那些超商給我的啟示

第11組

r08921a03 洪國喨、b06505022 袁肇謙、r08942088 李政旻、b06901188 李宗倫

**動機**

便利超商在台灣的密集度居世界第二[1]，可見其與台灣人的生活息息相關。縱然密度如此之高，其在地理位置上的分布卻極度不均。台灣主要有四家便利超商：7-11、全家、萊爾富、OK四家便利超商表面上提供的服務差不多，實際上他們經營的策略、地域的分布，卻都略有不同[2][3]。因此，我們認為可以透過挖掘某一地區"特定品牌分布比例"、"特定品牌家數"、"總超商家數"等和其他地域性資料的關係，觀察到一些有趣的現象。

**主要資料集**

全國4大超商資料集[4]，來源為政府開放資料平台。

欄位屬性 (column attribute)：

公司統一編號、公司名稱、分公司統一編號、分公司名稱、分公司地址、分公司狀態

根據資料集的欄位說明，分公司狀態包含 [01 核准設立]，[02 停業]，[03 撤銷/廢止]，  
但有 4 家超商狀態是 [04]，實際在 google 地圖上搜尋，結果皆為歇業。

**資料清理**

1. 先保留核准設立 (01) 部分的超商來分析，撤銷及廢止部分暫時先不考慮。

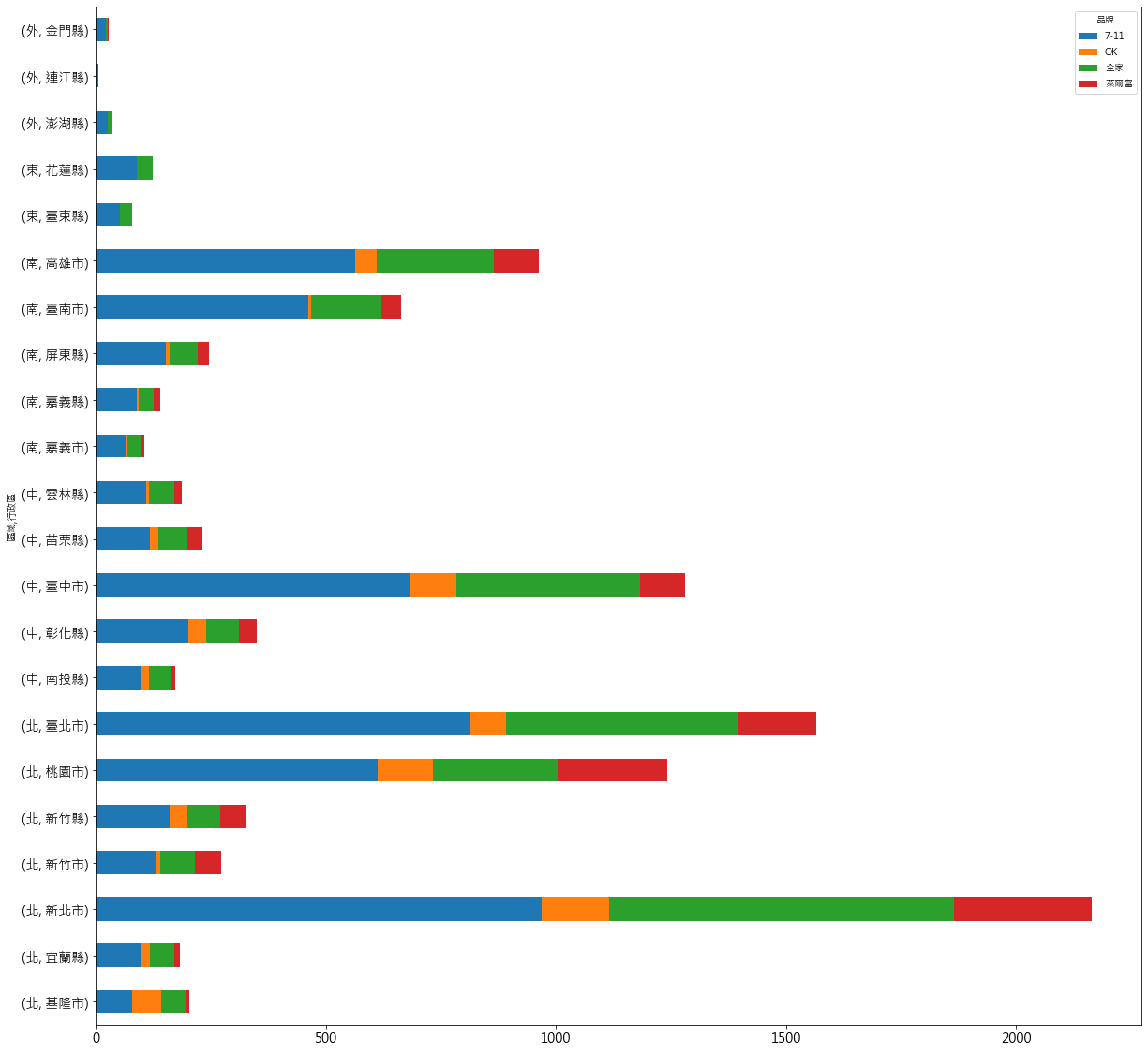
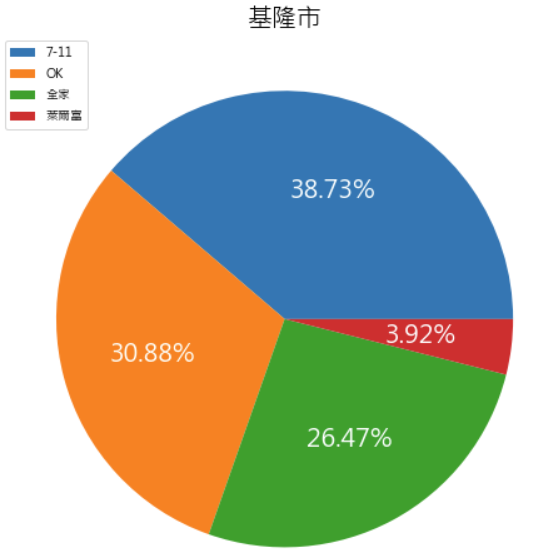
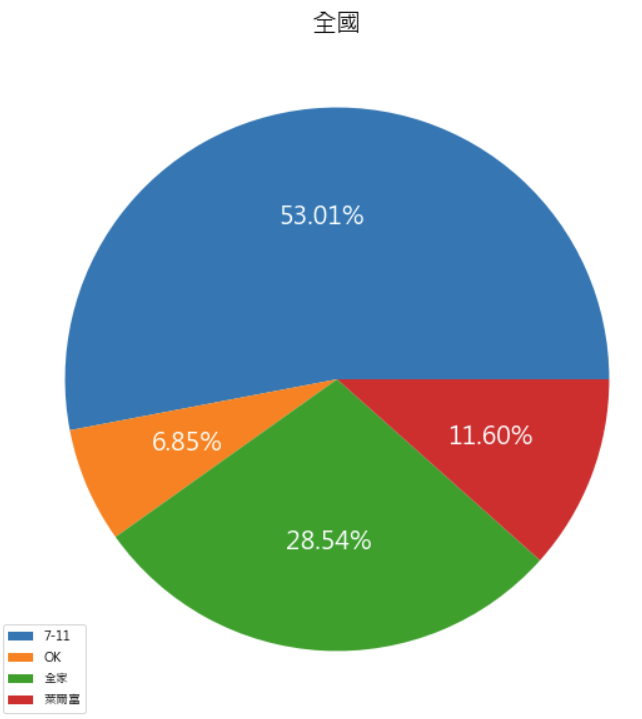
2. 將同義字但字元編碼不同的文字統一，如：'巿' v.s. '市'；'台' v.s. '臺'。

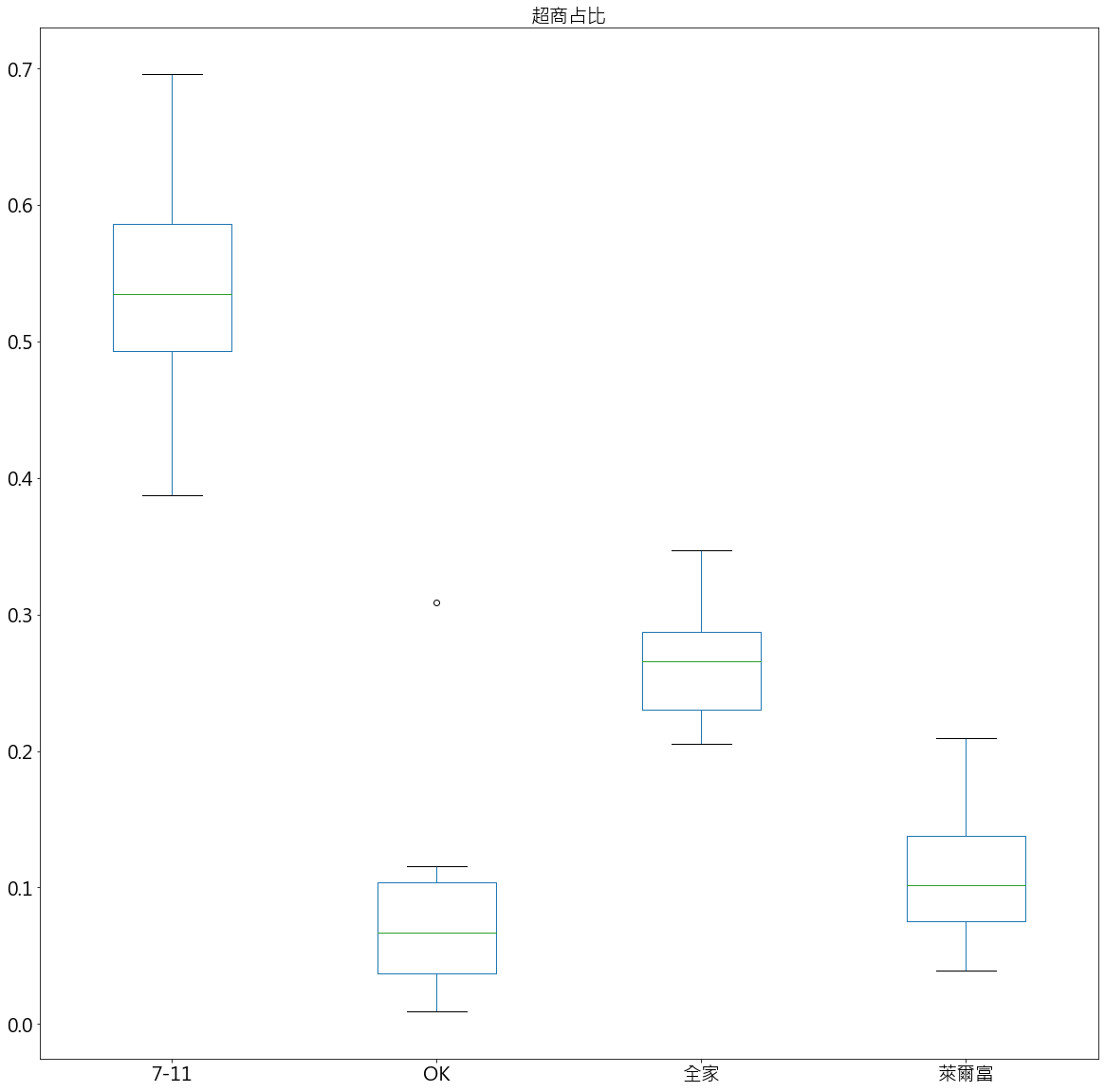
3. 改制後縣市合併，如：高雄縣合併至高雄市。

4. 去除地址重複的店家，共有五筆重複的地址。

5. 建立分區關係 (北台灣、中台灣、南台灣、東台灣、外島)。

**觀察與發現**

1. 將 7-11 店家數與 ibon 官網提供的店面資料進行交叉比對，結果十分接近[5]。
2. 比較各縣市超商數量的長條圖：
3. 各超商品牌全國平均佔比的圓餅圖，以及比例分佈特殊的基隆市：
4. 各超商品牌在各縣市佔比的箱型圖，各家都在一定的區間，但ok在基隆市的佔比特別高：



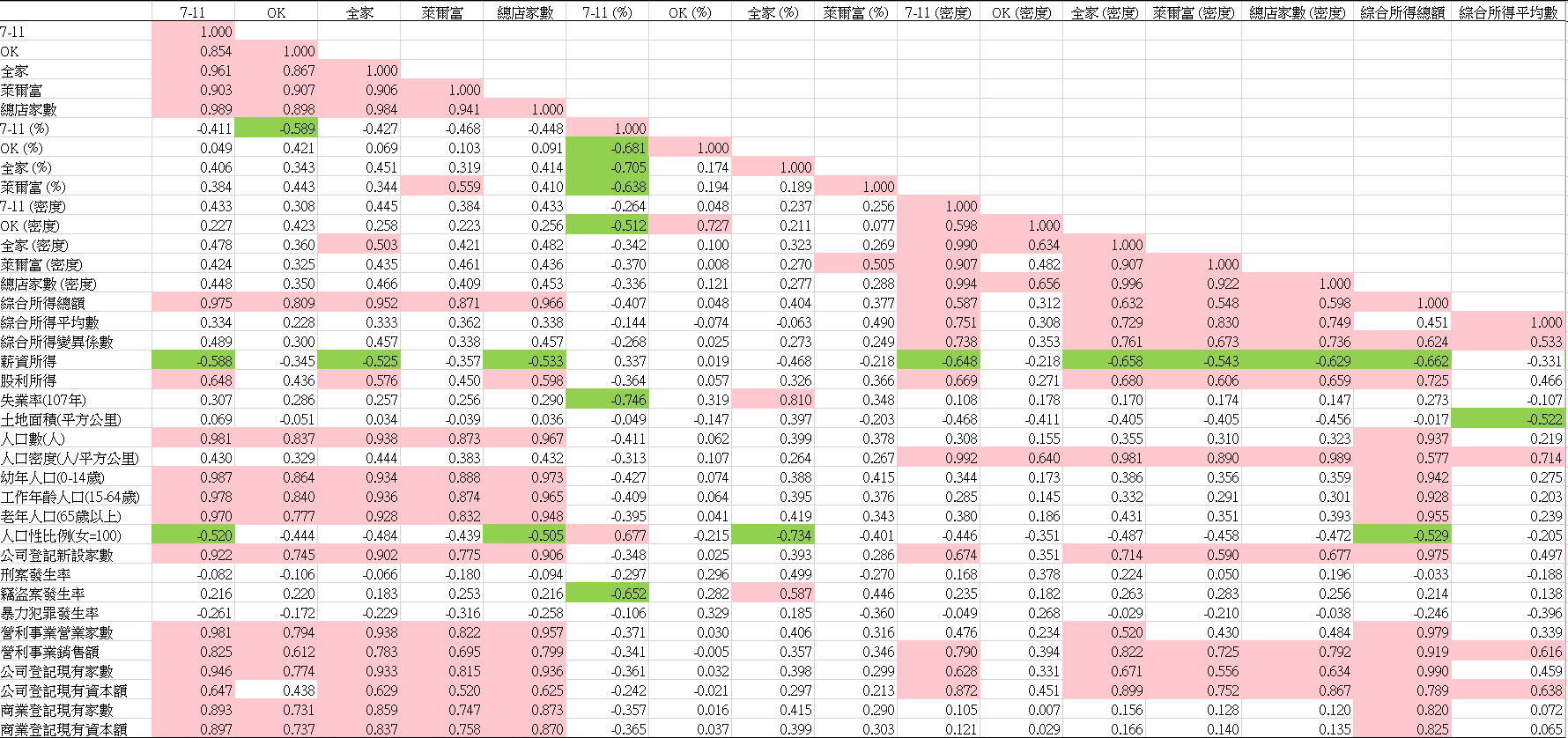
**縣市層級統計資料**

* 土地、人口方面[6][7]  
  土地面積(平方公里)、人口數(人)、人口密度(人/平方公里)、男性人口(人)、女性人口(人)、人口性別比例、 幼年人口(0-14歲)、工作年齡人口(15-64歲)、老年人口(65歲以上)
* 所得方面[8]  
  綜合所得總額、綜合所得平均數、綜合所得變異係數、薪資所得、股利所得
* 工商業方面[6]  
  公司登記新設家數、商業登記現有家數、公司登記現有資本額、營利事業銷售額
* 就業方面[9]  
  失業率
* 犯罪方面[6]  
  刑案發生率、竊盜案發生率、暴力犯罪發生率

**相關矩陣**

將各超商品牌在各縣市的數量、佔比，以及各縣市的統計資料做出相關係數矩陣。其中可以發現某些特性：

1. 超商數量與人口、所得、多數未經標準化的數值，呈高度正相關。
2. 7-11主宰各地超商數量，因此占比會與其他家超商呈負相關。
3. OK 是四大超商中最特別的一個，其從基隆開始發展、擴張[10]，所以在基隆具有非常高的市占率。
4. 從性別比和超商比的關係，可發現7-11占比與男性比例正相關；全家占比則與女性比例正相關。
5. 超商比例與多數地理資料分布相關係數均偏低，可知四大超商之間並沒有顯著的差異。



**資料分析**

Tools：

* numpy
* pandas
* scikit-learn
* pydotplus
* googlemaps：一個 python 套件，會呼叫 Google 地圖提供的 API，用來將地址轉換為 GPS 座標。
* gmplot：一個 python 套件，用以快速繪製 Google 地圖，提供一個類似 matplotlib 的操作介面，方便使用者快速上手，輸出HTML 或 javascript 檔案以將座標或網格資訊呈現在 Google 地圖上，提供多種繪製方法，可快速有效呈現圖資訊息。

Methodology：

* Binary Decision Tree：使用 sklearn package 產生 binary decision tree
* 超商百分比：以三家超商品牌的佔比為三維座標軸的基底，將各縣市的資料分群
* GPS 經緯度座標：KMeans、L2 Norm Spectrum

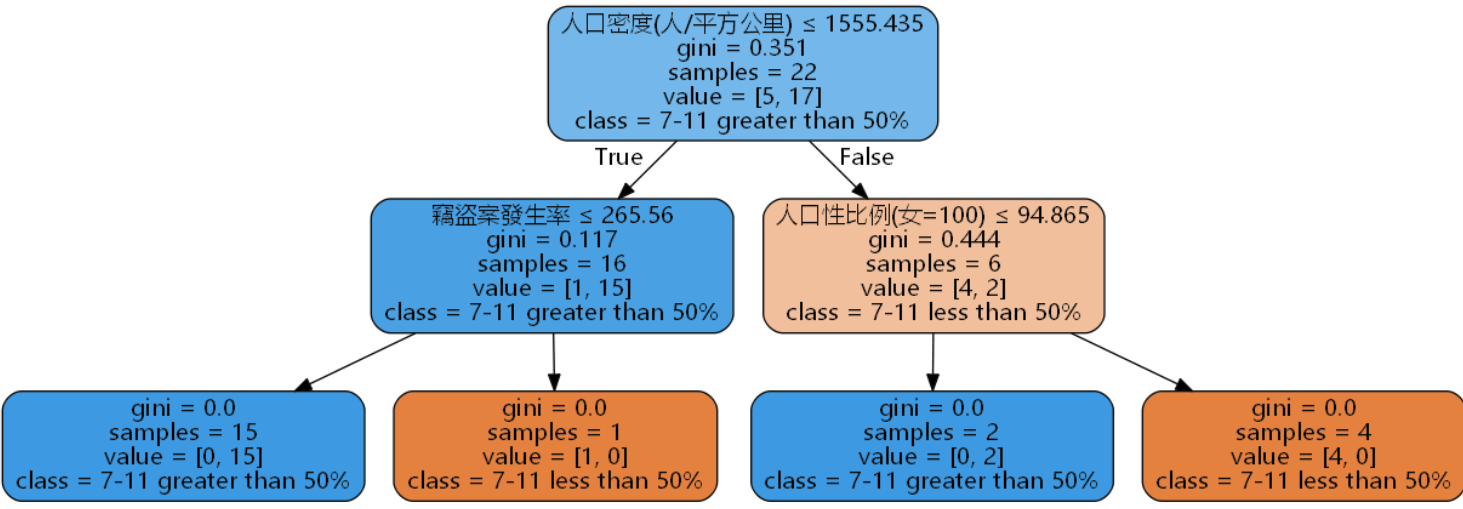
**Binary Decision Tree**

Input：全國4大超商資料集算出7-11、其他縣市層級資料

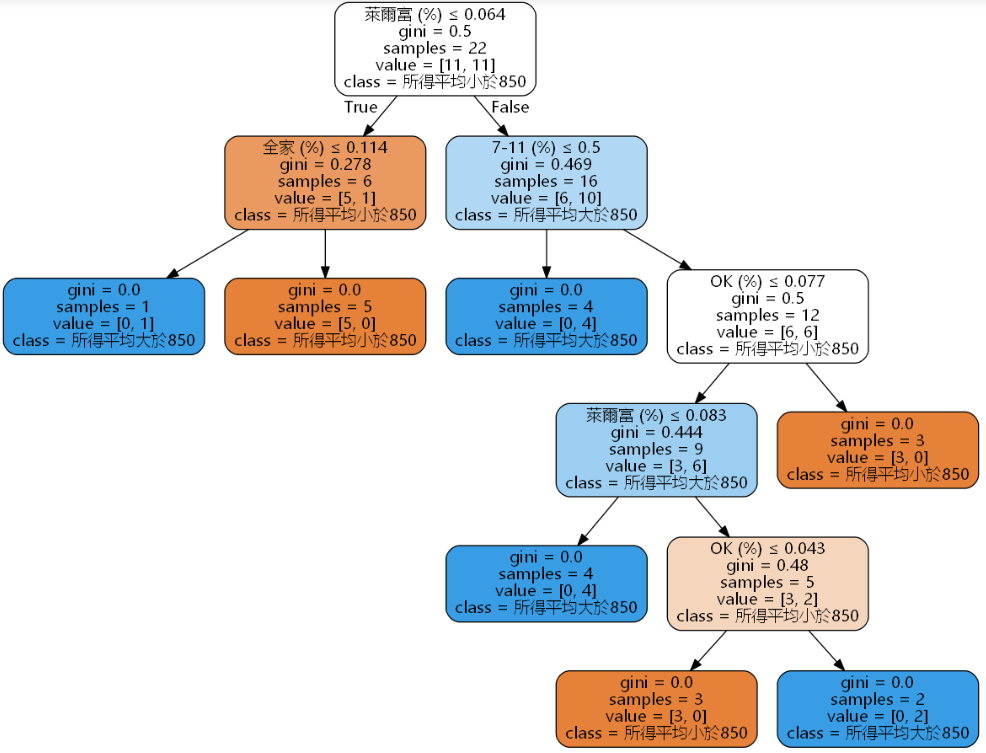
Implementation：將data讀入後，先決定class的分法以及所用的attributes，利用sklearn中tree，計算各種attribute的gini index，取minimum生成decision tree

|  |  |
| --- | --- |
| class | result attribute |
| 7-11 > 50%？  Y = raw.iloc[:, 6] > 0.5 | 人口密度、竊盜案發生率、人口性別比 |
| 全家 > 25%？  Y = raw.iloc[:, 8] > 0.25 | 綜合所得變異係數、幼年人口、公司登記現有家數、綜合所得總額 |
| OK > 10%？  Y = raw.iloc[:, 7] > 0.1 | 暴利犯罪發生率、竊盜案發生率、綜合所得總額、綜合所得變異數、人口密度 |
| 萊爾富 > 8%？  Y = raw.iloc[:, 9] > 0.08 | 營利事業銷售額、竊盜案發生率、刑案發生率、人口密度 |
| 所得平均 > 850K (所得平均median)？  Y = raw.iloc[:, 11] > 850 | 四大超商佔比 |
| 性別比 > 102 (性別比median)？  Y = raw.iloc[:, 24] > 102 |

Finding：

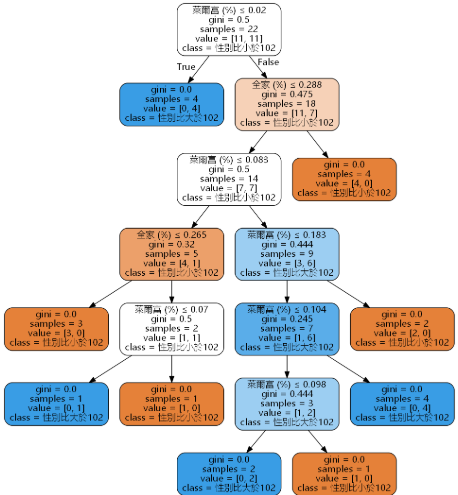


7-11 是否大於 50%



所得平均是否大於 850k

從上面 2 棵 decision trees 有個蠻有意思的發現，各品牌的便利商店設點策略直觀上應該是在人口密度愈密或所得越高的區域，設點越多。不過在第 1 棵 tree，可以看到 7-11 再人口密度較低的地方，反而佔比愈大；第 2 棵 tree 的第二層，可以發現 7-11 在所得較高的地方，反而比例較低，我推測 7-11 的設點策略應該是廣設門市、以量取勝。



性別比是否大於 102

由上圖中的 tree 可發現，影響性別比的 dominate factors 為萊爾富和全家的佔比，這兩個品牌在女性比例相對高 (性別比低) 處，愈傾向於設點。

**Naive Bayes Classification**

此種分類方式不須建構，只在資料輸入後進行分類，所以如果要拿來分析已經存在的資料集，就必須假設某些條件，接著驗證這個假設是否成立，比如人口數多於幾百萬人的地區同時犯罪率在某一個程度下，全家的占比會高於25%。不過這樣假設條件的方式非常沒有依據，不如我們直接使用decision tree，透過一定標準去切割這些資料。

**K-Means**

Input：全國4大超商資料集算出7-11、全家及萊爾富在各縣市佔比

Implementation：由資料集計算出各縣市超商品牌佔比後，首先以各區域作為分群標準， 接著以Sklearn的kmeans方法進行分群，將超商佔比相近的縣市分類，再將3d圖畫出。

Finding：

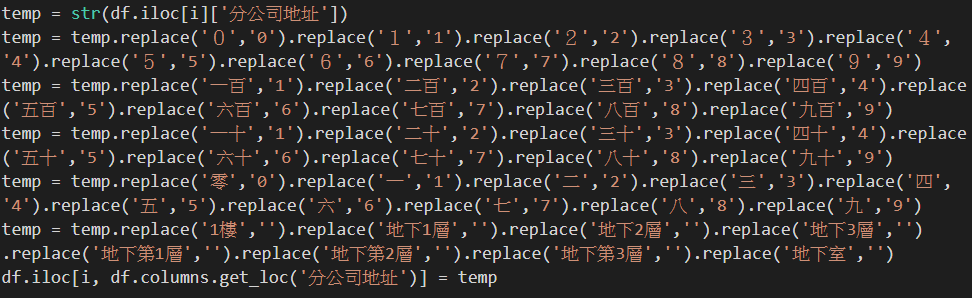
|  |  |
| --- | --- |
|  |  |
| 上方的左邊三張圖是 三大超商在各縣市的佔比散佈圖   * 北：藍 * 中：紅 * 南：灰 * 東：綠 * 外島：棕 | 上方的右邊三張圖是三大超商在各縣市的佔比kmeans分群圖  比較左方散佈圖及右方kmeans分群後，可以發現同個區域的點也會被分在同個kmeans的分群內，故可知各分區（北、中、南等）內縣市三大超商佔比有相關性，推斷超商會依據不同地理區域而有不同設點策略。 |

**GPS 經緯度座標**

Input：全國4大超商資料集

Implementation：

1. 地址轉換



首先，除了對地址做基礎的前處理 (改制縣市名稱、編碼錯誤)，我們還發現地址會有全形半形數字、數字大小寫、中文羅馬數字夾雜、地址中有地下一樓、地下室等錯誤，因此也需要做修正。

透過以下的程式碼呼叫 Google 地圖的 API：

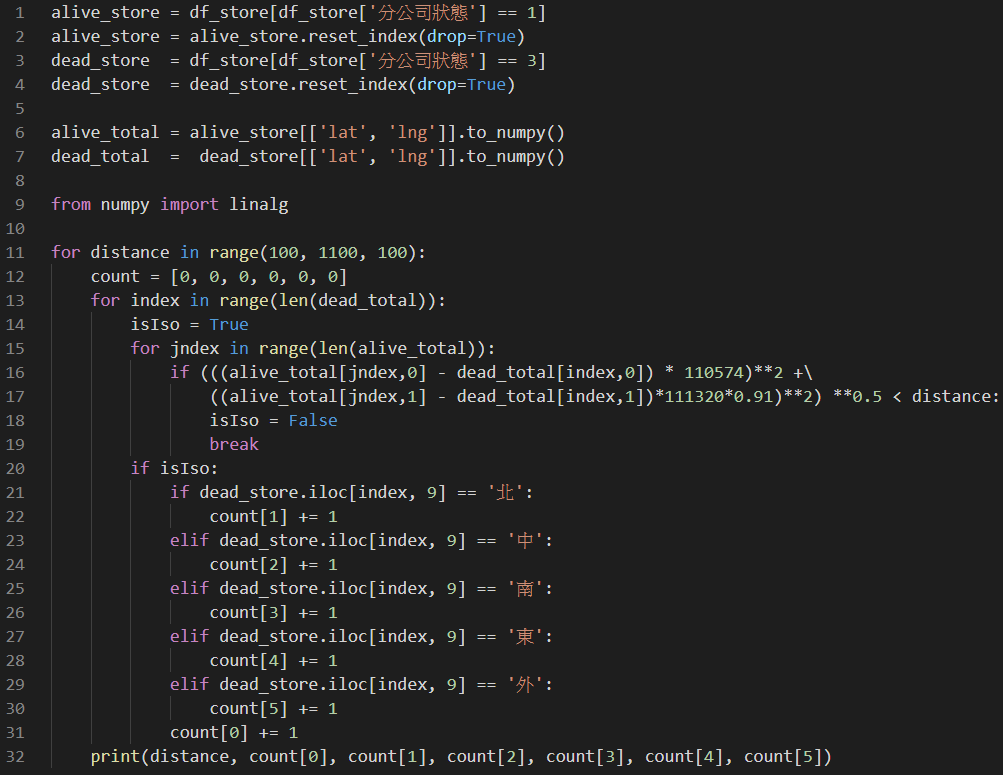
geocode\_result = googlemaps.geocode(處理後的地址字串)

其中 geocode\_result 就是從 Google 服務器回傳的完整圖資，其為 json 格式，透過 ['geometry']['location']['lat'] 和 ['geometry']['location']['lng'] 取出緯度和精度 (單位：degrees)，將資料寫回 csv 檔即完成轉換。

2. KMeans

針對超商的 GPS 座標，對 [全部超商、7-11、全家、萊爾富、OK] 分別進行群集 = [1~20] 的 KMeans 運算。

3. 觀察已歇業超商和開業中超商之關係



將已歇業店家和開業中店家之經緯度座標整理為兩個二維 numpy 陣列，利用

((緯度差 x 110574) ^ 2 + (經度差 x 111320 x cos(緯度)) ^ 2) ^ 0.5

之公式，求出兩兩超商間的距離 (單位為公尺)。針對每家已歇業的超商，假如其周圍  
[100, 200, … , 1000] 公尺內沒有任何營業中的超商，記錄其屬於台灣的哪個區域。

Finding

1. 我們發現 Google 地圖有個 BUG，當輸入"臺中市XXX至善路XXX號"，例如：

"臺中市大里區至善路188號"、"臺中市大里區至善路139號"、"臺中市西屯區至善路61號"

他就會自動變成導航模式，這應該是 "至" 這個字的 parsing error。

2. 結果如下：

灰色代表所有的超商；綠色代表 7-11；藍色代表 全家；紅色代表 萊爾富；粉色代表 OK。





k = 1 即為所有座標點的平均，可發現大約都落在苗栗縣山區，比臺灣地理中心 (位在南投縣) 的緯度高了不少。此外，可以發現 7-11 因為在南部展店的比例較高，所以平均點最南，OK 因為在基隆市、台北市、新北市設點多，所以落在最北。

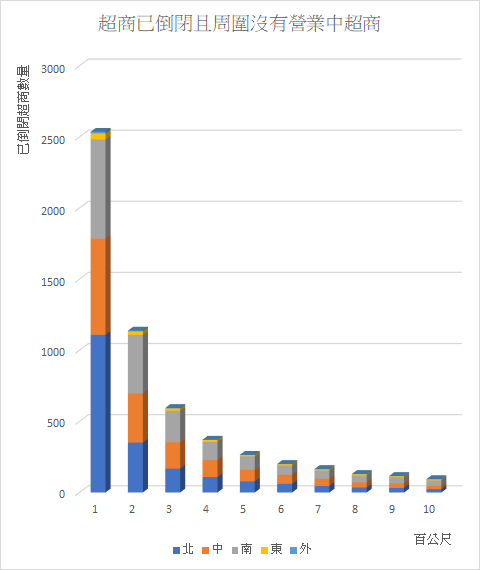
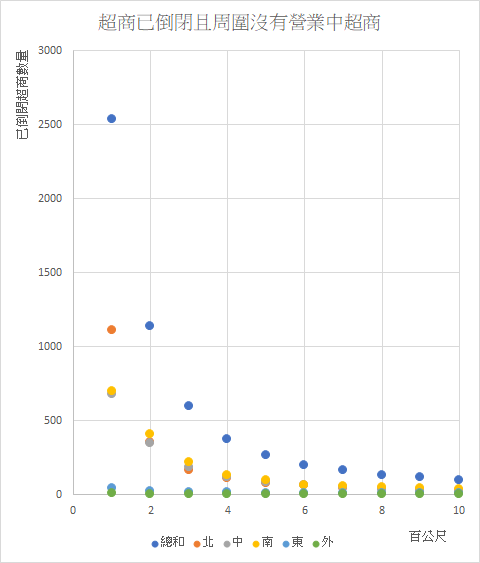
k = 2 時，點落在新北桃園交界和嘉義。

k = 3 時，點落在新北桃園交界、台中、高雄，標記出了台灣三大都會區中心位置。

由於已知超商和人口分布呈高度正相關，我們可以發現隨著 k 越大，就能越精確地呈現出人口的分布。

有個有趣的現象是，雖然金門超商數遠低於澎湖，但是大約 k = 5、6 時，金門就會出現mean point，但是澎湖要到 k = 12 時才會出現，這裡呈現出了 Kmeans 演算法對於距離的敏感程度。

3. 已歇業



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 超商已倒閉且半徑 d 公尺內沒有營業中超商之區域統計表 | | | | | | |
| 距離 d (公尺) | 家數  總和 | 家數  北部地區 | 家數  中部地區 | 家數  南部地區 | 家數  東部地區 | 家數  外島 |
| 100 | 2536 | 1108 | 676 | 698 | 43 | 11 |
| 200 | 1138 | 351 | 346 | 411 | 25 | 5 |
| 300 | 594 | 168 | 185 | 222 | 16 | 3 |
| 400 | 371 | 109 | 119 | 128 | 13 | 2 |
| 500 | 265 | 78 | 81 | 95 | 9 | 2 |
| 600 | 201 | 61 | 64 | 65 | 9 | 2 |
| 700 | 165 | 46 | 51 | 58 | 8 | 2 |
| 800 | 130 | 35 | 36 | 49 | 8 | 2 |
| 900 | 116 | 32 | 32 | 43 | 7 | 2 |
| 1000 | 94 | 24 | 23 | 38 | 7 | 2 |

上表所呈現的是，某一家已倒閉的超商，其周圍半徑 d 公尺內，都沒有營業中的超商之統計數量，日常生活中，通常某家超商倒閉後，競爭對手就會在原址重開，或者是同一家店只是搬到一條街的對面，因此若倒閉後周圍沒有還繼續營業的超商，表示該地點的超商難以經營，這對於生活在該地的人們可以說是噩耗。上表呈現的是，假如你家樓下本來有開了一家超商，卻不幸倒閉，你至少需要走多遠才可以找到第二家超商。

有趣的是，每一個區域內，被孤立的倒閉超商之數量都是隨著距離呈現**"指數下降"**，並且和區域沒有明顯的關係傾向，全台灣倒閉後，半徑一公里內沒有替代者的商店數量不到一百家，因此只要曾經開過便利超商，該區域通常經濟發展程度具有一定水準，所以不用擔心找不到第二家便利超商。

**Contribution**

有別於其他組使用 Kaggle 已標記的現成 Data set，我們想透過自己收集實際資料，來找到一些有趣的發現。我們從最接近日常生活的超商下手，嘗試用一些簡單的政府公開資料，輔以一些課堂所學的資料科學方法，以及一些從網路上學到的工具，來找出可能藏在這些資料背後的資訊。我們得到一些有趣的發現，諸如：

1. 女性比例較高的地區，可發現全家的佔比較高。
2. 7-11 的設點策略為以量取勝，基本上只要有人，不論多寡，就會有 7-11 的蹤跡，且 7-11 佔比從北部區域往中南部直到花東外島逐漸上升，推論是因為經濟發展程度相對較低的地區只有7-11這種大品牌能夠順利經營。
3. 超商會依據不同的地理分區（北、中、南等）而有不同設點策略，同區域內各超商的佔比傾向相同。
4. 超商數量和人口數量之相關係數極高。
5. 人們對於超商的選擇很大一部份取決於"距離"。
6. 即便有超商倒閉，要找到取代的超商是十分容易的 (一公里內)。

這份報告綜合了課堂所學，並讓我們充分複習到相關演算法的原理和實作方式，可謂獲益良多。

**Reference**

[1] 當前經濟情勢概況(專題：零售業發展現況及國際比較)。  
<https://www.moea.gov.tw/Mns/dos/bulletin/Bulletin.aspx?kind=23&html=1&menu_id=10212&bull_id=6124>

[2] 台灣超商龍頭爭霸戰：全家、7-11、萊爾富。<https://journal.eyeprophet.com/%E8%B6%85%E5%95%86%E9%BE%8D%E9%A0%AD%E7%88%AD%E9%9C%B8%E6%88%B0/?fbclid=IwAR0rSh9-4PRBINSNfrOEFcF_-LVXatplyhYU0J7EFPOv-JexPfHknqnUWow>

[3] 不怕統一、全家夾擊！超商老三萊爾富不開24小時照樣拼出營收百億。<https://udn.com/news/story/6839/3988181?fbclid=IwAR2WIIaL8TSqecnq0fN7KrUvsa4oDRHqwo5iUnvgbzMULVjD8rcNMcoN1nI>

[4] 全國4大超商資料集。<https://data.gov.tw/dataset/32086>

[5] ibon便利生活站門市查詢。<http://www.ibon.com.tw/retail_inquiry.aspx#gsc.tab=0>

[6]中華民國統計資料網。<https://statdb.dgbas.gov.tw/pxweb/Dialog/statfile9.asp>

[7]內政統計查詢網。<https://statis.moi.gov.tw/micst/stmain.jsp?sys=100>

[8]106年度綜合所得稅申報初步核定統計專冊。<https://www.fia.gov.tw/News_Content.aspx?n=7769B1BD01306B45&sms=BD450CA810662F3D&s=6C2B8374CBFF550E>

[9]勞動統計查詢網。<https://statdb.mol.gov.tw/statis/jspProxy.aspx?sys=210&kind=21&type=1&funid=q020741&rdm=l30eabte>

[10] Re: [問題] 基隆市的ok超商是不是特別多？。<https://www.ptt.cc/bbs/CVS/M.1482819043.A.682.html?fbclid=IwAR2AwcaipxK1VuLisu-PYAcAeLWdKTQkfeCMeU_ZOFxq9unBOBo3iV7duys>