**教育部補助行動寬頻尖端技術課程推廣計畫**

課程模組：巨量資料分析技術與應用

**期末作業報告**

學校系所：明新科技大學資訊管理系

課程名稱：巨量交易分析與應用

開課班級：四資四乙

組 別：第 十 組

組 員：B03090016 蕭兆軒

B03090073 黃宜廉

B03090084 呂韋志

B03090099 高翔廷

中華民國　107 年　1 月　4 日

**索 引**

**作業一：計程車資料大數據分析 3**

1-1: 最熱門的載客與下車地點為何 5

1-2: 幾點最容易搭到計程車 6

1-3: 天氣會影響到搭乘計程車的意願嗎 7

1-4: 長距離載客會得到較多的小費嗎 8

**作業二：字數計算大數據應用 9**

2-1:撰寫一word count程式，計算範例檔案各字詞出現的頻率 10

2-2:計算範例一其付款方式(Payment\_type)欄位，各種付款方式的總次數 11**作業三：美國大聯盟MLB資料分析 12**

3-1: 2016年的全壘打王是誰 13

3-2: 列出歷年來的得分王 14

**作業四：美國航班資料分析 15**

4-1: 2008年的平均航班延遲時間 17

4-2: 哪些機場出發的飛機，會延遲兩倍平均航班延遲時間以上 17

**作業五：WEKA與資料探勘 18**

5-1: WEKA的安裝過程 18

5-2: 利用決策樹進行天氣探勘 19

5-3: 利用貝式網路進行天氣因素探勘 21

**作業六：Tableau - 數據視覺化與應用 23**

6-1: Tableau的安裝過程 23

6-2: 超級市場範例數據資訊化呈現 25

6-3: 數據地圖化呈現 29

6-4: Dashboard整合型呈現 30

6-5: Tableau雲端發佈 31

**作業七：R語言大數據實戰 33**

7-1: R與R-Studio的安裝 33

7-2: R語言圖形Demo 35

7-3: IRIS圖形統計呈現 36

7-4: GoogleVis水果數據動態呈現 38

**作業八：SPSS數據統計分析 40**

8-1: Kent State Univ. Chi-Square範例 40

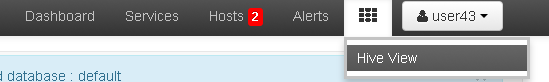
8-2: Math Achievement One-Way ANOVA範例 45

**作業一：計程車資料大數據分析**

**資料表新增**

網址輸入 h7.im.must.edu.tw:8080

Hive View 進入 Wortsheet



在Worksheet輸入以下指令，建立資料表

use user43;

CREATE TABLE taxi (

VendorID STRING,

tpep\_pickup\_datetime TIMESTAMP,

tpep\_dropoff\_datetime TIMESTAMP,

passenger\_count SMALLINT,

trip\_distance DOUBLE,

RatecodeID SMALLINT,

store\_and\_fwd\_flag STRING,

PULocationID STRING,

DOLocationID STRING,

payment\_type STRING,

fare\_amount DOUBLE,

extra DOUBLE,

mta\_tax DOUBLE,

tip\_amount DOUBLE,

tolls\_amount DOUBLE,

improvement\_surcharge DOUBLE,

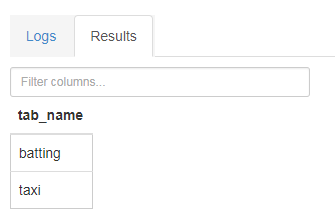
total\_amount DOUBLE)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' tblproperties("skip.header.line.count"="2");

觀察是否建立成功

use user43;

show tables;

****

導入資料，用 hdfs dfs -ls 確認有taxi.csv這個檔案

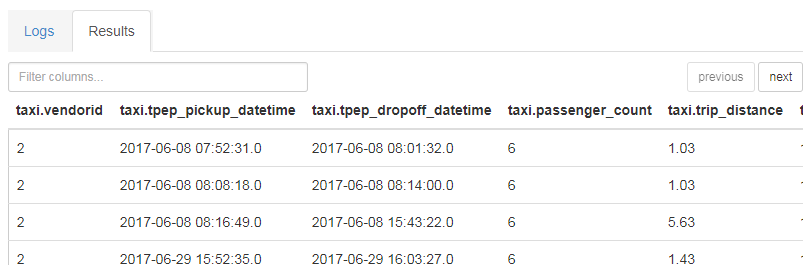
use user43;

LOAD DATA INPATH '/user/user43/taxi.csv' OVERWRITE INTO TABLE taxi;

確認導入資料是否成功

use user43;

select \* from taxi LIMIT 100;



**1-1: 最熱門的載客與下車地點為何**

最熱門載客地點

use user43;

select pulocationid, count(pulocationid) as pucount

from taxi

group by pulocationid

order by pucount desc;

結果

地點pulocationid 237

次數pucount 379701

最熱門下車地點

use user43;

select dolocationid, count(dolocationid) as docount

from taxi

group by dolocationid

order by docount desc;

結果

地點dolocationid 161

次數docount 364226

**1-2: 幾點最容易搭到計程車**

詳細至秒數

use user43;

select tpep\_pickup\_datetime, count(passenger\_count) as passengerCount

from taxi

group by tpep\_pickup\_datetime

order by passengerCount desc;

結果

時間tpep\_pickup\_datetime 2017-06-01 21:26:02.0

次數passengercount 20

詳細至小時

use user43;

select puhour, count(\*) avg0 from (

select hour(taxi.tpep\_pickup\_datetime) puhour, \*

from taxi ) x

group by puhour

order by avg0;

結果

時間(小時)puhour 18

次數 avg0 586108

**1-3: 天氣會影響到搭乘計程車的意願嗎**

無法得知，因為資料表無紀錄天氣資料。

**1-4: 長距離載客會得到較多的小費嗎**

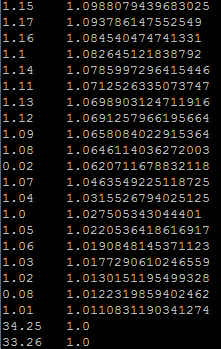
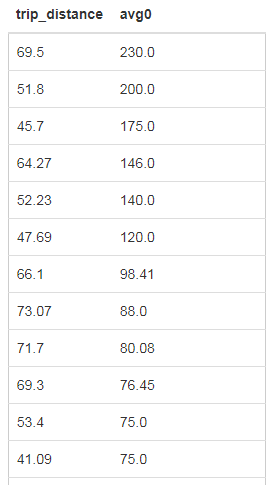
use user43;

select trip\_distance, avg(tip\_amount) avg0 from taxi

group by trip\_distance

order by avg0 desc;

結果，取自小費最多和小費趨近於1比較，大致上看起來長距離的確會獲得較多的小費(排除較極端數據)。



**作業二：字數計算大數據應用**

**資料表新增**

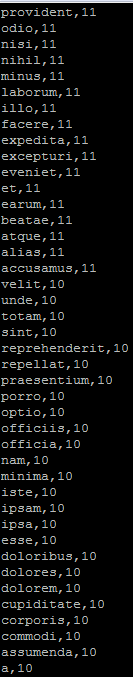
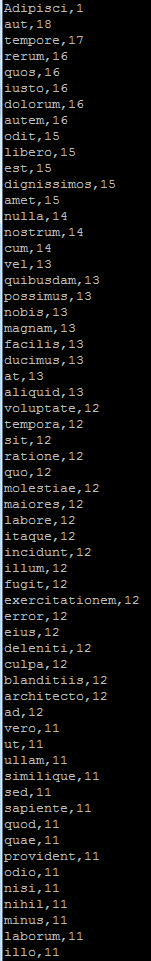
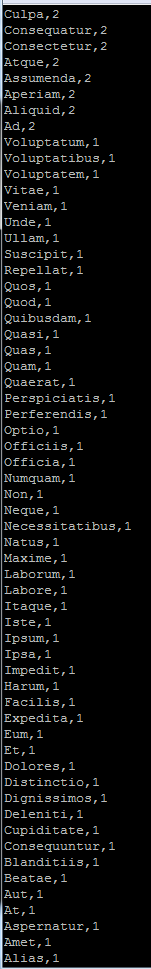
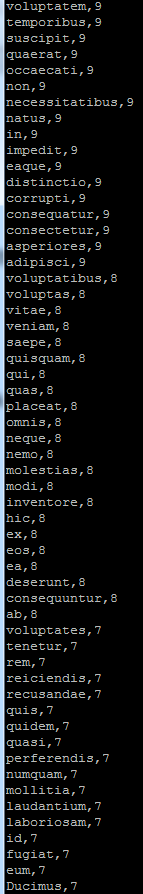
方法同上，作業一：計程車資料大數據分析，資料表新增。

已新增過。

**2-1:撰寫一word count程式，計算範例檔案各字詞出現的頻率**

使用範例檔 text.txt，輸入指令並儲存成一個txt檔

cat text.txt| tr -sc 'A-Za-z' '\n' | sort | uniq -c | sort -r | awk ' { t = $1; $1 = $2; $2 = t; print; } ' | tr ' ' ',' > final.txt



結果如上圖，按照出現最多的名詞次數排列。

**2-2:計算範例一其付款方式(Payment\_type)欄位，各種付款方式的總次數**

1= Credit card (信用卡)

2= Cash (現金)

3= No charge (免費)

4= Dispute (爭議)

5= Unknown (未知)

6= Voided trip(空車)

use user43;

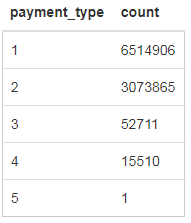
select payment\_type,count(payment\_type) count

from taxi

group by payment\_type

order by payment\_type;

結果

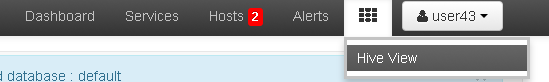


**作業三：美國大聯盟MLB資料分析**

**資料表新增**

網址輸入 h7.im.must.edu.tw:8080

Hive View 進入 Wortsheet



在Worksheet輸入以下指令，建立資料表

use user43;

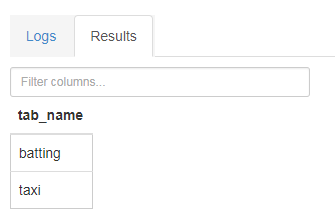
create table batting

(playerID STRING, yearID STRING, stint INT, teamID STRING, lgID STRING, G INT, AB INT, R INT, H INT, twoB INT, threeB INT, HR INT, RBI INT, SB INT, CS INT, BB INT, SO INT, IBB INT, HBP INT, SH INT, SF INT, GIDP INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' tblproperties("skip.header.line.count"="1");

觀察是否建立成功

use user43;

show tables;

****

導入資料，用 hdfs dfs -ls 確認有batting.csv這個檔案

use user43;

LOAD DATA INPATH '/user/user??/Batting.csv' OVERWRITE INTO TABLE batting;確認導入資料是否成功

use user43;

select \* from batting LIMIT 10;



**3-1: 2016年的全壘打王是誰**

use user43;

select T.playerid, T.yearid, T.hr from batting T where T.yearid=2016 and T.hr in (select max(hr) from batting where yearid=2016);

結果

球員trumbma01

年分2016

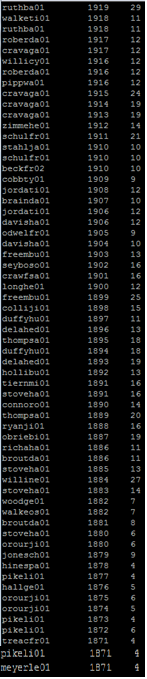
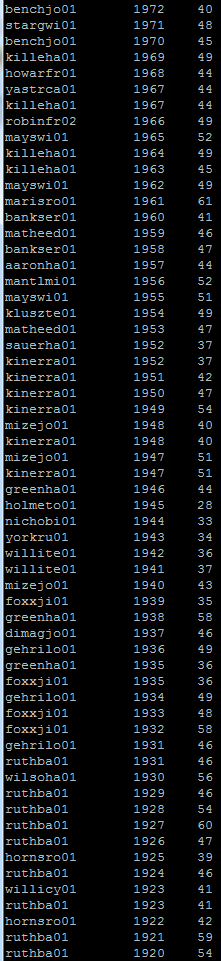
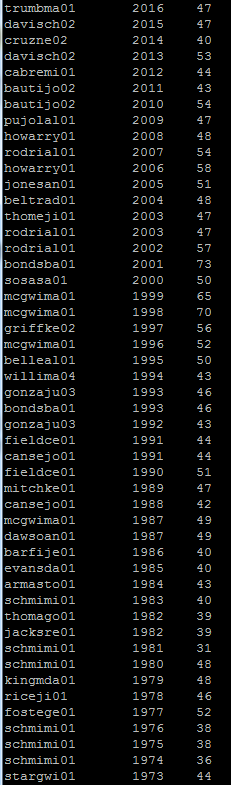
全壘打(count)47

**3-2: 列出歷年來的得分王**

use user43;

select T.playerid, T.yearid, T.hr from batting T join (select yearid, max(hr) as hr from batting group by yearid) a on T.yearid=a.yearid and T.hr=a.hr order by t.yearid desc;

結果

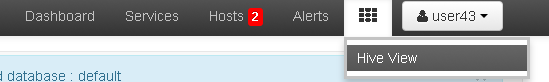


**作業四：美國航班資料分析**

**資料表新增**

網址輸入 h7.im.must.edu.tw:8080

Hive View 進入 Wortsheet



在Worksheet輸入以下指令，建立資料表

use user43;

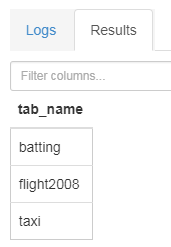
create table flight2008

(Year INT, Month INT, DayofMonth INT, DayOfWeek INT, DepTime INT, CRSDepTime INT, ArrTime INT, CRSArrTime INT, UniqueCarrier STRING, FlightNum STRING, TailNum STRING, ActualElapsedTime INT,CRSElapsedTime INT, AirTime INT,ArrDelay INT,DepDelay INT,Origin STRING,Dest STRING,Distance INT,TaxiIn INT,TaxiOut INT,Cancelled INT, CancellationCode INT,Diverted INT, CarrierDelay INT, WeatherDelay INT, NASDelay INT,SecurityDelay INT,LateAircraftDelay INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' tblproperties("skip.header.line.count"="1");

觀察是否建立成功

use user43;

show tables;

****

導入資料，用 hdfs dfs -ls 確認有2008.csv這個檔案

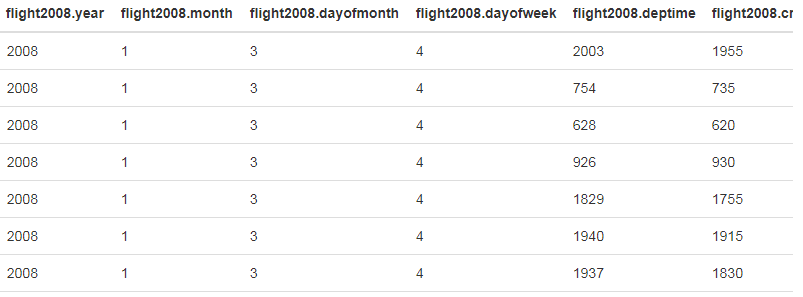
use user43;

LOAD DATA INPATH '/user/user43/2008.csv' OVERWRITE INTO TABLE flight2008;

確認導入資料是否成功

use user43;

select \* from flight2008 LIMIT 10;



**4-1: 2008年的平均航班延遲時間**

use user43;

select avg(depdelay) from flight2008;

結果

平均(分) 9.972570088930182

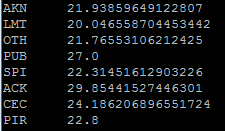
**4-2: 哪些機場出發的飛機，會延遲兩倍平均航班延遲時間以上**

兩倍時間=9.972570088930182\*2=19.94514017786036

use user43;

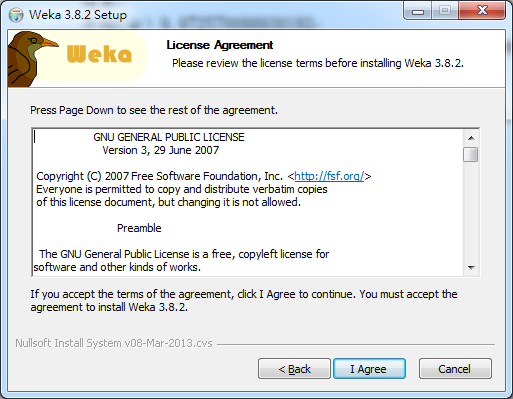
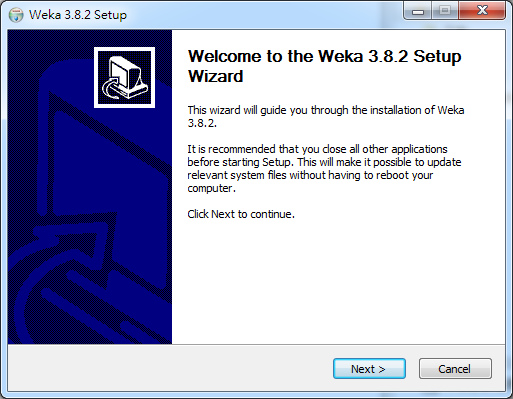
select origin, avg(depdelay) as delay from flight2008 group by origin having avg(depdelay) >19.94514017786036;

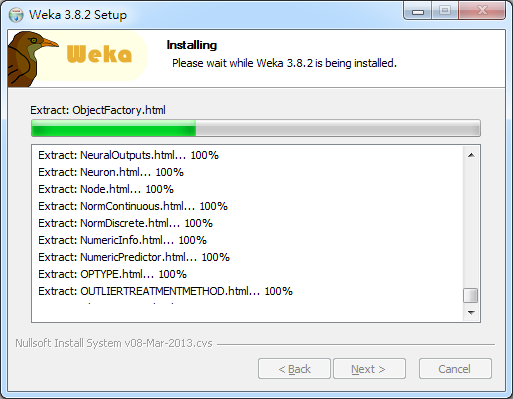
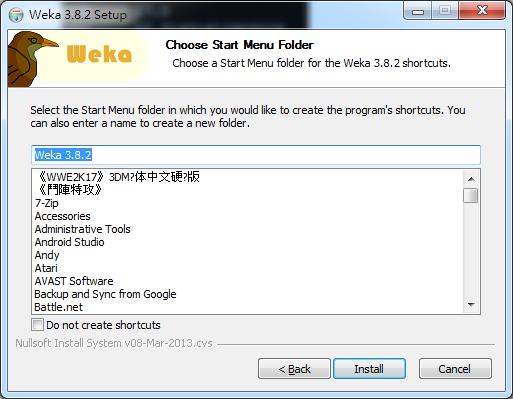
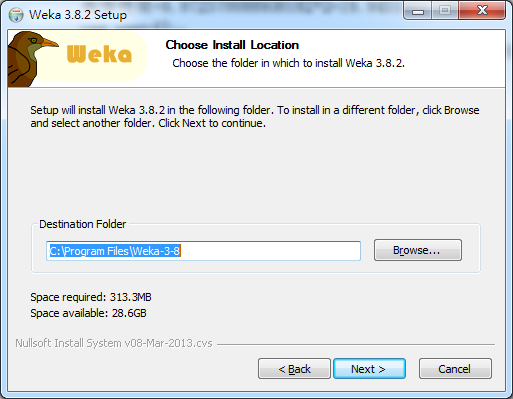
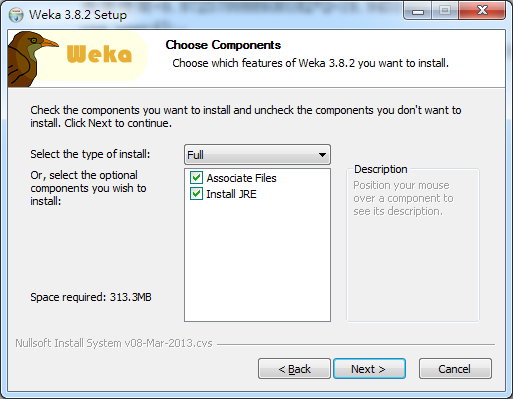
結果



**作業五：WEKA與資料探勘**

**5-1: WEKA的安裝過程**





改文件編碼

C:\Program Files\Weka-3-8

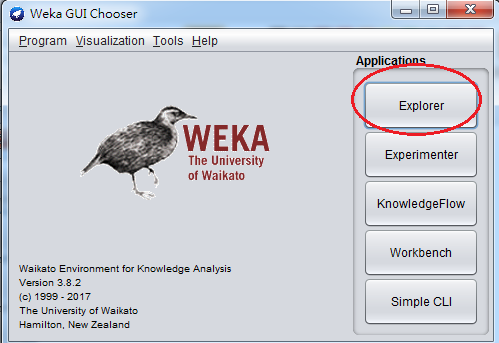
RunWeka.ini 記事本開啟

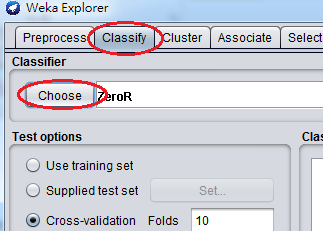
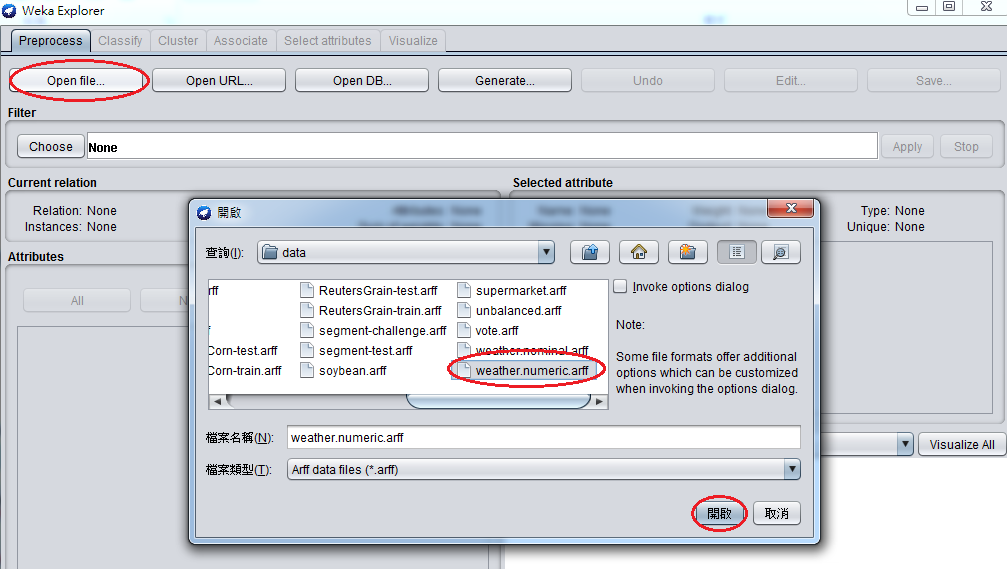
fileEncoding=Cp1252 →　fileEncoding=UTF-8

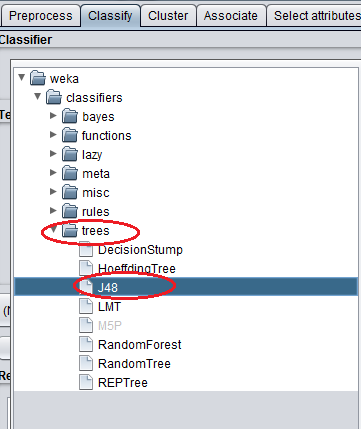
**5-2: 利用決策樹進行天氣探勘**

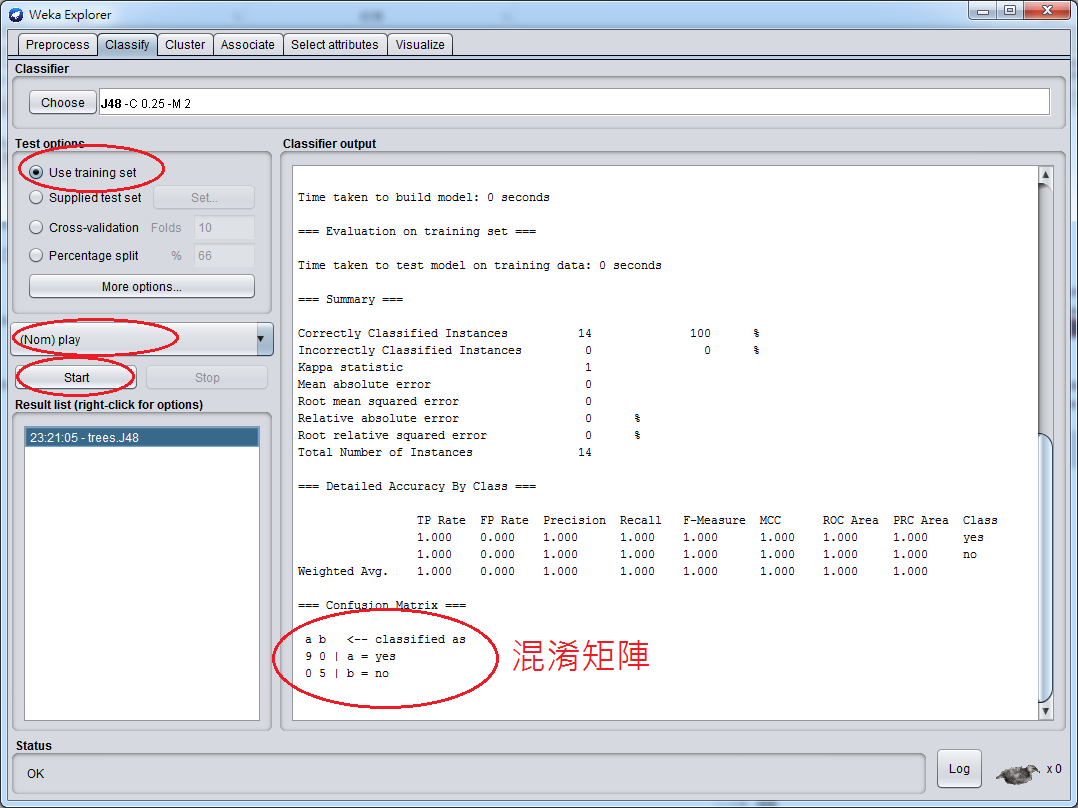
開啟範例檔

Explorer open file C:\Program Files\Weka-3-8\data weather.numeric.arff

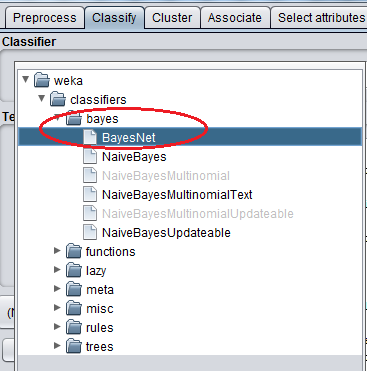


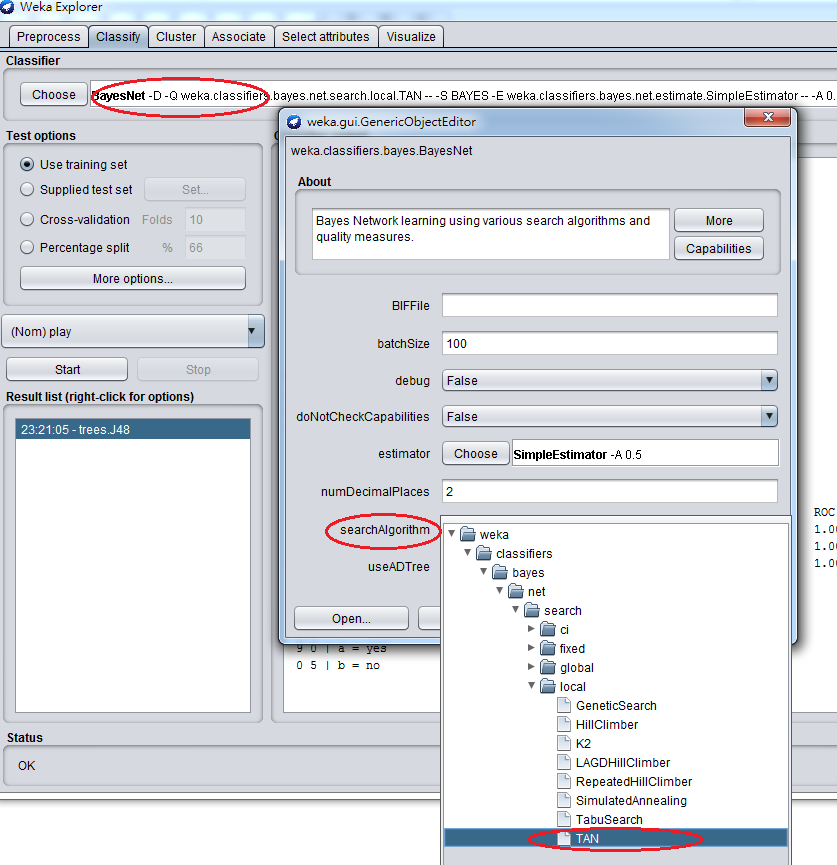




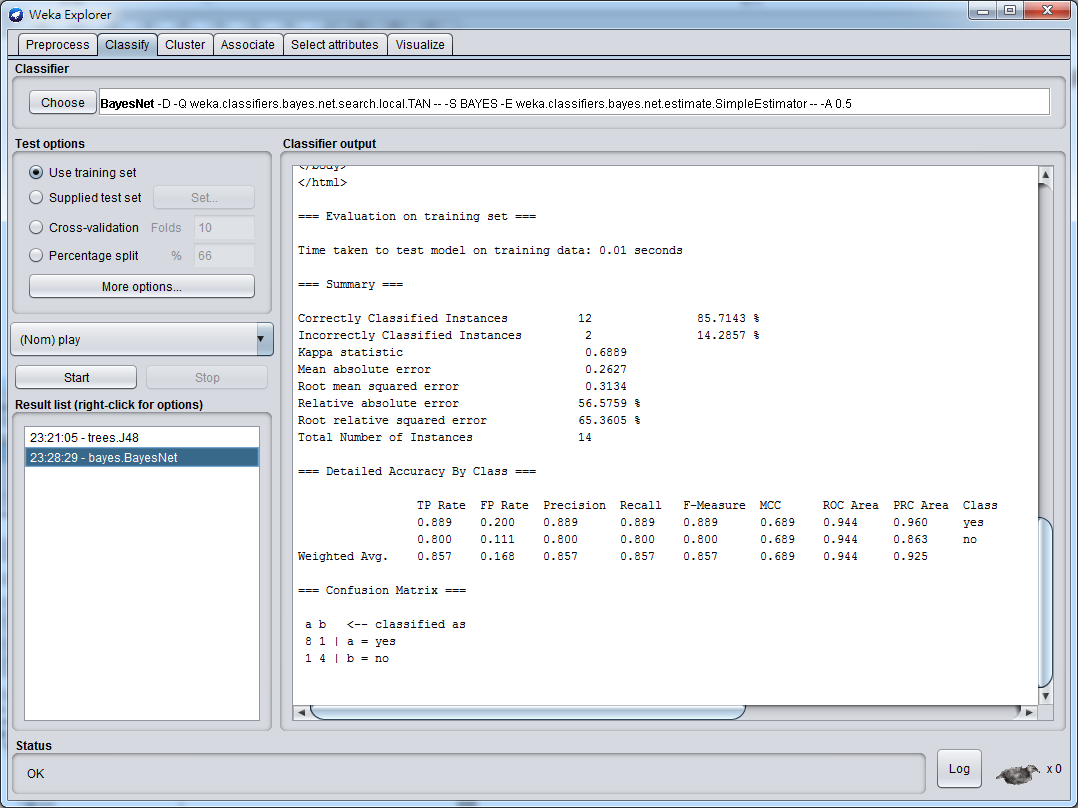


**5-3: 利用貝式網路進行天氣因素探勘**





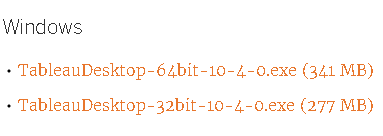
結果



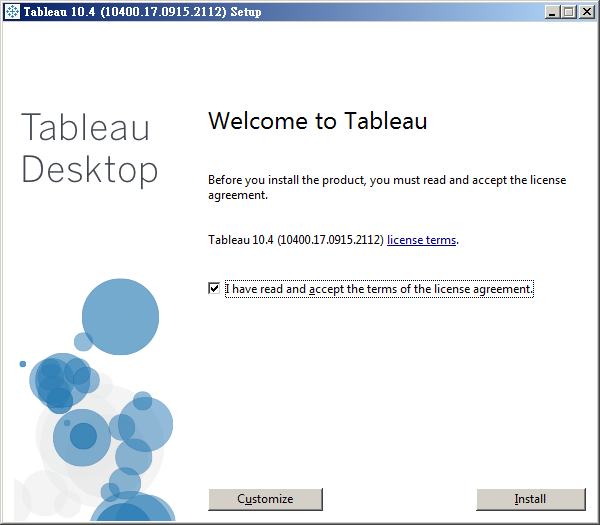
**作業六：Tableau - 數據視覺化與應用**

**6-1: Tableau的安裝過程**

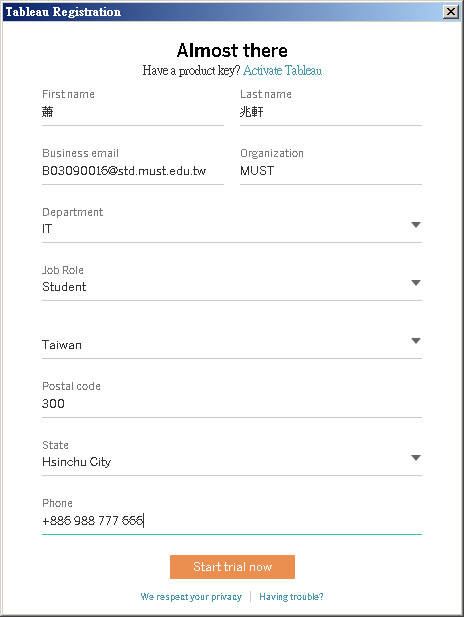
**搜尋、下載Tableau 10.4**



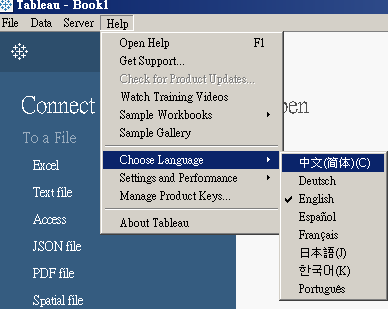
**勾選同意，安裝**



**註冊**

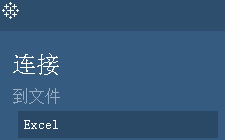


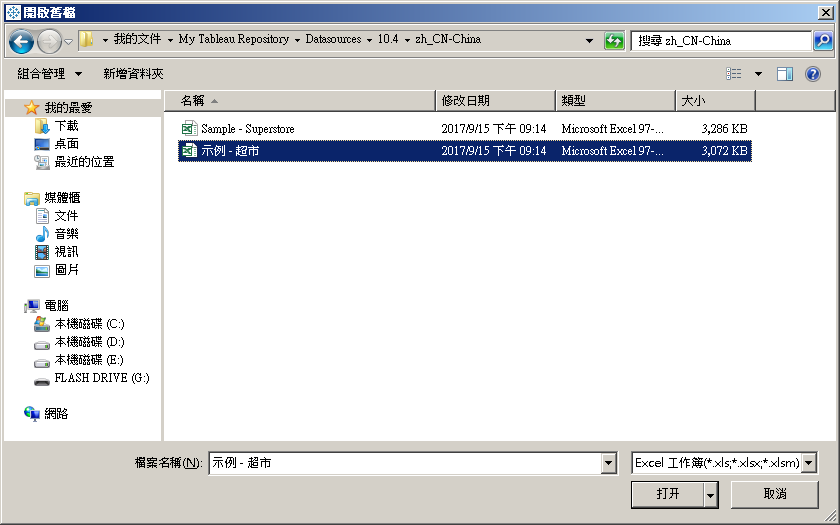
**改語言，然後重新啟動**



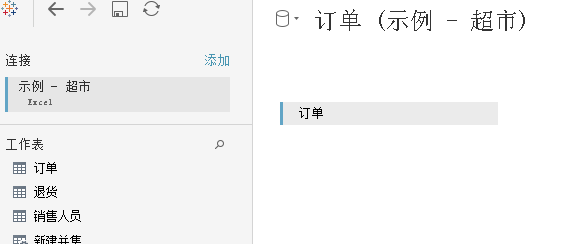
**6-2: 超級市場範例數據資訊化呈現**

**打開範例，連接，Excel，檔案預設在我的文件**





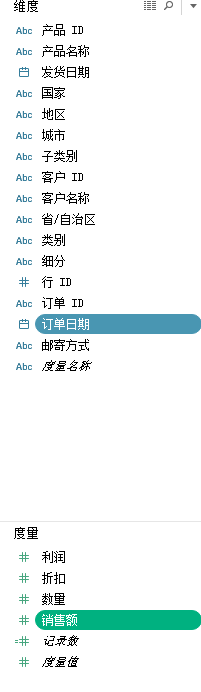
**選擇要的工作表拖曳至右方**



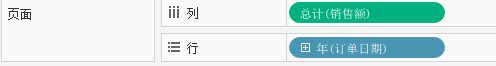
下方可轉換工作表、新增工作表



**左方用Ctrl作複選 右方選擇想做成的圖表**

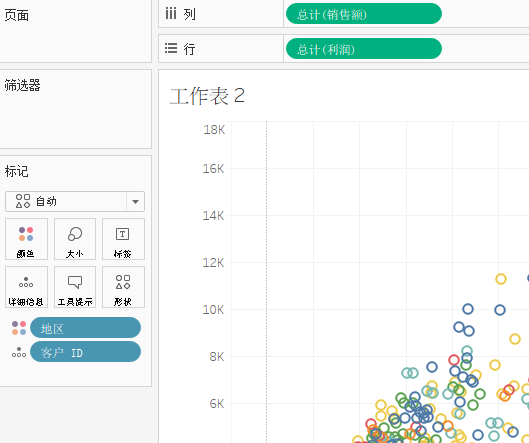


**可用拖曳改變行列**





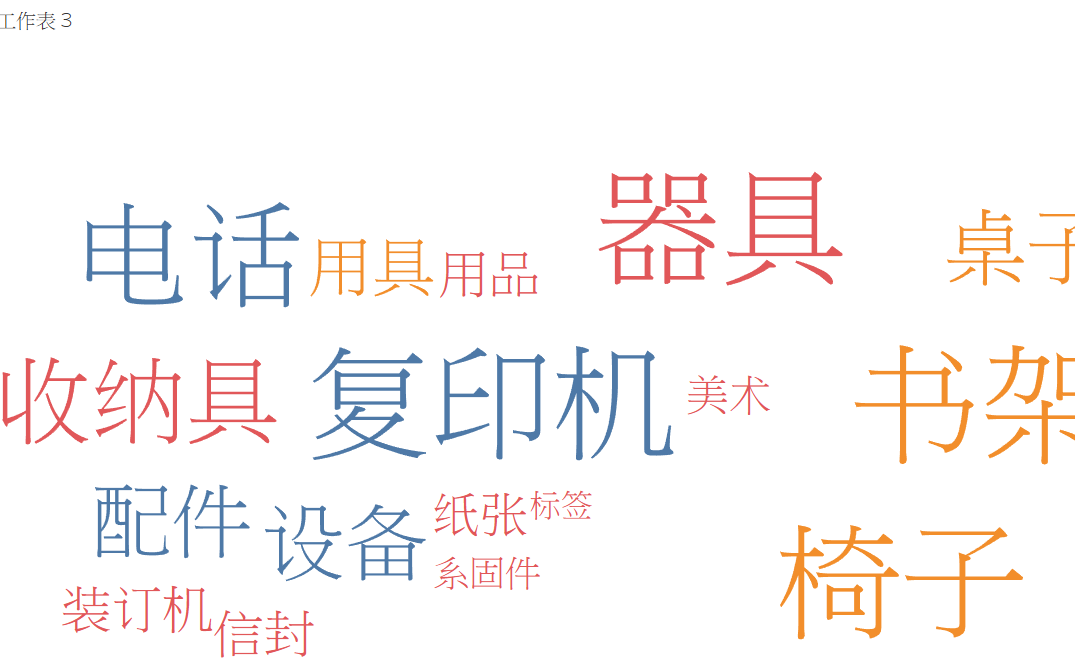
**或使用工具列直接交換行列**



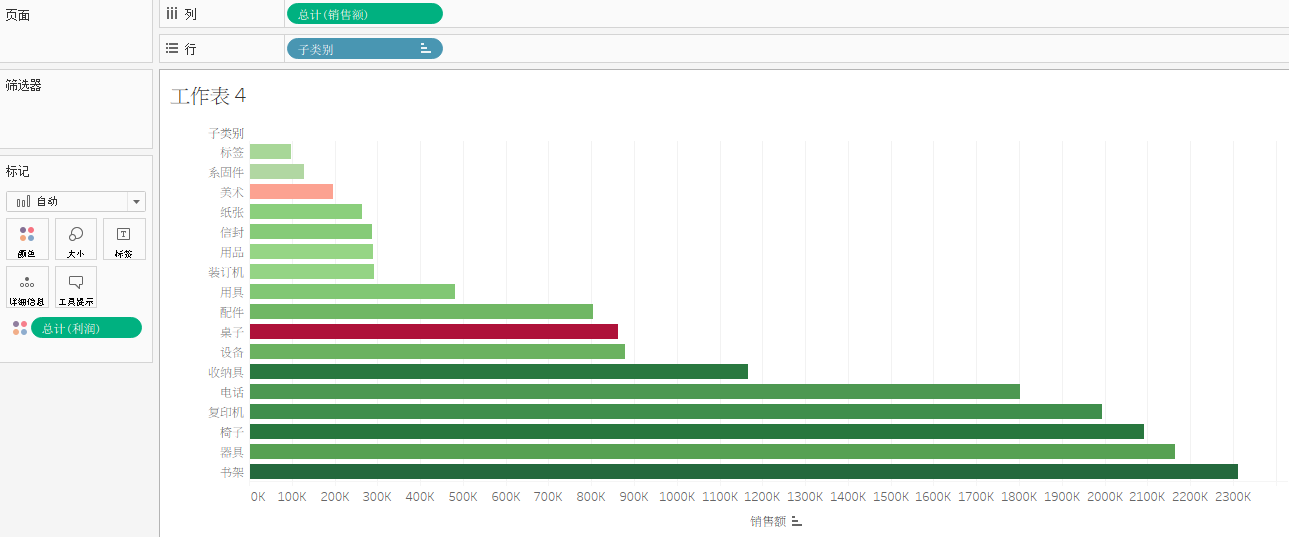
**利用雙擊或拖曳組合圖表，地區拖曳至顏色可改變呈現方式為顏色**

****

**將自動改成文本**

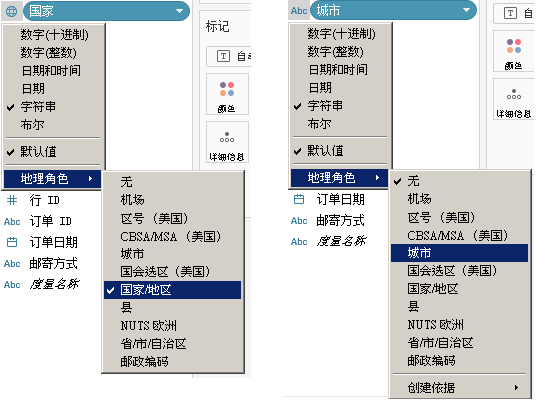
****

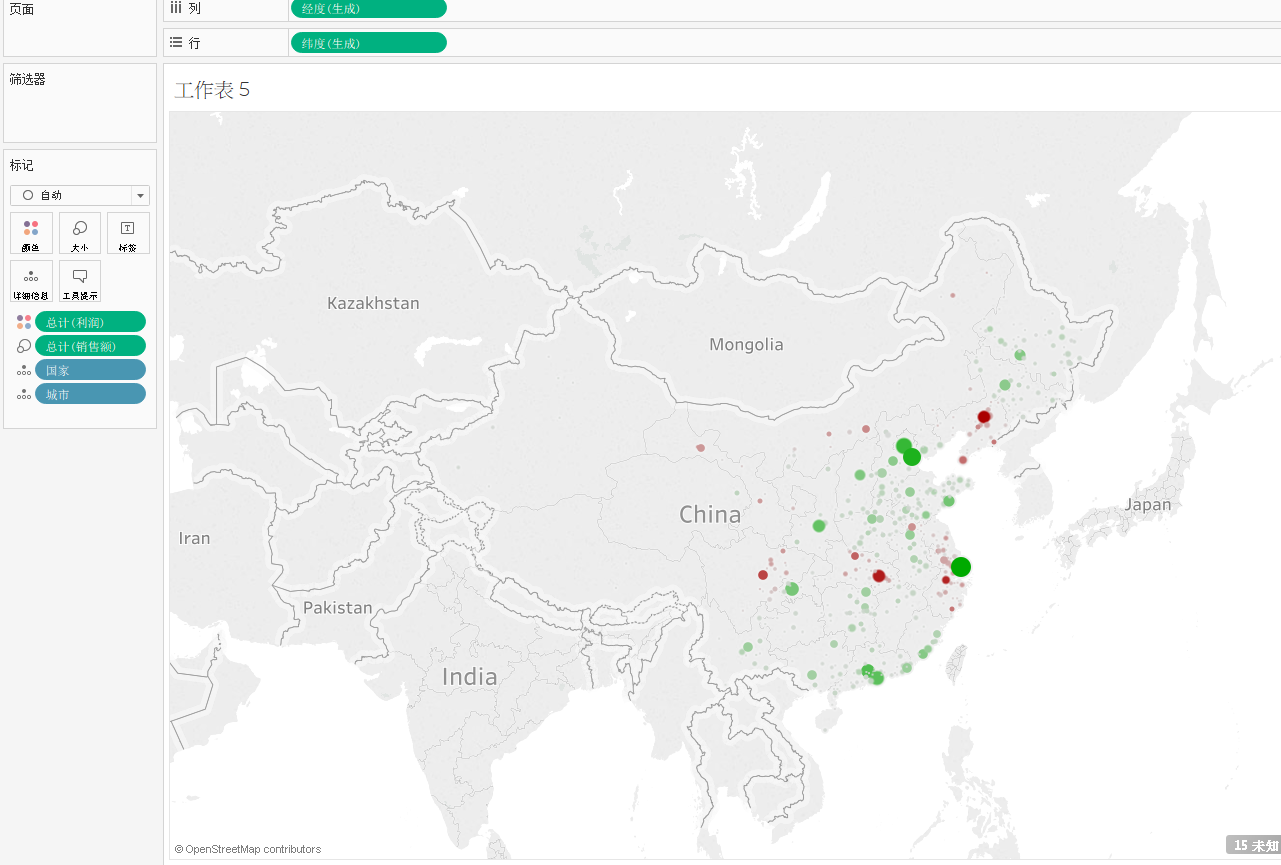
**工作表４，利用上方工具列作排序　遞增或遞減**

****

**6-3: 數據地圖化呈現**

**工作表５　地圖　先更改設定**

****

****

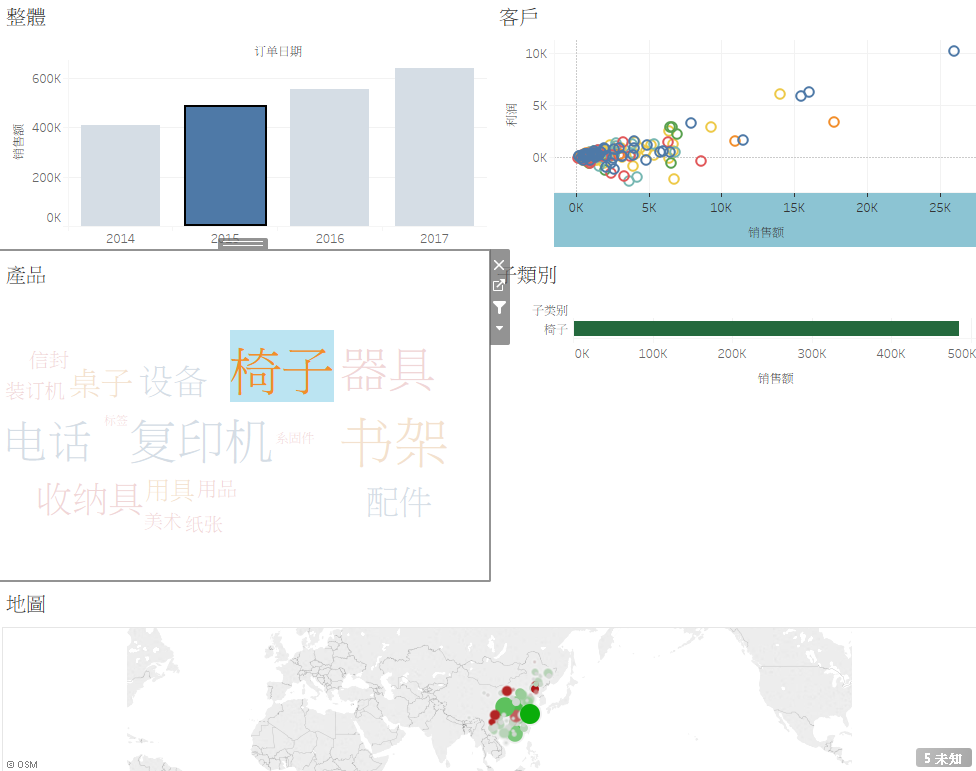
**6-4: Dashboard整合型呈現**

**新增儀表板**

****

****

**篩選器做連結其他圖表**

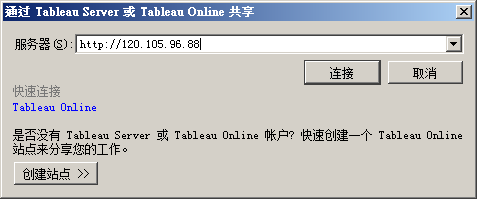
****

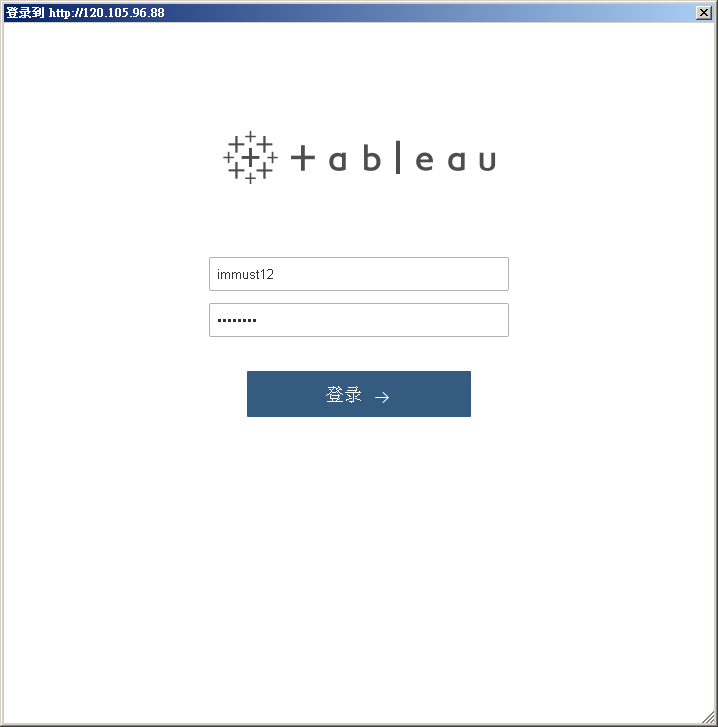
**成果**

**6-5: Tableau雲端發佈**

**上傳發布置Tableau Server　網址<http://120.105.96.88/>**

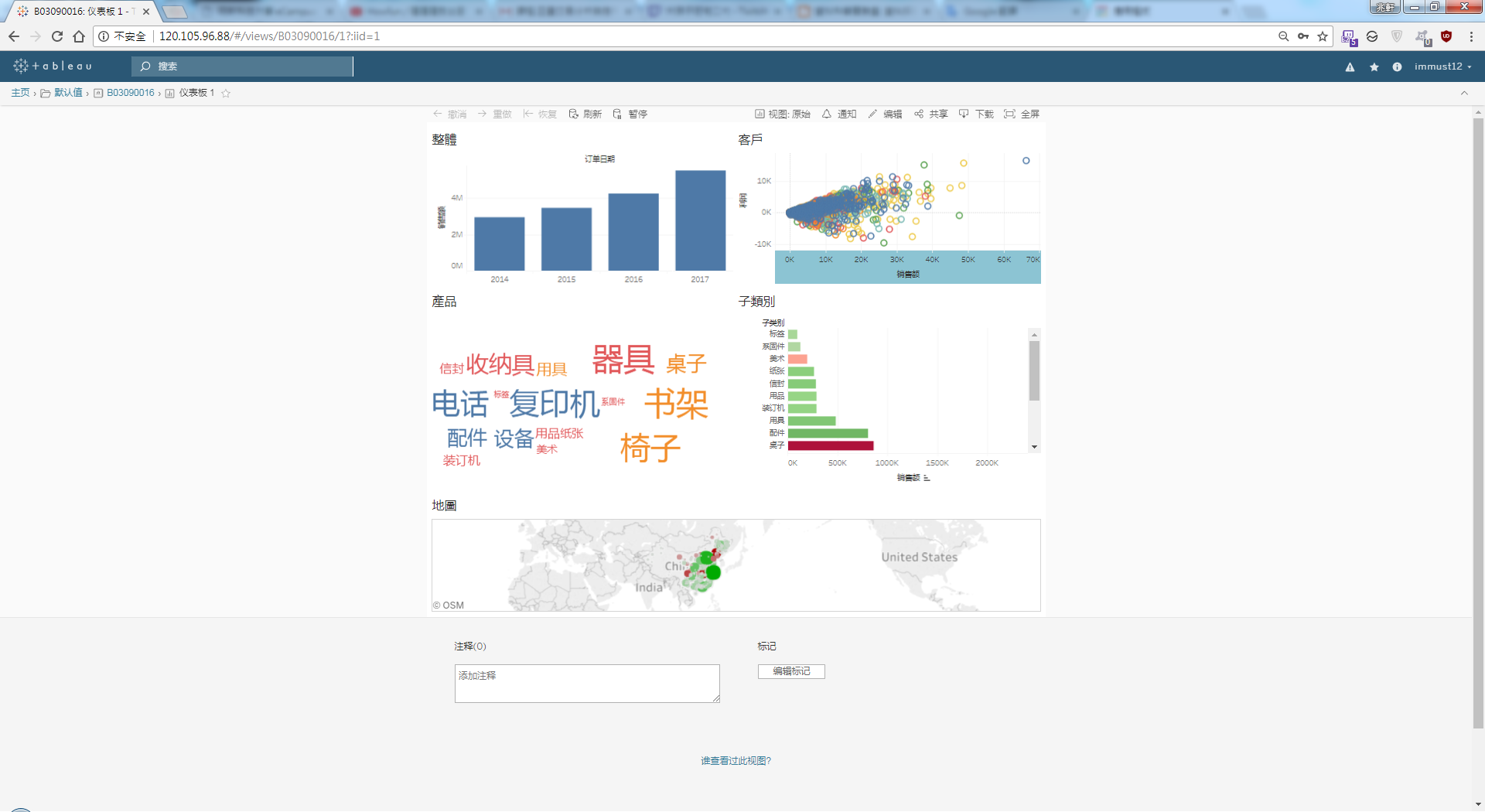
**帳密一樣**

**** 



****

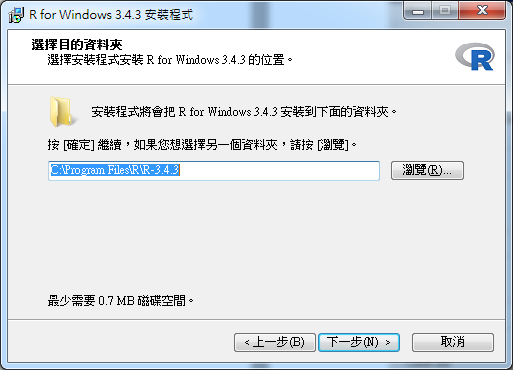
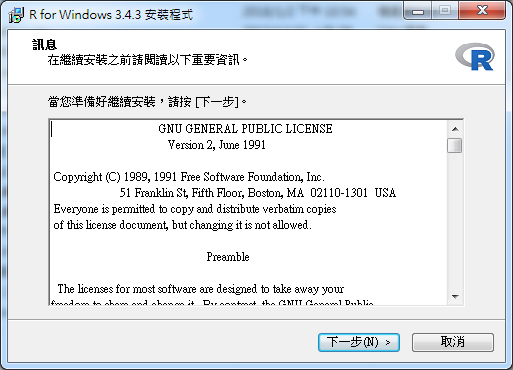


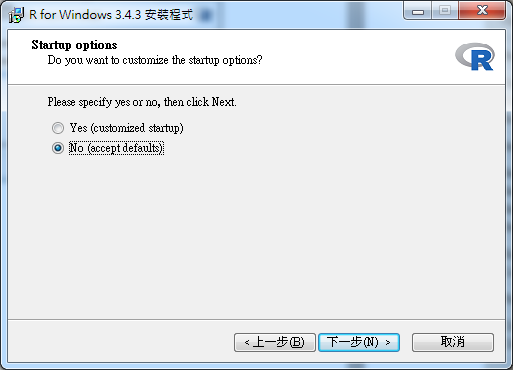
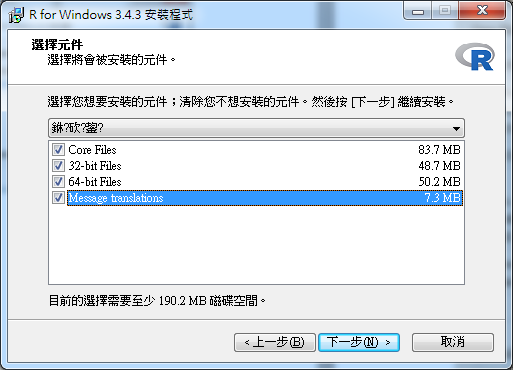


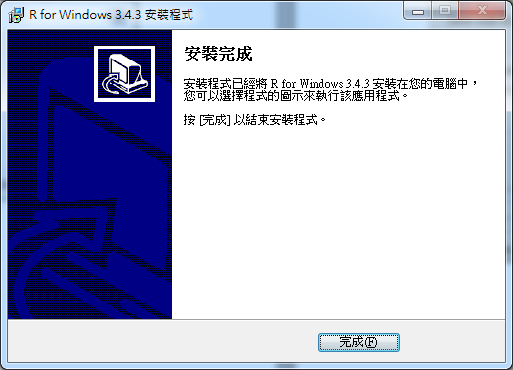
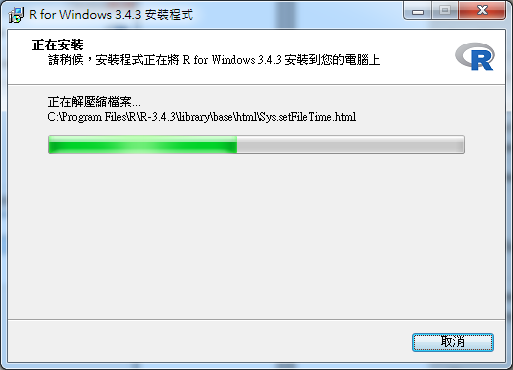
**作業七：R語言大數據實戰**

**7-1: R與R-Studio的安裝**

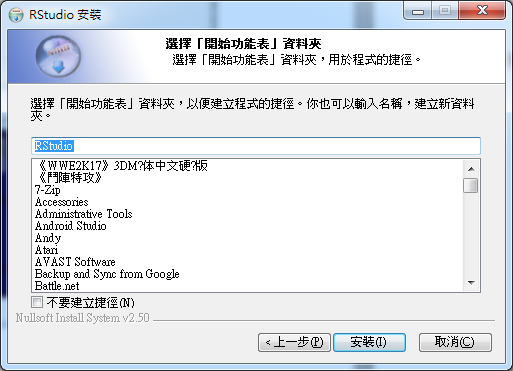
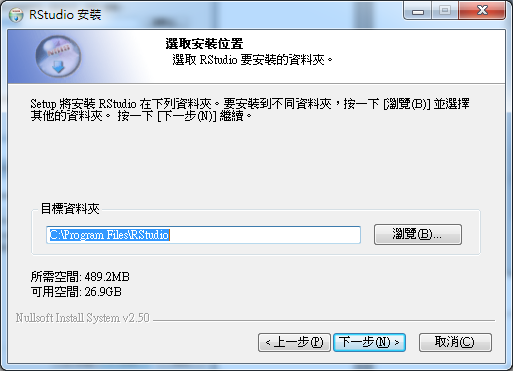
**R**

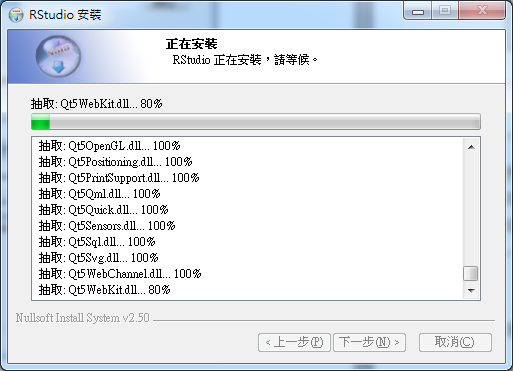
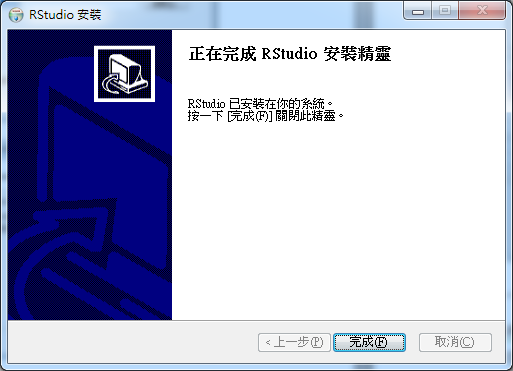




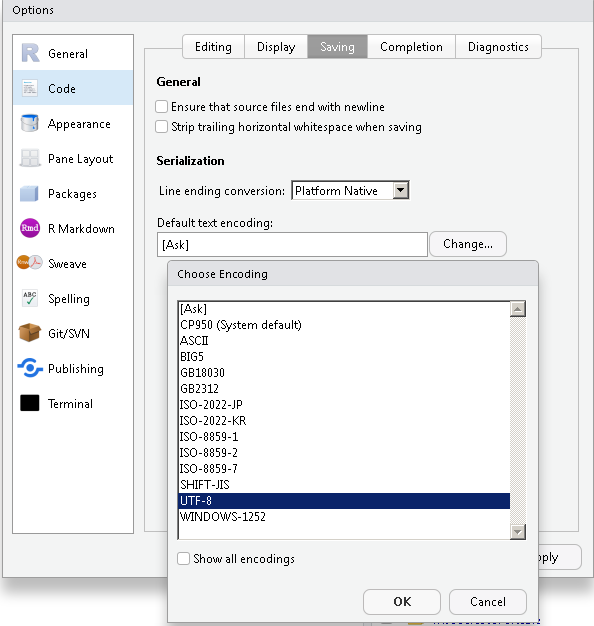


**R Studio**



環境設定



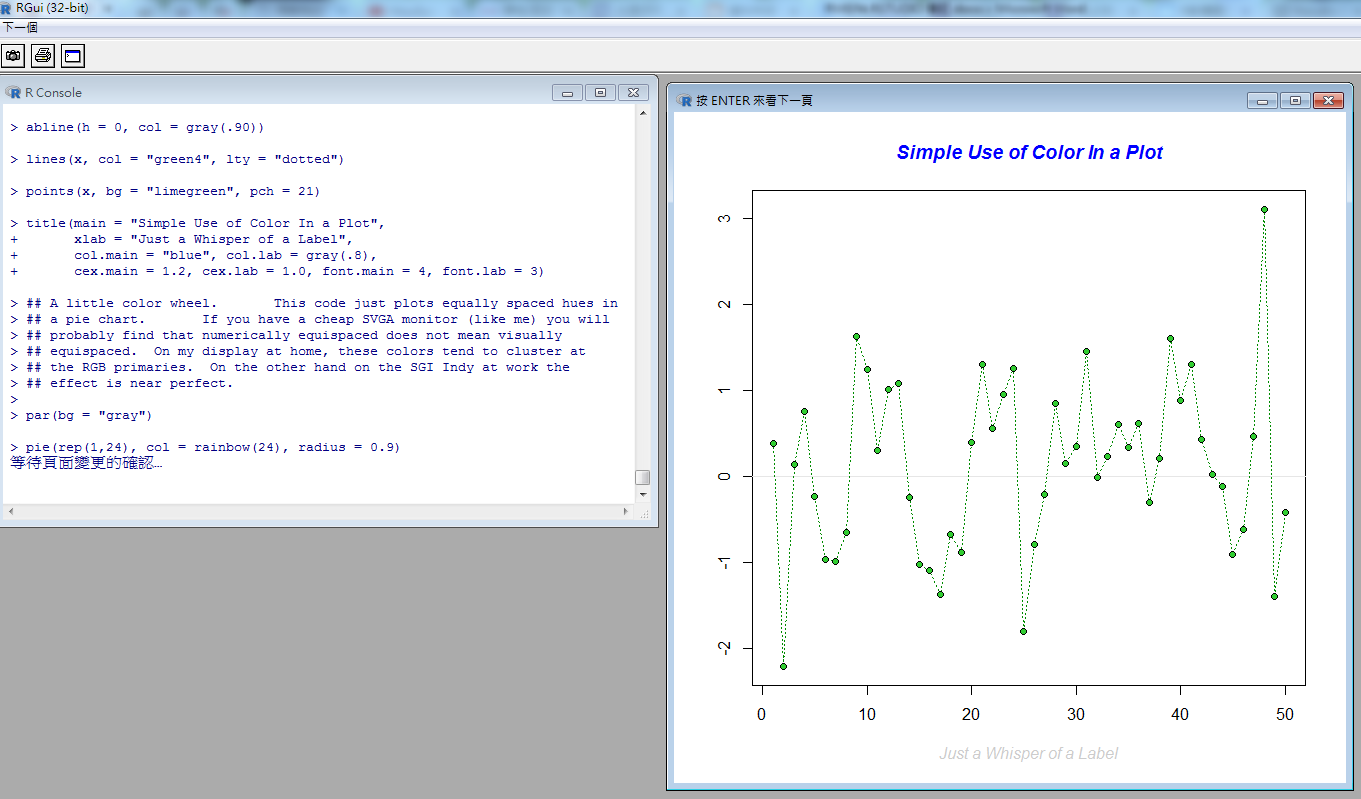
**7-2: R語言圖形Demo**

輸入

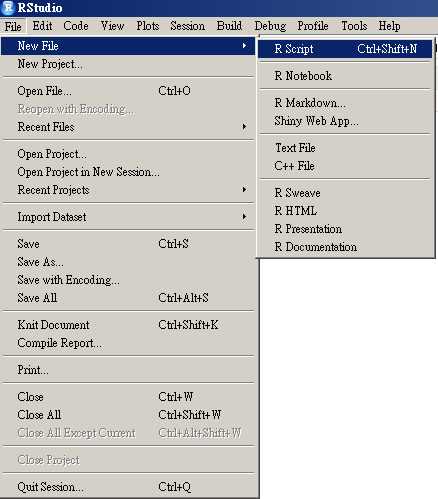
plot(runif(100),type=”l”)

demo(graphics)

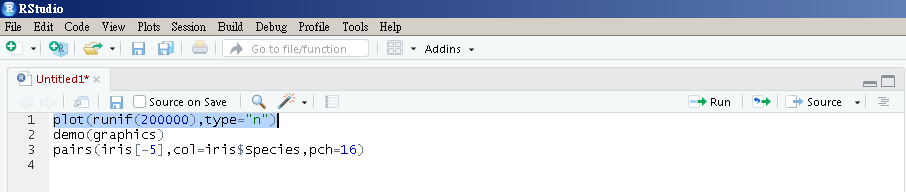
demo(persp)



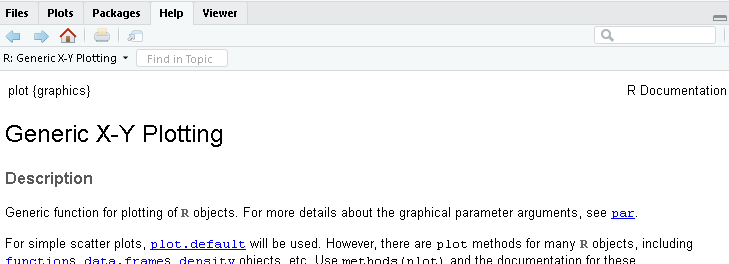
**7-3: IRIS圖形統計呈現**



反白要編譯的程式碼 點選Run編譯



反白指令按F1顯示幫助

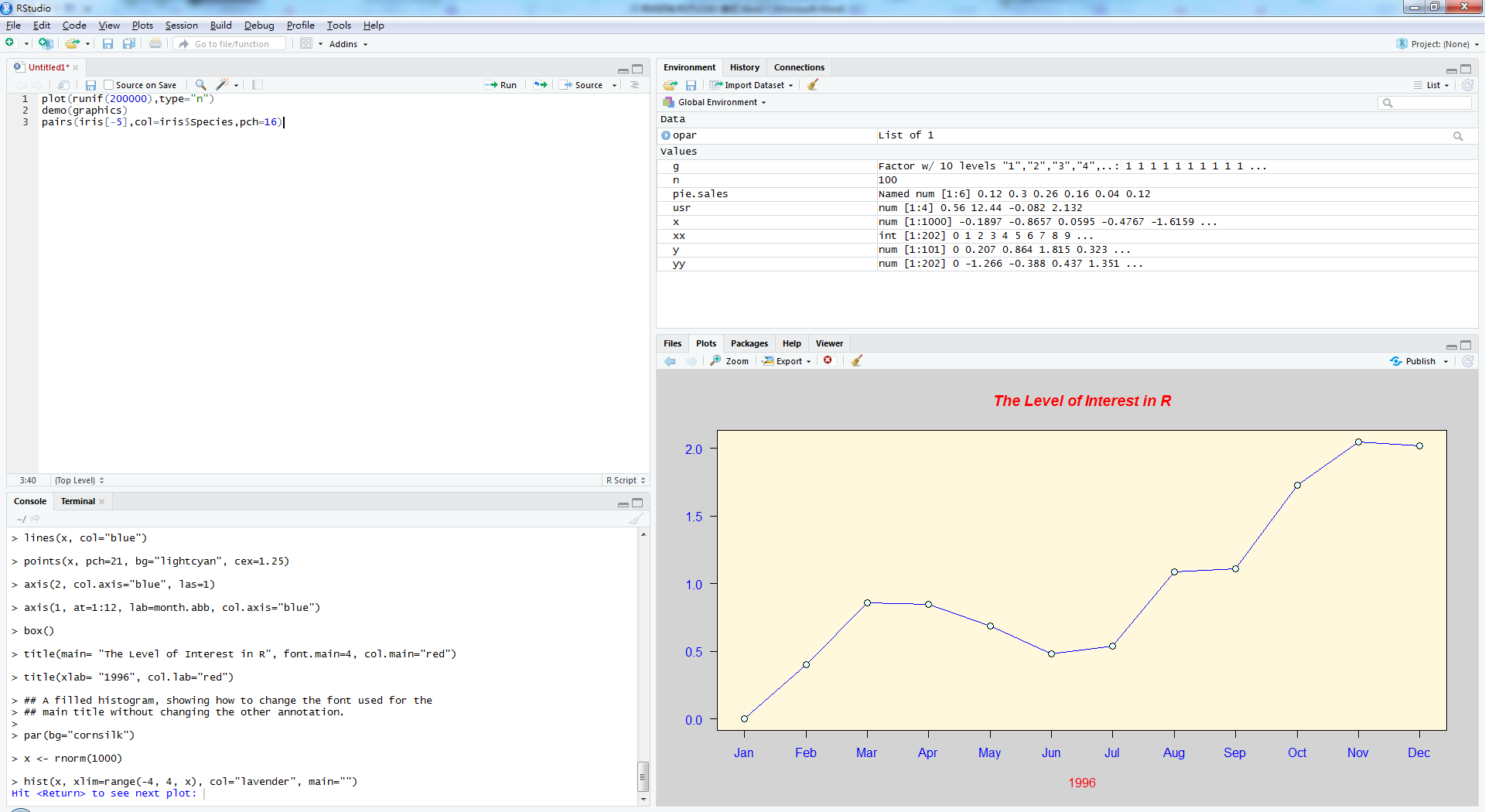


plot(runif(200000),type="n") n參數=不畫圖

demo(graphics)

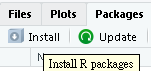
pairs(iris[-5],col=iris$Species,pch=16)

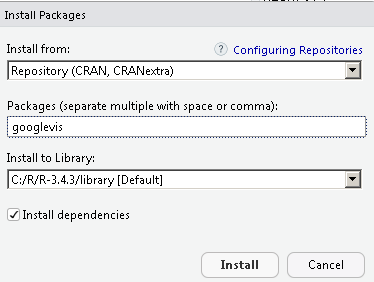
結果



**7-4: GoogleVis水果數據動態呈現**

安裝googlevis套件



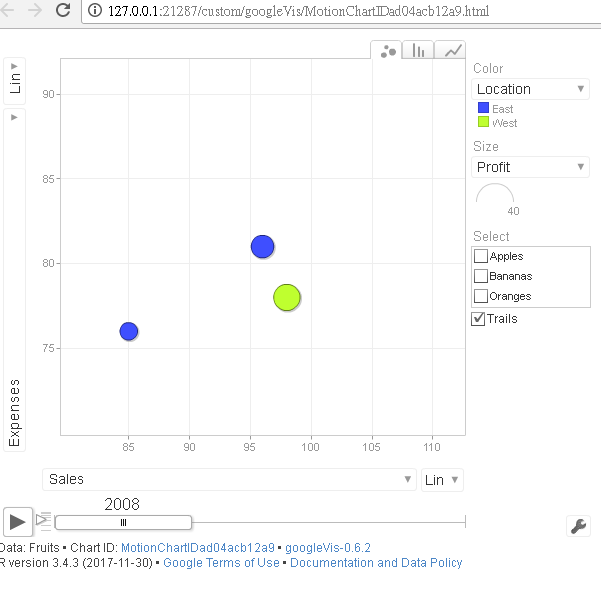


install.packages("googleVis")

library(googleVis)

M1 <- gvisMotionChart(Fruits, idvar="Fruit", timevar="Year")

plot(M1)



**作業八：SPSS數據統計分析**

**8-1: Kent State Univ. Chi-Square範例**

講義<http://120.105.96.69/Lesson%20SPSS-ANOVA.pdf>

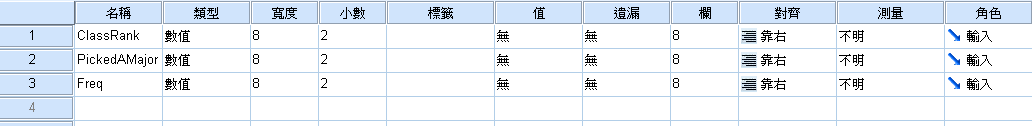
<https://libguides.library.kent.edu/SPSS/WeightCases>

範例檔案<http://120.105.96.69/hsbdataNew.sav>

切換至變數視圖

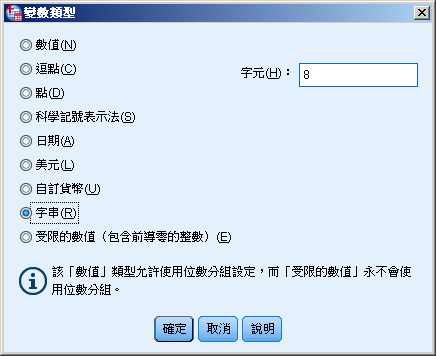


定義欄位名稱



點選類型 更改







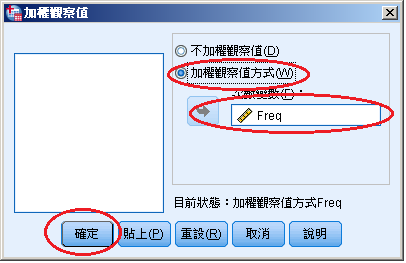
切換至資料視圖 KEY資料

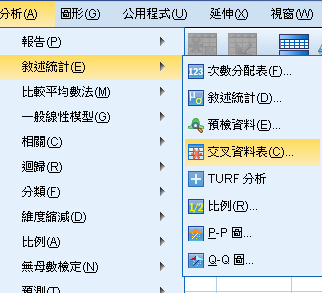


切換至變數視圖 將Freq小數改為0

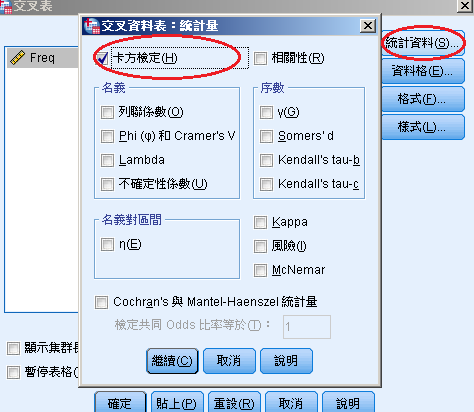




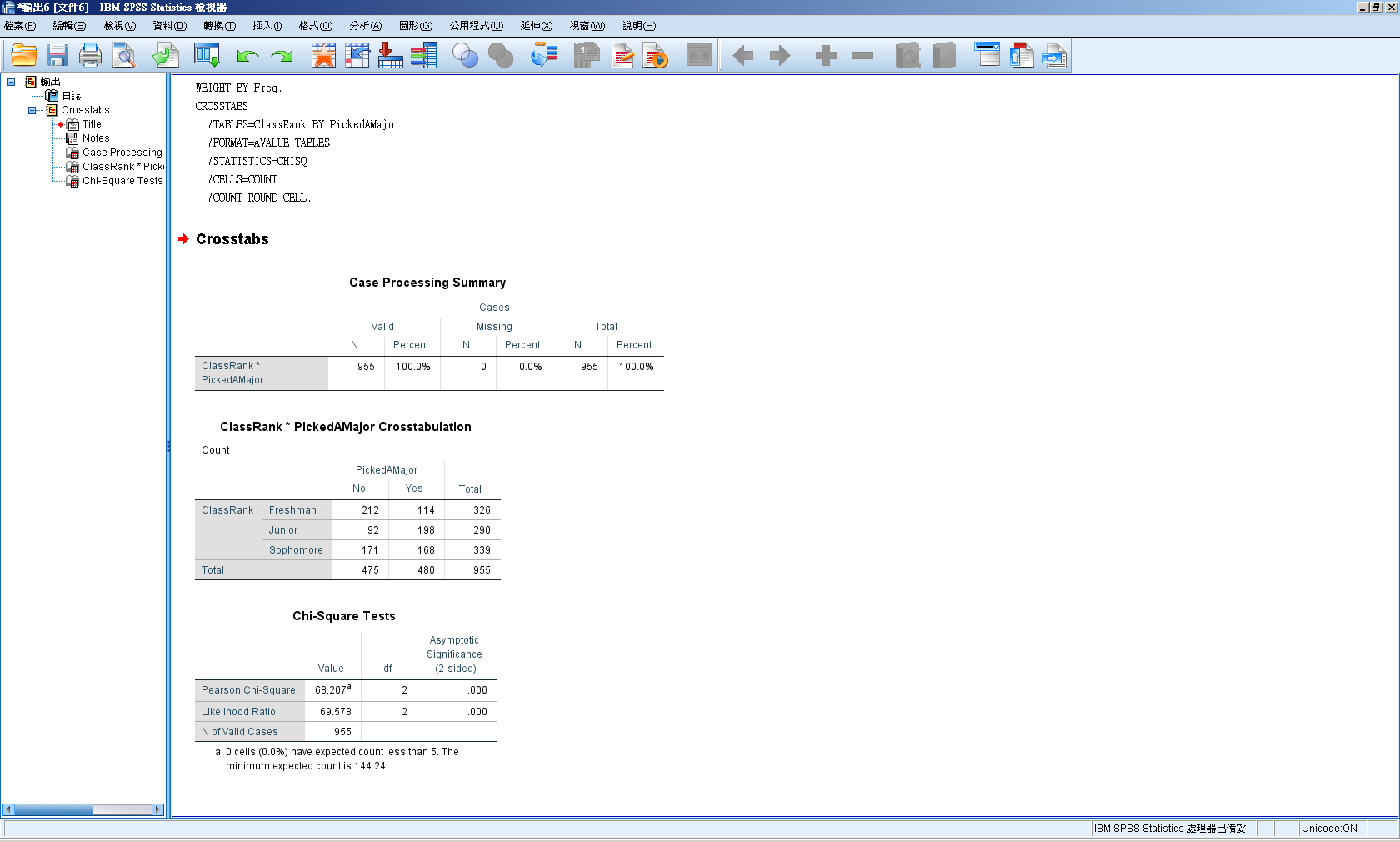








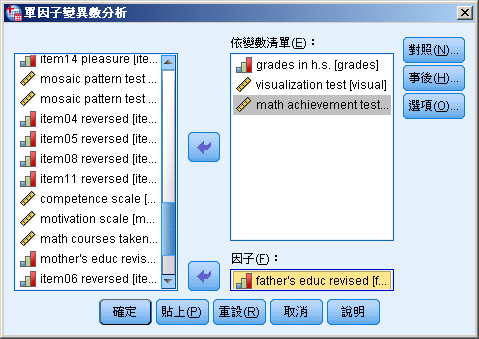
結果

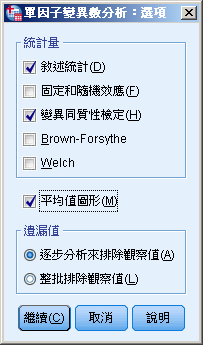


**8-2: Math Achievement One-Way ANOVA範例**

分析父親教育程度對數學成績的影響



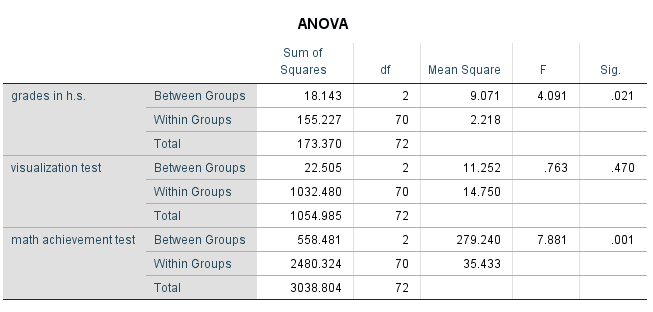




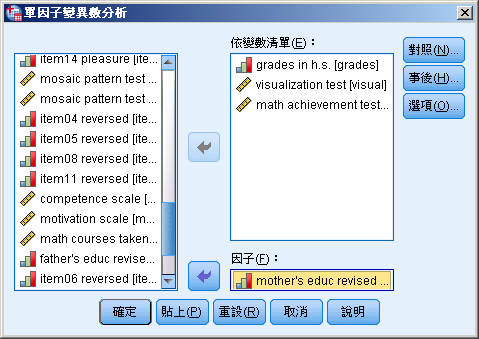
結果

看Sig . <0.05 為顯著

表示父親的教育程度對於學生的高中成績及數學成績有影響



分析母親教育程度對學生的數學成績影響 是否有差別



結果為

表示母親的教育程度對於學生的數學成績有影響

