Clustering Application Using DBSCAN for 911 Hotspot analysis

1. 資料描述

緊急事故 (911) 電話的資料，總共有三種：火災 (Fire)、車禍 (Traffic) 與其他緊急事件 (EMS) ，電話是賓州 Montgomery 州政府的開源資料

Acknowledgements: *Data provided by montcoalert.org*

Data Set Characteristics	Multivariate
Attribute Characteristics	Real
Associated Tasks	Cluster
Number of Instances	150416
Number of Attributes	6
Missing Values	6
Area	Government
Date Donated	Real time updated

2. 資料屬性與類別

lat : String variable, Latitude
lng: String variable, Longitude
desc: String variable, Description of the Emergency Call
zip: String variable, Zipcode
title: String variable, Title
timeStamp: String variable, YYYY-MM-DD HH:MM:SS
twp: String variable, Township
addr: String variable, Address
e: String variable, Dummy variable (always 1)

3. 資料前處理

觀察資料分布與特性時需要進行資料前處理：

(1) 將描述(desc)、地址(addr)與虛擬變數(e)移除。

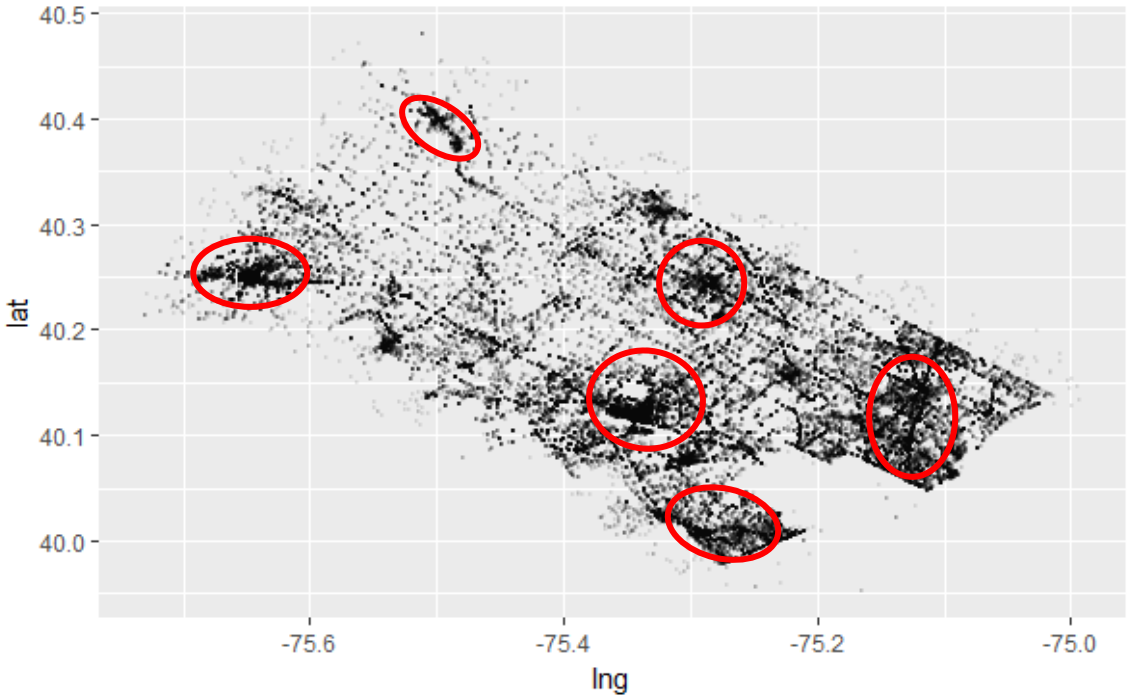
(2) 將時間(timestamp)轉換成時間格式，在分成兩個欄位，分別是日期(date)與時間(time)。

(3) 將緊急事件種類(title)分成主要種類(category)與副種類(subcat)。

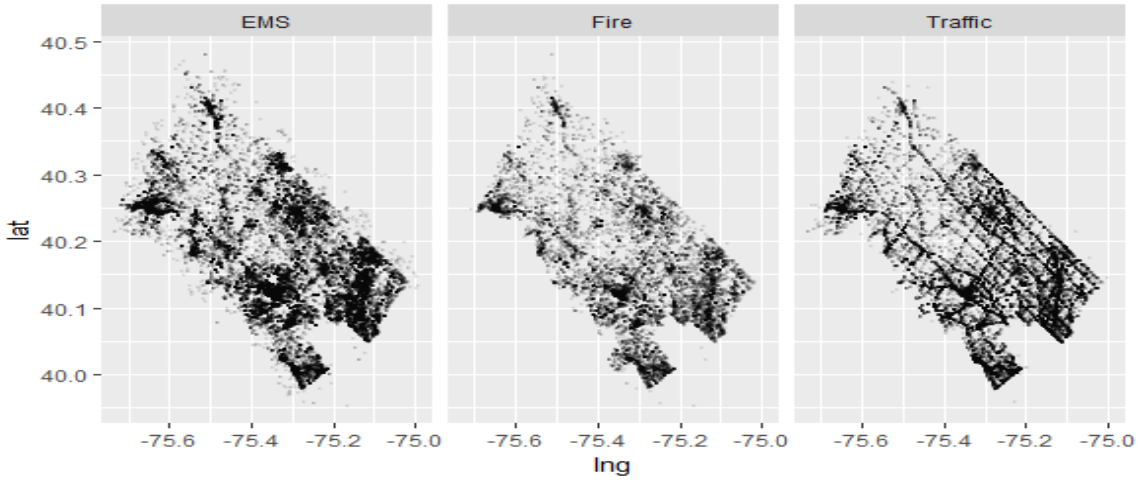
(4) 去除經緯度的離群值，我的方法是將經緯度資料統計四分位距，個別超過四分位距的兩倍為離群值。

(5) 依緊急事件種類分三類，分別是 EMS、Traffic 與 Fire

以所有空間資料畫出散布圖，可以用肉眼觀察到密集之處，為可能的熱點位置。



以類別分成三的圖來比較，密度高的區域大致有相同的結果。

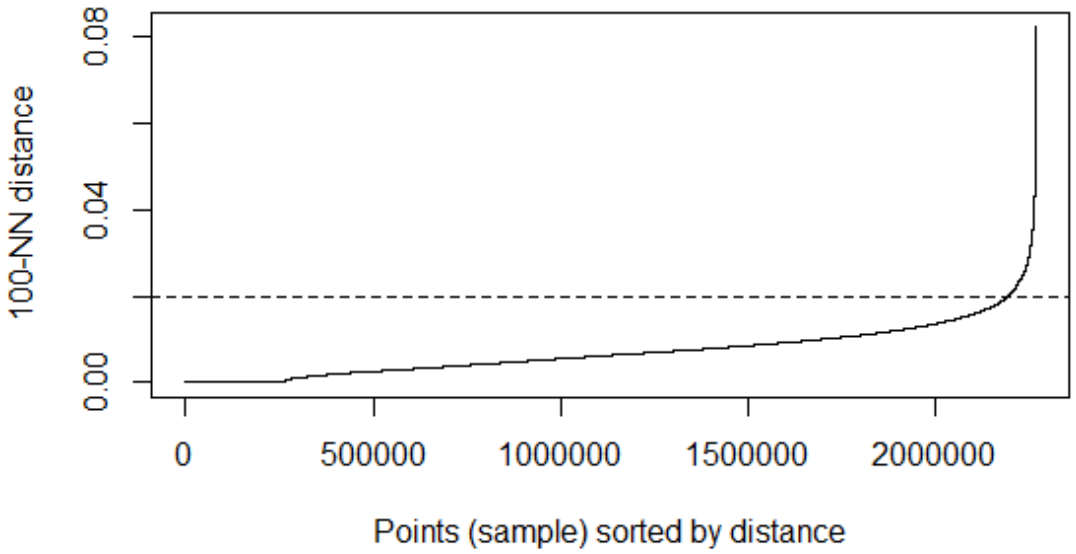


4. 系統目標

找出各個緊急事件種類發生次數頻繁的「熱點」，評估分群效果並分析結果背後可能的現象，對於這些熱點加強人員部署，或是警力支援。

5. DBSCAN 分群方法

執行 DBSCAN 時，需要決定兩個係數，半徑 (EPS) 與最少點數 (MinPts) ，因此選擇最適合的係數是十分重要的。經過研究，有兩種可行的執行 DBSCAN 分群方法，其一是依照 domain knowledge 設定一數 k 為資料中每一點與其最近的 k 個點的平均距離，並畫出 k-distance 圖，找出斜率急劇變化的點，對應到的平均距離設為 EPS。以火災類別的緊急電話為例，若將 k 設為 100，畫出 k-distance 圖，把 EPS 設為 0.02 將是一個合理的參數，見下圖。



而另一個方法是依照 domain knowledge，固定一個合理的 EPS，在決定最少點數，決定點數方法有一可行方法與前面相似：以每一點為中心，以先設定好的 EPS 為半徑畫圓，統計點的個數並且作密度圖，觀察圖中急劇變化的點數當作 MinPts。

最後，我選擇以固定 EPS 參數，改變 MinPts 參數。評量方法為在分一樣的群數下，計算 SSE，以不同的 MinPts 參數下計算群內點跟點的平均距離為多少。另外，計算不同群數下的 SSE 變化，是否與認知相同。

6. Code

以 R language 執行資料前處理、開發 dbscan 分群方法，以分群類別畫分布圖，計算平均點與點的距離、還有依經緯度畫分區域，定位密集事故的區域。

(1) 資料前處理、個分布圖: dmprjt_pre.R

(2) Dbscan 分群，以火災資料設定 Minpts = 400 為例: 400minpts_fir.R

(3) SSE 統計圖表: SSEplot.R

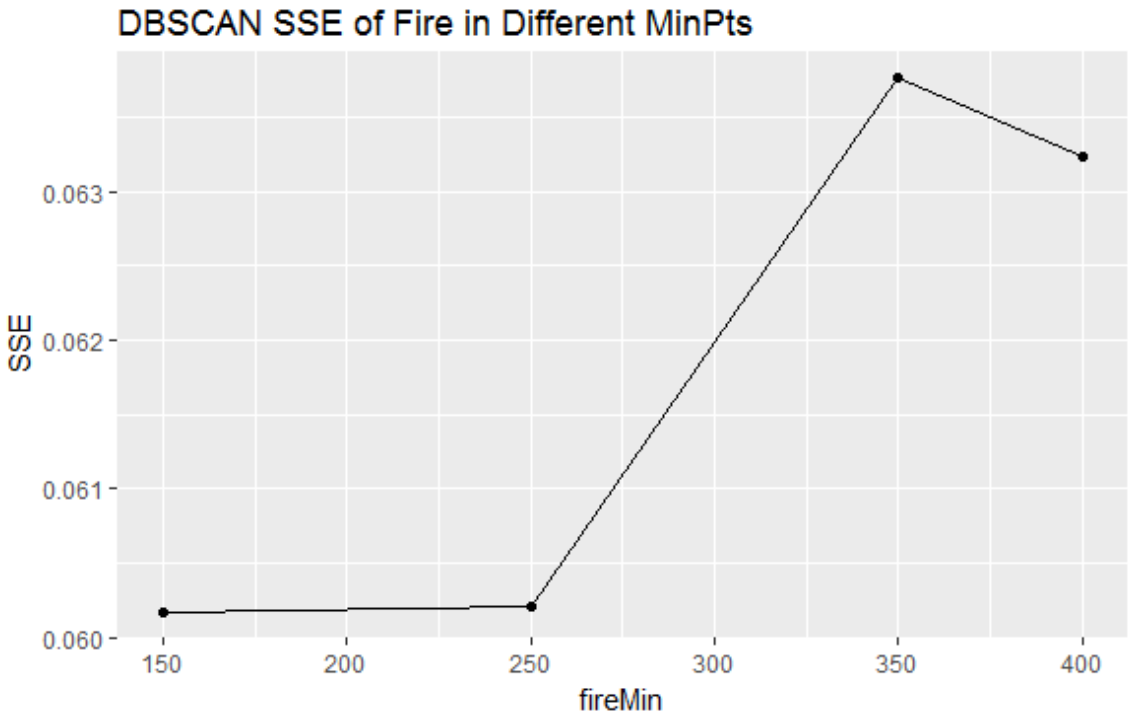
(4) 以區域化分的密度矩陣: DensityMatrix_ll.R

7. 結果與討論

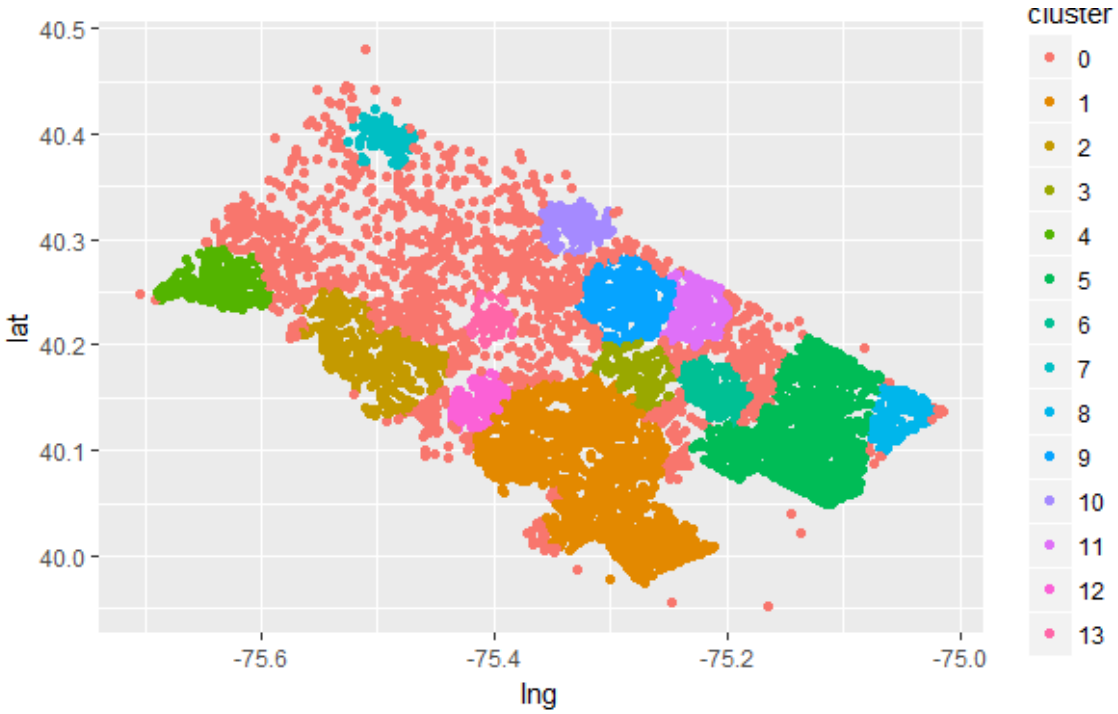
以 Fire、EMS、Traffic 個別呈現探勘結果。再將 SSE 最好的分群畫分布圖，可以發現在實驗前密度高的區域都有被覆蓋到。

將探勘結果整理成表格，可以發現 Minpts 越小分出來的群數越多，在每群中平均點數也越少，而平均 SSE 大致呈現越多群越小的趨勢。因為計算的是點跟點平均距離，因此分群結果看起來越小群 SSE 越小，與當初想的一樣。

Fire

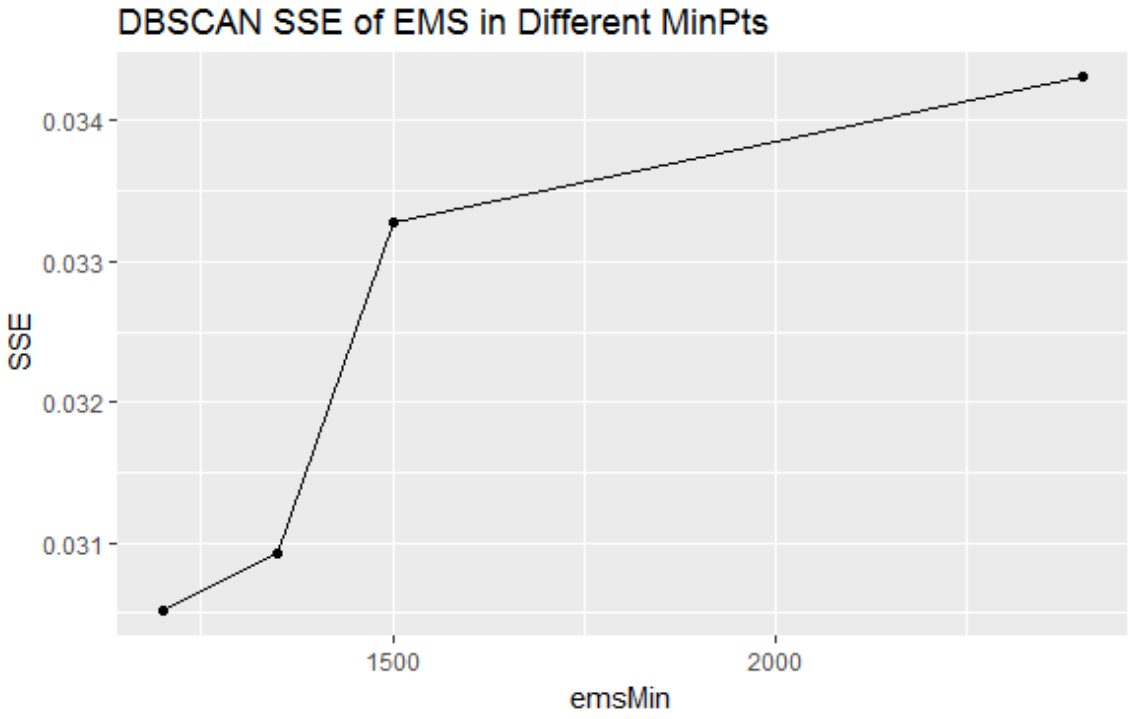


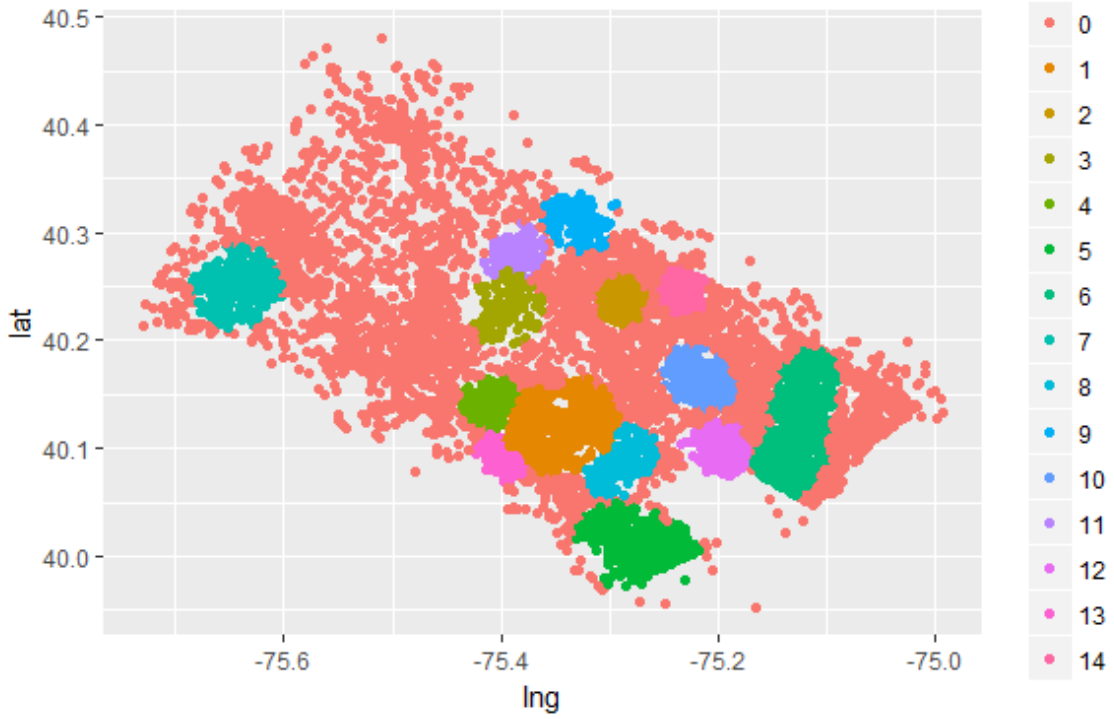
Parameters	Value of input parameters	Evaluation Criteria		
		No. of Clusters	Average number of points per cluster	Average SSE
EPS, Minimum number of points	EPS=0.021728,MinPts=400	5	3012	0.06323832
	EPS=0.021728,MinPts=350	7	2426	0.06376638
	EPS=0.021728,MinPts=250	11	1685	0.06021208
	EPS=0.021728,MinPts=150	13	1638	0.06017087



EMS

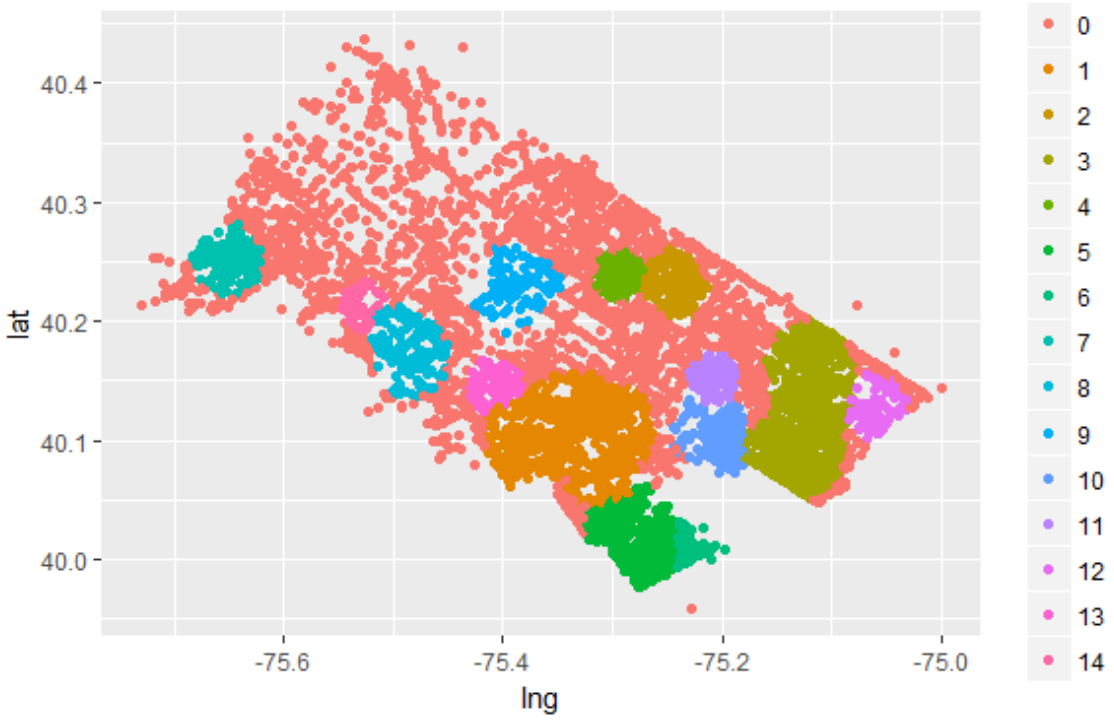
Parameters	Value of input parameters	Evaluation Criteria		
		No. of Clusters	Average number of points per cluster	Average SSE
EPS, Minimum number of points	EPS=0.021728,MinPts=2400	5	6454	0.03431355
	EPS=0.021728,MinPts=1500	7	5695	0.03327764
	EPS=0.021728,MinPts=1350	11	4058	0.03092017
	EPS=0.021728,MinPts=1200	14	3606	0.03052173


DBSCAN SSE of EMS in Different MinPts
SSE
0.034
0.033
0.032
0.031
1500
2000
emsMin


lat
40.5
40.4
40.3
40.2
40.1
40.0
-75.6
-75.4
-75.2
-75.0
lng
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14

Traffic

Parameters	Value of input parameters	Evaluation Criteria		
		No. of Clusters	Average number of points per cluster	Average SSE
EPS, Minimum number of points	EPS=0.021728,MinPts=1400	4	6636	0.04689432
	EPS=0.021728,MinPts=1100	7	4399	0.04321717
	EPS=0.021728,MinPts=1000	9	3943	0.04213072
	EPS=0.021728,MinPts=800	13	3068	0.03981473



另外，我想到另外一種偵測密度的方式，就是將所有空間分成等分空間計算在那格有幾個 911 電話，邊長是用前面算的 EPS，計算一個電話數的矩陣。見下圖可以發現幾個密度十分高的區域。(以火災資料為例。)

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17
1	0	0	0	2	4	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	1	26	228	33	0	0	0	0	0	0	0	0	0	0	0
3	0	0	5	20	73	126	25	19	3	0	0	0	0	0	0	0	0
4	0	38	172	38	27	60	40	39	288	72	0	0	0	0	0	0	0
5	47	266	141	72	37	95	80	227	132	148	94	4	0	0	0	0	0
6	147	860	318	148	223	166	65	235	216	618	411	174	17	1	0	0	0
7	0	0	0	338	333	270	105	71	76	257	217	180	151	388	73	0	0
8	0	0	0	0	112	276	269	371	434	382	212	485	203	819	237	170	1
9	0	0	0	0	2	33	215	646	1327	580	347	403	525	890	347	79	0
10	0	0	0	0	0	0	7	98	211	769	117	80	173	582	152	0	0
11	0	0	0	0	0	0	0	24	382	915	681	134	1	1	0	0	0
12	0	0	0	0	0	0	0	0	2	96	158	0	1	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

8. 心得

這次的專題讓我更瞭解資料探勘的實際操作，總體而言，探勘結果大致滿意，然而因為資料量龐大，R 在處理這麼大量的數據有一些限制，儘管我已經盡量使用向量化的運算以減少電腦的負擔，但還是有很大的限制，所以我自己寫的函數跟視覺化的表現沒辦法有非常好的表現，這是美中不足的地方。

在這學期紮實的作業，考試以及報告的洗禮下，我學到非常多知識，我要特別感謝老師耐心的教導，不會因為我是外系，基礎不好的關係而少了一點用心。除此之外，我要感謝同組的學長，常常傳達許多做研究的經驗與想法，幫助我學習。

在未來我仍然會繼續學習資料科學領域的知識，因為因為修了這門課讓我眼界大開，我覺得好玩的東西值得我花時間研究。