分散式資料庫期末報告

WEKA 分析-美國成人普查收入

分群分析

指導老師：蔡正發教授

組員：吳湄潔 B10756014

王諒桓 B10756055

壹、 資料預處理

☞ 研究一 以收入高低為主 ，來分群分析

- **先將文字資料轉換數字代號**

性別：男性是 1，女性則為 2

公司單位：私人為 1、聯邦政府為 2、州政府為 3、地方政府為 4、自僱收入為 5、自僱非收入為 6、無收入工作為 7

- **再進行分群分析，並將此資料集依據欄位(收入)，放於最後欄。**

	A	B	C	D	E
1	性別	年紀	公司單位	每週工作小時數	收入
2	2	82	1	18	<=50K
3	2	41	1	40	<=50K
4	2	54	1	40	<=50K
5	2	34	1	45	<=50K
6	1	38	1	40	<=50K
7	2	68	2	40	<=50K
8	2	74	3	20	>50K
9	2	45	1	35	>50K
10	1	32	1	55	>50K
11	1	34	1	50	>50K
12	1	37	1	40	>50K
13	1	38	6	45	>50K
14	1	45	1	76	>50K
15	1	46	1	40	>50K

☞ 研究二 以婚姻狀況為主，來分群分析

- **首先將文字資料轉換數字代號**

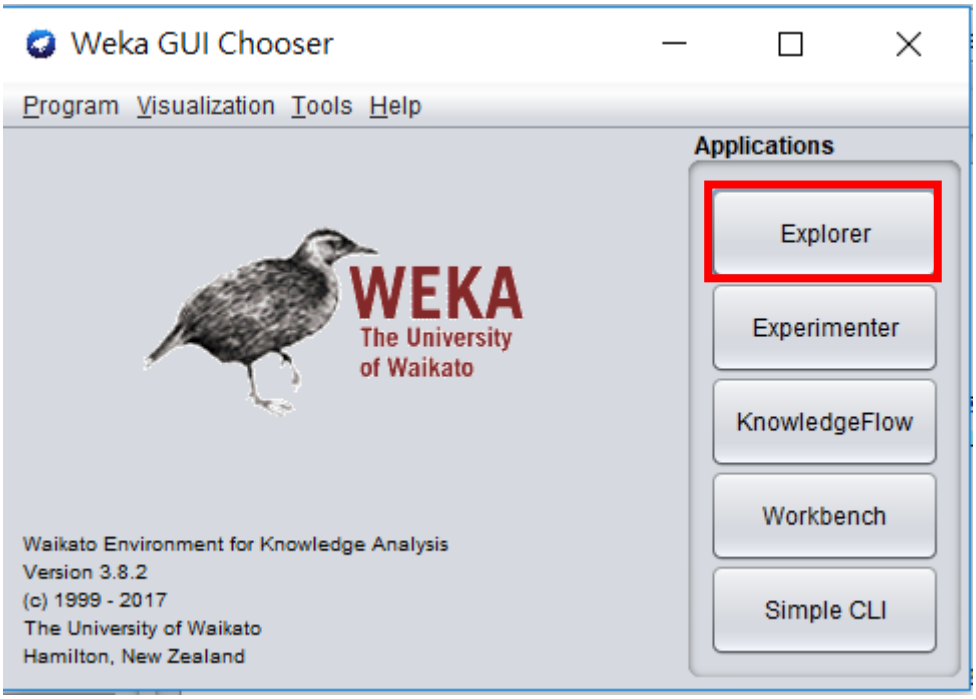
教育編號：未完成學業為１、高中畢業為２、就讀大學中為３、專科為４、高職為５、學士為６、碩士為７、教授為８、博士為９

公司單位：私人為１、聯邦政府為２、州政府為３、地方政府為４、自僱收入為５、自僱非收入為６、無收入工作為７

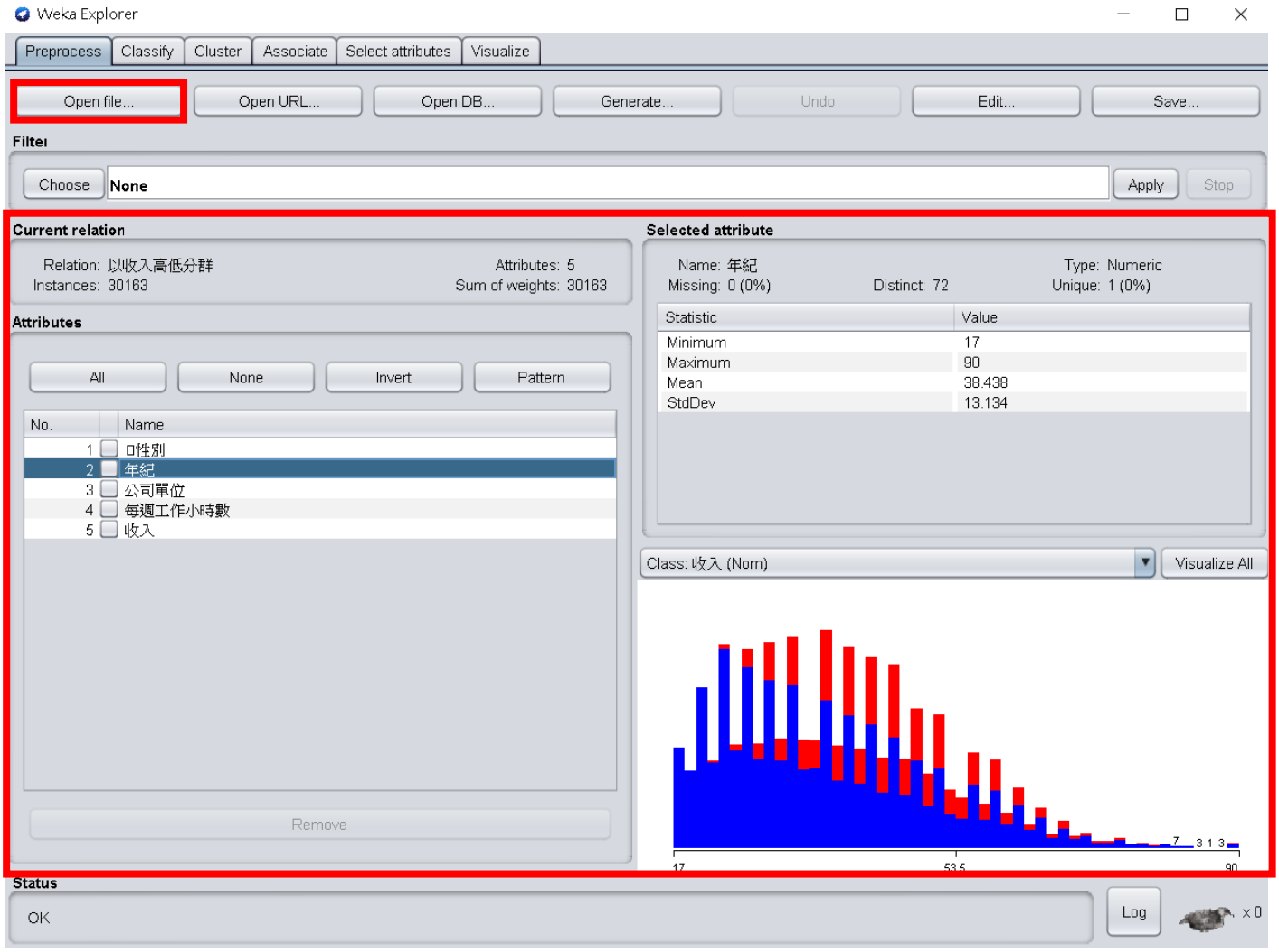
- **再進行分群分析，並將此資料集依據欄位(婚姻狀況)，放於最後欄。**

	A	B	C	D
1	每週工作小時數	教育編號	公司單位	婚姻狀況
2	20	9	3	未結婚
3	35	9	1	離婚
4	20	6	1	守寡
5	60	9	6	未結婚
6	50	8	1	未結婚
7	40	7	1	離婚
8	50	1	5	守寡
9	60	9	2	離婚
10	40	7	1	離婚
11	40	7	1	離婚
12	40	3	1	離婚
13	70	6	1	已婚配偶缺席(無奈分開彼此
14	45	6	1	未結婚
15	40	3	1	未結婚
16	72	2	1	守寡
17	40	2	6	已婚公民配偶
18	65	7	1	已婚公民配偶
19	55	8	5	已婚公民配偶
20	6	3	1	已婚公民配偶

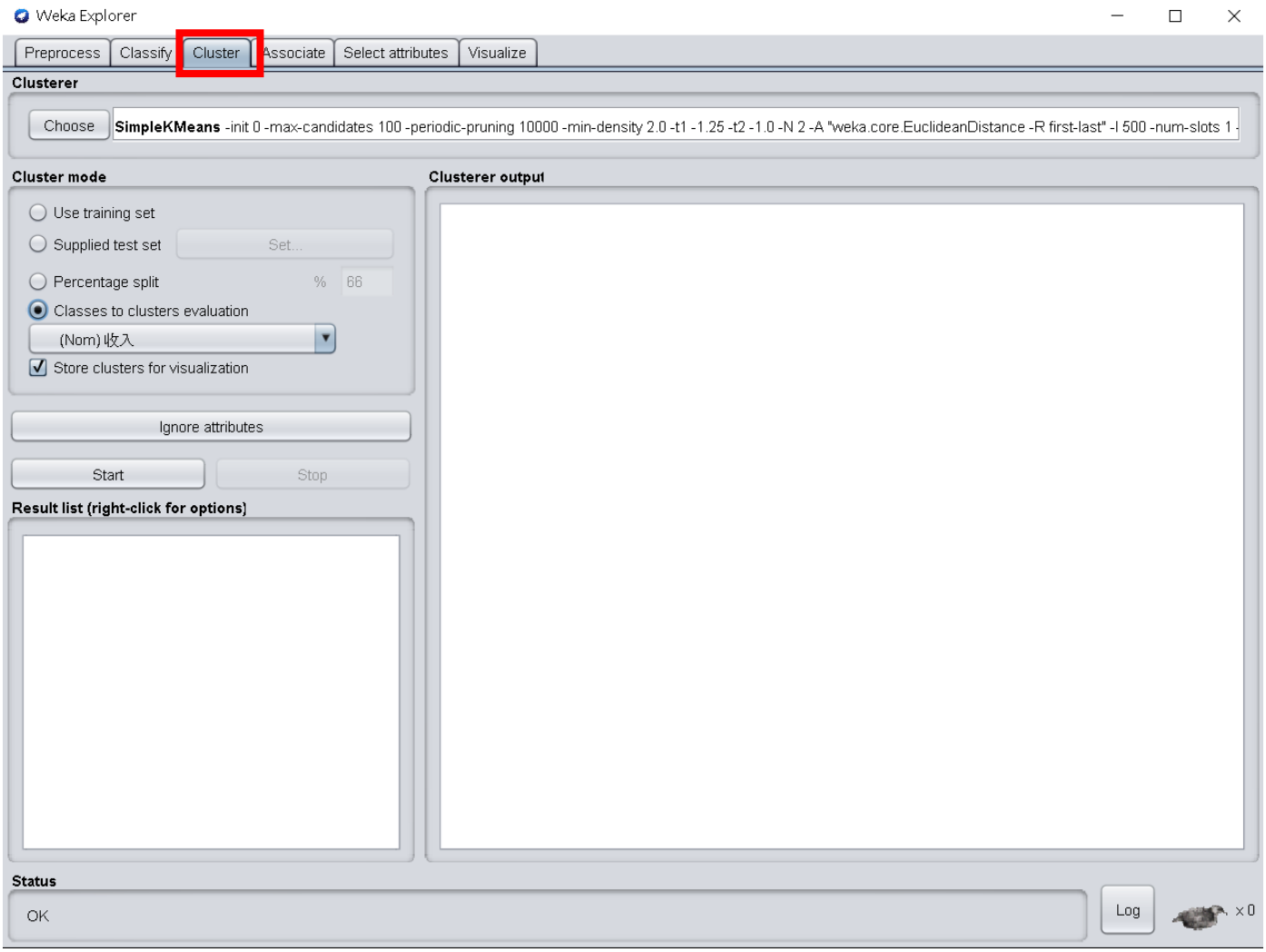
貳、 打開 weka 介面，點擊 **Explorer(探索者)**



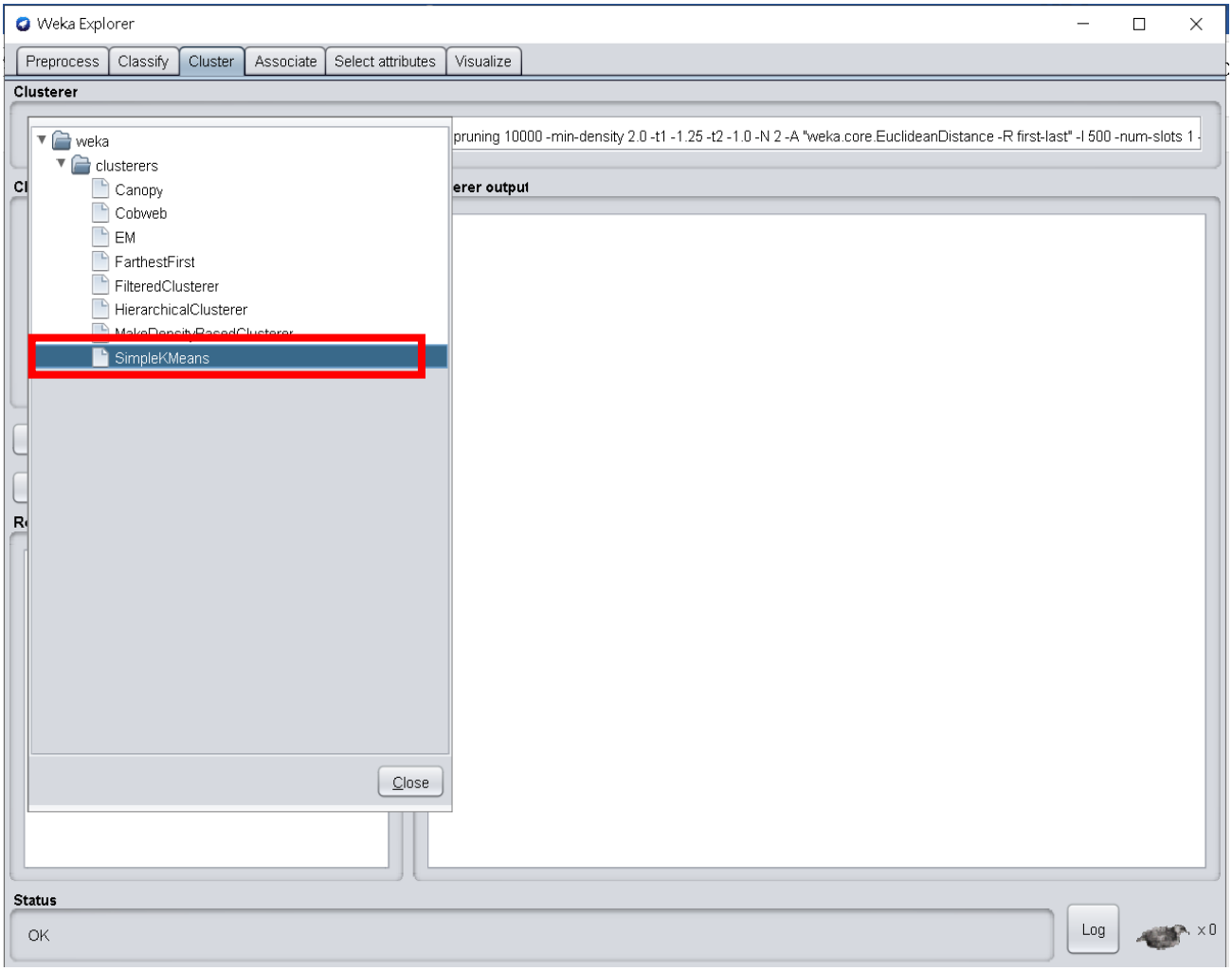
參、 點擊 **openfile** 匯入資料集，資料將會顯示在下面紅框



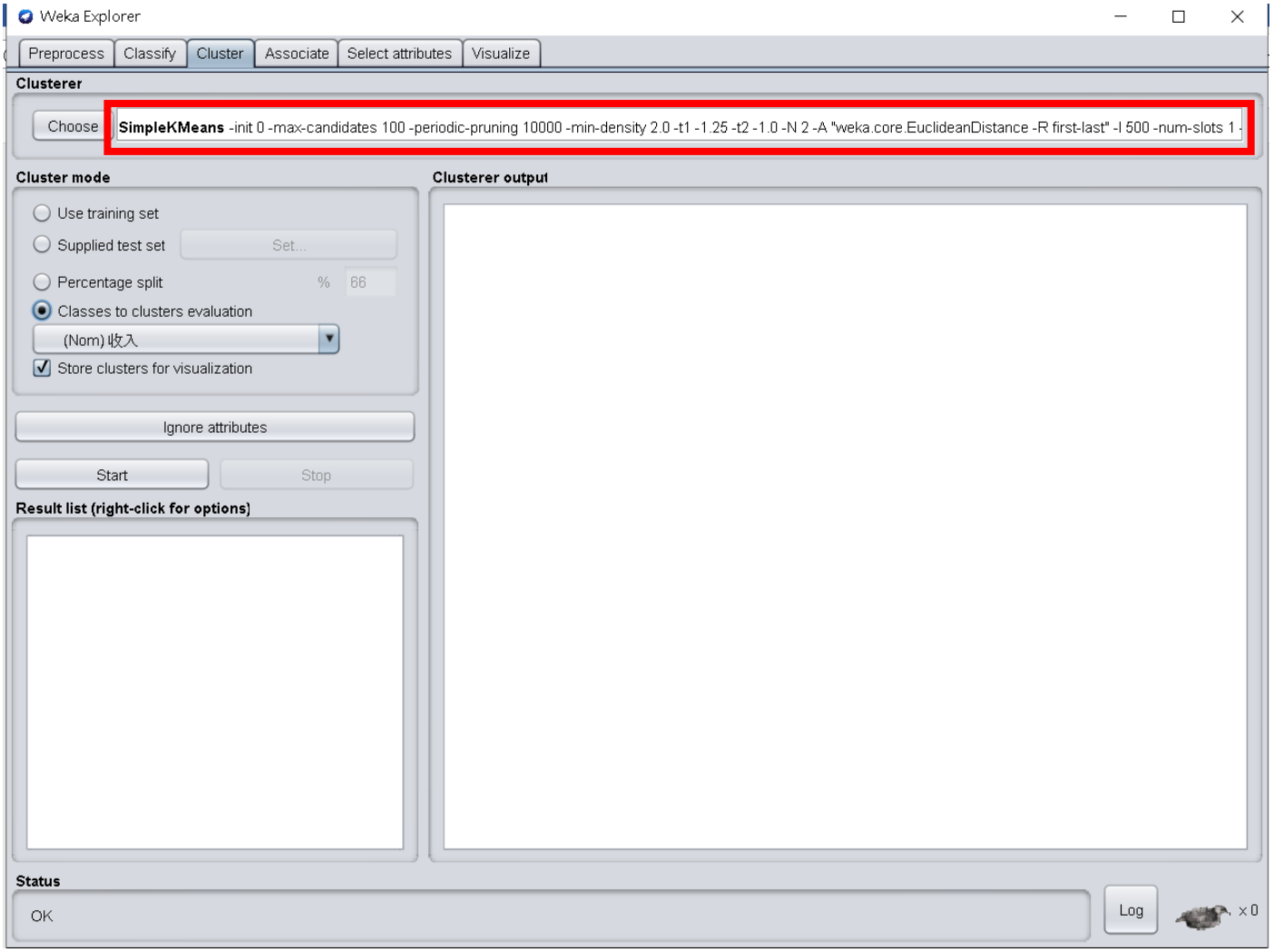
肆、　選擇 Cluster(分群)



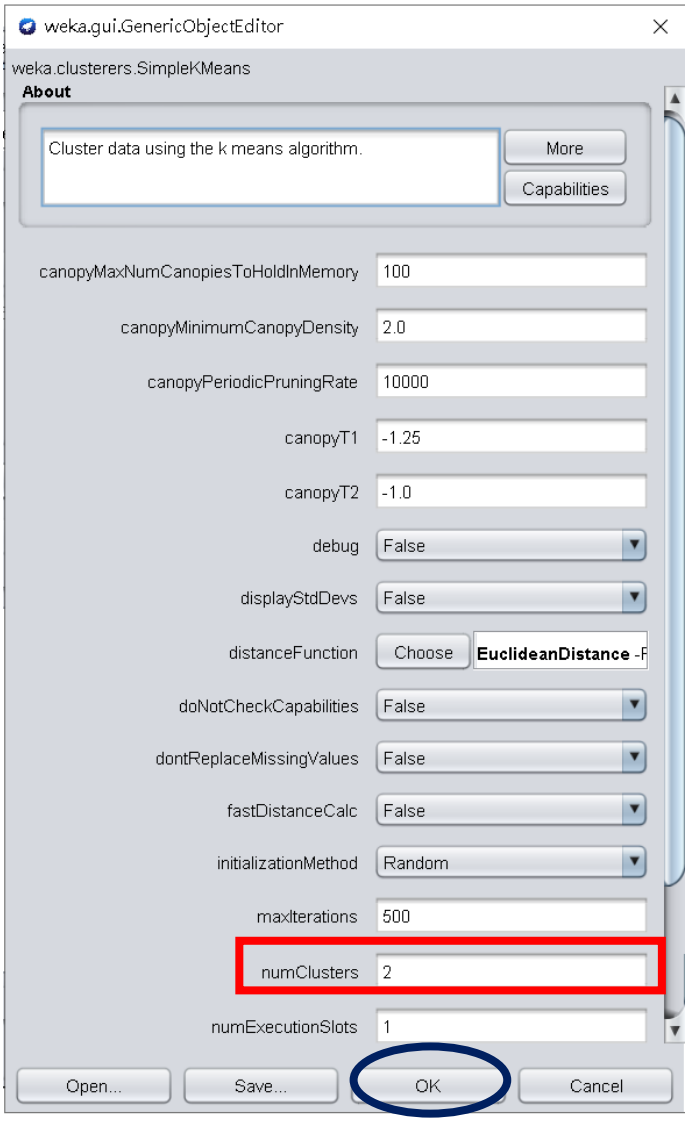
伍、 選擇 SimpleKmeans 演算法



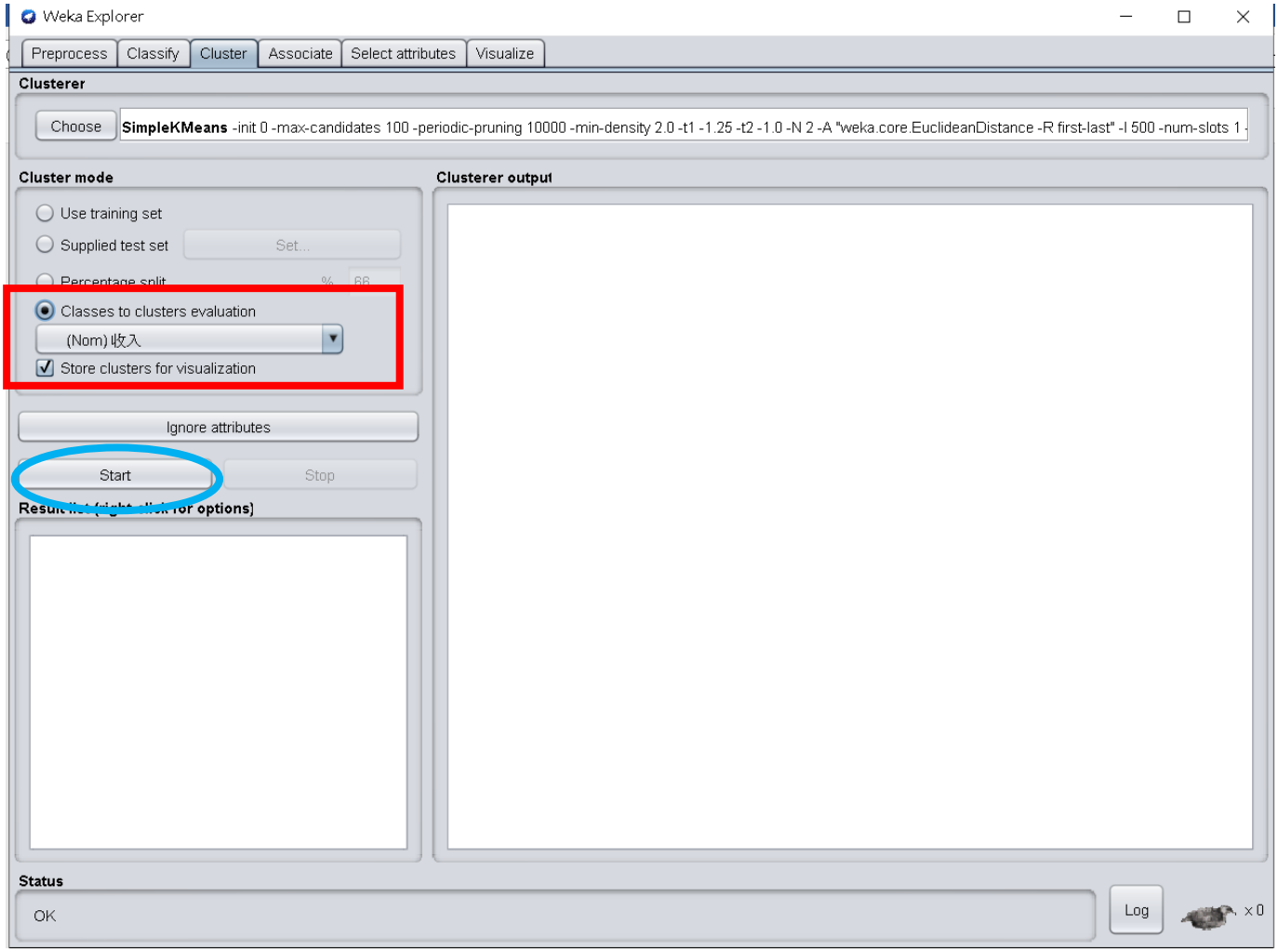
陸、　選擇完演算法後，點擊紅框



柒、　接著設定分群數量(紅框)



捌、　選擇分群的模式，再按下開始 Start 進行資料分析



- 模型選擇
  1. 訓練集　：使用原本的訓練集資料評估
  2. 測試集　：用測試集來預測
  3. 比例切割　：將比例 66%作為訓練集，其他為測試集來評估模型
  4. 用此欄位去分群評估　：檢驗分群結果與此欄位，兩者之間的關聯性強或弱。
- 「 classes to clusters evaluation」：檢驗哪個屬性和收入有較強的相關性。

因此會把「 classes to clusters evaluation」欄位屬性設定為「收入」。

玖、 研究一的分群說明

- **研究一 以收入高低分群，輸出結果**

20:07:06 - SimpleKMeans

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A
Relation:     以收入高低分群
Instances:    30163
Attributes:   5
              ㄖ性別
              年紀
              公司單位
              每週工作小時數
Ignored:
              收入
Test mode:    Classes to clusters evaluation on training data

=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 8074.224586275297

Initial starting points (random):

Cluster 0: 2,54,4,38
Cluster 1: 2,34,1,40

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute   Full Data         0         1
            (30163.0)  (5506.0) (24657.0)
=========================================
ㄖ性別            1.3243     1.073    1.3804
年紀             38.4378   43.8262   37.2345
公司單位            1.8813    4.9441    1.1973
每週工作小時數       40.9329   44.1963   40.2041
```

```
Time taken to build model (full training data) : 0.35 seconds

=== Model and evaluation on training set ===

Clustered Instances

0         5506 ( 18%)
1        24657 ( 82%)


Class attribute: 收入
Classes to Clusters:

     0     1  <-- assigned to cluster
  3494 19161 | <=50K
  2012  5496 | >50K

Cluster 0 <-- >50K
Cluster 1 <-- <=50K

Incorrectly clustered instances :        8990.0    29.8047 %
```

- **Number of iterations** 迭代次數

- **Within of** 評價模型好壞的標準

  數值越小，當然是越好，代表同一群實例的距離小，差距沒有很大。

- **Cluster 0**

  成立高收入(>50K)分群，性別多是男性，年紀平均數是 43.8626 歲左右，公司單位平均數為 4.9441(地方政府或自顧收入)，工作時數的平均數是 44.1963 小時

- **Cluster 1**

  年收小於或等於 50K，性別為男性，年紀 37 歲，公司單位平均數 1.1973(私人公司)，那每周工作時數要有 40 小時。

- **百分比比例** 不正確的分群百分比 29.8047% 正確百分比為 70.1953%

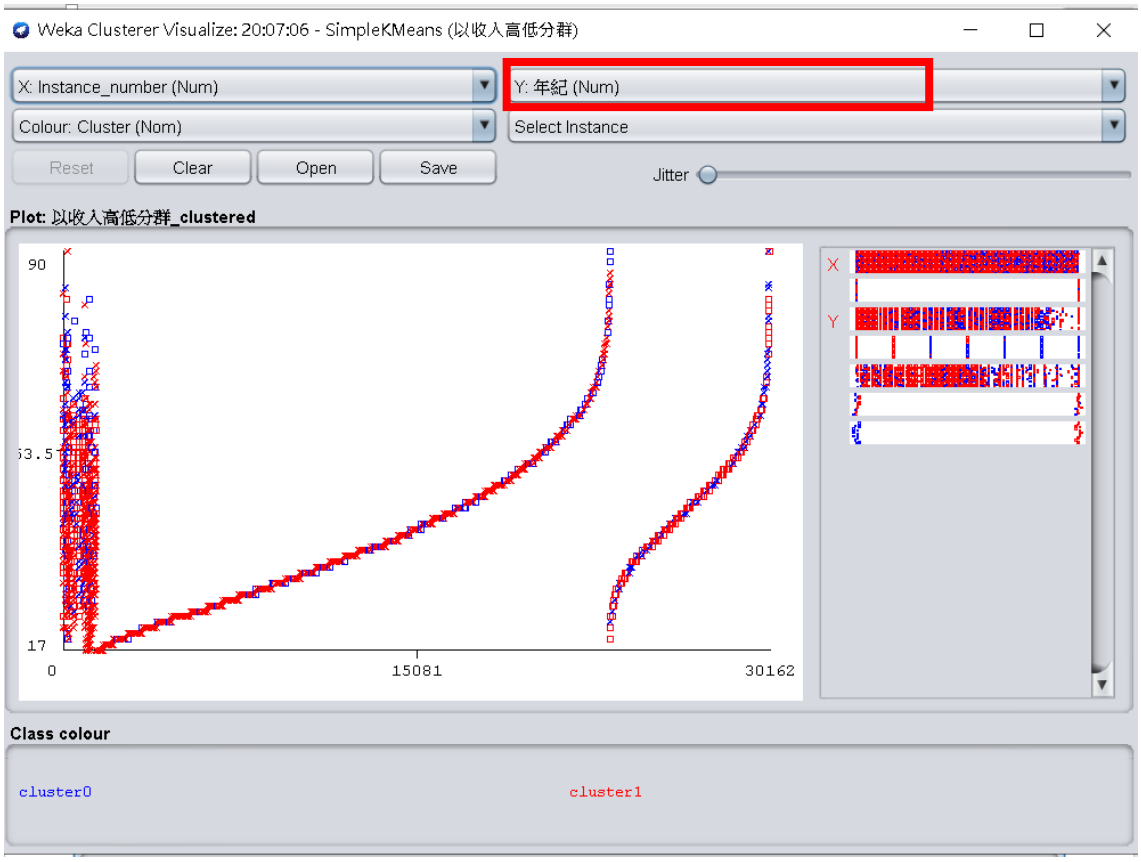
壹拾、 結論

不管年收大於或小於等於 50K 的人口都是占男性居多，符合對男生期待必須養家餬口的觀念。大於 50K 的男性中，幾乎都是 40 歲上的中年人，也可以大概推論這年紀的應該都在公司的管理階級的位置。

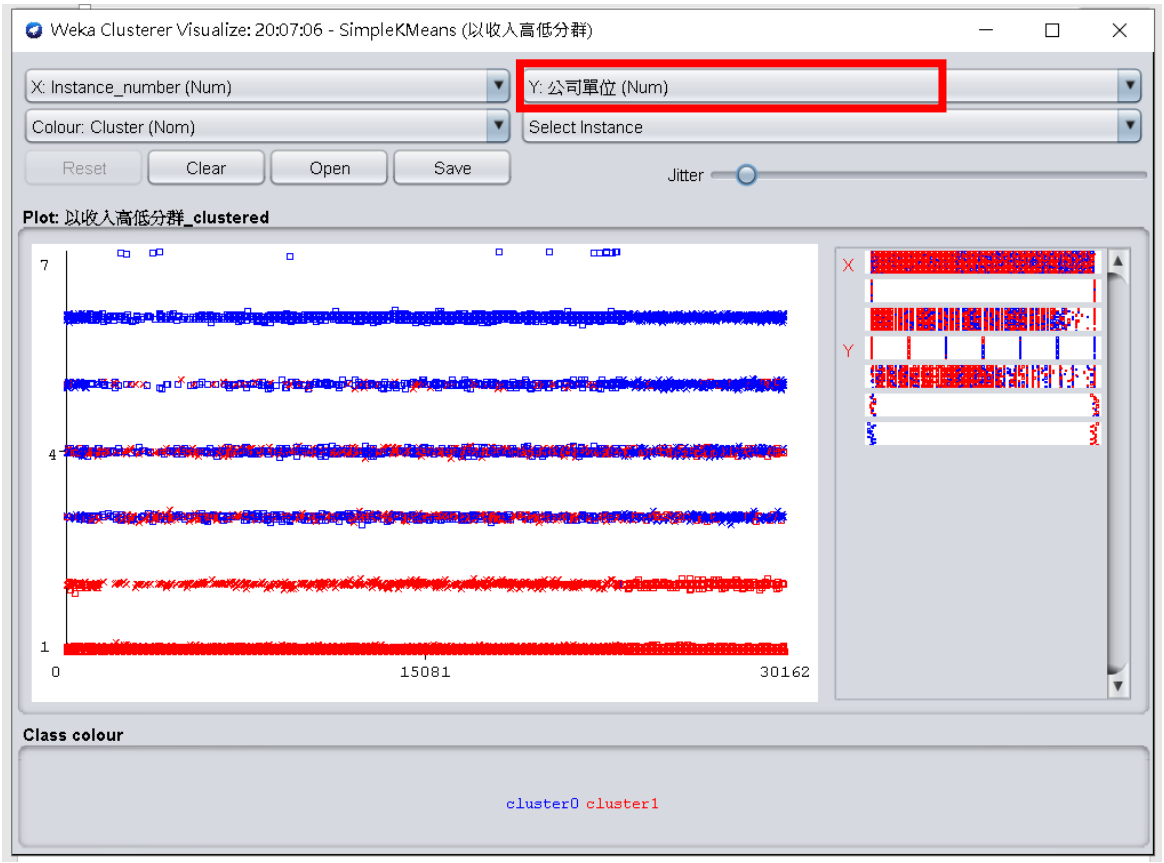
小於或等於 50K 的群體,推斷可能是公司單位的原因，其他公司沒有公家的福利和薪水，並證實高收入與公司單位息息相關，如果說學歷是和薪資密切度第一高，那公司單位絕對就是第二，沒人敢說第三，差不多是這種感覺。

壹拾壹、　分群圖說明

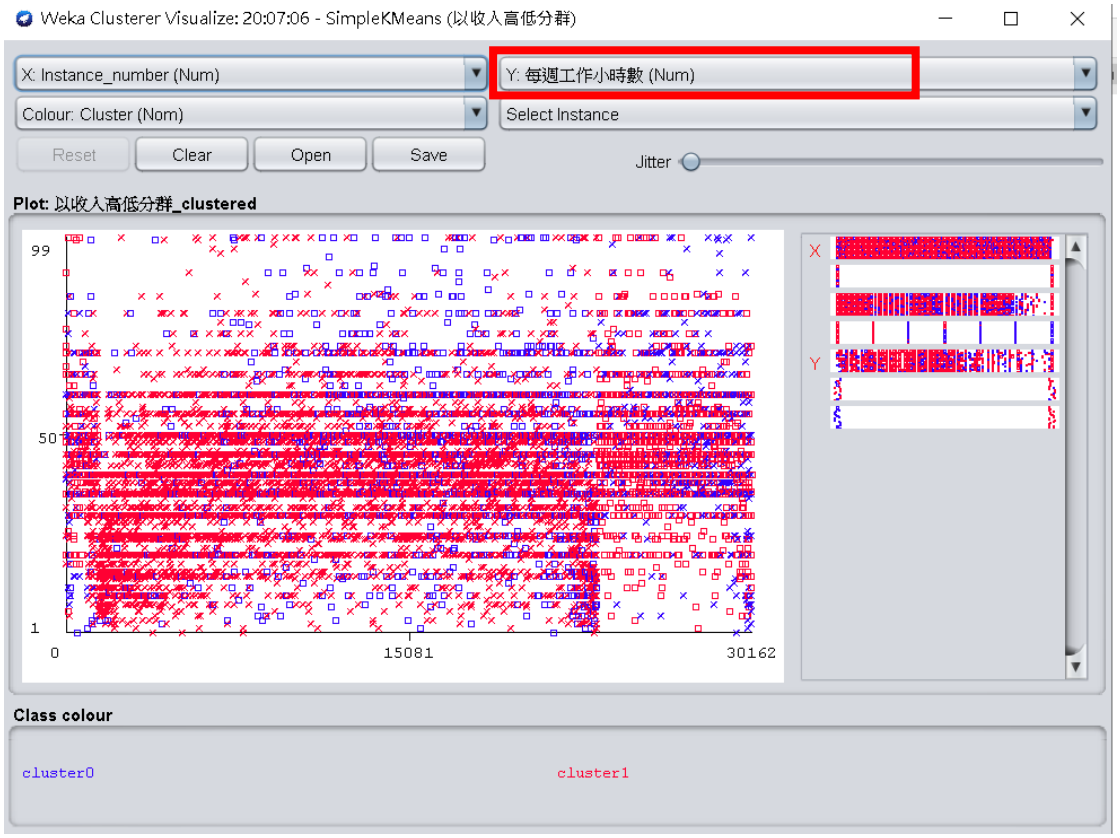
- X 軸為收入分佈，Y 軸為年紀



- X 軸為收入分佈，Y 軸為公司單位，藍色為高收入>50K，紅色為<=50K



- X 軸為收入分佈，Y 軸為每週工作時數，藍色為高收入>50K，紅色為<=50K



玖、 研究二分群說明

➢ 研究二 以高收入女性分群，輸出結果

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 7 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     以高收入女生分群
Instances:    1019524
Attributes:   4
              □每週工作小時數
              教育編號
              公司單位
Ignored:
              婚姻狀況
Test mode:    Classes to clusters evaluation on training data

=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 33.68548930161056

Initial starting points (random):

Cluster 0: 40.897482,4.84982,2.07464
Cluster 1: 40,6,4
Cluster 2: 40,7,5
Cluster 3: 80,9,1
Cluster 4: 40,5,1
Cluster 5: 40,7,1
Cluster 6: 65,8,2

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute     Full Data          0          1          2          3          4          5          6
            (1019524.0) (1018451.0)     (166.0)      (95.0)      (87.0)     (377.0)     (305.0)      (43.0)
=============================================================================================================
□每週工作小時數         40.8975    40.8975    42.9518    35.8947    51.5172    39.0053     40.377    42.3256
教育編號              4.8498     4.8498     6.5422     2.6526     7.8276     2.5915     6.0623     8.1163
公司單位              2.0746     2.0747      4.506        5.2     1.1494     1.1194     1.0033      3.186
```

```
Time taken to build model (full training data) : 3.46 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      1018451 (100%)
1          166 (  0%)
2           95 (  0%)
3           87 (  0%)
4          377 (  0%)
5          305 (  0%)
6           43 (  0%)


Class attribute: 婚姻狀況
Classes to Clusters:

   0   1   2   3   4   5   6  <-- assigned to cluster
   4  22   3  31  23  68  12 | 未結婚
   8  27   8  22  53  50   6 | 離婚
   1   5   9   0  17   9   0 | 守寡
   0   2   1   3   2   2   1 | 已婚配偶缺席(無奈分開彼此
  25 108  74  30 274 167  22 | 已婚公民配偶
   0   0   0   0   4   2   0 | 已婚-AF-配偶
   1   2   0   1   4   7   2 | 分居

Cluster 0 <-- No class
Cluster 1 <-- 離婚
Cluster 2 <-- 守寡
Cluster 3 <-- 已婚配偶缺席(無奈分開彼此
Cluster 4 <-- 已婚公民配偶
Cluster 5 <-- 未結婚
Cluster 6 <-- 分居

Incorrectly clustered instances :       729.0      0.0715 %
```

- ## Cluster 0

  婚姻狀況為已婚-AF-配偶，每週工作時數平均數為 40.1739，教育編號平均數為 2.6087 (高中畢業)，公司單位平均數為 2.413(私人公司)

- ## Cluster 1

  婚姻狀況為離婚，每週工作時數平均數為 44.4868，教育編號平均數為 5.8553 (學士)，公司單位平均數為 1.0526(私人公司)。

- ## Cluster 2

婚姻狀況為已婚配偶缺席，每週工作時數平均數為 42.5747，教育編號平均數為 6.8145 (碩士)，公司單位平均數為 4.1674 (地方政府)。

- **Cluster 3**

  婚姻狀況為分居，每週工作時數平均數為 23.3，教育編號平均數為 5.9 (學士)，公司單位平均數為 1.02(私人公司)。

- **Cluster 4**

  家庭狀況為不結婚，每週工作時數平均數為 46.7328，教育編號平均數為 7.5878 (碩士)，公司單位平均數為 1.1145(私人公司)。

- **Cluster 5**

  家庭狀況為已婚公民配偶為高收入女性中最多的一群，每週工作時數平均數為 39.0053，教育編號平均數為 2.6088 (高職畢業)，公司單位平均數為 1 (私人公司)。。

- **Cluster 6**

  婚姻狀況為守寡，每週工作時數平均數為 36.1458，教育編號平均數為 2.6667 (高中畢業)，公司單位平均數為 5.1875 (自顧收入)。

- **百分比比例** ➔ 不正確的分群百分比 31.47%　　正確百分比為 68.53%

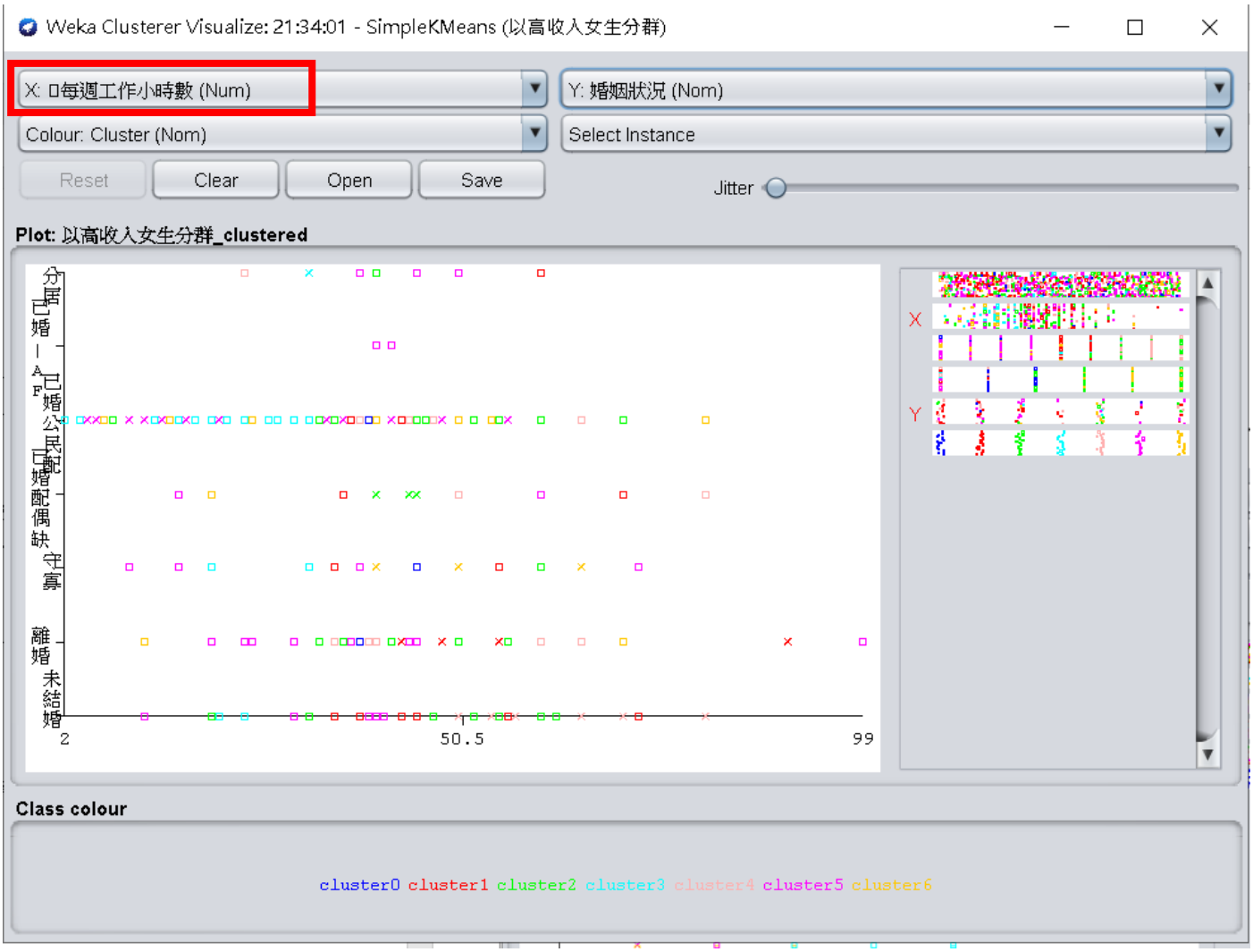
壹拾、 結論

在高收入(>50K)女性的群體中，每週的工作時數大多都落在 40 個小時，而學歷的部份他們幾乎都具有學士或是碩士的學位證明，而大多的工作單位也都落在私人公司居多。

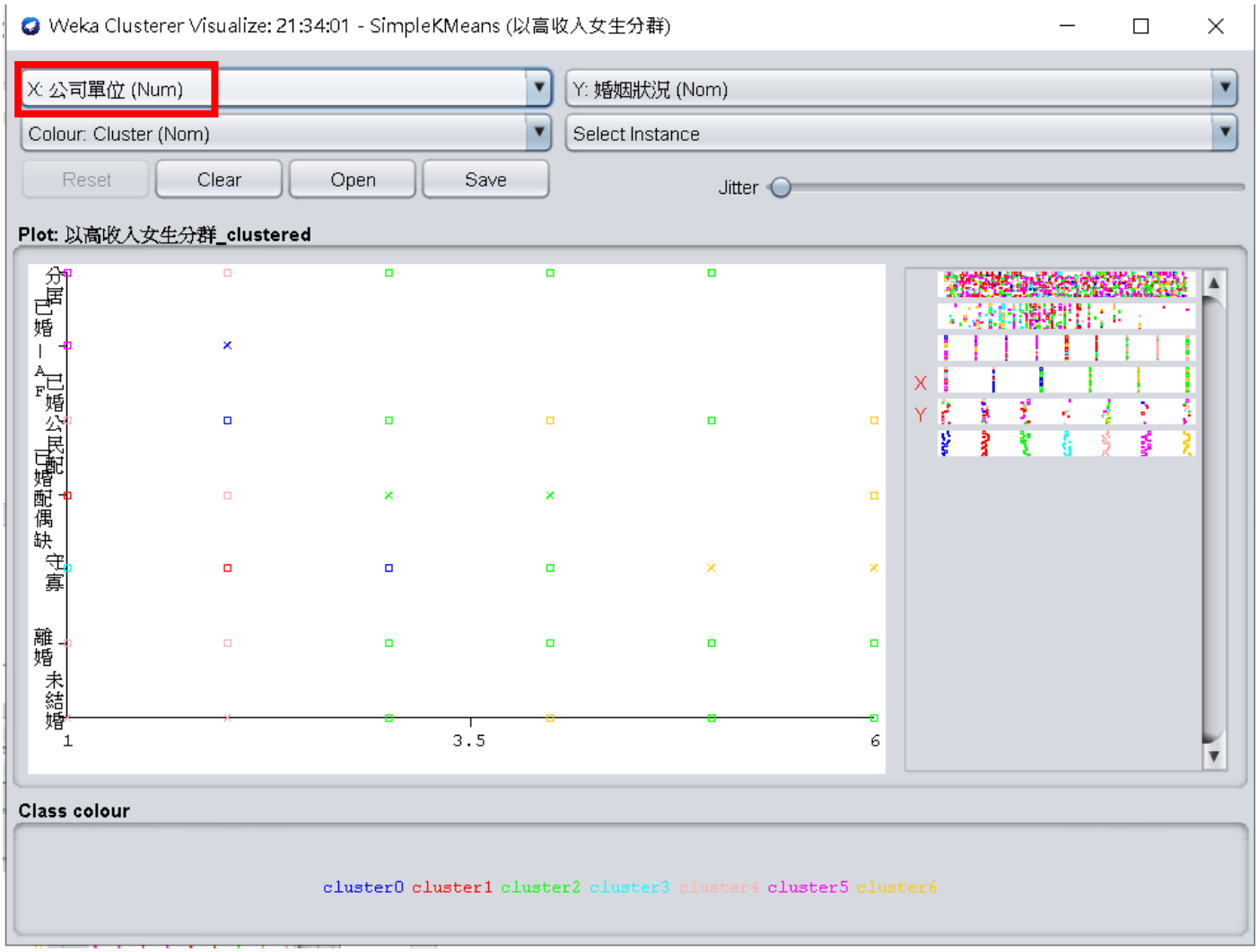
從此資料可以大概推斷這些女性幾乎都是具備高學歷的職業婦女，在每週他們花在工作上的時間就達到了 40 小時甚至更長，一周五天的時間三分之一的花都在工作上，可以想像這是相當不容易，也是因為幾乎都把時間貢獻給了工作，才能達到年收 50K。

壹拾壹、　分群圖說明

- X 軸為每周工作小時數，Y 軸為婚姻狀況



- X 軸為公司單位，Y 軸為婚姻狀況



- X 軸為教育編號，Y 軸為婚姻狀況