分散式資料庫期末報告

WEKA 分析-美國成人普查收入

關聯規則

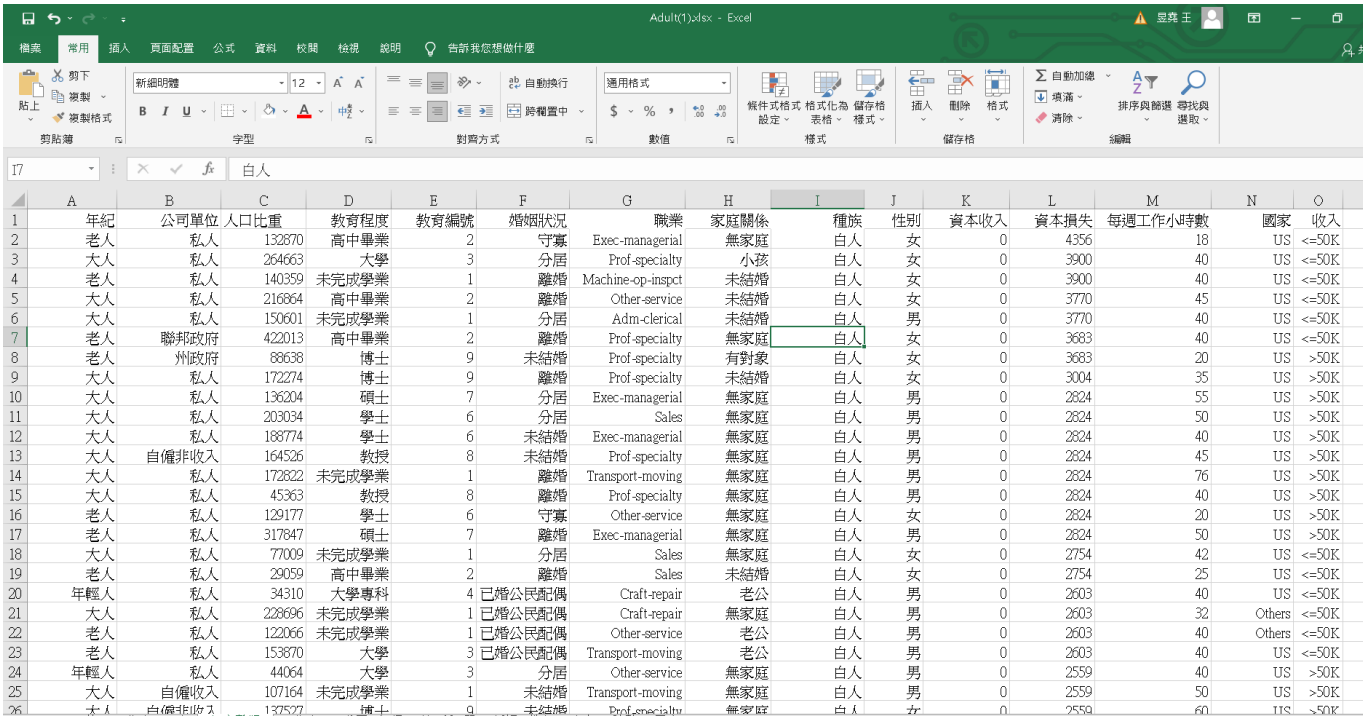
指導老師：蔡正發教授

組員：吳湄潔 B10756014

王諒桓 B10756055

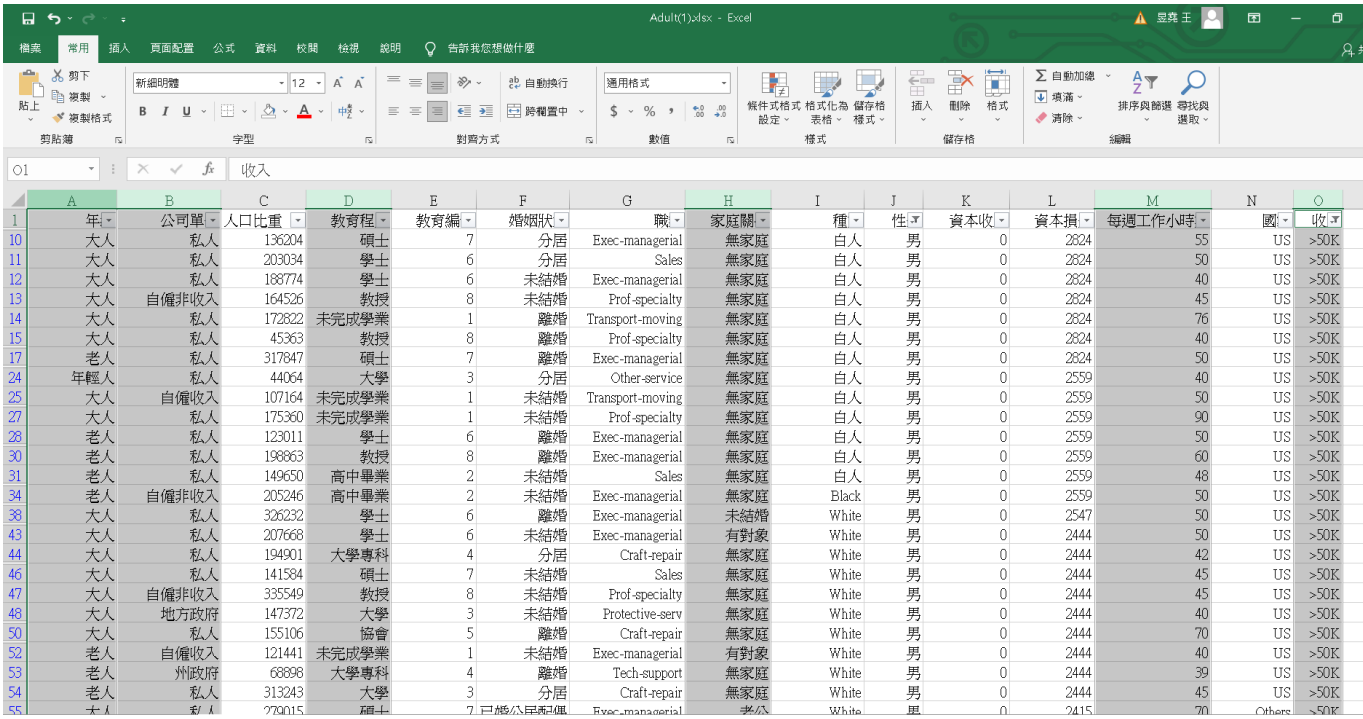
1. 資料處理

1.1 原始資料



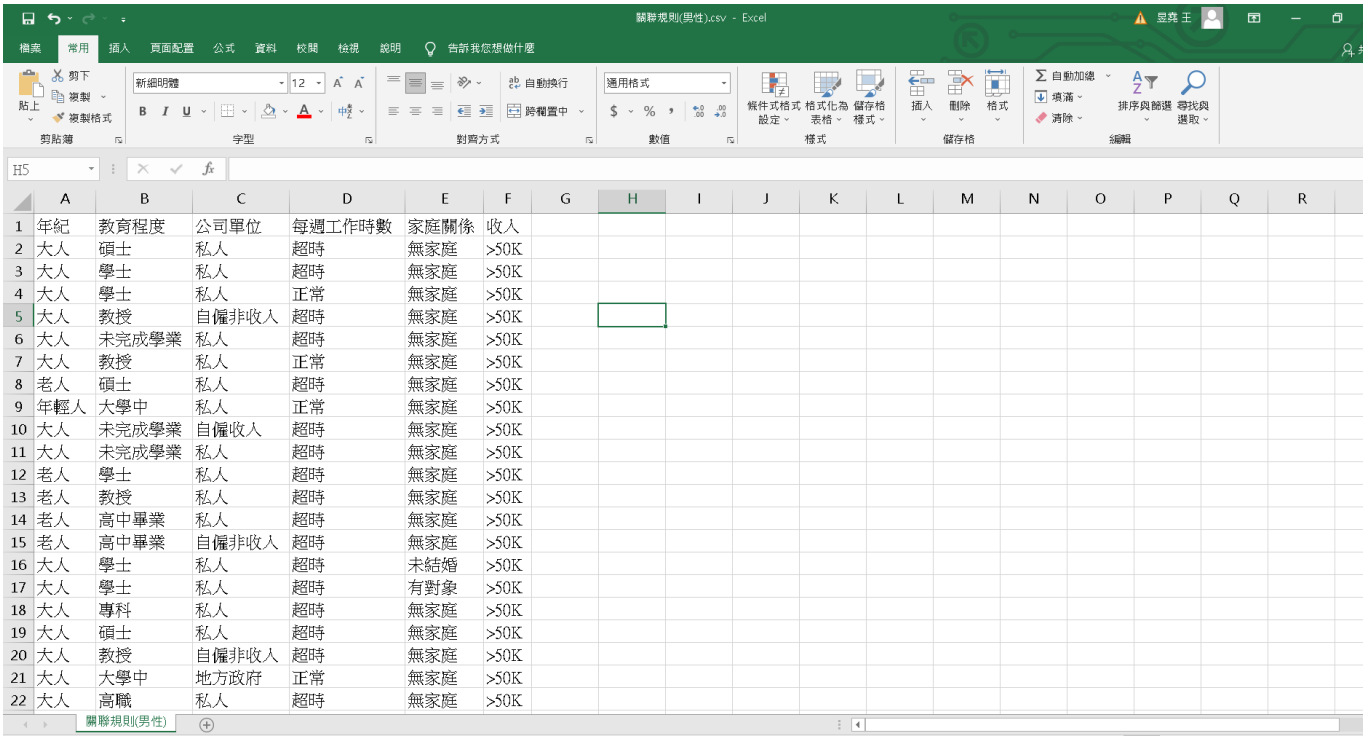
年紀	公司單位	人口比重	教育程度	教育編號	婚姻狀況	職業	家庭關係	種族	性別	資本收入	資本損失	每週工作小時數	國家	收入
老人	私人	132870	高中畢業	2	守寡	Exec-managerial	無家庭	白人	女	0	4356	18	US	<=50K
大人	私人	264663	大學	3	分居	Prof-specialty	小孩	白人	女	0	3900	40	US	<=50K
老人	私人	140359	未完成學業	1	離婚	Machine-op-inspct	未結婚	白人	女	0	3900	40	US	<=50K
大人	私人	216864	高中畢業	2	離婚	Other-service	未結婚	白人	女	0	3770	45	US	<=50K
大人	私人	150601	未完成學業	1	分居	Adm-clerical	未結婚	白人	男	0	3770	40	US	<=50K
老人	聯邦政府	422013	高中畢業	2	離婚	Prof-specialty	無家庭	白人	女	0	3683	40	US	<=50K
老人	州政府	88638	博士	9	未結婚	Prof-specialty	有對象	白人	女	0	3683	20	US	>50K
大人	私人	172274	博士	9	離婚	Prof-specialty	未結婚	白人	女	0	3004	35	US	>50K
大人	私人	136204	碩士	7	分居	Exec-managerial	無家庭	白人	男	0	2824	55	US	>50K
大人	私人	203034	學士	6	分居	Sales	無家庭	白人	男	0	2824	50	US	>50K
大人	私人	188774	學士	6	未結婚	Exec-managerial	無家庭	白人	男	0	2824	40	US	>50K
大人	自僱非收入	164526	教授	8	未結婚	Prof-specialty	無家庭	白人	男	0	2824	45	US	>50K
大人	私人	172822	未完成學業	1	離婚	Transport-moving	無家庭	白人	男	0	2824	76	US	>50K
大人	私人	45363	教授	8	離婚	Prof-specialty	無家庭	白人	男	0	2824	40	US	>50K
老人	私人	129177	學士	6	守寡	Other-service	無家庭	白人	女	0	2824	20	US	>50K
老人	私人	317847	碩士	7	離婚	Exec-managerial	無家庭	白人	男	0	2824	50	US	>50K
大人	私人	77009	未完成學業	1	分居	Sales	無家庭	白人	女	0	2754	42	US	<=50K
老人	私人	29059	高中畢業	2	離婚	Sales	未結婚	白人	女	0	2754	25	US	<=50K
年輕人	私人	34310	大學專科	4	已婚公民配偶	Craft-repair	老公	白人	男	0	2603	40	US	<=50K
大人	私人	228696	未完成學業	1	已婚公民配偶	Craft-repair	無家庭	白人	男	0	2603	32	Others	<=50K
老人	私人	122066	未完成學業	1	已婚公民配偶	Other-service	老公	白人	男	0	2603	40	Others	<=50K
老人	私人	153870	大學	3	已婚公民配偶	Transport-moving	老公	白人	男	0	2603	40	US	<=50K
年輕人	私人	44064	大學	3	分居	Other-service	無家庭	白人	男	0	2559	40	US	>50K
大人	自僱收入	107164	未完成學業	1	未結婚	Transport-moving	無家庭	白人	男	0	2559	50	US	>50K

1.2 利用篩選功能，劃分出男性且收入達到>50K 的資料，資料內容選取年紀、公司單位、教育程度、家庭關係、每周工作時數、收入。



年紀	公司單位	人口比重	教育程度	教育編號	婚姻狀況	職業	家庭關係	種族	性別	資本收入	資本損失	每週工作小時數	國家	收入
大人	私人	136204	碩士	7	分居	Exec-managerial	無家庭	白人	男	0	2824	55	US	>50K
大人	私人	203034	學士	6	分居	Sales	無家庭	白人	男	0	2824	50	US	>50K
大人	私人	188774	學士	6	未結婚	Exec-managerial	無家庭	白人	男	0	2824	40	US	>50K
大人	自僱非收入	164526	教授	8	未結婚	Prof-specialty	無家庭	白人	男	0	2824	45	US	>50K
大人	私人	172822	未完成學業	1	離婚	Transport-moving	無家庭	白人	男	0	2824	76	US	>50K
大人	私人	45363	教授	8	離婚	Prof-specialty	無家庭	白人	男	0	2824	40	US	>50K
老人	私人	317847	碩士	7	離婚	Exec-managerial	無家庭	白人	男	0	2824	50	US	>50K
年輕人	私人	44064	大學	3	分居	Other-service	無家庭	白人	男	0	2559	40	US	>50K
大人	自僱收入	107164	未完成學業	1	未結婚	Transport-moving	無家庭	白人	男	0	2559	50	US	>50K
大人	私人	175360	未完成學業	1	未結婚	Prof-specialty	無家庭	白人	男	0	2559	90	US	>50K
老人	私人	123011	學士	6	離婚	Exec-managerial	無家庭	白人	男	0	2559	50	US	>50K
老人	私人	198863	教授	8	離婚	Exec-managerial	無家庭	白人	男	0	2559	60	US	>50K
老人	私人	149650	高中畢業	2	未結婚	Sales	無家庭	白人	男	0	2559	48	US	>50K
老人	自僱非收入	205246	高中畢業	2	未結婚	Exec-managerial	無家庭	Black	男	0	2559	50	US	>50K
大人	私人	326232	學士	6	離婚	Exec-managerial	未結婚	White	男	0	2547	50	US	>50K
大人	私人	207668	學士	6	未結婚	Exec-managerial	有對象	White	男	0	2444	50	US	>50K
大人	私人	194901	大學專科	4	分居	Craft-repair	無家庭	White	男	0	2444	42	US	>50K
大人	私人	141584	碩士	7	未結婚	Sales	無家庭	White	男	0	2444	45	US	>50K
大人	自僱非收入	335549	教授	8	未結婚	Prof-specialty	無家庭	White	男	0	2444	45	US	>50K
大人	地方政府	147372	大學	3	未結婚	Protective-serv	無家庭	White	男	0	2444	40	US	>50K
大人	私人	155106	協會	5	離婚	Craft-repair	無家庭	White	男	0	2444	70	US	>50K
老人	自僱收入	121441	未完成學業	1	未結婚	Exec-managerial	有對象	White	男	0	2444	40	US	>50K
老人	州政府	68898	大學專科	4	離婚	Tech-support	無家庭	White	男	0	2444	39	US	>50K
老人	私人	313243	大學	3	分居	Craft-repair	無家庭	White	男	0	2444	45	US	>50K

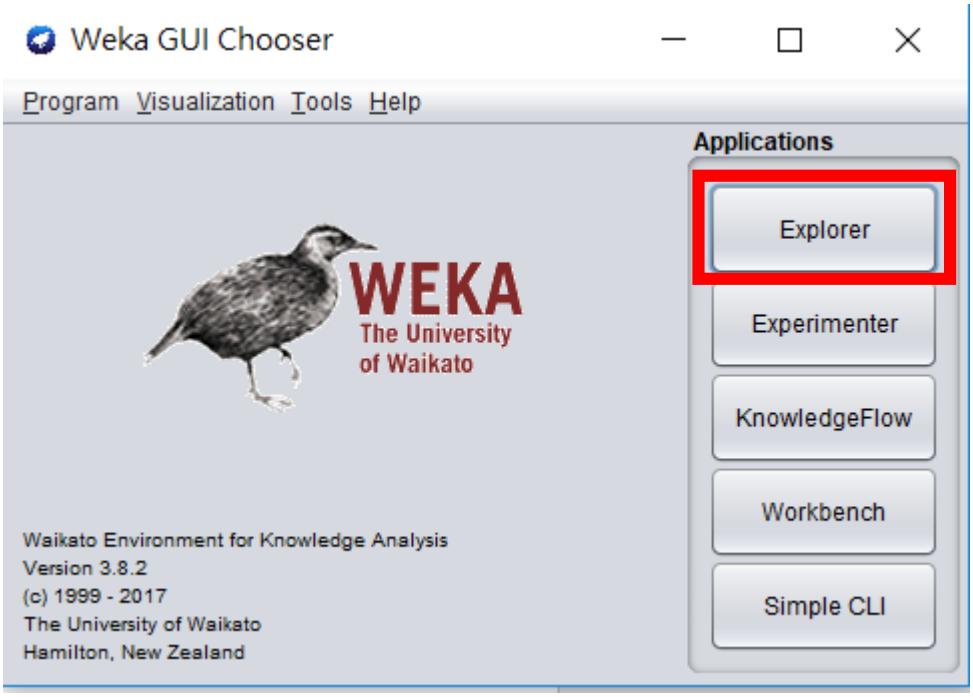
1.3 處理好的資料。

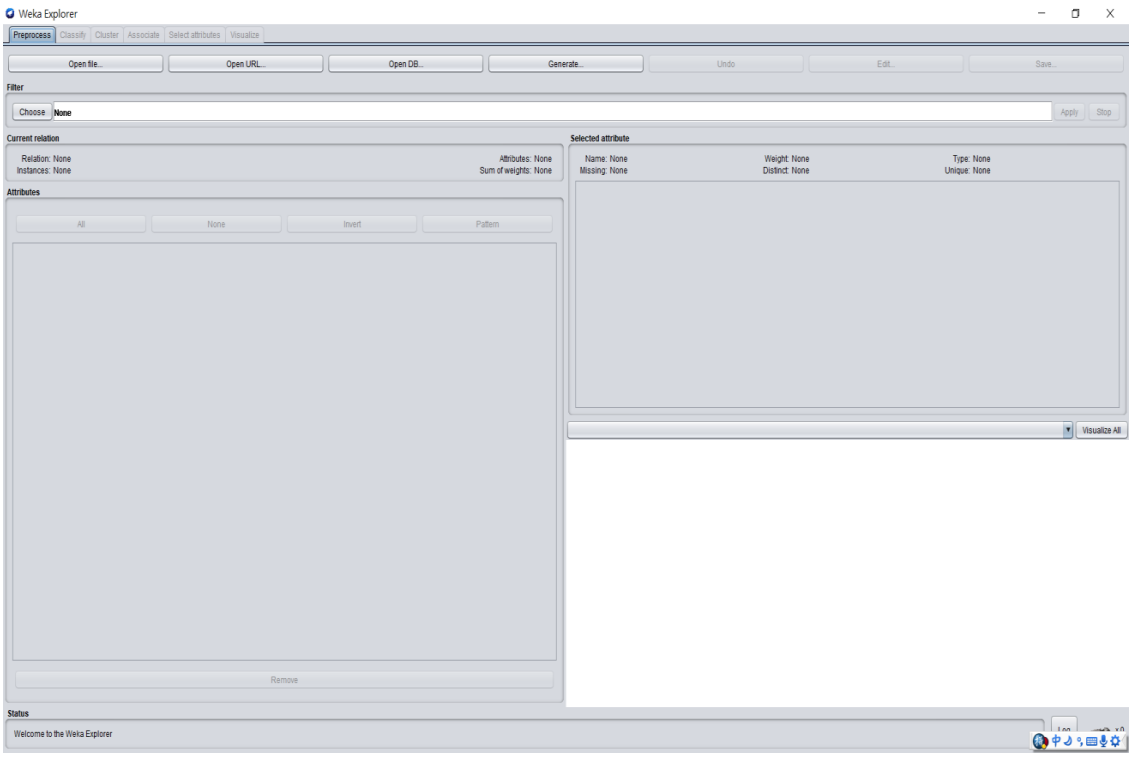


	A	B	C	D	E	F
1	年紀	教育程度	公司單位	每週工作時數	家庭關係	收入
2	大人	碩士	私人	超時	無家庭	>50K
3	大人	學士	私人	超時	無家庭	>50K
4	大人	學士	私人	正常	無家庭	>50K
5	大人	教授	自僱非收入	超時	無家庭	>50K
6	大人	未完成學業	私人	超時	無家庭	>50K
7	大人	教授	私人	正常	無家庭	>50K
8	老人	碩士	私人	超時	無家庭	>50K
9	年輕人	大學中	私人	正常	無家庭	>50K
10	大人	未完成學業	自僱收入	超時	無家庭	>50K
11	大人	未完成學業	私人	超時	無家庭	>50K
12	老人	學士	私人	超時	無家庭	>50K
13	老人	教授	私人	超時	無家庭	>50K
14	老人	高中畢業	私人	超時	無家庭	>50K
15	老人	高中畢業	自僱非收入	超時	無家庭	>50K
16	大人	學士	私人	超時	未結婚	>50K
17	大人	學士	私人	超時	有對象	>50K
18	大人	專科	私人	超時	無家庭	>50K
19	大人	碩士	私人	超時	無家庭	>50K
20	大人	教授	自僱非收入	超時	無家庭	>50K
21	大人	大學中	地方政府	正常	無家庭	>50K
22	大人	高職	私人	超時	無家庭	>50K

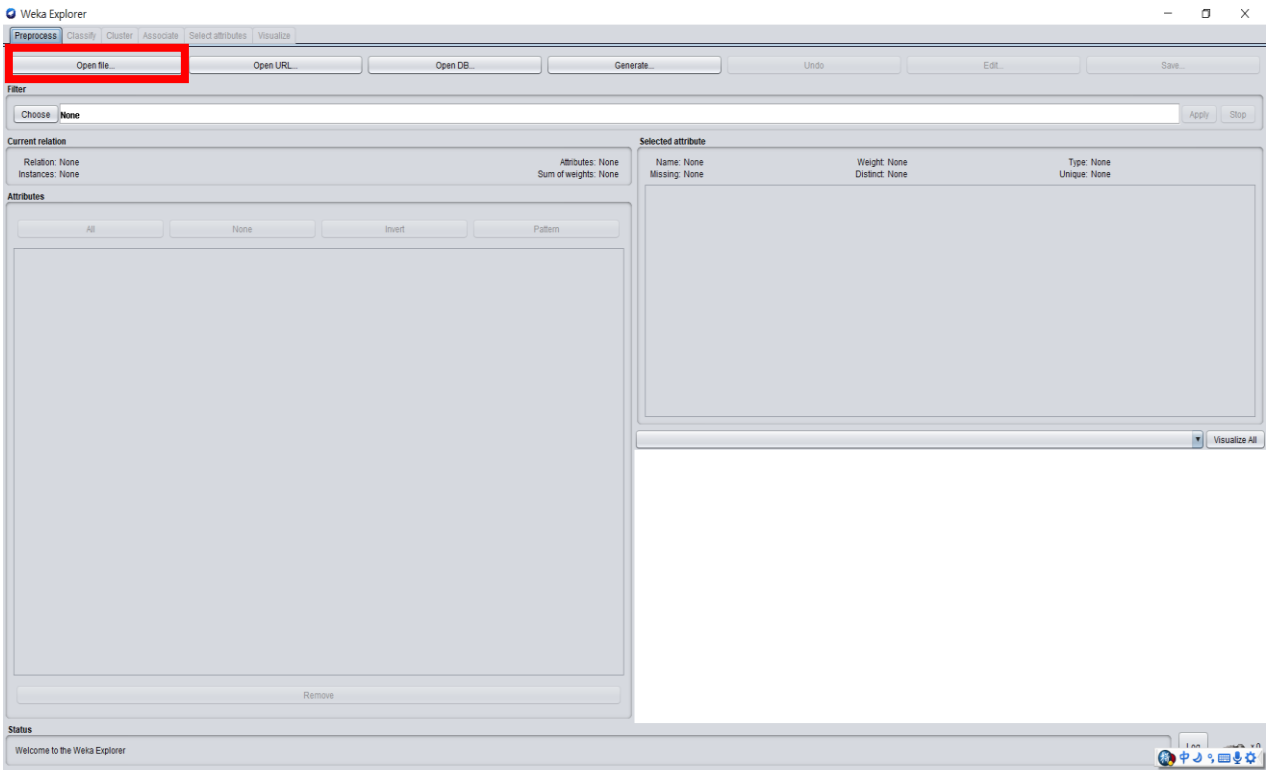
2. 資料分析

2.1 打開 weka 開啟探索者界面

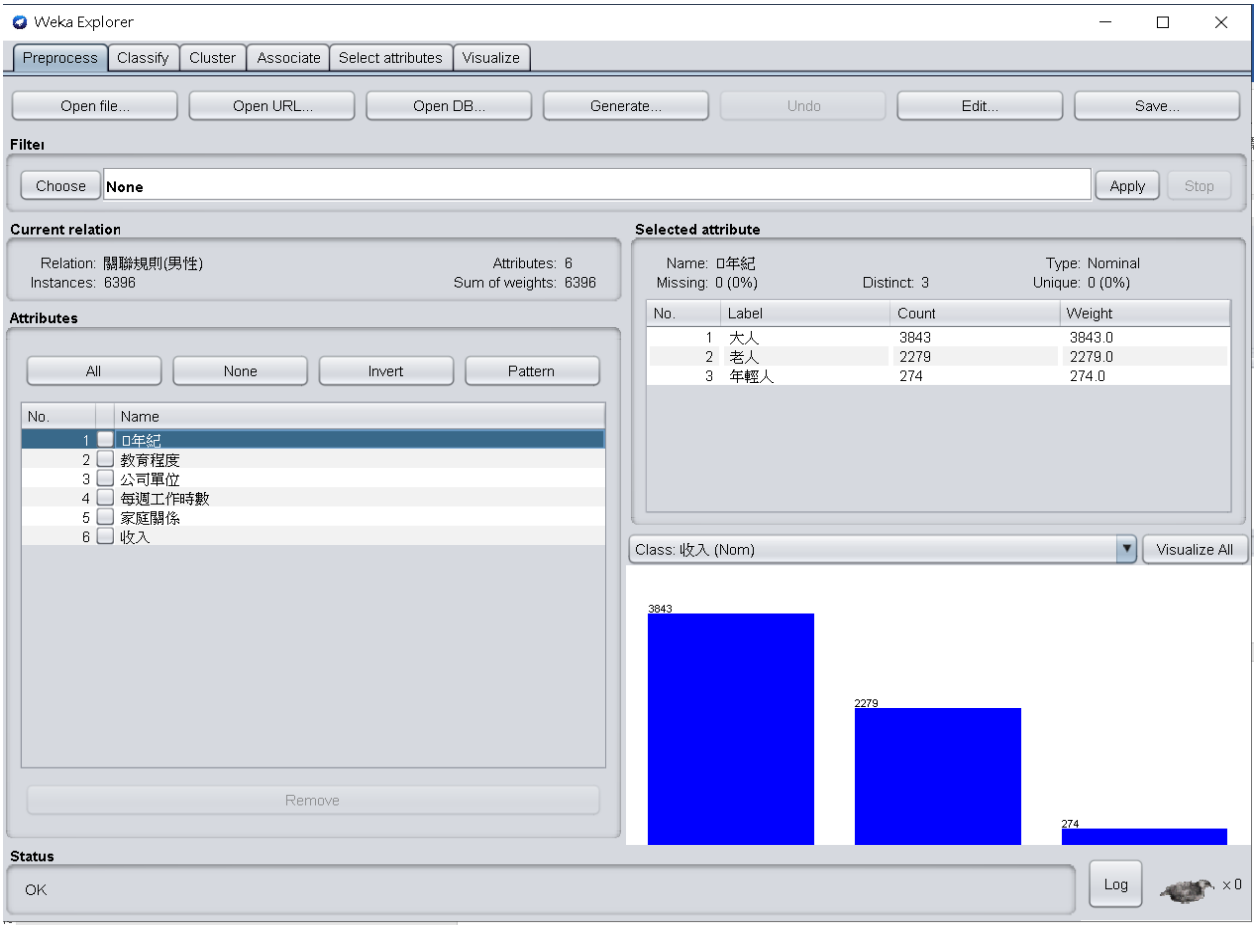



Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: None
Instances: None
Attributes: None
Sum of weights: None
Selected attribute
Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None
Attributes
All
None
Invert
Pattern
Remove
Visualize All
Status
Welcome to the Weka Explorer

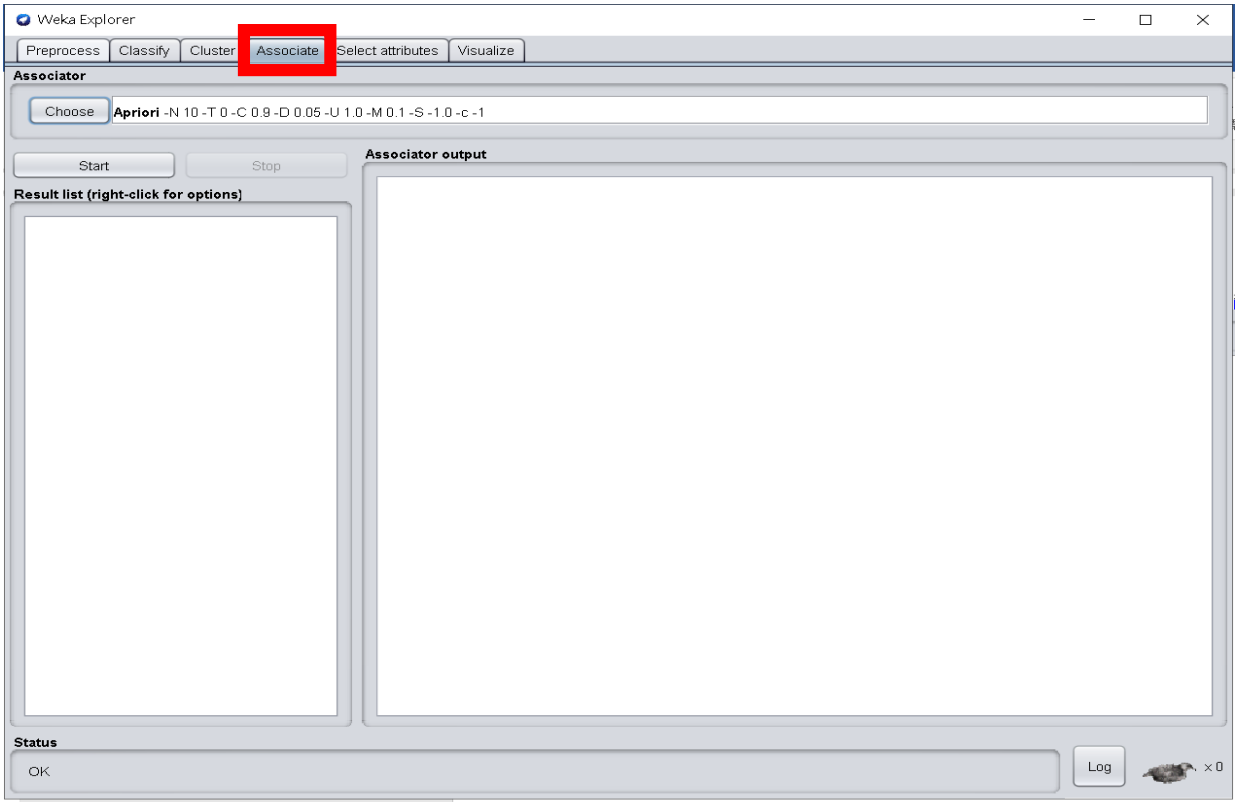
2.2 匯入資料


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: None
Instances: None
Attributes: None
Sum of weights: None
Selected attribute
Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None
Attributes
All
None
Invert
Pattern
Remove
Visualize All
Status
Welcome to the Weka Explorer

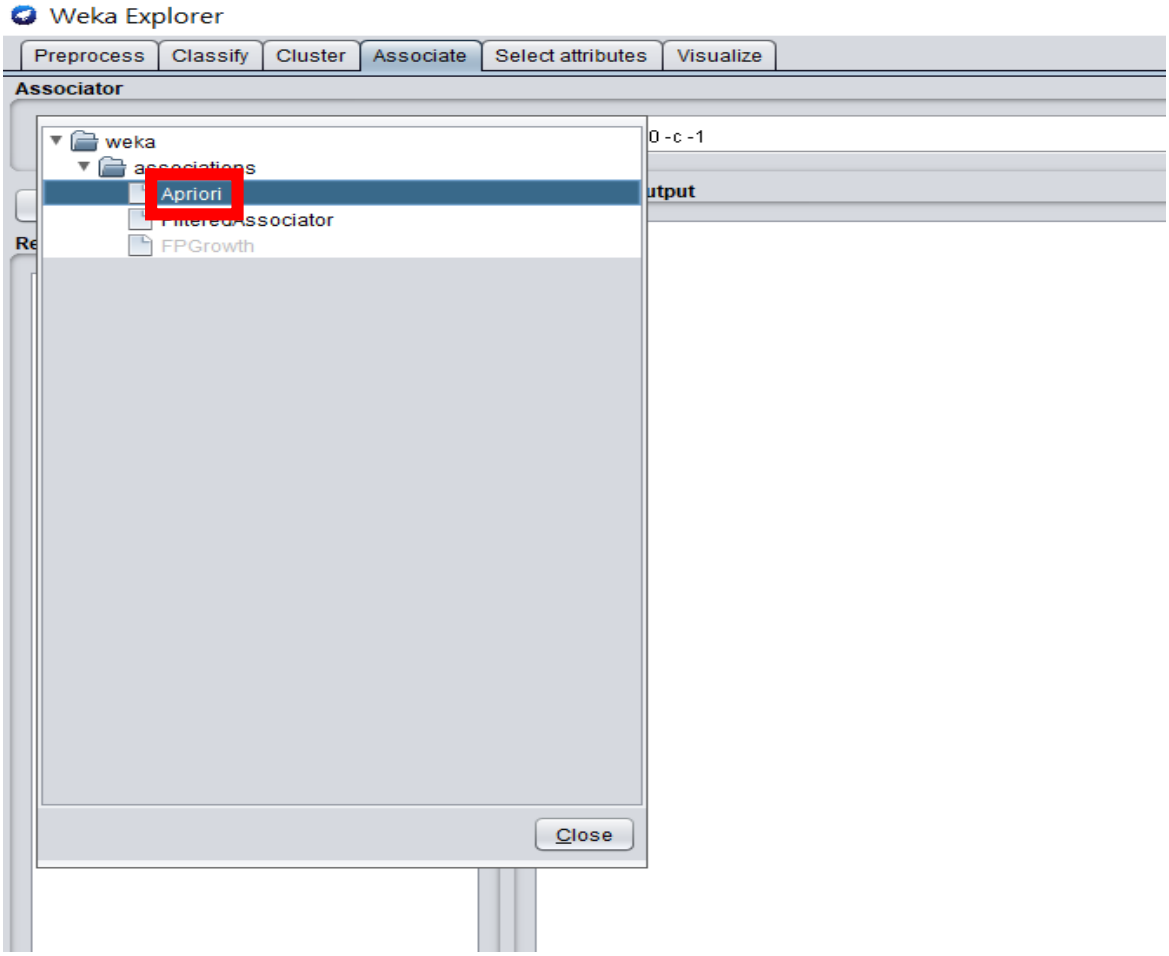
匯入後顯示結果



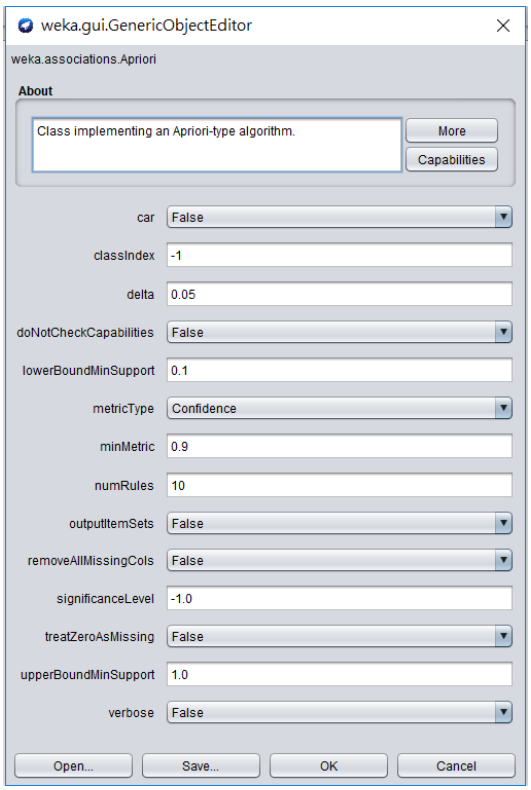
2.3 點擊 Associate(關聯)分析資料



2.4 選擇 Apriori 演算法



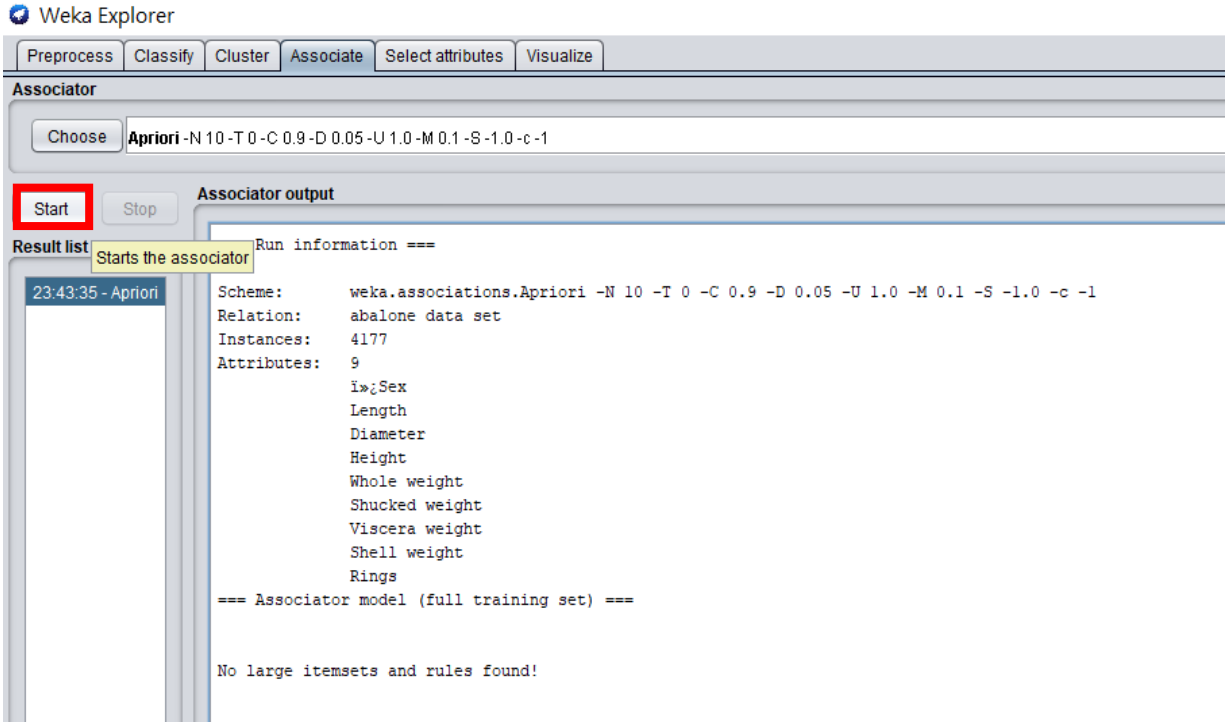
2.5 Aprion 參數介紹



- N(numRules):需要顯示結果的筆數,預設值為 10 筆
- C(minMetric):信心度(confidence)的最小值,預設值為 0.9
- D(delta):每次執行支持度(support)所遞減值,支持度會先從設定的最大值(即 -U 參數的設定)開始計算,如果沒有符合的值,則根據此參數來遞減,直到可 以得到所要的值或低於 -M 的設定值為止,預設值為 0.05
- U(upperBoundMinsupport):支持度的上限,預設值為 1.0
- M(lowerBoundMinsupport):支持度的下限,預設值為 0.1

  例:以上述的預設值設定表示需要從設定的資料找出十筆符合信心度 90%以上且 支持度為 10%到 100%之間的資料

2.6 點擊 Start 執行關聯分析



3.關聯分析結果

3.1 總結果和運行結果說明

```
15:42:00 - Apriori
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     關聯規則(男性)
Instances:    6396
Attributes:   6
              □年紀
              教育程度
              公司單位
              每週工作時數
              家庭關係
              收入
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.4 (2558 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 10

Size of set of large itemsets L(3): 5

Best rules found:

 1. 家庭關係=老公 5679 ==> 收入=>50K 5679    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. 公司單位=私人 4155 ==> 收入=>50K 4155    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. □年紀=大人 3843 ==> 收入=>50K 3843    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. 公司單位=私人 家庭關係=老公 3685 ==> 收入=>50K 3685    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. □年紀=大人 家庭關係=老公 3407 ==> 收入=>50K 3407    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. 每週工作時數=超時 3380 ==> 收入=>50K 3380    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 7. 每週工作時數=超時 家庭關係=老公 2944 ==> 收入=>50K 2944    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 8. 每週工作時數=正常 2905 ==> 收入=>50K 2905    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 9. 每週工作時數=正常 家庭關係=老公 2634 ==> 收入=>50K 2634    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. □年紀=大人 公司單位=私人 2626 ==> 收入=>50K 2626    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
```

- scheme:說明使用何種演算法及參數設定
- instance:資料的筆數有 6396 筆
- Attributes: 共有 6 個欄位及欄位的名稱
- Minimum support:最小的支持度為 40%,共找到 2558 筆資料
- Minimum metric<Confidence>:最小信心度為 90%
- Number of cycles performed :執行了 12 次，代表 run 了 12 輪才找到符合設定的
- 不同的 Minium support(最小的支持度)的值和不同的 Minimum metric<Confidence>(最小信心度)的值會呈現不同的關聯規則
- 測試無數次的 Minium support(最小的支持度)和 Minimum metric<Confidence>(最小信心度)終於在 Minium support=40%且 Minimum metric<Confidence>=90%找到最好的關聯。

3.2 關聯分析說明

```
Apriori
=======

Minimum support: 0.4 (2558 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 10

Size of set of large itemsets L(3): 5

Best rules found:

 1. 家庭關係=老公 5679 ==> 收入=>50K 5679    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. 公司單位=私人 4155 ==> 收入=>50K 4155    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. □年紀=大人 3843 ==> 收入=>50K 3843    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. 公司單位=私人 家庭關係=老公 3685 ==> 收入=>50K 3685    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. □年紀=大人 家庭關係=老公 3407 ==> 收入=>50K 3407    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. 每週工作時數=超時 3380 ==> 收入=>50K 3380    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 7. 每週工作時數=超時 家庭關係=老公 2944 ==> 收入=>50K 2944    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 8. 每週工作時數=正常 2905 ==> 收入=>50K 2905    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 9. 每週工作時數=正常 家庭關係=老公 2634 ==> 收入=>50K 2634    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. □年紀=大人 公司單位=私人 2626 ==> 收入=>50K 2626    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
```

Minium support(最小的支持度)=40%和 Minimum metric<Confidence>(最小信心度)=90%
Generated sets of large itemsets (頻繁項集):
Size of set of large itemsets L(1): 6(有 6 各大小為 1 的項集)
Size of set of large itemsets L(2): 10(有 10 各大小為 2 的項集)
Size of set of large itemsets L(3): 5(有 5 各大小為 2 的項集)

3.3 Best rules found(關聯規則)

- 家庭關係=老公 5679 ==> 收入=>50K 5679 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
- 公司單位=私人 4155 ==> 收入=>50K 4155 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
- 年紀=大人 3843 ==> 收入=>50K 3843 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
- 公司單位=私人 家庭關係=老公 3685 ==> 收入=>50K 3685 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
- 年紀=大人 家庭關係=老公 3407 ==> 收入=>50K 3407 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
- 每週工作時數=超時 3380 ==> 收入=>50K 3380 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
- 每週工作時數=超時 家庭關係=老公 2944 ==> 收入=>50K 2944 <conf:(1)> lift:(1)

lev:(0) [0] conv:(0)

- 每週工作時數=正常 2905 ==> 收入=>50K 2905 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
- 每週工作時數=正常 家庭關係=老公 2634 ==> 收入=>50K 2634 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
- 年紀=大人 公司單位=私人 2626 ==> 收入=>50K 2626 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

4. 結論

我們分析關聯規則的資料選取是想說現今還是以男生在外工作的居多，所以想分析年收入有>50K 的男性都有什麼關聯性，他們大多都身為老公，且公司單位幾乎都是私人公司，工作時數多數還是正常和超時的，這個分析結果也很符合當初我們的假設，在多數年收入>50K 的男性都是落在壯年期，這個年齡的男性幾乎都已經身為了公司的管理階層位置，且他們多半都身為家中的經濟來源，必須努力工作來養家餬口，所以這次的分析可以說是成功的驗證了我們的看法。