

108-1 巨量資料分析應用與實作期末專題報告

全台毒品犯罪資料分析

指導教授: 江傳文 副教授

組員 : 0451030 黃柏皓

0551076 黃泰源

目錄

1. 研究目的
2. 資料介紹與處理
3. 方法介紹
4. 分析過程
5. 結論與未來展望

研究目的

研究目的

隨著科技的進步，新型毒品的製作層出不窮，更有無良毒品商人經常將毒品包裝成食品或其他日常用品來逃過警察的攔查。為此我們於內政部的網站中找到了毒品犯罪資料，希望藉由這門課所學習到的分析方法可以在這份資料分析有關各地區場所每月的毒品犯罪資料。

下列幾項是我們的研究目的：

1. 尋找毒品數量與地區分布存在之關聯，增加搜查效率並減少警力分配不均。
2. 找出供應鏈上游，擒賊先擒王。
3. 觀察犯罪趨勢，預估罪犯動向。

資料介紹與處理

資料介紹

本報告所使用的資料集為內政部警政署提供的毒品犯罪資料，共85031筆案件紀錄。

no	type	oc_dt	oc_addr	oc_p1	oc_p2	oc_p3	proc_no	kind	weight_g
編號	案類	發生日期	發生地點	發生場所一	發生場所二	發生場所三	嫌疑犯人數	毒品品項	數量淨重_克
1	毒品	107/01/01	桃園市桃園區	普通住宅(一、二層樓)			1	甲基安非他命	0.4
2	毒品	107/01/01	桃園市桃園區	公路(省、縣、鄉、鎮道			1	海洛因	0.95
3	毒品	107/01/01	新北市新莊區	普通住宅(一、二層樓)			1	甲基安非他命	0
4	毒品	107/01/01	台南市安平區	公路(省、縣、鄉、鎮道			1	安非他命	0
5	毒品	107/01/01	高雄市楠梓區	普通住宅(一、二層樓)			1	鴉片(阿片)	0

資料處理(1/2)

經過整理後，毒品的種類多達115種，且單筆案件經常只出現其中一種的毒品，我們將115種毒品劃分為9個種類，並以數字代號代稱。

毒品名稱	分類
美沙冬	抵癮藥品
古柯鹼	一級毒品
甲基安非他命	二級毒品
K他命	三級毒品
阿普唑他(蝴蝶片)	四級毒品
罌粟種子	一級毒品原料
麻黃鹼	二級毒品原料
鹽酸羥亞胺	三級毒品原料
苯基丙酮	四級毒品原料

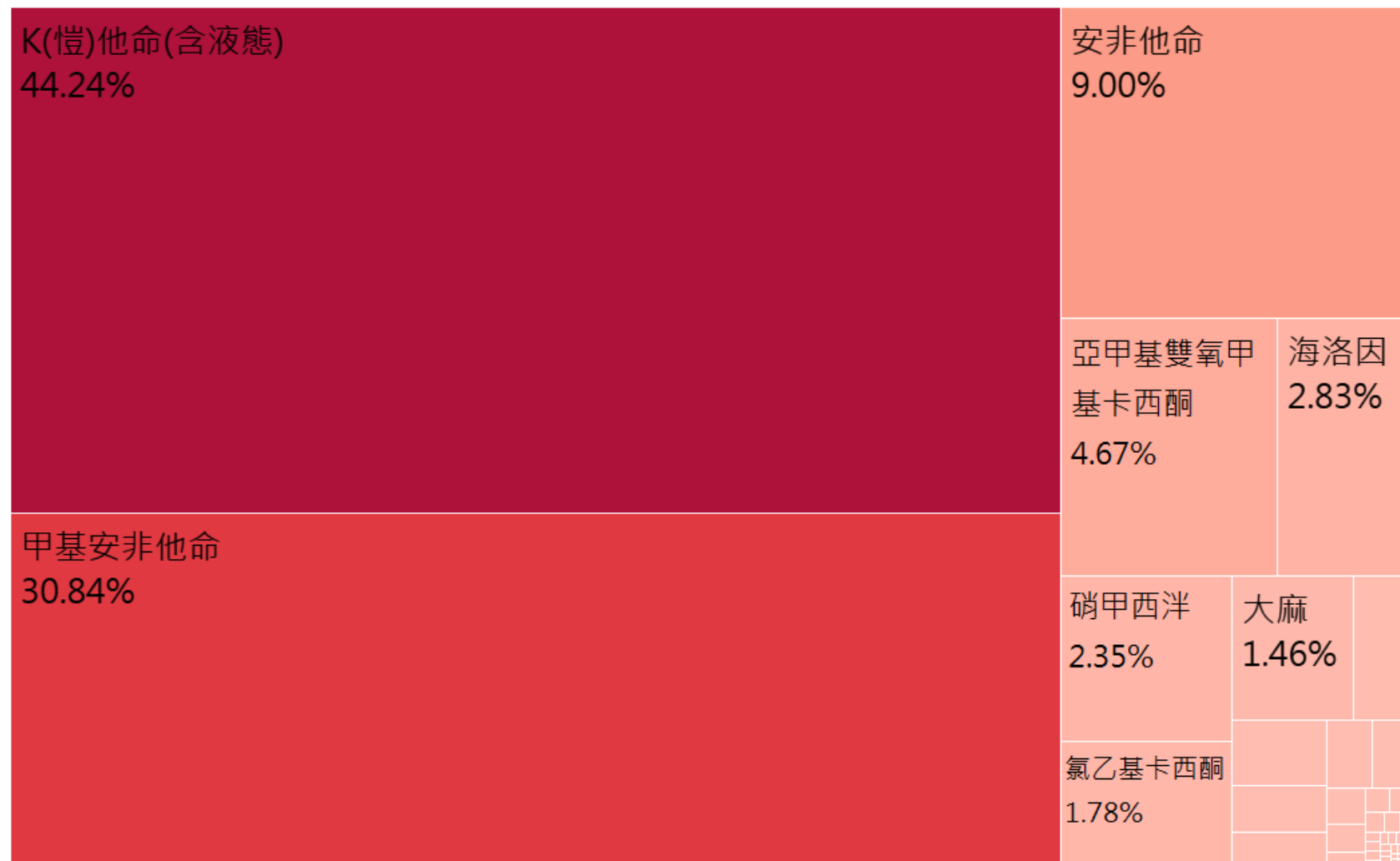
種類	代號
一級毒品	1
二級毒品	2
三級毒品	3
四級毒品	4
一級原料	5
二級原料	6
三級原料	7
四級原料	8
抵癮藥品	0

資料處理(2/2)

我們將毒品犯罪資料依案件將各屬性數量加總起來並得到以下結果

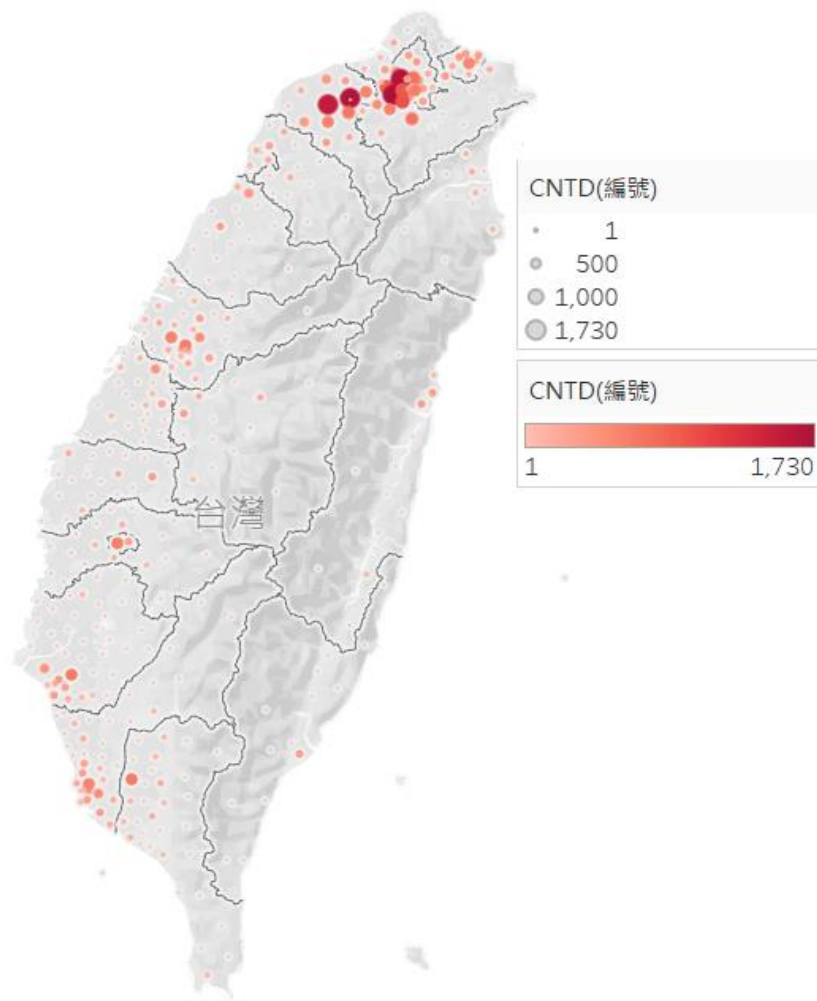
no	0	1	2	3	4	5	6	7	8
1	0	0	0.4	0	0	0	0	0	0
2	0	0.95	0	0	0	0	0	0	0
3	0	0	0.01	0	0	0	0	0	0
4	0	0	0.01	0	0	0	0	0	0
5	0	0.01	0	0	0	0	0	0	0

資料視覺化呈現(1/3)

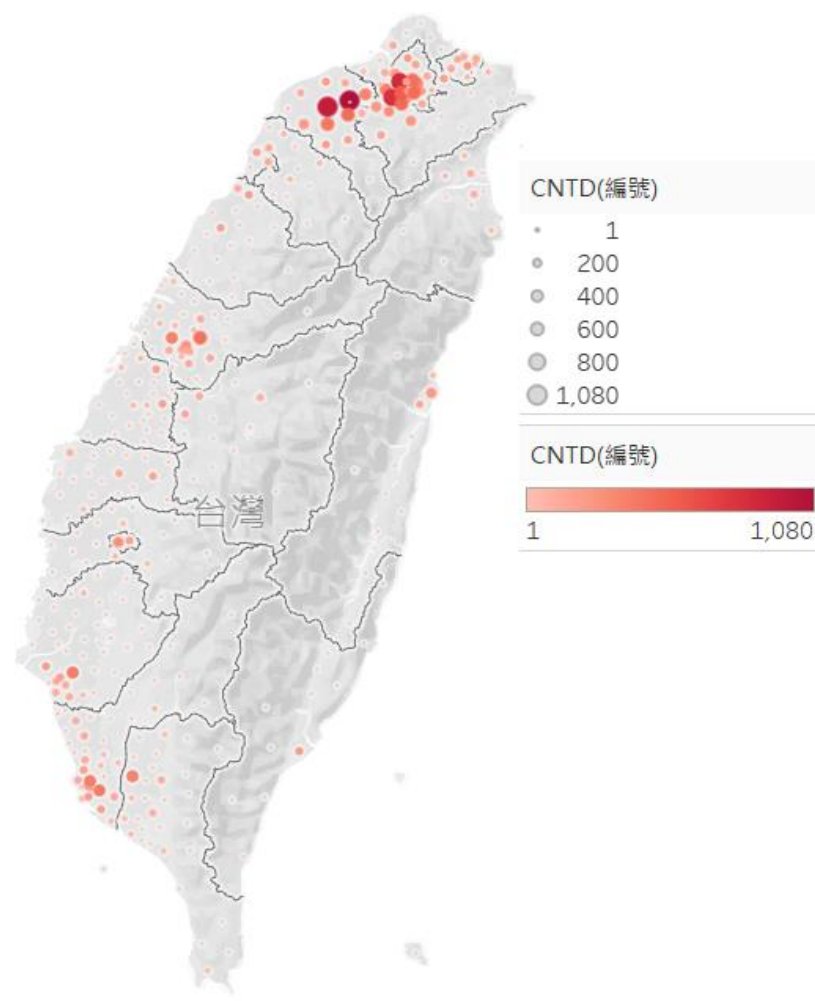


資料視覺化呈現(2/3)

107年毒品案件分布圖

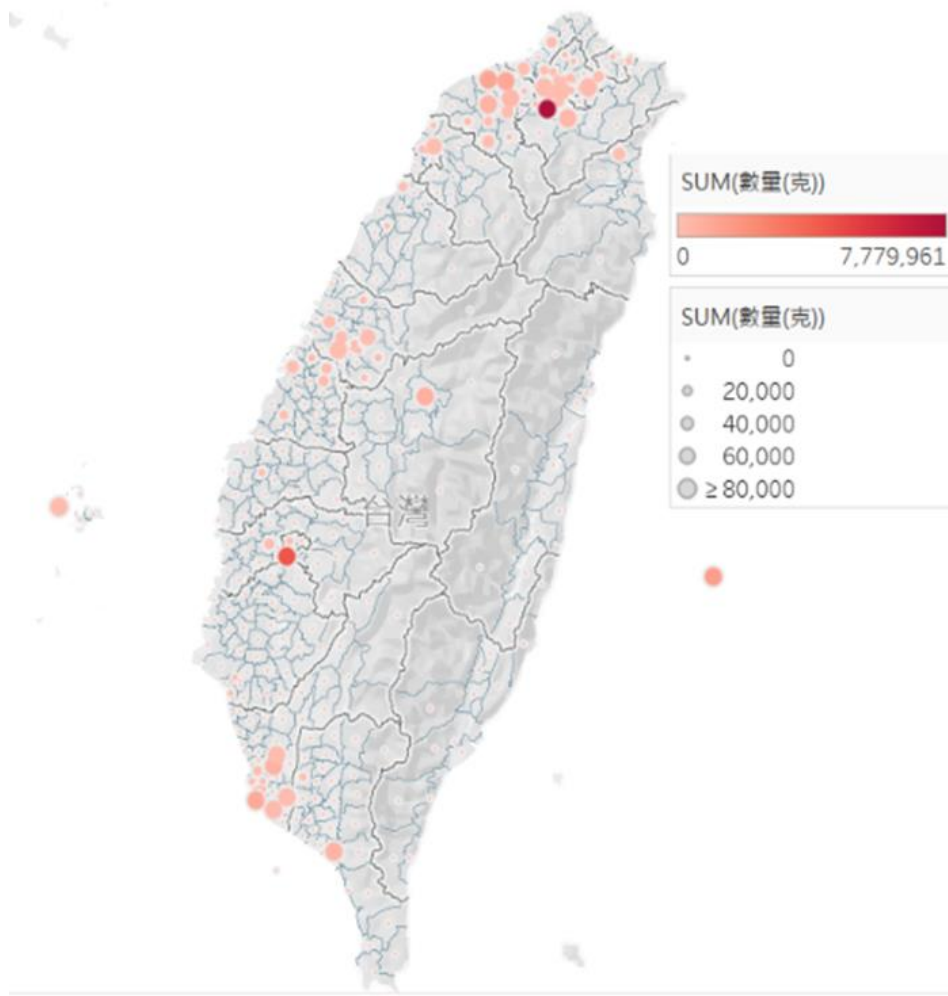


108年毒品案件分布圖

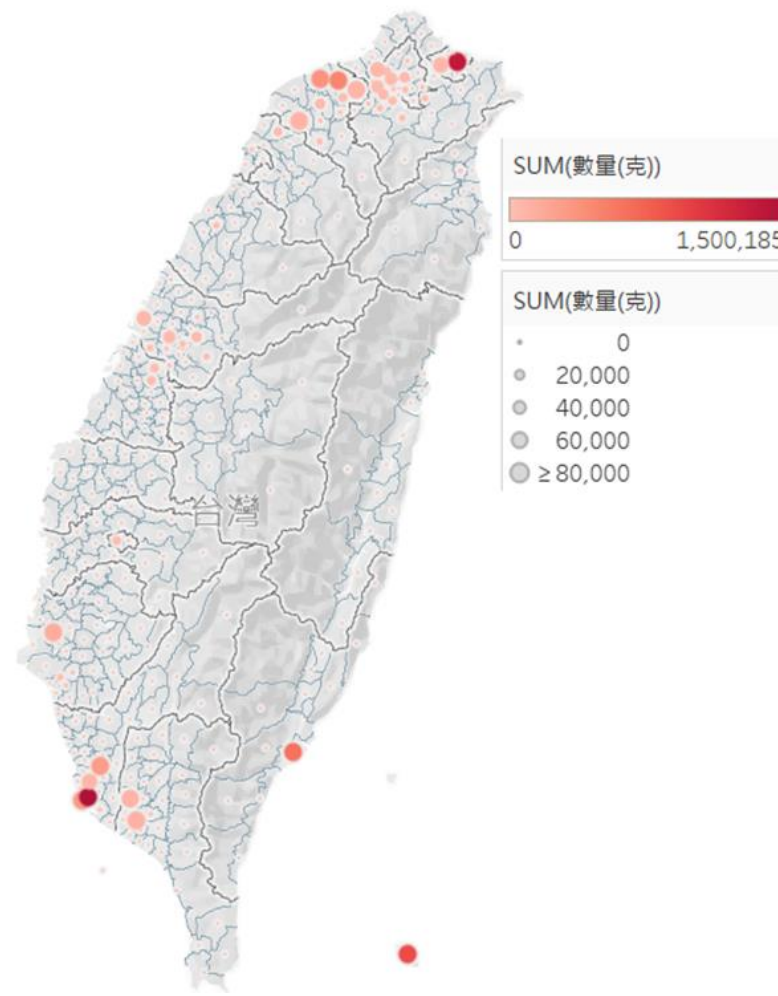


資料視覺化呈現(3/3)

107年毒品數量分布圖



108年毒品數量分布圖



方法介紹

方法介紹

DBSCAN (Density-based spatial clustering of applications with noise)，這個算法是以密度為本的：給定某空間裡的一個點集合，這算法能把附近的點分成一組（有很多相鄰點的點），並標記出位於低密度區域的局外點（最接近它的點也十分遠）。其中主要考量的參數為：

1. Eps：指定半徑，決定區域範圍。
2. MinPts：區域最小密度。

根據以上兩參數，空間中的各筆資料可被區分為下列三種：

1. 核心點 (Core points)：於指定半徑中具有大於最小密度的點，存在於密集區域內部。
2. 邊界點 (Border points)：於指定半徑中不具有最小密度的點，鄰近於核心點。
3. 雜訊點 (Noise points)：於指定半徑中不具有最小密度的點，且不鄰近於任一核心；既非核心點，亦非邊界點，則稱為雜訊點。

演算法步驟：

1. 區分核心、邊界和雜訊點。
2. 清除雜訊點。
3. 利用指定半徑，由所有核心點構成之邊緣。
4. 連結鄰近之核心點成群。
5. 將邊界點納入最相近的核心點之群。

分析過程

分析過程(1/4)

Step1. 將毒品犯罪資料轉換為各案件根據屬性總和數量的表格。

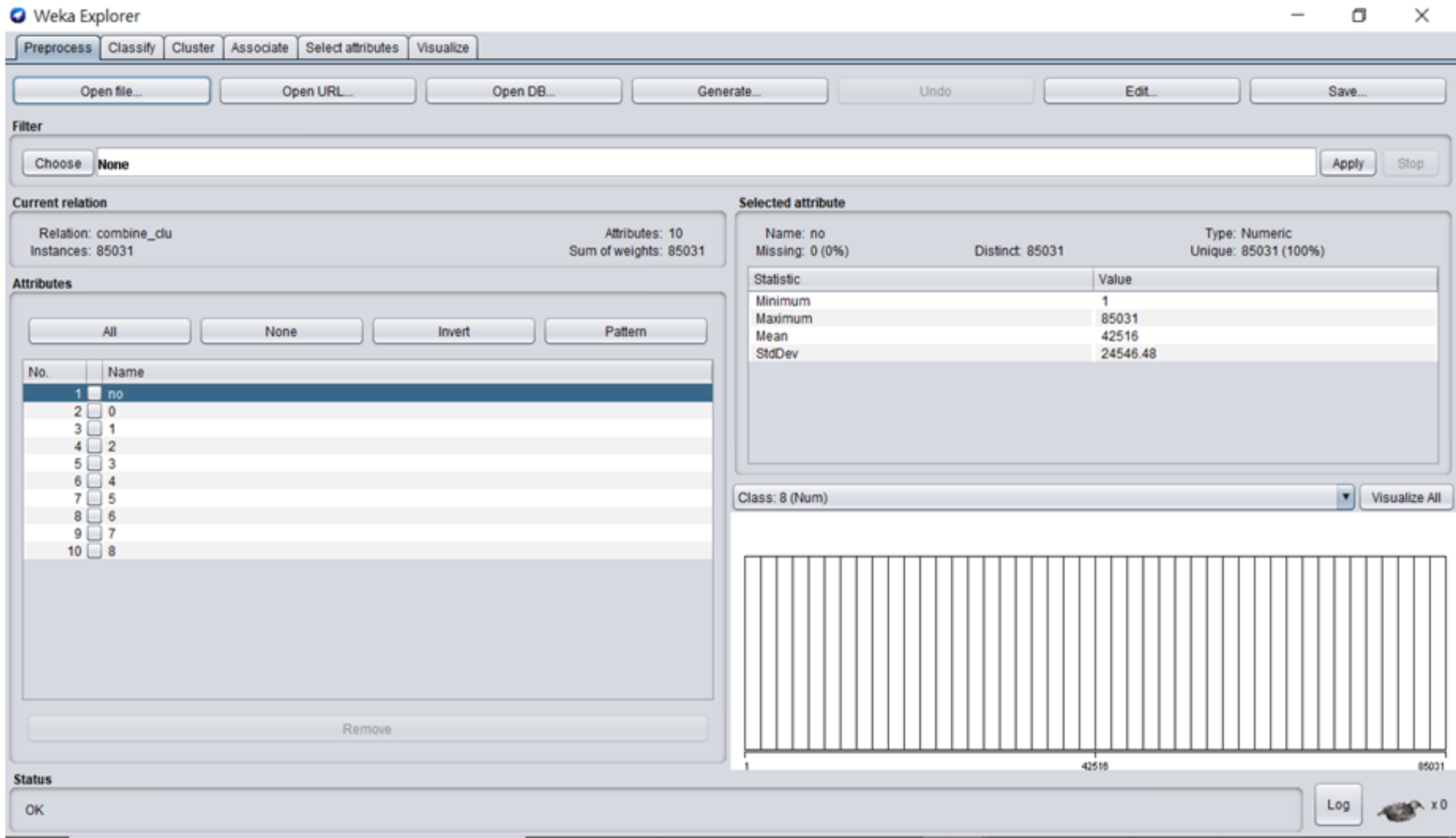
轉換前:

no	type	oc_dt	oc_addr	oc_p1	oc_p2	oc_p3	proc_no	kind	weight_g
編號	案類	發生日期	發生地點	發生場所一	發生場所二	發生場所三	嫌疑犯人數	毒品品項	數量淨重_克
1	毒品	107/01/01	桃園市桃園區	普通住宅(一、二層樓)			1	甲基安非他命	0.4
2	毒品	107/01/01	桃園市桃園區	公路(省、縣、鄉、鎮道			1	海洛因	0.95
3	毒品	107/01/01	新北市新莊區	普通住宅(一、二層樓)			1	甲基安非他命	0
4	毒品	107/01/01	台南市安平區	公路(省、縣、鄉、鎮道			1	安非他命	0
5	毒品	107/01/01	高雄市楠梓區	普通住宅(一、二層樓)			1	鴉片(阿片)	0

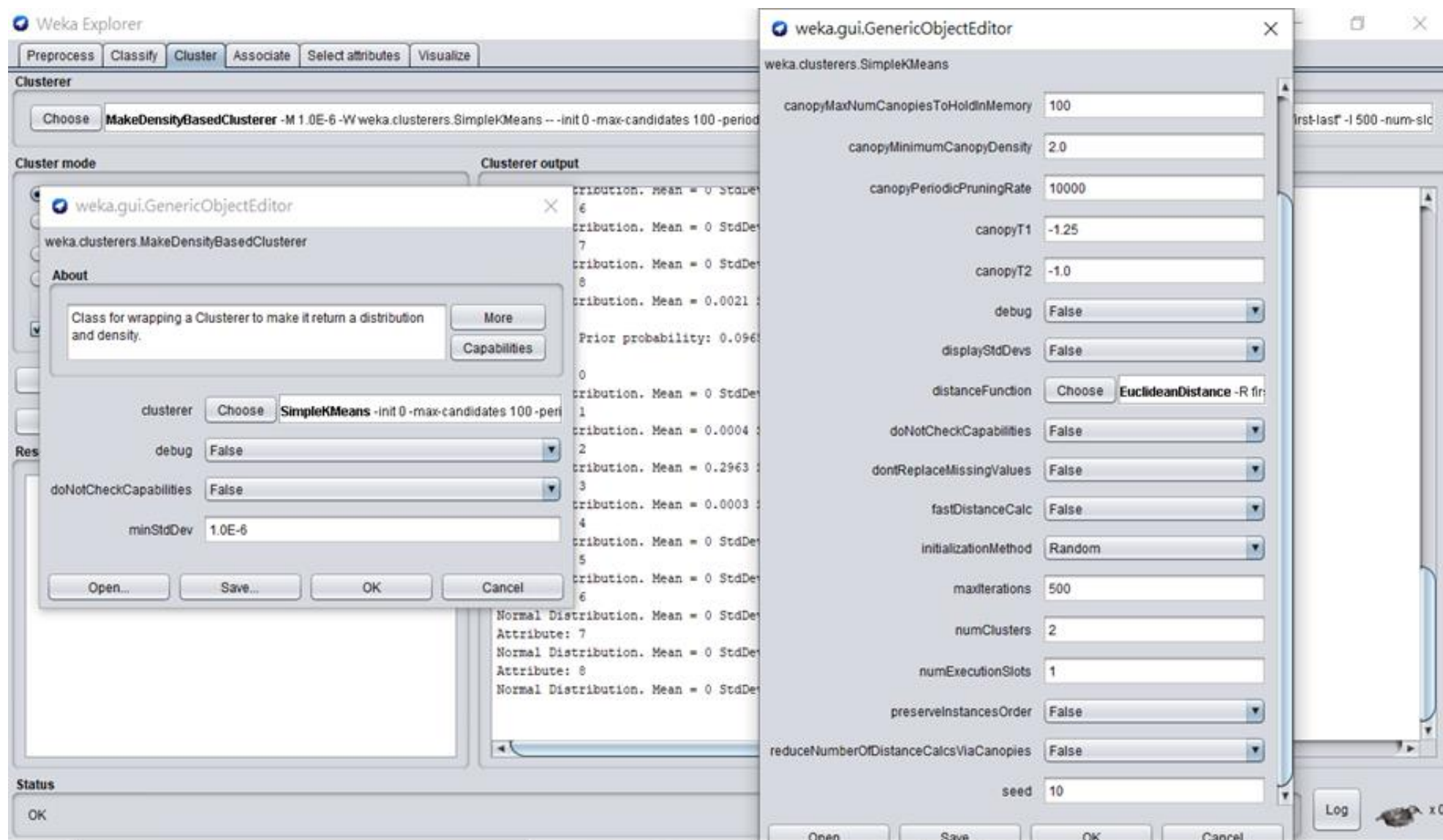
轉換後：

no	0	1	2	3	4	5	6	7	8
1	0	0	0.4	0	0	0	0	0	0
2	0	0.95	0	0	0	0	0	0	0
3	0	0	0.01	0	0	0	0	0	0
4	0	0	0.01	0	0	0	0	0	0
5	0	0.01	0	0	0	0	0	0	0

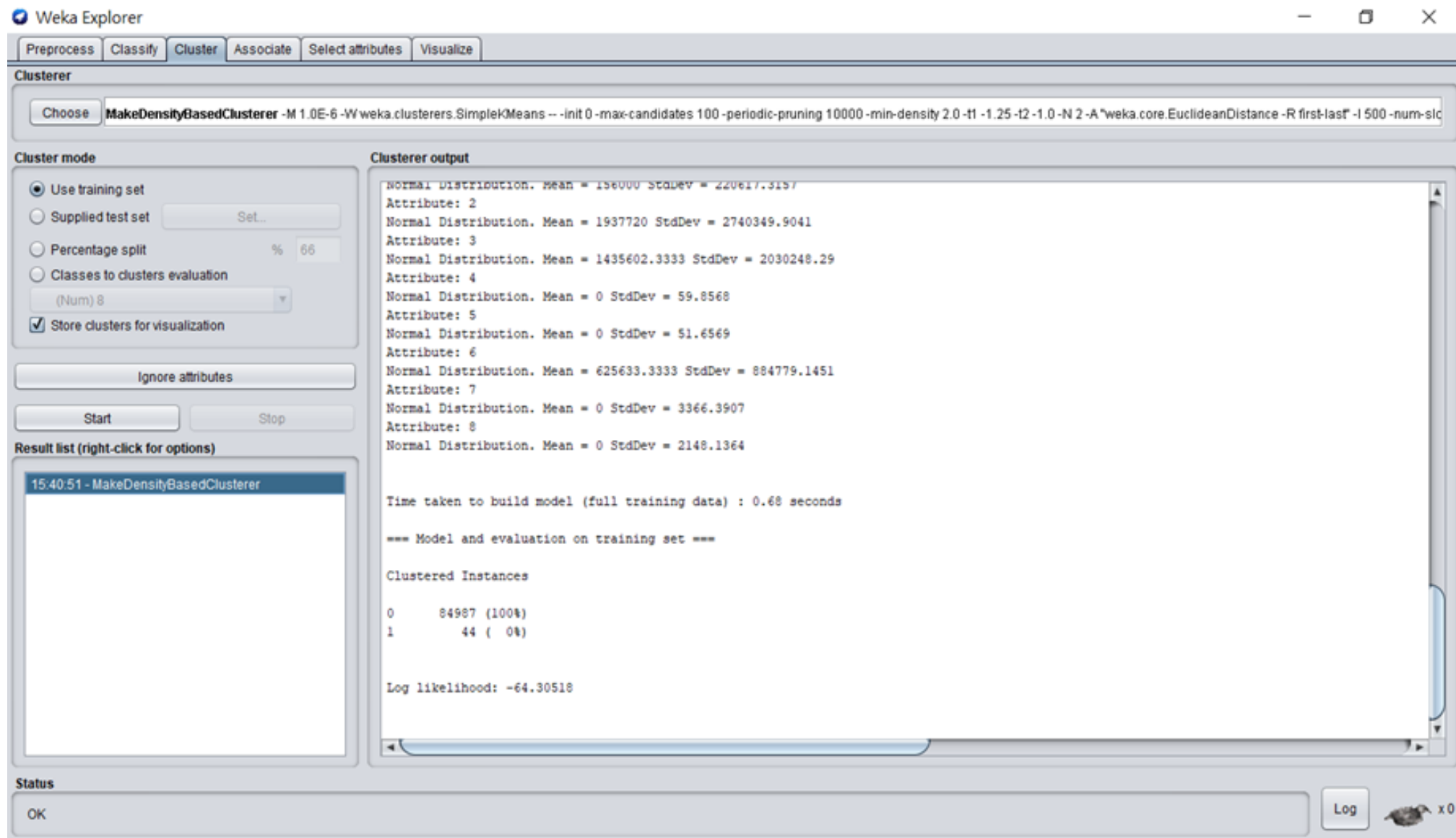
Step2.打開WEKA軟體，點取Explorer再點選Preprocessor標籤，按下Openfile按鈕後，開啟想要分群的檔案，以下是操作完成的畫面。



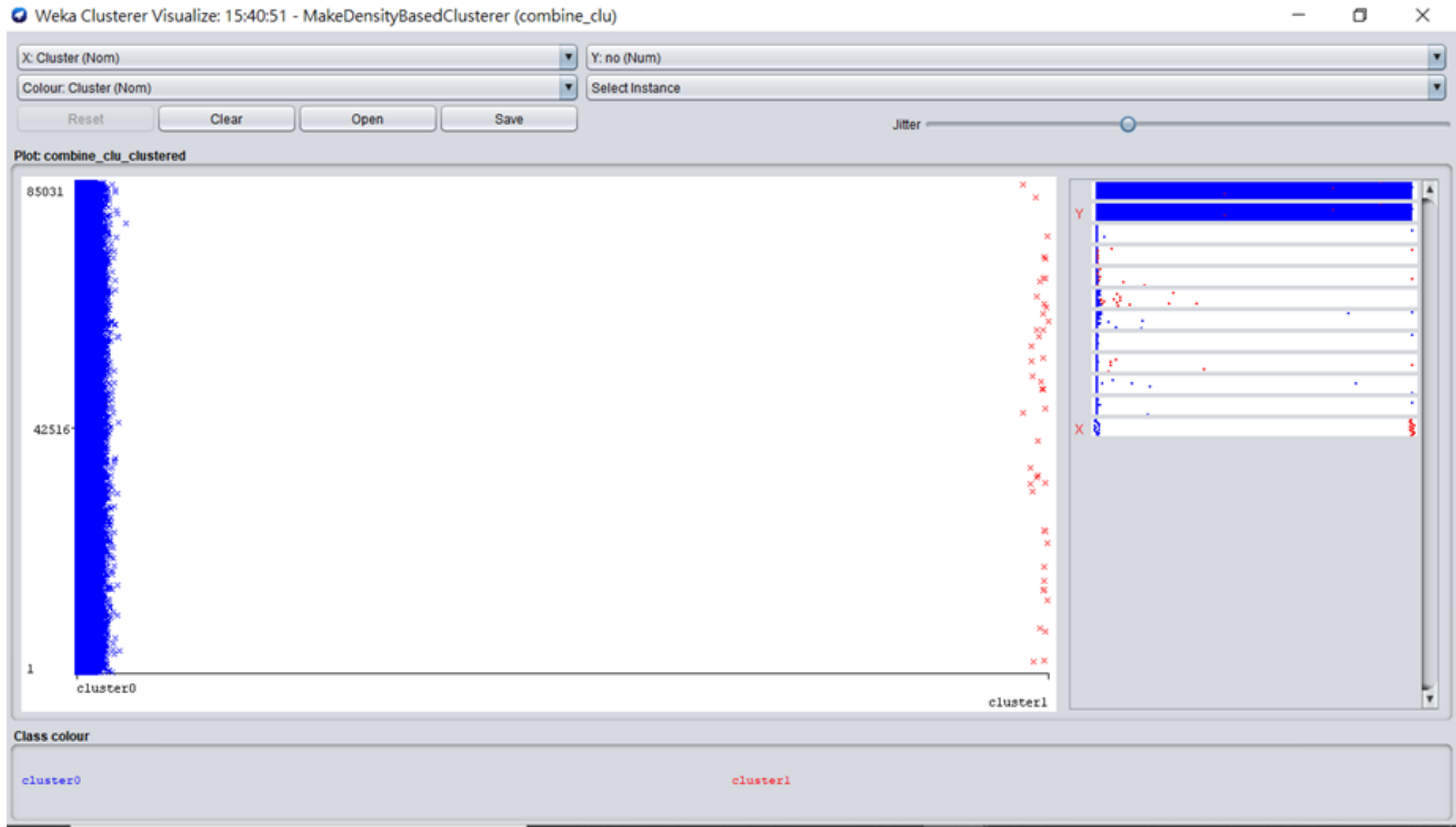
Step3. 點選Cluster標籤並選擇分類器MakeDensityBasedCluster並選擇所包裹的聚類演算法，這邊我們使用的是SimpleKMeans，選擇初始點方式與分群數目。



Step4. 在按下Start按鈕前，按下Ignore按鈕選取不列入計算的屬性，本此分析為no（案件編號），然後按下Start開始分析，過一段時間後就能得分析結果。



Step5. 藉由右方報表所顯示的Centroid Point、SSE、likelihood數值我們可以藉此判斷聚類模型的好壞，並在Result list針對該筆紀錄按下右鍵點選Visualize cluster Assignment觀察各案件毒品屬性與賦予的類別關係。



分析過程(2/4)

在第一次聚類後，有少數數據被分出來，大部分數據還是被分為一類。如下：

28	1	57	84945
----	---	----	-------

經過觀察後，我們發現少數數據多為數量龐大的極值：

[illegible]

分析過程(3/4)

在儲存結果之後，
我們將極值取出並
將84945筆資料，
如此重複上述步驟
之後，所得的分類
結果如右圖，我們
總共分了24個群。

combine	C1 28	C2 1	C3 57	84945				
combine_2	C4 12	C5 210	C6 26	C7 622	C8 42	C9 102	C10 9	83922
combine_3	C11 3101	C12 2461	C13 320	C14 1110	C15 275	C16 191	C17 621	75863
combine_4	C18 3521		C19 1525		C20 4283		66534	
combine_5	C21 9201		C22 13977		C23 38162			C24 5194

分析過程(4/4)

在同一步驟分出的群我們將它們劃為同一層，一共分為5層。上層平均數量最多，之後依層數遞減。我們認定上兩層為所謂的毒品供應製造層，而下面的層數則是所謂的毒品消費層。

layer_1

C1	C2	C3
28	1	57

layer_2

C4	C5	C6	C7	C8	C9	C10
12	210	26	622	42	102	9

layer_3

C11	C12	C13	C14	C15	C16	C17
3101	2461	320	1110	275	191	621

layer_4

C18	C19	C20
3521	1525	4283

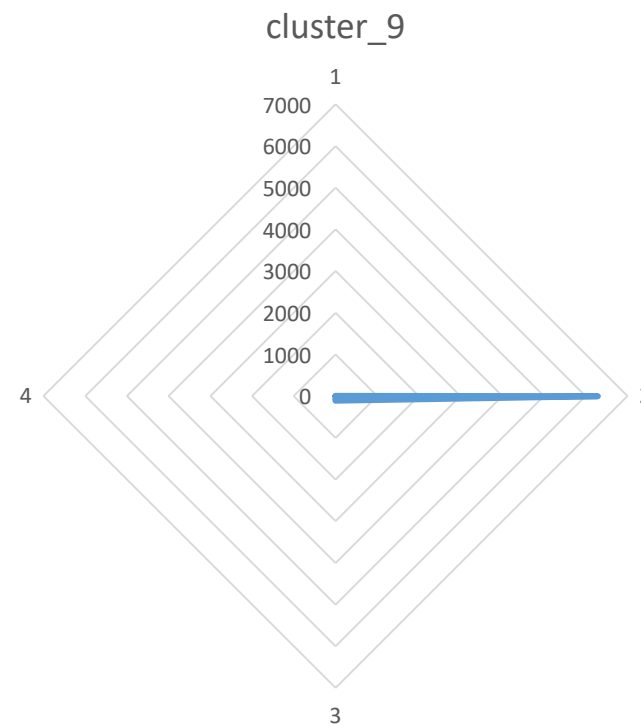
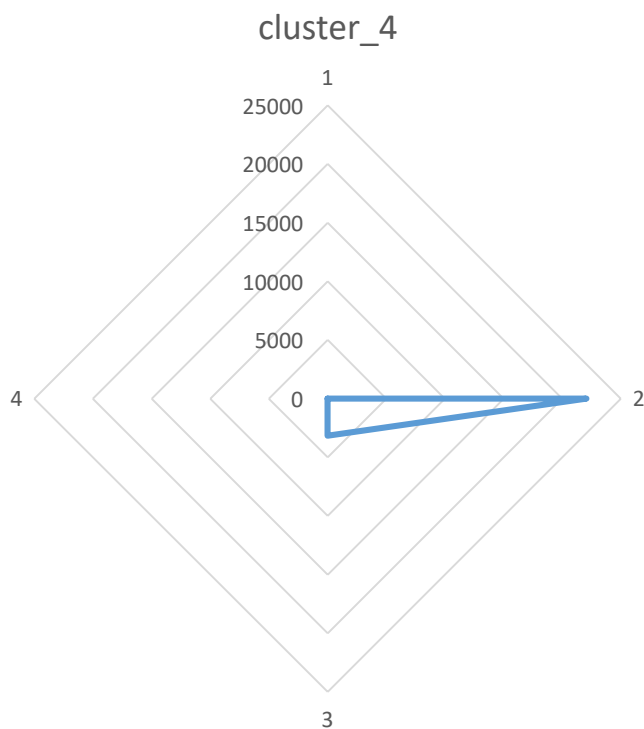
layer_5

C21	C22	C23	C24
9201	13977	38162	5194

結果呈現

結果呈現(1/8)

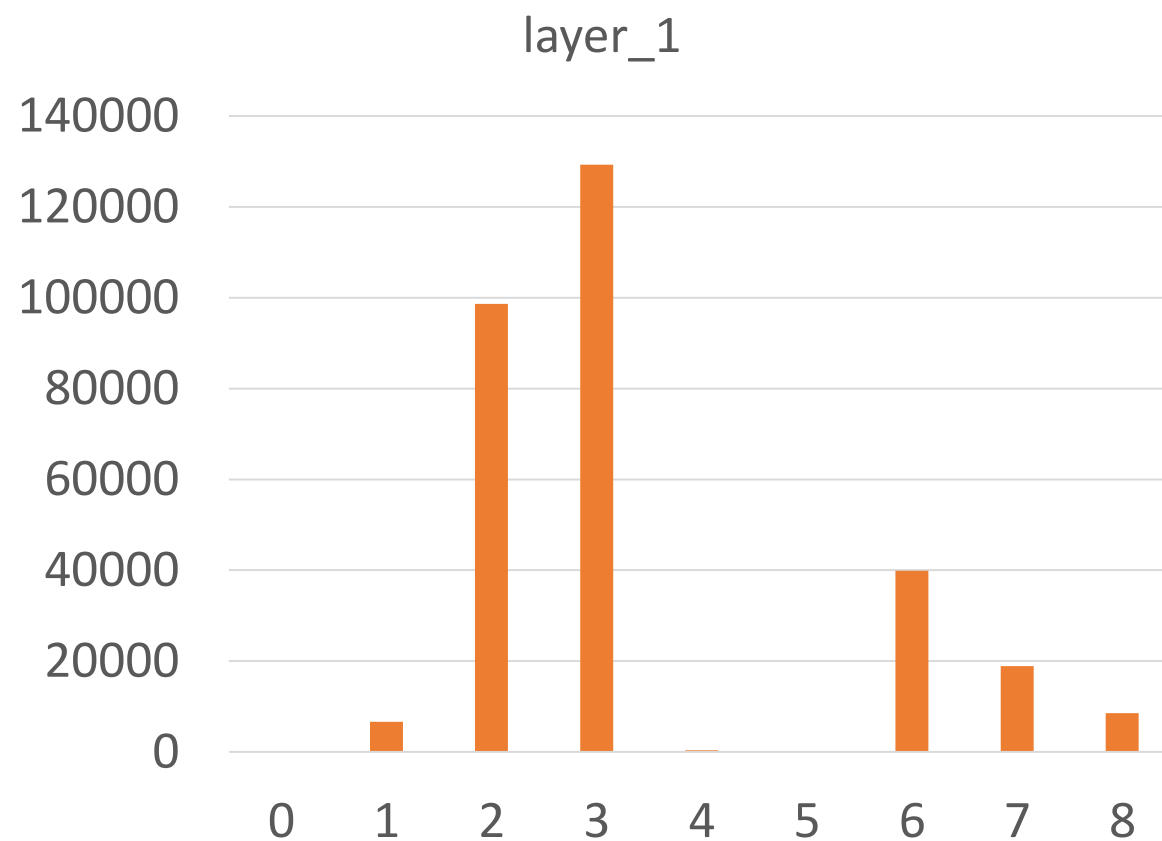
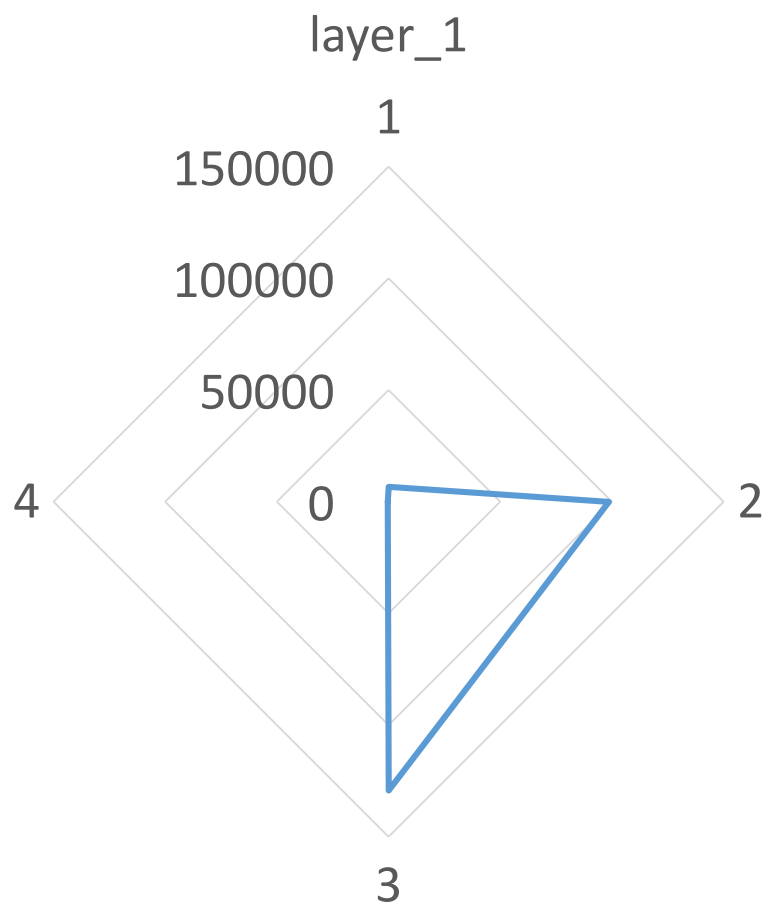
利用DBSCAN將Data Set分成24個clusters後，依毒品分級(0~8)的毒品數量(單位為公克)，將各群的Centroid Point以雷達圖及長條圖的方式呈現，以便觀察各群的分群特性。



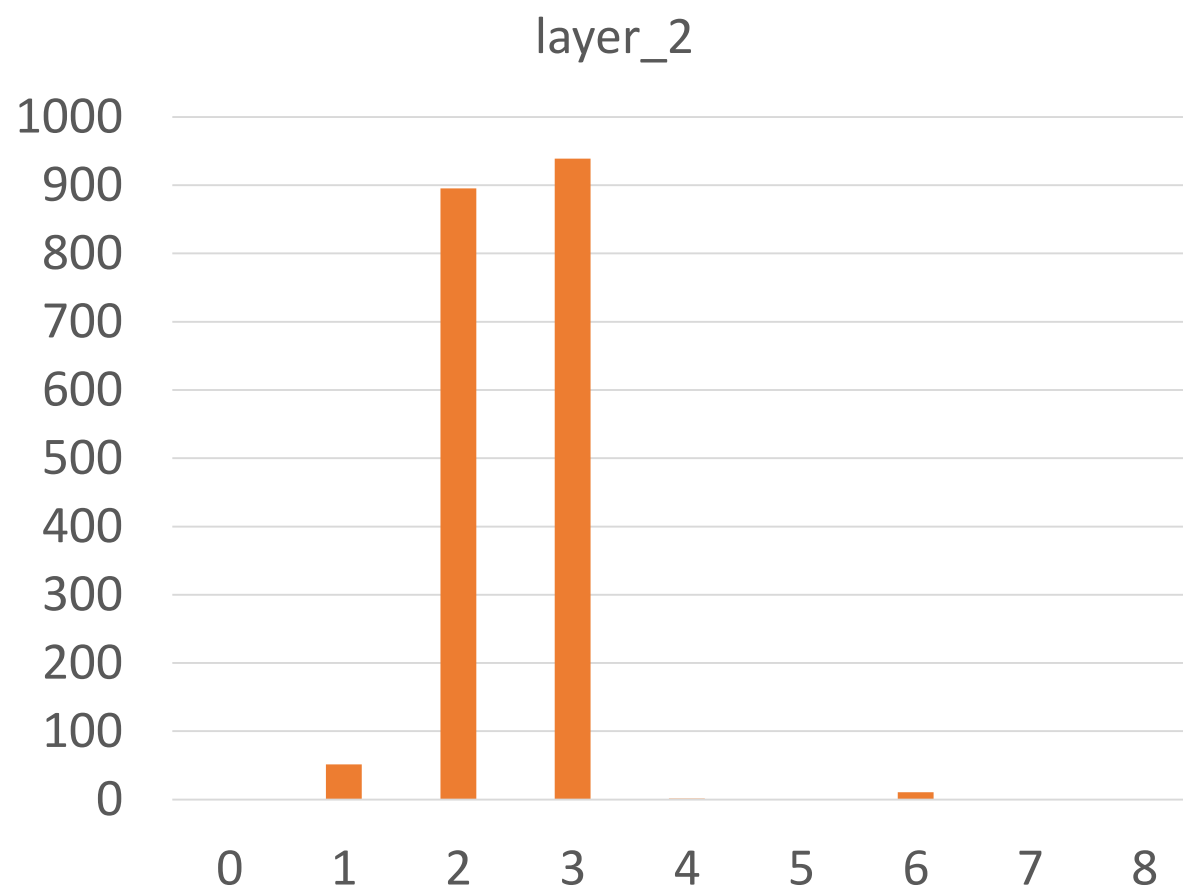
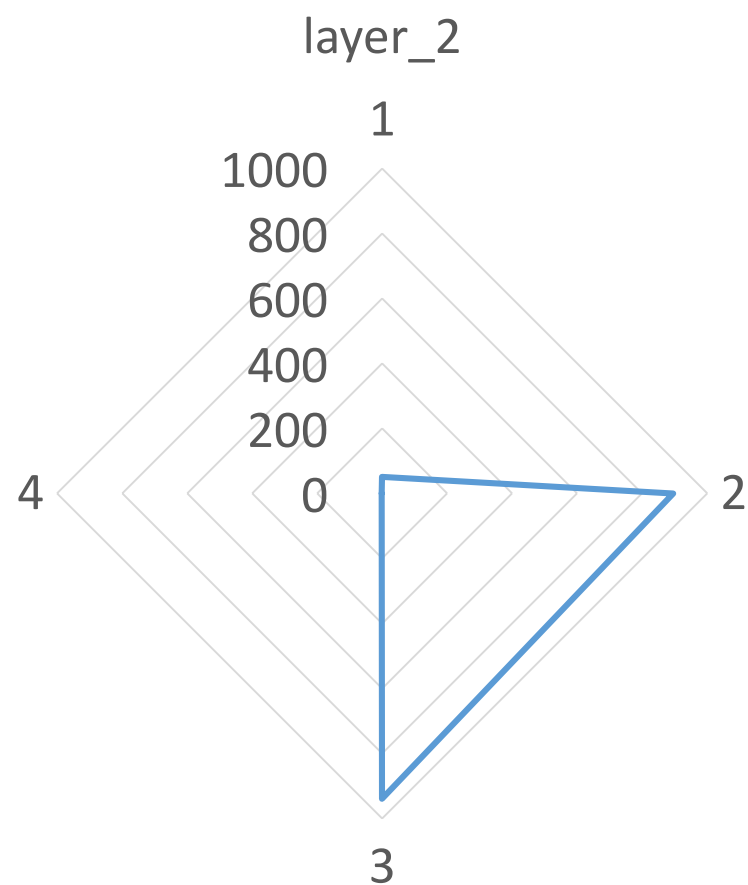
結果呈現(2/8)

在觀察各群的特性後，我們發現分群主要大多會依據1、2、3、4屬性與毒品數量進行分群，對於這樣的結果我們認為較無太大的意義；但我們發現越早分出來的群體得毒品數量越多，具有較多的6、7、8、9屬性，但案件數較少，所以我們將各層屬性平均並進行視覺化呈現。

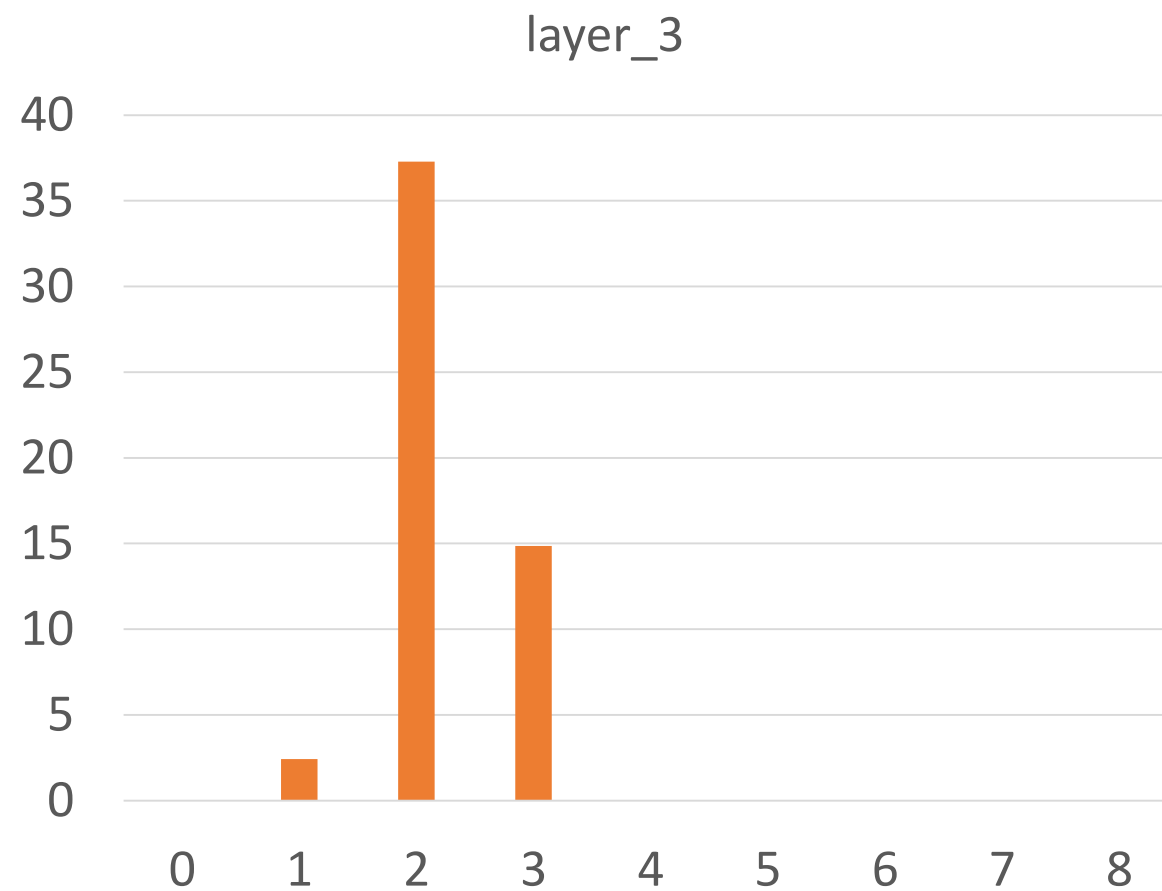
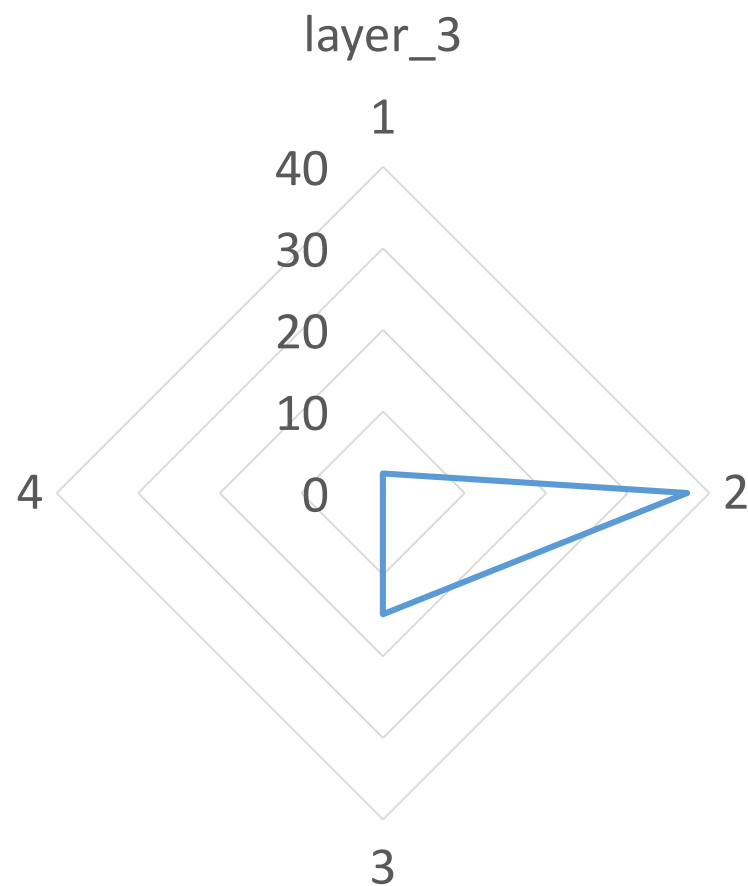
結果呈現(2/8)



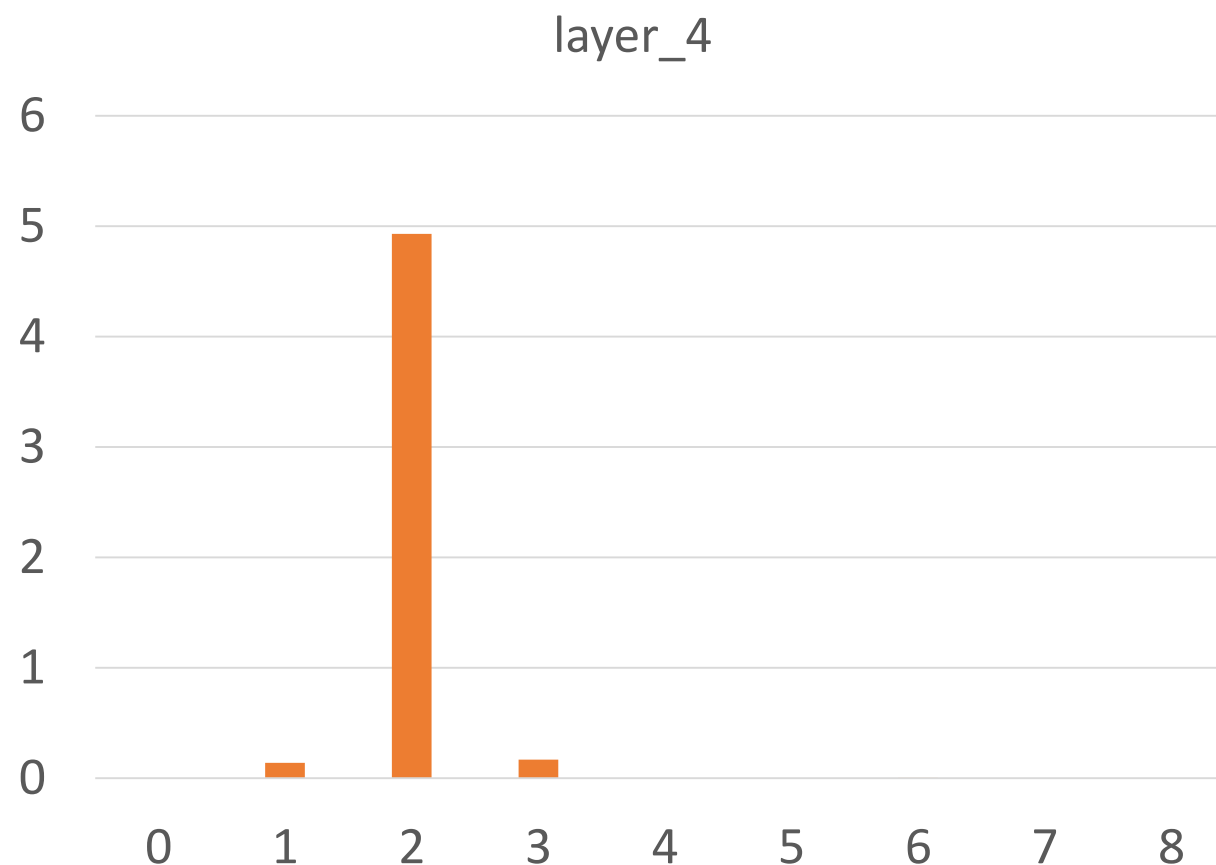
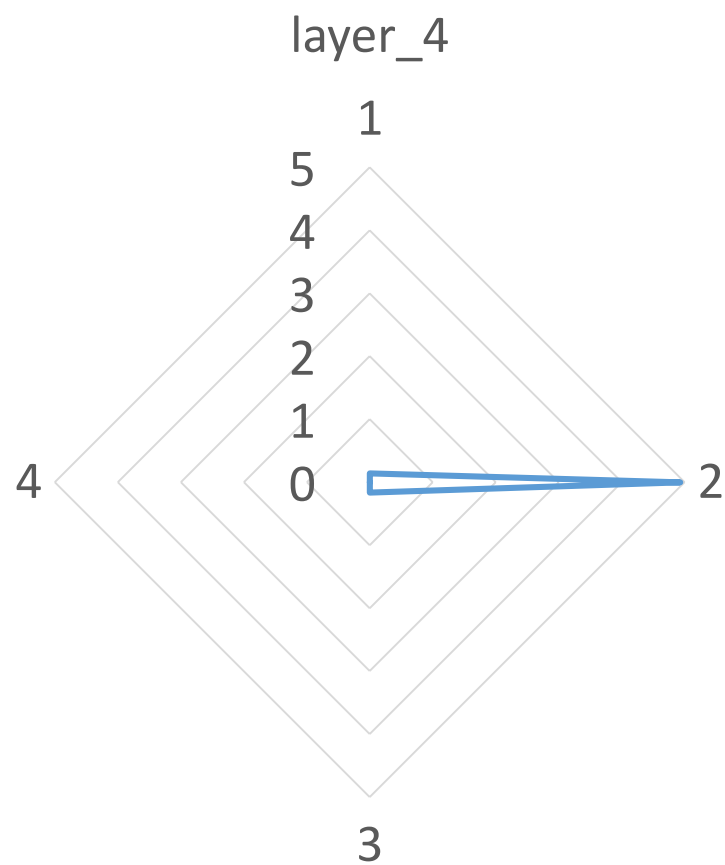
結果呈現(2/8)



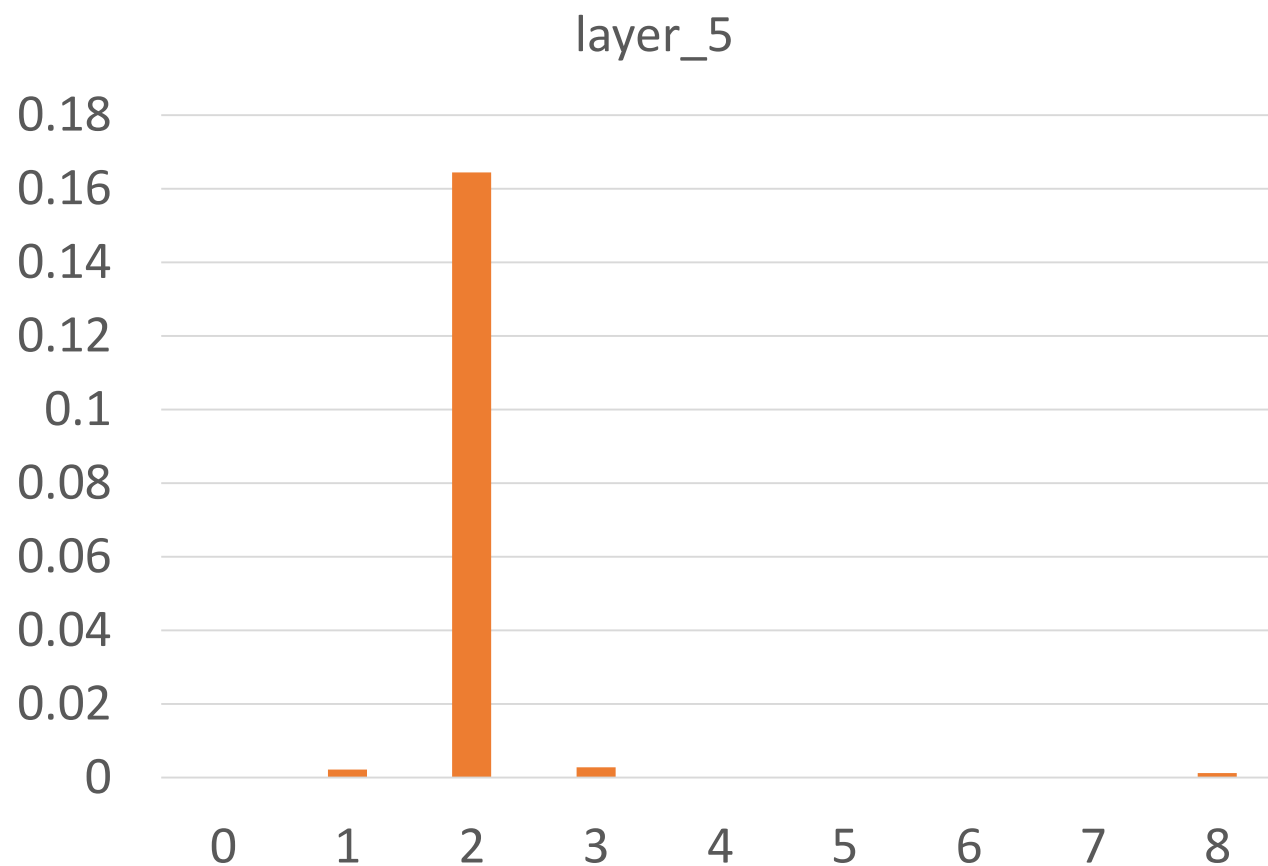
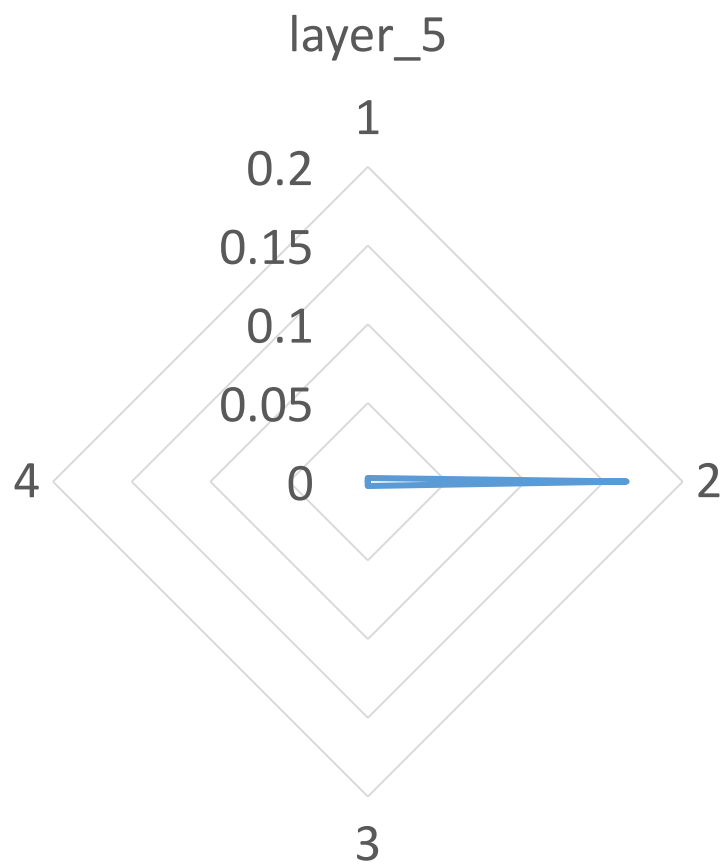
結果呈現(2/8)



結果呈現(2/8)

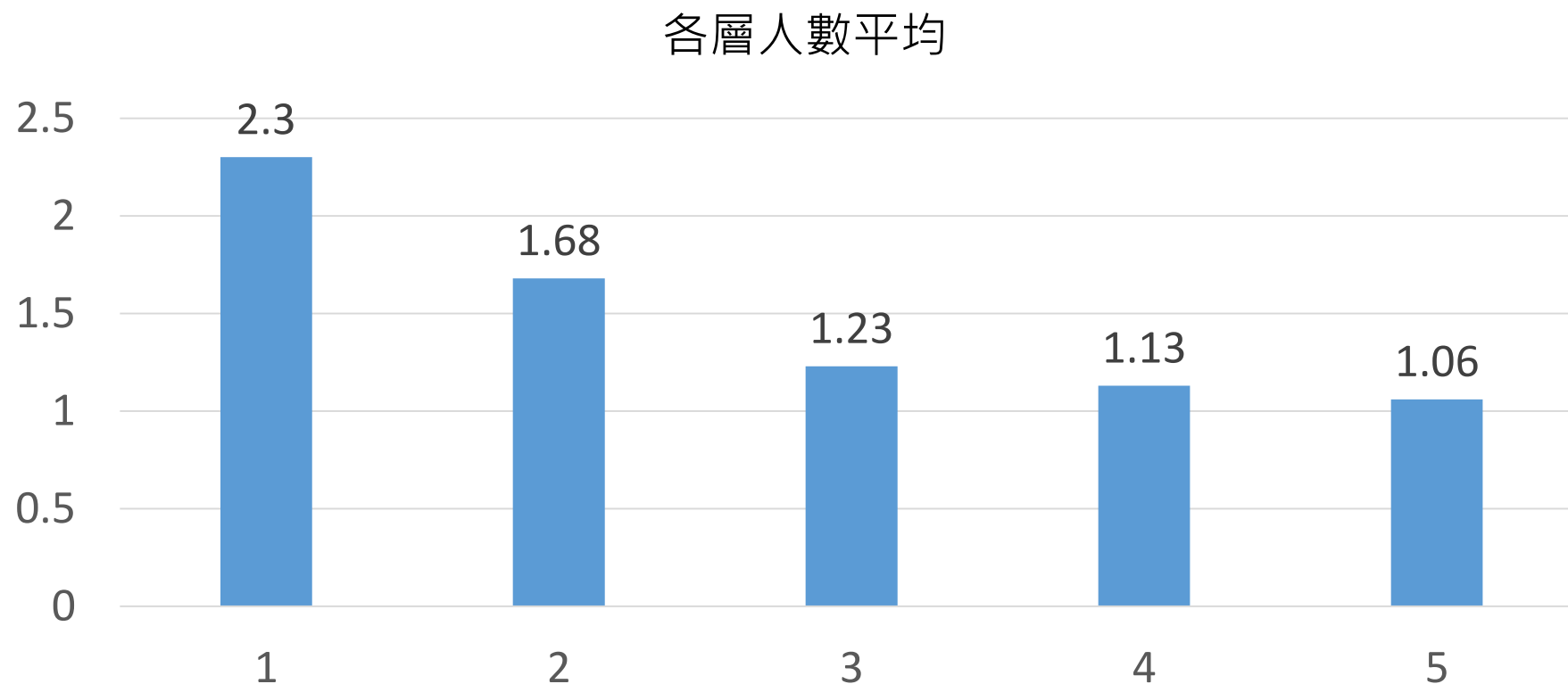


結果呈現(2/8)



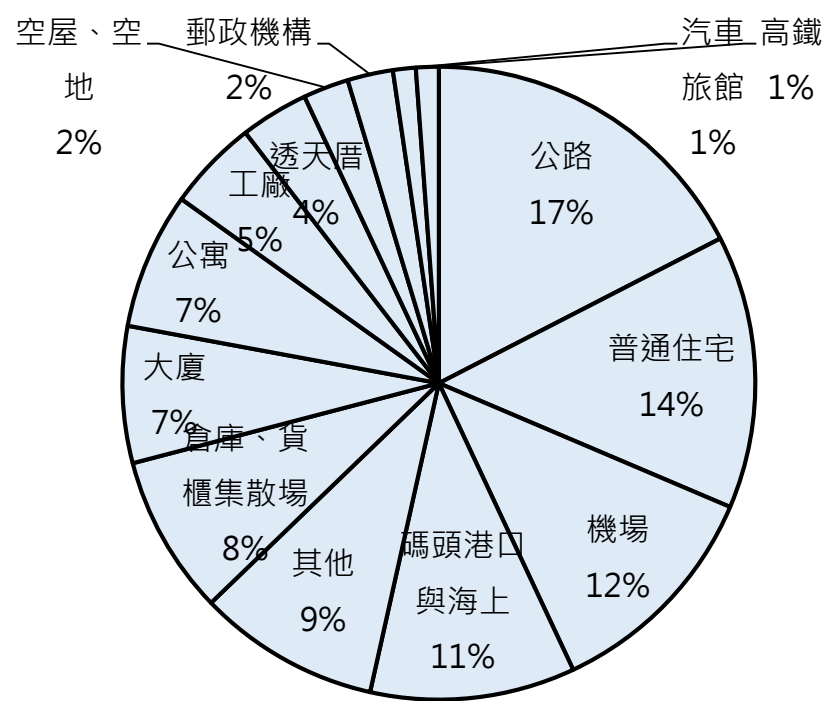
結果呈現(2/8)

另外我們想了解各層嫌疑犯的人數，所以作了以下的呈現。

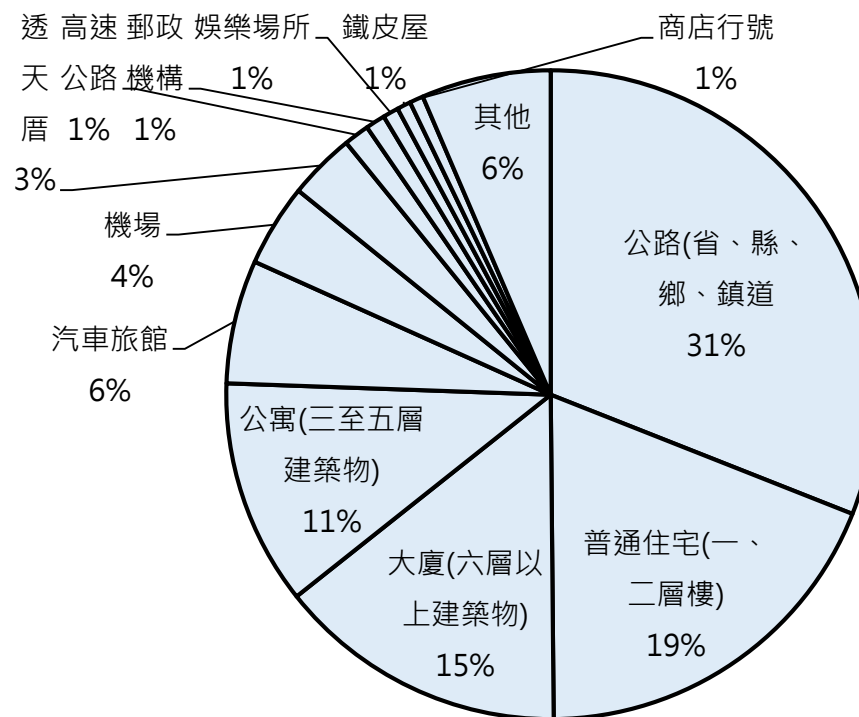


結果呈現(3/8)

另外我們也十分好奇各層發生的場所，所以作了以下的呈現。

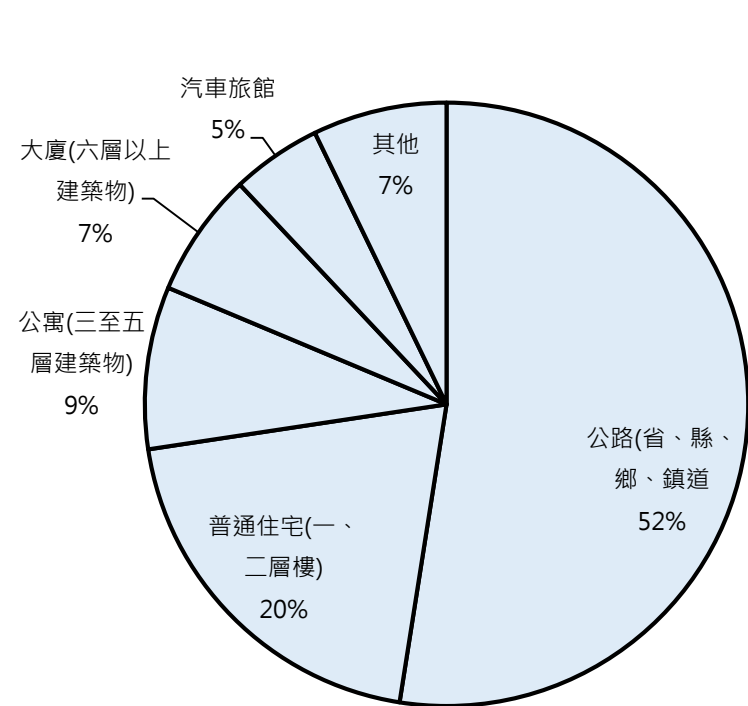


第一層發生場所圓餅圖

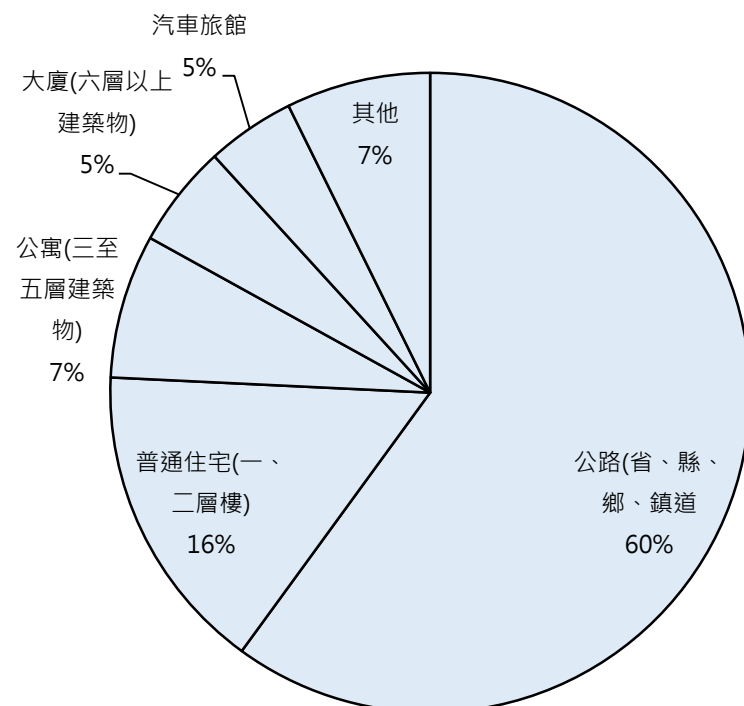


第二層發生場所圓餅圖

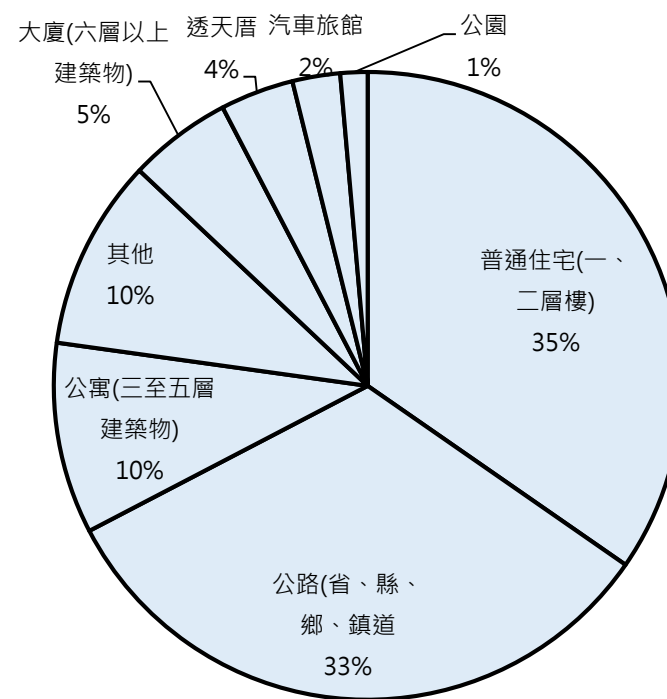
結果呈現(4/8)



第三層發生場所圓餅圖



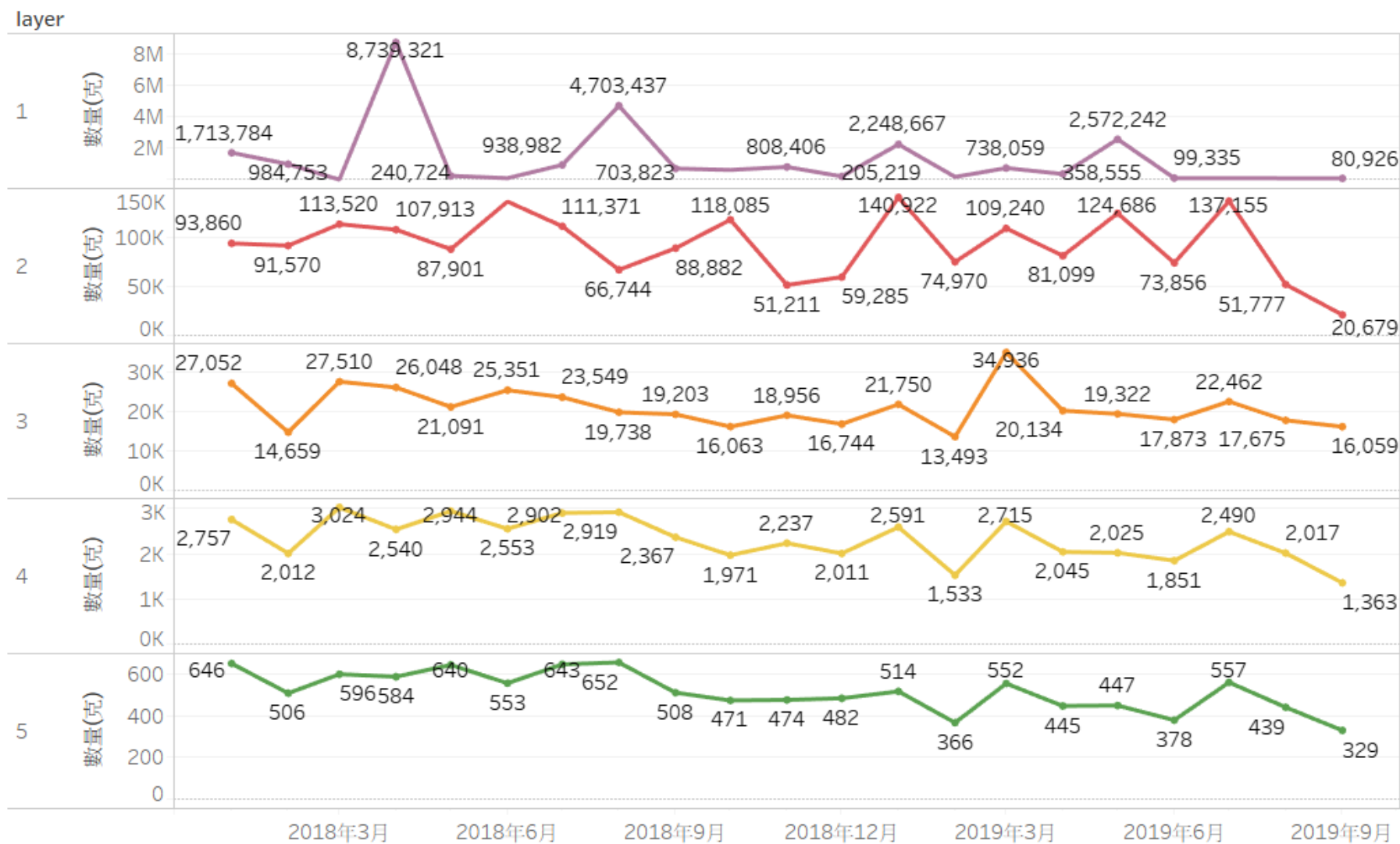
第四層發生場所圓餅圖



第五層發生場所圓餅圖

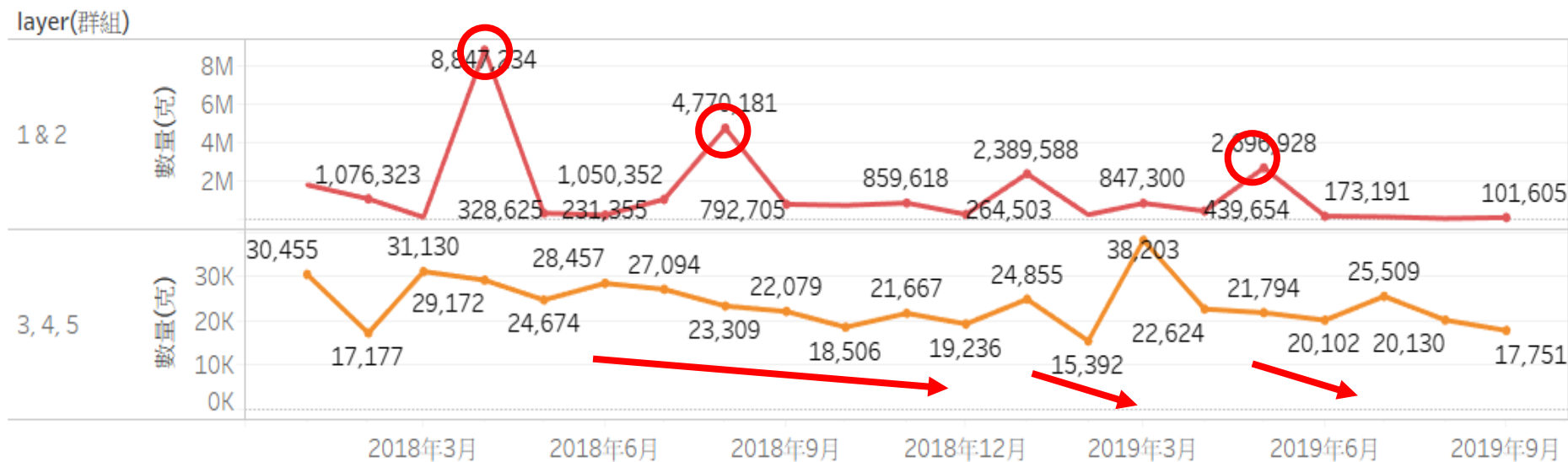
結果呈現(5/8)

為觀察上層毒品數量與下層毒品數量的關係，我們依月分將毒品數量以折線圖呈現。



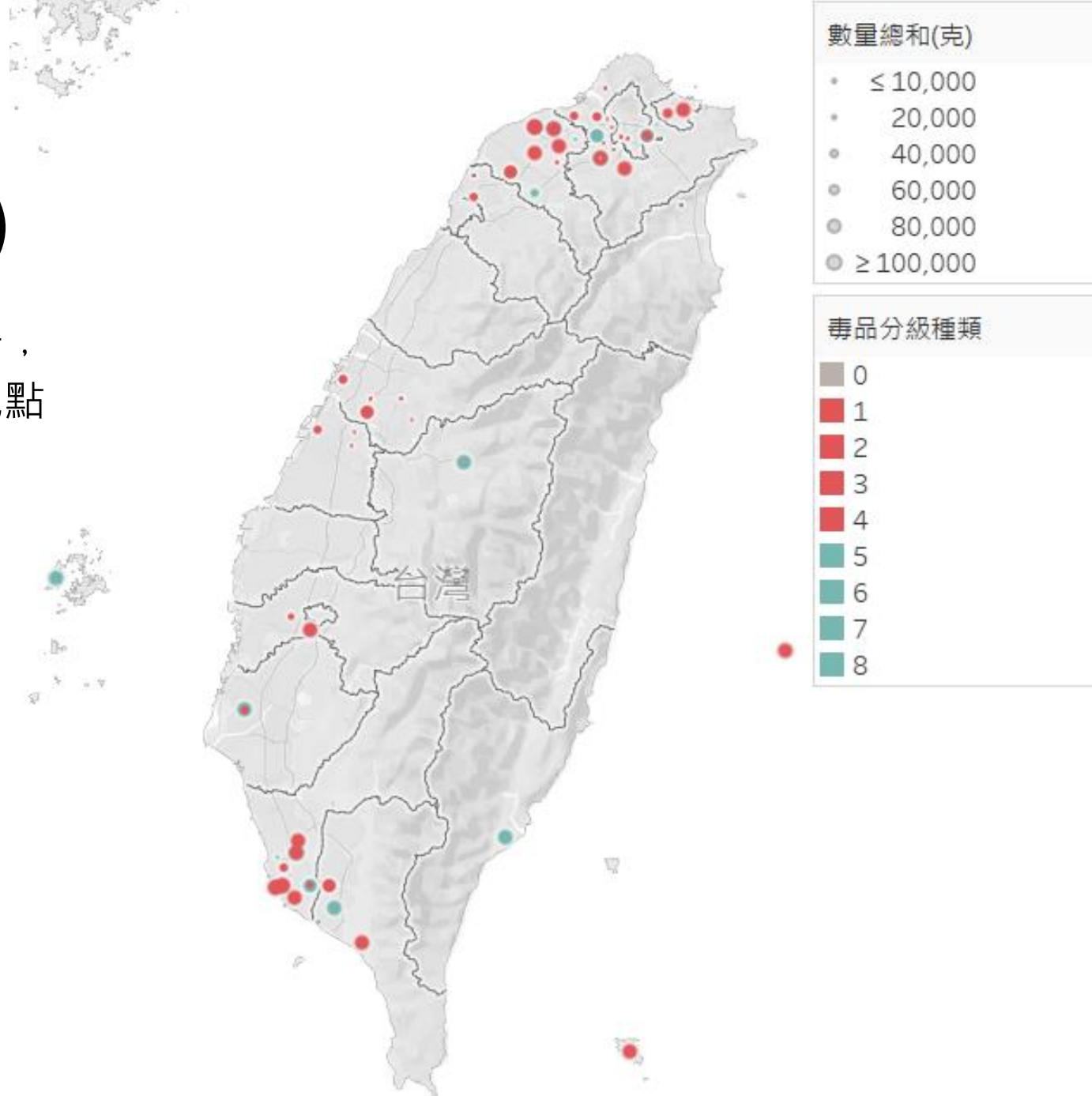
結果呈現(5/8)

我們發現3、4、5層的走勢較為相似，而1、2層與其他層數較不相似，我們將1、2層合併，3、4、5層合併並進行呈現



結果呈現(6/8)

為了解毒品供應與製造的地點，我們將第一層的案件依發生地點進行呈現。

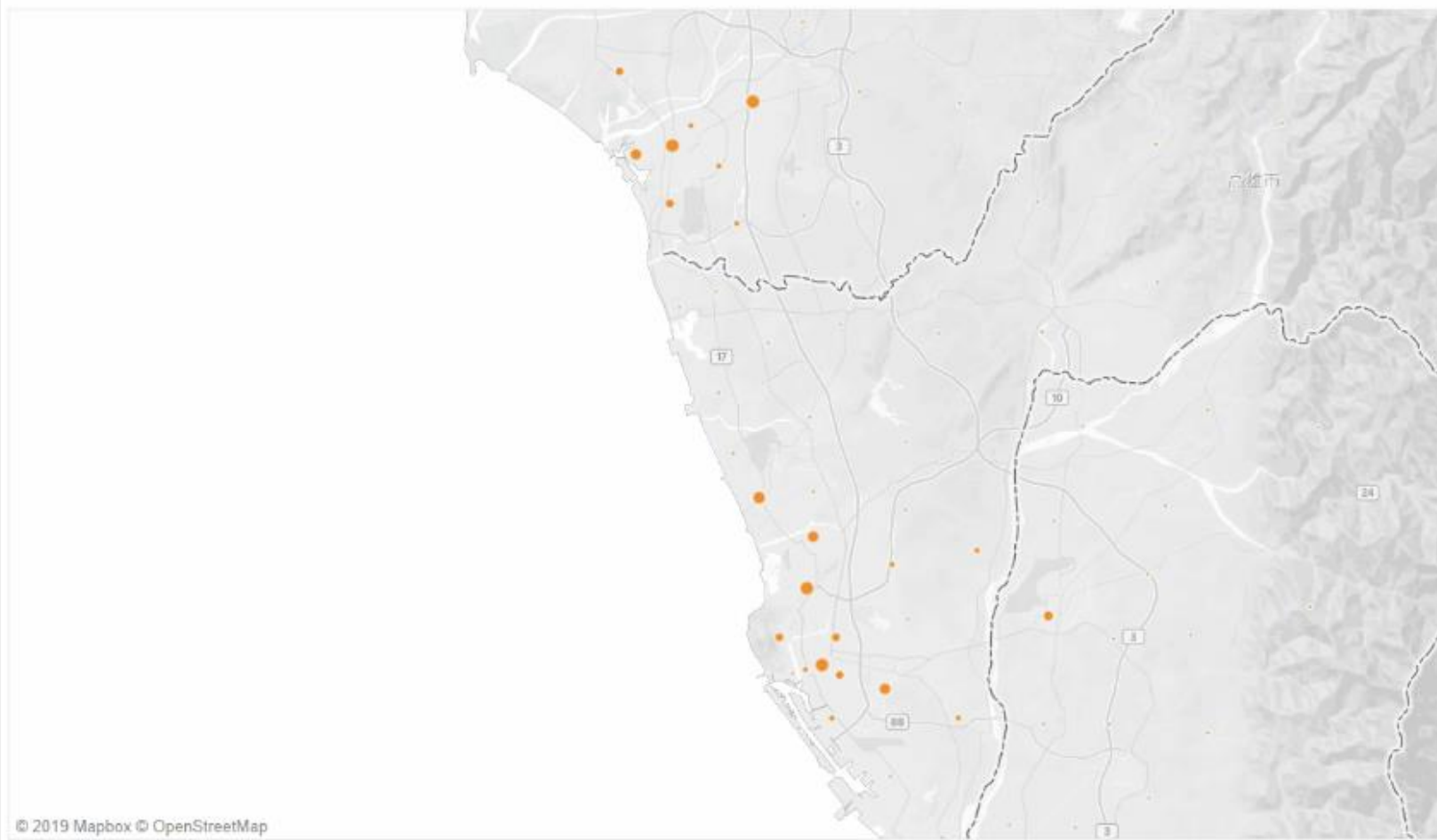


結果呈現(7/8)

在觀察地區性的分布之後，我們觀察到在人口眾多的鬧區3、4、5層的毒品數量會隨著月份有震盪的現象，我們猜想可能是警察在抓獲毒品之後的一、兩個月後該地區的毒品數量確實會減少，但過一段時間之後又會增加。以下是以高雄地區為例進行視覺化呈現。

結果呈現(7/8)

3+4+5南區毒品分布 - 2018年1月



MONTH(發生日期)

< 2018年1月 >

○

◀ ▶

顯示歷程記錄

數量(克)

• 0

● 500

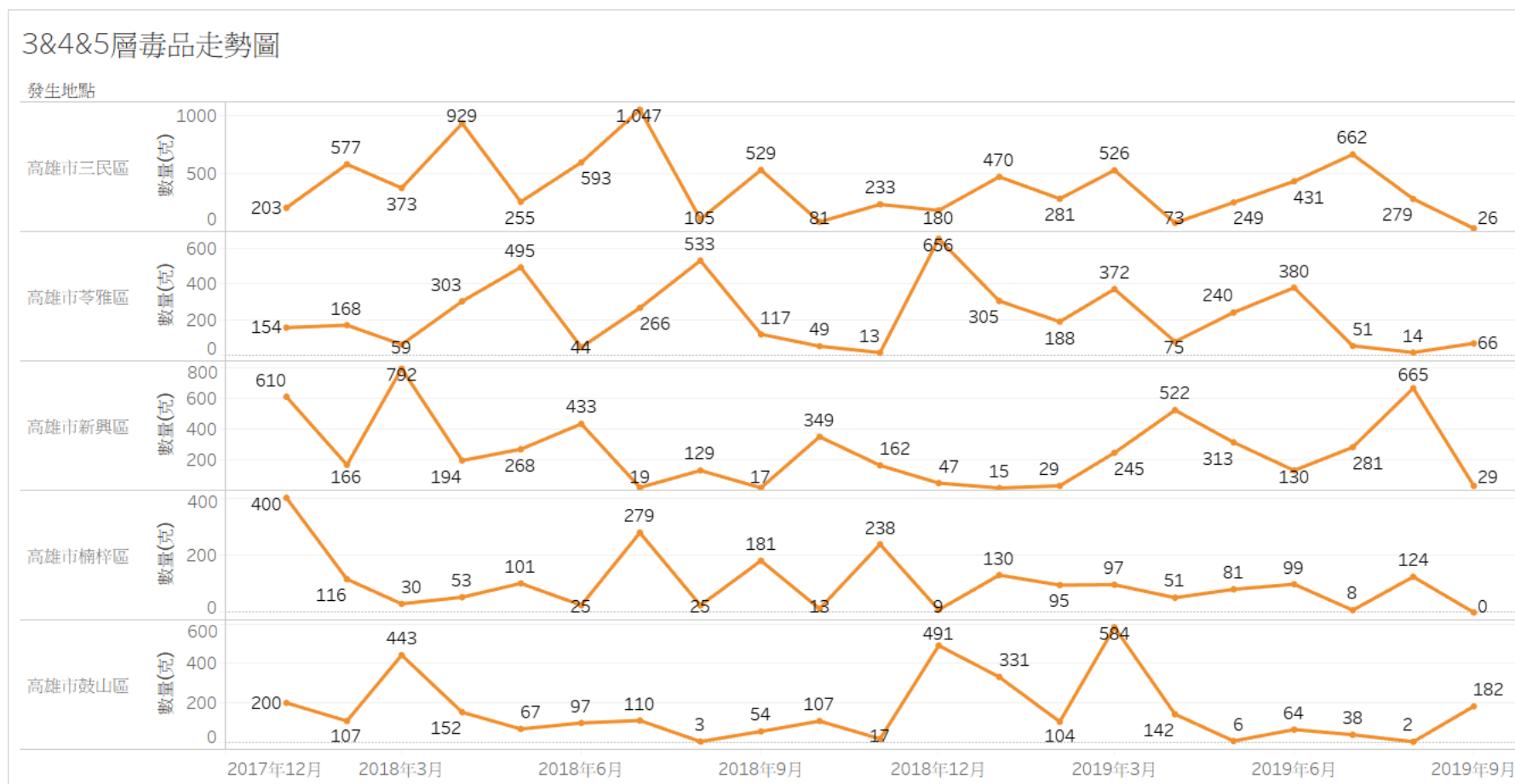
● 1,000

● 1,500

● 2,172

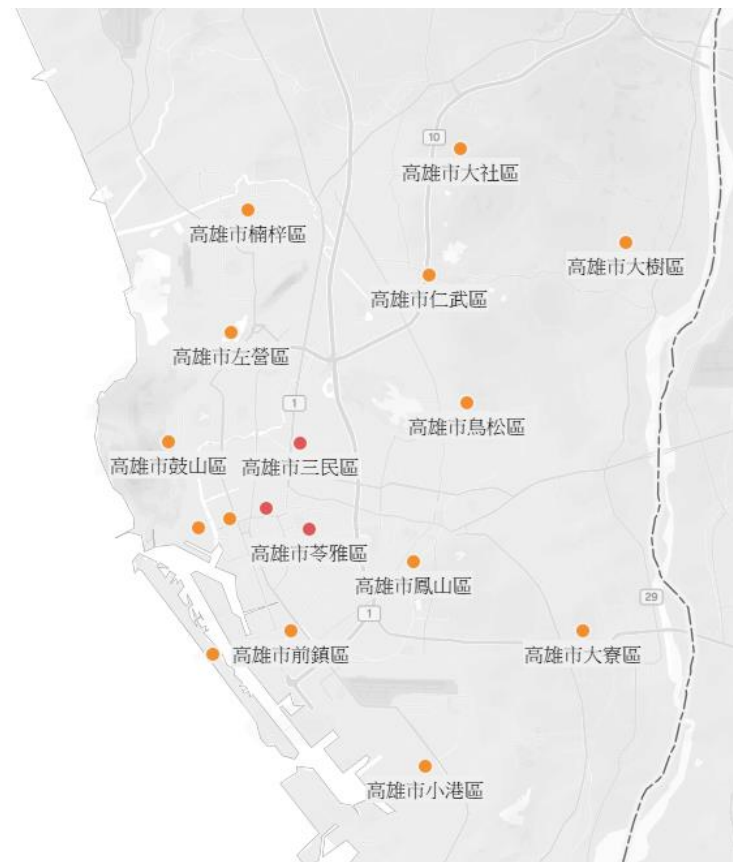
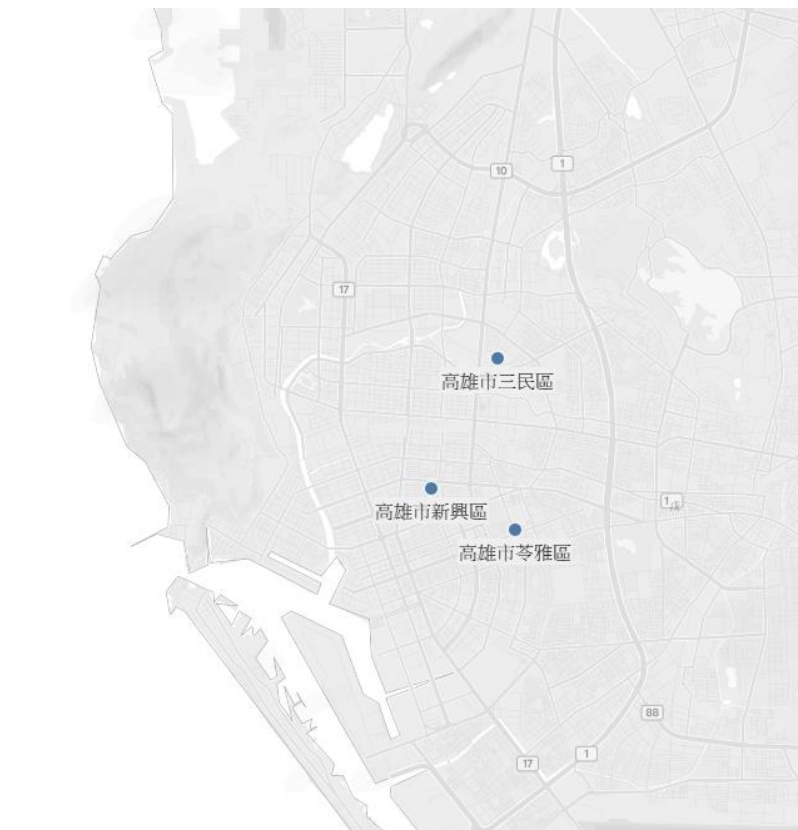
結果呈現(7/8)

由於在地圖上觀察較不明顯，我們將其中幾個行政區取出，依月分對毒品數量3+4+5層用折線圖進行呈現。



結果呈現(8/8)

由於觀察到上述現象，我們希望觀察鬧區與周遭地區的毒品數量的趨勢。
以下以高雄市三民區、新興區、苓雅區為例。

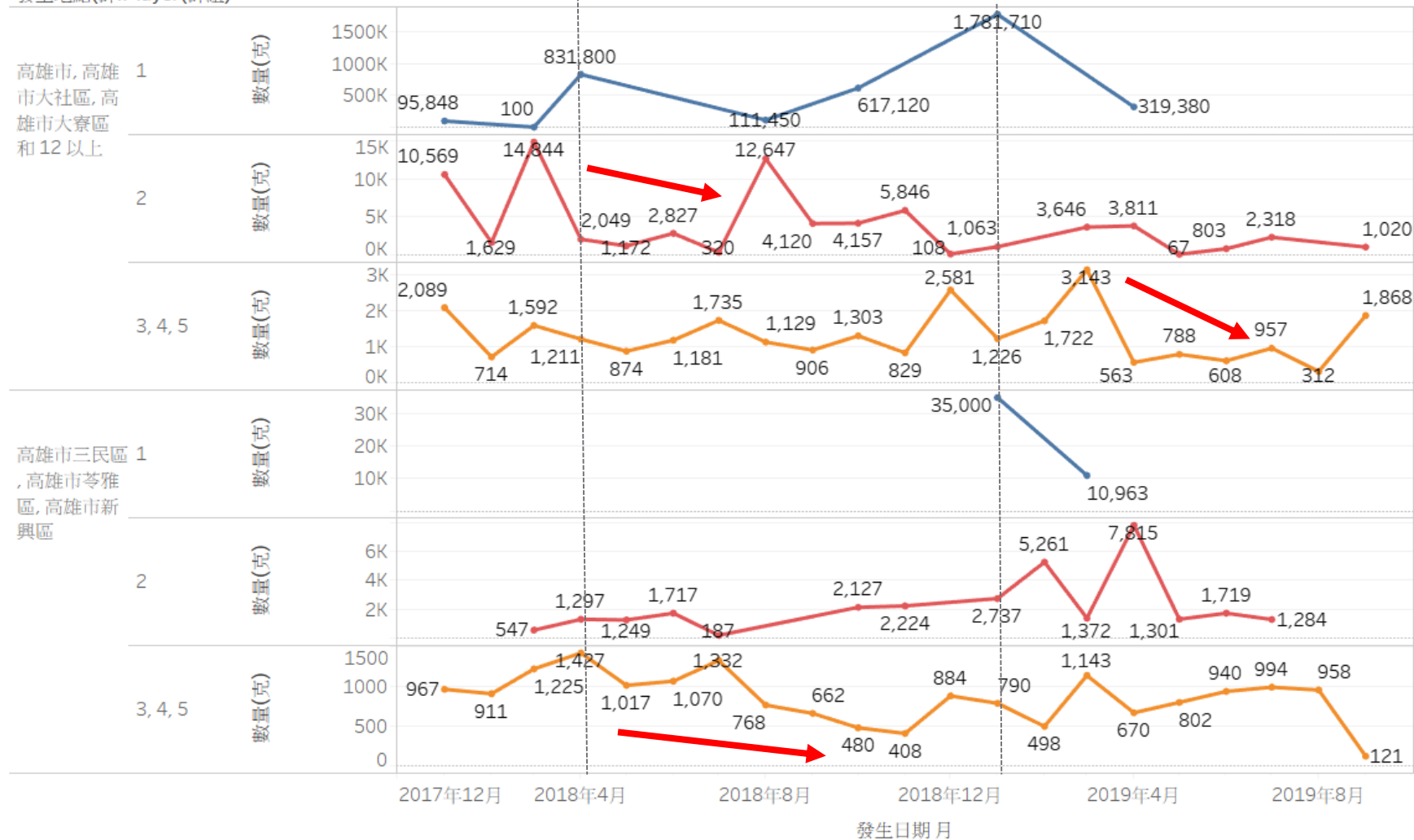


結果呈現(8/8)



結果呈現(8/8)

發生地點(群.. layer(群組))



結論與未來展望

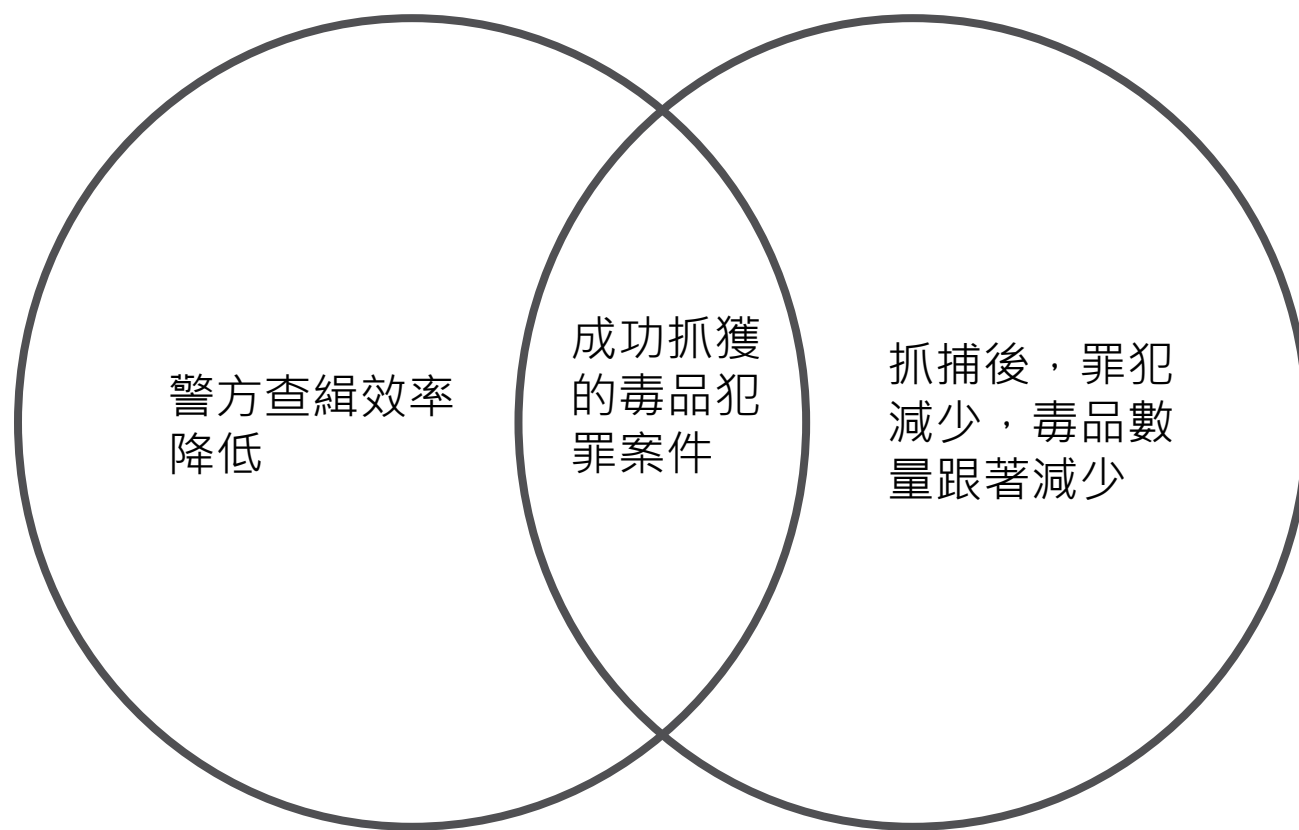
結論與未來展望(1/3)

藉由這次的分析，我們發現了毒品犯罪分布存在三項特性：

- 週期性：同一地點的犯罪數量會隨著時間的推移，一旦某個月捕獲的數量提升，下個月便會明顯下降，而後又上升，形成一個規律、近似於弦波的曲線。
- 流動性：每個月查緝到的大宗毒品地點皆不盡相同。此消彼長，只要打擊某一區域的毒品集團，另一區域便會崛起持續供應下游吸毒人口
- 擴散性：毒品從上層擴散到下層，但只要上層遭受打擊，下層多少也會遭受波及。

結論與未來展望(2/3)

對於毒品犯罪的週期性，我們有以下的猜想：



結論與未來展望(3/3)

為了有效地遏止毒品犯罪的氣焰，我們根據觀察總結出以下幾點：

- 根據統計數據改變巡邏重點
- 提前預估，並依照此預測分配臨檢地點與排班方式。
- 在大數據的幫助下，可以將警力合理地分配於各個地區。

參考資料(1/2)

資料:

- 107年1月~108年9月毒品犯罪資料
- 行政區經緯度資料
- 107年各縣市行政區人口密度統計資料

參考資料(2/2)

網站:

- 政府資料開放平台

[https : //data.gov.tw/](https://data.gov.tw/)

- 全國法規資料庫

[https : //law.moj.gov.tw/LawClass/LawAll.aspx?PCode=C0000008](https://law.moj.gov.tw/LawClass/LawAll.aspx?PCode=C0000008)

- 反毒大本營

<https://antidrug.moj.gov.tw/mp-4.html>

書籍:

- 課堂講義