大數據基本演算

探索性分析

分群與異常偵測

政治大學圖檔所博士候選人

陳勇汀

pudding@nccu.edu.tw

2020年

陳勇汀 (布丁) 講者簡介

學歷:

- 政治大學 圖書資訊與檔案學研究所 博士候選人
- 政治大學 圖書資訊與檔案學研究所 碩士
- 輔仁大學 圖書資訊學系 學士

專長:

- 數位學習
- 數位典藏與數位圖書館
- 資料探勘

布丁布丁吃什麼？
http://blog.pulipuli.info

關於本週課程

課程大綱

1. 認識Weka
2. 準備Weka
3. 分群
4. 異常偵測
5. 學習單作業的說明

課程目標

- 能夠自行安裝Weka與所需環境
- 能夠使用Weka找出隱含在資料中的模式 (類別)
- 能夠使用Weka找出資料中特別的案例

Part 1.

認識Weka

Weka的出生地
紐西蘭懷卡託大學

https://www.waikato.ac.nz/

開放原始碼工具
Weka

- 紐西蘭懷卡托大學機器學習實驗室專為學習資料探勘所開發的Java軟體，可用於研究、教學、應用等各種用途
- 包含完整的資料探勘處理流程，含括資料前處理工具、機器學習演算法、成效評估方法、資訊視覺化報表摘要
- 兼具圖形化使用者介面與指令列應用工具
  - 易於比較不同演算法的分析結果
  - 模組化設計，能夠擴充不同的演算法
- 跨平臺：Windows、Mac OS、Linux
- 1993年開發初版，至今最新版本是2018年發佈的3.9.4

https://www.cs.waikato.ac.nz/ml/weka/

Weka命名的由來

- Weka是指紐西蘭秧雞 (Gallirallus australis)
- 紐西蘭地區的一種不會飛的特有種鳥類

Waikato Environment for Knowledge Analysis

↓

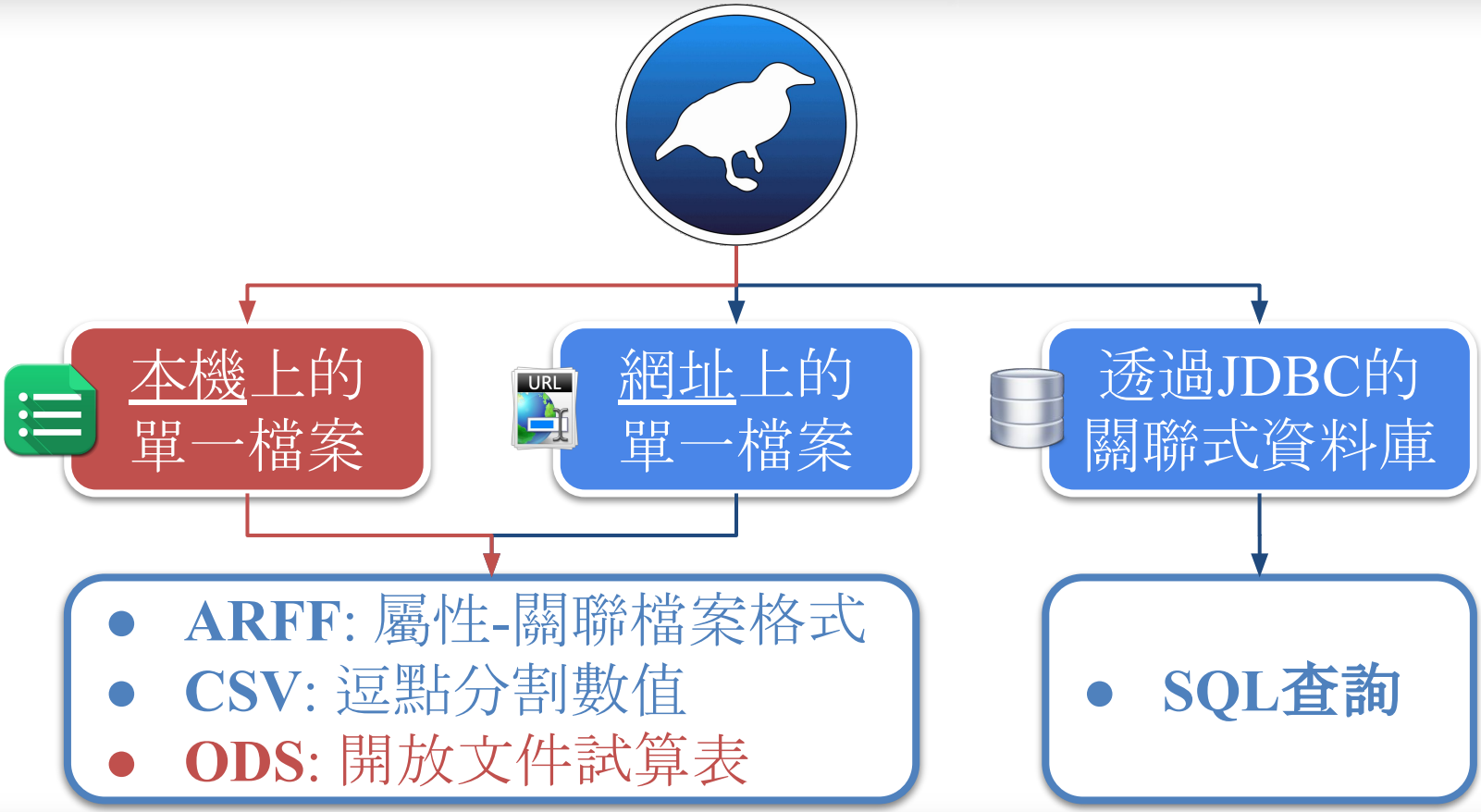
Weka

http://blog.pulipuli.info/2017/07/weka-spin-animation-of-weka-bird.html

Weka對於資料探勘的支援

http://www.mobileprivacy.org/2013/06/digital-privacy-isnt-taken-it-is-given-away/

Weka可接受的資料來源

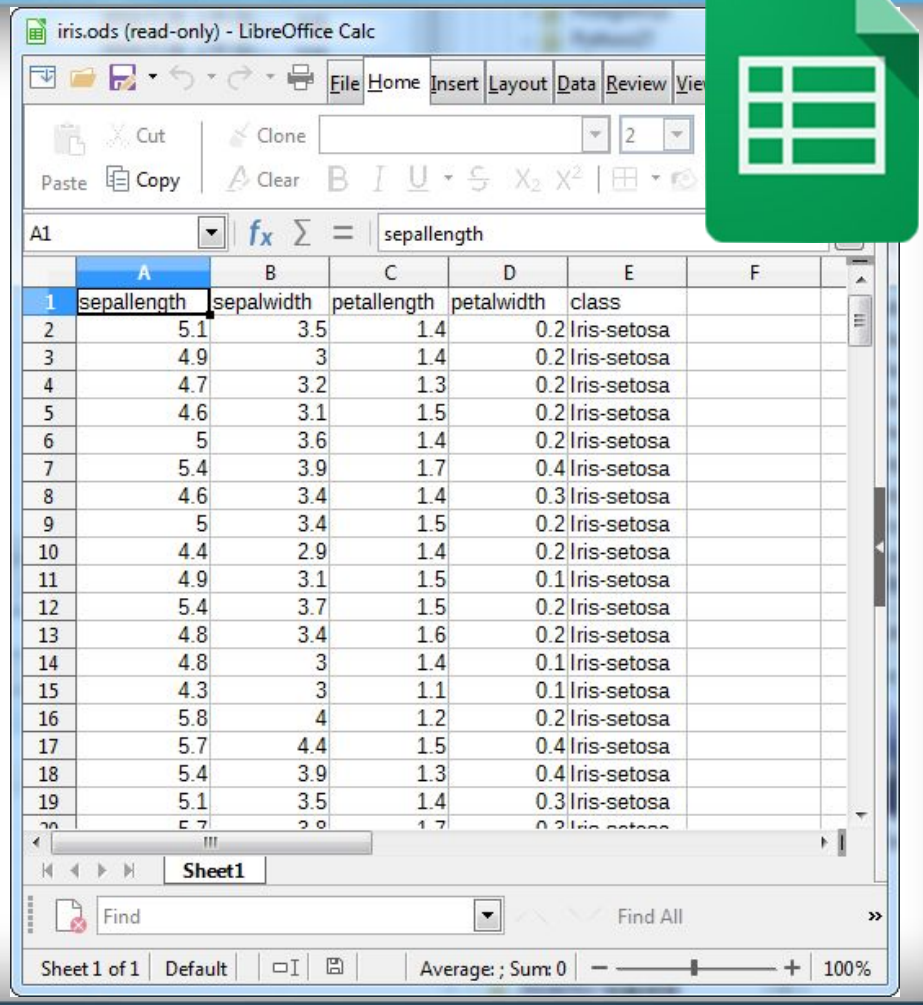


單一檔案格式
ODS

OpenDoucment Spreadsheet

開放文件試算表

1. **屬性**：每一直欄為一屬性
2. **屬性標題**：第一橫列為屬性標題
3. **資料案例**：第二列之後的每一列為一案例
4. **資料類型**：依細格資料類型設定
5. 主要使用LibreOffice編輯



	A	B	C	D	E
1	sepallength	sepalwidth	petallength	petalwidth	class
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	5	3.6	1.4	0.2	Iris-setosa
7	5.4	3.9	1.7	0.4	Iris-setosa
8	4.6	3.4	1.4	0.3	Iris-setosa
9	5	3.4	1.5	0.2	Iris-setosa
10	4.4	2.9	1.4	0.2	Iris-setosa
11	4.9	3.1	1.5	0.1	Iris-setosa
12	5.4	3.7	1.5	0.2	Iris-setosa
13	4.8	3.4	1.6	0.2	Iris-setosa
14	4.8	3	1.4	0.1	Iris-setosa
15	4.3	3	1.1	0.1	Iris-setosa
16	5.8	4	1.2	0.2	Iris-setosa
17	5.7	4.4	1.5	0.4	Iris-setosa
18	5.4	3.9	1.3	0.4	Iris-setosa
19	5.1	3.5	1.4	0.3	Iris-setosa

辦公室套裝的自由軟體
LibreOffice

- LibreOffice辦公室套裝軟體的試算表工具
- LibreOffice是跨平臺的開放自由軟體，是編輯開放文件格式(ODF)的最佳選擇
- 開放文件格式包含文件(ODT)、試算表(ODS)、投影片(ODP)等多種類型格式
- 開放文件格式是我國政府的主要通用格式

https://zh-tw.libreoffice.org/download/libreoffice-fresh/

相關詞彙定義
案例、屬性 (1/2)

案例 (Instance)

抽樣對象、個案、
觀察值個體

屬性(Attribute)

特徵、變項、觀察值

案例1

屬性

- 名字:豪快綠
- 攻擊力:9
- 防禦力:5

案例2

屬性

- 名字:猛牛紫
- 攻擊力:3
- 防禦力:12

相關詞彙定義
案例、屬性 (2/2)

屬性

屬性標題

案例

名字	攻擊力	防禦力
豪快綠	9	5
猛牛紫	3	12

Attribute Type
屬性的資料類型

資料類型分類		舉例
主要 資料類型	Nominal 類別型	• male • 臺南
	Numeric 數值型	• 1 • 0.75
特殊 資料類型	String 字串型 (文字型) 需要額外處理	• This is a pen • 這是一隻筆
	Boolean 布林值 (是或否)	• t • f (全部以類別型取代)
缺失資料	Missing Value 缺失值/未知值	? (空白)

Part 2.

準備Weka
下載、安裝與設定

Weka的下載

(本教學是用3.8.1版本)

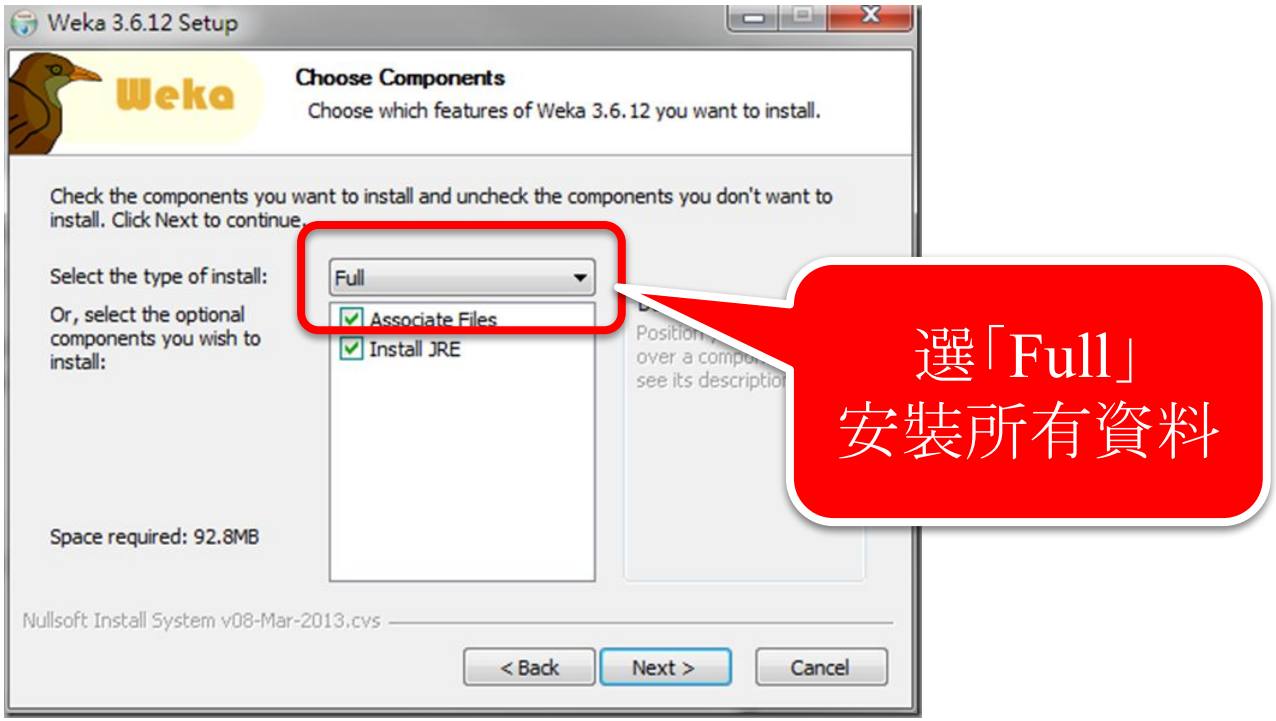
https://www.cs.waikato.ac.nz/ml/weka/downloading.html

請下載
includes Java VM
版本
(有分64 bit或32 bit)

所需環境:

Weka的安裝

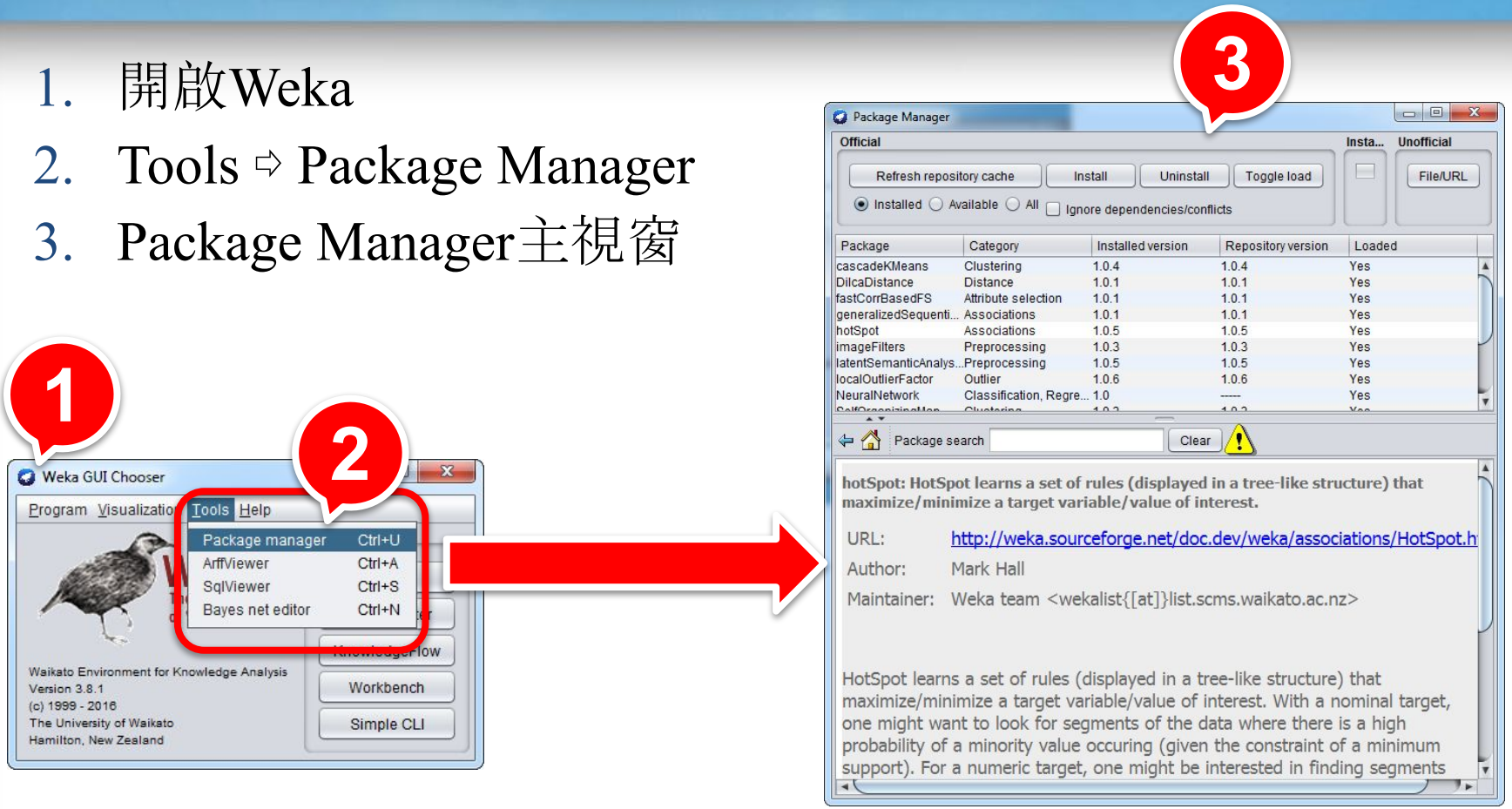
安裝精靈，容易上手



https://me1237guy.pixnet.net/blog/post/59757613-%E5%AE%89%E8%A3%9Dweka

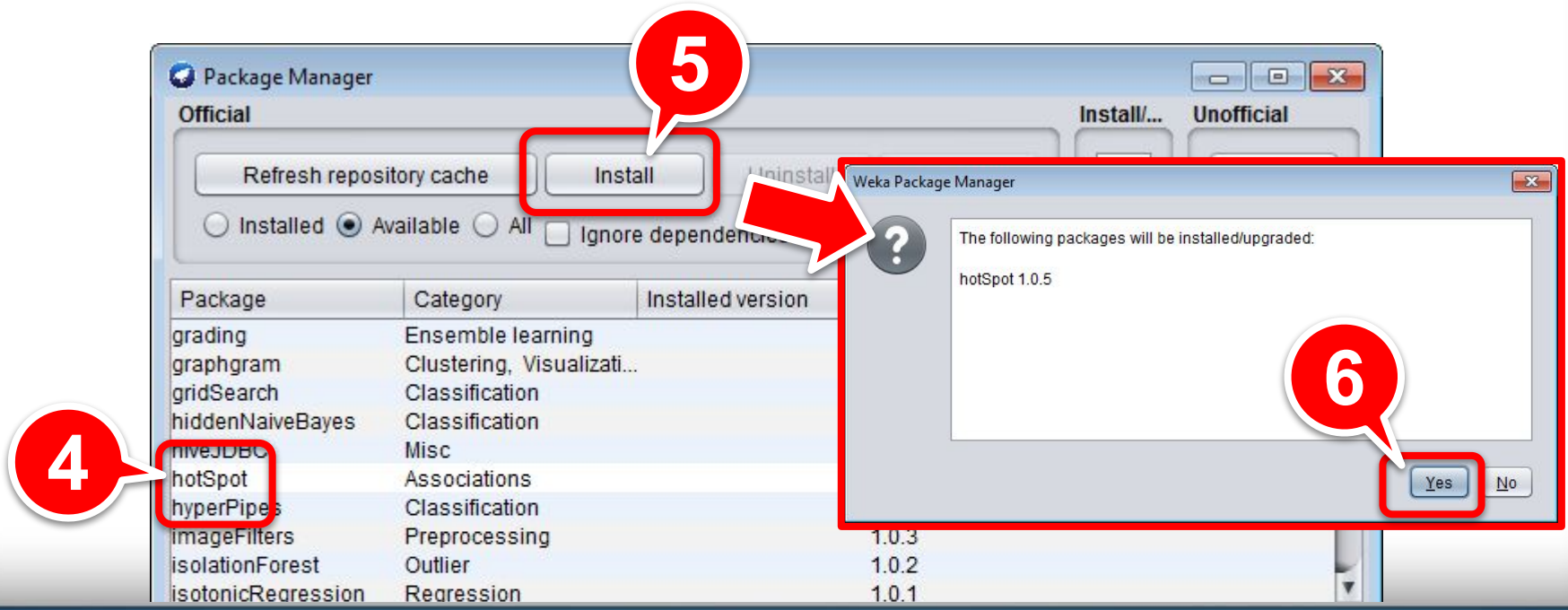
STEP 1. 安裝官方套件 (1/3)

1. 開啟Weka
2. Tools ⇨ Package Manager
3. Package Manager主視窗



STEP 1. 安裝官方套件 (2/3)

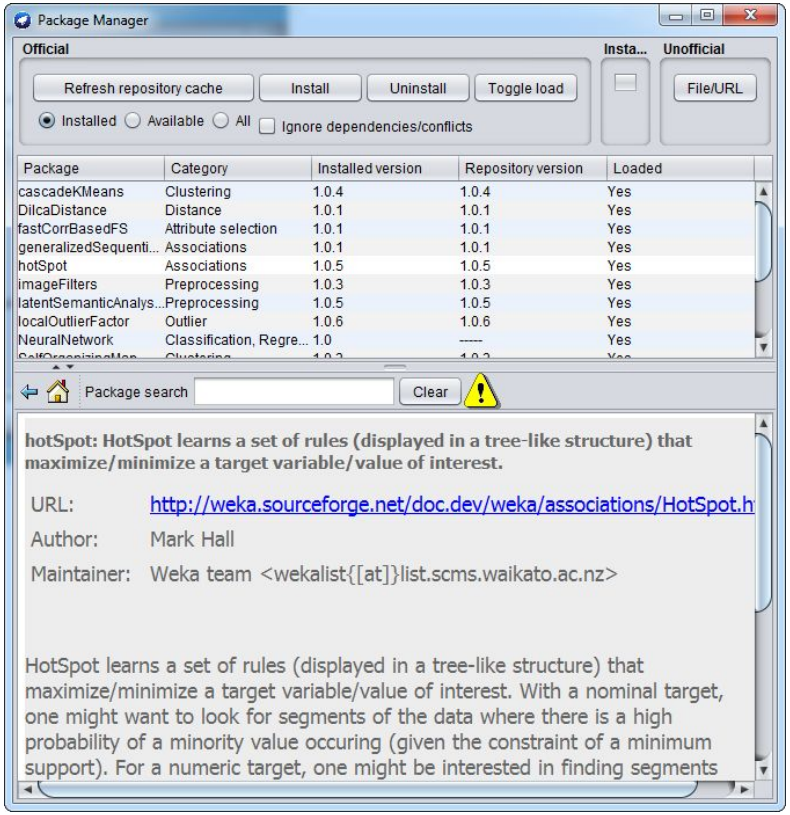
4. 找到要安裝的套件，例如cascadeKMeans
5. Install
6. 確認安裝，Yes



STEP 1. 安裝官方套件 (3/3)

請按照以上步驟, 安裝以下套件吧:

- **cascadeKMeans**
  分群演算法
- **localOutlierFactor**
  異常偵測演算法



STEP 2. 安裝自製套件 (1/2)

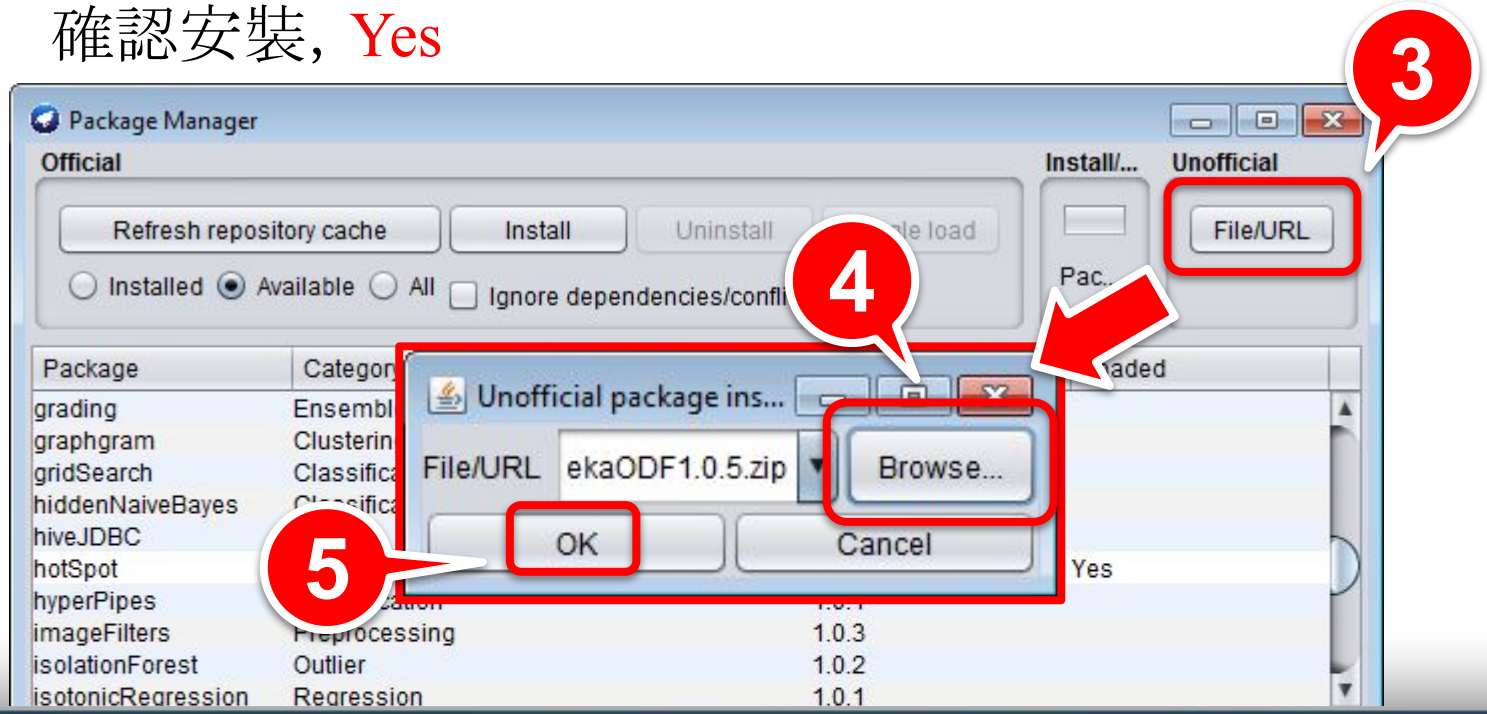
WekaODF

1. 在課程網頁上，下載WekaODF1.0.5.zip檔案

STEP 2. 安裝自製套件 (2/2)

WekaODF

2. 開啟Package Manager主視窗
3. 在右上角Unofficial按下File/URL
4. 按Browse，選擇剛剛下載的WekaODF1.0.5.zip檔案
5. 確認安裝，Yes



STEP 3. 關閉Weka，再重新啟動

因為安裝了套件

請重新啟動Weka吧

安裝套件

1. 安裝官方套件
   - cascadeKMeans
     分群演算法
   - localOutlierFactor
     異常偵測演算法
2. 安裝自製套件
   - WekaODF
3. 關閉Weka, 再重新啟動

(只有Windows作業系統需要設定)

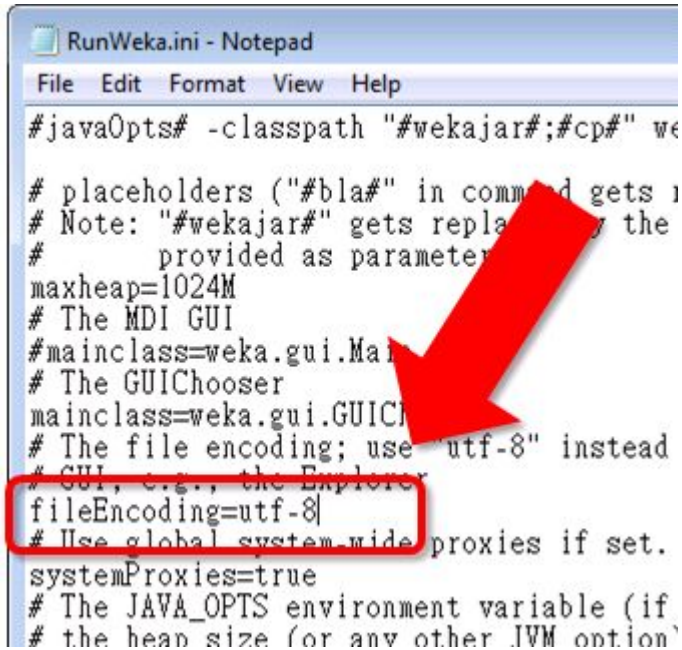
讓Weka能夠讀取中文

1. 開啟Weka安裝目錄, 預設為 C:\Program Files\Weka-[版本號]
2. 用文字編輯器開啟RunWeka.ini
3. 將以下設定

   fileEncoding=**Cp1252**

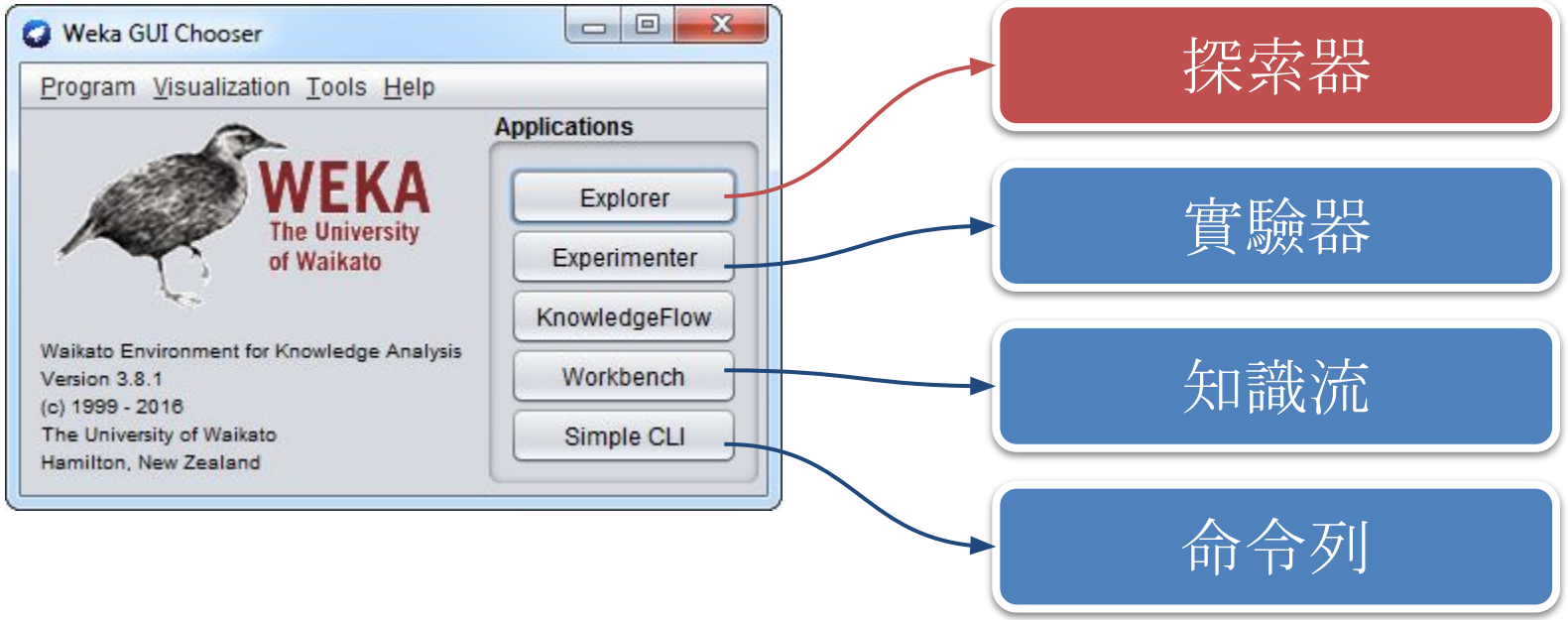
   改成

   fileEncoding=**utf-8**
4. 儲存, 重新啟動Weka



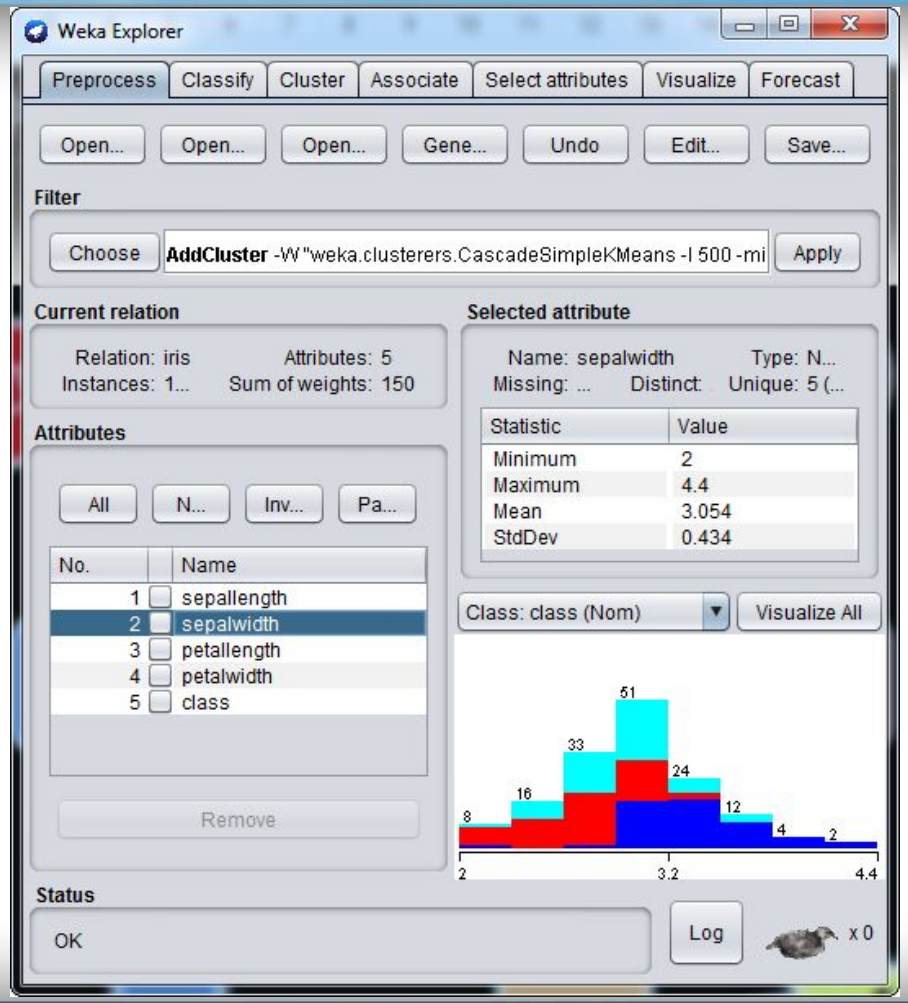
詳細操作請看Blog: 如何在Weka中顯示中文

Weka的功能架構


Weka GUI Chooser
Program Visualization Tools Help
Applications
WEKA
The University of Waikato
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
Waikato Environment for Knowledge Analysis
Version 3.8.1
(c) 1999 - 2016
The University of Waikato
Hamilton, New Zealand
探索器
實驗器
知識流
命令列

Explorer
探索器

- Weka主要的圖形化使用者介面
- 以頁籤、下拉式選單、欄位設定等表單元件，讓使用者輕易進行資料分析與探勘
- 直接提供各種視覺化圖表，展現分析結果
- 一次只能分析一份資料集

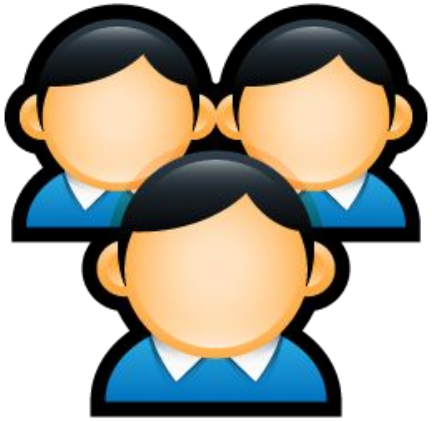


Part 3.

分群

實作資料集
學生成績資料集 (1/3)

- 學生成績資料集是Cortez等人(2008)年從兩所葡萄牙學校蒐集649位學生、33種屬性的開放資料集
- 屬性包括學生個人資料、家庭狀況、就學狀況、學校生活、課堂表現



※ 本教學取其資料集內容, 因應教學內容而作調整

http://archive.ics.uci.edu/ml/datasets/Student+Performance

實作資料集
學生成績資料集 (2/3)

Nominal Type

類別型屬性

共15種

- 性別
- 就學理由
- 是否補習
- 學校

Numeric Type

數值型屬性

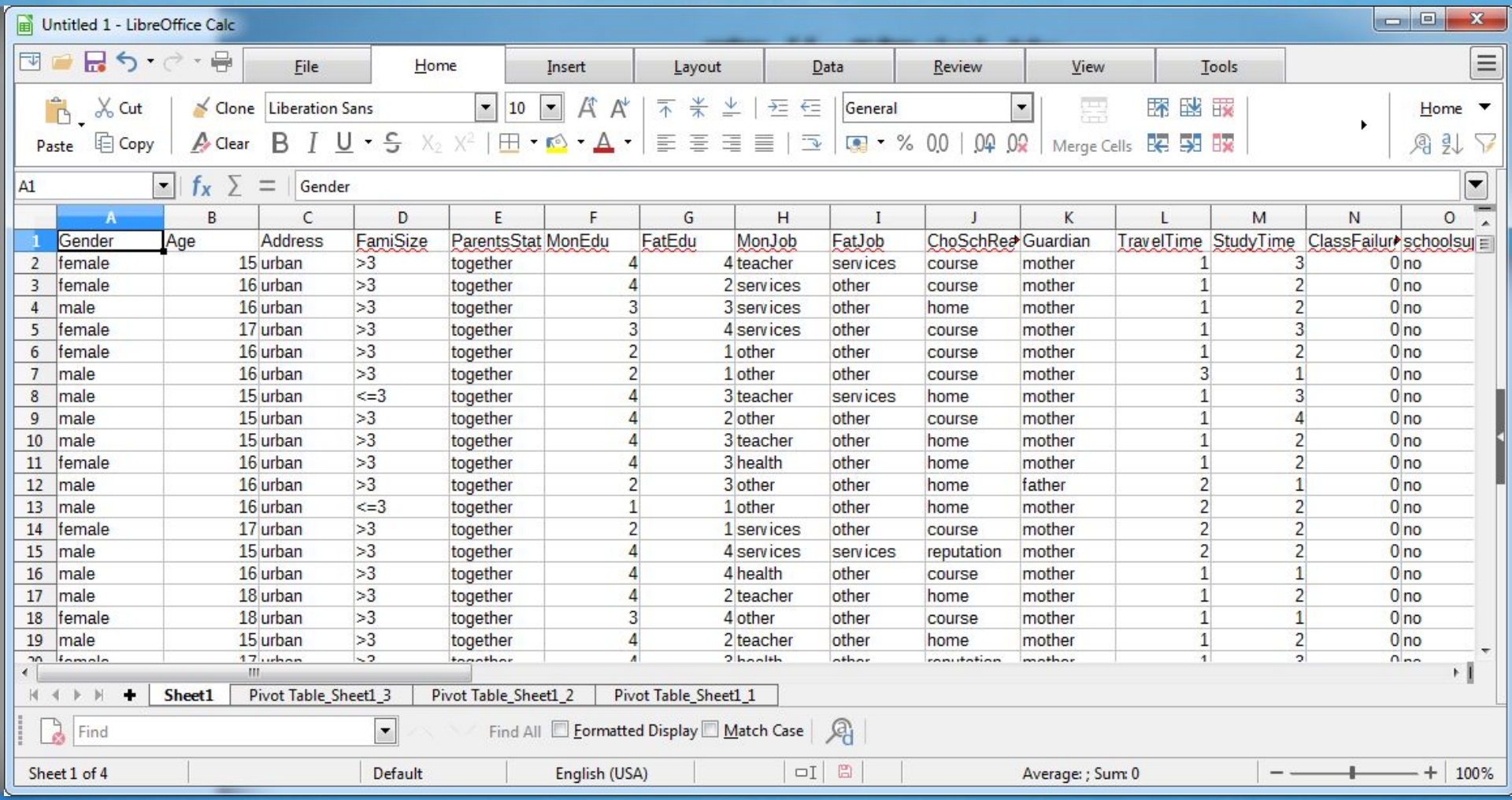
共18種

- 年齡
- 雙親教育程度
- 缺席次數
- 課堂成績

※ 完整的屬性說明請看論文

http://archive.ics.uci.edu/ml/datasets/Student+Performance

實作資料集
學生成績資料集 (3/3)



Untitled 1 - LibreOffice Calc

File Home Insert Layout Data Review View Tools

Cut Copy Paste Clone Clear Liberation Sans 10 General Merge Cells Home

A1 Gender

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Gender	Age	Address	FamiSize	ParentsStat	MonEdu	FatEdu	MonJob	FatJob	ChoSchRea	Guardian	TravelTime	StudyTime	ClassFailur	schoolsu
2	female	15	urban	>3	together	4	4	teacher	services	course	mother	1	3	0	no
3	female	16	urban	>3	together	4	2	services	other	course	mother	1	2	0	no
4	male	16	urban	>3	together	3	3	services	other	home	mother	1	2	0	no
5	female	17	urban	>3	together	3	4	services	other	course	mother	1	3	0	no
6	female	16	urban	>3	together	2	1	other	other	course	mother	1	2	0	no
7	male	16	urban	>3	together	2	1	other	other	course	mother	3	1	0	no
8	male	15	urban	<=3	together	4	3	teacher	services	home	mother	1	3	0	no
9	male	15	urban	>3	together	4	2	other	other	course	mother	1	4	0	no
10	male	15	urban	>3	together	4	3	teacher	other	home	mother	1	2	0	no
11	female	16	urban	>3	together	4	3	health	other	home	mother	1	2	0	no
12	male	16	urban	>3	together	2	3	other	other	home	father	2	1	0	no
13	male	16	urban	<=3	together	1	1	other	other	home	mother	2	2	0	no
14	female	17	urban	>3	together	2	1	services	other	course	mother	2	2	0	no
15	male	15	urban	>3	together	4	4	services	services	reputation	mother	2	2	0	no
16	male	16	urban	>3	together	4	4	health	other	course	mother	1	1	0	no
17	male	18	urban	>3	together	4	2	teacher	other	home	mother	1	2	0	no
18	female	18	urban	>3	together	3	4	other	other	course	mother	1	1	0	no
19	male	15	urban	>3	together	4	2	teacher	other	home	mother	1	2	0	no

Sheet1 Pivot Table_Sheet1_3 Pivot Table_Sheet1_2 Pivot Table_Sheet1_1

Find Find All Formatted Display Match Case

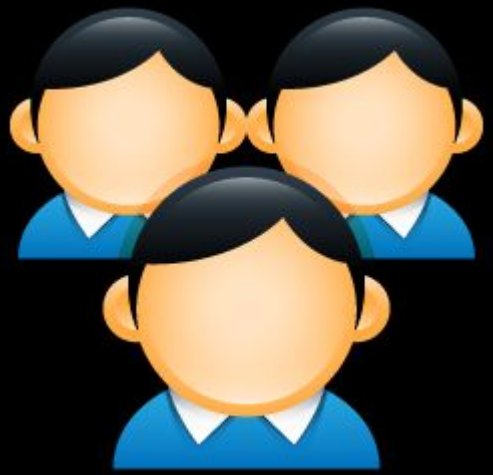
Sheet 1 of 4 Default English (USA) Average: ; Sum: 0 100%

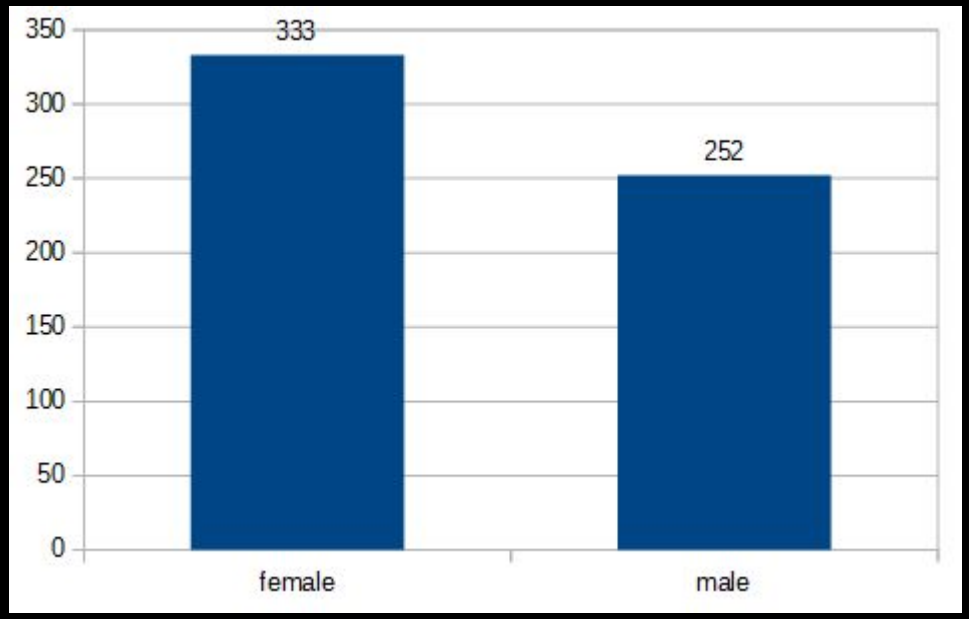
http://archive.ics.uci.edu/ml/datasets/Student+Performance

老闆交代的任務

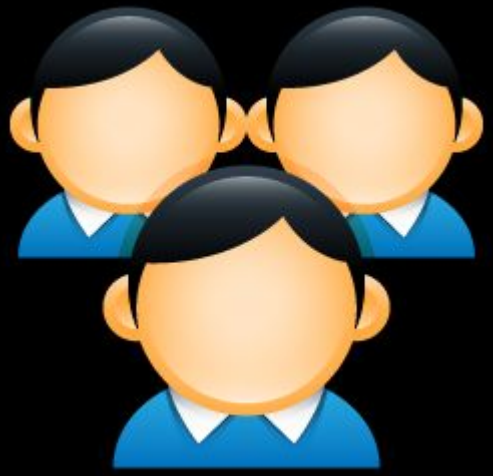
「你能簡單地描述一下這群學生嗎？」

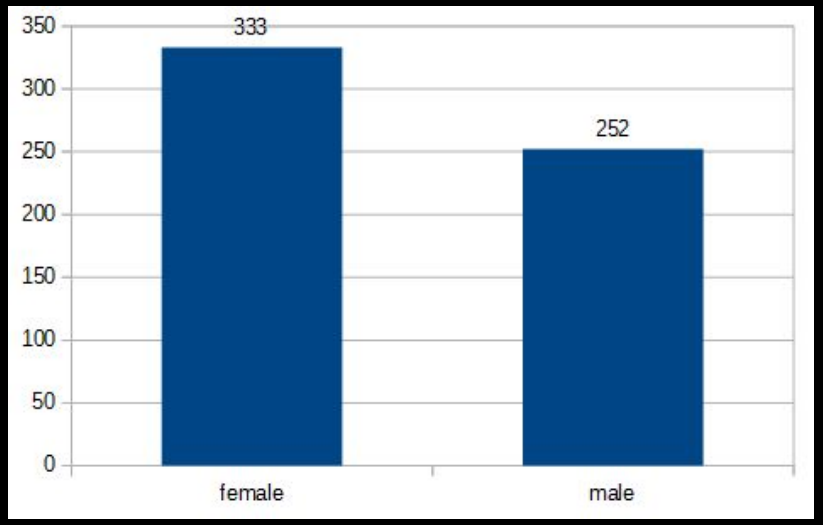
如何描述所有學生的
性別屬性？

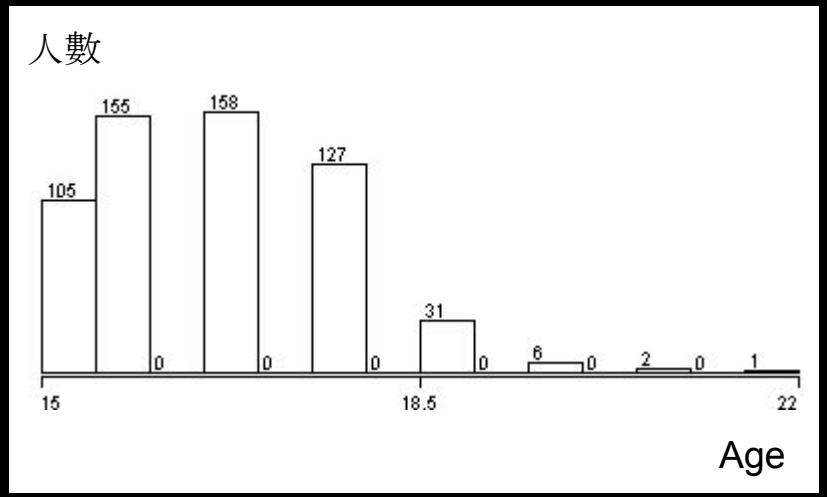




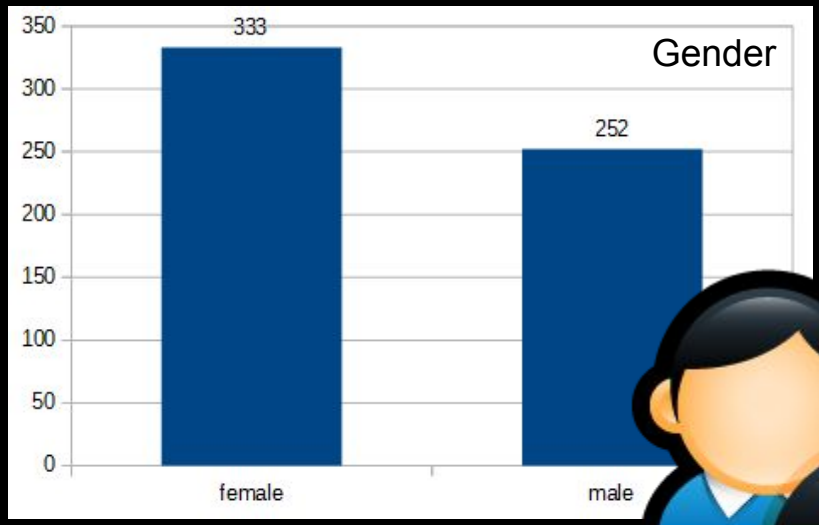
如何描述所有學生的
性別與年齡屬性？

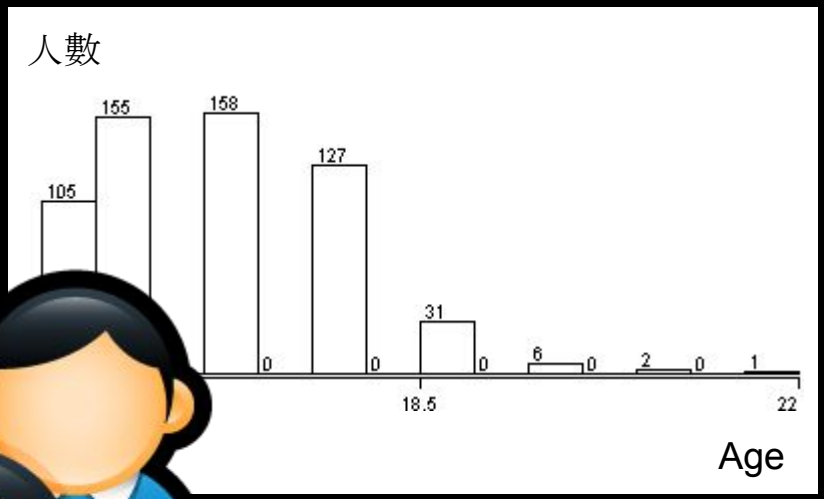


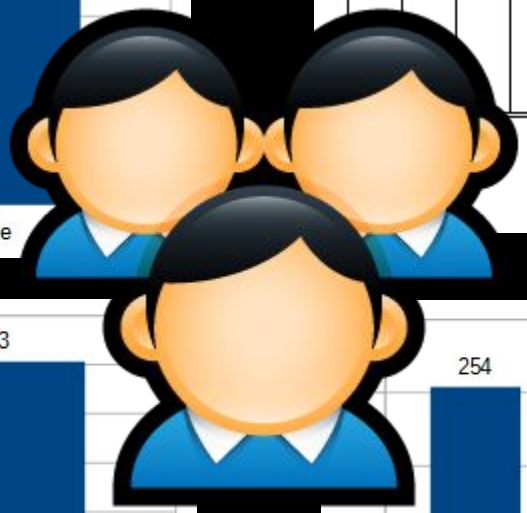


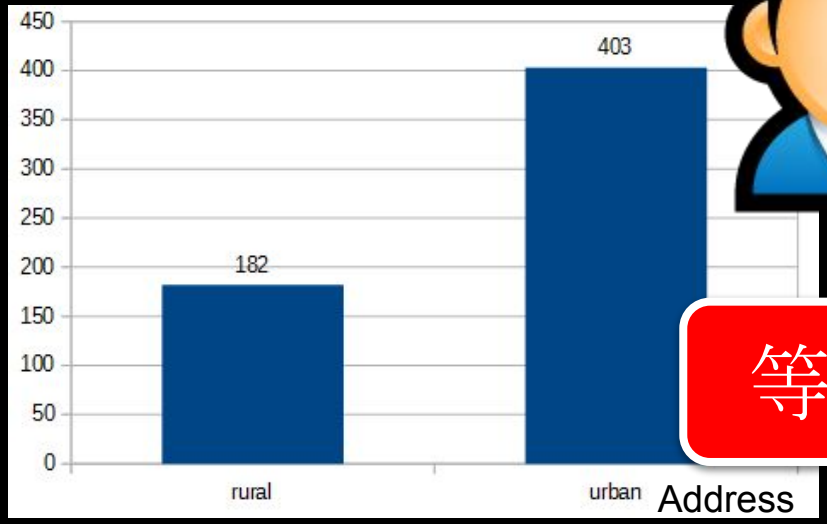


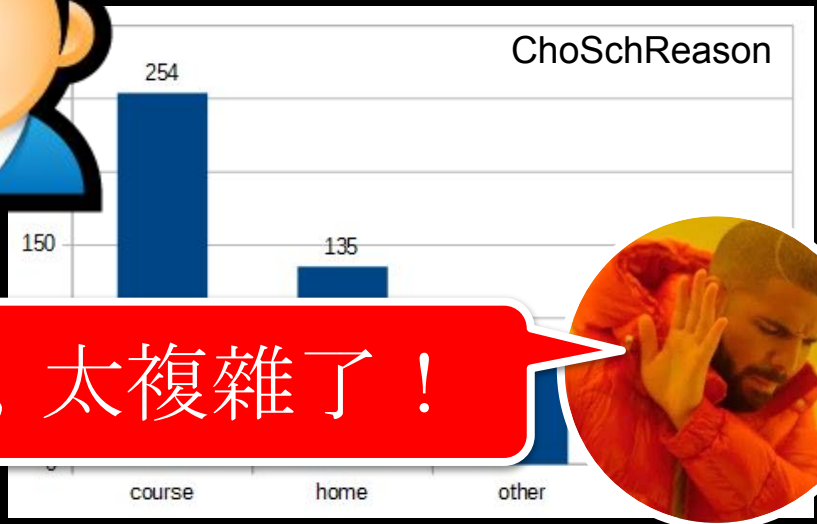
如何描述所有學生的
性別、年齡、住處等 30種屬性？

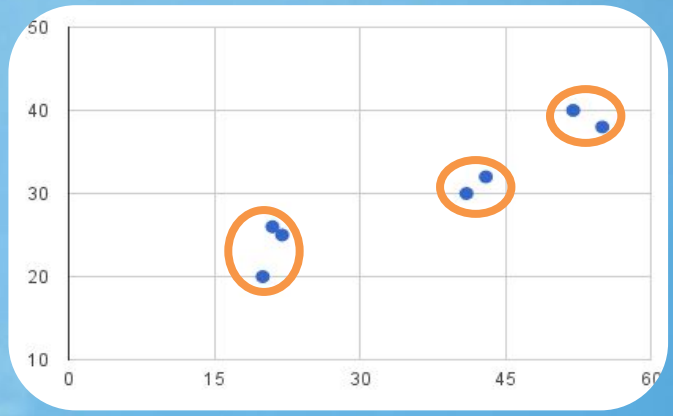










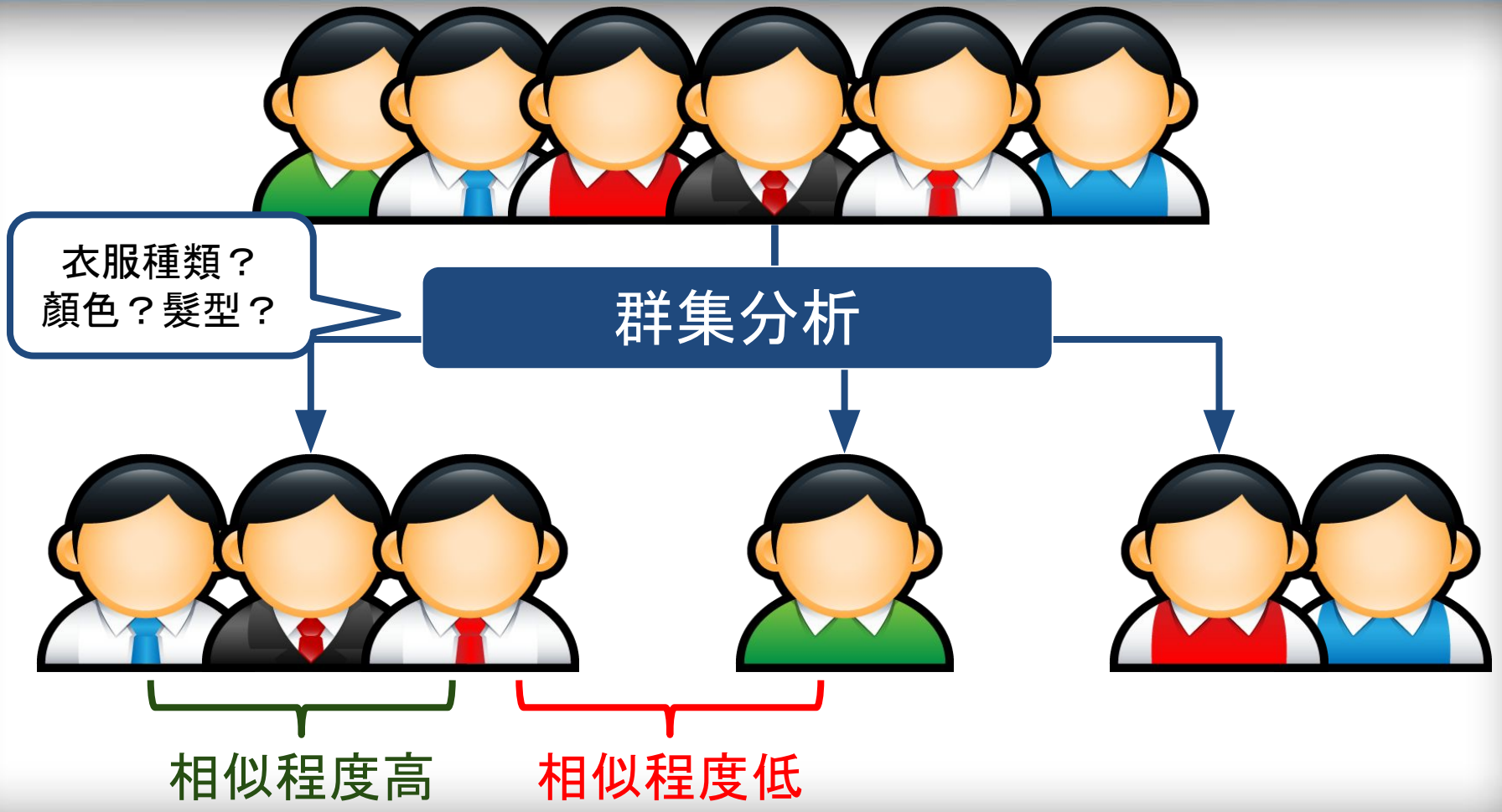


分群演算法

K Means

K平均法

發掘消費者之間的隱藏關聯



線上購物網站的使用者族群與消費能力

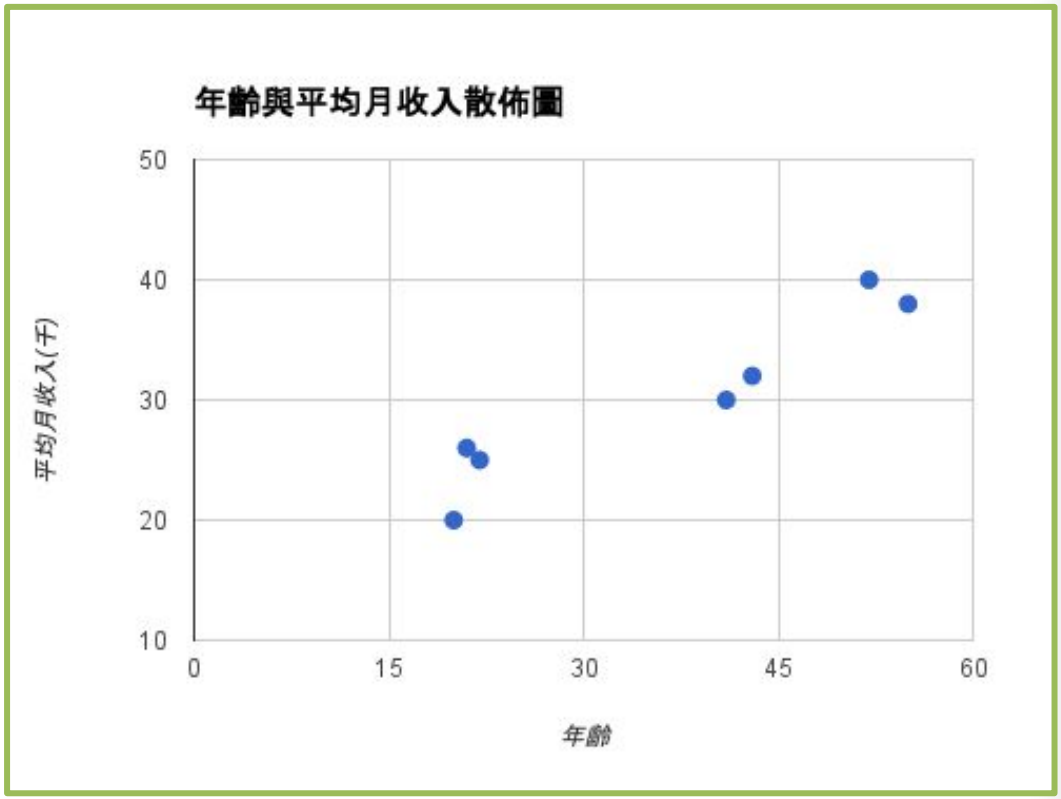
會員編號	年齡	平均月收入 (千)
1	20	20
2	21	26
3	22	25
4	41	30
5	43	32
6	52	40
7	55	38

要怎麼描述這群人呢？

能不能從這些資料看出什麼特徵呢？

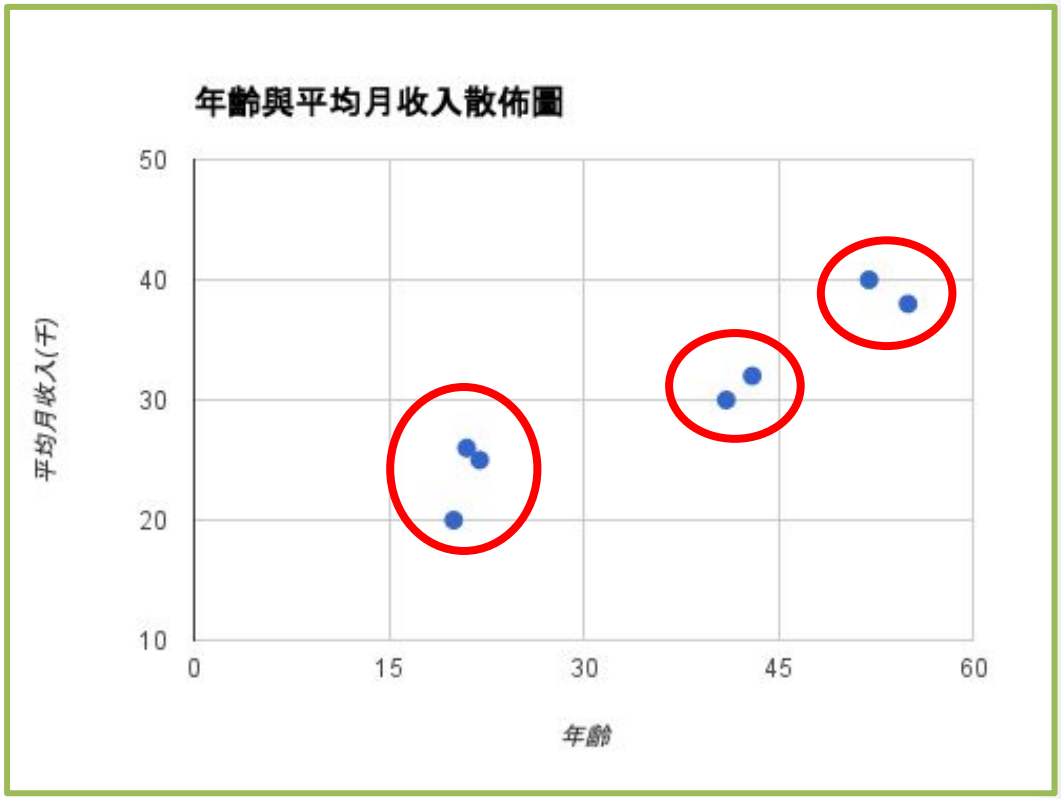
散佈圖探索性分析 (1/2)

會員	年齡	平均月收入(千)
1	20	20
2	21	26
3	22	25
4	41	30
5	43	32
6	52	40
7	55	38



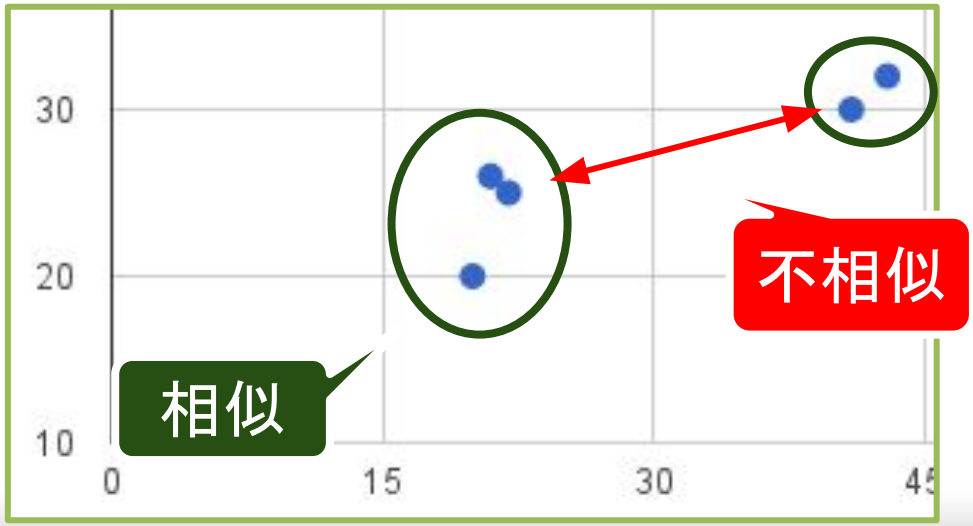
散佈圖探索性分析 (2/2)

會員	年齡	平均月收入(千)
1	20	20
2	21	26
3	22	25
4	41	30
5	43	32
6	52	40
7	55	38



群集分析的概念與目的

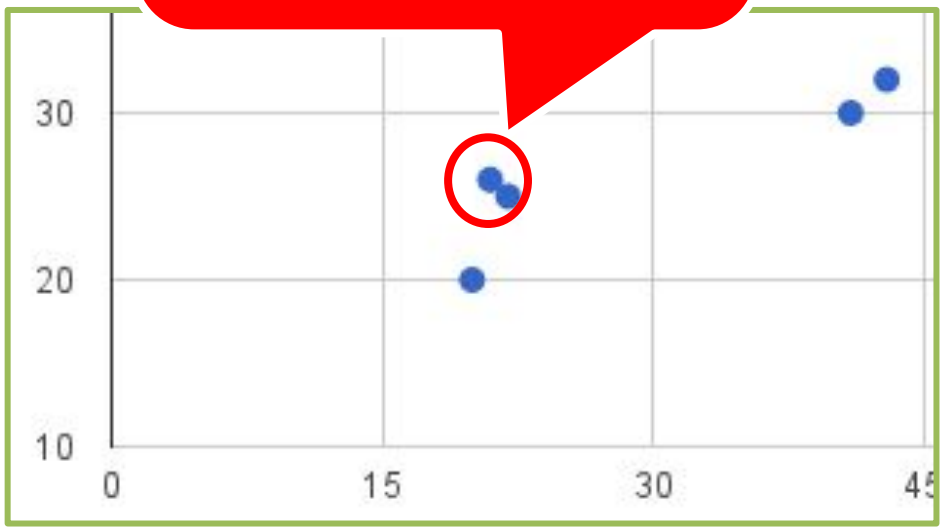
- 將資料集合中的資料記錄(資料點)加以分群成數個群集(cluster)
- 使得每個群集中的資料點間相似程度高於與其它群集中資料點的相似程度
- 從群集結果推論出有用、隱含、令人感興趣的特性和現象



資料點的表示法

會員	年齡	平均月收入(千)
1	20	20
2	21	26
3	22	25
4	41	30
5	43	32
6	52	40
7	55	38

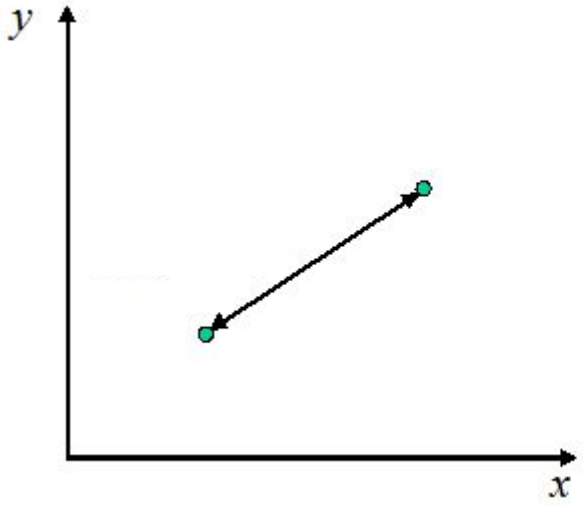
資料點
x_2=<21,26>



相似度的計算

歐基里得距離 (Euclidean distance)

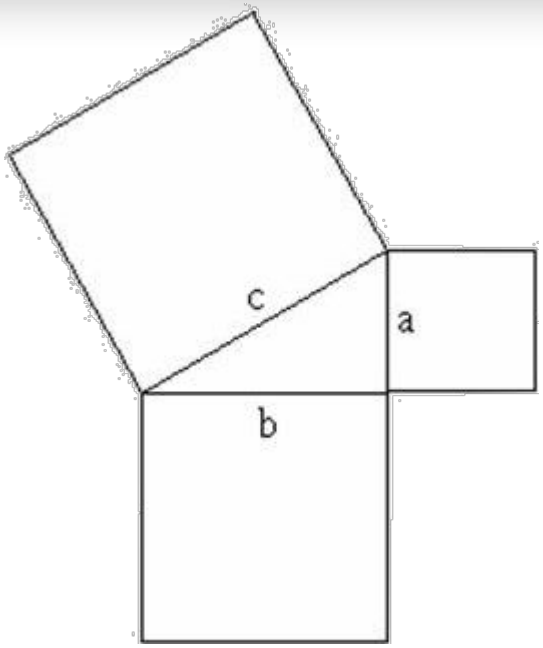
資料點$x_i=<x_{i1},x_{i2},\ldots,x_{ik}>$和資料點$x_j=<x_{j1},x_{j2},\ldots,x_{jk}>$之間的歐基里得距離算法

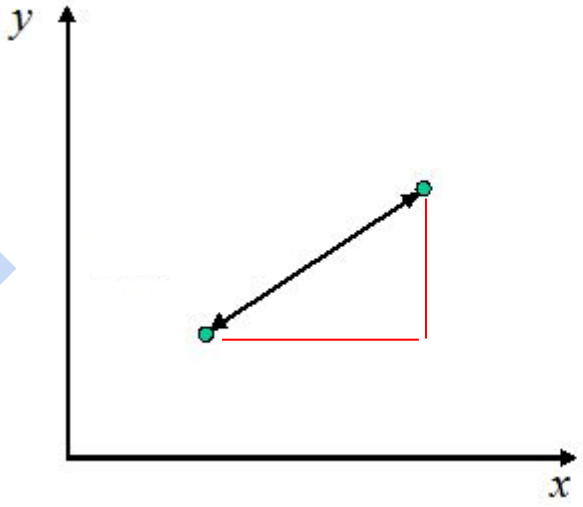


$$d_2(x_i, x_j) = \left(\sum_{d=1}^{k} |x_{id} - x_{jd}|^2\right)^{1/2}$$

$$= \|x_i - x_j\|_2 \left(\|x_i - x_j\|\right)$$

畢氏定理⇨歐基里得距離

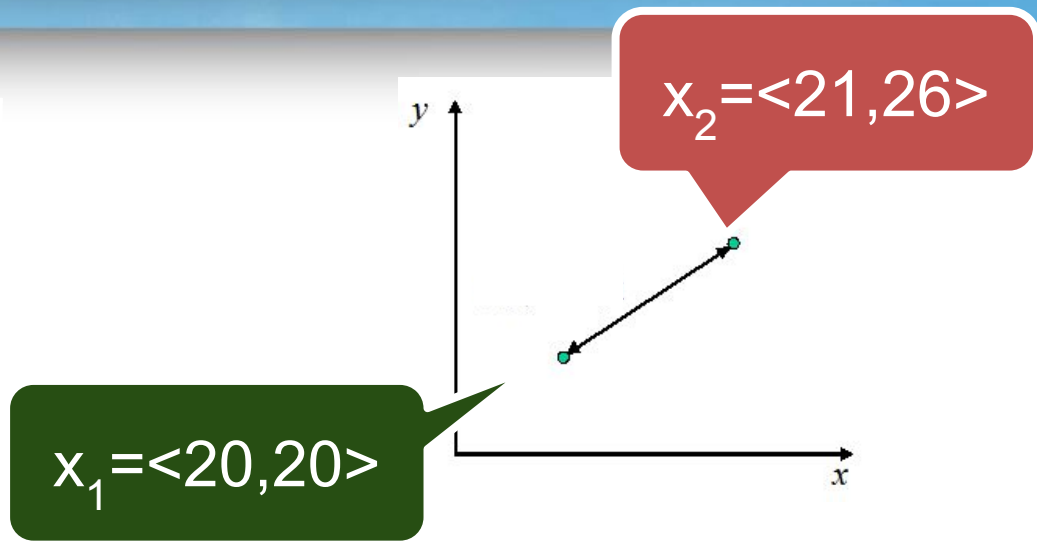




$c^2=a^2+b^2$

歐基里得距離計算

會員	年齡	平均月收入(千)
1	20	20
2	21	26



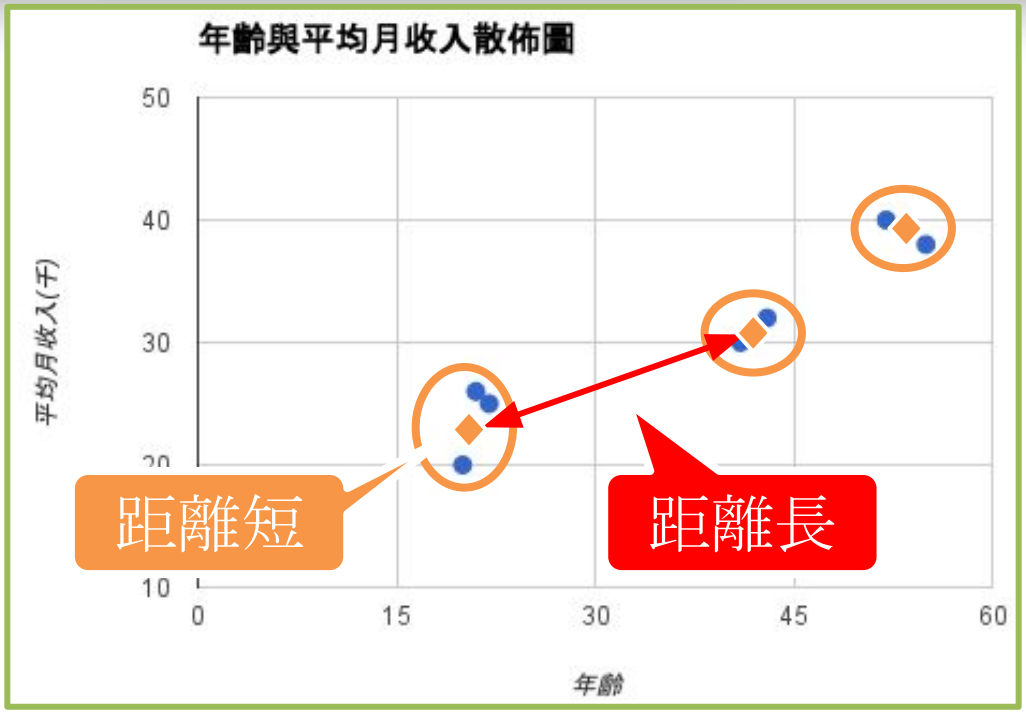
會員x_1=<20, 20> 與

會員x_2=<21, 26>

之間的歐基里得距離為：

$$d_2(x_1, x_2) = \sqrt{(21-20)^2 + (26-20)^2} \approx 6$$

K平均法
演算法目標

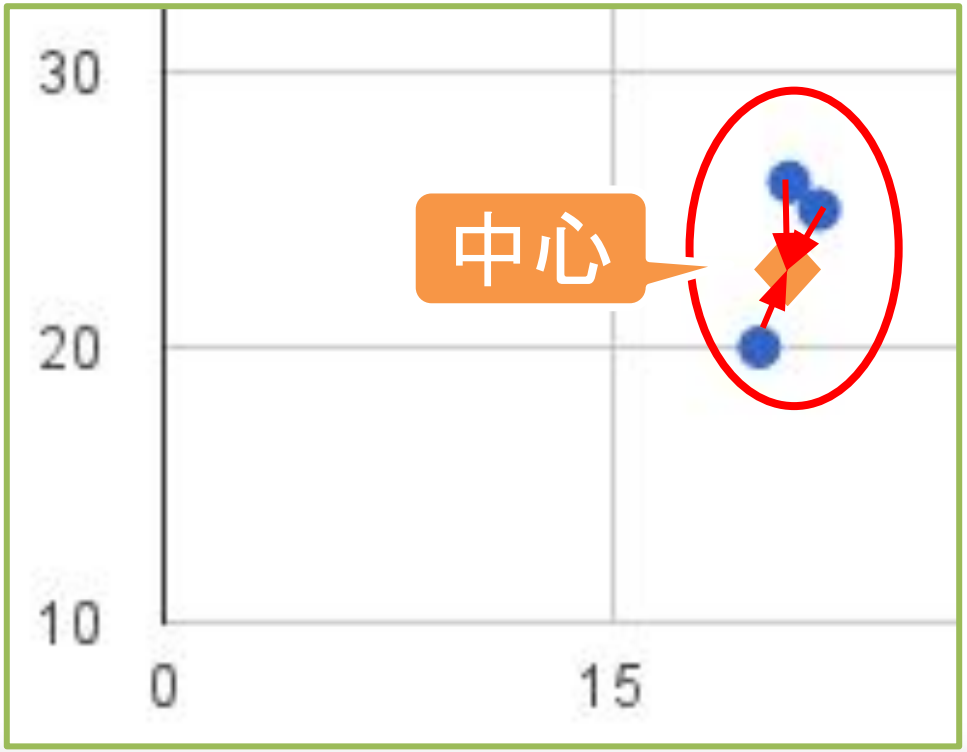


將資料點分成k個分群

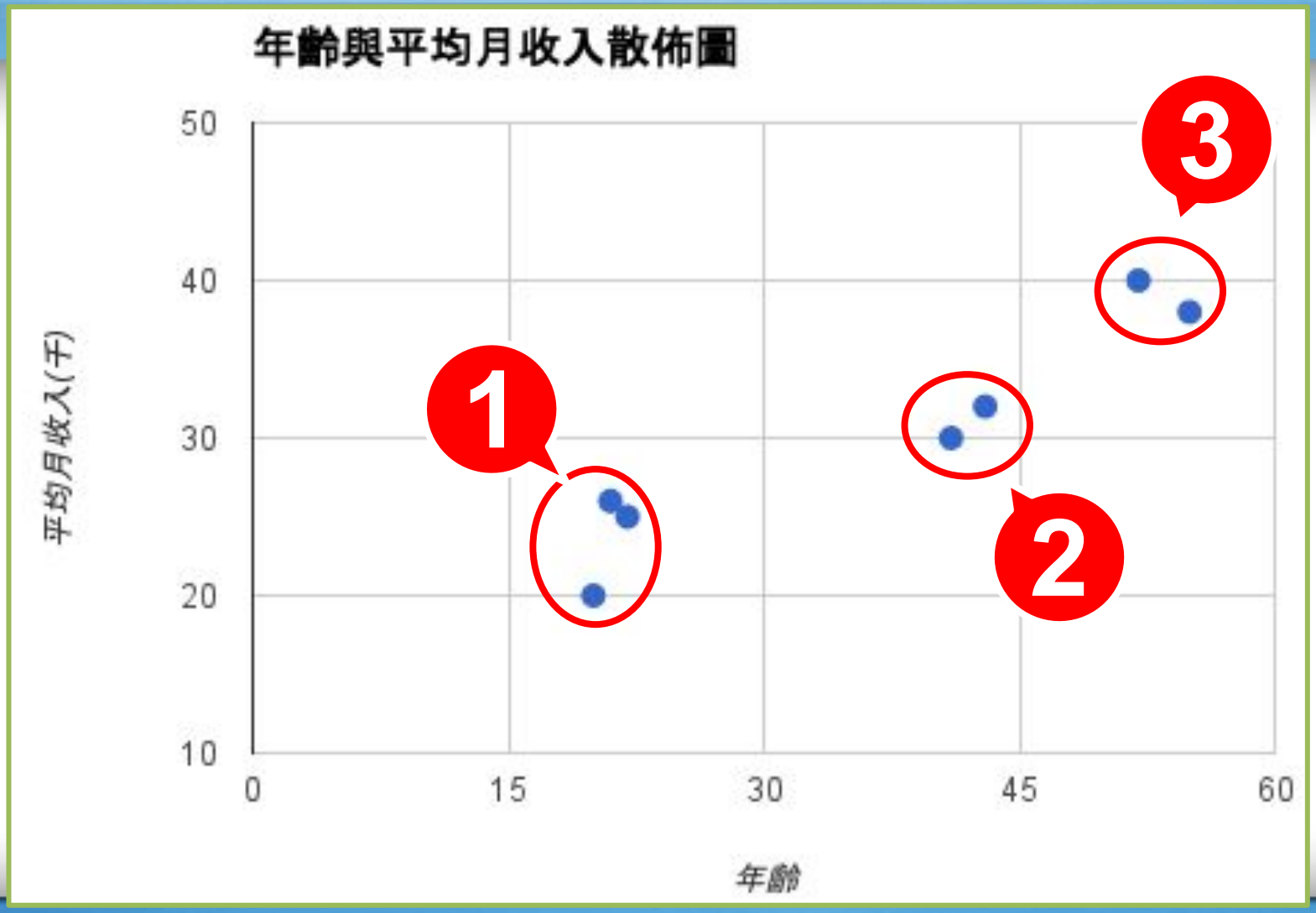
各分群的群集中心與其資料點距離最短、群集之間距離最長

群集中心

群集中心即是每個分群的**平均數**



「k」=3，3個分群



K平均法
演算法流程 (1/2)

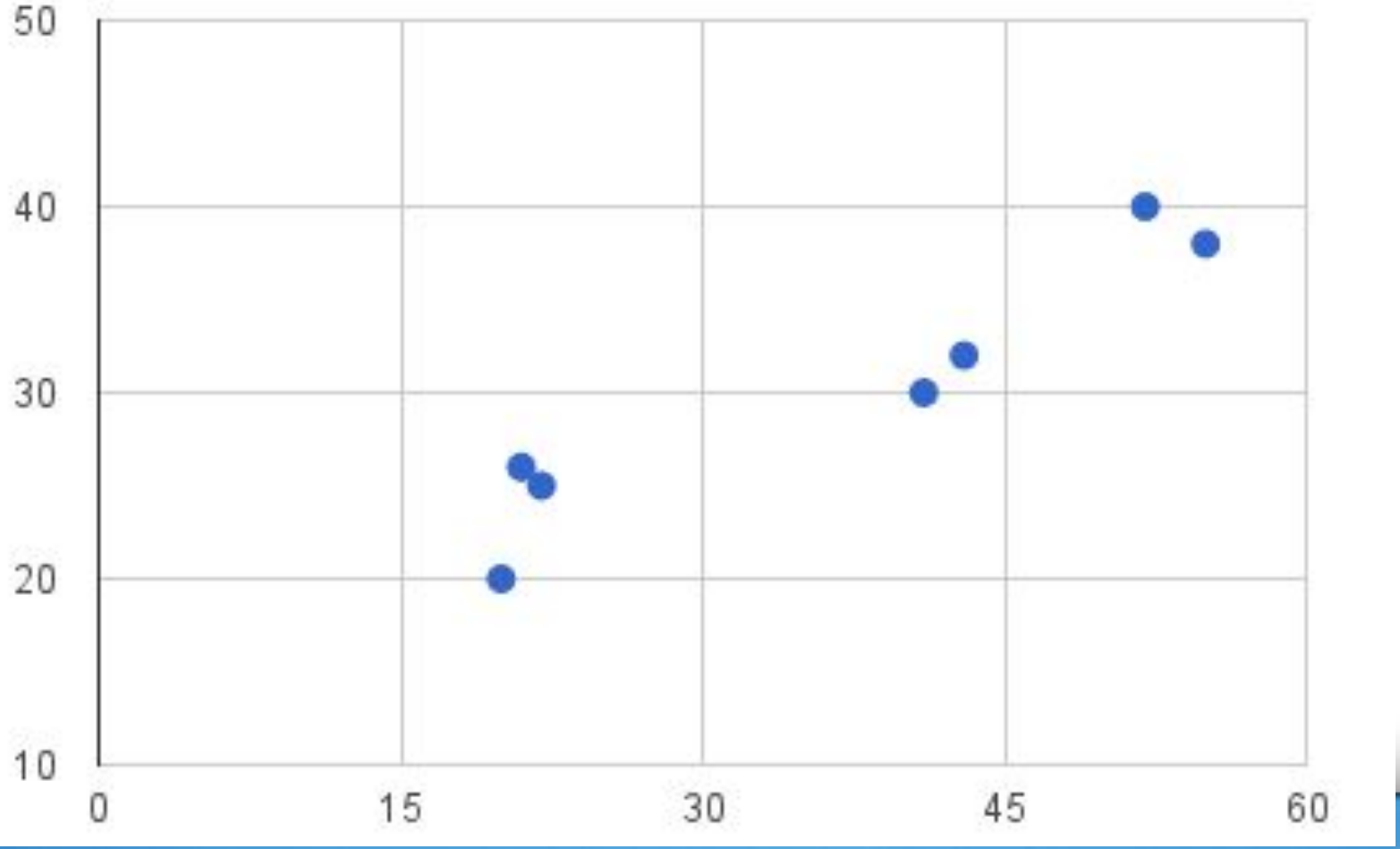
輸入：**資料集合、使用者定義之群集數量 *k***

輸出：***k*個互不交集的群集**

1. 隨機從資料集合中選擇任*k*個資料點當作起始*k*群的群集中心
2. 利用相似度計算公式，將資料點分別歸屬到距其最近之群集中心所屬的群集，形成*k*個群集。
3. 利用各群集中所含的資料點，重新計算各群集之群集中心點
4. 條件判斷：
   a. 假如由步驟3所得到各群之群集中心與之前所計算之群集中心相同，則表示分群結果已穩定，並結束此處理程序並輸出各群結果
   b. 否則回到步驟2繼續執行

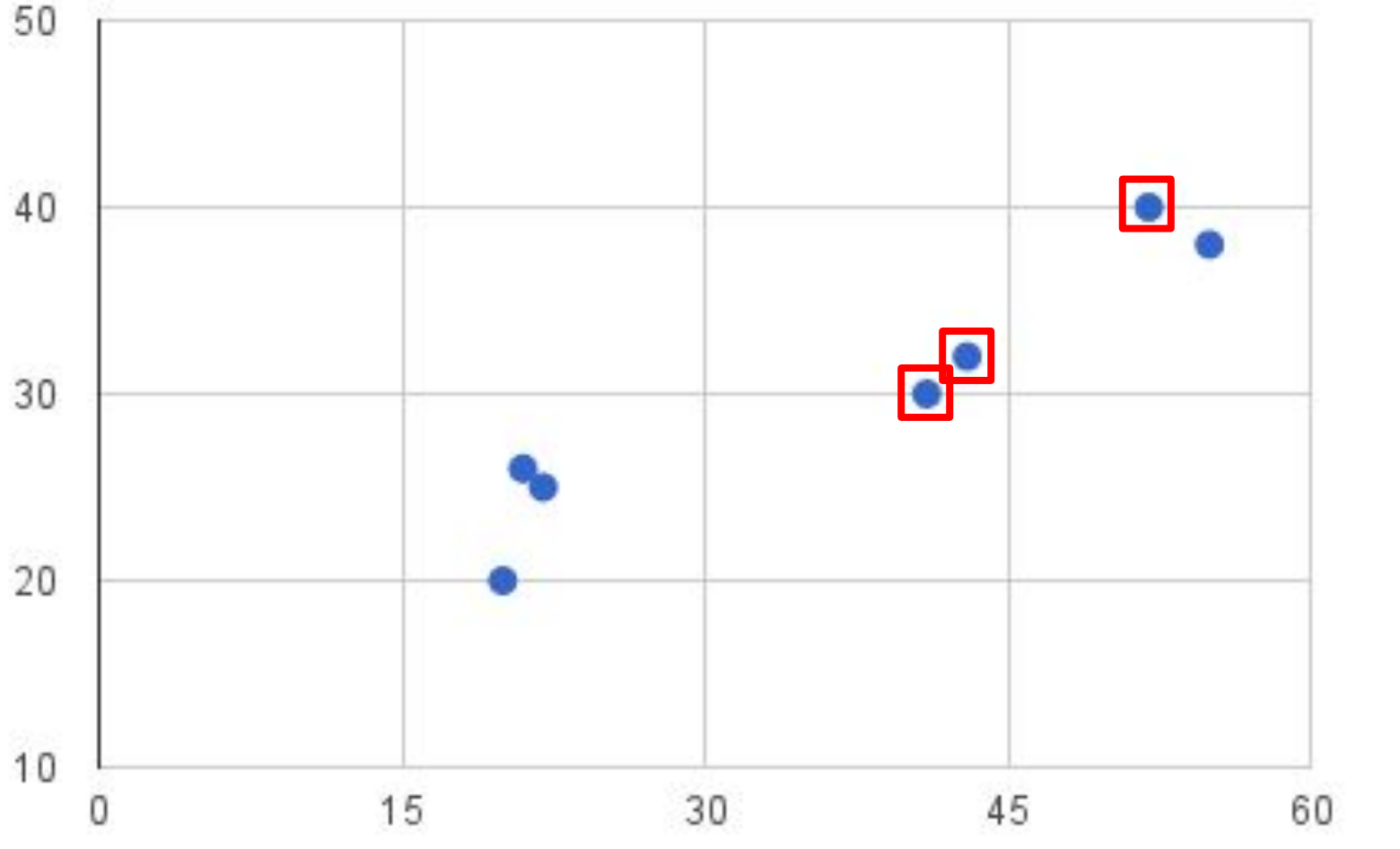
起始輸入

- 年齡與月收入資料集合
- 使用者定義之群集數量k = 3



1-1. 起始設置

隨機選擇任k個資料點當作起始k群的群集中心

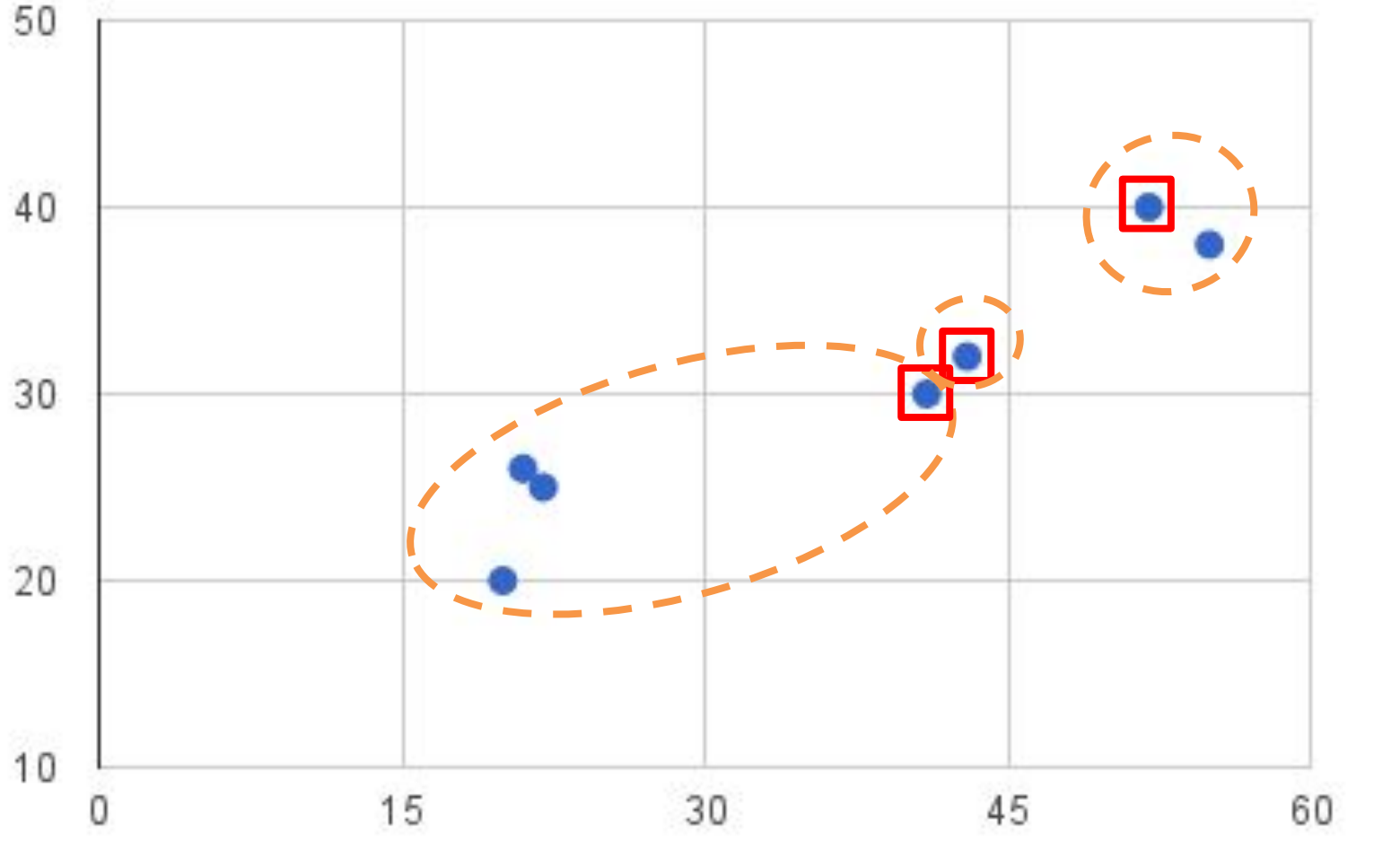


1-2. 形成群集

利用相似度計算公式,

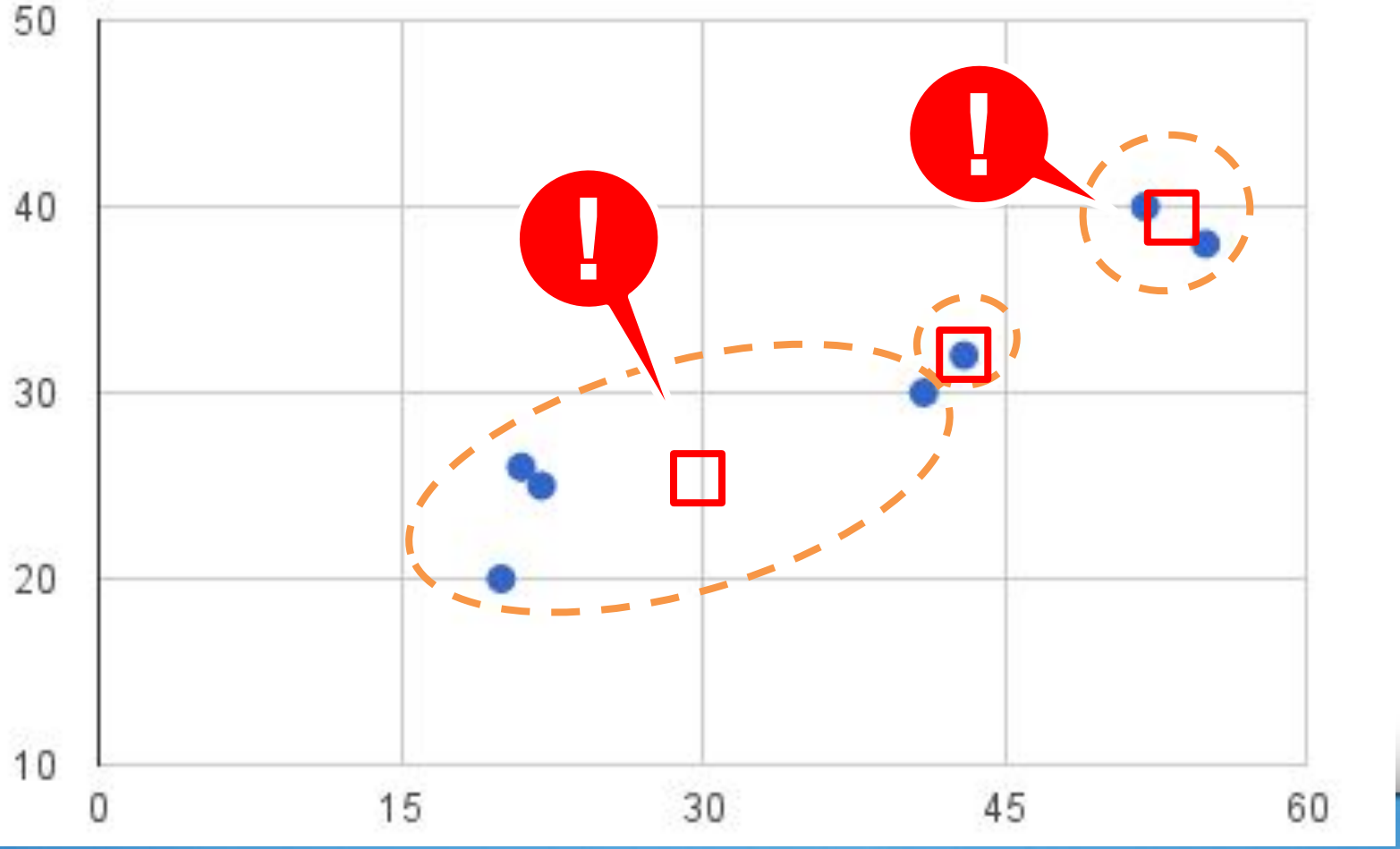
將資料點分別歸屬到距其最近之群集中心所屬的群集,

形成k=3個群集。



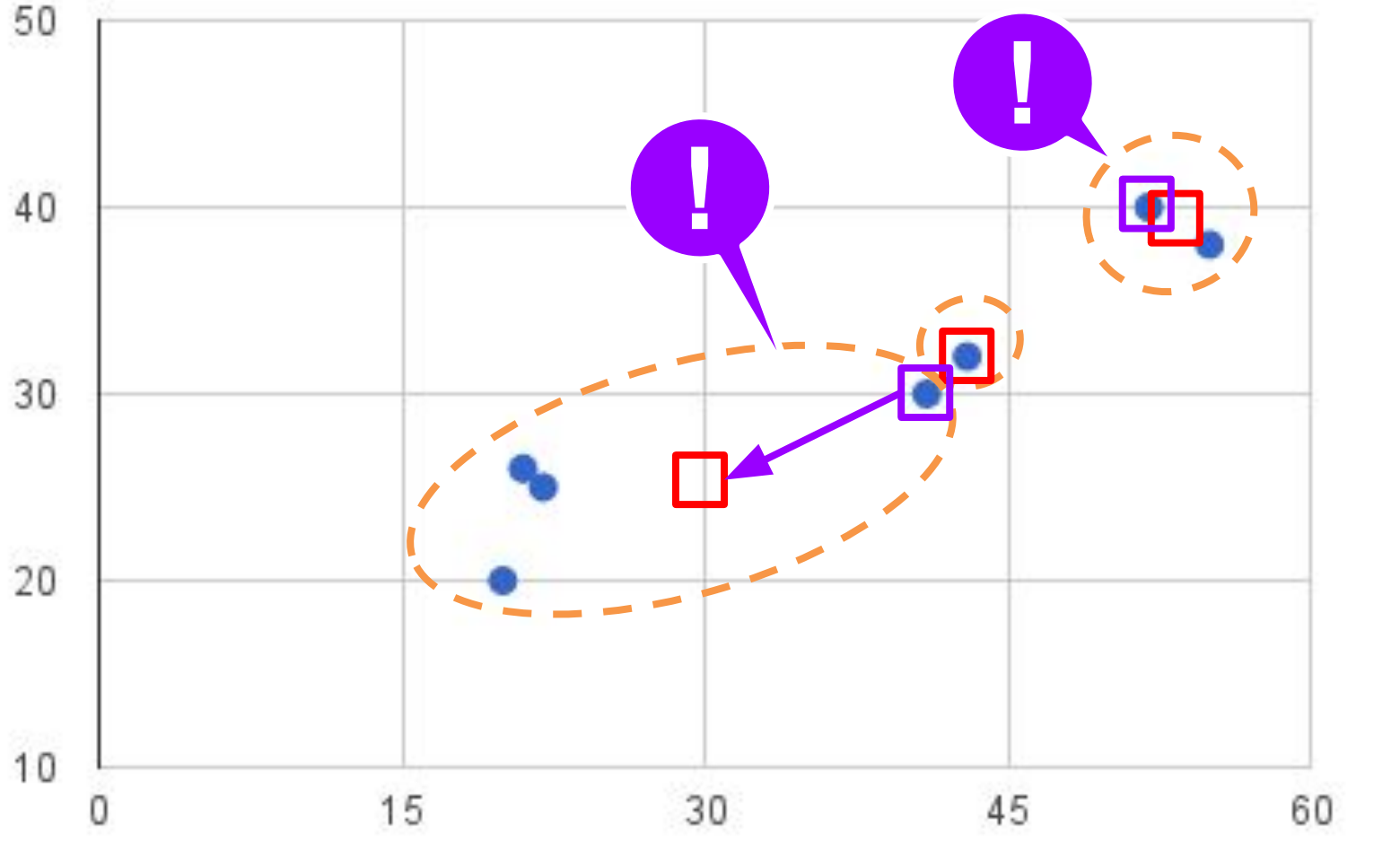
1-3. 計算群集中心

利用各群集中所含的資料點，重新計算各群集之群集中心點



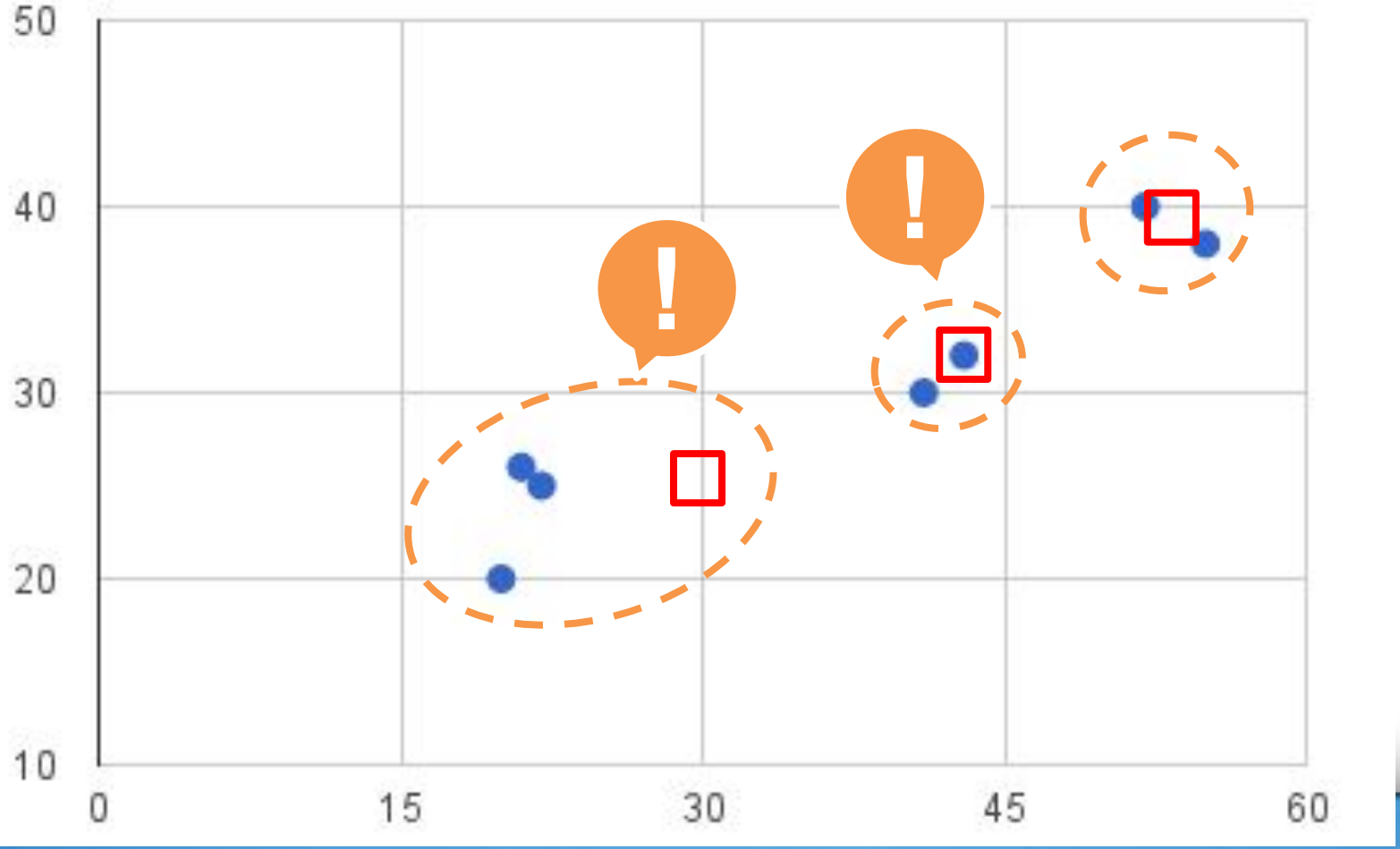
1-4. 結束條件判斷

現在的群集中心與之前的群集中心不同，故繼續執行



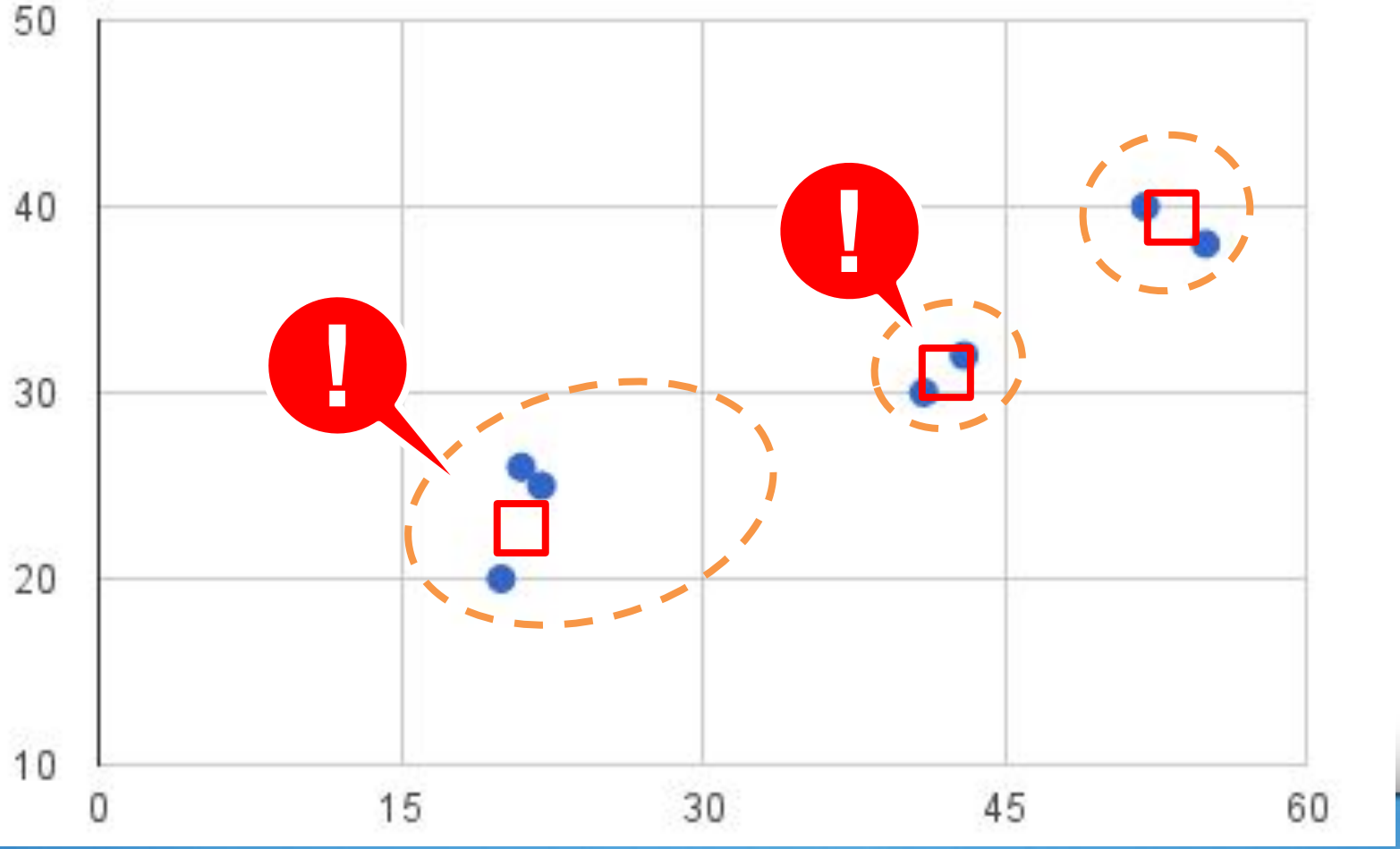
2-2. 形成群集

利用相似度計算公式,
將資料點分別歸屬到距其最近之群集中心所屬的群集,
形成k=3個群集。



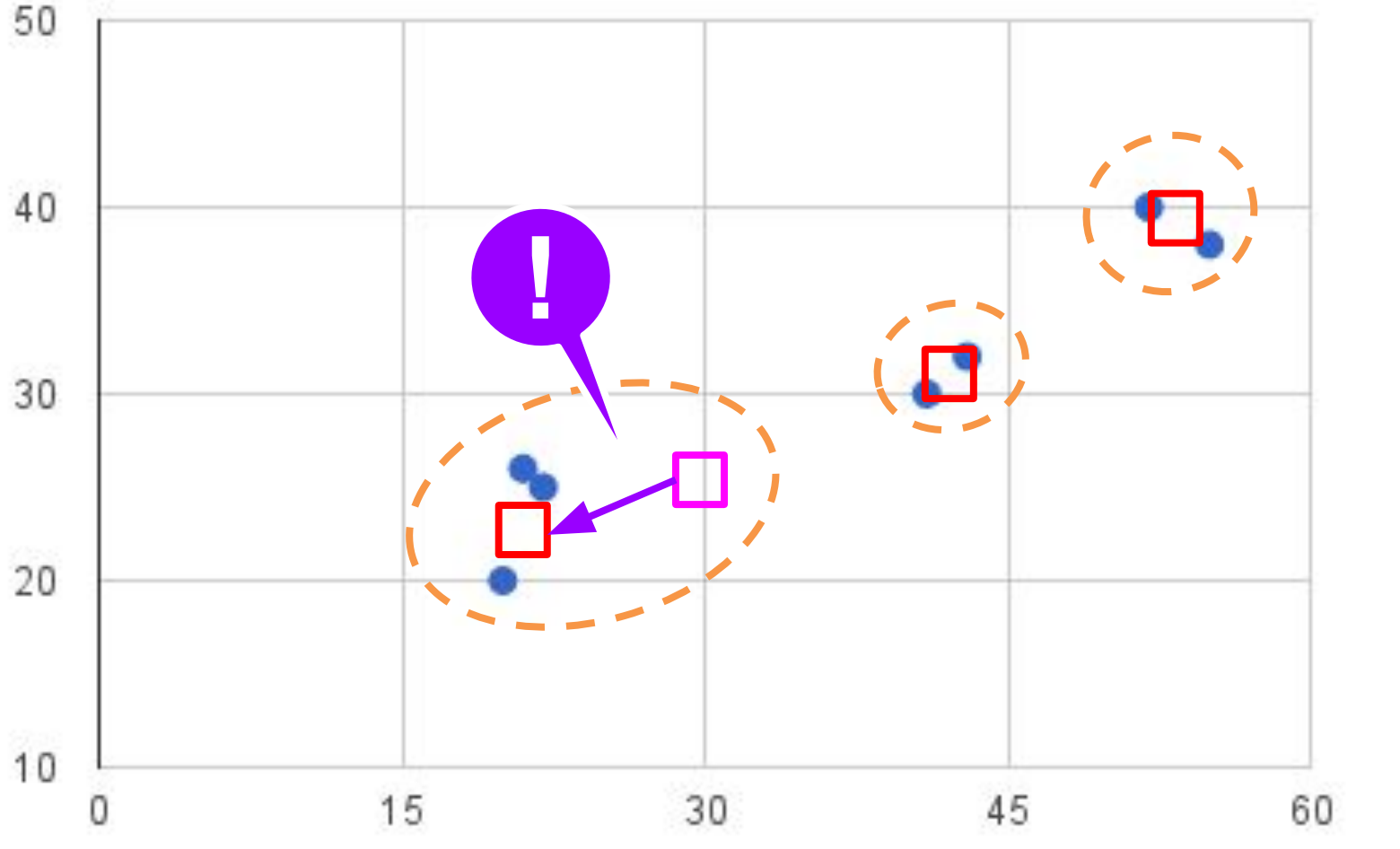
2-3. 計算群集中心

利用各群集中所含的資料點，重新計算各群集之群集中心點



2-4. 結束條件判斷

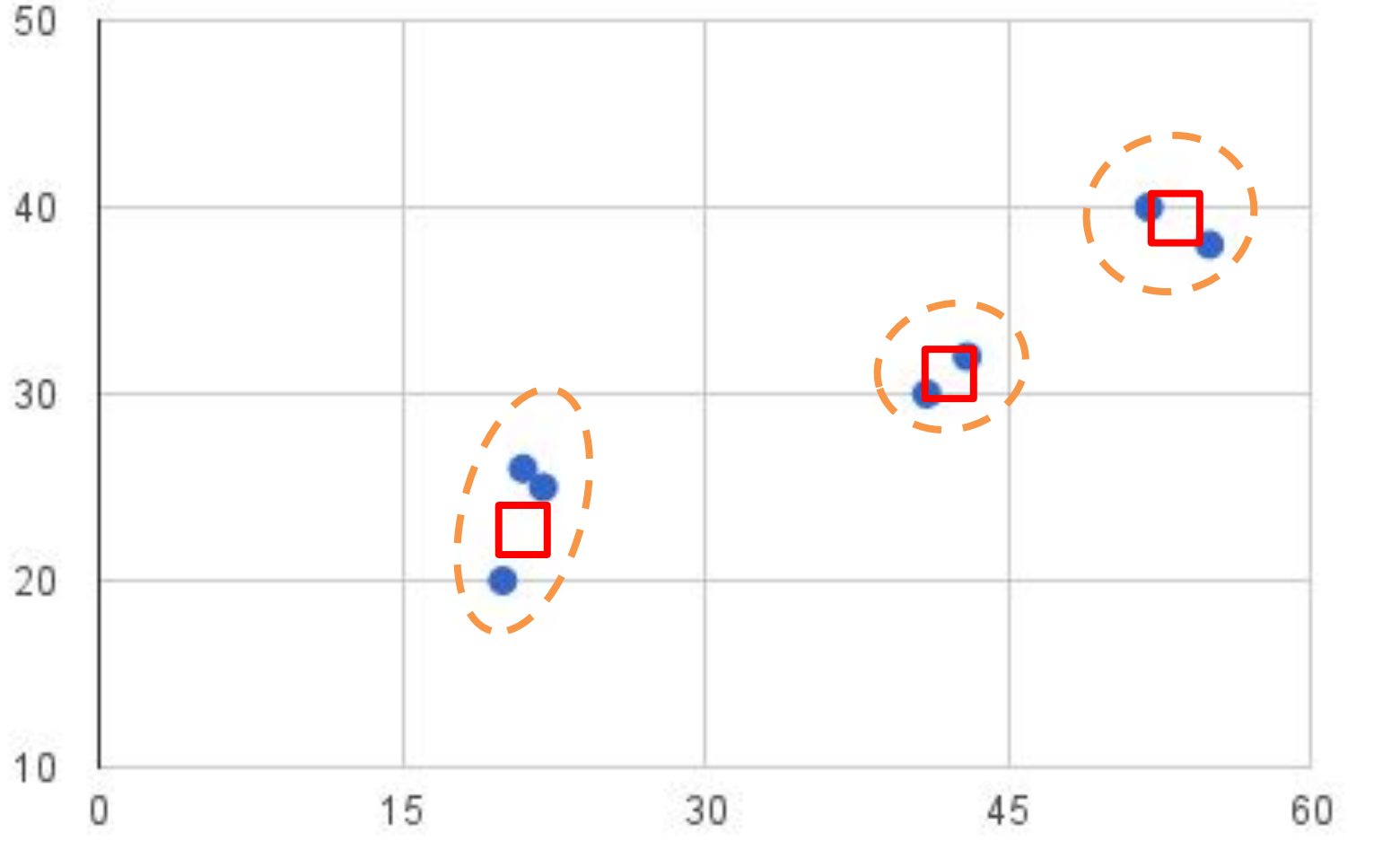
現在的群集中心與之前的群集中心不同，故繼續執行



3-3. 計算群集中心

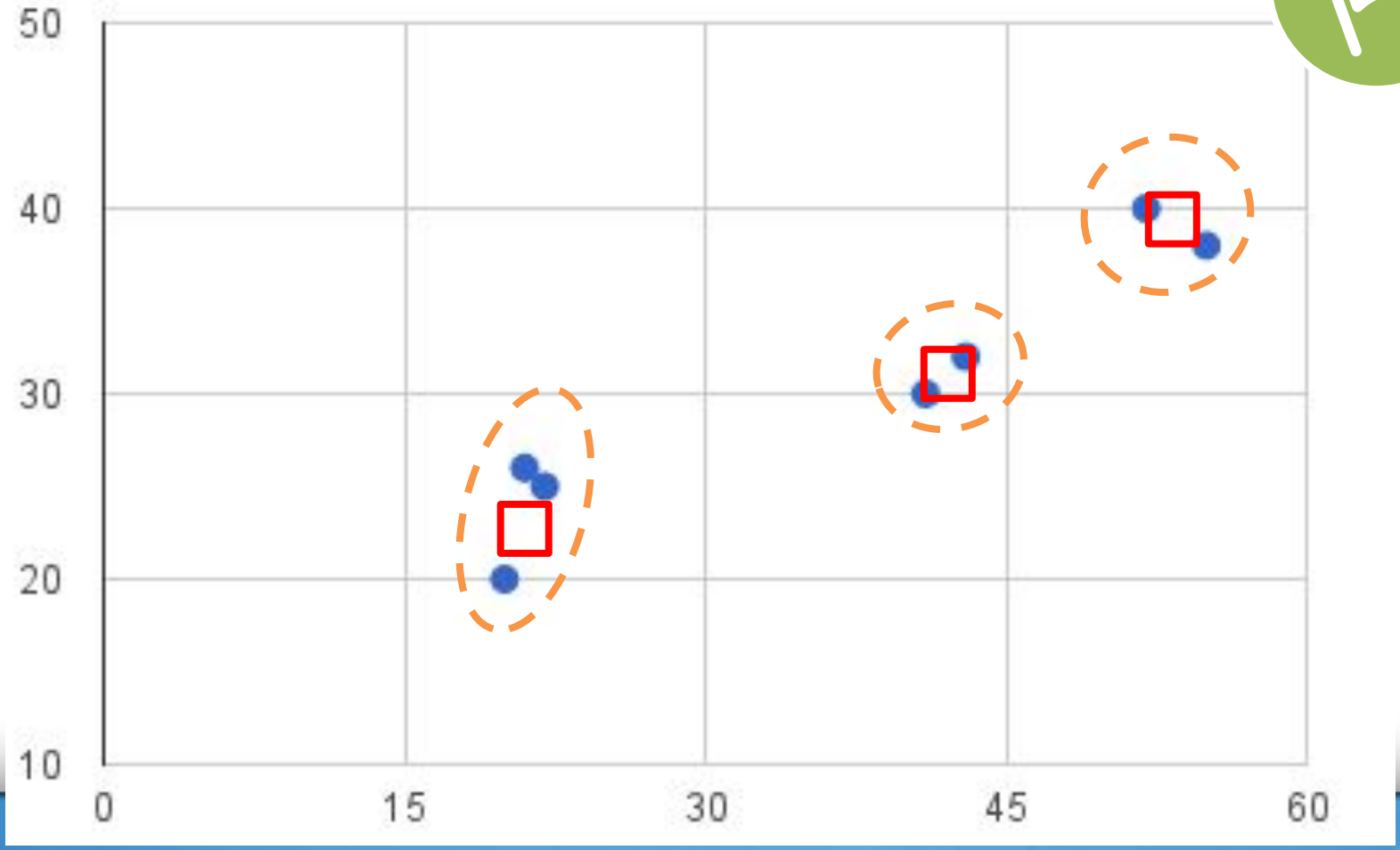
利用各群集中所含的資料點，重新計算各群集之群集中心點

(此階段沒改變)



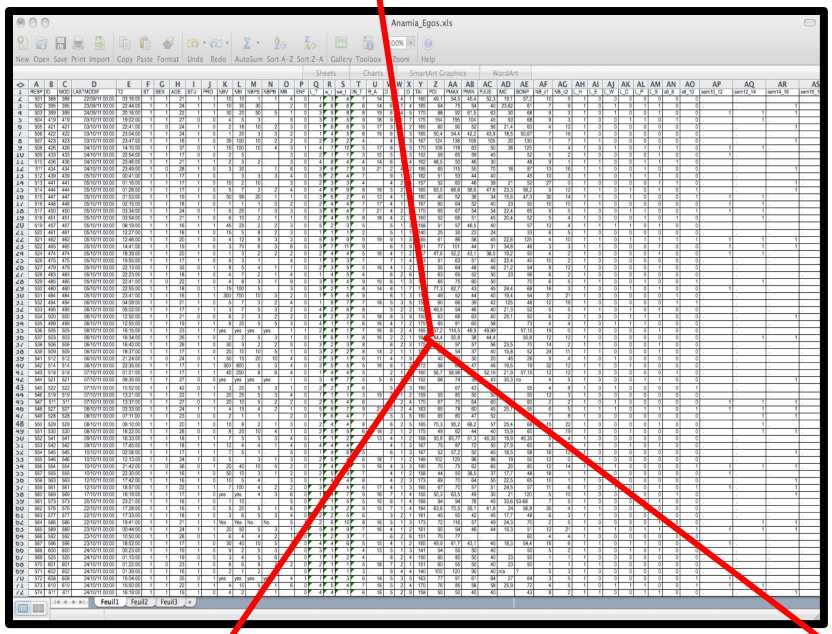
3-4. 結束條件判斷

現在的群集中心與之前的群集中心相同，分群任務全部完成



關鍵就是那個

k=3



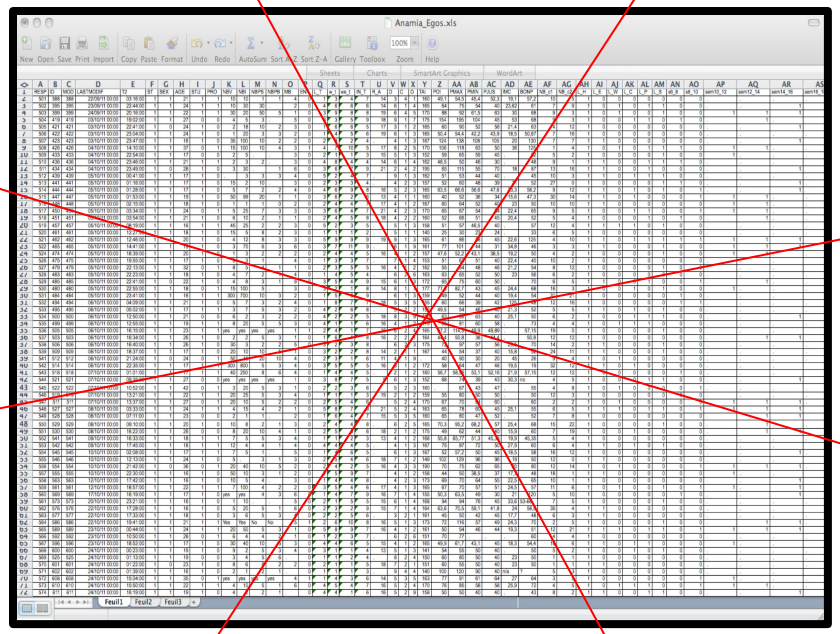
分群演算法

Cascade Simple K Means

層疊式K平均法

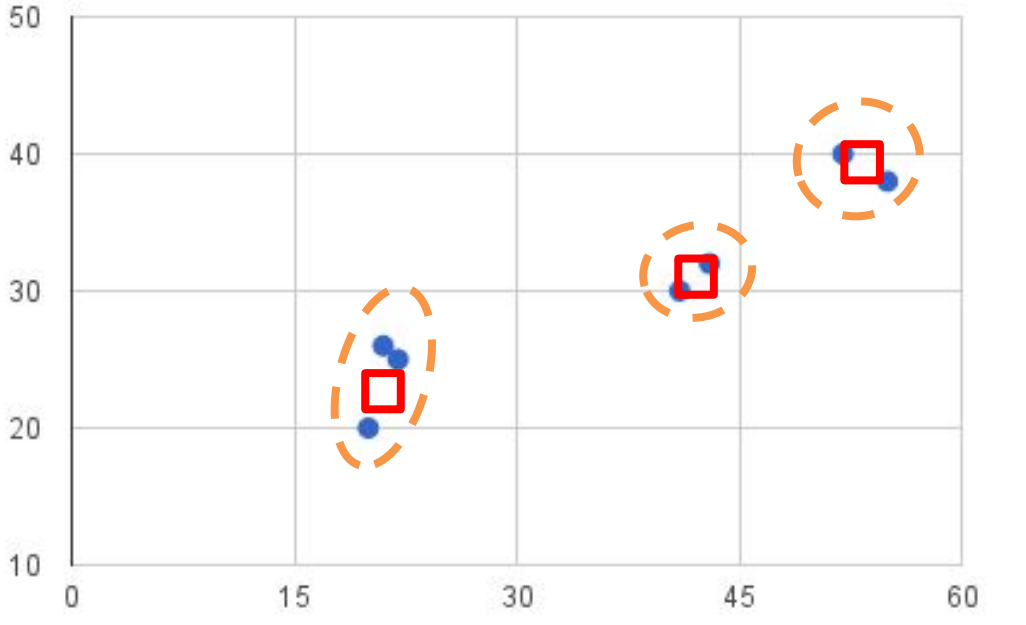
(T. Caliński & J. Harabasz, 1974)

如何選擇分群數量 K？



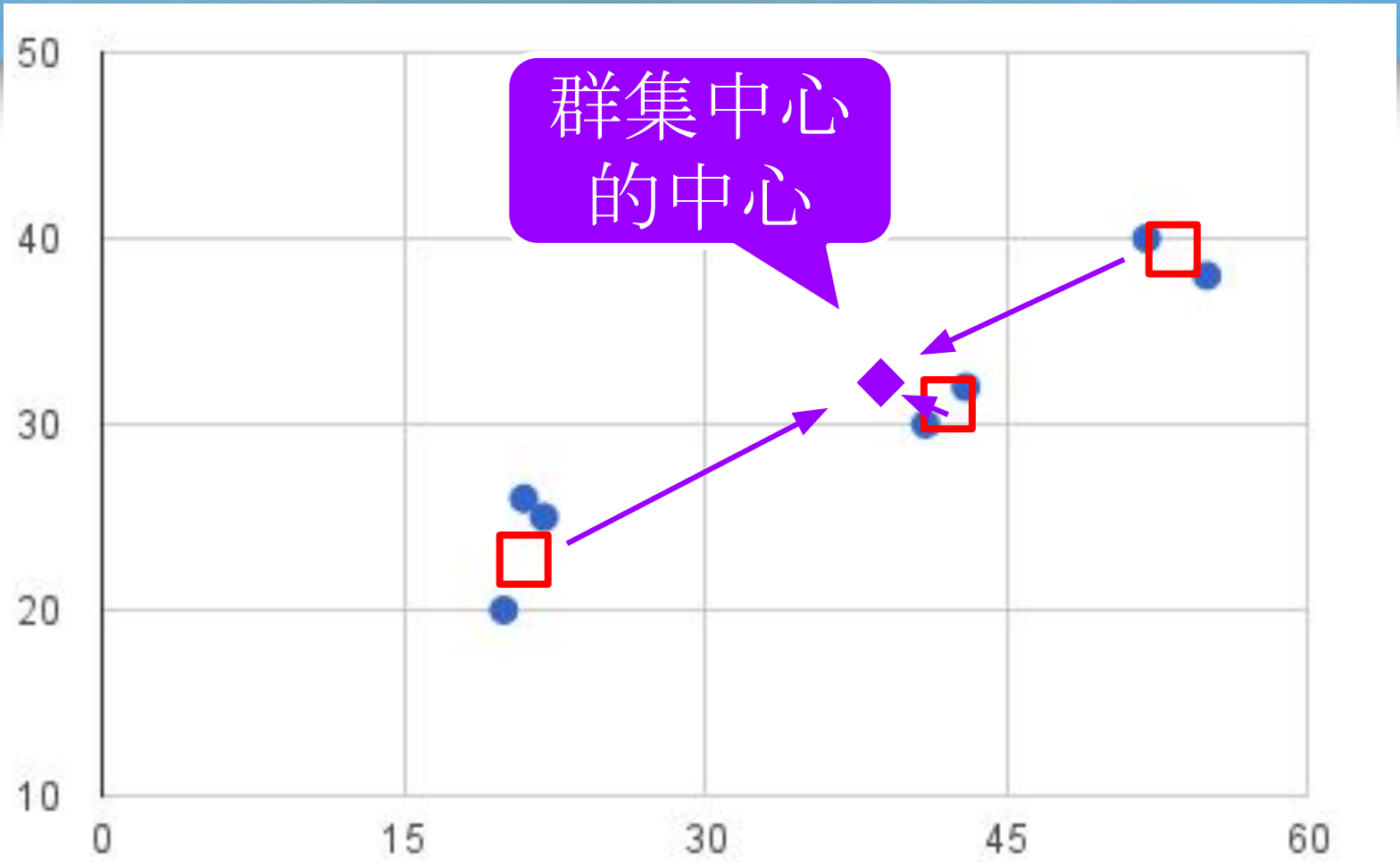
評估分群品質
CH指標 (Calinski-Harabasz)

$$CH(K) = \frac{[\text{trace } \mathbf{B}/K - 1]}{[\text{trace } \mathbf{W}/N - K]} \quad \text{for} \quad K \in \mathbb{N}$$



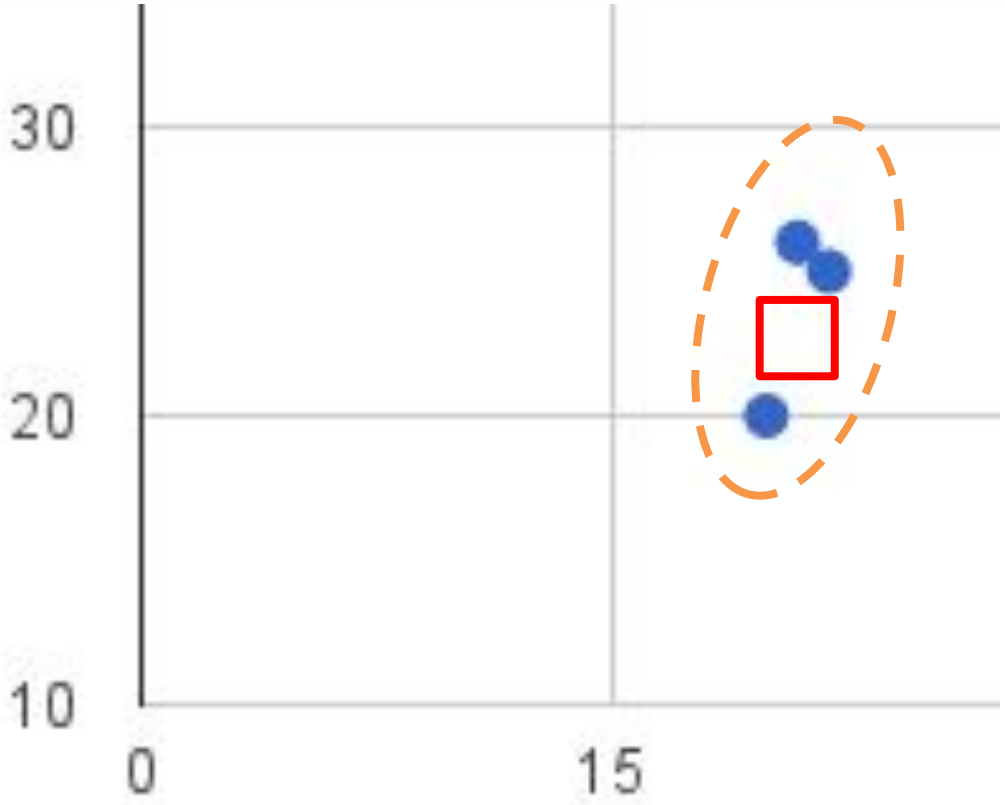
(T. Caliński & J. Harabasz, 1974)

CH指標
trace B:各群之間的距離 (越大越好)

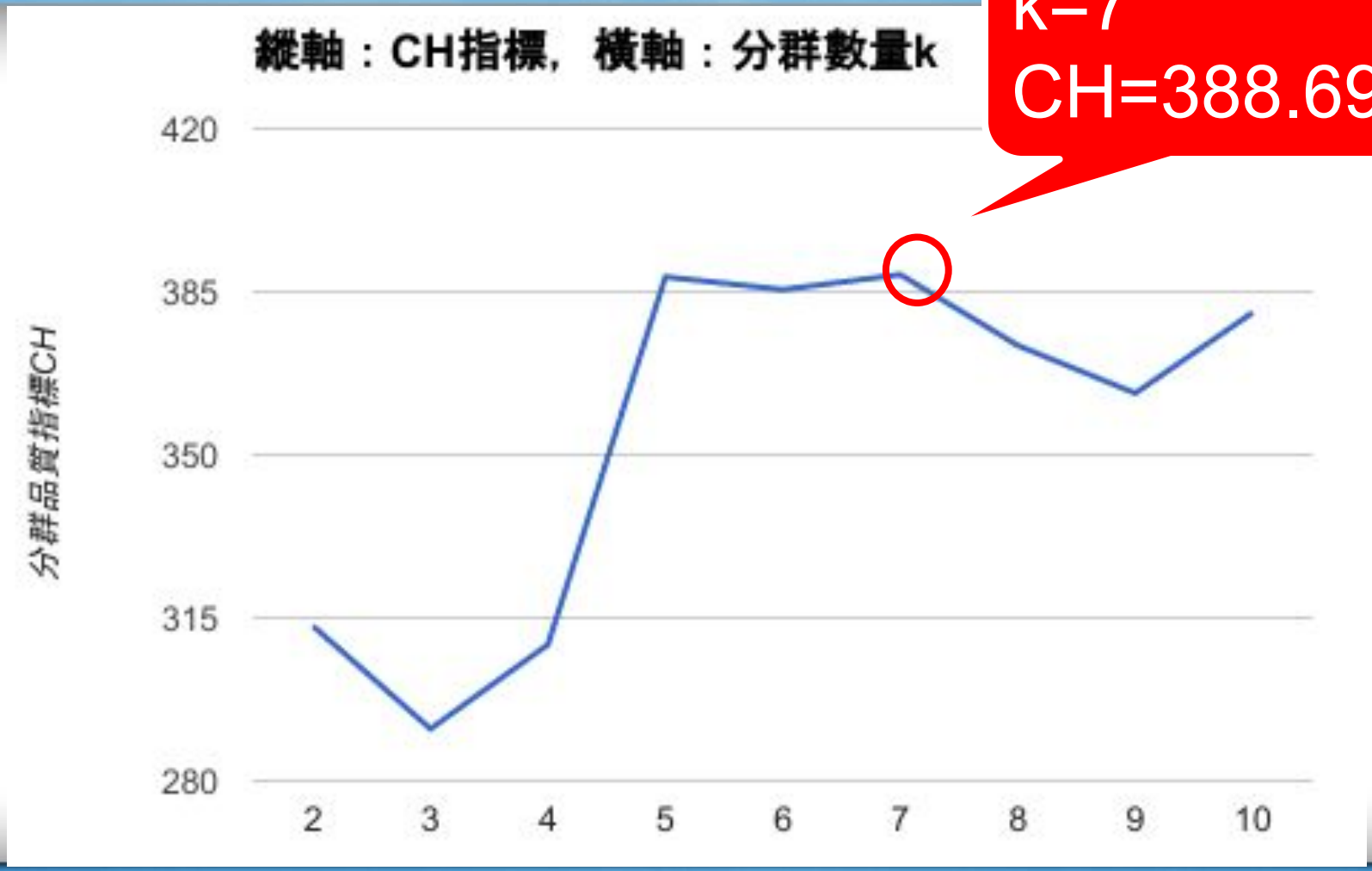


CH指標

trace W:群內各點的距離 (越小越好)



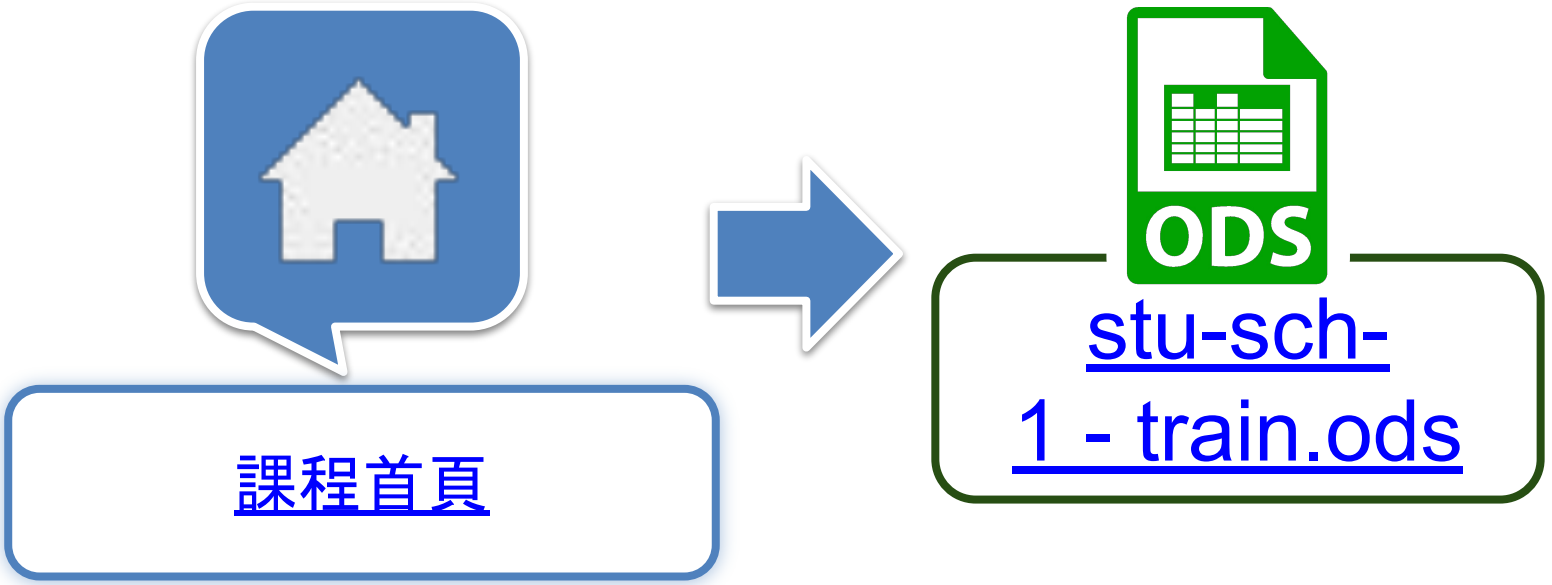
分群數量k與CH指標的變化



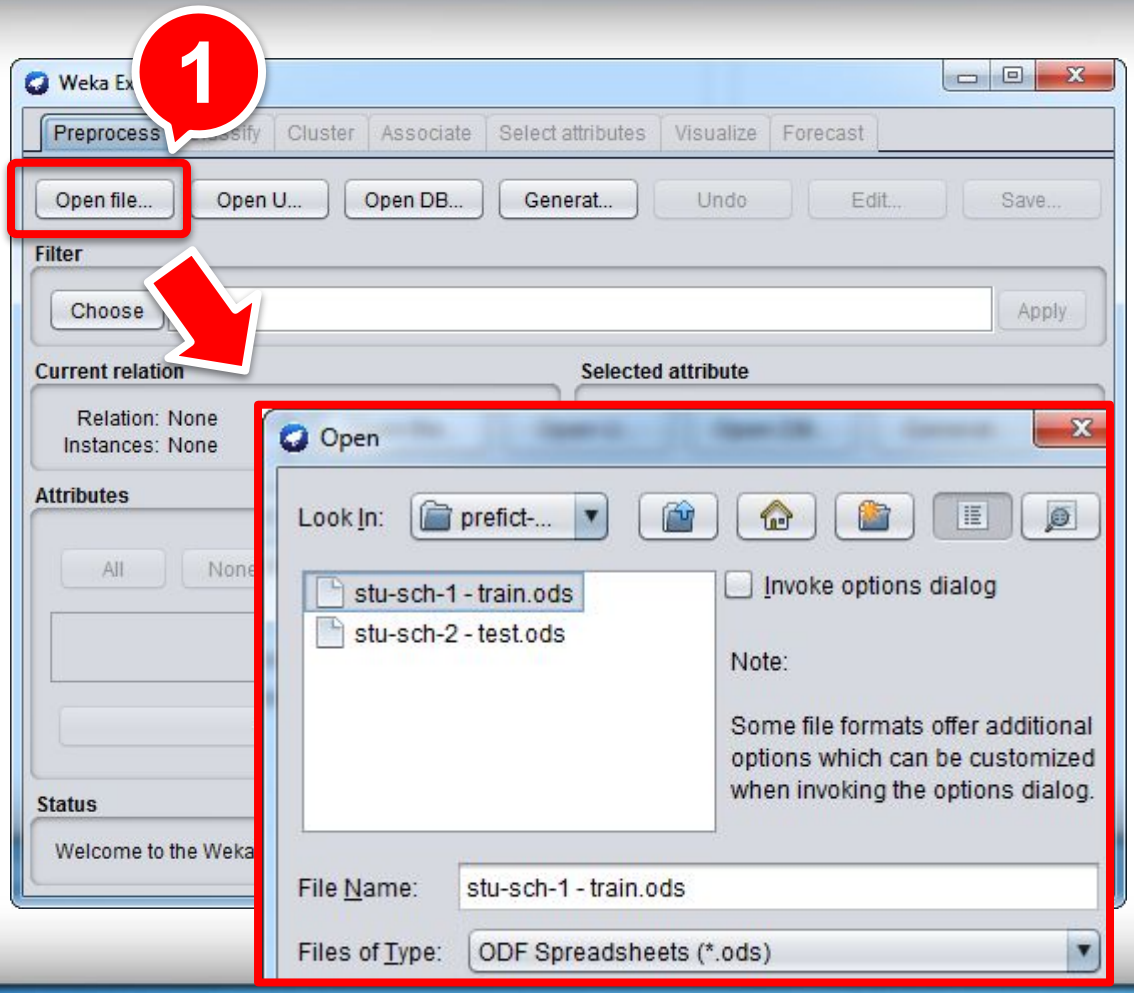
分群:層疊式K平均法
實作步驟

1. 下載與開啟檔案
2. 資料前處理:
   a. 關閉目標屬性
   b. NominalToBoolean
   c. 再度關閉目標屬性
3. 執行分群:AddCluster → CascadeSimpleKMeans
4. 檢視探勘結果:
   Weka分群結果分析器

STEP 1. 下載與開 啟檔案 (1/4)

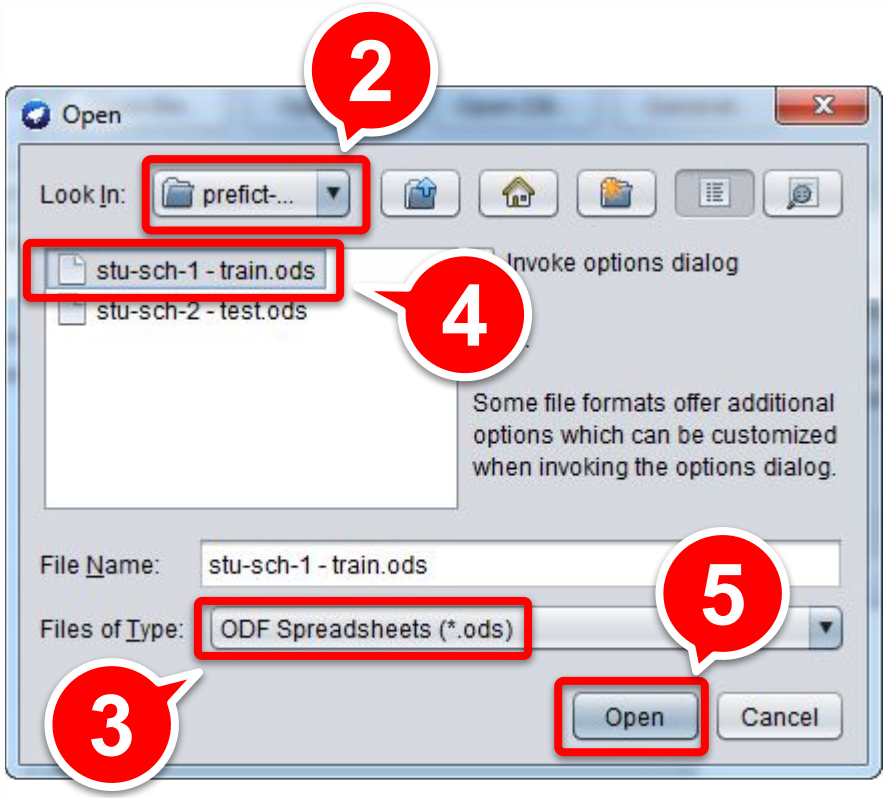


STEP 1. 下載與開啟檔案 (2/4)



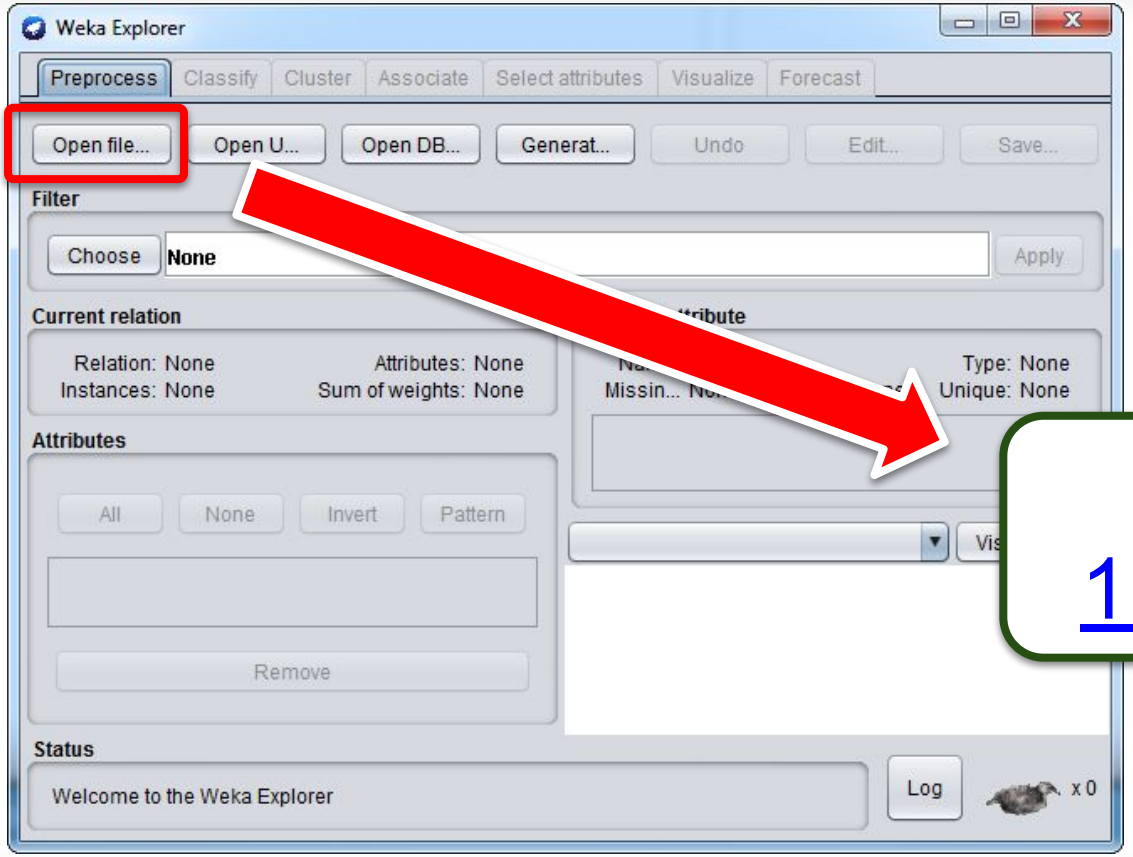
1. Open file…
   開啟檔案

STEP 1. 下載與開啟檔案 (3/4)



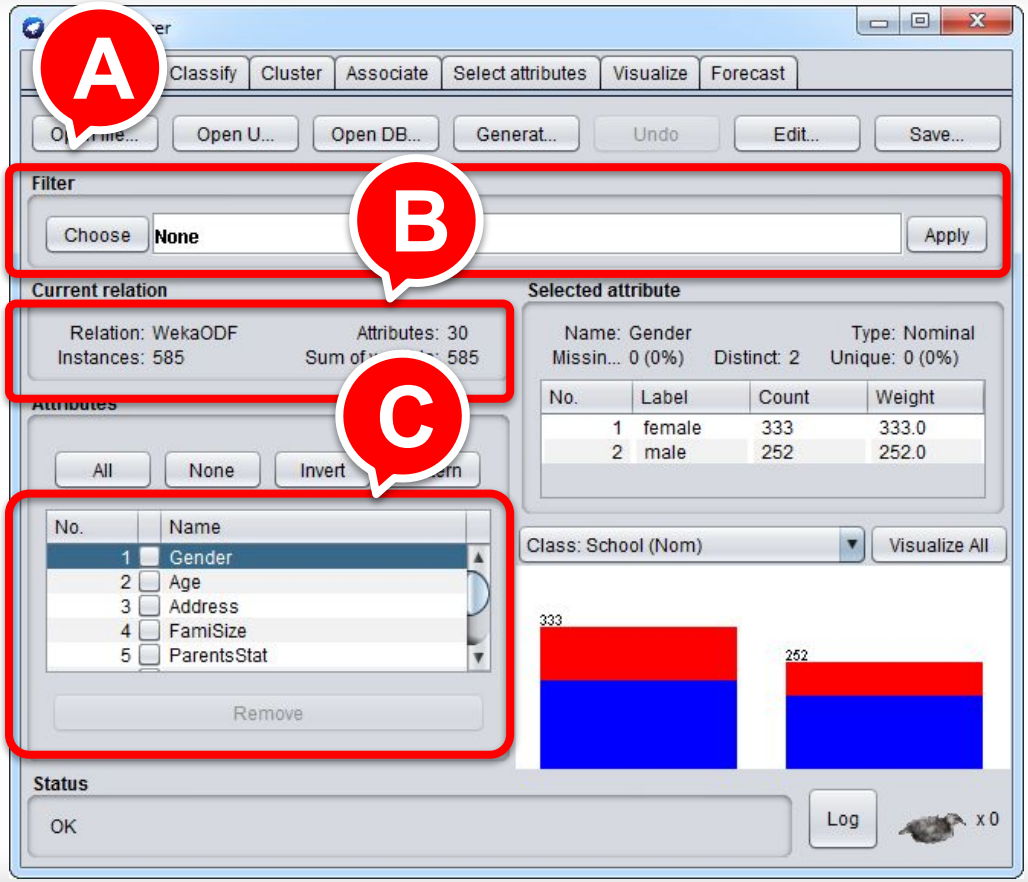
2. Look in:
   移動到下載資料夾
3. Files of Type:
   ODF Spreadsheets (*.ods)
   開啟ODF檔案類型
   ※ 需安裝套件WekaODF
4. 選擇檔案
   stu-sch-1 - train.ods
5. Open 開啟檔案

STEP 1. 下載與開啟檔案 (4/4)



stu-sch-1 - train.ods

探索器介面說明
前處理 (Preprocess) (1/2)

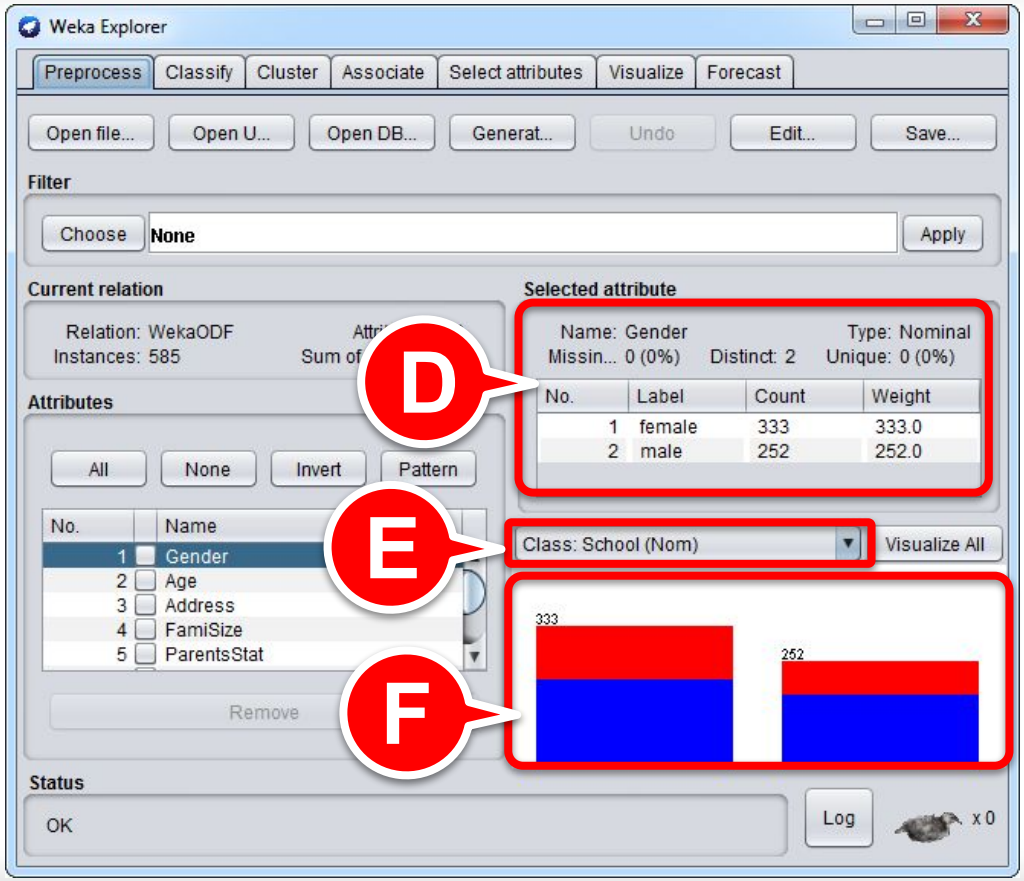


A. Filter 過濾器

B. Current relation
資料整體狀況

C. Attributes
屬性列表

探索器介面說明
前處理 (Preprocess) (2/2)

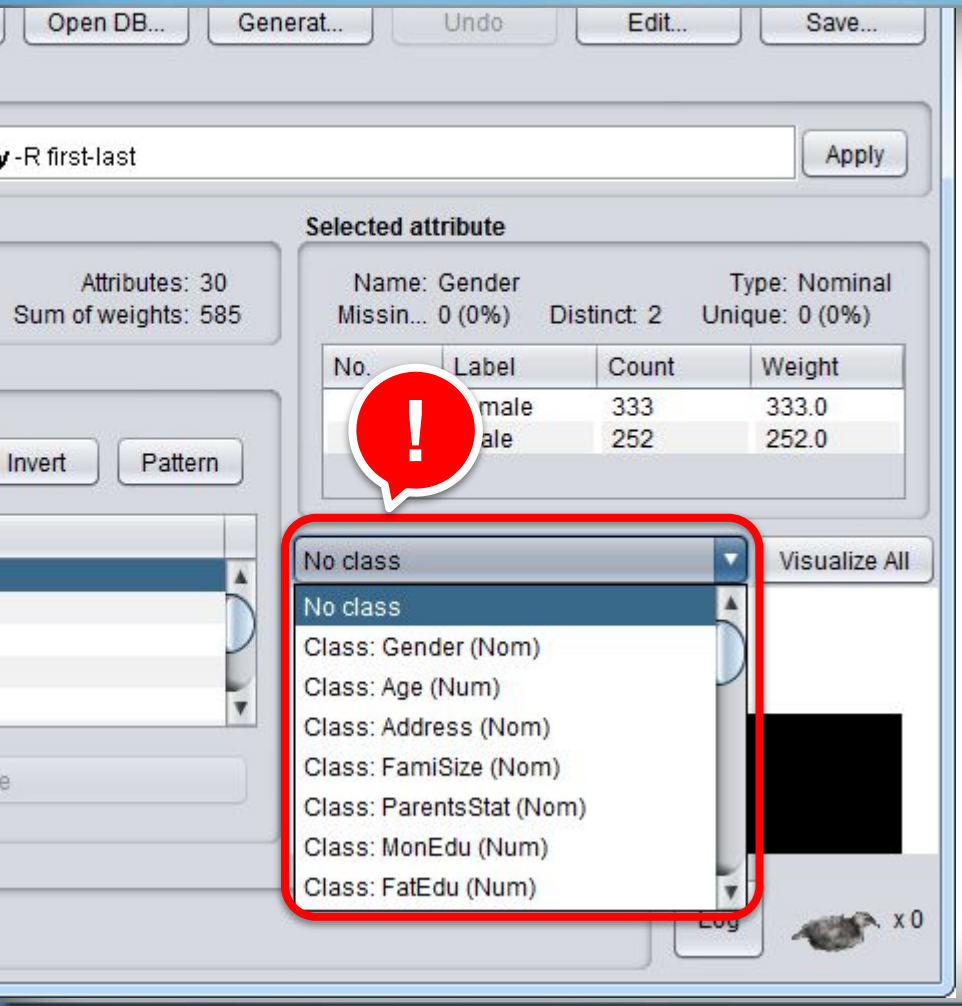


D. Selected atttribute 所選屬性的資料分佈

E. Class 目標屬性

F. 所選屬性的視覺化圖表

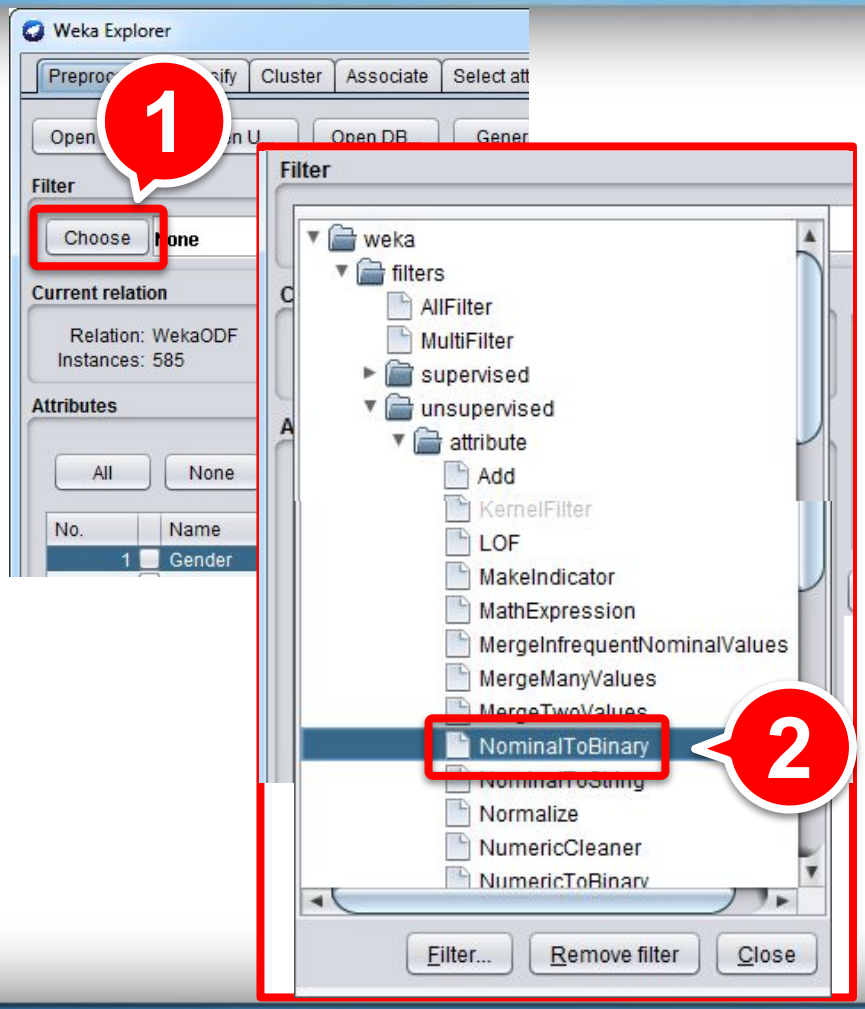
STEP 2a. 資料前處理
關閉目標屬性



- 將目標屬性Class
  改選為No class

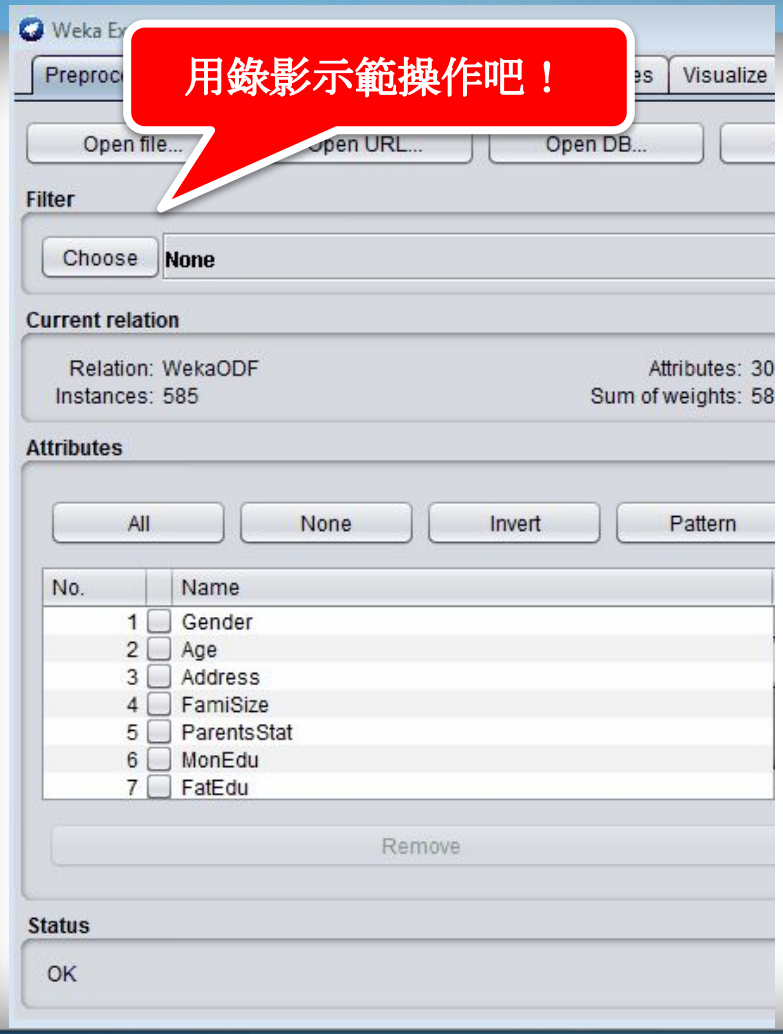
※ 探索性分析不使用目標屬性

STEP 2b. 資料前處理
類別轉虛擬變項 (1/5)



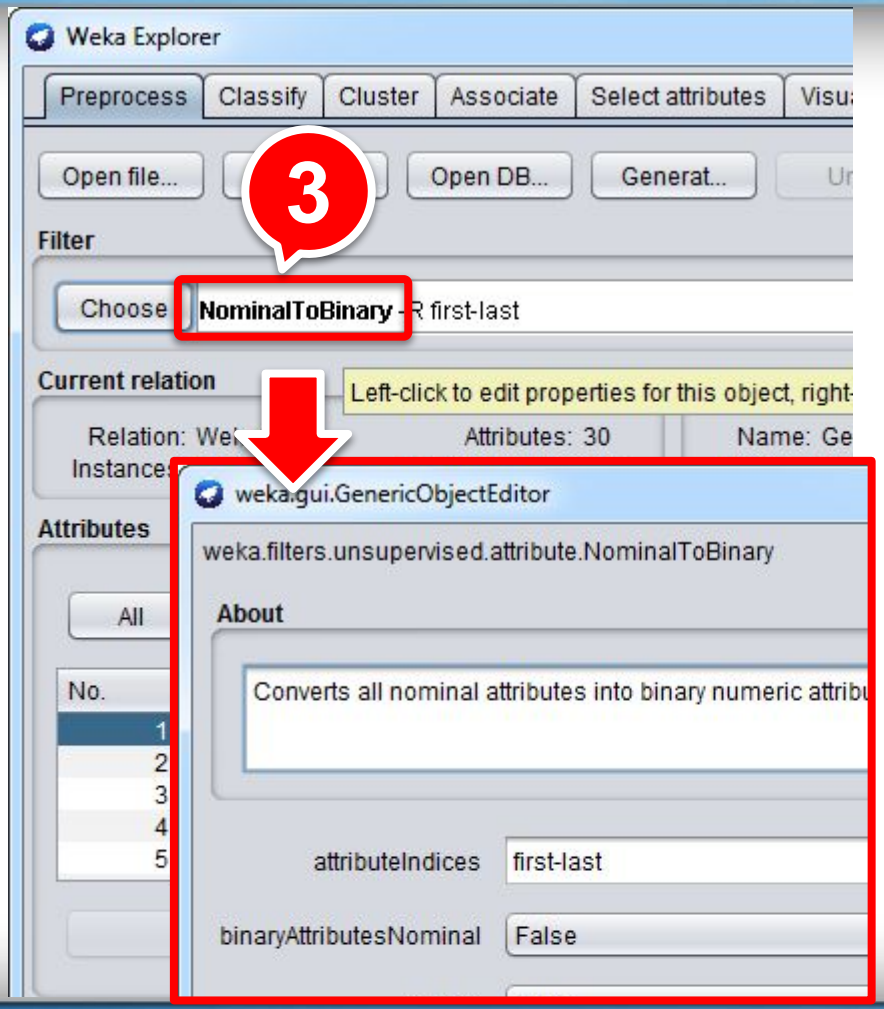
1. 按Filter 底下的 Choose 選擇篩選器
2. 找到篩選器 weka.filters.unsupervised.attribute.**NominalToBinary**

STEP 2b. 資料前處理
類別轉虛擬變項 (1/5)



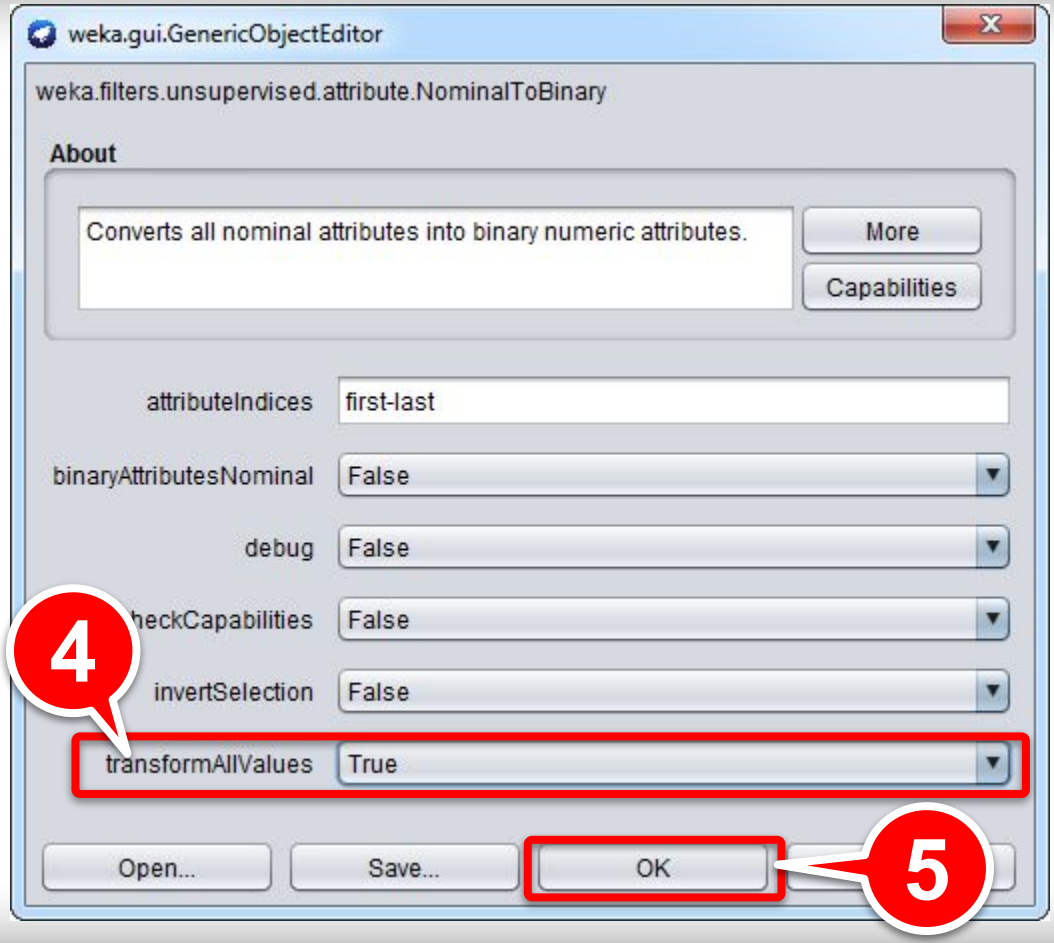
1. 按Filter 底下的 Choose 選擇篩選器
2. 找到篩選器 weka.filters.unsupervised.attribute.**NominalToBinary**

STEP 2b. 資料前處理
類別轉虛擬變項 (2/5)



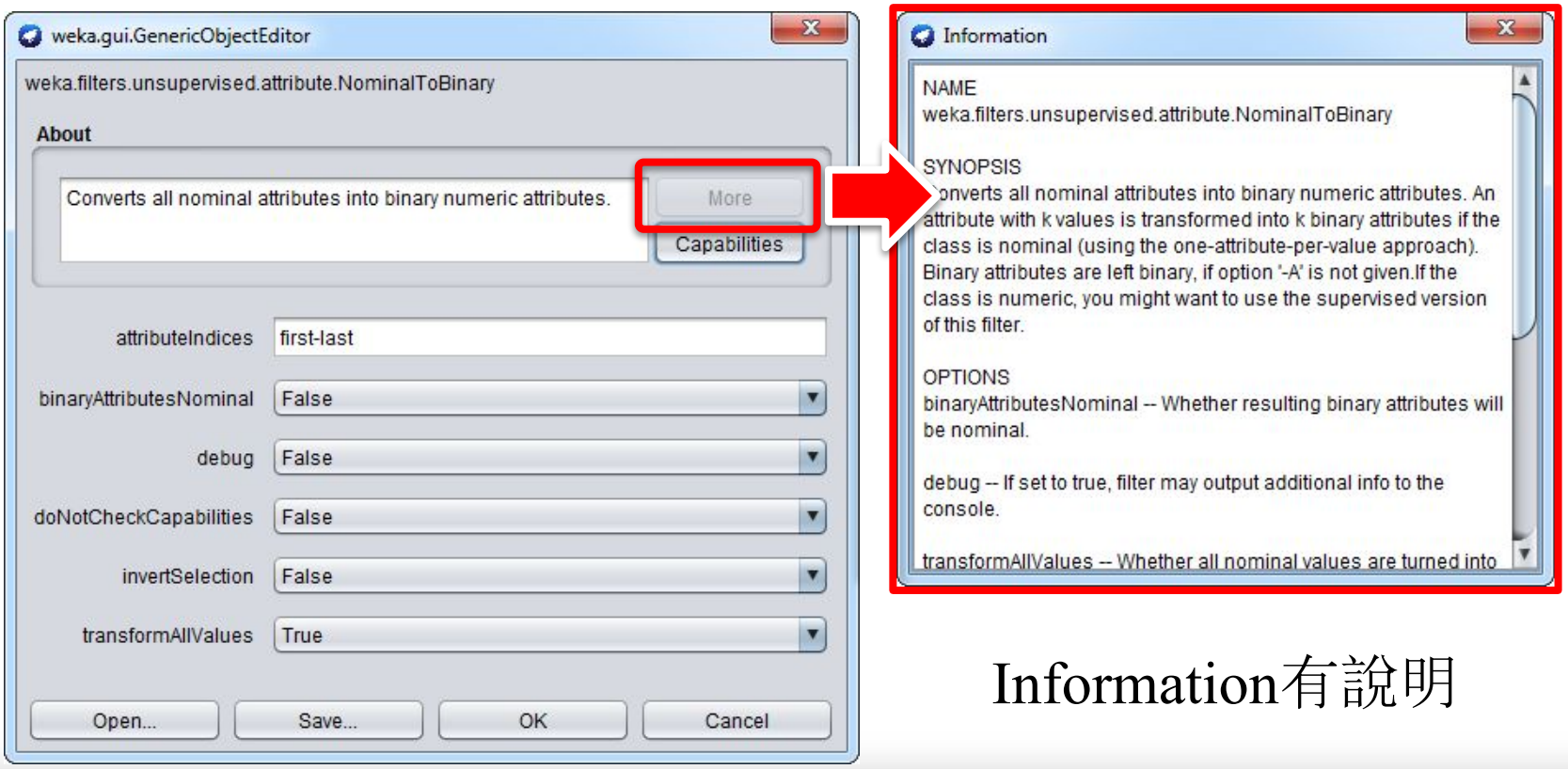
3. 按下粗體字的篩選器名稱 **NominalToBinary** 開啟進階設定

STEP 2b. 資料前處理
類別轉虛擬變項 (3/5)



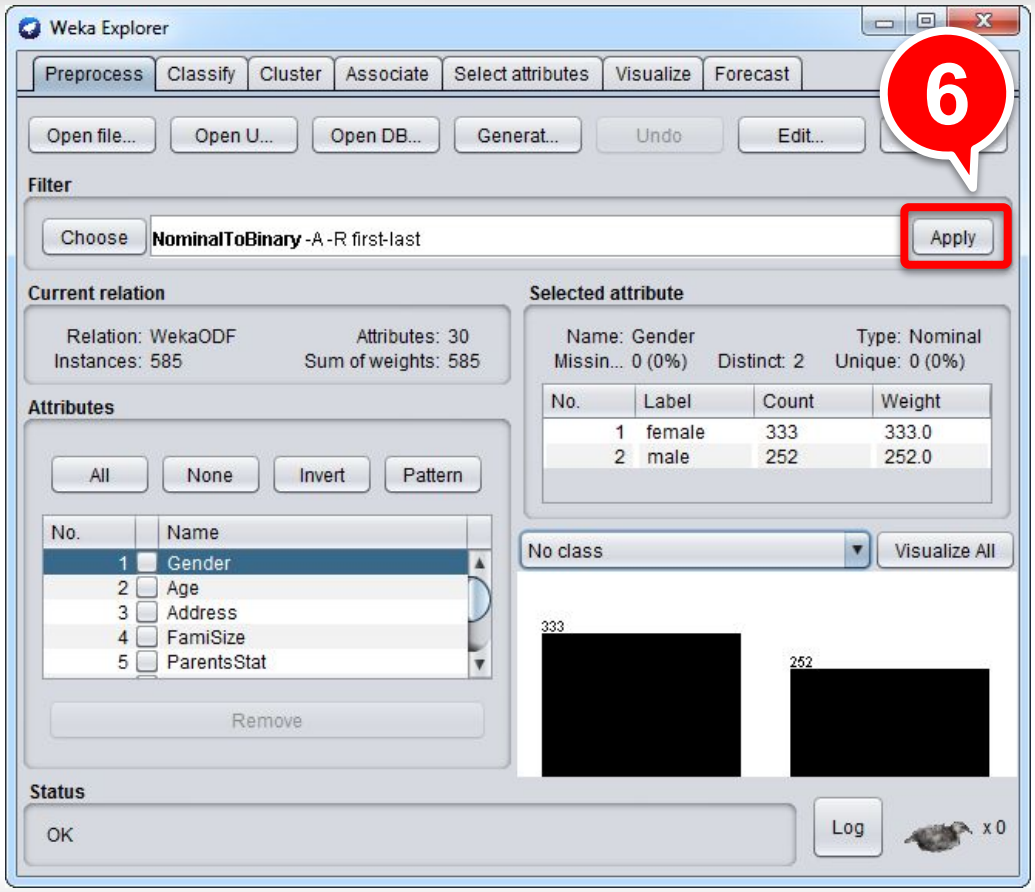
4. 將transformAllValues 設為True 執行虛擬變項轉換
5. OK 離開進階設定

想知道進階設定每個欄位的意思？



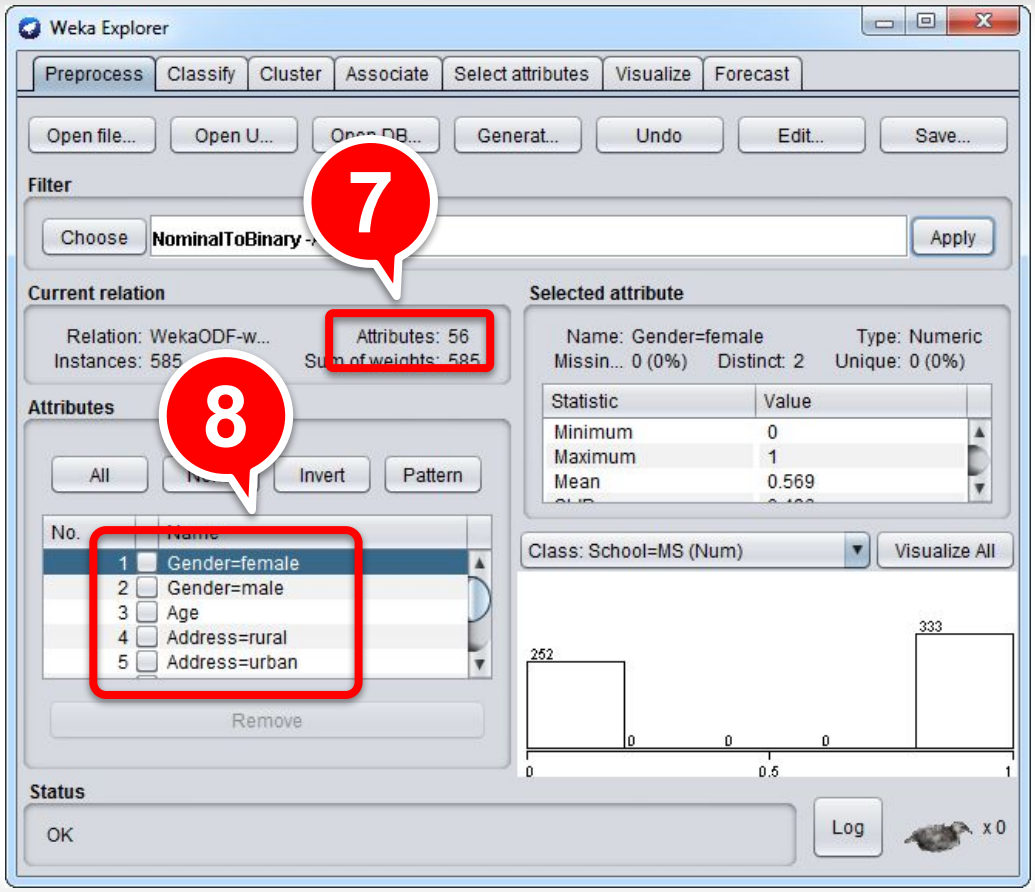
Information有說明

STEP 2b. 資料前處理
類別轉虛擬變項 (4/5)



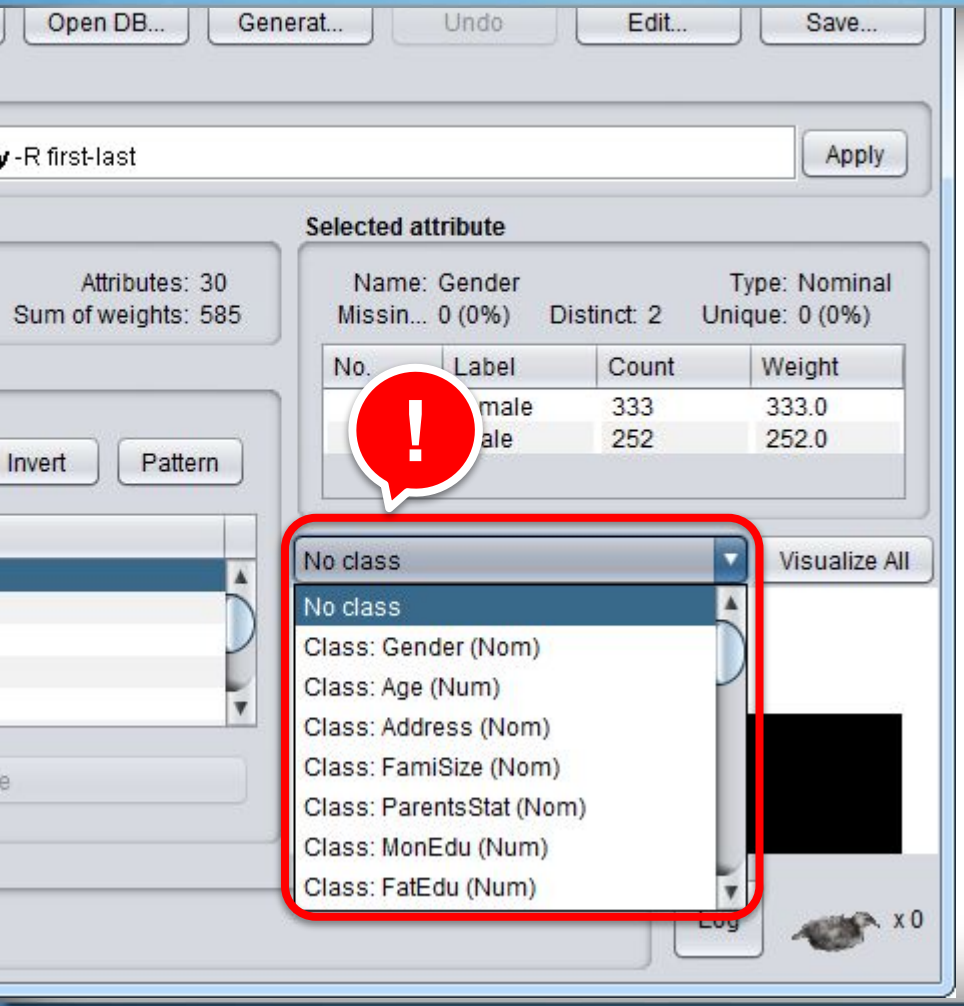
6. 按下Apply
套用篩選器

STEP 2b. 資料前處理
類別轉虛擬變項 (5/5)



7. Attributes: 56
   屬性數量增加
   從30變成56個
8. 類別型屬性Gender
   被轉換成兩個數值
   型屬性
   a. Gender=female
   b. Gender=male

STEP 2c. 資料前處理
再次關閉目標屬性



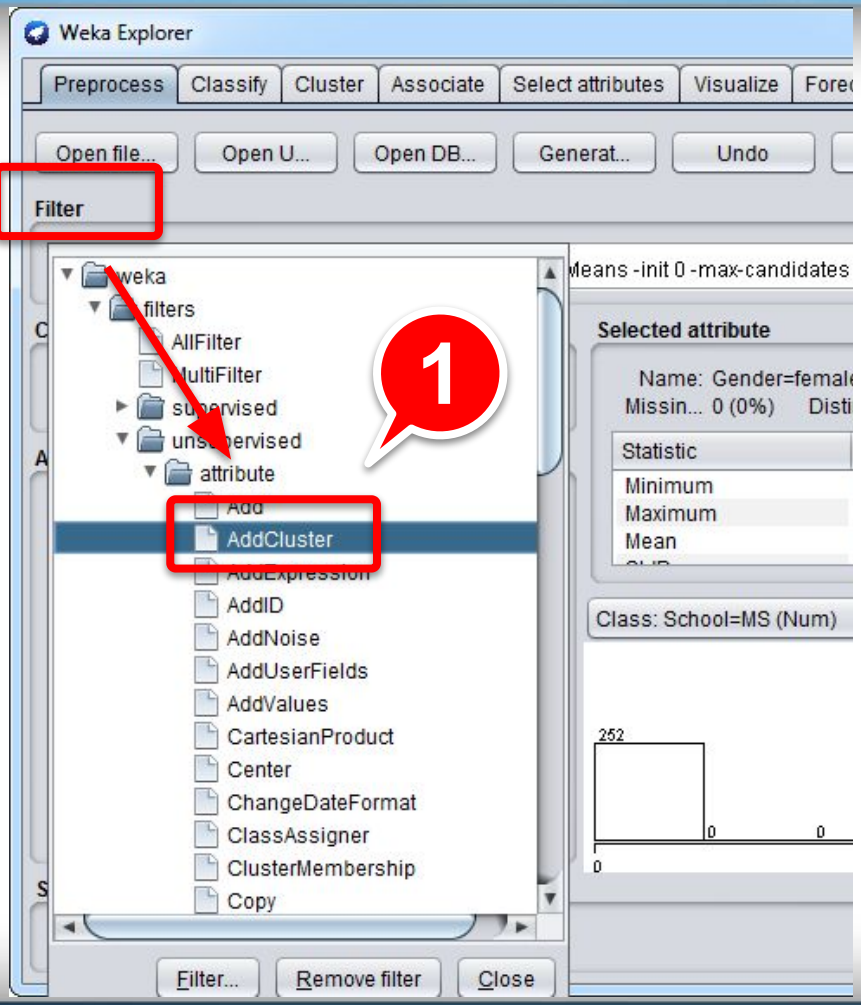
因為屬性修改了，Weka會自動幫你選擇目標屬性

這時候要記得關掉它

- 將目標屬性Class改選為No class

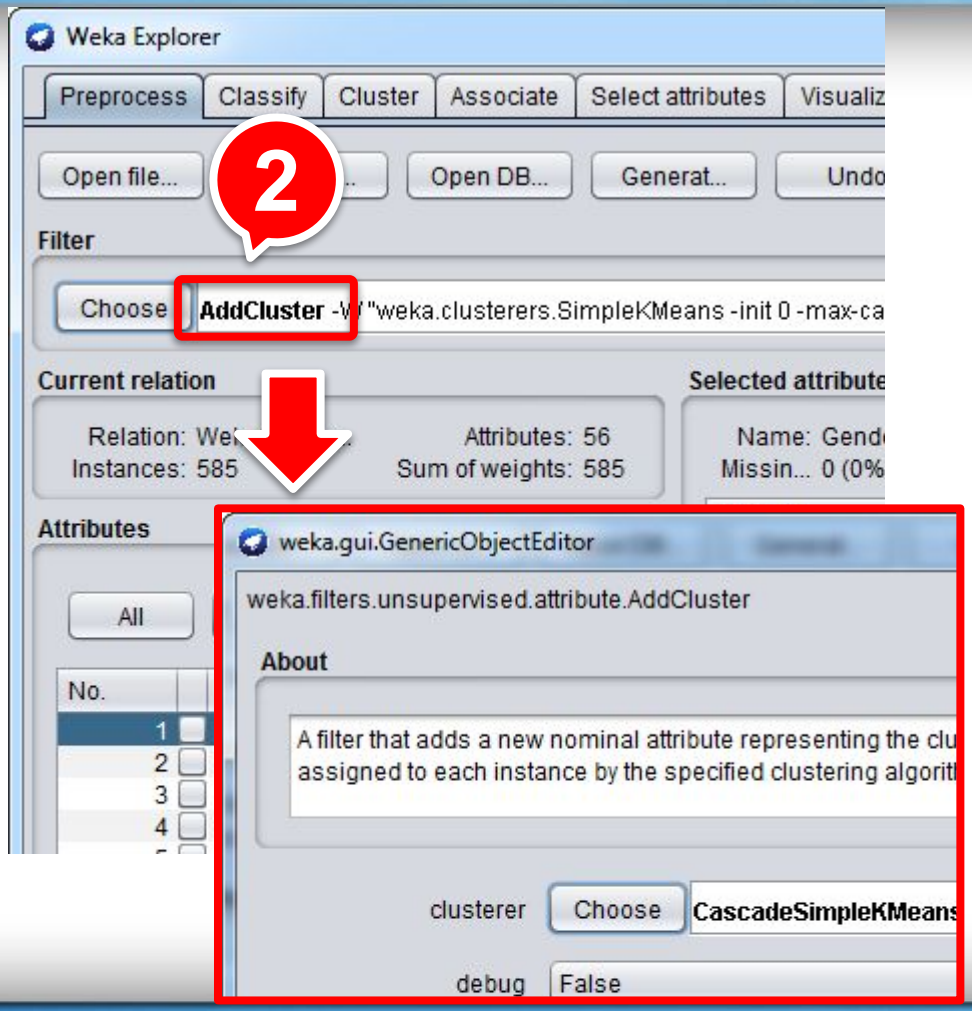
※ 探索性分析不使用目標屬性

STEP 3. 執行分群 (1/7)



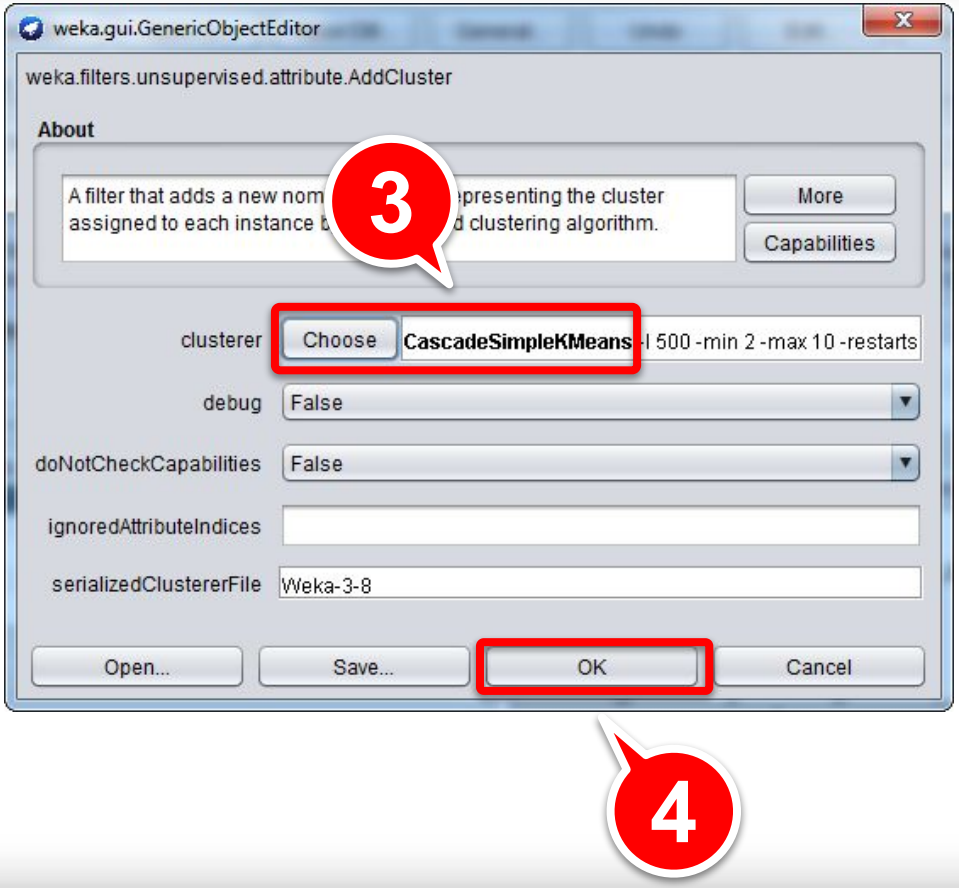
1. Filter ⇨ Choose
   選擇篩選器
   weka.filters.unsupervised
   .attribute.**AddCluster**

STEP 3. 執行分群 (2/7)



2. 按下粗體字的篩選器名稱
   AddCluster
   開啟進階設定

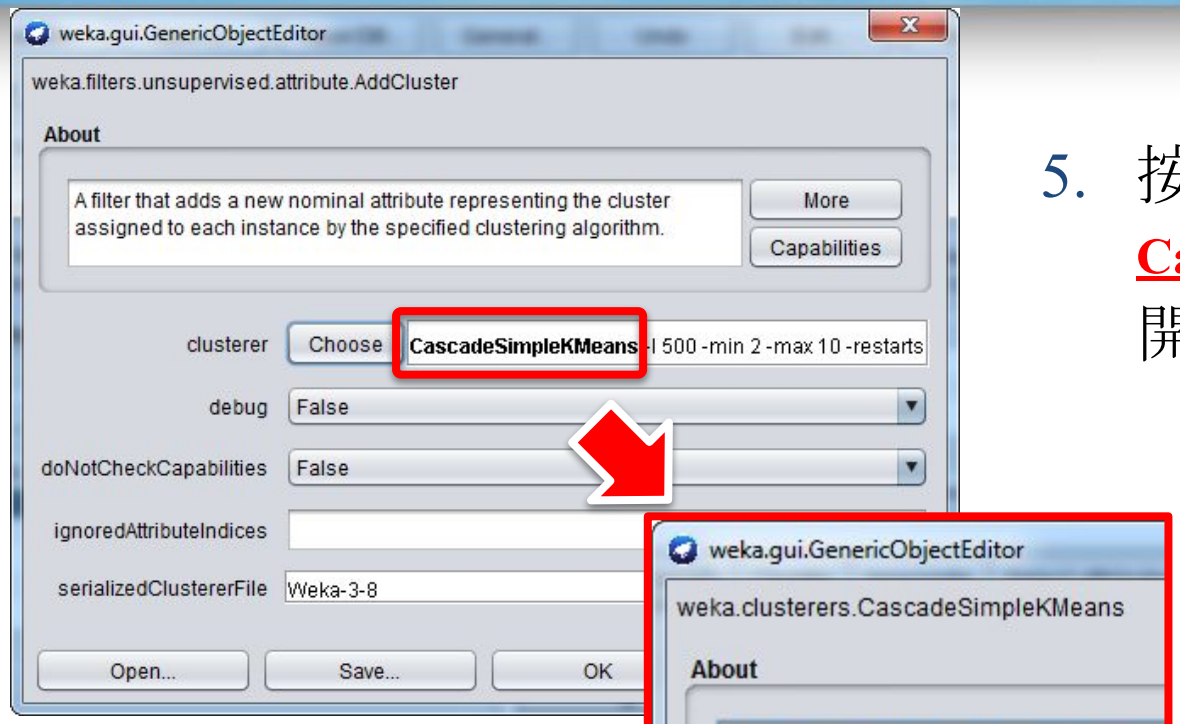
STEP 3. 執行分群 (3/7)

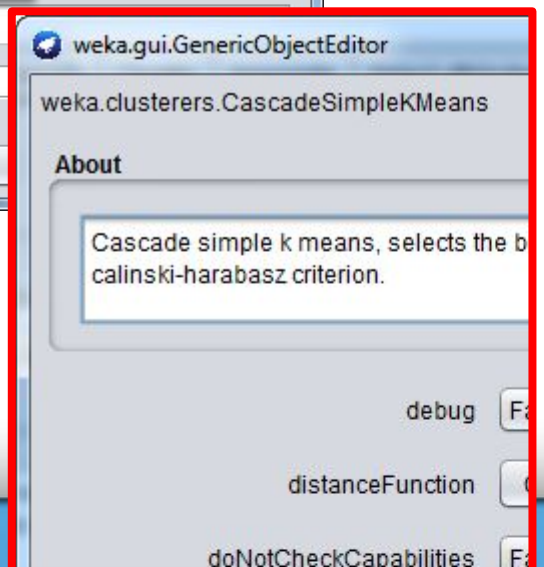


3. 將 clusterer 分群演算法選擇 weka.clusterers.**CascadeSimpleKMeans**
4. OK 離開進階設定

※ 需安裝套件cascadeKMeans

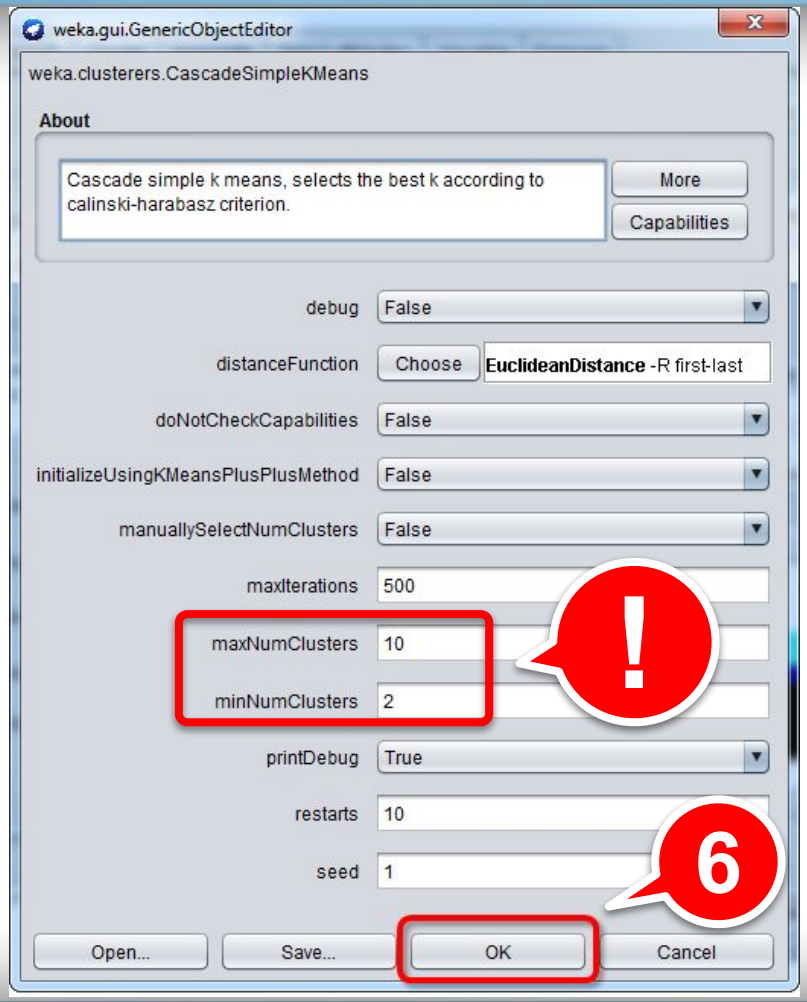
STEP 3. 執行分群 (4/7)





5. 按粗體字 **CascadeSimpleKMeans** 開啟進階設定

STEP 3. 執行分群 (5/7)

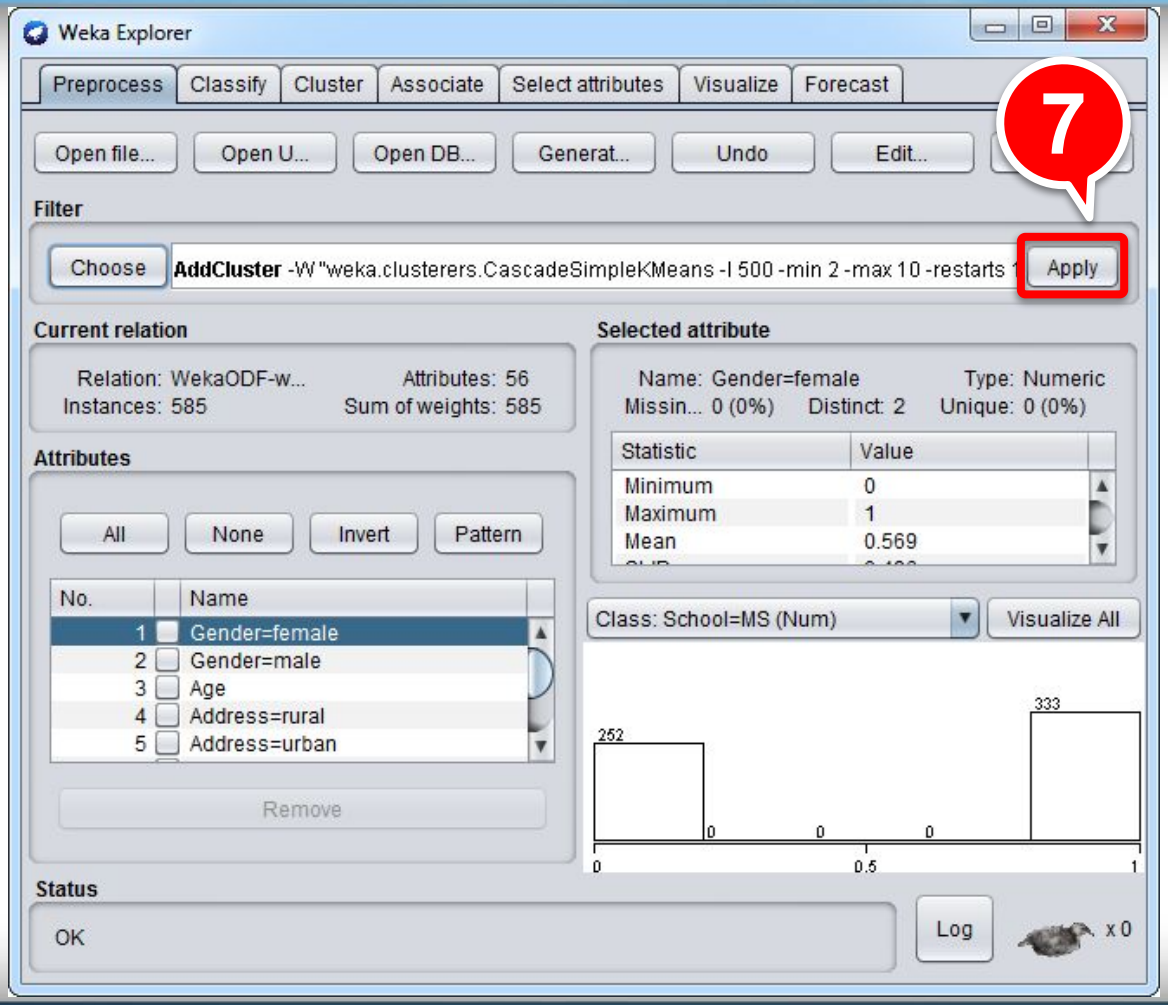


在maxNumClusters跟minNumClusters裡面可以設定最多和最少的分群數量。預設值會讓分群數量介於2至10之間。

如果沒有特別要修改的話，

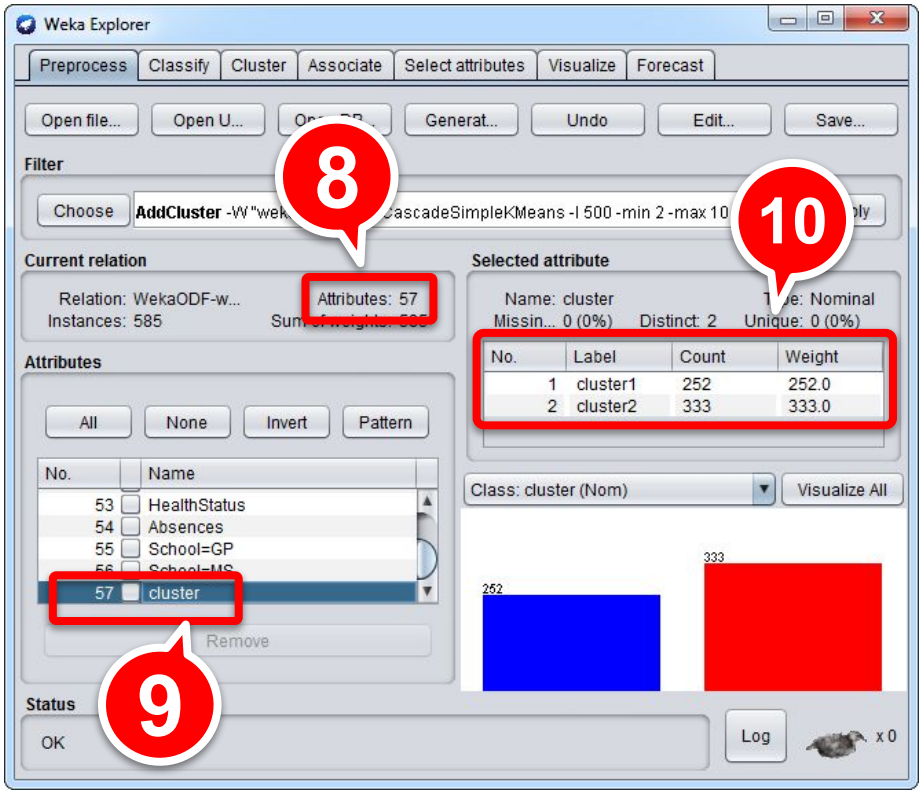
6. OK 離開進階設定

STEP 3. 執行分群 (6/7)



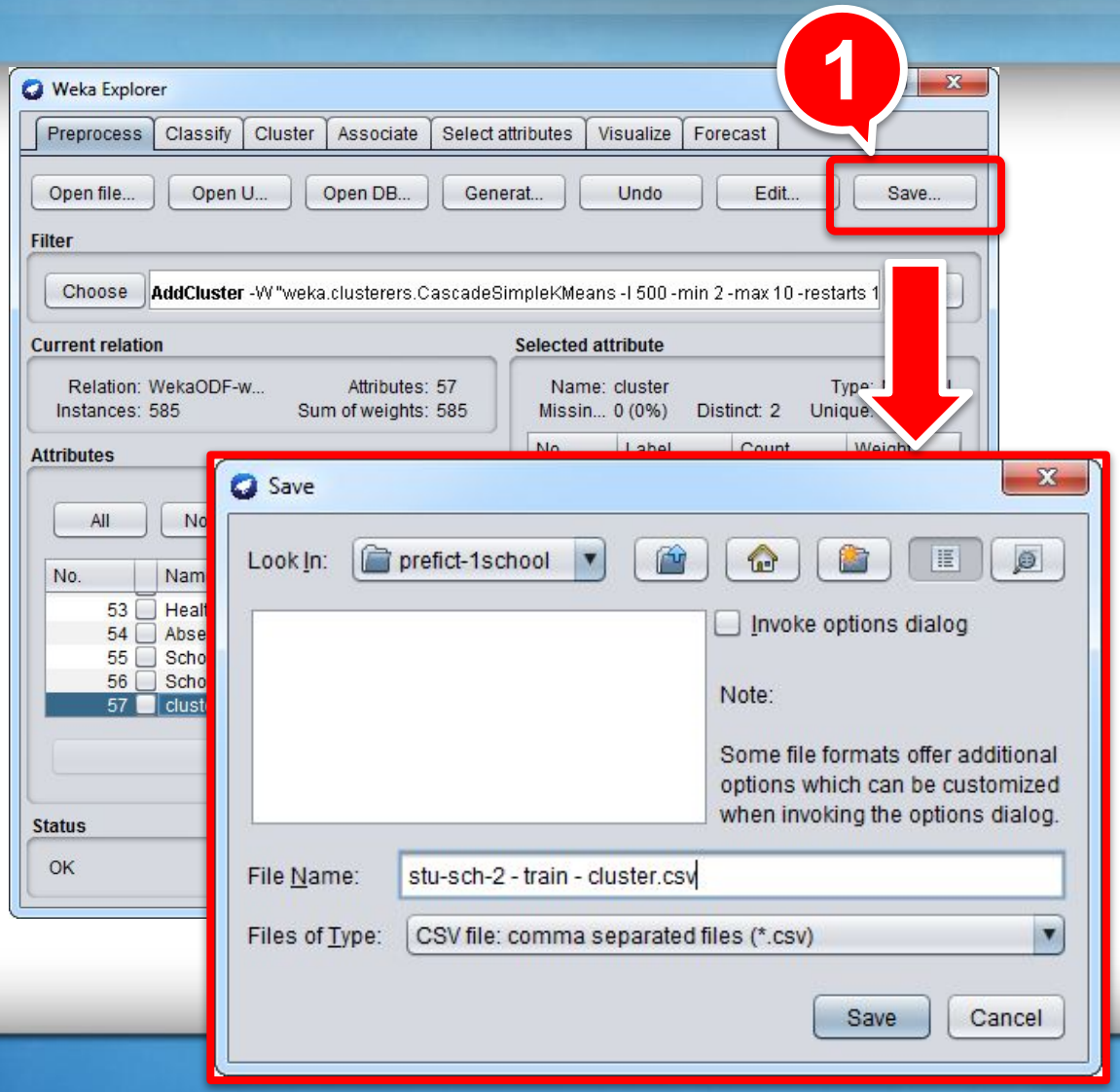
7. 按下Apply
套用篩選器

STEP 3. 執行分群 (7/7)



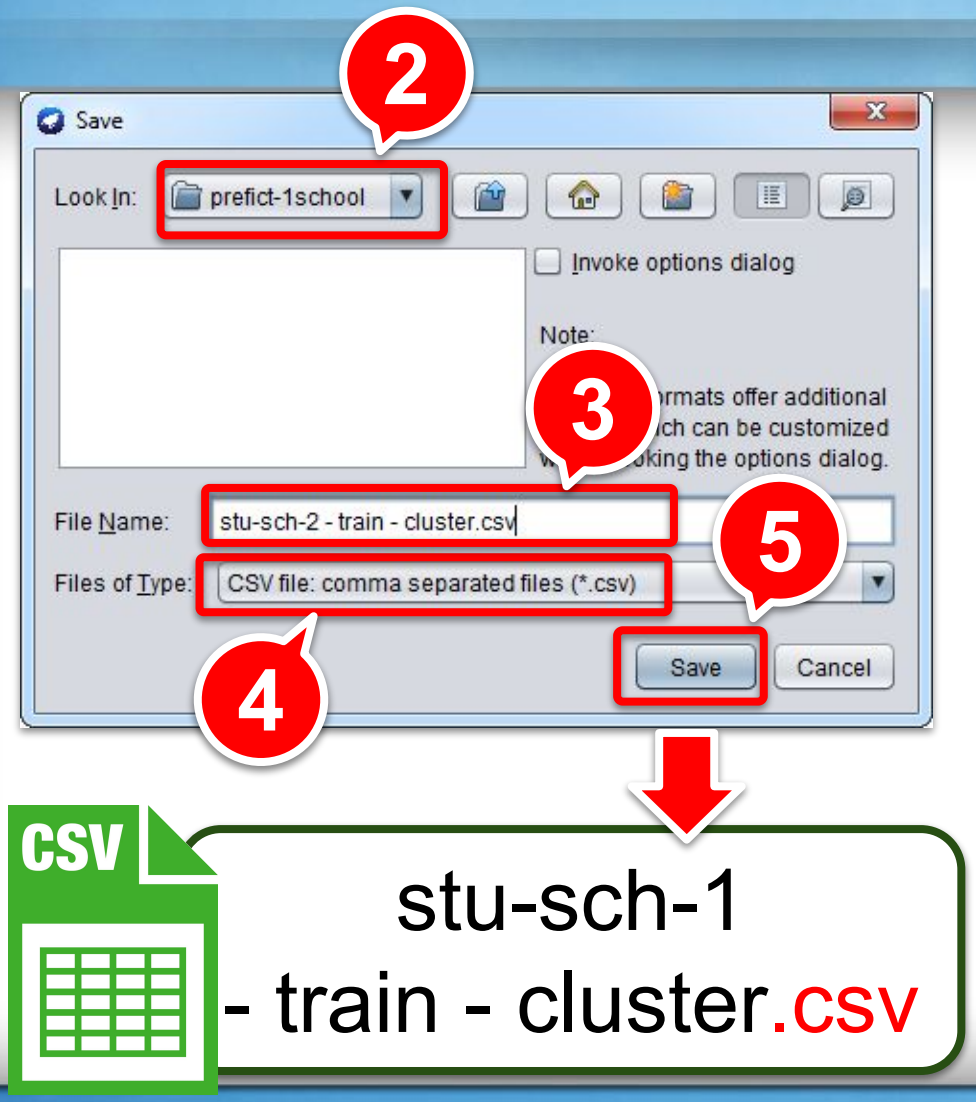
8. Attributes: 57
   屬性數量增加
   從56變成57個
9. 新增了cluster類別型屬性
10. 所有資料被分成兩群
    cluster1: 共252筆
    cluster2: 共333筆

STEP 4. 檢視探勘結果 (1/6)



1. Save 儲存檔案

STEP 4. 檢視探勘結果 (2/6)



2. Look in:
   移動到下載資料夾
3. File Name 檔案命名

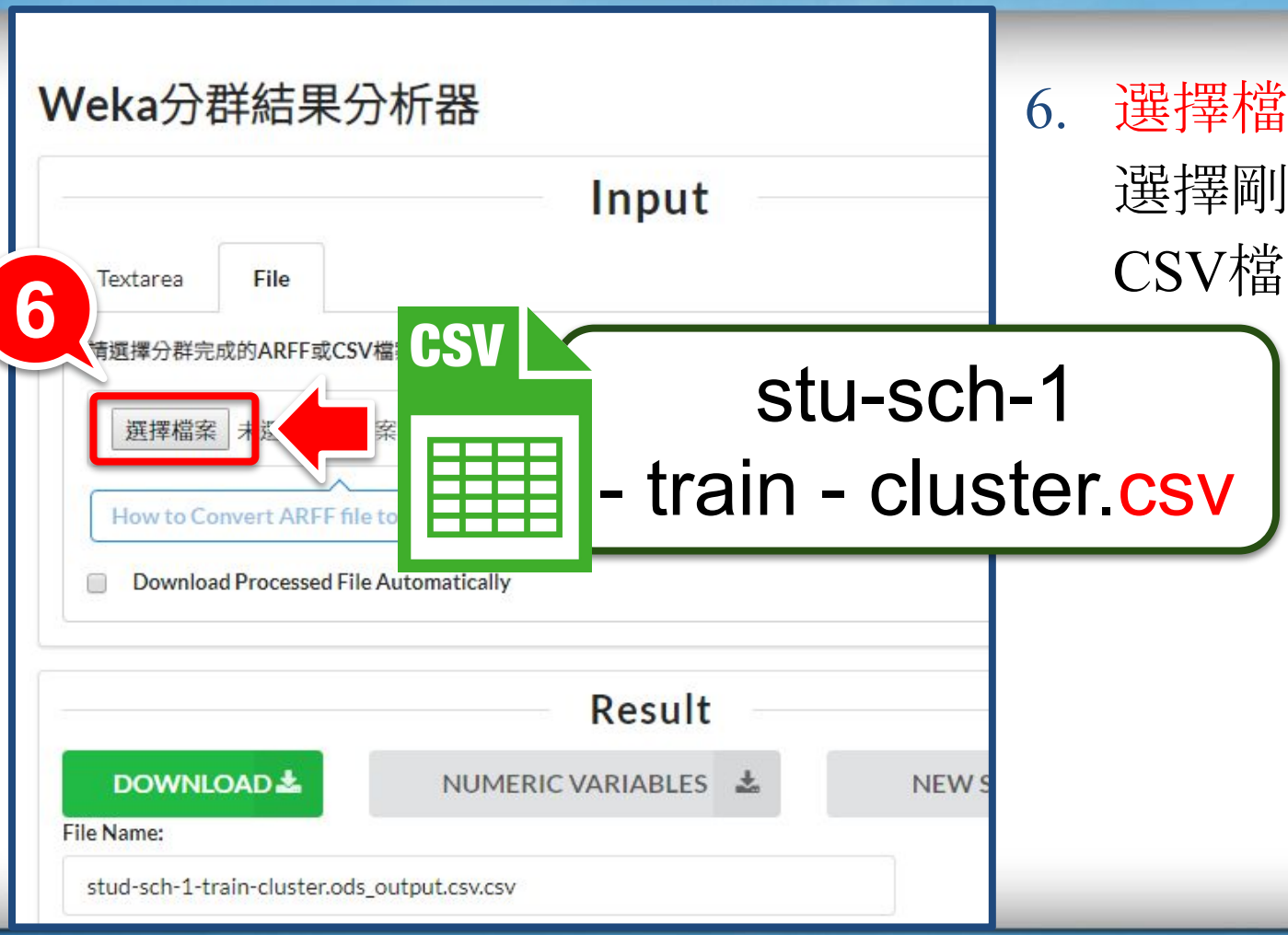
stu-sch-1 - train -cluster.csv

4. Files of Type:
   CSV file: comma separated files (*.csv)
   以CSV檔案類型儲存
5. Save 儲存檔案
   此資料夾就會產生CSV檔案

STEP 4. 檢視探勘結果 (3/6)

5. 開啟Weka分群結果分析器

STEP 4. 檢視探勘結果 (4/6)



6. 選擇檔案

選擇剛剛儲存的CSV檔案

STEP 4. 檢視探勘結果 (6/6)

7. 分群比較表
   查看各分群大於和小於全部資料均值的屬性
8. 舉例：
   - 第2群的Age大於平均值
   - 表示第2群的年齡較大

STEP 4. 檢視探勘結果 (6/6)

分群結果 / Statistic Result:

變項	全部資料	第1群(252) 下載	第2群(333) 下載
筆數	585 (平均: 292.5)	252	333
var0: Gender=female (Avg.)	0.5692	0	1
var0: Gender=female (Std.)	0.4952	0	
var1: Gender=male (Avg.)	0.4308	1	
var1: Gender=male (Std.)	0.4952	0	0
var2: Age (Avg.)	16.7521	16.6786	16.8078
var2: Age (Std.)	1.2411	1.277	1.2102
var3: Address=rural (Avg.)	0.3111	0.3016	0.3183
var3: Address=rural (Std.)	0.4629	0.4589	0.4658
var4: Address=urban (Avg.)	0.6889	0.6984	0.6817
var4: Address=urban (Std.)	0.4629	0.4589	0.4658
var5: FamiSize=<=3 (Avg.)	0.3179	0.3571	0.2883
var5: FamiSize=<=3 (Std.)	0.4657	0.4792	0.453
var6: FamiSize=>3 (Avg.)	0.6821	0.6429	0.7117
var6: FamiSize=>3 (Std.)	0.4657	0.4792	0.453
var7: ParentsStat=apart (Avg.)	0.1368	0.1032	0.1622

9. 分群結果
   查看各屬性在各群中的平均值和標準差
10. 舉例:
    - 第2群的Age大於平均值為16.8
    - 比第1群的16.6還要大一點

我們可以把相似的學生分成兩群

第1群

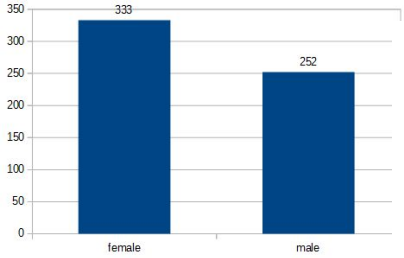
- 大多為男性
- 更多是住市區
- 家庭成員偏3人以下
- 雙親大多同住
- 母親教育程度相對較高

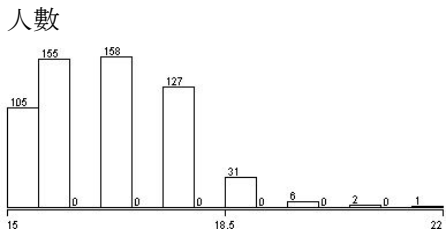
……

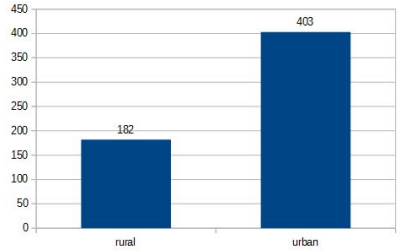
第2群

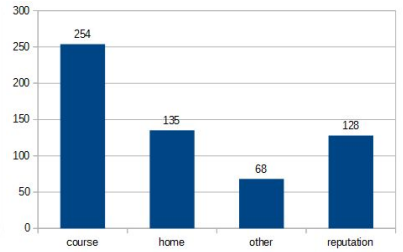
- 大多為女性
- 年齡較大
- 更多是住鄉村
- 家庭成員偏大於3人
- 雙親大多分居
- 較多母親在家工作

……









第1群(252) 下載

var1: Gender=male*
var4: Address=urban*
var5: FamiSize=<=3*
var8: ParentsStat=together*
var9: MonEdu*
var20: FatJob=teacher*
var22: ChoSchReason=home*
var23: ChoSchReason=other*

第1群

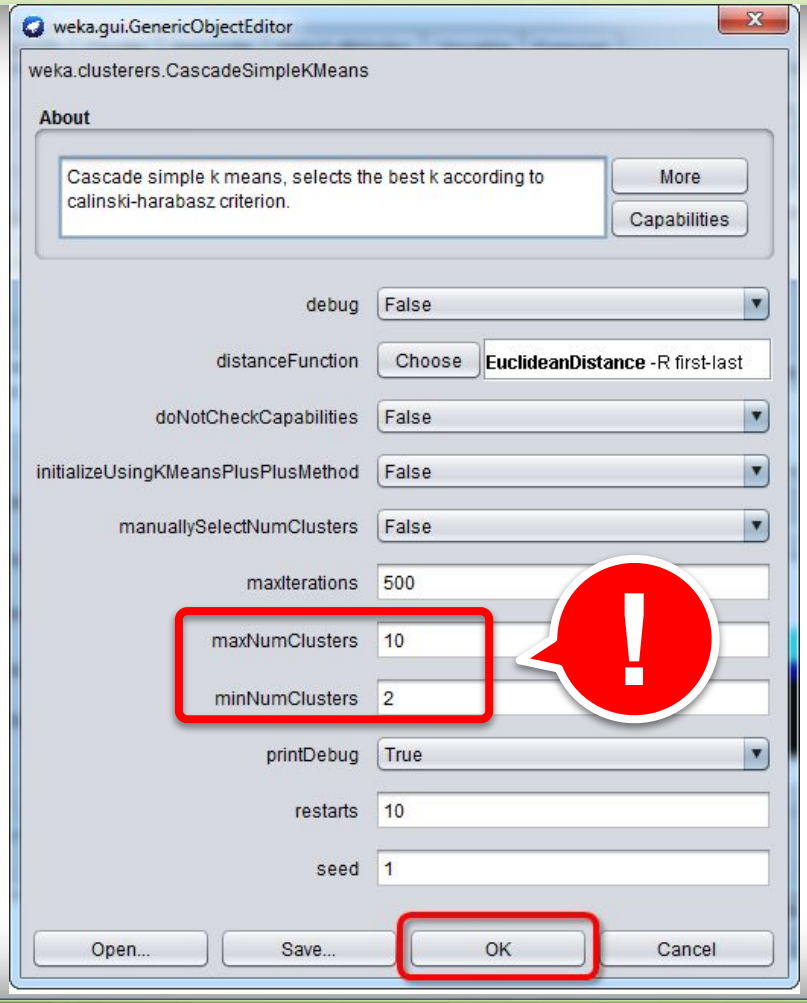
第2群(333) 下載

筆數*
var0: Gender=female*
var2: Age*
var3: Address=rural*
var6: FamiSize=>3*
var18: FatJob=other*
var21: ChoSchReason=course*
var24: ChoSchReason=reputation*

第2群

https://www.kapwing.com/explore/drake-reaction-meme-maker

能否調整為容易詮釋的分群數量？



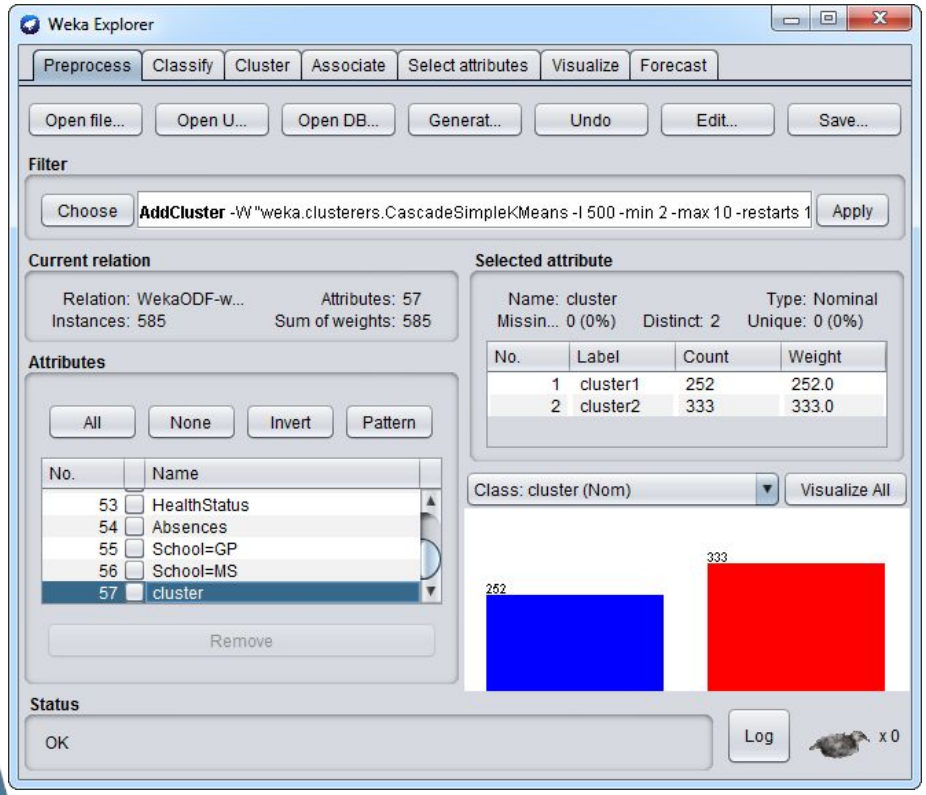
在前面 STEP 3.
CascadeSimpleKMeans
進階設定中

- maxNumClusters:
  分群數量上限
- minNumClusters:
  分群數量下線

建議設定為7~3比較好解釋

分群:層疊式K平均法

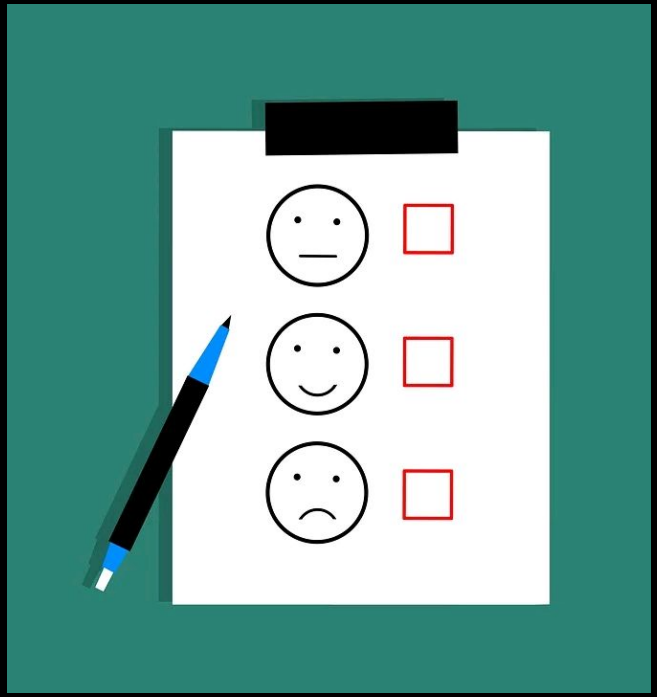

上機啦！


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Forecast
Open file...
Open U...
Open DB...
Generat...
Undo
Edit...
Save...
Filter
Choose
AddCluster -W "weka.clusterers.CascadeSimpleKMeans -I 500 -min 2 -max 10 -restarts 1
Apply
Current relation
Relation: WekaODF-w...
Attributes: 57
Instances: 585
Sum of weights: 585
Selected attribute
Name: cluster
Type: Nominal
Missin... 0 (0%)
Distinct: 2
Unique: 0 (0%)
No. Label Count Weight
1 cluster1 252 252.0
2 cluster2 333 333.0
Attributes
All
None
Invert
Pattern
No. Name
53 HealthStatus
54 Absences
55 School=GP
56 School=MS
57 cluster
Remove
Class: cluster (Nom)
Visualize All
252
333
Status
OK
Log
x 0

Part 4.

異常偵測

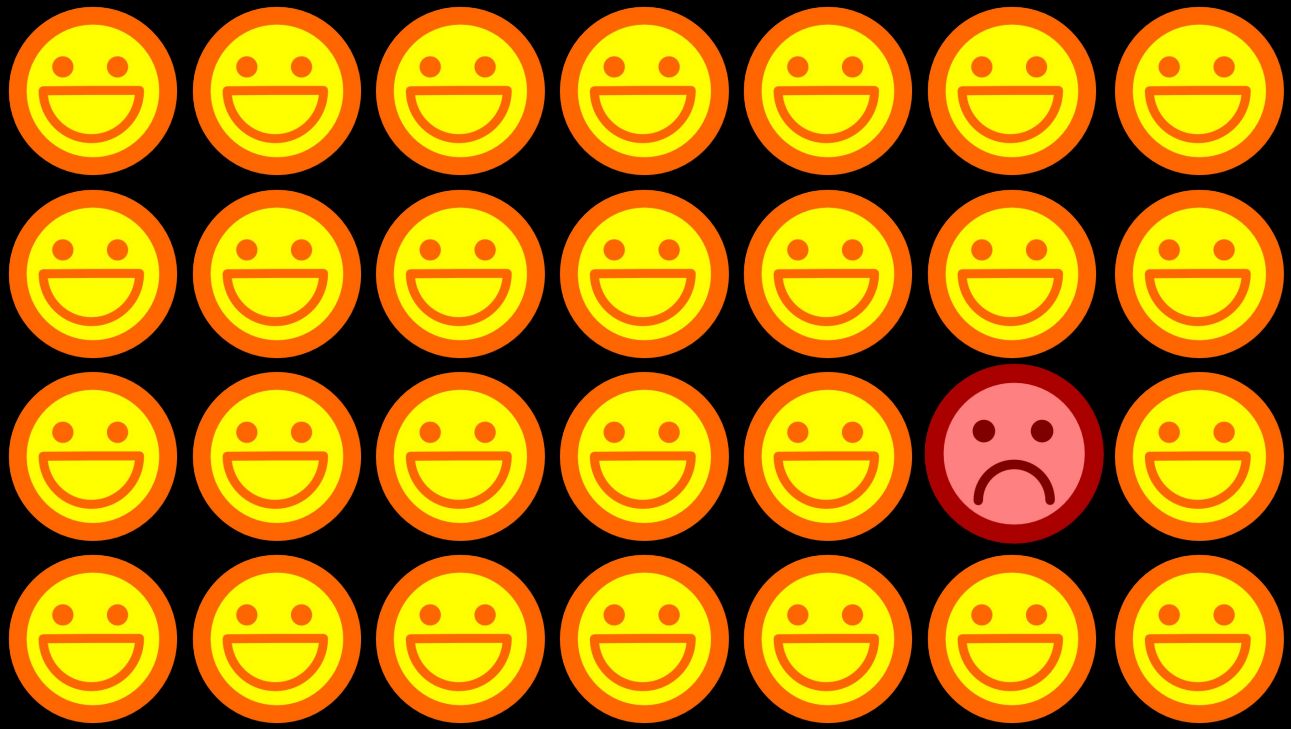
可以幫我做個學術研究嗎？



做完有抽禮券喔

https://pixabay.com/zh/illustrations/feedback-survey-questionnaire-3239454/

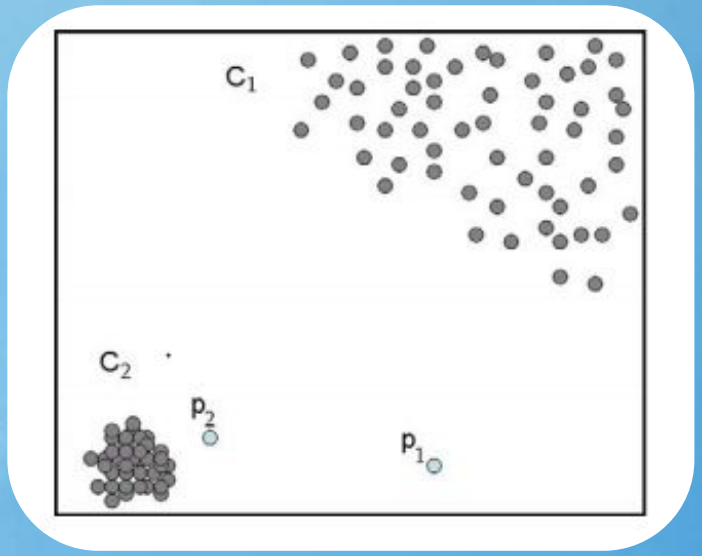
你是如何做問卷調 查的邏輯檢查？



裡面是不是有人怪怪的？

你確定螢幕後面的是人嗎？

https://www.daskeyboard.com/blog/cats-and-keyboards-a-guide-to-protecting-your-keyboard-from-feline-companions/



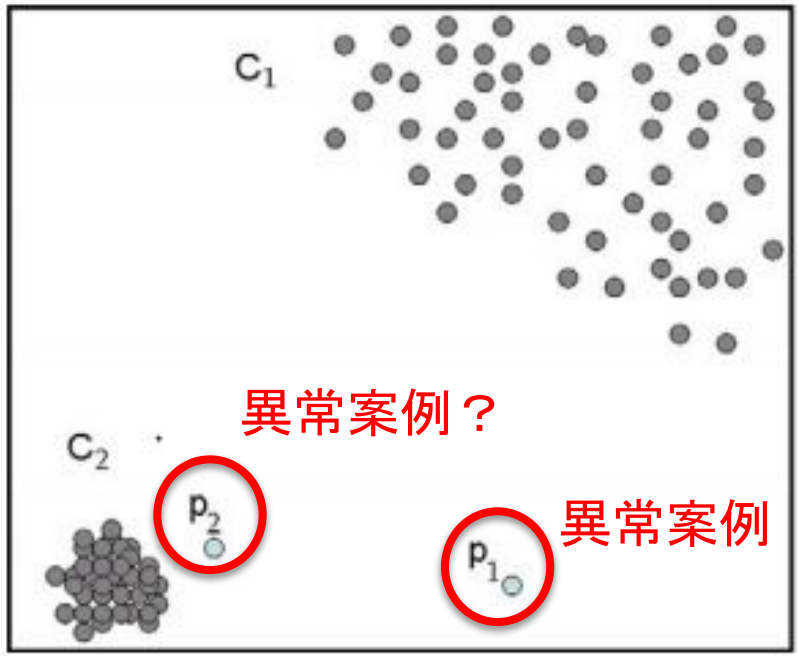
異常偵測演算法

Local Outlier Factor
區域異數因素

(Breunig, et al., 2000)

局部異數因素
演算法目標

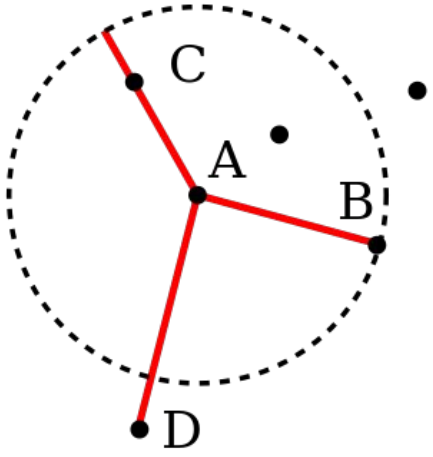
LOF = 相較於它周圍的鄰居，它異常的程度



https://medium.com/@arunm8489/local-outlier-factor-13784dc1992a

LOF計算公式

計算A的異常程度 (LOF) =
比較 A所在的的密度
和 A的鄰居所在的密度

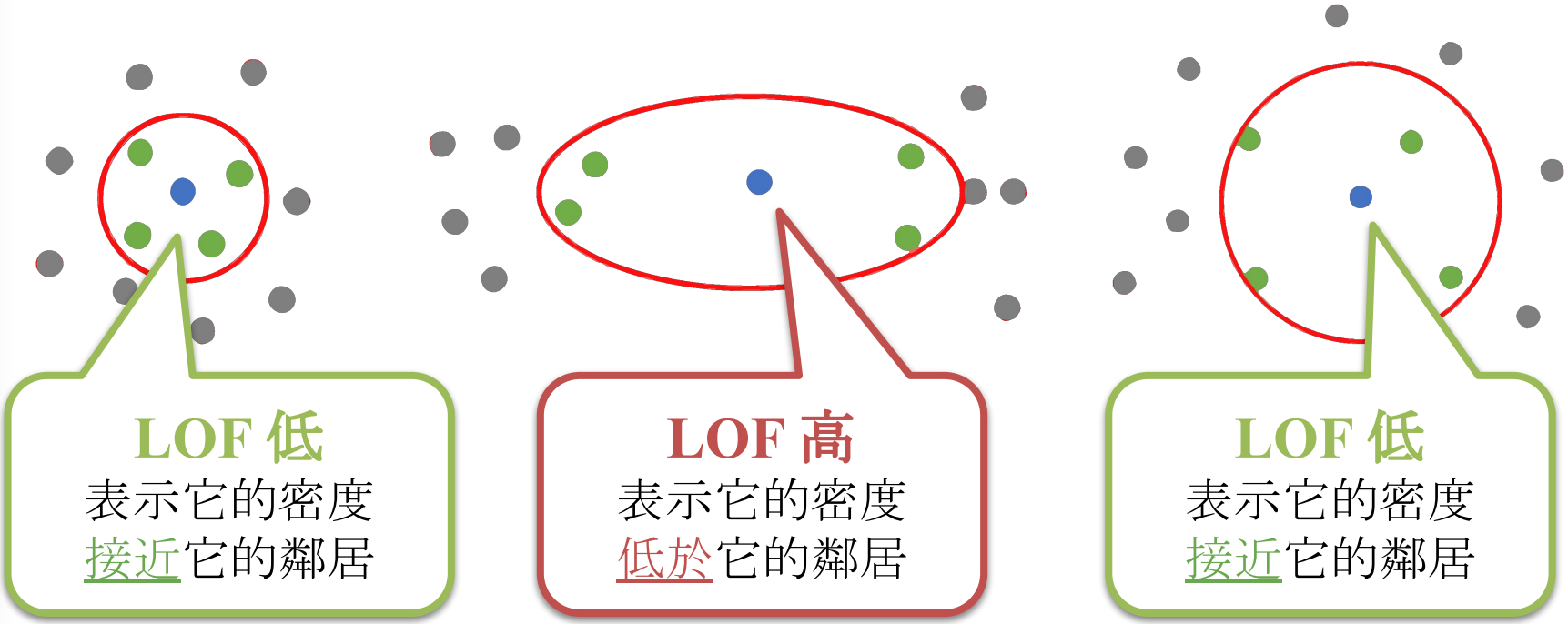


$$\mathrm{LOF_k} := \frac{\sum_{B \in N_k(A)} \frac{lrd(B)}{lrd(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} lrd(B)}{|N_k(A)|} \div lrd(A)$$

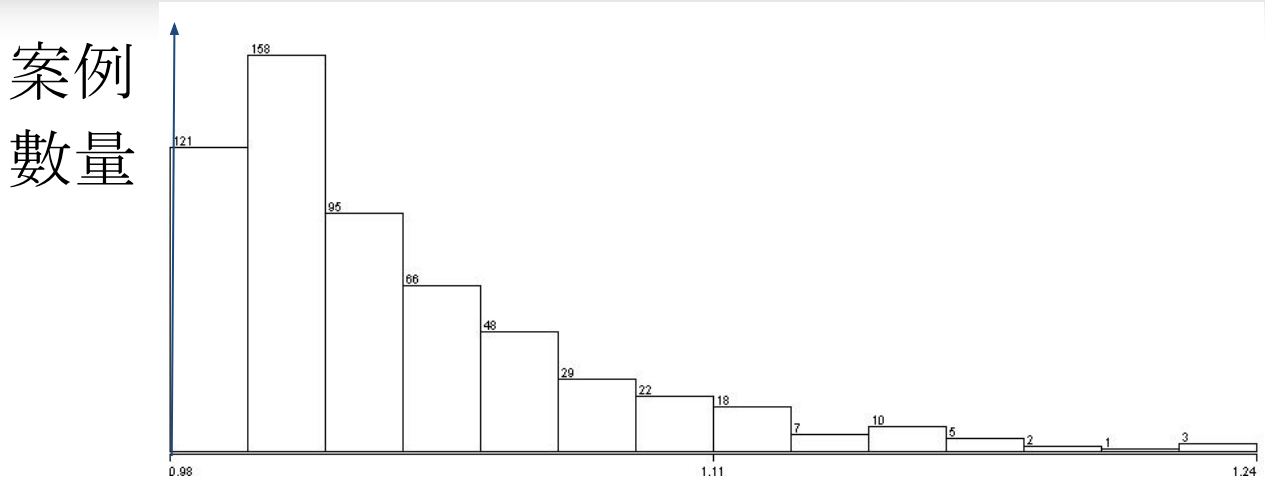
$$lrd_k(A) := 1 \Big/ \left(\frac{\sum_{B \in N_k(A)} reachability\text{-}distance_k(A,B)}{|N_k(A)|} \right)$$

$$reachability\text{-}distance_k(A,B) = max\{k\text{-}distance(B), d(A,B)\}$$

不同資料分佈的 LOF



LOF與次數分配圖



LOF < 1
表示它的密度
高於它的鄰居

LOF ~ 1
表示它的密度
接近它的鄰居

LOF > 1
表示它的密度
低於它的鄰居

https://www.wikiwand.com/en/Local_outlier_factor

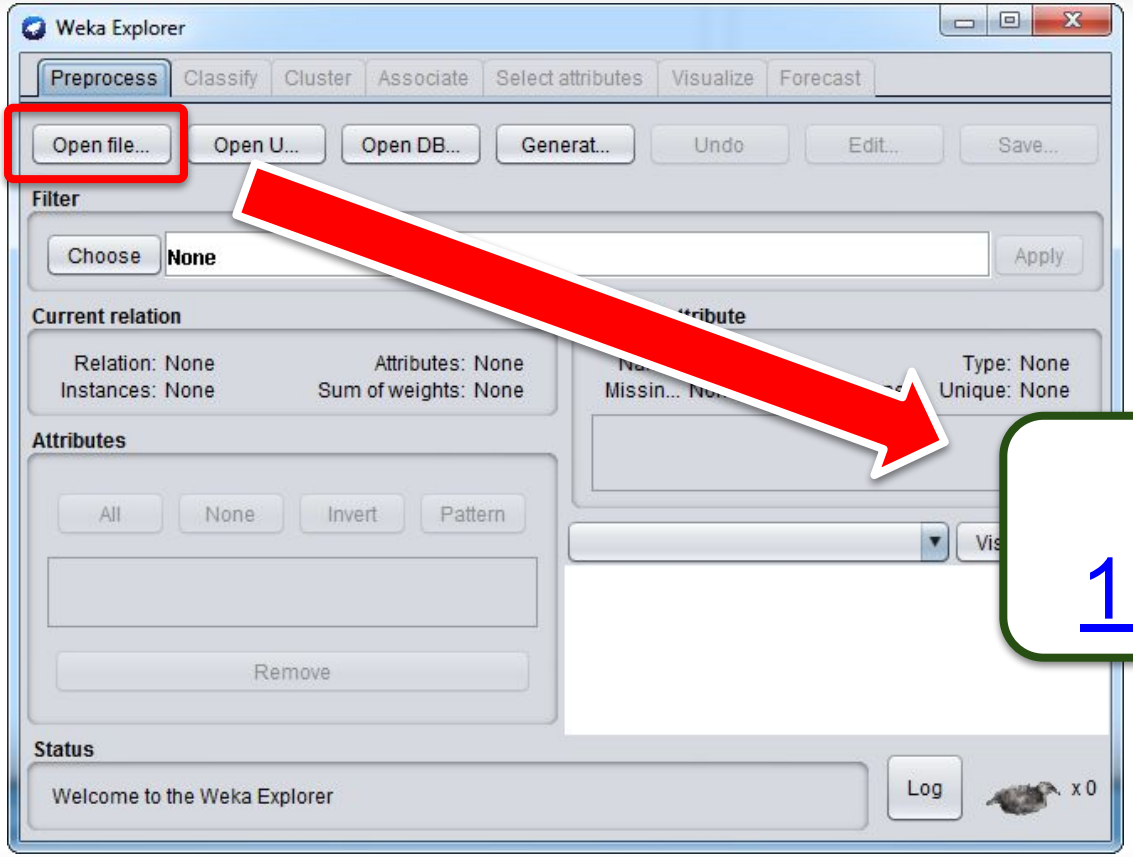
異常偵測:區域異數因素
實作步驟

1. 下載與開啟檔案
2. 資料前處理:關閉目標屬性
3. 執行異常偵測:LOF
4. 檢視探勘結果:LibreOffice Calc
   → AutoFilter

STEP 1. 下載與開啟檔案 (1/2)

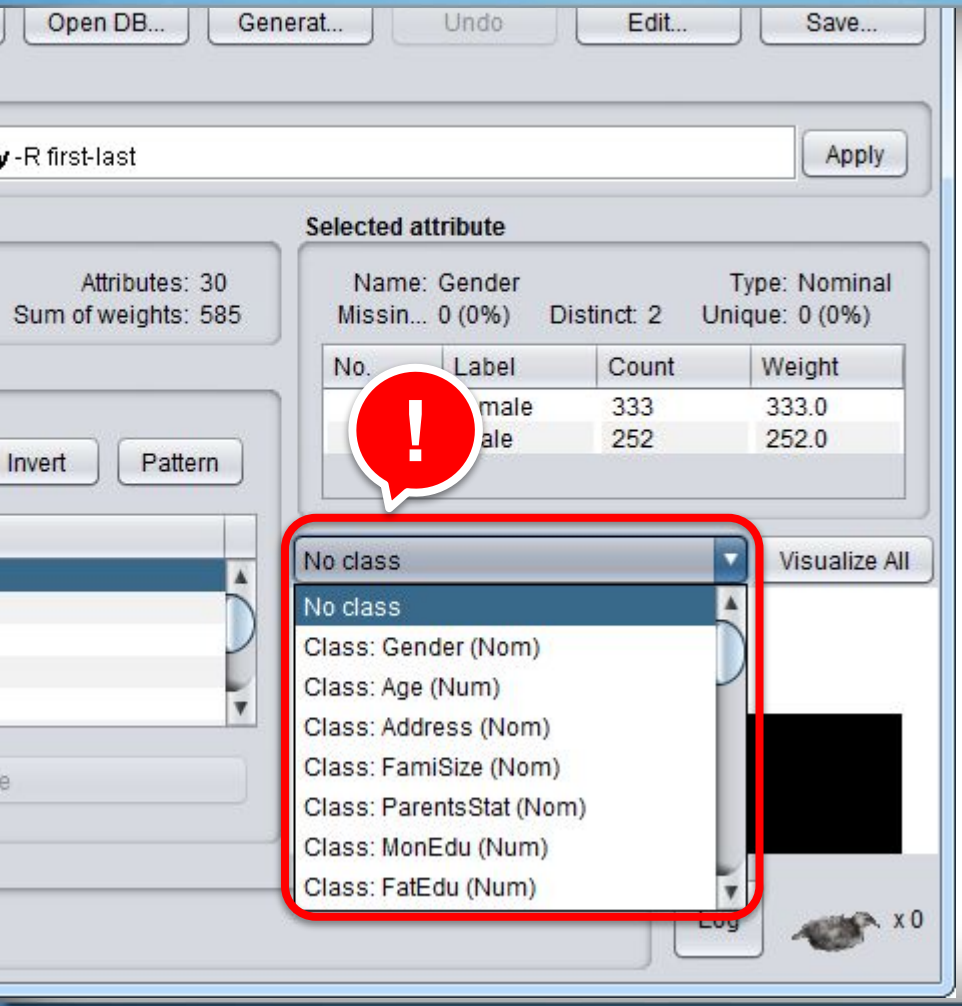
※ 跟前面是同一個檔案

STEP 1. 下載與開啟檔案 (2/2)



stu-sch-1 - train.ods

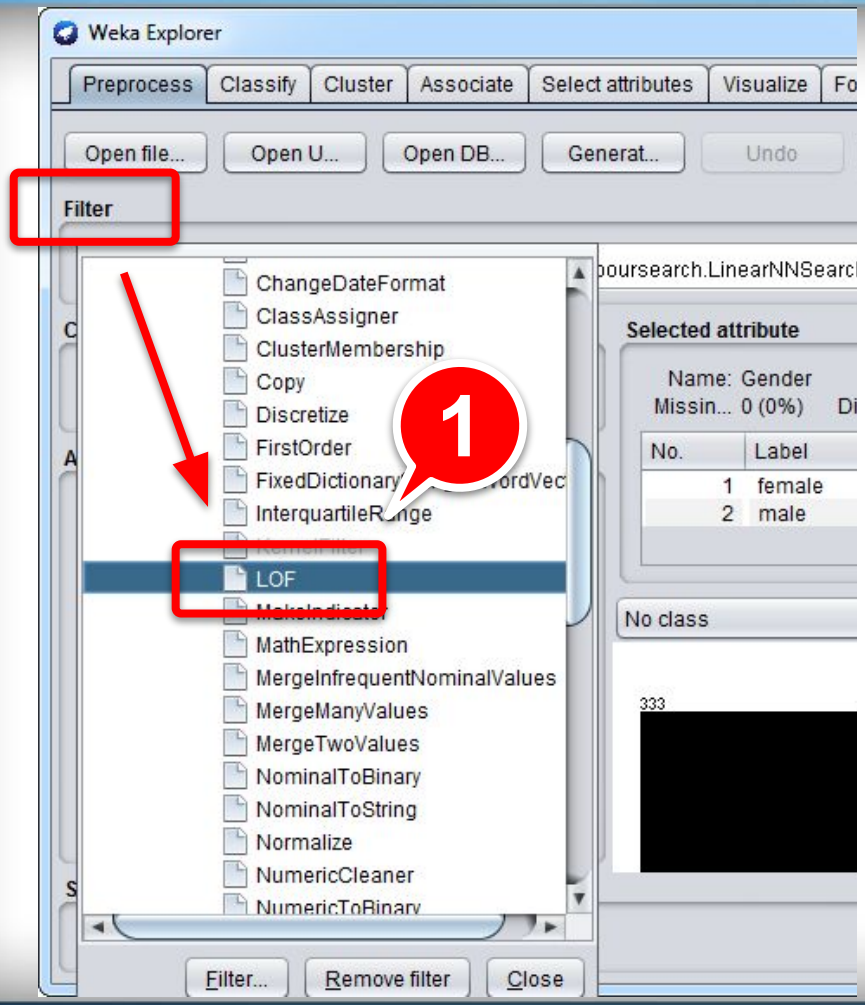
STEP 2. 資料前處理
關閉目標屬性



- 將目標屬性Class
  改選為No class

※ 探索性分析不使用目標屬性

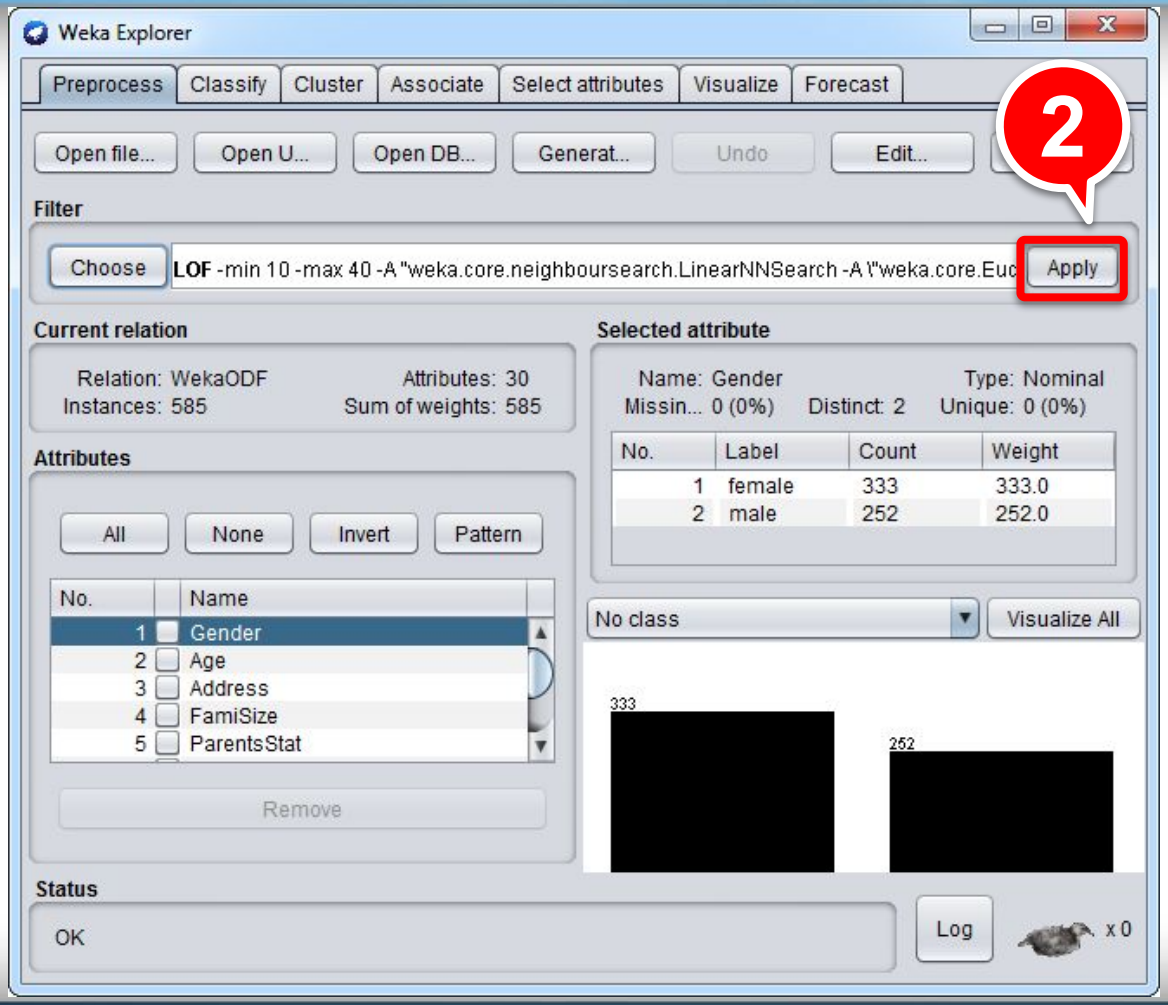
STEP 3. 執行異常偵測 (1/3)



1. Filter ⇨ Choose
   選擇篩選器
   weka.filters.unsupervised
   .attribute.**LOF**

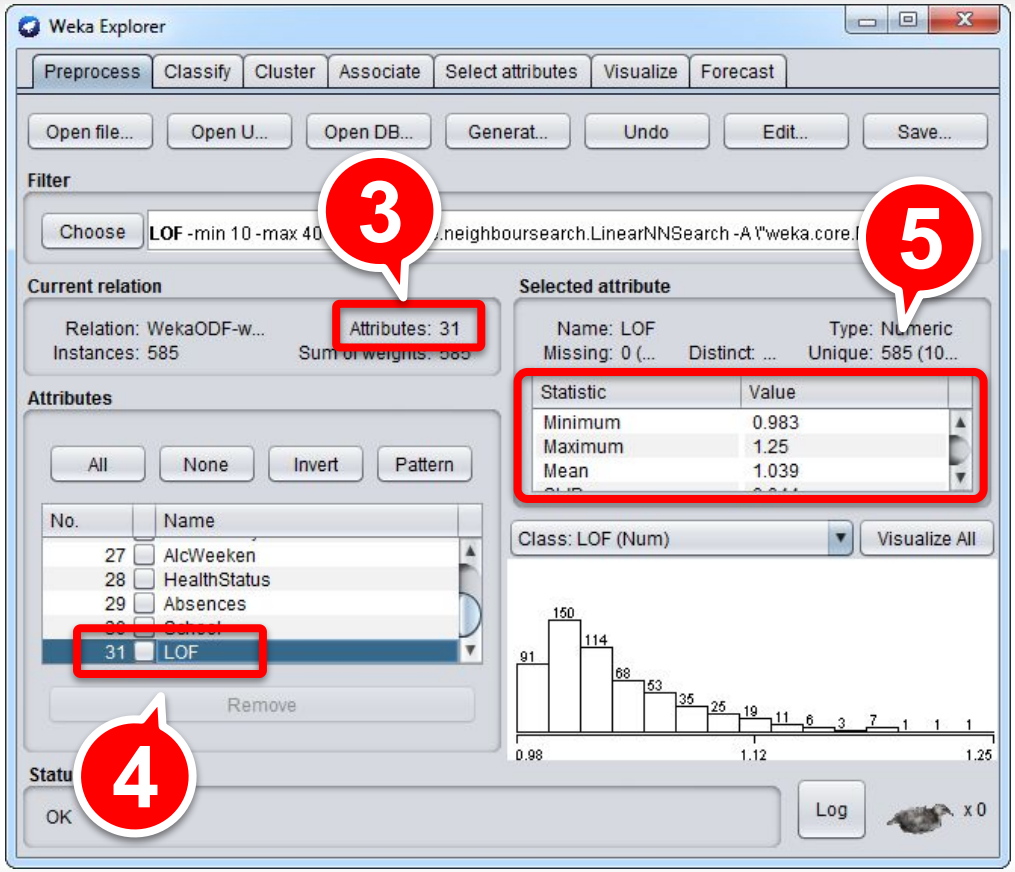
※ 需安裝套件localOutlierFactor

STEP 3. 執行異常偵測 (2/3)



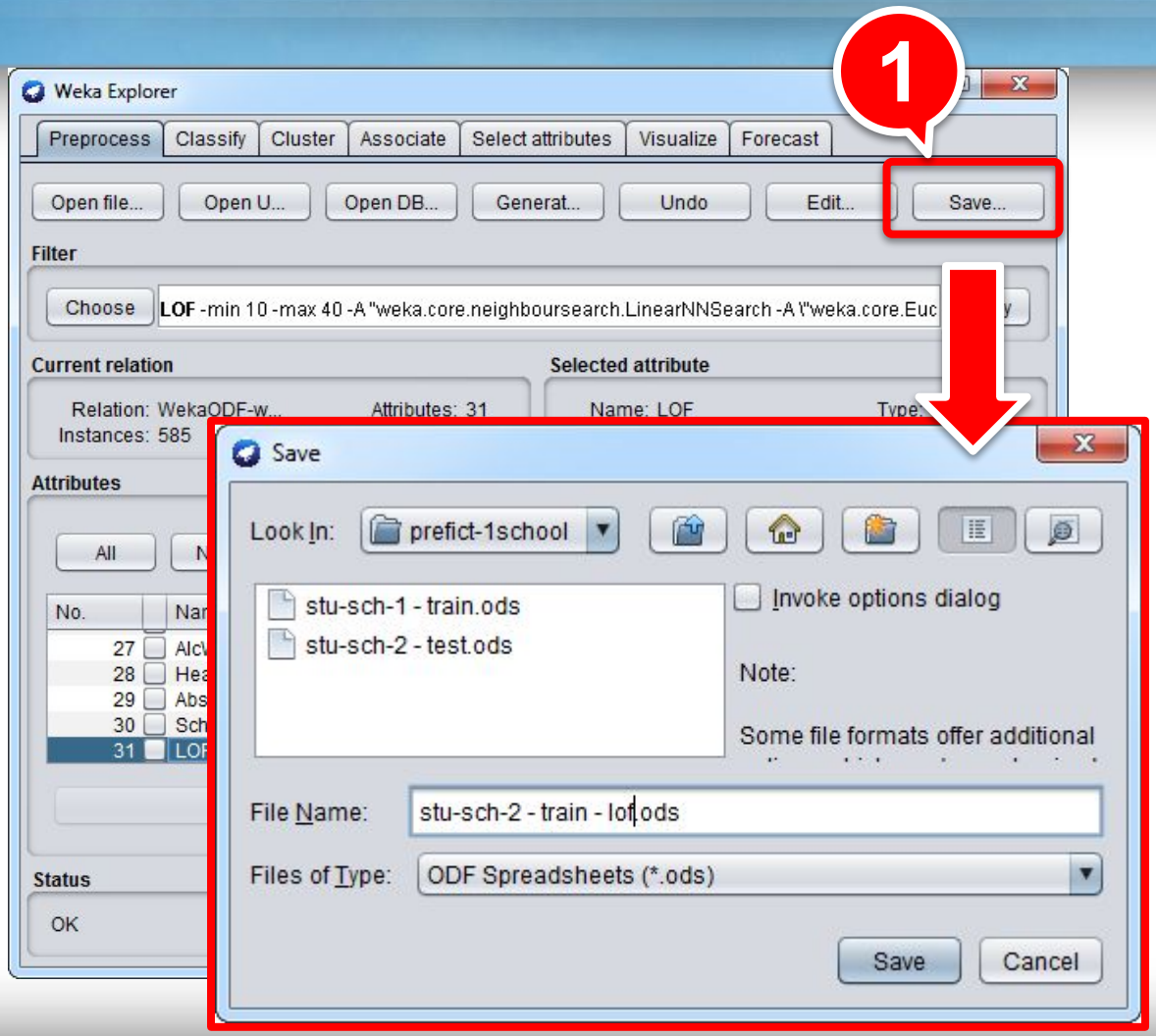
2. 按下Apply
套用篩選器

STEP 3. 執行異常偵測 (3/3)



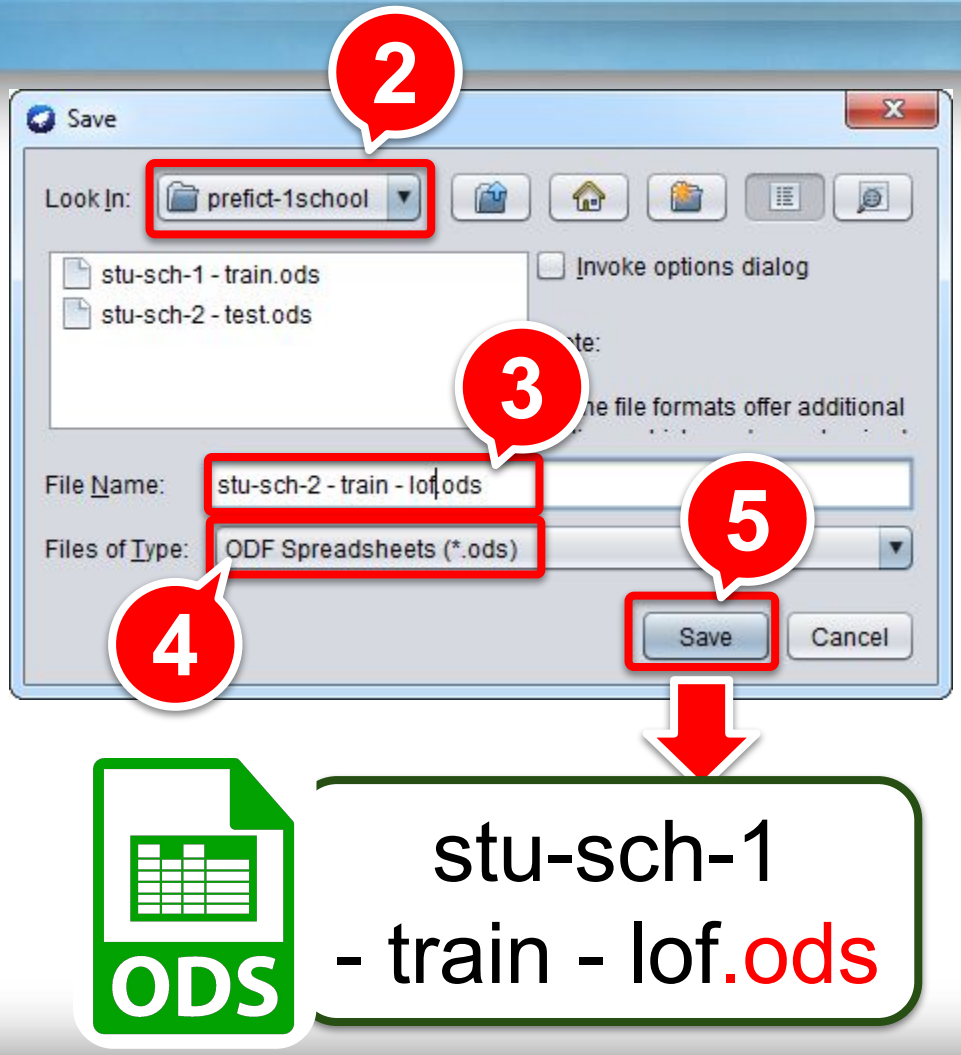
3. Attributes: 31
   屬性數量增加
   從30變成31個
4. 新增了LOF數值型屬性
5. 查看LOF資料分佈
   - 最小值 0.983
   - 最大值 1.25

STEP 4. 檢視探勘結果 (1/7)



1. Save 儲存檔案

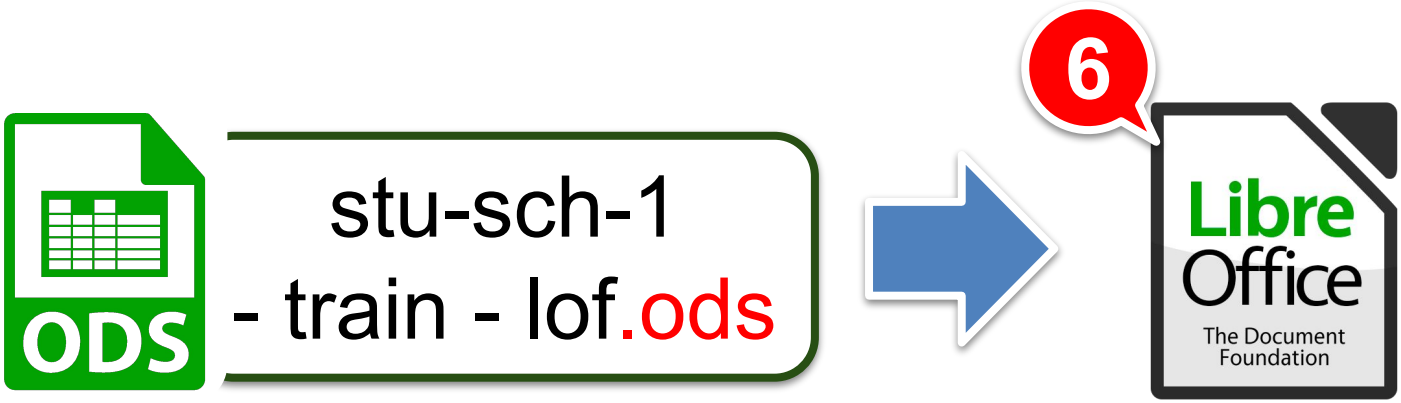
STEP 4. 檢視探勘結果 (2/7)



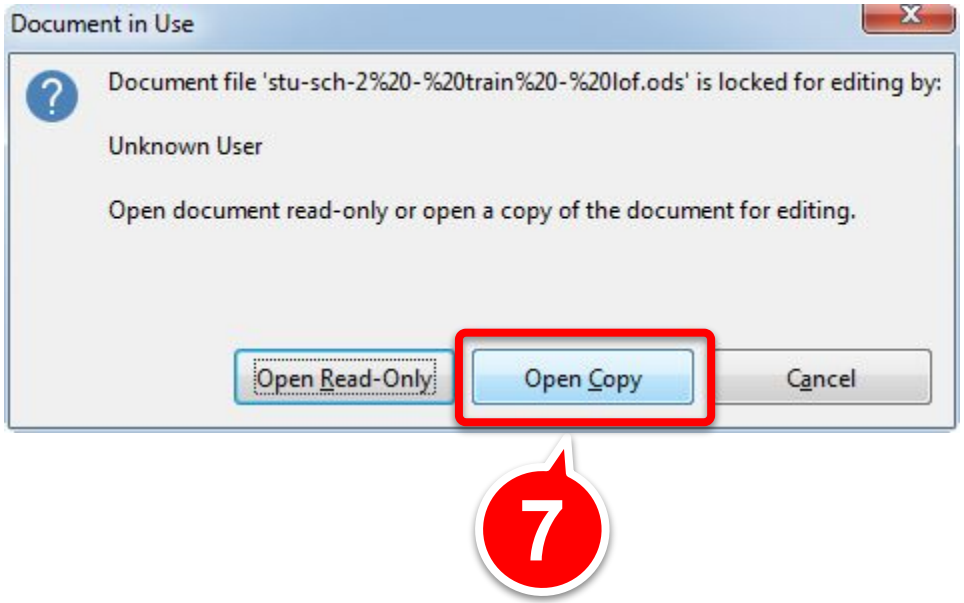
2. Look in:
   移動到下載資料夾
3. File Name: 檔案命名
   stu-sch-2 - train - lof.ods
4. Files of Type:
   ODF Spreadsheets (*.ods)
   以ODS檔案類型儲存
5. Save 儲存檔案
   此資料夾就會產生ODS檔案

STEP 4. 檢視探勘結果 (3/7)

6. 用LibreOffice開啟ODS類型檔案



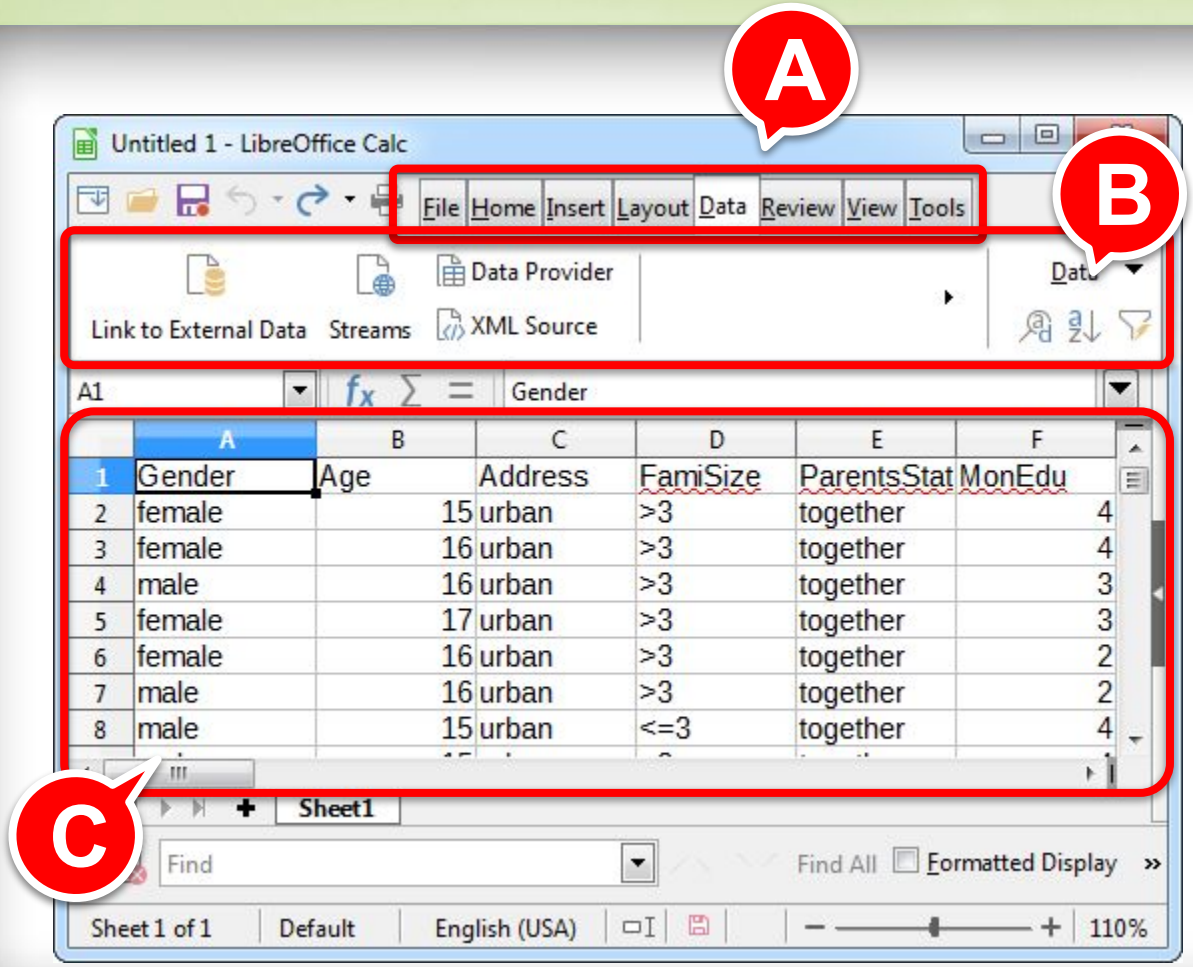
STEP 4. 檢視探勘結果 (4/7)



7. Open Copy
以副本模式開啟

(因為Weka程式
鎖定了原本的ODS檔案)

LibreOffice Calc介面說明



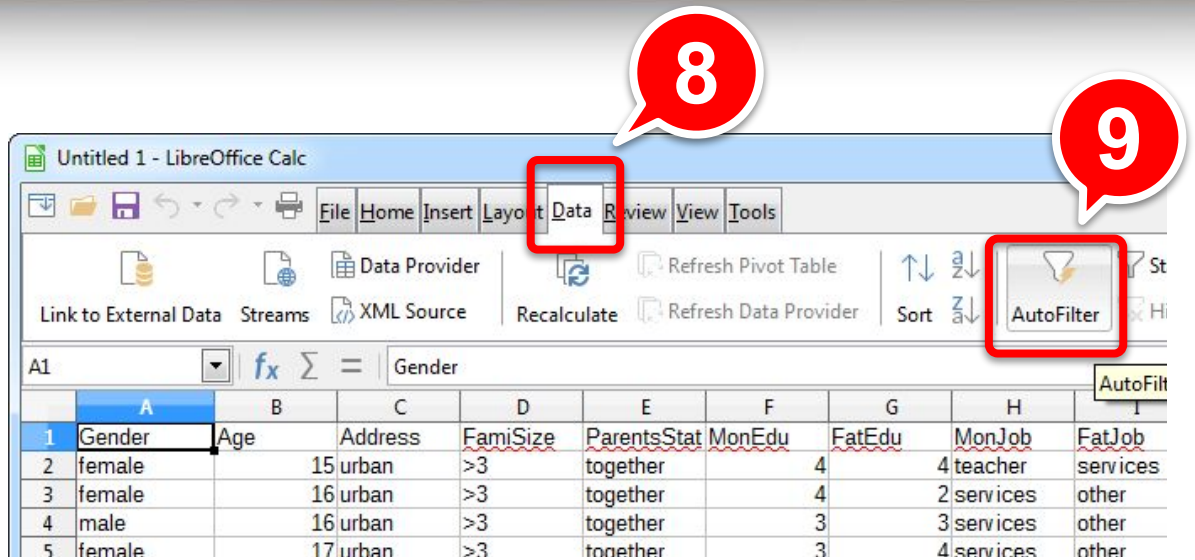
	A	B	C	D	E	F
1	Gender	Age	Address	FamiSize	ParentsStat	MonEdu
2	female	15	urban	>3	together	4
3	female	16	urban	>3	together	4
4	male	16	urban	>3	together	3
5	female	17	urban	>3	together	3
6	female	16	urban	>3	together	2
7	male	16	urban	>3	together	2
8	male	15	urban	<=3	together	4

A. 功能群組頁籤

B. 功能按鈕

C. 資料表

STEP 4. 檢視探勘結果 (5/7)



	A	B	C	D	E	F	G	H	I
1	Gender	Age	Address	FamiSize	ParentsStat	MonEdu	FatEdu	MonJob	FatJob
2	female	15	urban	>3	together	4	4	teacher	services
3	female	16	urban	>3	together	4	2	services	other
4	male	16	urban	>3	together	3	3	services	other
5	female	17	urban	>3	together	3	4	services	other

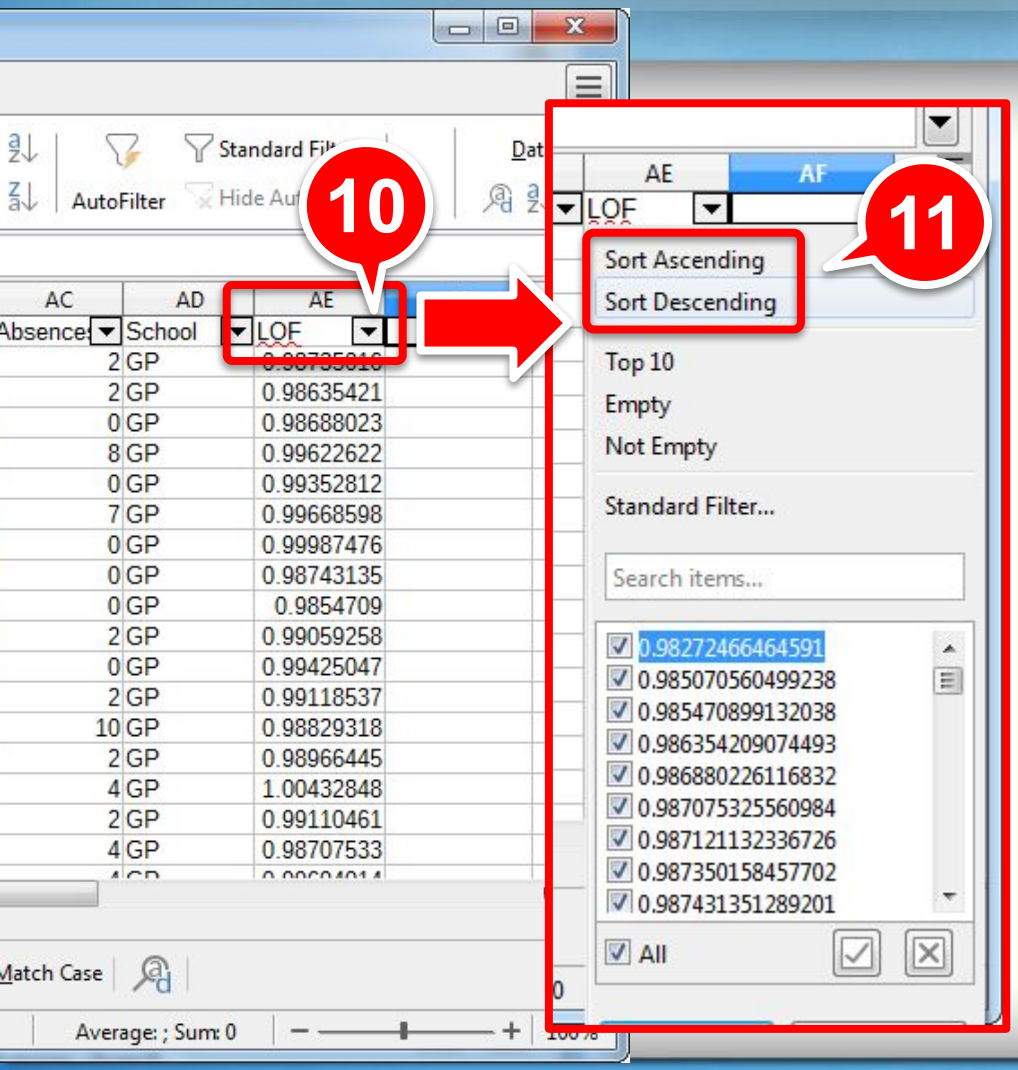
8. Data

   開啟資料的功能群組頁籤

9. AutoFilter

   啟動自動篩選功能

STEP 4. 檢視探勘結果 (6/7)



10. 找到最後一個直欄
    LOF
    點下右邊的下拉選單
    按鈕
11. 選擇排序
    - Sort Ascending
      由小到大排序
    - Sort Descending
      由大到小排序

STEP 4. 檢視探勘結果 (7/7)

LOF由大到小排序結果:異常案例

	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	RomanRe	FamiRela	Freetime	GoOut	AlcWorkd	AlcWeek	HealthSta	Absence	School	LOF
2	yes	5	4	5	5	5	1	12	GP	1.24976386
3	yes	5	5	5	5	5	5	0	GP	1.22100878
4	yes	5	5	5	5	5	5	2	MS	1.20700789
5	yes	1	3	3	5	5	3	0	GP	1.19240557
6	no	5	5	5	3	4	5	4	GP	1.18985426
7	no	5	5	3	1	1	5	0	GP	1.18712501
8	yes	1	2	1	1	1	1	4	MS	1.18555381
9	yes	1	3	5	3	5	1	8	GP	1.18345662
10	yes	1	5	5	4	3	5	12	GP	1.18262757
11	yes	1	3	2	2	3	1	24	GP	1.17995623

LOF由小到大排序結果:普通案例

	AC	AD	AE
1	Absence	School	LOF
2	2	GP	0.98735016
3	2	GP	0.98635421
4	0	GP	0.98688023
5	8	GP	0.99622622
6	0	GP	0.99352812
7	7	GP	0.99668598
8	0	GP	0.99987476
9	0	GP	0.98743135
10	0	GP	0.9854709
11	2	GP	0.99059258

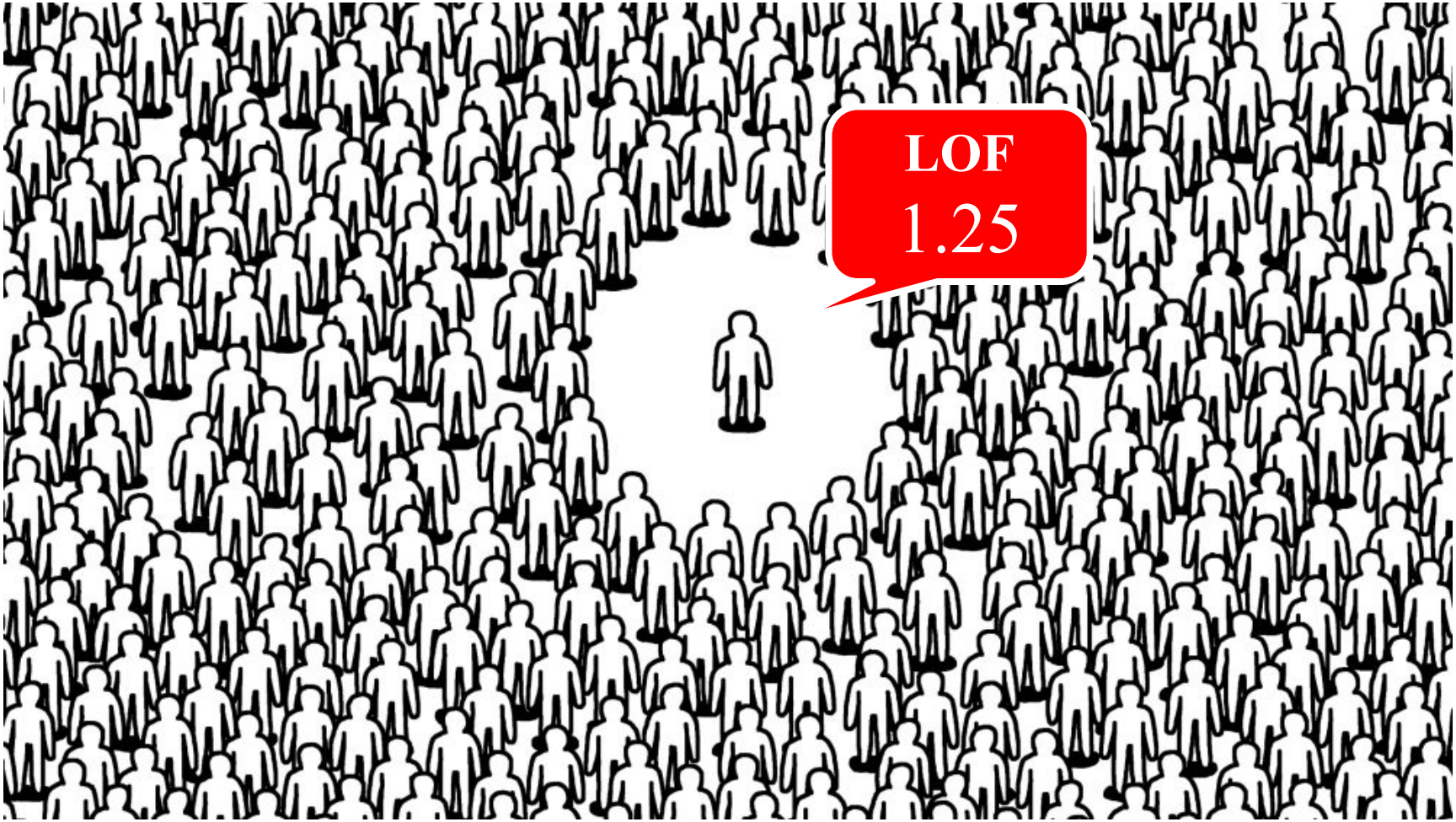
LOF	家庭關係 Fami Relation (1-5)	自由程度 Freetime (1-5)	出外程度 GoOut (1-5)	平日飲酒程度 Alc Workday (1-5)	週末飲酒程度 Alc Weeken (1-5)	健康狀況 Health Status (1-5)
1.25	5	4	4	5	5	1
1.22	5	5	5	5	5	5
0.99	4	2	2	1	1	5
0.99	4	2	3	1	1	5

是不是有人喝太多了?

LOF分數大於1:表示異常

LOF分數接近1:表示普通

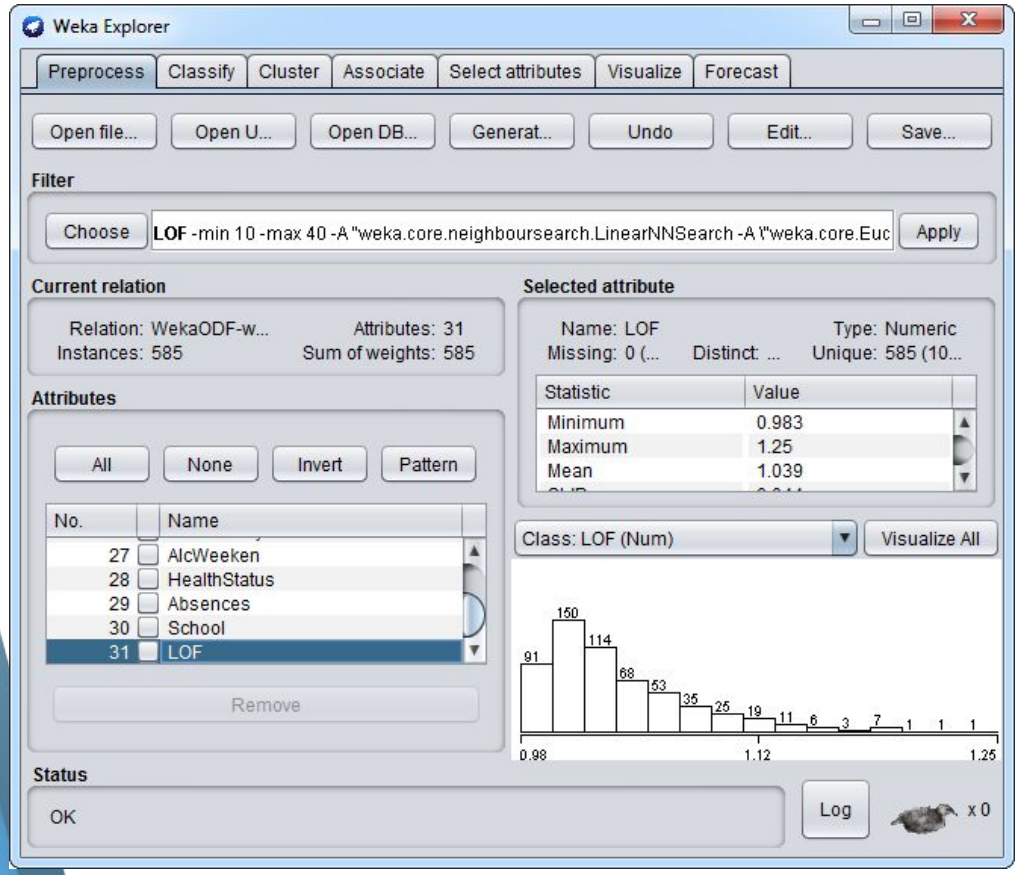
https://www.roblox.com/library/1664931194/THINKING-EMOJI-IS-A-MEME-thinking



Gotcha!

異常偵測:區域異數因素


上機啦！

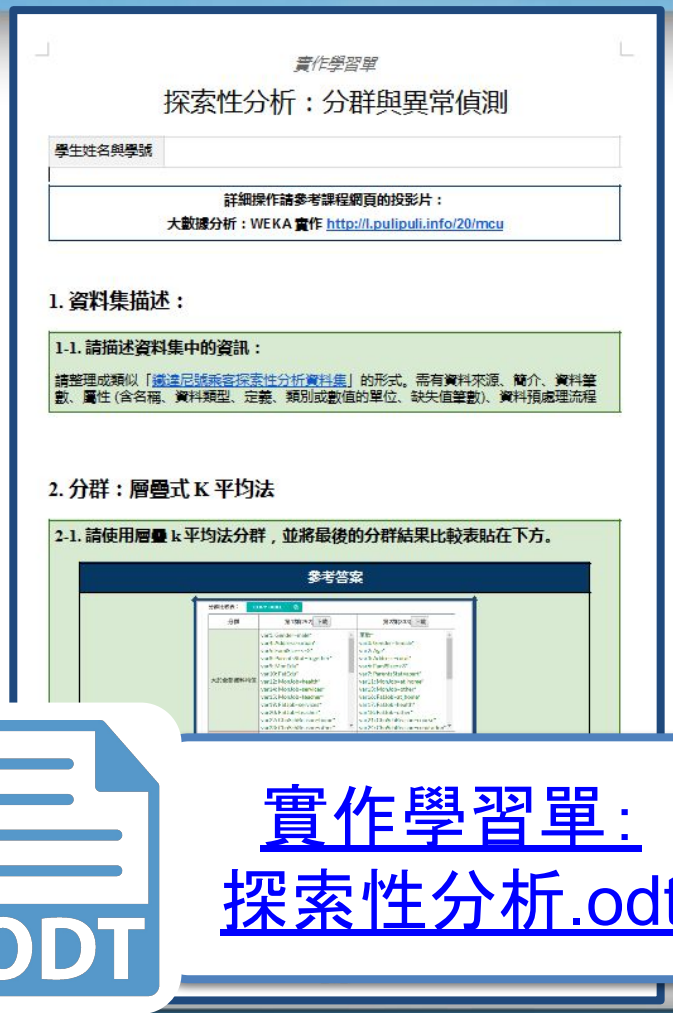

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Forecast
Open file...
Open U...
Open DB...
Generat...
Undo
Edit...
Save...
Filter
Choose
LOF -min 10 -max 40 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.Euc
Apply
Current relation
Relation: WekaODF-w...
Attributes: 31
Instances: 585
Sum of weights: 585
Selected attribute
Name: LOF
Type: Numeric
Missing: 0 (...
Distinct: ...
Unique: 585 (10...
Statistic
Value
Minimum
0.983
Maximum
1.25
Mean
1.039
Attributes
All
None
Invert
Pattern
No.
Name
27
AlcWeeken
28
HealthStatus
29
Absences
30
School
31
LOF
Remove
Class: LOF (Num)
Visualize All
150
114
91
68
53
35
25
19
11
6
3
7
1
1
1
0.98
1.12
1.25
Status
OK
Log
x 0

Part 5.

試著分析看看吧

學習單作業

作業說明



實作學習單

探索性分析：分群與異常偵測

學生姓名與學號

詳細操作請參考課程網頁的投影片：
大數據分析：WEKA 實作 http://l.pulipuli.info/20/mcu

1. 資料集描述：

1-1. 請描述資料集中的資訊：

請整理成類似「鐵達尼號乘客探索性分析資料集」的形式。需有資料來源、簡介、資料筆數、屬性（含名稱、資料類型、定義、類別或數值的單位、缺失值筆數）、資料預處理流程

2. 分群：層疊式 K 平均法

2-1. 請使用層疊 k 平均法分群，並將最後的分群結果比較表貼在下方。

參考答案

實作學習單：
探索性分析.odt

選一個資料集，
然後完成學習單吧！

1. 資料集的描述
2. 分群：層疊式K平均法
3. 異常偵測：區域異數因素
4. 綜合比較

銀行推銷了

那些客戶呢？

[銀行行銷資料集]

同學們的教學意見
有什麼模式呢？

教學意見回饋資料集

請同學記得填
教學意見回饋喔

美國公民

有什麼共同模式？

收入普查
資料集

大家是怎麼瀏覽
線上購物網站的呢？

線上購物
資料集

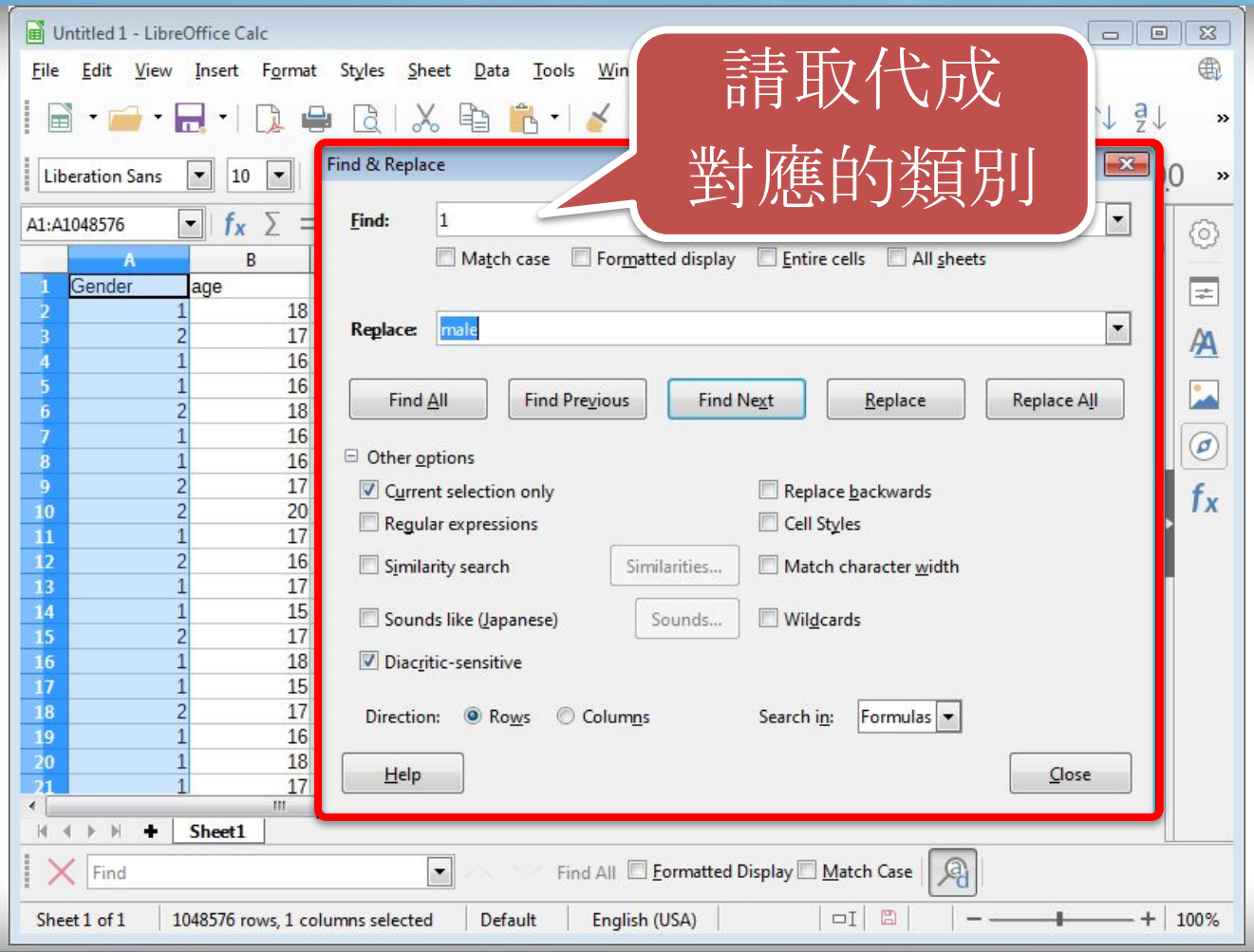
鐵達尼號乘客
有那些呢？

[鐵達尼號乘客資料集]

關於學習單的說明

1. 可選擇其他資料集, 不過注意以下條件
   a. 屬性儘量為類別型或數值型, 避免包含日期型或字串型等需要額外資料預處理的資料類型
2. 如果分析結果難以解釋, 不妨重新整理資料本身
   a. 試著刪除屬性、調整的值、或是想辦法填補缺漏值
   b. 如果有對資料本身做預處理, 請在學習單中說明做法
3. 學習單大部分都沒有標準答案, 請發揮觀察力和想象力, 試著分析看看吧

屬性是類別資料卻記成數 值的話？



- 性別
- 地區

布丁布丁吃什麼？
http://blog.pulipuli.info