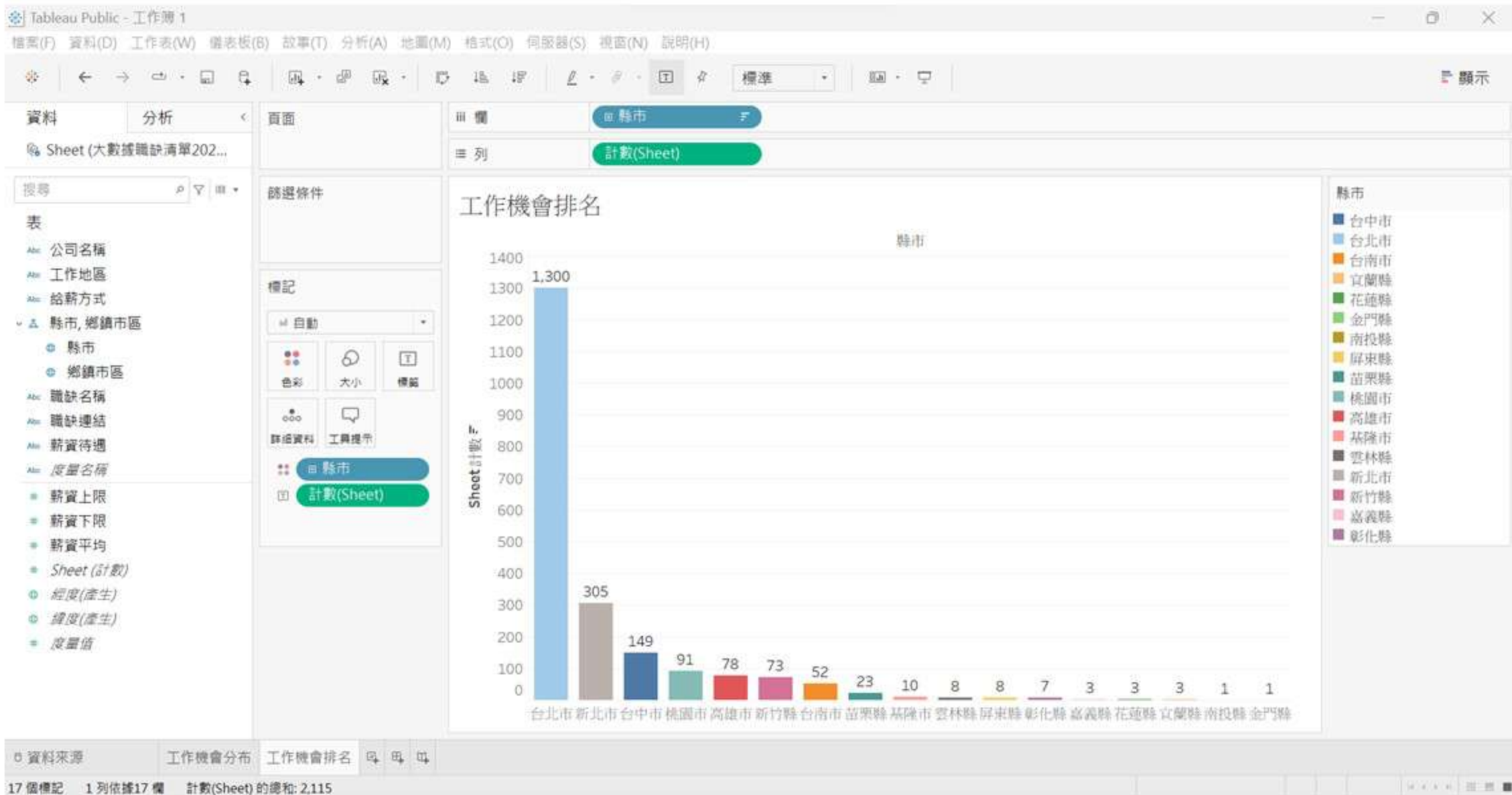
表

- Abc 公司名稱
- Abc 工作地區
- Abc 給薪方式
- 縣市, 鄉鎮市區
- 縣市
- 鄉鎮市區
- Abc 職缺名稱
- Abc 職缺連結
- Abc 薪資待遇
- Abc 度量名稱
- # 薪資上限
- # 薪資下限
- # 薪資平均
- # Sheet (計數)
- 經度(產生)
- 緯度(產生)
- # 度量值

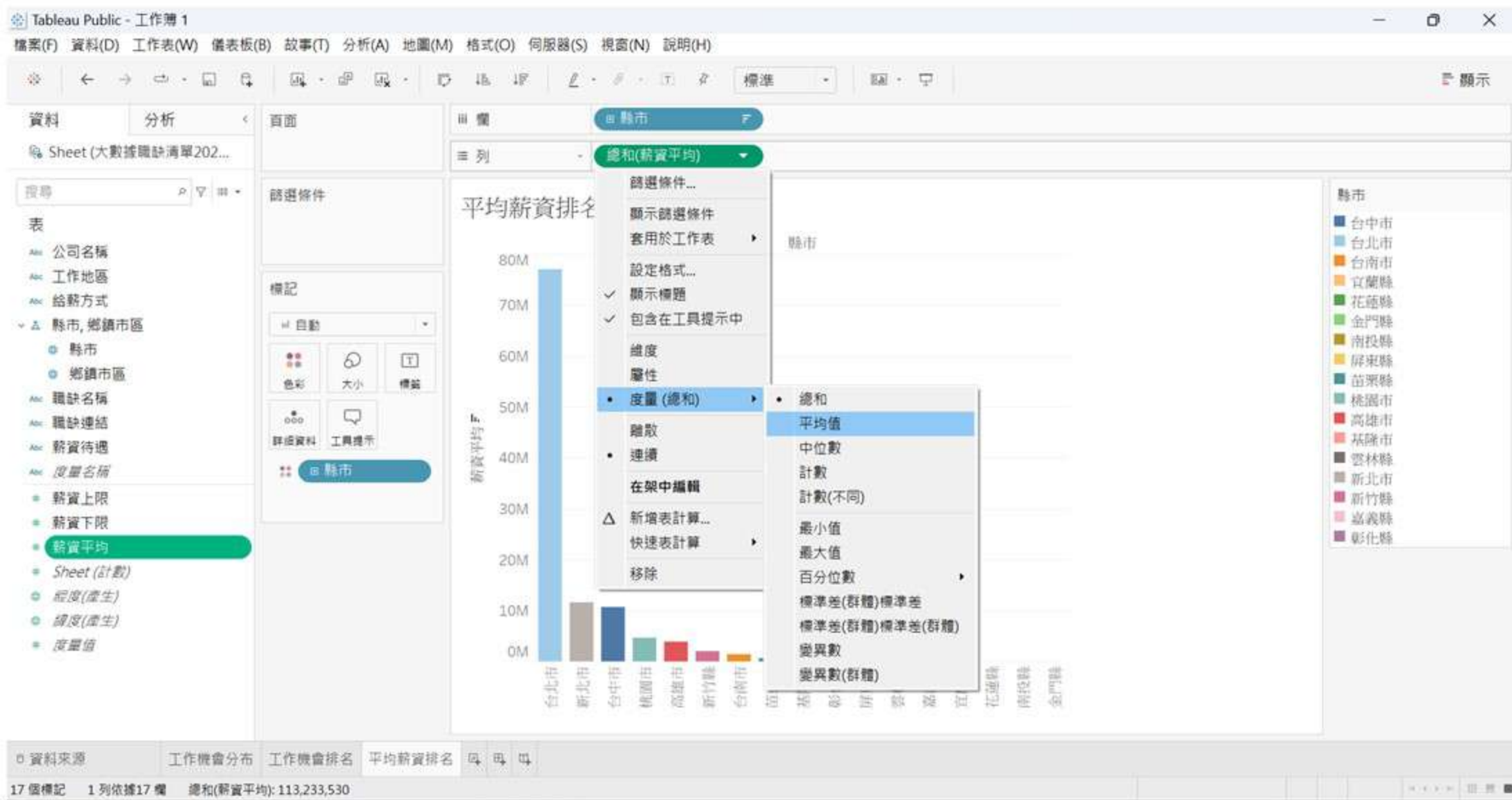
# 工作機會分布圖



# 工作機會排名

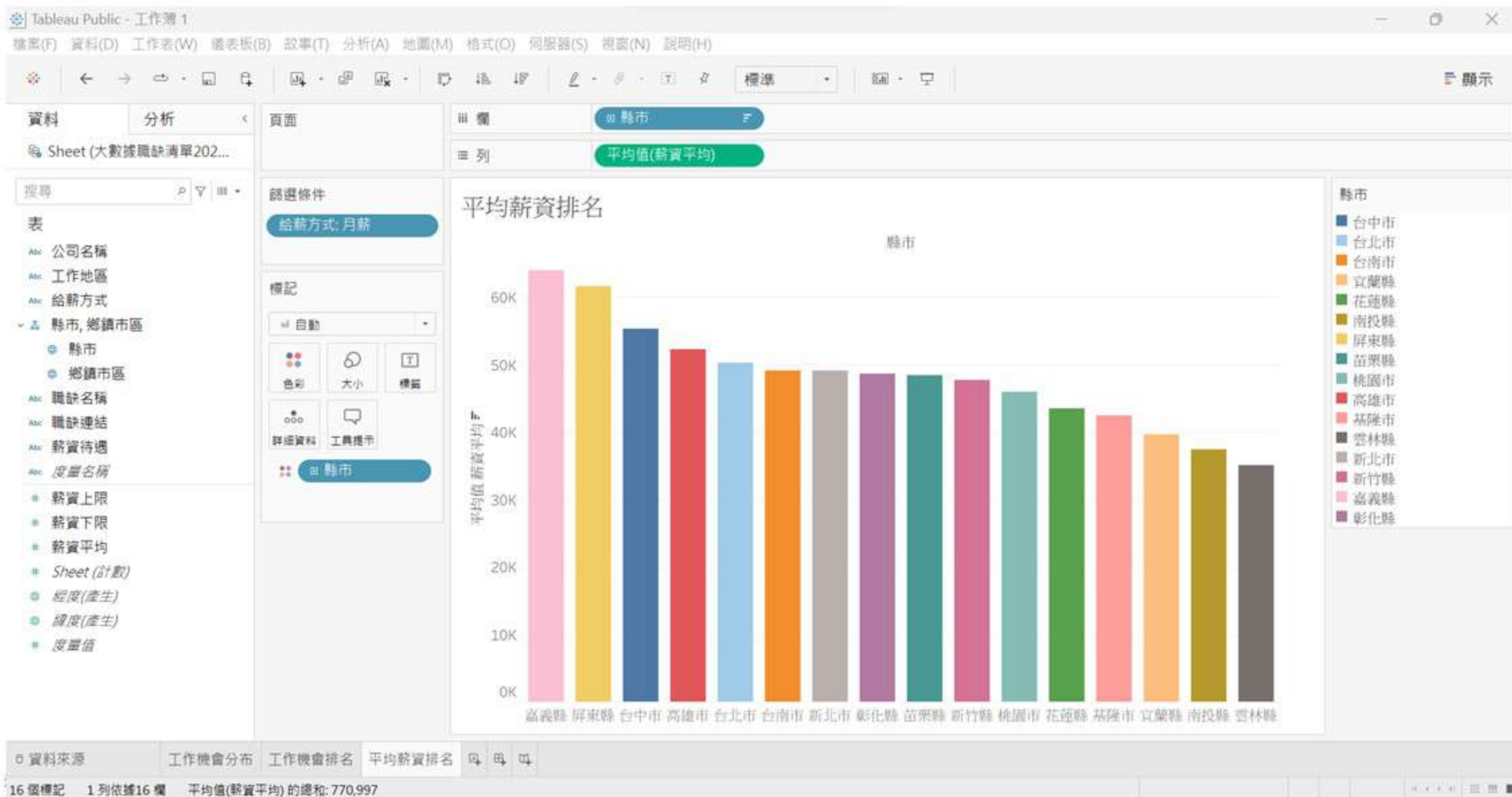


# 平均薪資排名



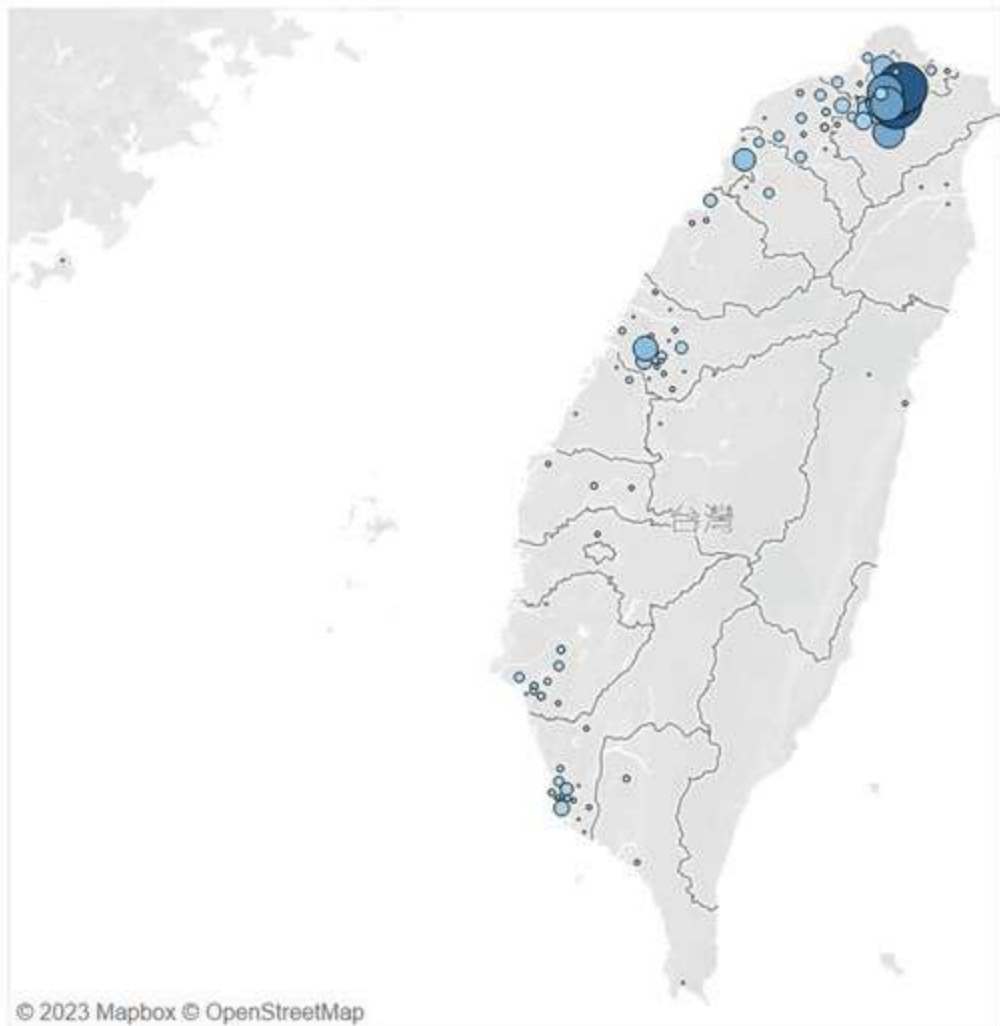


# 平均薪資排名



# 104大數據職缺分布視覺化呈現

工作機會分布



工作機會排名



平均薪資排名



