數據科學概論 作業2

挑選資料集：**Mice Protein Expression**

UCI: http://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression

Kaggle: https://www.kaggle.com/ruslankl/mice-protein-expression

資料集內的特徵

[1] 老鼠 ID，例如 309_6 表 309 號老鼠在做第 6 個實驗。
[2:78] 77 種蛋白質表現程度，*For example: DYRK1A_n*
[79] Genotype（基因型）: control (c) or trisomy (t，染色體三體，即唐氏症基因)
[80] Treatment: memantine (m，藥物) or saline (s，生理食鹽水)
[81] Behavior: context-shock (CS, 給予學習刺激) or shock-context (SC, 無刺激)
[82] Class，由以上三個變項分為 8 個 class：
c-CS-s, c-CS-m, c-SC-s, c-SC-m, t-CS-s, t-CS-m, t-SC-s, t-SC-m

由 [2] 提供的表格可知一些其資料集 attribute 的一些意義。

先將資料由 class 分類，並挑選一些 attribute 來繪圖觀察。

```
control <- subset(protein, class=="c-CS-m" | class=="c-CS-s" |
                                class=="c-SC-m" | class=="c-SC-s")

tri <- subset(protein, class=="t-CS-m" | class=="t-CS-s" |
                                class=="t-SC-m" | class=="t-SC-s")

tri_m <- subset(tri, class=="t-CS-m" | class=“t-SC-m”)

tri_CS <- subset(protein, class=="t-CS-m" | class=“t-CS-s")
```

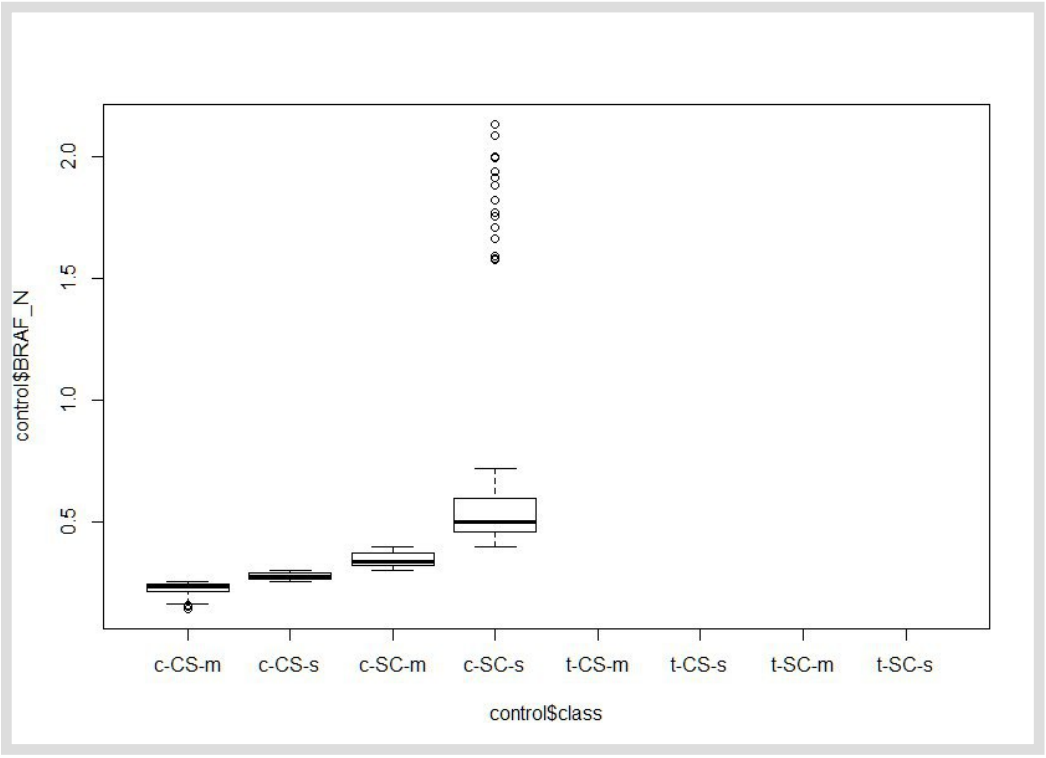
資工系 0416235 劉昱劭 | c2525441@gmail.com

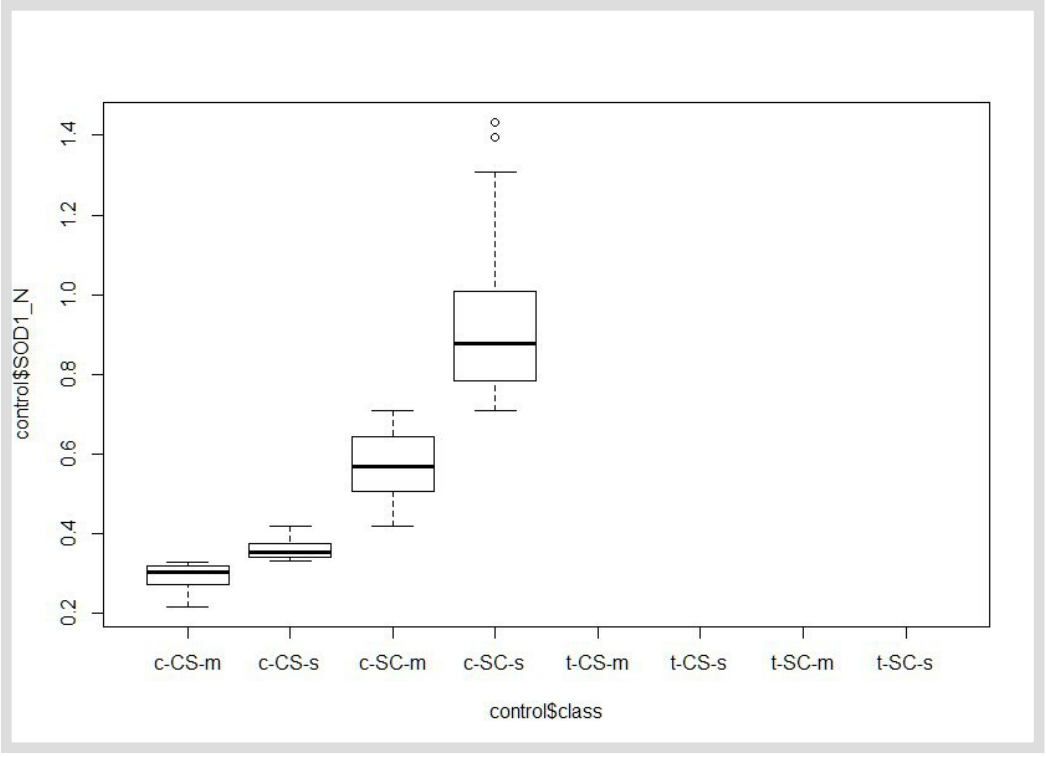
挑選 ATTRIBUTE 繪圖分析

1. 對 control (對照組) 和一些蛋白質畫 qqplot，找到能夠辨別出 CS 和 SC 的蛋白質 attribute，即找出和「是否給予學習刺激」相關性較高的 attribute。

```
ggplot(data = control, aes(x = BRAF_N, y = SOD1_N, colour = class)) +
 geom_point()
```

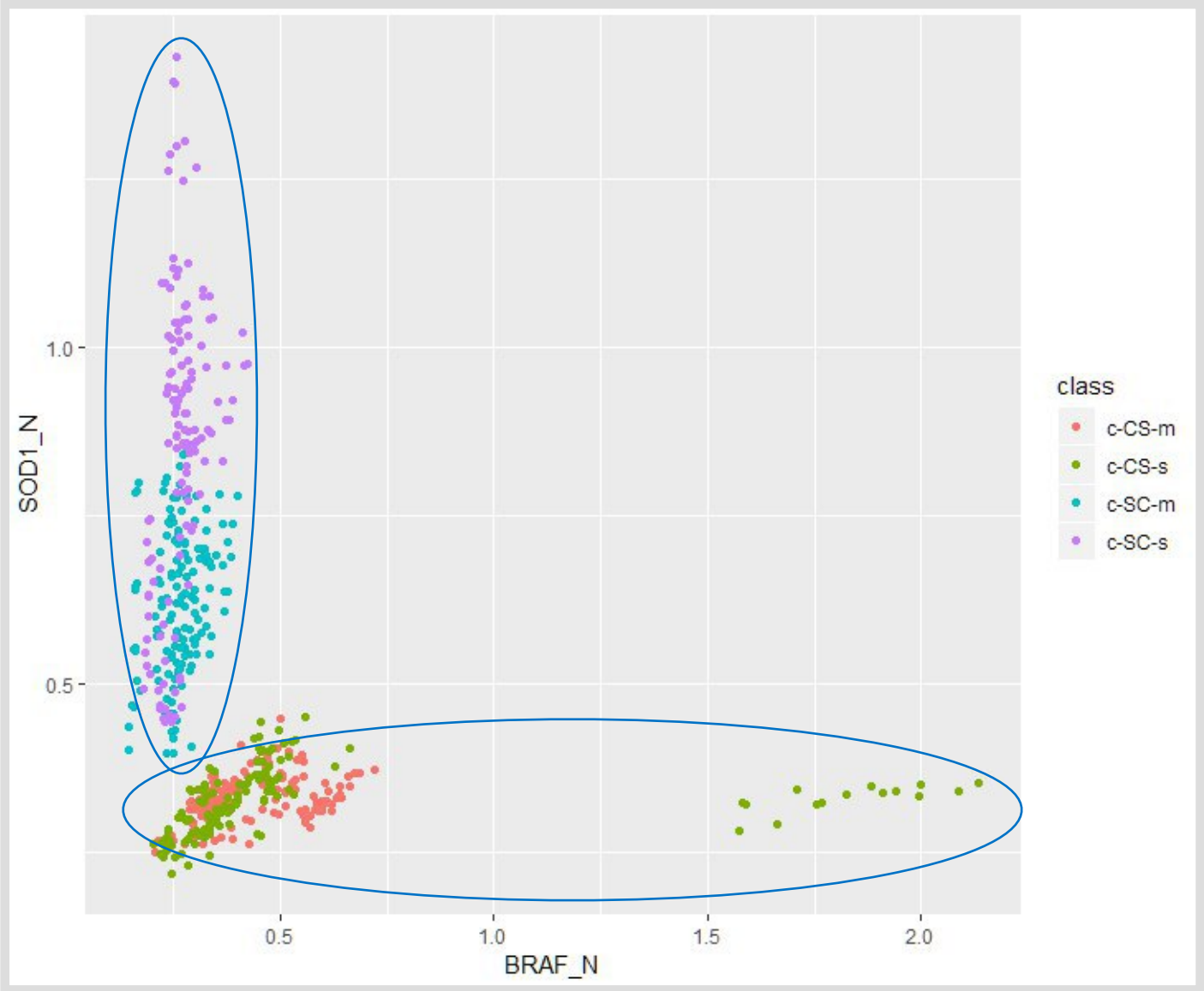
例如 SOD1 和 BRAF：





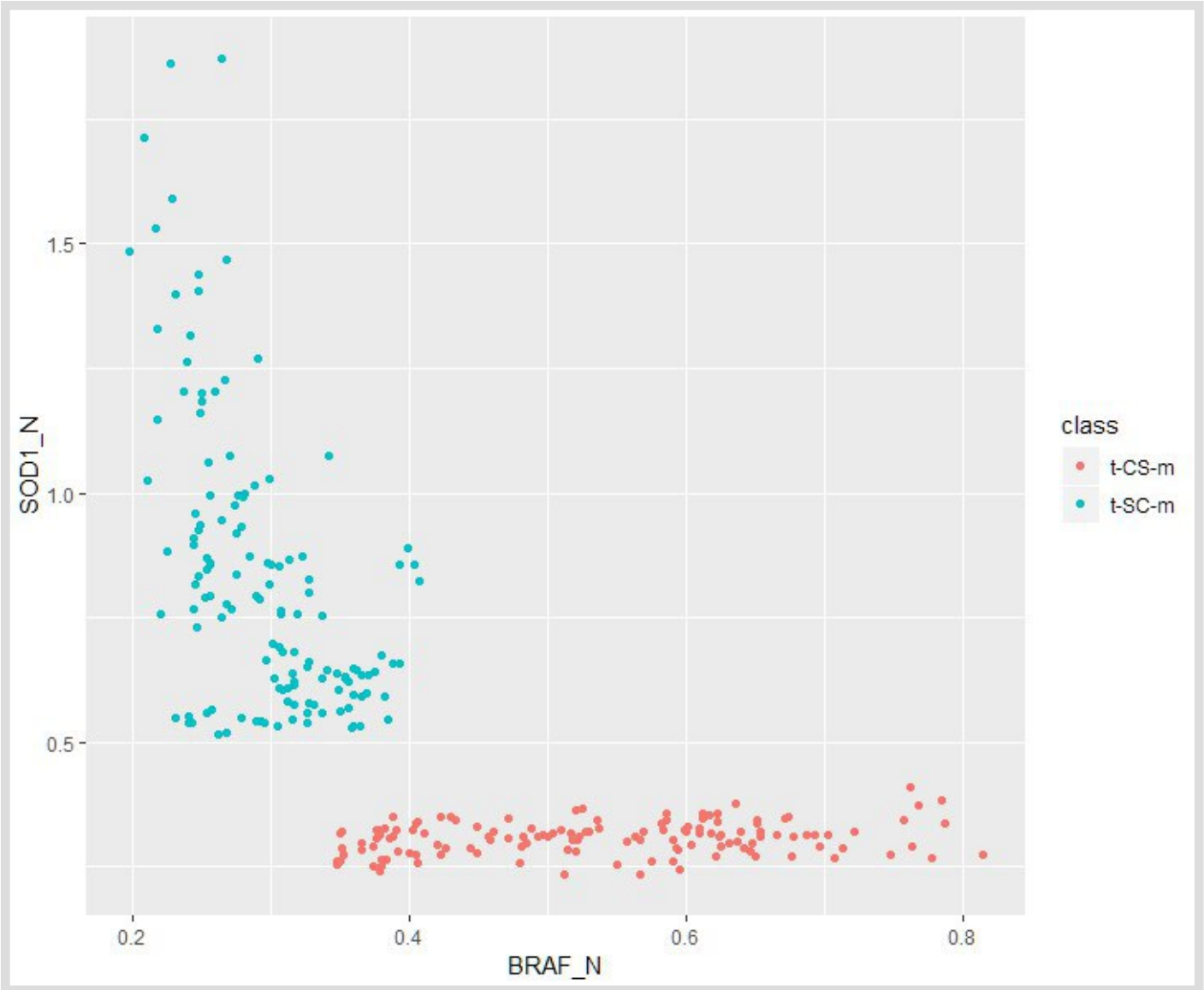
2. 或可用 ggplot 畫出點圖，看起來也能分出 CS 和 SC

```
ggplot(data = control, aes(x = BRAF_N, y = SOD1_N, colour = class)) +
 geom_point()
```



3. 比較 t-CS-m 和 t-SC-m 兩種 class，一組有學習刺激，另一組沒有，若能夠分得出這兩組，表示學習刺激對這兩種蛋白質可能呈反向的相關性。

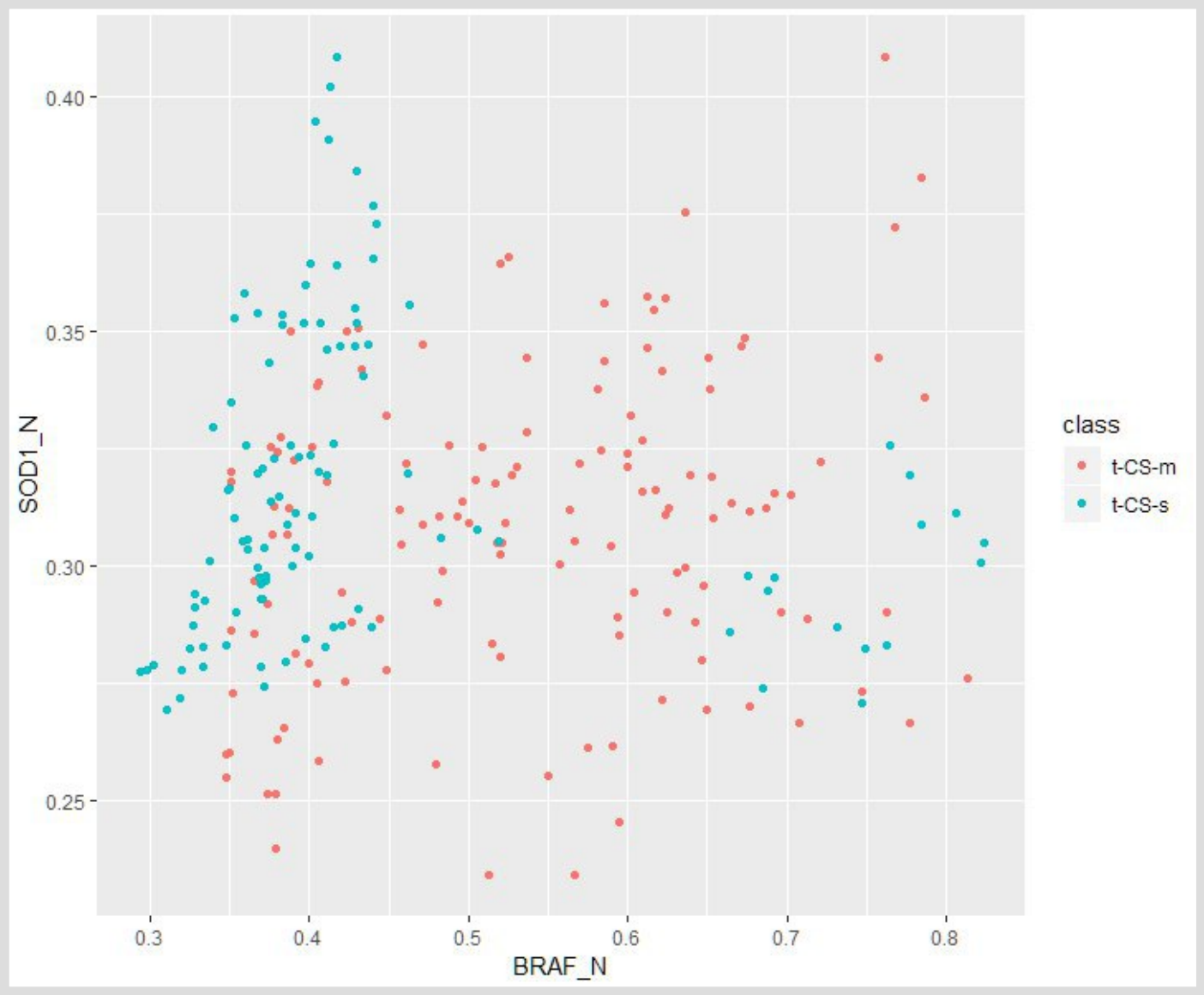
```
ggplot(data = tri_m, aes(x = BRAF_N, y = SOD1_N, colour = class)) +
 geom_point()
```



結果分的出來，表示「給予學習刺激」應會使 BRAF 蛋白質表現較強，使 SOD1 蛋白質表現較弱。

4. 比較 t-CS-m 和 t-CS-s 兩種 class，可看 memantine 對兩種蛋白質的影響。

```
ggplot(data = tri_CS, aes(x = BRAF_N, y = SOD1_N, colour = class)) +
 geom_point()
```



結果兩個 class 混在一起，可見 memantine 的效果並非影響該兩種蛋白質。
由 [2] 提供的表格可知，此兩種蛋白質都並非屬於神經傳導類型的蛋白質，而 memantine 是抑制過度神經傳導的藥物，所以以此當標準分不出來是合理的。

相關論文

1. [WEB LINK]
HIGUERA C, GARDINER KJ, CIOS KJ (2015)
SELF-ORGANIZING FEATURE MAPS IDENTIFY PROTEINS CRITICAL TO LEARNING IN A MOUSE MODEL OF DOWN SYNDROME.
PLOS ONE 10(6): E0129126. JOURNAL.PONE.0129126

2. [WEB LINK]
AHMED MM, DHANASEKARAN AR, BLOCK A, TONG S, COSTA ACS, STASKO M, ET AL. (2015)
PROTEIN DYNAMICS ASSOCIATED WITH FAILED AND RESCUED LEARNING IN THE TS65DN

MOUSE MODEL OF DOWN SYNDROME. PLOS ONE 10(3): E0119491.