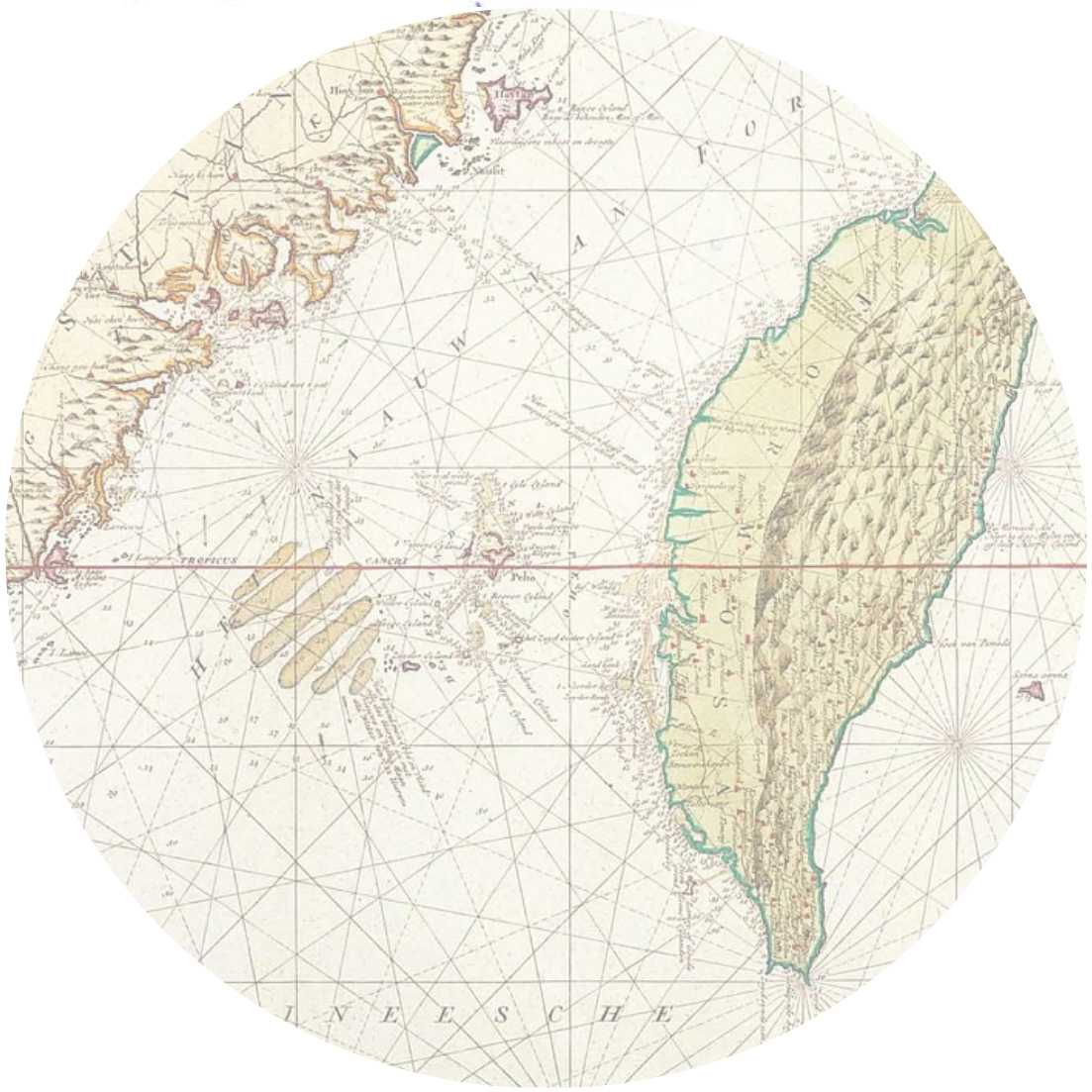
# MACHINE LEARNING FOR PRESIDENTAL ELECTION PREDICTION




**Dr. Chun-Hsiang Chan**

Department of Geography
National Taiwan Normal University

# OUTLINE

- Objectives
- Data Collection
- Mindset in Data Preprocessing
- Data Preprocessing (Cleaning)
- Data Integration
- Summary

# OBJECTIVES

- 我們先以2024年臺灣總統大選作為練習資料集：
  從{**臺北市|新北市|臺中市|高雄市**}的社會經濟資料與投票率來預測{**民進黨|國民黨|民眾黨**}於2024年總統大選的{**得票數|是否在該選區(村里)當選**}?
- 從這個角度出發，我們就需要收集你們感興趣的縣市之社會經濟資料與投票相關資料集。。。

# DATA COLLECTION

- 在社會科學中，我們常以社會經濟特性來了解某一地區對於某一現象的影響，因為我們假設來自相同或是類似社會經濟地位的人會有類似的行為特徵，所以這次的練習會先結合社會經濟相關變相，來了解究竟在不同縣市社會經濟對於總統大選的各政黨候選人的偏好的預測力。
- 如果你想知道究竟不同社會經濟地位對於政黨候選人的偏好，以後你可以利用階層式回歸模型或是廣義線性模型來回答。

# DATA COLLECTION

- 為了資料筆數可以足夠進行模型訓練，我們以村里作為所有的空間解析度，資料時間可以盡量貼近2024年為基準去做收集。
- 資料筆數對於模型的訓練成果至關重要，因為資料筆數越多代表樣本所涵蓋的資料多樣性也越高，可以增進模型的穩健度跟未來預測的準確性。
- 以臺北市為例，總共有456里(筆)可以作為我們訓練模型之用的資料；也是因為如此，我們目前就先以村里數量比較高的六都，同時又可以呈現臺灣北中南屬性的四個縣市做為代表。

# DATA COLLECTION

- 首先，我們先從社會經濟資料集先下手，社會人口相關的變相資料，我們可以透過社會經濟資料服務平台下載到：
  1) 113年6月行政區**人口統計**_村里_{縣市}
  2) 113年6月行政區**人口指標**_村里_{縣市}
  3) 113年6月行政區**五歲年齡組**性別人口統計_村里_{縣市}
  4) 112年行政區**15歲以上人口教育程度**統計_村里_{縣市}

  *以下用{桃園市}作為示範案例，請各位依照自己所分配到縣市下載資料。

# DATA COLLECTION

- 而**經濟相關**的變相資料，我們透過政府資料開放平臺下載：

  1) 110年度**綜稅綜合所得總額**各縣市鄉鎮村里統計分析表-{縣市}

  *以下用{桃園市}作為示範案例，請各位依照自己所分配到縣市下載資料。

---

**綜稅綜合所得總額各縣市鄉鎮村里統計分析表-桃園市** CSV

提供綜合所得稅之綜合所得總額及相關統計值(平均數、中位數等資訊)於桃園市的鄉鎮村里統計分析表 單位：金額(千元) 本項統計資料來源為各年度綜合所得稅申報核定資(...詳內)

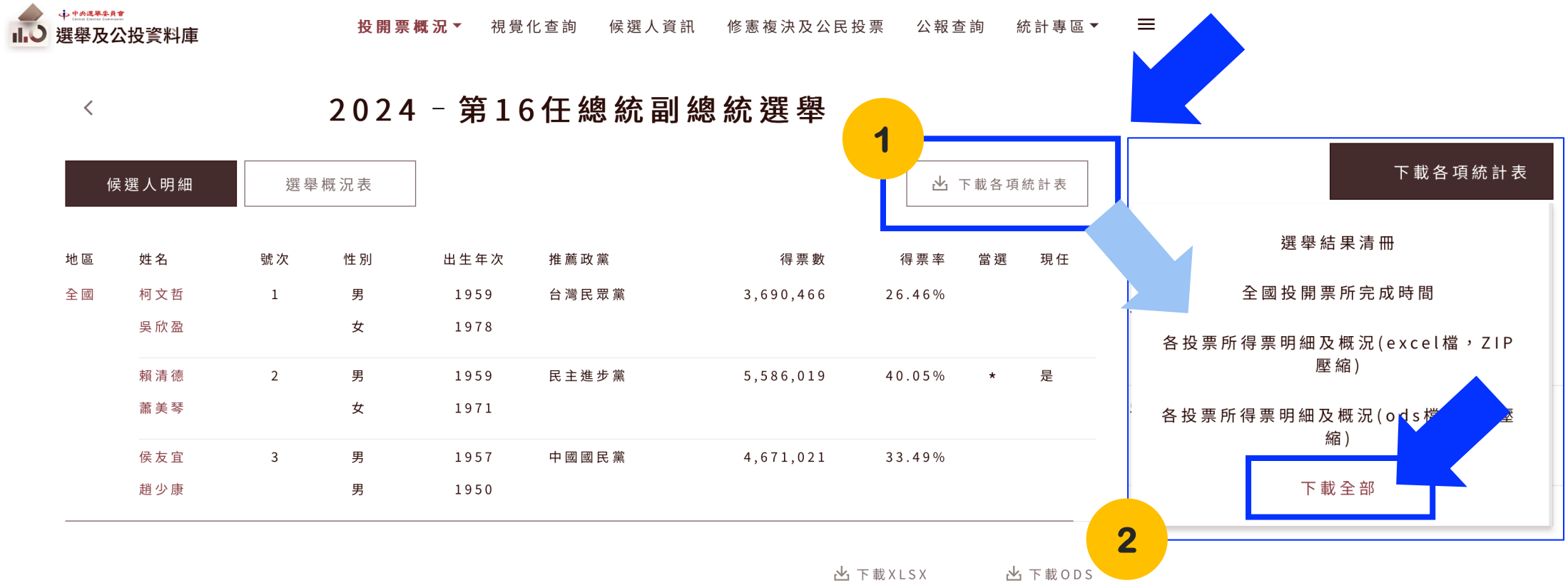
**主要欄位說明:** 鄉鎮市區、村里、納稅單位(戶)、綜合所得總額、平均數、中位數、第一分位數、第三分位數、標準差、變異係數

**資料提供屬性:** 檔案資料

財政部財政資訊中心 2024-08-08 16:50:03 更新 5882 1276 2

# DATA COLLECTION

- 投票結果的資料可以從中央選舉委員會下載:




將下載下來的資料夾解壓縮{4d83…c800}/總統-各投票所得票明細及概況(Excel檔)/
**總統-A05-4-候選人得票數一覽表-各投開票所({縣市}).xlsx**

# MINDSET IN DATA PREPROCESSING

- 在資料前處理的過程中，我們的目標就是將不同來源的資料整併成單一的資料集，但是會因為不同來源的資料及建立方式不同，會有不同的潛在問題，例如：
  1) 資料欄位名稱相同但定義不同：縣市與村里尺度的人口數資料都會用相同的名稱(人口數)，但是空間統計範圍不同
  2) 資料具有遺漏值：可能會以NaN、Null、-9999或是-1出現
  3) 資料格式不一致：台北市與臺北市
  4) 資料解析度不一致：村里資料與鄉鎮市區資料整併時
  5) 資料值有亂碼：「\u3000龍潭區」的前方有亂碼，有些會是有隱藏的空格等

# DATA PREPROCESSING (CLEANING)

- 因我們需要將不同資料及合併，故我們需要先觀察每個資料集的欄位內容，尋找可以作為索引值的欄位，以利之後進行資料合併作業：
  1) 113年6月行政區**人口統計**_村里_**{縣市}**
  2) 113年6月行政區**人口指標**_村里_**{縣市}**
  3) 113年6月行政區**五歲年齡組**性別人口統計_村里_**{縣市}**
  4) 112年行政區**15歲以上人口教育程度**統計_村里_**{縣市}**
  5) 110年度**綜稅綜合所得總額**各縣市鄉鎮村里統計分析表-**{縣市}**
  6) 總統-A05-4-**候選人得票數一覽表-各投開票所**(**{縣市}**)

# DATA PREPROCESSING (CLEANING)

## 1) 113年6月行政區**人口統計**_村里_**{縣市}**

| | 縣市代碼 | 縣市名稱 | 鄉鎮市區代碼 | 鄉鎮市區名稱 | 村里代碼 | 村里名稱 | 戶數 | 人口數 | 男性人口數 | 女性人口數 | 資料時間 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-001 | 文化里 | 714 | 1277 | 553 | 724 | 113Y06M |
| **1** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-002 | 文明里 | 800 | 1763 | 823 | 940 | 113Y06M |
| **2** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-003 | 武陵里 | 1047 | 1942 | 869 | 1073 | 113Y06M |
| **3** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-004 | 大林里 | 3334 | 9108 | 4421 | 4687 | 113Y06M |
| **4** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-005 | 忠義里 | 1728 | 4276 | 2022 | 2254 | 113Y06M |

# DATA PREPROCESSING (CLEANING)

## 2) 113年6月行政區**人口指標**_村里_**{縣市}**

| | 縣市代碼 | 縣市名稱 | 鄉鎮市區代碼 | 鄉鎮市區名稱 | 村里代碼 | 村里名稱 | 性比例 | 戶量 | 人口密度 | 扶養比 | 扶幼比 | 扶老比 | 老化指數 | 資料時間 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-001 | 文化里 | 76.38 | 1.79 | 10807.68 | 44.620612 | 17.780294 | 26.840317 | 150.955413 | 113Y06M |
| **1** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-002 | 文明里 | 87.55 | 2.20 | 24856.94 | 56.571936 | 17.584369 | 38.987567 | 221.717171 | 113Y06M |
| **2** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-003 | 武陵里 | 80.99 | 1.85 | 12098.55 | 47.344461 | 13.808801 | 33.535660 | 242.857142 | 113Y06M |
| **3** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-004 | 大林里 | 94.32 | 2.73 | 15361.96 | 44.182365 | 27.623872 | 16.558493 | 59.942693 | 113Y06M |
| **4** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-005 | 忠義里 | 89.71 | 2.47 | 12866.23 | 37.271268 | 17.720706 | 19.550562 | 110.326087 | 113Y06M |

# DATA PREPROCESSING (CLEANING)

## 3) 113年6月行政區**五歲年齡組**性別人口統計_村里_**{縣市}**

| | 縣市代碼 | 縣市名稱 | 鄉鎮市區代碼 | 鄉鎮市區名稱 | 村里代碼 | 村里名稱 | 0-4歲人口數 | 0-4歲男性人口數 | 0-4歲女性人口數 | 5-9歲人口數 | ... | 90-94歲人口數 | 90-94歲男性人口數 | 90-94歲女性人口數 | 95-99歲人口數 | 95-99歲男性人口數 | 95-99歲女性人口數 | 100歲以上人口數 | 100歲以上男性人口數 | 100歲以上女性人口數 | 資料時間 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-001 | 文化里 | 50 | 26 | 24 | 57 | ... | 6 | 2 | 4 | 3 | 2 | 1 | 1 | 1 | 0 | 113Y06M |
| **1** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-002 | 文明里 | 65 | 34 | 31 | 69 | ... | 16 | 6 | 10 | 4 | 1 | 3 | 0 | 0 | 0 | 113Y06M |
| **2** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-003 | 武陵里 | 53 | 25 | 28 | 67 | ... | 14 | 5 | 9 | 6 | 2 | 4 | 1 | 0 | 1 | 113Y06M |
| **3** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-004 | 大林里 | 429 | 212 | 217 | 743 | ... | 10 | 4 | 6 | 3 | 1 | 2 | 0 | 0 | 0 | 113Y06M |
| **4** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-005 | 忠義里 | 128 | 59 | 69 | 213 | ... | 6 | 2 | 4 | 5 | 1 | 4 | 0 | 0 | 0 | 113Y06M |

# DATA PREPROCESSING (CLEANING)

## 4) 112年行政區15歲以上人口教育程度統計_村里_{縣市}

| | 縣市代碼 | 縣市名稱 | 鄉鎮市區代碼 | 鄉鎮市區名稱 | 村里代碼 | 村里名稱 | 博士人口數 | 碩士人口數 | 大學院校人口數 | 專科人口數 | 高中職人口數 | 國中初職人口數 | 小學人口數 | 自修人口數 | 不識字人口數 | 資料時間 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-001 | 文化里 | 18 | 192 | 434 | 147 | 217 | 60 | 39 | 3 | 1 | 112Y |
| **1** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-002 | 文明里 | 10 | 174 | 592 | 233 | 354 | 117 | 79 | 2 | 3 | 112Y |
| **2** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-003 | 武陵里 | 20 | 130 | 582 | 235 | 473 | 175 | 106 | 1 | 5 | 112Y |
| **3** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-004 | 大林里 | 40 | 629 | 2547 | 858 | 2070 | 652 | 498 | 8 | 32 | 112Y |
| **4** | 68000 | 桃園市 | 68000010 | 桃園區 | 68000010-005 | 忠義里 | 15 | 284 | 1298 | 489 | 1069 | 349 | 203 | 5 | 10 | 112Y |

# DATA PREPROCESSING (CLEANING)

5) 110年度**綜稅綜合所得總額**各縣市鄉鎮村里統計分析表-**{縣市}**

| | 縣市別 | 村里 | 納稅單位(戶) | 綜合所得總額 | 平均數 | 中位數 | 第一分位數 | 第三分位數 | 標準差 | 變異係數 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 桃園市桃園區 | 文化里 | 464 | 532454 | 1148 | 628 | 242 | 1431 | 1604.19 | 139.80 |
| 1 | 桃園市桃園區 | 武陵里 | 584 | 446796 | 765 | 490 | 237 | 908 | 1108.21 | 144.85 |
| 2 | 桃園市桃園區 | 文昌里 | 273 | 187910 | 688 | 468 | 201 | 917 | 721.26 | 104.79 |
| 3 | 桃園市桃園區 | 長美里 | 341 | 231145 | 678 | 426 | 217 | 841 | 815.35 | 120.29 |
| 4 | 桃園市桃園區 | 永興里 | 658 | 499064 | 758 | 462 | 218 | 887 | 1486.14 | 195.94 |

# DATA PREPROCESSING (CLEANING)

## 6) 總統-A05-4-候選人得票數一覽表-各投開票所({縣市})

第16任總統副總統選舉候選人在桃園市各投開票所得票數一覽表

| 鄉(鎮、市、區)別 | 村里別 | 投開票所別 | 各組候選人得票情形 (1) 柯文哲 吳欣盈 | (2) 賴清德 蕭美琴 | (3) 侯友宜 趙少康 | 有效票數A A=1+2+...+N | 無效票數B | 投票數C C=A+B | 已領未投票數 D D=E-C | 發出票數E E=C+D | 用餘票數F | 選舉人數G G=E+F | 投票率H H=C÷G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 總　計 | | | 413,528 | 476,441 | 460,823 | 1,350,792 | 8,897 | 1,359,689 | 41 | 1,359,730 | 522,862 | 1,882,592 | 72.22 |
| 蘆竹區 | | | 29,127 | 38,760 | 29,192 | 97,079 | 608 | 97,687 | 1 | 97,688 | 36,881 | 134,569 | 72.59 |
| | 新興里 | 0001 | 297 | 583 | 260 | 1,140 | 4 | 1,144 | 0 | 1,144 | 369 | 1,513 | 75.61 |
| | 新興里 | 0002 | 275 | 494 | 226 | 995 | 3 | 998 | 0 | 998 | 356 | 1,354 | 73.71 |
| | 中福里 | 0003 | 314 | 503 | 221 | 1,038 | 7 | 1,045 | 0 | 1,045 | 419 | 1,464 | 71.38 |
| | 中福里 | 0004 | 293 | 494 | 213 | 1,000 | 2 | 1,002 | 0 | 1,002 | 342 | 1,344 | 74.55 |
| | 中福里 | 0005 | 247 | 420 | 223 | 890 | 10 | 900 | 0 | 900 | 319 | 1,219 | 73.83 |
| | 上興里 | 0006 | 376 | 543 | 413 | 1,332 | 11 | 1,343 | 0 | 1,343 | 438 | 1,781 | 75.41 |
| | 上興里 | 0007 | 444 | 453 | 364 | 1,261 | 3 | 1,264 | 0 | 1,264 | 435 | 1,699 | 74.40 |
| | 中興里 | 0008 | 351 | 427 | 350 | 1,128 | 5 | 1,133 | 0 | 1,133 | 351 | 1,484 | 76.35 |

# DATA PREPROCESSING (CLEANING)

- **目標：將五個資料集同一個村里的不同資料集進行合併，變成一張大的資料表。**
- 在觀察完剛剛那五個資料集，有沒有發現到有四個資料集都有「縣市代碼、縣市名稱、鄉鎮市區代碼、鄉鎮市區名稱、村里代碼、村里名稱」的資訊。換句話說，就可以依照這些資訊把不同的資料集進行對位，做橫向合併資料表。
- **問題：如果只是要用村里合併，為什麼要需要縣市與鄉鎮市區的資訊呢?**（…自己想想看）

# DATA PREPROCESSING (CLEANING)

- 但是除了合併問題外，那所得收入資料的欄位與其他不同該怎麼做呢?

| | 縣市別 | 村里 | 納稅單位(戶) | 綜合所得總額 | 平均數 | 中位數 | 第一分位數 | 第三分位數 | 標準差 | 變異係數 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 桃園市桃園區 | 文化里 | 464 | 532454 | 1148 | 628 | 242 | 1431 | 1604.19 | 139.80 |
| 1 | 桃園市桃園區 | 武陵里 | 584 | 446796 | 765 | 490 | 237 | 908 | 1108.21 | 144.85 |
| 2 | 桃園市桃園區 | 文昌里 | 273 | 187910 | 688 | 468 | 201 | 917 | 721.26 | 104.79 |
| 3 | 桃園市桃園區 | 長美里 | 341 | 231145 | 678 | 426 | 217 | 841 | 815.35 | 120.29 |
| 4 | 桃園市桃園區 | 永興里 | 658 | 499064 | 758 | 462 | 218 | 887 | 1486.14 | 195.94 |

# DATA PREPROCESSING (CLEANING)

- 我們可以用Excel的「資料剖析」，將縣市名稱與鄉鎮市區名稱拆開來，就可以變成下方的樣子：

| | 縣市名稱 | 鄉鎮市區名稱 | 村里名稱 | 納稅單位(戶) | 綜合所得總額 | 平均數 | 中位數 | 第一分位數 | 第三分位數 | 標準差 | 變異係數 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 桃園市 | 桃園區 | 文化里 | 464 | 532454 | 1148 | 628 | 242 | 1431 | 1604.19 | 139.80 |
| 1 | 桃園市 | 桃園區 | 武陵里 | 584 | 446796 | 765 | 490 | 237 | 908 | 1108.21 | 144.85 |
| 2 | 桃園市 | 桃園區 | 文昌里 | 273 | 187910 | 688 | 468 | 201 | 917 | 721.26 | 104.79 |
| 3 | 桃園市 | 桃園區 | 長美里 | 341 | 231145 | 678 | 426 | 217 | 841 | 815.35 | 120.29 |
| 4 | 桃園市 | 桃園區 | 永興里 | 658 | 499064 | 758 | 462 | 218 | 887 | 1486.14 | 195.94 |

# DATA PREPROCESSING (CLEANING)

## (1) 資料合併會遇到的問題:

- **欄位名稱與想像中的不同:**
- 收入所得資料中的縣市標頭: **\ufeff**縣市別
- 投票資料中的鄉鎮市區資訊: **\u3000**蘆竹區

  * 就要靠自己手動刪除

# DATA PREPROCESSING (CLEANING)

## (2) 資料合併會遇到的問題:

- 投票資料的標頭有太多儲存格合併，需要進行拆分。
- 行政區別的地方要把總計、各區總計(意指蘆竹區那列的數值就是該區的加總)刪掉，因為每一筆的資料間必須獨立，不能有相關或是階層式的關係。

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 第16任總統副總統選舉候選人在桃園市各村(里)得票數一覽表 | | | | | | | | | | | | |
| 2 | | | 各組候選人得票情形 | | | | | | | | | | |
| 3 | | | (1) | (2) | (3) | | | | 已領未投票數 | | | | |
| 4 | | | 柯文哲 | 賴清德 | 侯友宜 | 有效票數A | | 投票數C | D | 發出票數E | | 選舉人數G | 投票率H |
| 5 | 行政區別 | 村里別 | 吳欣盈 | 蕭美琴 | 趙少康 | A=1+2+...+N | 無效票數B | C=A+B | D=E-C | E=C+D | 用餘票數F | G=E+F | H=C÷G |
| 6 | 總　計 | | 413,528 | 476,441 | 460,823 | 1,350,792 | 8,897 | 1,359,689 | 41 | 1,359,730 | 522,862 | 1,882,592 | 72.22 |
| 7 | 蘆竹區 | | 29,127 | 38,760 | 29,192 | 97,079 | 608 | 97,687 | 1 | 97,688 | 36,881 | 134,569 | 72.59 |
| 8 | | 新興里 | 572 | 1,077 | 486 | 2,135 | 7 | 2,142 | 0 | 2,142 | 725 | 2,867 | 74.71 |
| 9 | | 中福里 | 854 | 1,417 | 657 | 2,928 | 19 | 2,947 | 0 | 2,947 | 1,080 | 4,027 | 73.18 |
| 10 | | 上興里 | 820 | 996 | 777 | 2,593 | 14 | 2,607 | 0 | 2,607 | 873 | 3,480 | 74.91 |

# DATA PREPROCESSING (CLEANING)

## (2) 資料合併會遇到的問題:

- 處理完後應該會長這樣

| | 鄉鎮市區名稱 | 村里名稱 | 民眾黨 | 民進黨 | 國民黨 | 有效票數 | 無效票數 | 投票數 | 已領未投票數 | 發出票數 | 用餘票數 | 選舉人數 | 投票率 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 蘆竹區 | 新興里 | 572 | 1077 | 486 | 2135 | 7 | 2142 | 0 | 2142 | 725 | 2867 | 74.71 |
| **1** | 蘆竹區 | 中福里 | 854 | 1417 | 657 | 2928 | 19 | 2947 | 0 | 2947 | 1080 | 4027 | 73.18 |
| **2** | 蘆竹區 | 上興里 | 820 | 996 | 777 | 2593 | 14 | 2607 | 0 | 2607 | 873 | 3480 | 74.91 |
| **3** | 蘆竹區 | 中興里 | 1333 | 1397 | 1084 | 3814 | 20 | 3834 | 0 | 3834 | 1381 | 5215 | 73.52 |
| **4** | 蘆竹區 | 新莊里 | 1157 | 1640 | 961 | 3758 | 27 | 3785 | 0 | 3785 | 1296 | 5081 | 74.49 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **511** | 復興區 | 長興里 | 107 | 83 | 192 | 382 | 1 | 383 | 0 | 383 | 275 | 658 | 58.21 |
| **512** | 復興區 | 奎輝里 | 142 | 58 | 268 | 468 | 4 | 472 | 0 | 472 | 261 | 733 | 64.39 |
| **513** | 復興區 | 高義里 | 135 | 60 | 224 | 419 | 3 | 422 | 0 | 422 | 372 | 794 | 53.15 |
| **514** | 復興區 | 三光里 | 114 | 48 | 238 | 400 | 5 | 405 | 0 | 405 | 309 | 714 | 56.72 |
| **515** | 復興區 | 華陵里 | 182 | 183 | 364 | 729 | 4 | 733 | 0 | 733 | 552 | 1285 | 57.04 |

# DATA PREPROCESSING (CLEANING)

## [關於資料橫向合併的大小事]

- 在這次的資料集中，也會遇到上述的一些問題:
- 基本上，在資料的整併過程中，你們可以利用Excel內建的函數(vlookup)進行點對點(村里對村里)的資料合併，這部分請各位自行查閱網路資料或是以下的連結進行實作:

  VLOOKUP, 函數

  Excel, VLOOKUP, 函數教學：按列搜尋表格，自動填入資料

# DATA PREPROCESSING (CLEANING)

- 在資料都成功合併完之後，我們一般的習慣會把Y放在左邊，X的變數放在右邊，以符合「**Y=aX+b**」的概念。
- 但為了可讀性(readibility)，我們會把村里相關資訊放在最左邊；因此順序就會變成為：**村里屬性、Y變數{某黨的候選人得票數}、X變數{教育程度、年齡結構、人口指標、人口統計、收入所得}**整理起來。
- 詳細範例可參閱下一頁的截圖。

# DATA PREPROCESSING (CLEANING)

- 合併完的資料應該會長類似這樣：

| | 鄉鎮市區名稱 | 村里名稱 | 民眾黨 | 民進黨 | 國民黨 | 有效票數 | 無效票數 | 投票數 | 已領未投票數 | 發出票數 | ... | 男性人口數 | 女性人口數 | 納稅單位(戶) | 綜合所得總額 | 平均數 | 中位數 | 第一分位數 | 第三分位數 | 標準差 | 變異係數 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 蘆竹區 | 新興里 | 572 | 1077 | 486 | 2135 | 7 | 2142 | 0 | 2142 | ... | 1746.0 | 1602.0 | 1012.0 | 732438.0 | 724.0 | 436.0 | 224.0 | 814.0 | 1378.35 | 190.45 |
| **1** | 蘆竹區 | 中福里 | 854 | 1417 | 657 | 2928 | 19 | 2947 | 0 | 2947 | ... | 2514.0 | 2347.0 | 1440.0 | 968263.0 | 672.0 | 472.0 | 240.0 | 823.0 | 822.34 | 122.30 |
| **2** | 蘆竹區 | 上興里 | 820 | 996 | 777 | 2593 | 14 | 2607 | 0 | 2607 | ... | 2178.0 | 2241.0 | 1351.0 | 1202702.0 | 890.0 | 563.0 | 278.0 | 1125.0 | 1020.80 | 114.67 |
| **3** | 蘆竹區 | 中興里 | 1333 | 1397 | 1084 | 3814 | 20 | 3834 | 0 | 3834 | ... | 3336.0 | 3664.0 | 2210.0 | 1899205.0 | 859.0 | 614.0 | 318.0 | 1070.0 | 890.08 | 103.57 |
| **4** | 蘆竹區 | 新莊里 | 1157 | 1640 | 961 | 3758 | 27 | 3785 | 0 | 3785 | ... | 3142.0 | 3071.0 | 1834.0 | 1327289.0 | 724.0 | 480.0 | 240.0 | 899.0 | 936.39 | 129.39 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **509** | 復興區 | 長興里 | 107 | 83 | 192 | 382 | 1 | 383 | 0 | 383 | ... | 444.0 | 353.0 | 81.0 | 47749.0 | 589.0 | 458.0 | 255.0 | 642.0 | 563.62 | 95.61 |
| **510** | 復興區 | 奎輝里 | 142 | 58 | 268 | 468 | 4 | 472 | 0 | 472 | ... | 527.0 | 482.0 | 133.0 | 69037.0 | 519.0 | 447.0 | 209.0 | 716.0 | 399.39 | 76.94 |
| **511** | 復興區 | 高義里 | 135 | 60 | 224 | 419 | 3 | 422 | 0 | 422 | ... | 573.0 | 512.0 | 120.0 | 53276.0 | 444.0 | 393.0 | 245.0 | 609.0 | 309.20 | 69.64 |
| **512** | 復興區 | 三光里 | 114 | 48 | 238 | 400 | 5 | 405 | 0 | 405 | ... | 502.0 | 448.0 | 106.0 | 47175.0 | 445.0 | 372.0 | 158.0 | 597.0 | 385.10 | 86.53 |
| **513** | 復興區 | 華陵里 | 182 | 183 | 364 | 729 | 4 | 733 | 0 | 733 | ... | 885.0 | 732.0 | 187.0 | 91839.0 | 491.0 | 372.0 | 208.0 | 580.0 | 493.06 | 100.39 |

# DATA PREPROCESSING (CLEANING)

**[問題]** 看到這邊，如果你是桃園人，應該會知道桃園有516個里，但這裡只有514個里?

- 因為有兩個里含有缺值，所以這兩個里的資料整筆刪除。不論是統計模型還是機器學習的模型，都不能接受有遺漏值資料集。

| | 鄉鎮市區名稱 | 村里名稱 | 民眾黨 | 民進黨 | 國民黨 | 有效票數 | 無效票數 | 投票數 | 已領未投票數 | 發出票數 | ... | 男性人口數 | 女性人口數 | 納稅單位(戶) | 綜合所得總額 | 平均數 | 中位數 | 第一分位數 | 第三分位數 | 標準差 | 變異係數 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **164** | 大園區 | 菓林里 | 1375 | 1854 | 1311 | 4540 | 38 | 4578 | 0 | 4578 | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **212** | 新屋區 | 槺榔里 | 178 | 323 | 207 | 708 | 10 | 718 | 0 | 718 | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

# DATA PREPROCESSING (CLEANING)

- 理論上會有108個欄位:

**村里屬性** 縣市名稱, 縣市代碼, 鄉鎮市區名稱, 鄉鎮市區代碼, 村里名稱, 村里代碼,

**投票資料** 民眾黨, 民進黨, 國民黨, 有效票數, 無效票數, 投票數, 已領未投票數, 發出票數, 用餘票數, 選舉人數, 投票率,

**教育程度** 博士人口數, 碩士人口數, 大學院校人口數, 專科人口數, 高中職人口數, 國中初職人口數, 小學人口數, 自修人口數, 不識字人口數,

**年齡結構** 0-4歲人口數, 0-4歲男性人口數, 0-4歲女性人口數, 5-9歲人口數, 5-9歲男性人口數, 5-9歲女性人口數, 10-14歲人口數, 10-14歲男性人口數, 10-14歲女性人口數, 15-19歲人口數, 15-19歲男性人口數, 15-19歲女性人口數, 20-24歲人口數, 20-24歲男性人口數, 20-24歲女性人口數, 25-29歲人口數, 25-29歲男性人口數, 25-29歲女性人口數, 30-34歲人口數, 30-34歲男性人口數, 30-34歲女性人口數, 35-39歲人口數, 35-39歲男性人口數, 35-39歲女性人口數, 40-44歲人口數, 40-44歲男性人口數, 40-44歲女性人口數, 45-49歲人口數, 45-49歲男性人口數, 45-49歲女性人口數, 50-54歲人口數, 50-54歲男性人口數, 50-54歲女性人口數, 55-59歲人口數, 55-59歲男性人口數, 55-59歲女性人口數, 60-64歲人口數, 60-64歲男性人口數, 60-64歲女性人口數, 65-69歲人口數, 65-69歲男性人口數, 65-69歲女性人口數, 70-74歲人口數, 70-74歲男性人口數, 70-74歲女性人口數, 75-79歲人口數, 75-79歲男性人口數, 75-79歲女性人口數, 80-84歲人口數, 80-84歲男性人口數, 80-84歲女性人口數, 85-89歲人口數, 85-89歲男性人口數, 85-89歲女性人口數, 90-94歲人口數, 90-94歲男性人口數, 90-94歲女性人口數, 95-99歲人口數, 95-99歲男性人口數, 95-99歲女性人口數, 100歲以上人口數, 100歲以上男性人口數, 100歲以上女性人口數,

**人口指標** 性比例, 戶量, 人口密度, 扶養比, 扶幼比, 扶老比, 老化指數,

**人口統計** 戶數, 人口數, 男性人口數, 女性人口數,

**收入所得** 納稅單位(戶), 綜合所得總額, 平均數, 中位數, 第一分位數, 第三分位數, 標準差, 變異係數

# DATA INTEGRATION

- 在整理後資料之後，就要該開始思考資料是不是要合併，因為很多欄位的資料可能太細，或是欄位之間存在不獨立(或階層式)的關係，這樣的資料無法訓練模型。
- **譬如說：**年齡結構資料，每五歲一組對於模型訓練沒有太大的幫助，同時低於18歲的人沒有投票權，因此放這些資料也不能解釋投票的結果➔所以思考這些X是否具有解釋力？是否有代表性？如果沒有，是否要刪除資料或是整併資料？像是區分為青年、壯年、老年人口。

# DATA INTEGRATION

- 或是說，你想要做的主題是哪個黨得票數最多，那就要先將每個里的各政黨票數取最大值，看是哪一個政黨為最高票:

| | 最高票黨_三黨 | 最高票黨_三黨_索引值 | 最高票黨_兩黨 | 最高票黨_兩黨_索引值 |
|---|---|---|---|---|
| 20 | 民進黨 | 1 | 民進黨 | 0 |
| 21 | 民進黨 | 1 | 民進黨 | 0 |
| 22 | 民進黨 | 1 | 民進黨 | 0 |
| 23 | 民進黨 | 1 | 民進黨 | 0 |
| 24 | 民進黨 | 1 | 民進黨 | 0 |
| 25 | 民進黨 | 1 | 民進黨 | 0 |
| 26 | 國民黨 | 2 | 國民黨 | 1 |
| 27 | 國民黨 | 2 | 國民黨 | 1 |
| 28 | 民眾黨 | 0 | 民進黨 | 0 |
| 29 | 民進黨 | 1 | 民進黨 | 0 |
| 30 | 國民黨 | 2 | 國民黨 | 1 |

**三黨的話就是多元分類問題：**
通常機器學習在多元分類的模型表現效果都會比較差；一般來說，我們的X很難可以同時反映不同類別的特性。

**二黨問題屬於二元分類：**
通常會得到比較好的模型預測結果。

# DATA INTEGRATION

- **[重複性]** 還有如果我們放了年齡結構男女人口數資料，那麼該里的總人口數、男性、女性都不應該被列入X中。
- **[太過細分]** 一樣的問題在教育程度也會出現，建議可以參考民調中心區分的方式，將你的資料進行整併。
- **[欄位的抉擇]** 同時收入所得的部分，遇到的問題會是同樣的資料有許多不同的統計量，你可以用你學過的統計學(這邊指的是高中程度)，就可以知道該怎麼做！答案不是只有一種，但是要靠自己思考，建立屬於你自己的資料集。

* 為了培養各位的獨立思考能力，資料整併與欄位選擇的部分，我並沒有提供相關的示範資料。

# DATA INTEGRATION

- 我用一些簡單的邏輯進行篩選，但是每個人可以有自己的想法去創立出自己的資料集，所以以下僅供參考：

Y-numeric　Y-multiclass　Y-binary

| | 縣市名稱 | 縣市代碼 | 鄉鎮市區名稱 | 鄉鎮市區代碼 | 村里名稱 | 村里代碼 | 民眾黨 | 民進黨 | 國民黨 | 最高票黨_三黨_索引值 | 最高票黨_兩黨_索引值 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 桃園市 | 68000.0 | 蘆竹區 | 68000050.0 | 新興里 | 68000050-001 | 572 | 1077 | 486 | 1 | 0 |
| 1 | 桃園市 | 68000.0 | 蘆竹區 | 68000050.0 | 中福里 | 68000050-002 | 854 | 1417 | 657 | 1 | 0 |
| 2 | 桃園市 | 68000.0 | 蘆竹區 | 68000050.0 | 上興里 | 68000050-037 | 820 | 996 | 777 | 1 | 0 |
| 3 | 桃園市 | 68000.0 | 蘆竹區 | 68000050.0 | 中興里 | 68000050-036 | 1333 | 1397 | 1084 | 1 | 0 |
| 4 | 桃園市 | 68000.0 | 蘆竹區 | 68000050.0 | 新莊里 | 68000050-003 | 1157 | 1640 | 961 | 1 | 0 |

X-dataset

| | 投票率 | 性比例 | 老化指數 | 中位數 | 變異係數 | 青壯年_女性人口數 | 青壯年_男性人口數 | 中年_男性人口數 | 中年_女性人口數 | 老年_男性人口數 | 老年_女性人口數 | 高學歷人口數 | 非高學歷人口數 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 74.71 | 108.99 | 187.887323 | 436.0 | 190.45 | 538.0 | 645.0 | 540.0 | 463.0 | 311.0 | 356.0 | 1320.0 | 1697.0 |
| 1 | 73.18 | 107.12 | 114.218750 | 472.0 | 122.30 | 880.0 | 960.0 | 719.0 | 662.0 | 363.0 | 368.0 | 1918.0 | 2355.0 |
| 2 | 74.91 | 97.19 | 78.443114 | 563.0 | 114.67 | 827.0 | 851.0 | 627.0 | 672.0 | 231.0 | 293.0 | 2085.0 | 1653.0 |
| 3 | 73.52 | 91.05 | 41.724403 | 614.0 | 103.57 | 1414.0 | 1324.0 | 912.0 | 1087.0 | 239.0 | 303.0 | 3194.0 | 2449.0 |
| 4 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 1134.0 | 1243.0 | [illegible] | 942.0 | 449.0 | 434.0 | [illegible] | 2899.0 |

投票　人口指標　所得收入　年齡結構　教育程度

# SUMMARY

- 這一份簡報，就是以2024年總統大選作為一個範例資料，介紹我們在資料前處理跟分析上，應該需要注意哪些小細節，如何將資料夾正確無誤的方式整併起來。
- 政府公開資料有非常多的問題，唯有你自己去用過才會知道有什麼樣的疑慮、特性或是代表性。

WE BETTER

YOUR LIFE

# The End

THANK YOU FOR YOUR ATTENTION!