巨量資料與統計分析

政治大學統計系余清祥

2016年9月28日

第三週：SQL和R軟體介紹

http://csyue.nccu.edu.tw

SQL介紹、使用方法

Agenda

- 認識SQL與資料庫
- Database Basic, Data Input
- SQL的基本指令
- 實例操作

CH1 認識SQL與資料庫

1.1 SQL簡介

- 結構化查詢語言(Structured Query Language)，簡稱SQL，為專門用於關聯式資料庫的一種查詢語言。
- 可用來定義資料庫結構、建立表格、指定欄位型態與長度；也能新增、異動或查詢資料。
- 統計分析軟體結合SQL的程式能力為必備技能。

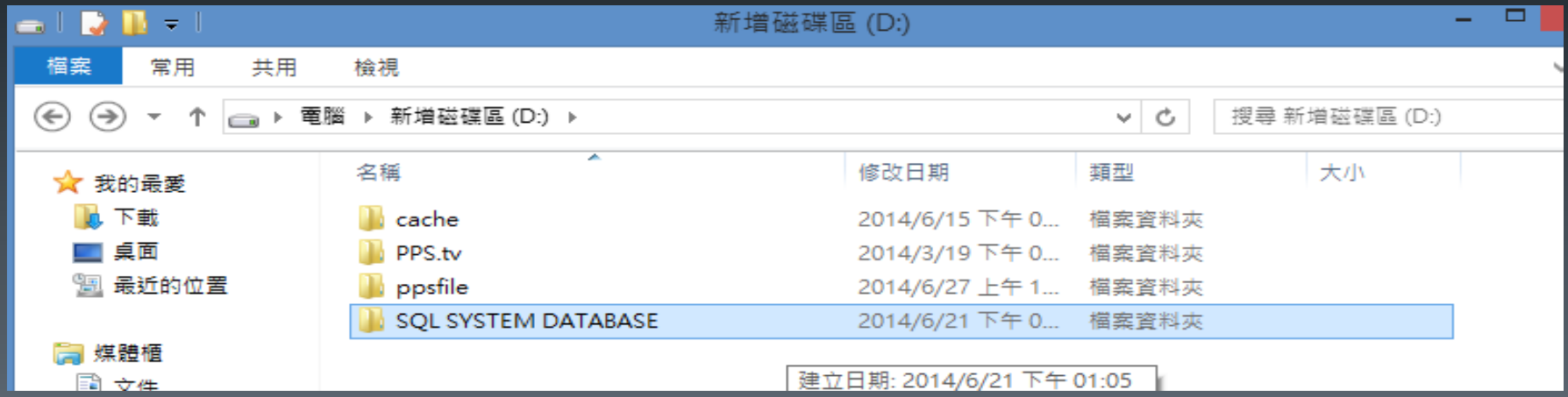
1.2 資料庫簡介

- 資料庫(Data Base)為一存放大量資料的地方，由各式各樣的資料匯集而成。
- 資料庫管理系統(Database Management System)提供使用者一個環境，使其能有效率且方便地對資料庫進行管理。
- 透過SQL語法，順利達到資料庫之間的溝通與管理資料。

CH2 Database Basic & Input

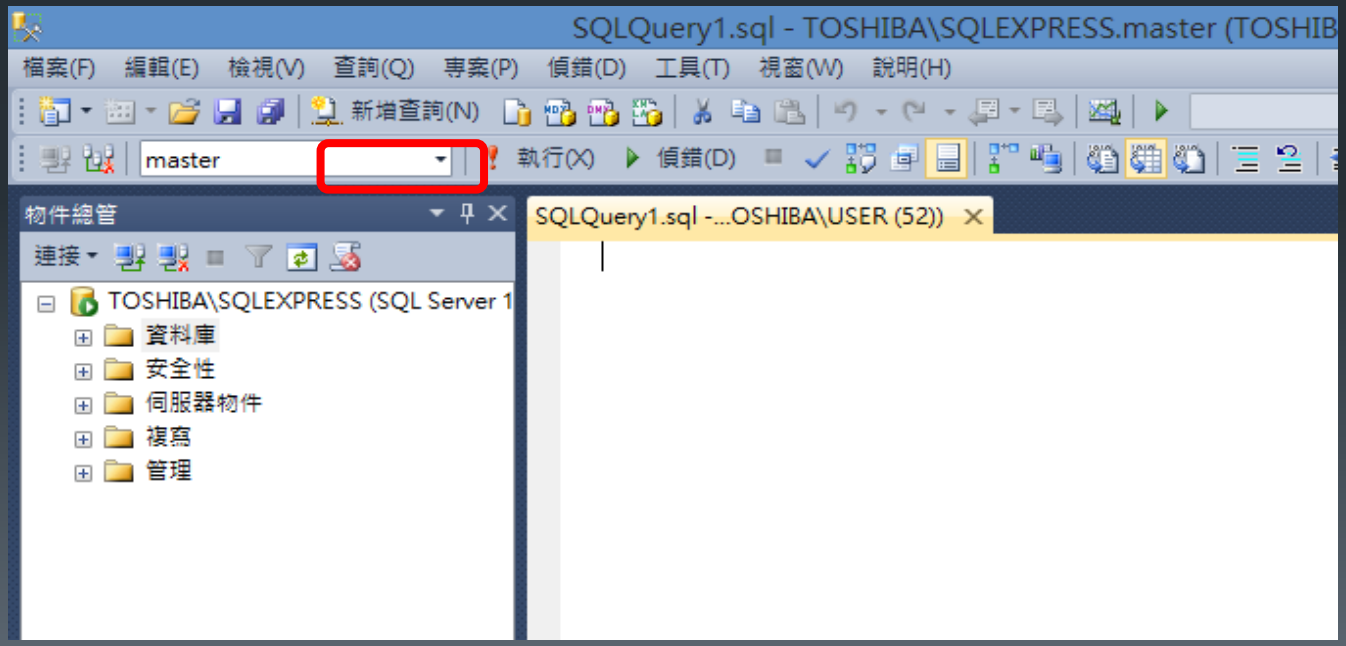
2.1 Database Basic

- 使用資料庫一開始，需在電腦內建立一個資料夾來作為存放SQL之Database的檔案路徑。
- 如下圖，先在D槽，新增一個資料夾SQL SYSTEM DATABASE。



2.1 Database Basic

- 開啟SQL Server Management Studio，並點選連結，接著點選新增查詢，就會出現可輸入程式碼的工作區，如下圖：



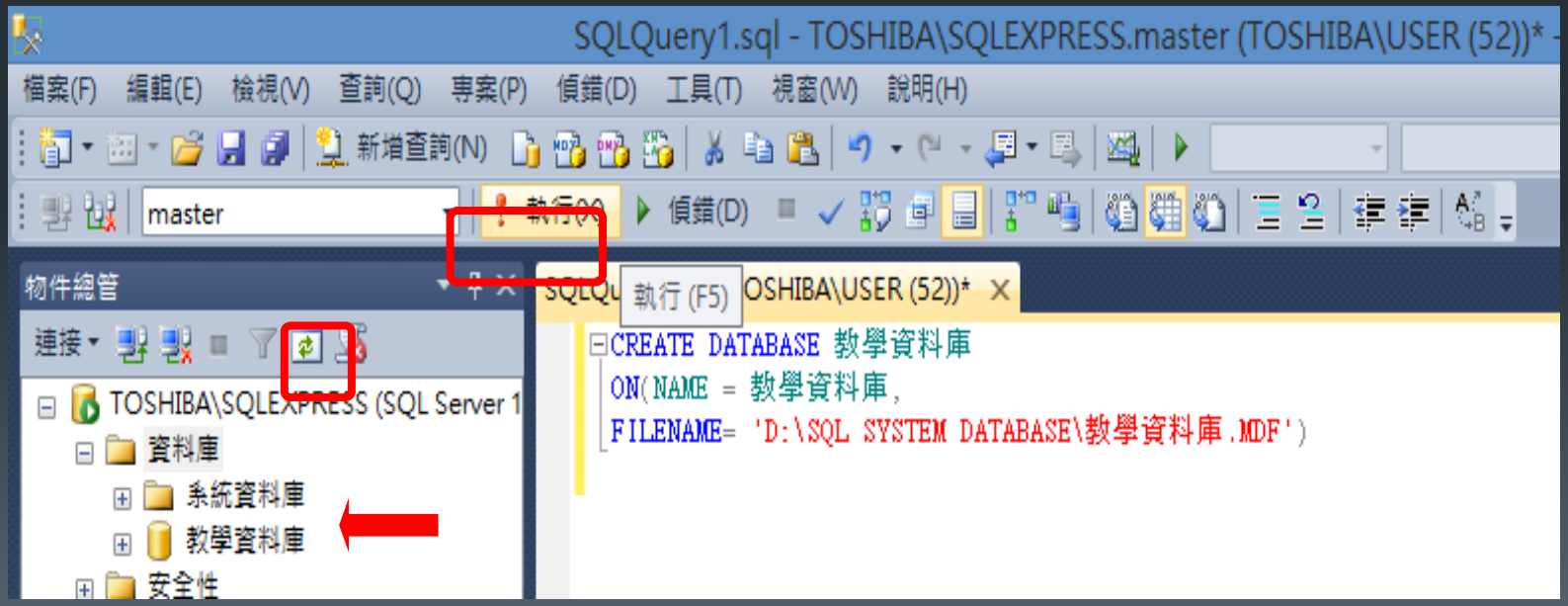
2.1 Database Basic

- 在工作區，輸入下列程式碼，並按執行，以建立資料庫。
- 程式碼：

```
CREATE DATABASE 教學資料庫
ON(NAME = 教學資料庫,
FILENAME= 'D:\SQL SYSTEM DATABASE\教學資料庫.MDF')
```

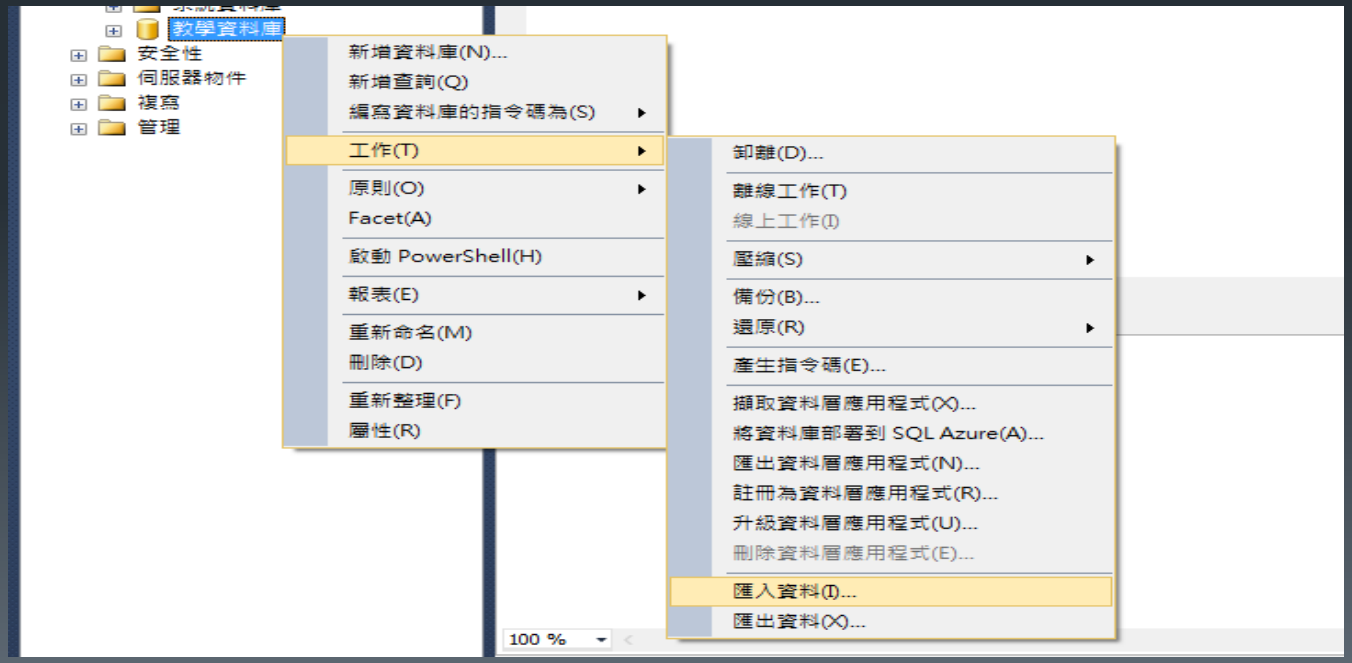
2.1 Database Basic

- 執行完成後，在左邊物件總管上方，先按重新整理，再點開資料庫，就會看到剛才所建立的教學資料庫。



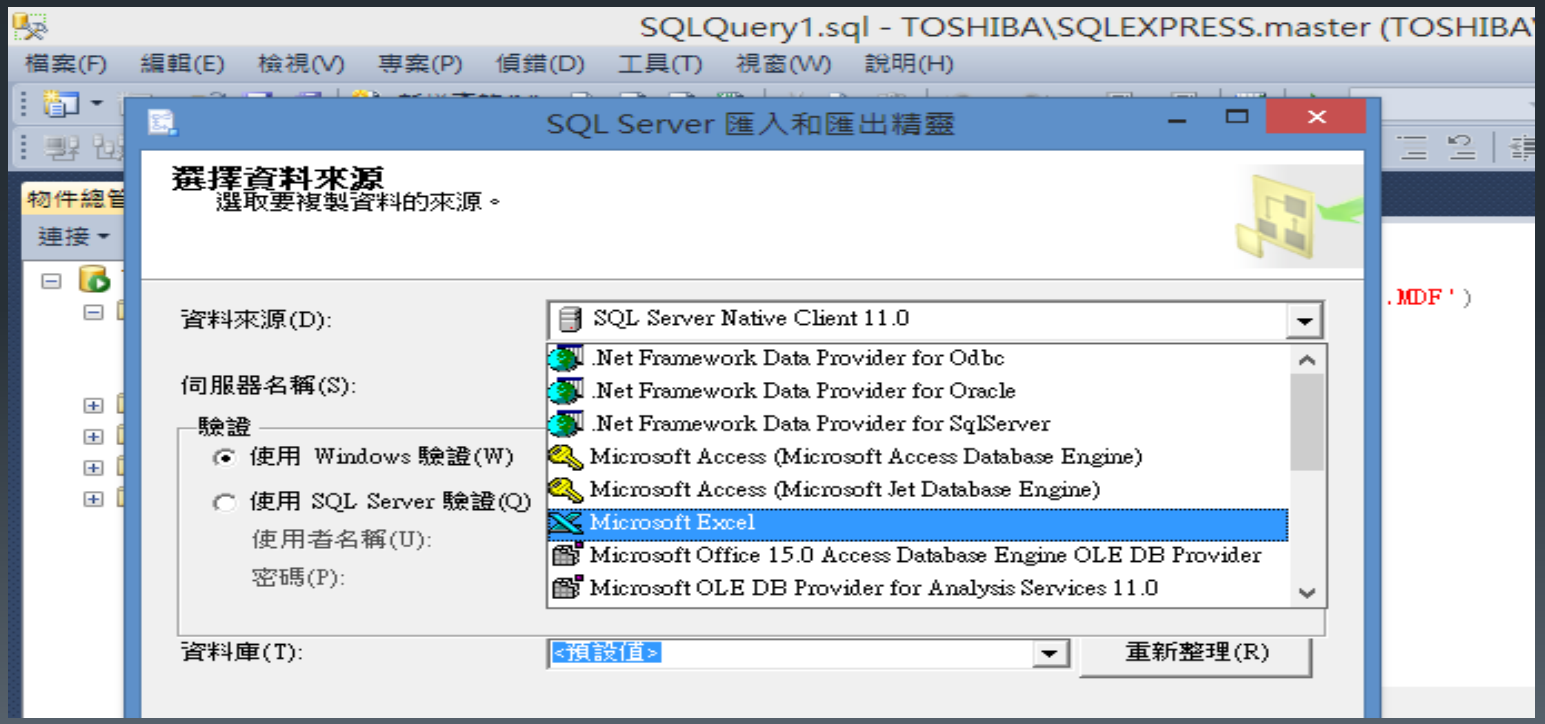
2.2 Data Input

- 首先，將Excel檔匯入資料庫成資料表。
- Step1：在欲加入資料表之資料庫按右鍵→工作→匯入資料。



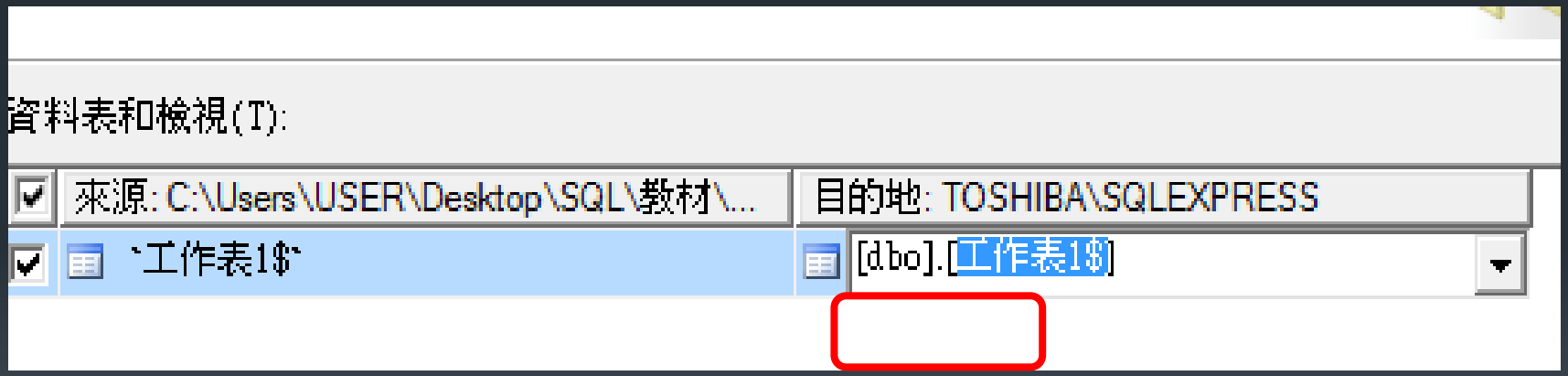
2.2 Data Input

- Step2：選取資料來源→設定檔案路徑和版本→瀏覽選取要匯入的檔案。



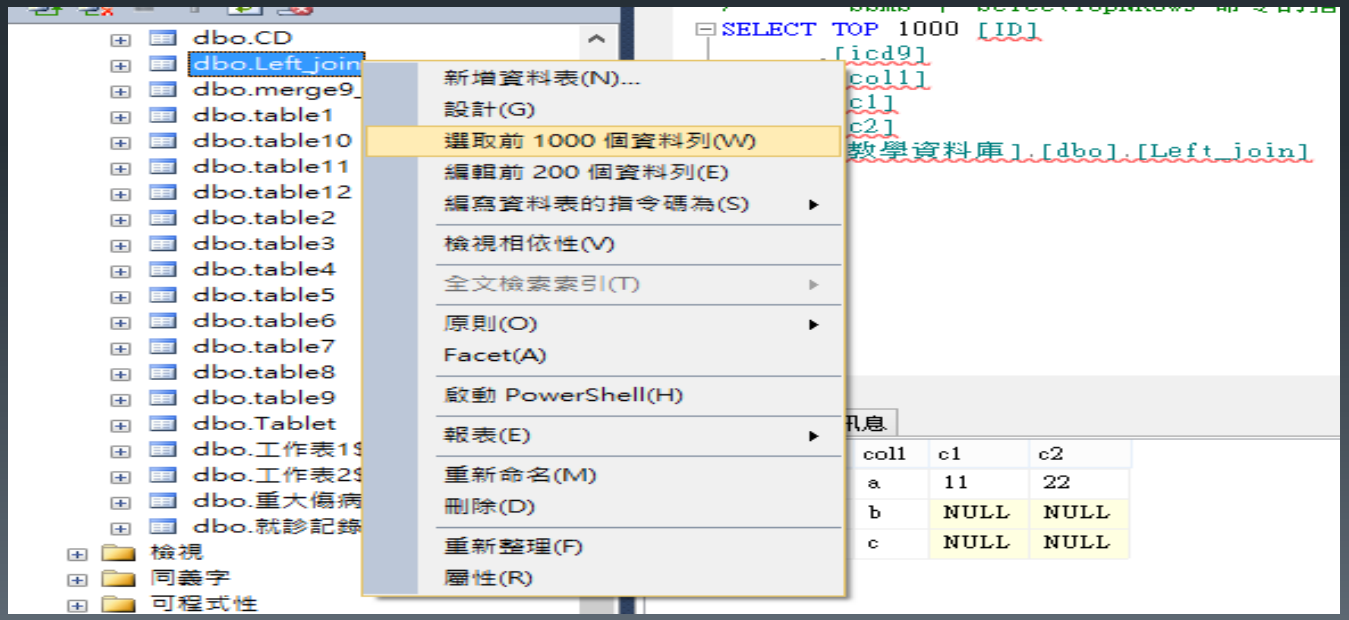
2.2 Data Input

- Step3：將框線處(工作表1$)改成欲在SQL內顯示之資料表名稱(table1)→完成。



2.2 Data Input

- Step4：執行完成後，在左邊物件總管上方，先按重新整理，再點開教學資料庫，就會看到剛才所匯入的資料表。在資料表按右鍵，點選選取前1000個資料列(W)，就可以看到資料輸入後的型態。



2.2 Data Input

- NOTE：
  若匯入EXCEL檔有問題，請至以下網站下載：
  2007 Office system 驅動程式：資料連線元件https://www.microsoft.com/zh-tw/download/confirmation.aspx?id=23734

2.2 Data Input

- 練習：
  在Excel輸入以下五個資料表，並匯入SQL Server。

A：

icd9	c1	c2
101	11	22

B1：

ID	SEX	AGE	Disease
A	F	1	A5
B	F	2	B6
C	M	3	C7

2.2 Data Input

B2 :

ID	SEX	AGE	Disease
A	F	1	A4
D	M	2	D5
E	M	5	E4

B3 :

ID	SEX	AGE	Disease
C	M	3	C7
F	F	4	F2
G	M	6	G3

CD :

ID	icd9	col1
1	101	a
2	102	b
3	103	c

CH3 SQL的基本指令

3.1 Learning Concept

- SQL的資料欄位，主要分為數值欄位、字串欄位兩大類。

- 數值欄位：
  int：整數，範圍從 -2^31~ 2^31
  bigint：整數，範圍從 -2^63 ~ 2^63 -1
  float：近似小數資料的資料類型，
  範圍從 -1.79E+308~1.79E+308
  real：近似小數資料的資料類型，
  範圍從 -3.04E+38~3.04E+38

3.1 Learning Concept

- 字串欄位：
  char：固定大小浪費空間，所需計算時間少，只能儲存英文字元。
  varchar：不固定長度，必須要花費較多的CPU計算時間，只能儲存英文字元。
  nchar、nvarchar：與前兩者的差異，可儲存其他字元(中文)。
- NOTE：數值欄位常用float，字串欄位則是nvarchar。若欄位名稱是由中文與數字組成，則在指令中需加上[ ] (ex.[資料行0])。

3.2 Basic Command

- 3.2.1 SQL基本符號介紹：
- *：ALL
- ''：中間輸入特定值，主要是用於加入字串
- --：後面可輸入註解，不會被程式語言讀入
- /* */：中間可輸入註解，不會被程式語言讀入

3.2 Basic Command

- 3.2.2 SQL基本語法：
- SELECT 欄位名稱
  INTO 表名
  FROM 表名
  WHERE 條件敘述
  GROUP BY 欄位名稱
  ORDER BY 欄位名稱

3.2 Basic Command

- EX1：將CD表中，col1欄位是a的所有欄位，挑出並存到新的資料表CD1中。
- EX2：將CD表中，欄位順序調整，並存到新的資料表CD2中。
- 有問題都可以互相討論噢！

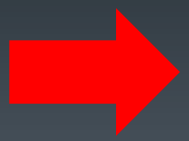
3.2 Basic Command

- <Ans1> :

```
SELECT *          /* 挑出所有欄位的資料 */
INTO CD1          /* 存到新的資料表CD1中
*/FROM CD           /* 從CD表抓資料 */
WHERE col1 = 'a'   /* 選取資料的條件為：
col1欄位是a */
```

結果 訊息

	ID	icd9	col1
1	1	101	a
2	2	102	b
3	3	103	c



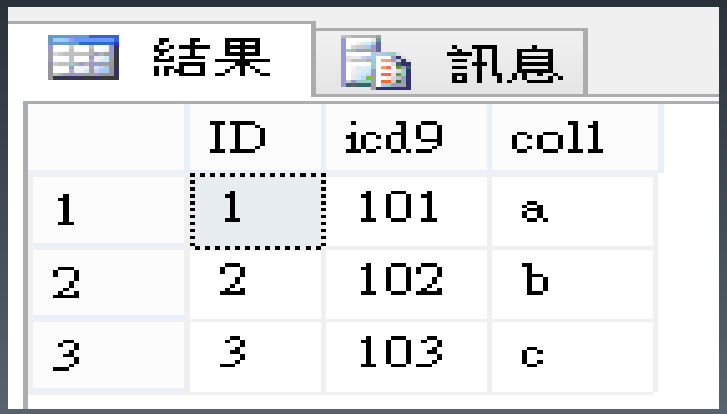
結果 訊息

	ID	icd9	col1
1	1	101	a

3.2 Basic Command

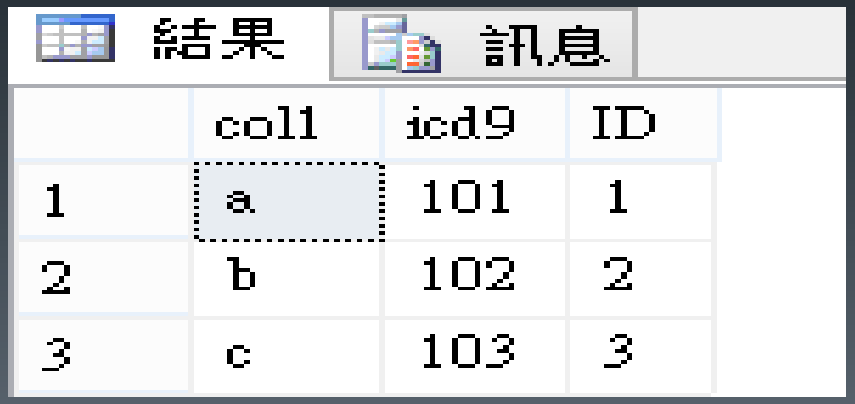
- <Ans2> :

```
select col1,icd9,ID    /* 依序挑出col1,icd9,ID欄
位的資料 */
into CD2               /* 存到新的資料表CD1中 */
from CD                /* 從CD表抓資料 */
```



結果 訊息

	ID	icd9	col1
1	1	101	a
2	2	102	b
3	3	103	c



結果 訊息

	col1	icd9	ID
1	a	101	1
2	b	102	2
3	c	103	3

3.2 Basic Command

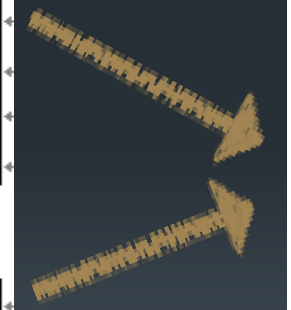
- 3.2.3 資料表合併：將B1與B2合併成B

B1：

ID	SEX	AGE	Disease
A	F	1	A5
B	F	2	B6
C	M	3	C7

B2:

ID	SEX	AGE	Disease
A	F	1	A4
D	M	2	D5
E	M	5	E4



B：

ID	SEX	AGE	Disease
A	F	1	A5
B	F	2	B6
C	M	3	C7
A	F	1	A4
D	M	2	D5
E	M	5	E4

3.2 Basic Command

- Code：

```
SELECT *          /* 挑出所有欄位的資料 */
INTO B            /* 存到新的資料表B */
FROM B1           /* 從B1表抓資料 */
UNION ALL
SELECT *
FROM B2           /* 再從B2表抓資料 */
```

3.2 Basic Command

- EX3：將B1與B3合併成B4

B1：

ID	SEX	AGE	Disease
A	F	1	A5
B	F	2	B6
C	M	3	C7

B3：

ID	SEX	AGE	Disease
C	M	3	C7
F	F	4	F2
G	M	6	G3

B4：

ID	SEX	AGE	Disease
A	F	1	A5
B	F	2	B6
C	M	3	C7
F	F	4	F2
G	M	6	G3

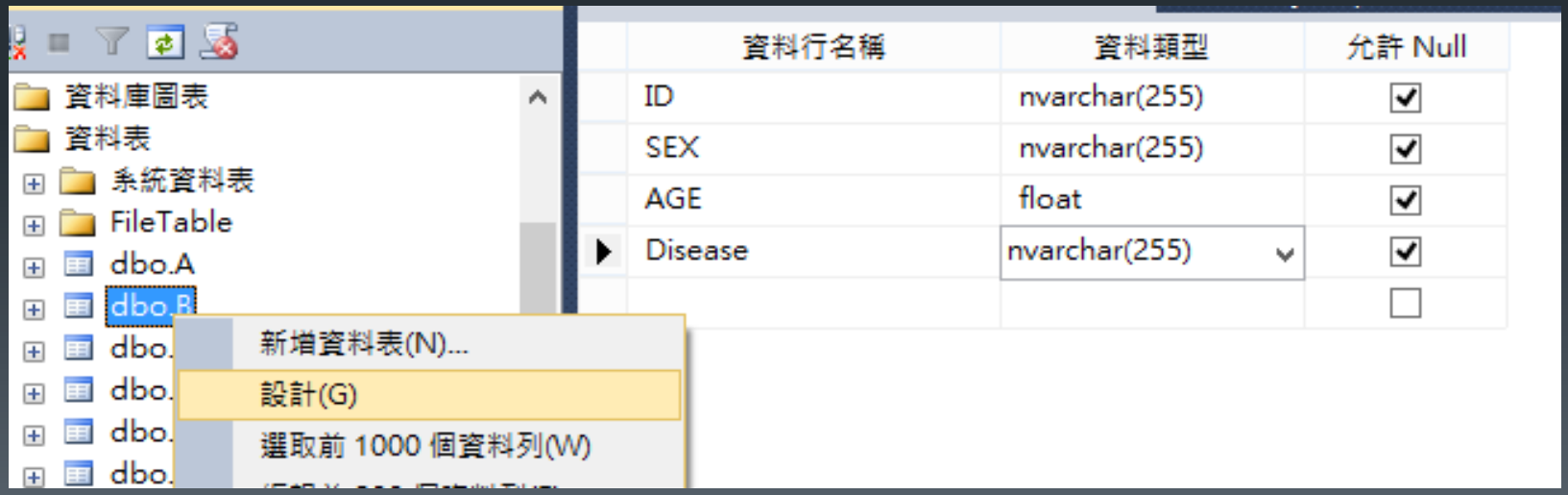
3.2 Basic Command

- <Ans3>：

```
SELECT *          /* 挑出所有欄位的資料 */
INTO B4           /* 存到新的資料表B4 */
FROM B1           /* 從B1表抓資料 */
UNION
SELECT *
FROM B3           /* 再從B3表抓資料 */
```

3.2 Basic Command

- 3.2.4 修改欄位屬性
- Step1：先在要修改欄位所屬的資料表按右鍵，點選設計(G)，可以看到目前個欄位的資料類型。



3.2 Basic Command

- Step2：現在想把AGE的欄位格式，從float轉換成nvarchar(15)。
- Code：

```
ALTER TABLE B   /*欲修改的欄位在B表中*/
ALTER COLUMN AGE nvarchar(15)
 /*將AGE欄位格式，從原本的float轉換成nvarchar(15)*/
```

資料行名稱	資料類型	允許 Null
ID	nvarchar(255)	☑
SEX	nvarchar(255)	☑
AGE	nvarchar(15)	☑
Disease	nvarchar(255)	☑
		☐

3.2 Basic Command

- 3.2.5 修改欄位名稱：
  將Disease欄位名稱改為ID_Disease。
- Code：

```
SP_RENAME 'B.Disease' , 'ID_Disease' ,
'COLUMN'
/*將B表中之Disease欄位名稱修改成
ID_Disease */
```


結果 訊息

	ID	SEX	AGE	ID_Disease
1	A	F	1	A5
2	B	F	2	B6
3	C	M	3	C7
4	A	F	1	A4
5	D	M	2	D5
6	E	M	5	E4

3.2 Basic Command

- 3.2.6 新增欄位：新增出生年份欄位。
- Code：

```
ALTER TABLE B
ADD 出生年份 NVARCHAR(15) /*將B表加入'出生年份'，欄位格式為NVARCHAR(15) */
```

	ID	SEX	AGE	ID+Disease	出生年份
1	A	F	1	A5	NULL
2	B	F	2	B6	NULL
3	C	M	3	C7	NULL
4	A	F	1	A4	NULL
5	D	M	2	D5	NULL
6	E	M	5	E4	NULL

3.2 Basic Command

- 3.2.7 刪除欄位：刪除出生年份欄位。
- Code：

```
ALTER TABLE B
DROP COLUMN 出生年份  /*將B表中出生年份欄位刪除 */
```

結果 訊息

	ID	SEX	AGE	ID_Disease
1	A	F	1	A5
2	B	F	2	B6
3	C	M	3	C7
4	A	F	1	A4
5	D	M	2	D5
6	E	M	5	E4

3.2 Basic Command

- 3.2.8 填入欄位
- Step1：在B表中新增欄位ID1。
- Code：

```
ALTER TABLE B
ADD ID1 NVARCHAR(15)  /*在B表
中新增ID1欄位*/
```

3.2 Basic Command

- Step2：將ID_Disease欄位的第一個英文字母，填入ID1。
- Code：

```
UPDATE B /*欲填入的欄位，位於資料表B中*/
SET ID1 = SUBSTRING (ID_Disease,1,1)
/*將ID_Disease欄位的第一個英文字母，填入ID1*/
```

	ID	SEX	AGE	ID_Disease	ID1
1	A	F	1	A5	A
2	B	F	2	B6	B
3	C	M	3	C7	C
4	A	F	1	A4	A
5	D	M	2	D5	D
6	E	M	5	E4	E

3.2 Basic Command

- NOTE：SUBSTRING(str,pos,len)，是由<str>中的第<pos>位置開始，選出接下去的<len>個字元。
- EX4：在B4表中新增ID_SEX欄位，並將ID與SEX合併填入其中。

ID	SEX	AGE	Disease	ID_SEX
A	F	1	A5	AF
B	F	2	B6	BF
C	M	3	C7	CM
F	F	4	F2	FF
G	M	6	G3	GM

3.2 Basic Command

- <Ans4>：

  Step1：

```
ALTER TABLE B4
ADD ID_SEX NVARCHAR(15)
```

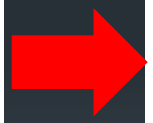
  Step2：

```
UPDATE B4
SET ID_SEX = ID+SEX
```

3.2 Basic Command

- 3.2.9 進階條件指令(WHERE IN)：從B表中抓出ID有出現在B2表中的資料(A、D、E)。

	ID	SEX	AGE	ID_Disease	ID1
1	A	F	1	A5	A
2	B	F	2	B6	B
3	C	M	3	C7	C
4	A	F	1	A4	A
5	D	M	2	D5	D
6	E	M	5	E4	E



結果 訊息

	ID	SEX	AGE	ID_Disease	ID1
1	A	F	1	A5	A
2	A	F	1	A4	A
3	D	M	2	D5	D
4	E	M	5	E4	E

3.2 Basic Command

- Code：

```
select *                    /* 挑出所有欄位的資料 */
into where_in
/* 存到新的資料表 where_in 中 */
from B                      /* 從B表抓資料 */
where ID IN (SELECT ID FROM B2)
/* 條件為ID有出現在B2表中(A、D、E)資料 */
```

3.2 Basic Command

- 3.2.10 進階條件指令(LEFT JOIN)：
  以表CD為主，去表A中擷取在表CD中沒有的資料欄位，並用icd9去連結兩表。

CD:

ID	icd9	col1
1	101	a
2	102	b
3	103	c

A:

icd9	c1	c2
101	11	22

ID	icd9	col1	c1	c2
1	101	a	11	22
2	102	b	NULL	NULL
3	103	c	NULL	NULL

3.2 Basic Command

- Code：

```
select CD.ID,教學資料庫
.dbo.CD.icd9,col1,A.c1,c2  /*表CD挑出
ID,icd9,col1欄位的資料；表A挑出C1,C2欄位
的資料*/
into Left_join       /* 存到新的資料表 Left_join
中*/
from CD  LEFT JOIN  A   /*以表CD為主，去
表A中擷取在表CD中沒有的資料欄位*/
on  CD.icd9 = A.icd9    /*用欄位icd9去連結兩
個表*/
```

CH4 實例操作

4.1 表單合併

- Question：如何利用第二章的指令，將左邊兩個表單，整理成右邊的表單？

ID	Class	Region	Stat	Doe
A	1	Taipei	70	80
A	1	Taipei	70	0
B	1	Ilan	70	0

ID	Class	Region	Stat	Doe
C	2	Kaohsiung	90	0
C	2	Kaohsiung	0	60
D	2	Taipei	0	50
D	2	Taipei	0	70

ID	Class	ID_Region	Stat	Doe	Total
A	1	ATaipei	140	80	220
B	1	BIlan	70	0	70
C	2	CKaohsiun	90	60	150
D	2	DTaipei	0	120	120

4.1 表單合併

- <Answer2>：
- Step1：手動匯入 table9與table10。
- Step2：將table9與table10合併成table11。

```
SELECT *
INTO table11
FROM  table9
UNION ALL
SELECT *
FROM  table10
```

4.1 表單合併

- Step3：將table11，依據ID，求出Stat與Doe的總和，存入table12。

```
SELECT ID, Class, Region, SUM(Stat) as Stat,
SUM(Doe) as Doe
into table12
FROM table11
GROUP BY ID, Class, Region
```

- Step4：於table12新增欄位ID_Region，並填入ID加上Region。

a.
```
ALTER TABLE table12
ADD ID_Region nvarchar(15)
```

b.
```
UPDATE table12
SET ID_Region = ID+Region
```

4.1 表單合併

- Step5：於table12刪除欄位Region。

```
ALTER TABLE table12
DROP COLUMN Region
```

- Step6：重新排列table12的欄位，依序為ID, Class, ID_Region, Stat, Doe，並存入table13。

```
SELECT ID, Class, ID_Region, Stat, Doe
INTO table13
FROM table12
```

4.1 表單合併

- Step7：新增欄位Total，並填入Stat與Doe的加總

a.

```
ALTER TABLE table13
ADD Total float
```

b.

```
UPDATE table13
SET Total = Stat+Doe
```

結果 訊息

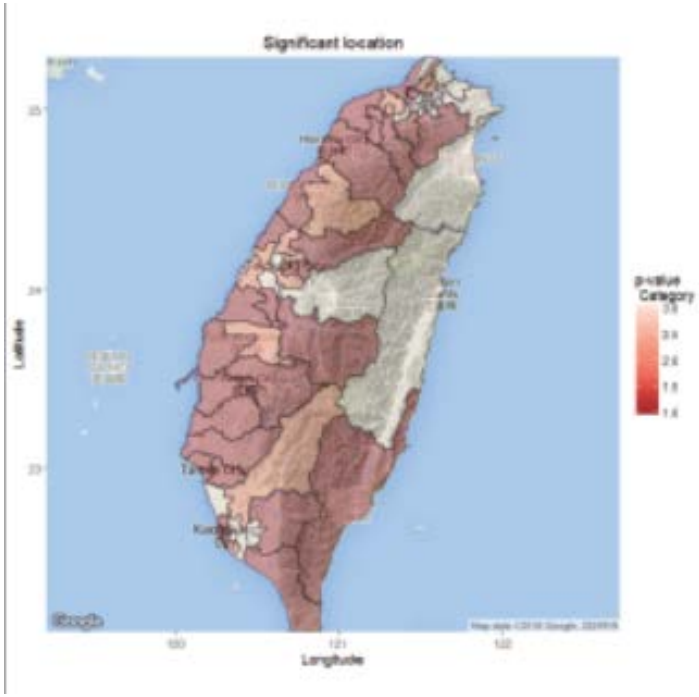
	ID	Class	ID_Region	Stat	Doe	Total
1	A	1	ATaipei	140	80	220
2	B	1	BIlan	70	0	70
3	C	2	CKaohsiung	90	60	150
4	D	2	DTaipei	0	120	120

Thanks for your listening !

R語言與文字採礦

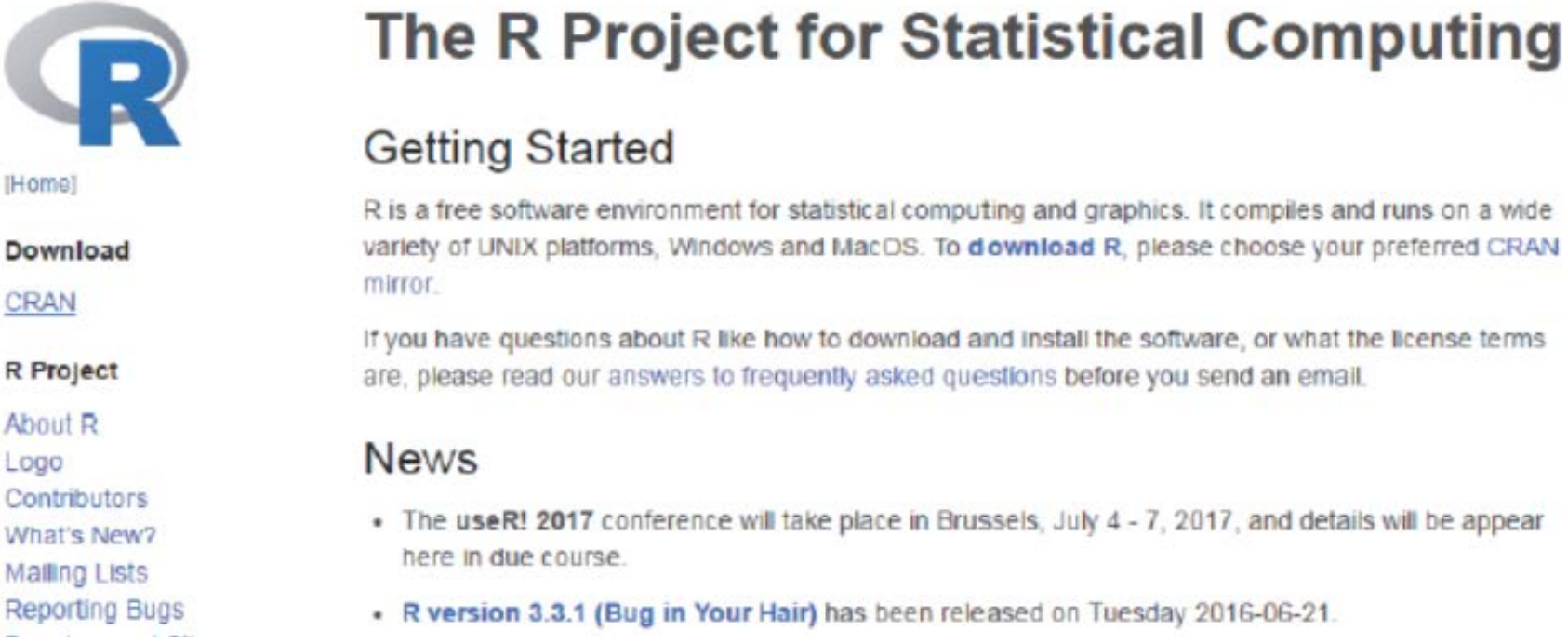
R語言簡介

- R語言自1993年問世，用於統計分析、繪圖、資料採礦、矩陣運算與機器學習等多個面向
- 兩大特色：免費下載、開放原始碼。
- 套件：ggmap, ggplot
- R Studio/R Pubs/GitHub



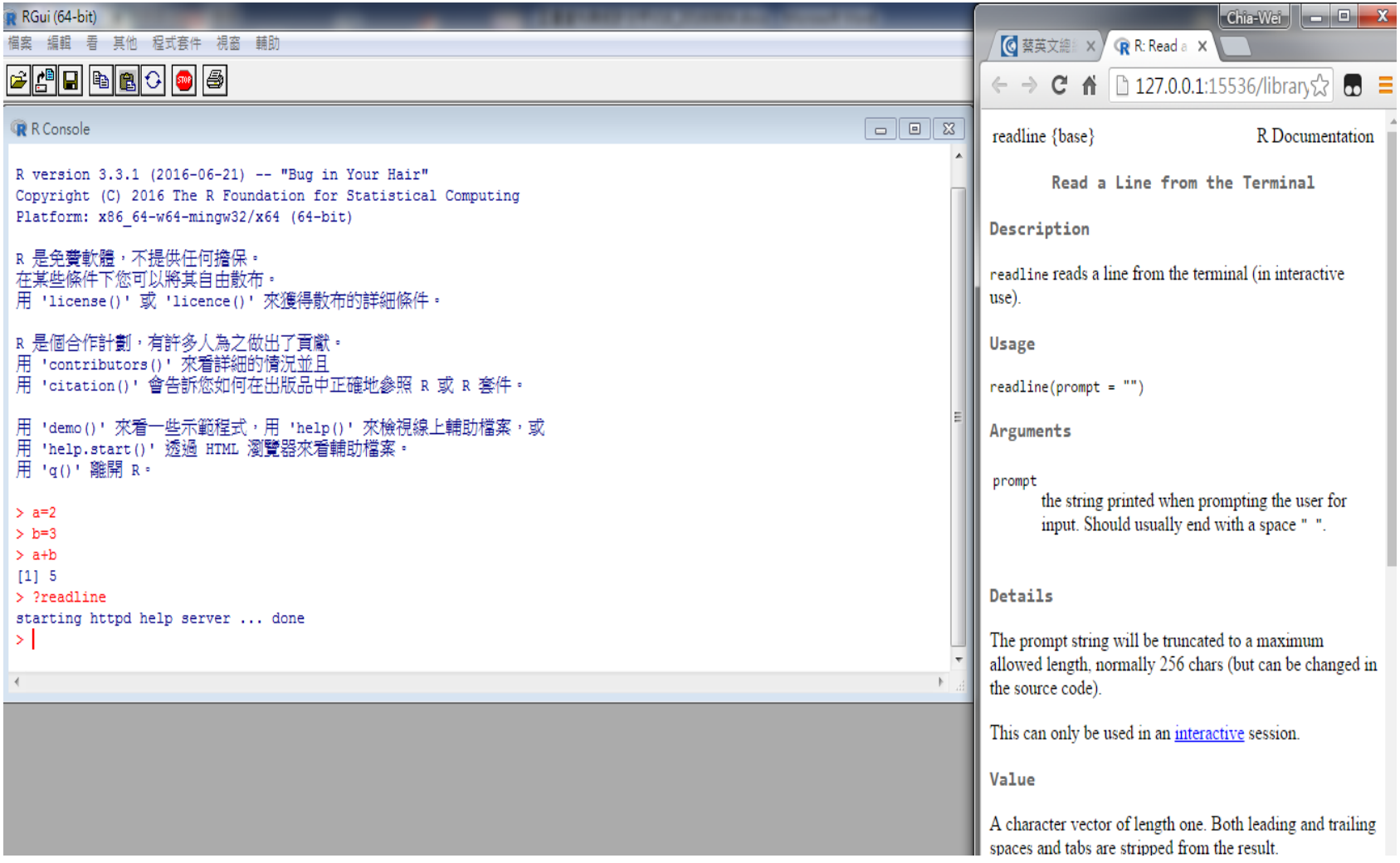
安裝軟體

- 請到R軟體的網站，https://www.r-project.org/
- 點左側「Download: CRAN」之後，搜尋「Taiwan」並任選一個下載點（臺灣大學或元智大學），再選擇作業系統「Download R for Windows」（以Windows為例），點選「base」即可下載安裝。



進入使用者介面

- 進入圖形使用者介面程式。進行簡單的程式練習

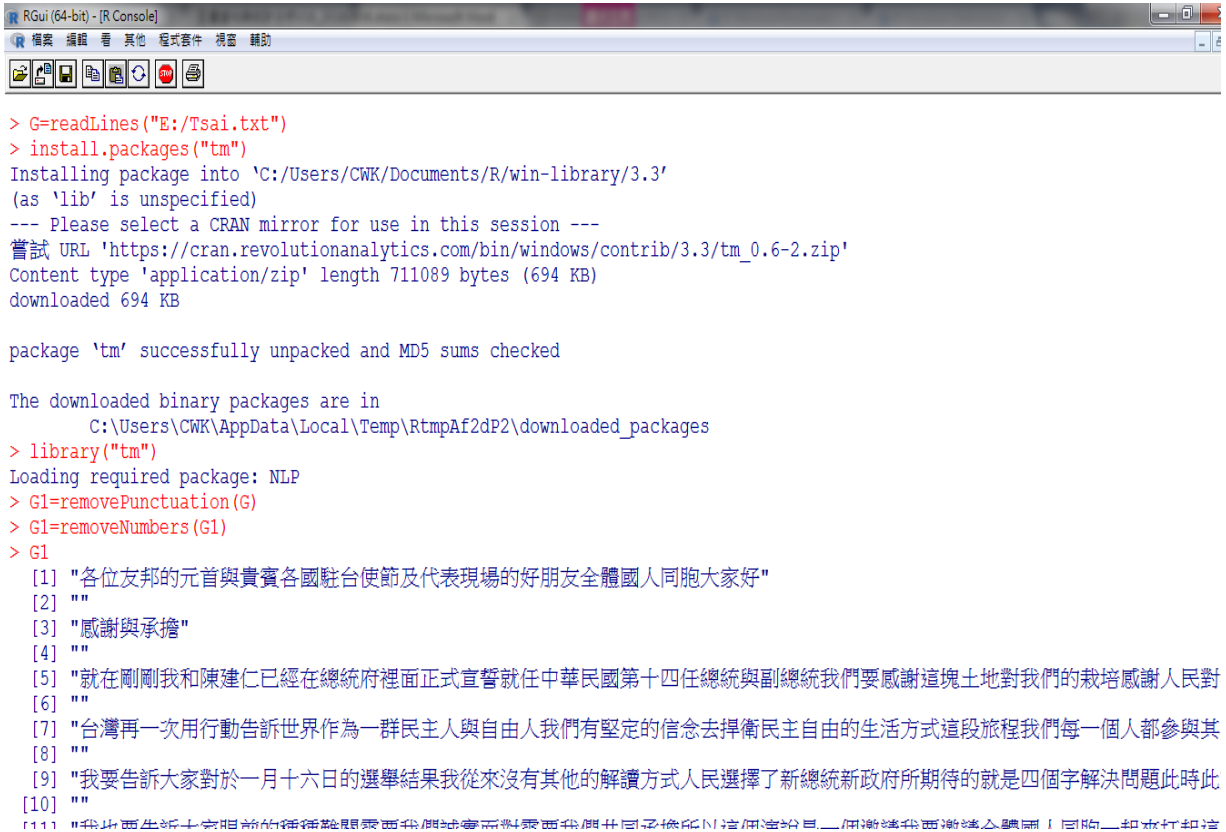


讀取文字資料

- 利用「readLine()」語法讀取文字檔

刪除標點符號與數字

- 安裝套件「tm」或「tmcn」後，輸入library ()語法載入套件，即可使用「removePunctuation()」、「removeNumbers()」等語法依序將標點符號、數字移除



將各句連結成一整個不分行的段落

中文斷詞

RGui (32-bit) - [R Console]

檔案 編輯 看 其他 程式套件 視窗 輔助

```
[141] "謝謝大家"
> row=ncol(t(G1))
> for (i in 1:row-1)
+ {
+ G1[i+1]=paste(G1[i],G1[i+1])
+ }
> Gf=gsub(" ","",G1)
> Gfinal=Gf[row]
> Gfinal
[1] "各位友邦的元首與貴賓各國駐台使節及代表現場的好朋友全體國人同胞大家好感謝與承擔就在剛剛我和陳建仁已經在總統府裡面正式宣
> word=NULL
> n2=nchar(Gfinal)
> ii=1
> n3=n2-ii
> for(i in 1:n3)
+ {
+ word=c(word,substr(Gfinal,i,i+ii))
+ }
> word
   [1] "各位" "位友" "友邦" "邦的" "的元" "元首" "首與" "與貴" "貴賓" "賓各"
  [11] "各國" "國駐" "駐台" "台使" "使節" "節及" "及代" "代表" "表現" "現場"
  [21] "場的" "的好" "好朋" "朋友" "友全" "全體" "體國" "國人" "人同" "同胞"
  [31] "胞大" "大家" "家好" "好感" "感謝" "謝與" "與承" "承擔" "擔就" "就在"
  [41] "在剛" "剛剛" "剛我" "我和" "和陳" "陳建" "建仁" "仁已" "已經" "經在"
  [51] "在總" "總統" "統府" "府裡" "裡面" "面正" "正式" "式宣" "宣誓" "誓就"
  [61] "就任" "任中" "中華" "華民" "民國" "國第" "第十" "十四" "四任" "任總"
  [71] "總統" "統與" "與副" "副總" "總統" "統我" "我們" "們要" "要感" "感謝"
  [81] "謝這" "這塊" "塊土" "土地" "地對" "對我" "我們" "們的" "的栽" "栽培"
  [91] "培感" "感謝" "謝人" "人民" "民對" "對我" "我們" "們的" "的信" "信任"
```

雙字詞出現次數統計

RGui (32-bit) - [R Console]

檔案 編輯 看 其他 程式套件 視窗 輔助

```
[133] ""                                                                                                   $
[134] "歷史會記得我們這個勇敢的世代這個國家的繁榮尊嚴團結自信和公義都有我們努力的痕跡歷史會記得我們的勇敢我們在年一起把國家帶向新的方向這塊土地上的每一個 人都因為參與$
[135] ""                                                                                                   $
[136] "剛才表演節目中的一首歌曲當中有一句讓我很感動的歌詞"                                                 $
[137] "台語現在是彼一天勇敢ㄟ台灣人"                                                                       $
[138] ""                                                                                                   $
[139] "各位國人同胞兩千三百萬的台灣人民等待已經結束現在就是那一天今天明天未來的每一天我們都要做一個守護民主守護自由守護這個國家的台灣人"    $
[140] ""                                                                                                   $
[141] "謝謝大家"                                                                                           $
>
> row=ncol(t(G1))
> for (i in 1:row-1)
+ {
+ G1[i+1]=paste(G1[i],G1[i+1])
+ }
> Gf=gsub(" ","",G1)
> Gfinal=Gf[row]
> Gfinal
[1] "各位友邦的元首與貴賓各國駐台使節及代表現場的好朋友全體國人同胞大家好感謝與承擔就在剛剛我和陳建仁已經在總統府裡面正式宣誓就任中華民國第十四任總統與副總統我們要感$
>
>
> word=NULL
> n2=nchar(Gfinal)
> ii=1
> n3=n2-ii
> for(i in 1:n3)
+ {
+ word=c(word,substr(Gfinal,i,i+ii))
+ }
>
>
> wordtable_i=table(word)
> wordtable=sort(wordtable_i,decreasing=TRUE)
> wordtable[1:10]
word
我們 台灣 政府 國家 一個 新政 經濟 這個 民主 社會
  86   41   37   32   29   27   27   25   24   22
>
```

安裝套件wordcloud─繪製文字雲