手把手教你 R 語言資料分析實務

中央研究院
陳柏亨 / 張毓倫

課程簡介

講者

陳柏亨

張毓倫

今天的流程

- 主題講解
- 實作 demo
- 練習
  - 公佈解答

- 課程投影片及需要的 data 與 code 皆可下載
  - Op1: http://goo.gl/bGLgbQ
  - Op2: https://goo.gl/zyUclg
  - Op3: https://goo.gl/YRFpwC

圖例

請打開 session_00_install_R_packages.R

→ 代表在 Rstudio 工作

請加入我們的課程討論區

→ 代表開啟網頁

Ex. https://www.facebook.com/events/677036392444252

今天課堂中的問題也可在此討論

Pkg “installr”

```
# installing/loading the package:
if(!require(installr)) {
install.packages("installr"); require(installr)}

# using the package:
updateR()
```

資料收集
Data Collection

Session A

資料：蘋果日報暖流版

蘋果暖流新聞

頭條要聞 〉 A3959 夫癌末 貧婦恐慌撐8口

分享到FB 分享到g+ 分享到Plurk 分享到Twitter

A3959 夫癌末 貧婦恐慌撐8口

2016年06月24日 傳送 讚 2 G+1 0 我要捐款 更多專欄文章

阿芬(左)照料丈夫灌食。

49歲楊景光2月罹患舌癌第四期，腫瘤切除後仍覺舌頭僵硬，且因化、放療疲累不適，咬字不清吃力地說：「責任還未了…一定要…好起來，再去工作。」

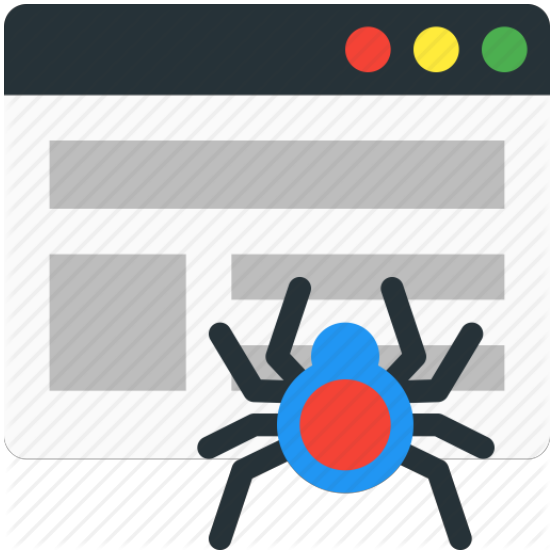
報導•攝影／仲芝蓉

46歲妻子阿芬(葉淑芬)將稀飯打泥，用鼻胃管協助丈夫灌食，並指著一旁流質營養品說，丈夫罹癌後體重掉10幾公斤，醫生說要補充營養，才有體力對抗癌細胞，「每天得灌6瓶營養品，每月開銷就要1萬元」。

採訪時，23歲在外地讀大三的長子阿軒返家探望爸爸，見10歲讀小5的么弟小宸從外回來，馬上叨唸弟弟：「衣服怎麼弄得溼答答？趕快去換，不要感冒了！」小宸一邊嘟囔著是去學校打籃球，流很多汗，一邊進房換衣服。

抓取網頁

- download.file
  - 請打開 session_A_DataCollection.R



抓取網頁

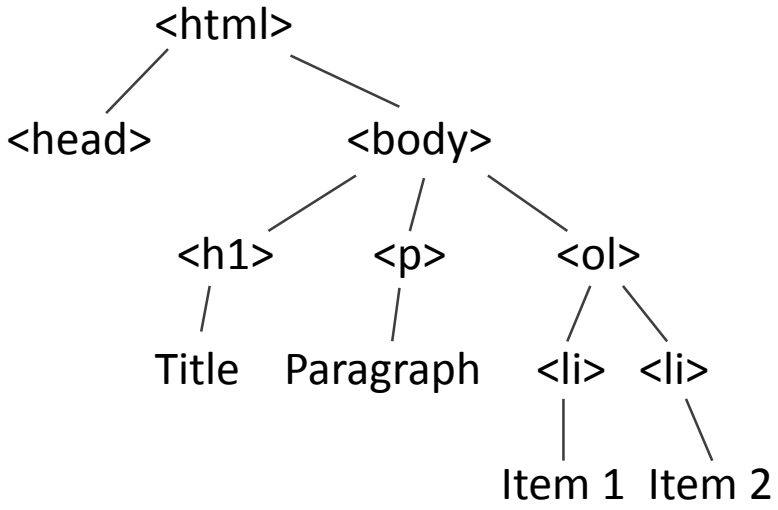
```
# Article url
url <-
“http://www.appledaily.com.tw/appledaily/article/headl
ine/20160704/37293645”
# save the page
download.file(url, “data/test.html”)
```

存好檔案，如何把資訊擷取出來？

網頁架構及語法

<標籤> 內容 </標籤>

```
<html>
   <head>
   </head>
   <body>
            <h1> Title </h1>
            <p> Paragraph </p>
            <ol>
                <li> Item 1 </li>
                <li> Item 1 </li>
            </ol>
    </body>
</html>
```



捐款進度報告

編號	報導標題				
A3883	單親父癌末 為3兒拼活				
A3882	女不捨半癱父「要當他的				
A3881	雙親亡 16歲孫苦養病嬤				
A3880	愛女突癱 病母：我虧欠				
A3879	夫肝病妻骨折「沒這麼難過」				
A3878	資優兒流感癱瘓 父：不放棄	2016/03/10	已結案	695471	明細
A3877	單親媽抗癌 牽掛獨生女	2016/03/09	已結案	1335829	明細

td | 59.5×42

```
▼<div id="charitysidebox3">
  ▼<div id="inquiry3">
    ▼<table>
      ▼<tbody>
        ▶<tr class="odd">…</tr>
        ▶<tr>…</tr>
        ▶<tr class="odd">…</tr>
        ▼<tr class="odd">
            <td>A3882</td> == $0
          ▶<td>…</td>
            <td id="wordcenter">2016/03/16</td>
          ▶<td id="wordcenter">…</td>
            <td id="wordcenter">757665</td>
          ▶<td id="wordcenter">…</td>
          </tr>
        ▶<tr class="odd">…</tr>
        ▶<tr class="odd">…</tr>
        ▶<tr class="odd">…</tr>
        ▶<tr class="odd">…</tr>
        ▶<tr class="odd">…</tr>
```

XPath

標記	意義
/	選取根節點
//	選取任何節點
@	選取屬性 (attribute)
*	選取所有節點
|	OR

Xpath = “//*[@id=‘inquiry3"]/table//tr[4]/td[1]”

捐款進度報告

編號	報導標題
A3883	單親父癌末 為3兒拼活
A3882	女不捨半癱父「要當他的腳
A3881	雙親亡 16歲孫苦養病嬤
A3880	愛女突癱 病母：我虧欠她
A3879	夫肝病妻骨折「沒這麼難受過」
A3878	資優兒流感癱瘓 父：不放棄
A3877	單親媽抗癌 牽掛獨生女

td | 59.5×42

```
▼<div id="charitysidebox3">
  ▼<div id="inquiry3">
    ▼<table>
      ▼<tbody>
        ▶<tr class="odd">…</tr>
        ▶<tr>…</tr>
        ▶<tr class="odd">…</tr>
        ▼<tr class="odd">
            <td>A3882</td> == $0
          ▶<td>…</td>
            <td id="wordcenter">2016/03/16</td>
          ▶<td id="wordcenter">…</td>
            <td id="wordcenter">757665</td>
          ▶<td id="wordcenter">…</td>
          </tr>
        ▶<tr class="odd">…</tr>
        ▶<tr class="odd">…</tr>
        ▶<tr class="odd">…</tr>
        ▶<tr class="odd">…</tr>
        ▶<tr class="odd">…</tr>
```

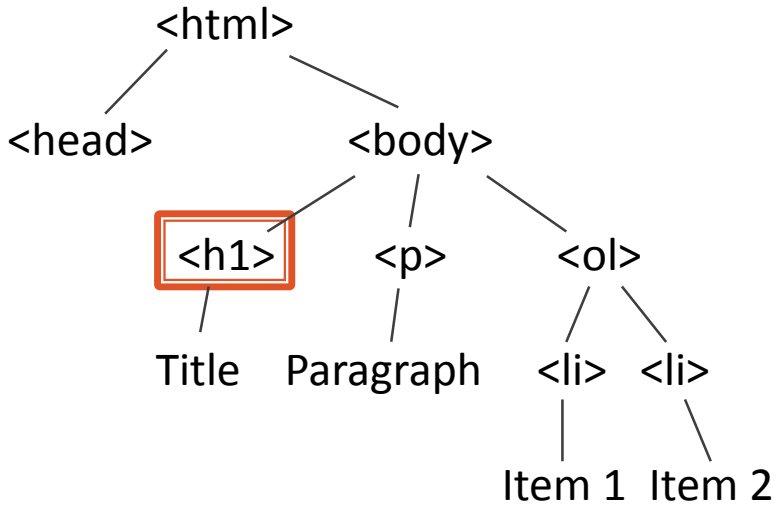
讀取網頁資訊 (pkg xml2)

```
library(xml2)

# set your target url
doc <- read_html(url)

# set the xpath of info needed
xpath <- “//*[@id=‘inquiry3’]/table//tr[4]/td[1]”
xml_find_all(doc, xpath) %>% xml_text()
```

讀取網頁資訊 (pkg xml2)

- xml2
  - 讀取網頁：`read_html; read_xml`
  - 選擇節點：`xml_find_all; xml_find_one`
  - 擷取資訊：`xml_text; xml_attrs`

```
<html>
  <head>
  </head>
  <body>
      <h1> Title </h1>
      <p> Paragraph </p>
      <ol>
          <li> Item 1 </li>
          <li> Item 1 </li>
```



讀取網頁資訊 (pkg xmlview)

```
# open the document to test your xpath
xml_view(doc, add_filter = T)
```

XPath:

```
<?xml version="1.0" encoding="utf-8" standalone="yes"?>
<!DOCTYPE html>
<!--[if lt IE 7 ]>
<html lang="zh-TW" class="ie6 ielt8">
    <![endif]-->
    <!--[if IE 7 ]>
    <html lang="zh-TW" class="ie7 ielt8">
        <![endif]-->
        <!--[if IE 8 ]>
        <html lang="zh-TW" class="ie8">
            <![endif]-->
            <!--[if (gte IE 9)|!(IE)]>
            <!-->
            <html lang="zh-TW">
                <!--
                <![endif]-->
                <head>
                    <meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>
                    <title>蘋果日報慈善基金會 | 蘋果日報</title>
                    <meta name="Title" content="蘋果日報慈善基金會 | 蘋果日報"/>
                    <meta name="description" content="蘋果日報慈善基金會 | 蘋果日報"/>
                    <meta name="keywords" content="蘋果日報慈善基金會 , 蘋果日報"/>
                    <meta name="Robots" content="INDEX,FOLLOW"/>
```

練習 A-01 (8 mins)

- 請觀察網頁，寫出可以擷取正確資訊的 Xpath
  - 1-1: 編號
  - 1-2: 欄位名稱
  - 1-3: 捐款文章連結
  - bonus: 共計頁數

練習 A-01 (8 mins)

1-1 1-3

1-2

編號	報導標題	刊登日期	狀態	累計(元)	捐款明細
A3883	單親父癌末 為3兒拼活	2016/03/17	已結案	993429	明細
A3882	女不捨半癱父「要當他的腳」	2016/03/16	已結案	757665	明細
A3881	雙親亡 16歲孫苦養病嬤	2016/03/15	已結案	2027619	明細
A3880	愛女突癱 病母：我虧欠她	2016/03/14	已結案	832743	明細
A3879	夫肝病妻骨折「沒這麼難受過」	2016/03/11	已結案	650769	明細
A3878	資優兒流感癱瘓 父：不放棄	2016/03/10	已結案	695471	明細
A3877	單親媽抗癌 牽掛獨生女	2016/03/09	已結案	1335829	明細

練習 A-01 (8 mins)

- 請觀察網頁，寫出可以擷取正確資訊的 Xpath
  - 1-1: 編號
  - 1-2: 欄位名稱
  - 1-3: 捐款文章連結
  - bonus: 共計頁數

練習 A-01 (解答)

session_A_ex01.R

練習 A-02 (10 mins)

- 請將捐款進度報告頁面轉存成 csv 檔

```
# 每位學員取第 n 頁
page <- sample(1:n.page, 1)
```

捐款進度報告

累計 3788 筆，共 190 頁 | 目前在 第2頁

編號	報導標題	刊登日期	狀態	累計(元)	捐款明細
A3949	母老姐憨 遮男「我得拾命活下去」	2016/06/10	未結案	739543	明細
A3948	夫腰傷子截肢 鬱婦萬元薪苦撐	2016/06/08	已結案	573972	明細
A3947	男憂腦傷妻中風母「生活慘慘慘」	2016/06/07	已結案	611042	明細
A3946	單親父遮歿 病嬤難顧3孫	2016/06/06	已結案	936188	明細
A3945	小姊弟憂父遮病 躲被偷哭	2016/06/03	已結案	754182	明細
A3944	母罹罕病失明 幸福變調	2016/06/02	已結案	665325	明細
A3943	父中風母罹遮 國一女怕來不及孝順	2016/06/01	已結案	884586	明細
A3942	撐家漢遮攤散已斷炊「措手不及」	2016/05/31	已結案	649365	明細

df_article_raw.csv

練習 A-02 (10 mins)

session_A_ex02.R

- 欄位
  - aid
  - case.closed
  - date.published
  - donation
  - title
  - url.article
  - url.detail

捐款進度報告

累計 3788 筆，共 190 頁 | 目前在 第2頁

編號	報導標題	刊登日期	狀態	累計(元)	捐款明細
A3949	母老姐憨 癌男「我得拚命活下去」	2016/06/10	未結案	739543	明細
A3948	夫腰傷子截肢 鬱婦萬元薪苦撐	2016/06/08	已結案	573972	明細
A3947	男憂腦傷妻中風母「生活慘慘慘」	2016/06/07	已結案	611042	明細
A3946	單親父癌歿 病變難顧3孫	2016/06/06	已結案	936188	明細
A3945	小姊弟憂父癌病 躲被偷哭	2016/06/03	已結案	754182	明細
A3944	母罹罕病失明 幸福變調	2016/06/02	已結案	665325	明細
A3943	父中風母罹癌 國一女怕來不及孝順	2016/06/01	已結案	884586	明細
A3942	撐家漢癌擴散已斷炊「措手不及」	2016/05/31	已結案	649365	明細

df_article_raw.csv

練習 A-03 (15 mins)

- 請規劃如何分析蘋果公益捐款資料，依設定的目標將需要的網頁資訊擷取轉存下來。

練習 A-03 (15 mins)

session_A_ex03.R

- Outcome
  - df_article_raw.csv
  - 文章.txt
  - 捐款明細.txt

下一堂課需要用到的資料欄位

- df_article.csv
  - aid
  - case.closed
  - date.published
  - donation
  - title
  - url.article
  - url.detail
  - donor
  - date.funded
  - journalist
  - n.fb.comment
  - n.fb.like
  - n.fb.share
  - n.fb.total
  - n.image
  - n.word

Data Manipulation

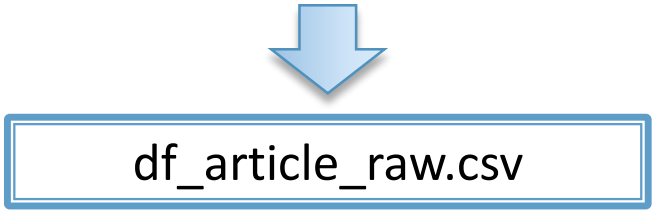
df_article.csv

- 欄位
  - aid
  - case.closed
  - date.published
  - donation
  - title
  - url.article
  - url.detail
  - donor
  - date.funded
  - journalist
  - n.fb.comme
  - n.fb.like
  - n.fb.share
  - n.fb.total
  - n.image
  - n.word

捐款進度報告

累計 3788 筆，共 190 頁 | 目前在 第2頁

編號	報導標題	刊登日期	狀態	累計(元)	捐款明細
A3949	母老姐憨 癌男「我得拚命活下去」	2016/06/10	未結案	739543	明細
A3948	夫腰傷子截肢 鬱婦萬元薪苦撐	2016/06/08	已結案	573972	明細
A3947	男憂腦傷妻中風母「生活慘慘慘」	2016/06/07	已結案	611042	明細
A3946	單親父癌歿 病臊難顧3孫	2016/06/06	已結案	936188	明細
A3945	小姊弟憂父癌病 躲被偷哭	2016/06/03	已結案	754182	明細
A3944	母罹罕病失明 幸福變調	2016/06/02	已結案	665325	明細
A3943	父中風母罹癌 國一女怕來不及孝順	2016/06/01	已結案	884586	明細
A3942	撐家漢癌擾數已斷炊「揖手不及」	2016/05/31	已結案	649365	明細



df_donation.csv

- In db_donation_txt.rar

捐款進度報告

《A3948捐款明細》

瀏覽相關報導：2016-06-08 夫腰傷子截肢 孀婦萬元薪苦撐　蘋果急難救助金(讀者不指定捐款) 25000元

筆數	捐款人姓名	累計(元)	捐款明細
1	蕭政豪	15000	2016/6/16
2	PANDORA Taiwan陳昱龍、黃庭姍	13600	2016/6/29
3	施炳宏	10000	2016/6/14
4	黃秉鈞	10000	2016/6/16
5	威潤科技股份有限公司	6000	2016/6/21
6	李榮宗	5400	2016/7/5
7	周珮詩	5400	2016/6/23
8	張坤銘	5153	2016/6/29
9	林長甫	5000	2016/6/9
10	財團法人台中市私立世貿社會福利慈善事業基金會	5000	2016/6/20
11	蔡裕靖	5000	2016/6/26
12	劉明宗	5000	2016/6/17
13	486團購	5000	2016/6/29
14	陳宜新	5000	2016/6/9

df_donation.csv

練習 A-04 (Homework)

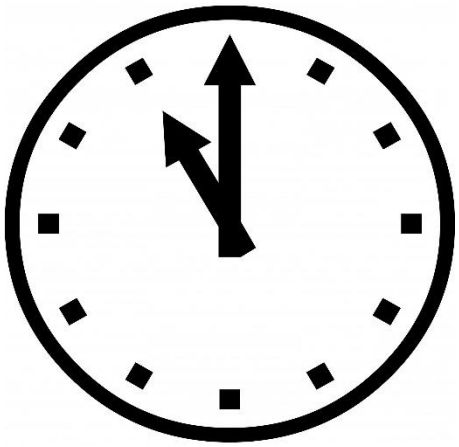
- 請將 crawl 下來的所有專案捐款明細表，整合成一張大表 df_donation.csv
- 取出必要資訊計算各專案捐款總人數

練習 A-04 (Homework)

- 範例

aid	case.closed	date.funded	date.published	donation	donor	journalist
A0001	已結案	2003/7/28	2003/6/29	15900	22	其他
A0002	已結案	2003/6/17	2003/5/3	108415	37	鋼鐵人
A0004	已結案	2003/5/18	2003/5/8	315	2	小丑女
A0005	已結案	2003/6/6	2003/5/4	33220	27	鋼鐵人
A0007	已結案	2003/5/29	2003/5/7	26965	21	鋼鐵人

comm	fb.lik	fb.sha	fb.tot	imag	n.word	title	url.article	url.detail
0	0	0	0	NA	NA	泰雅災戶盼能重建家園	http://www.appledaily.com.tw	http://searcl
0	0	0	0	1	934	夫肺癌末期妻無力謀生	http://www.appledaily.com.tw	http://searcl
0	0	0	0	1	853	中風殘胞無親人陪伴	http://www.appledaily.com.tw	http://searcl
0	0	0	0	1	1004	阿嬤體弱孫女多重障礙	http://www.appledaily.com.tw	http://searcl
0	0	0	0	0	915	單親啞母術後又罹病	http://www.appledaily.com.tw	http://searcl



Stay Tuned…… We'll be back soon!!

Next session starts at **AM 11:00**

Character encoding problem

- (Mac) 如果 read.csv 讀取中文出現亂碼
  - 透過 Sys.getlocale() 確認 locale 預設語言
  - 設定成英文
    - system("defaults write org.R-project.R force.LANG en_US.UTF-8")
  - 設定成繁體中文
    - system("defaults write org.R-project.R force.LANG zh_TW.UTF-8")
- 或是透過 read.csv 的 parameters 設定
  - fileEncoding = 'UTF-8'

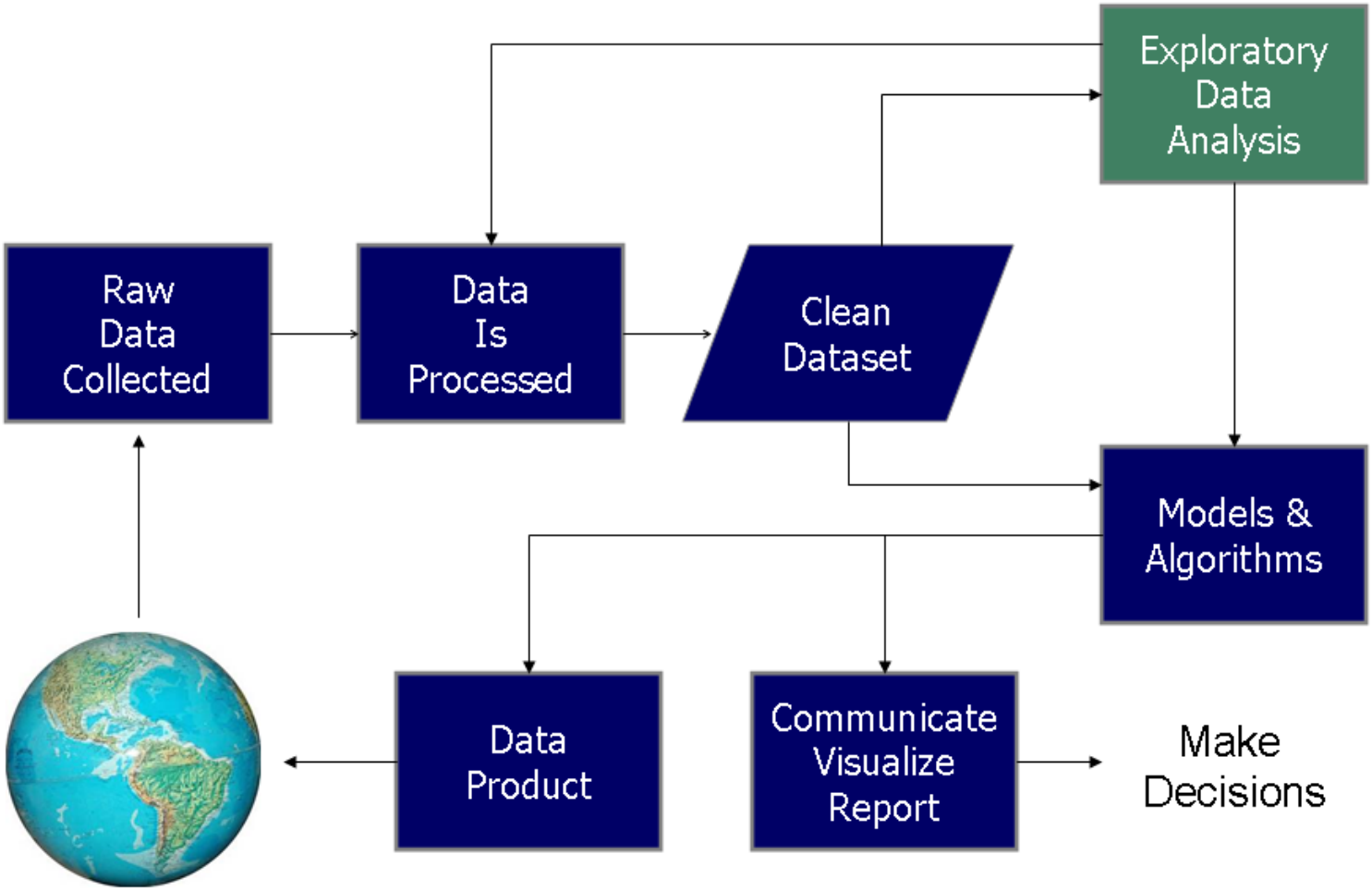
探索式資料分析
Explanatory Data Analysis

Session B

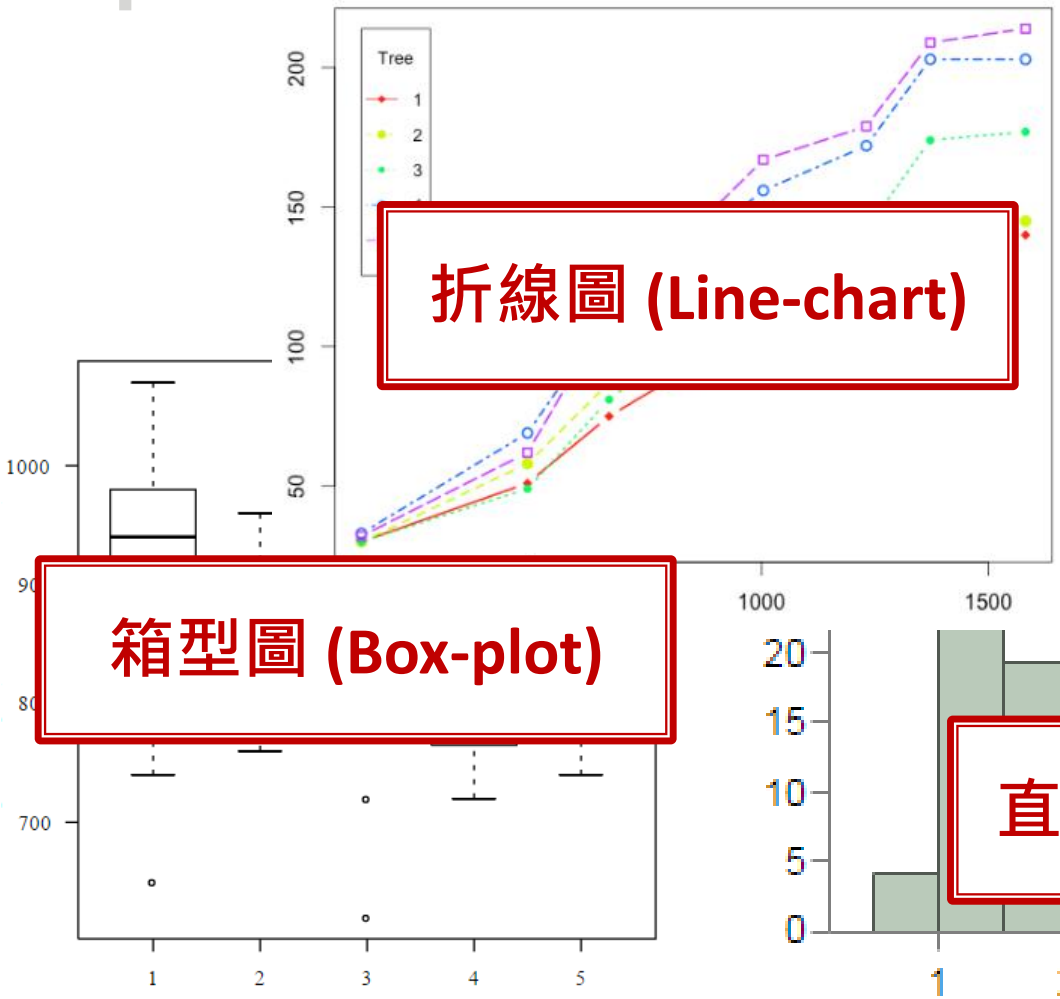
什麼是 EDA？

- EDA 是一種初步分析的方法 (或態度)，主要是透過畫圖的方式，達到三個主要的目的
  - 最大化對資料的了解
  - 找出重要的變數
  - 發現 outliers 或異常數值
- 不做過度假設地從原始數據看出隱含意義

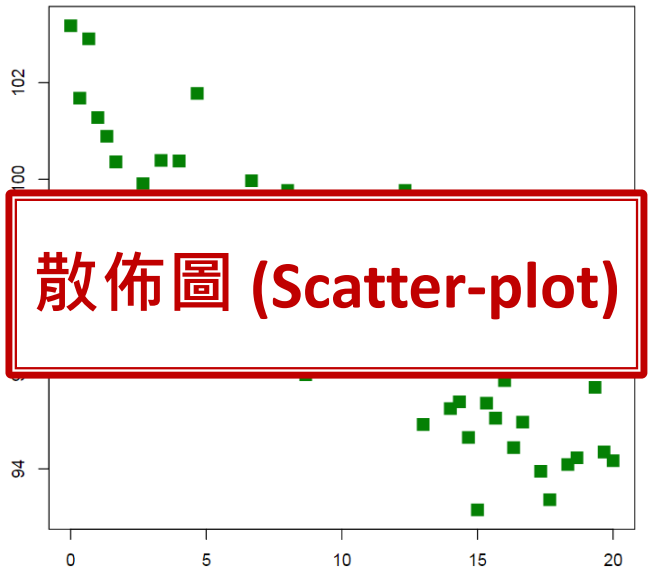
Data Science Process



EDA 常用的視覺化方式

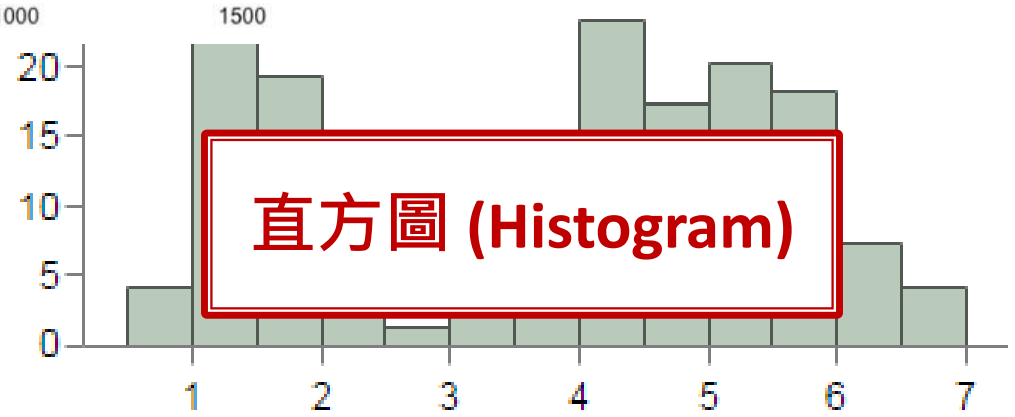


折線圖 (Line-chart)



散佈圖 (Scatter-plot)

箱型圖 (Box-plot)



直方圖 (Histogram)

Summary Functions in R

Function Name	Description
names()	Functions to get or set the names of an object
head(), tail()	Returns the first or last parts of a vector, matrix, table, data frame or function
str()	Compactly display the internal **str**ucture of an **R** object
summary()	Produce result summaries
dim()	Retrieve or set the dimension of an object
length()	Get or set the length of vectors
complete.cases()	Return a logical vector indicating which cases are complete, i.e., have no missing values
as.Data()	Convert between character representations and objects of class "Date" representing calendar dates

Visualization Functions in R

Function Name	Description
plot()	Generic function for plotting of **R** objects
boxplot()	Produce box-and-whisker plot(s) of the given (grouped) values
hist()	Computes a histogram of the given data values
barplot()	Creates a bar plot with vertical or horizontal bars
arrows()	Draw arrows between pairs of points
abline()	a, b: the intercept and slope, single values. y = [A] + [B]x
lines()	Join the corresponding points with line segments.

Function name and parameter 的縮寫解釋：

http://jeromyanglim.blogspot.tw/2010/05/abbreviations-of-r-commands-explained.html

讀入資料與看一看變數

session_B_eda.R

```
# load in apple daily article
> d <- read.csv("df_article.csv", fileEncoding = "UTF-8")

# use dim() to know data frame dimension
> dim(d)
[1]  3784    17

# check the column names
> names(d)
[1]  "aid"           "case.closed"     "circulation"
[4]  "date.funded"   "date.published"  "donation"
[7]  "donor"         "journalist"      "n.fb.comment"
[10] "n.fb.like"     "n.fb.share"      "n.fb.total"
[13] "n.image"       "n.word"          "title"
[16] "url.article"   "url.detail"
```

利用 str() 迅速了解資料格式

```
# use str() to have a brief data summary
> str(d)
'data.frame':   3750 obs. of  251 variables:
 $ aid              : Factor w/ 3759 levels "A0001","A0002",..: 2 3
 $ case.closed      : Factor w/ 1 level "已結案": 1 1 1 1 1 1 1 1 1 1
 $ circulation      : int  NA NA NA NA NA NA NA NA NA NA ...
 $ date.funded      : Factor w/ 1603 levels "2003-05-18","2003-05-2
 $ date.published   : Factor w/ 3718 levels "2003-05-03","2003-05-0
 $ donation         : int  108415 315 33220 26965 143200 38015 6565
 $ donor            : int  37 2 27 21 64 21 66 65 59 65 ...
 $ journalist       : Factor w/ 14 levels "小丑女","白皇后",..: 14 1 1
 $ n.fb.comment     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ n.fb.like        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ n.fb.share       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ n.fb.total       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ n.image          : int  1 1 1 0 1 1 1 1 1 1 ...
 $ n.word.x         : int  934 853 1004 915 912 890 879 1093 734 10
 $ title            : Factor w/ 3759 levels "「不怕死 只怕孩子沒人顧」"
```

將欄位轉至適合的格式

```
> d$date.published <- as.Date(d$date.published)
> d$title <- as.character(d$title)
```

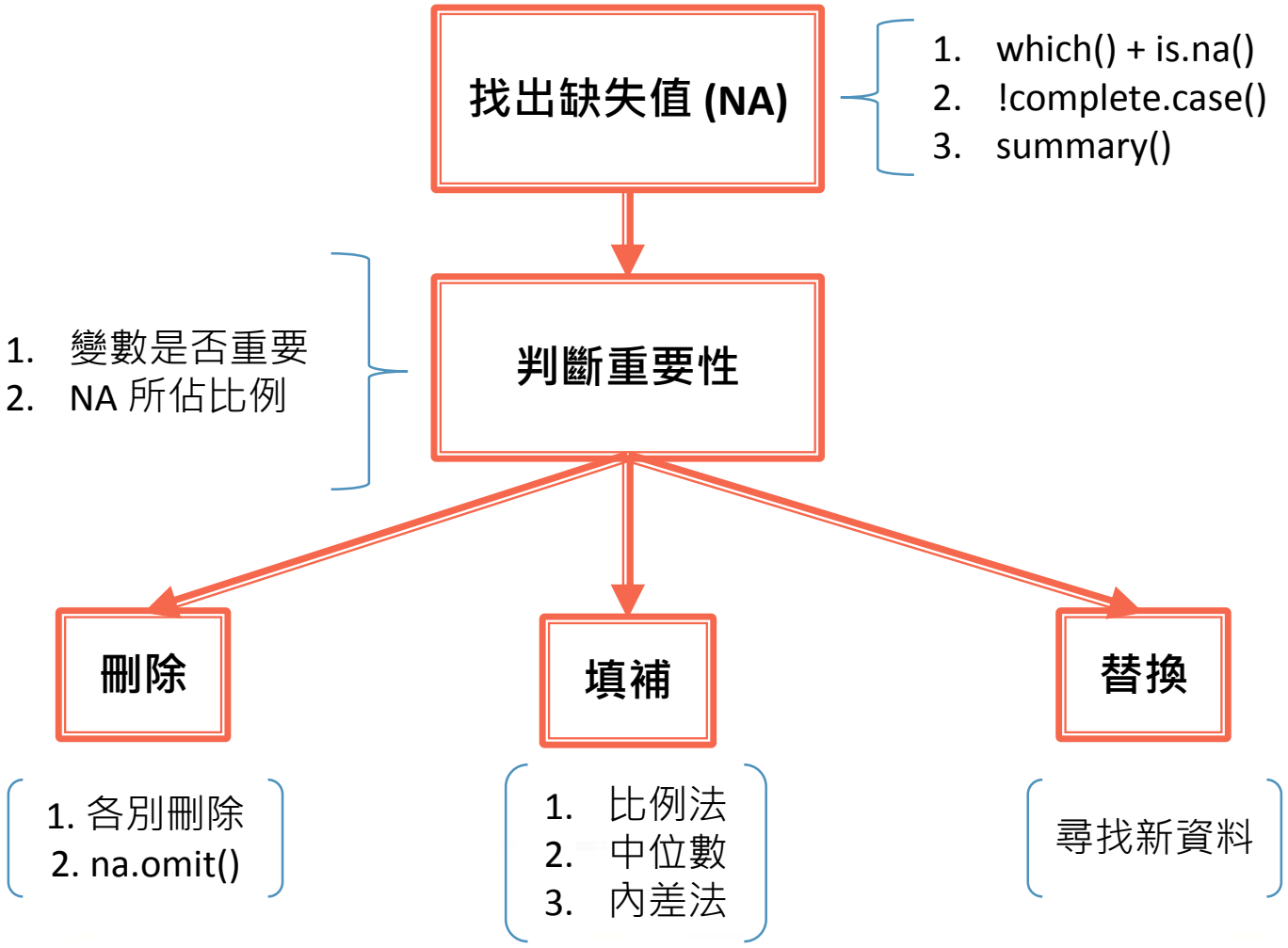
用 summary() 找出 NA

```
> summary(d)
```

```
     aid        case.closed    circulation          date.funded      date.published     donation       
A0001  :   1   已結案:3763   Min.   :355000   2016/06/14:  27   2003/08/23:   3   Min.   :     100  
A0002  :   1   未結案:  21   1st Qu.:456000   2016/01/08:  20   2003/10/03:   3   1st Qu.:  258856  
A0004  :   1                 Median :488000    2003/10/15:   9   2004/12/30:   3   Median :  408110  
A0005  :   1                 Mean   :482723    2010/03/15:   9   2003/05/08:   2   Mean   :  461454  
A0007  :   1                 3rd Qu.:508000    2003/09/11:   8   2003/05/10:   2   3rd Qu.:  611459  
A0008  :   1                 Max.   :750000    (Other)   :3693   2003/05/18:   2   Max.   :46537707  
(Other):3778                 NA's   :2209      NA's      :  18   (Other)   :3769   NA's   :1         
```

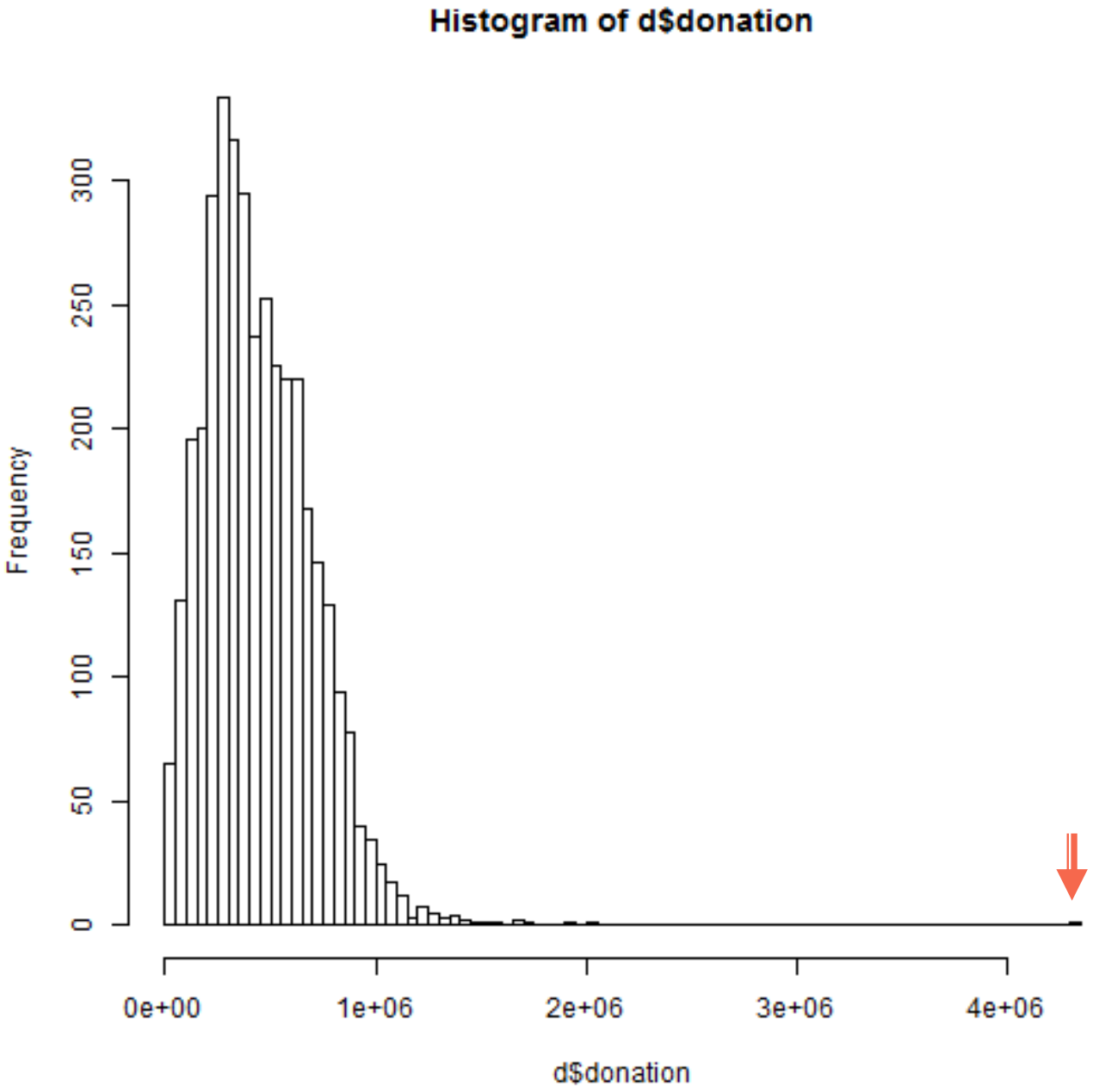
```
  n.fb.like          n.fb.share         n.fb.total          n.image          n.word      
Min.   :    0.00   Min.   :   0.000   Min.   :    0.00   Min.   :0.00   Min.   : 300.0  
1st Qu.:    0.00   1st Qu.:   0.000   1st Qu.:    0.00   1st Qu.:1.00   1st Qu.: 794.0  
Median :    0.00   Median :   0.000   Median :    0.00   Median :1.00   Median : 866.0  
Mean   :   35.34   Mean   :   5.744   Mean   :   50.18   Mean   :1.49   Mean   : 892.7  
3rd Qu.:    3.00   3rd Qu.:   3.000   3rd Qu.:    8.00   3rd Qu.:2.00   3rd Qu.: 951.0  
Max.   :26729.00   Max.   :2985.000   Max.   :30898.00   Max.   :6.00   Max.   :1858.0  
                                                          NA's   :27     NA's   :27      
```

缺失值 (NA) 處理


找出缺失值 (NA)
1. which() + is.na()
2. !complete.case()
3. summary()
判斷重要性
1. 變數是否重要
2. NA 所佔比例
刪除
1. 各別刪除
2. na.omit()
填補
1. 比例法
2. 中位數
3. 內差法
替換
尋找新資料

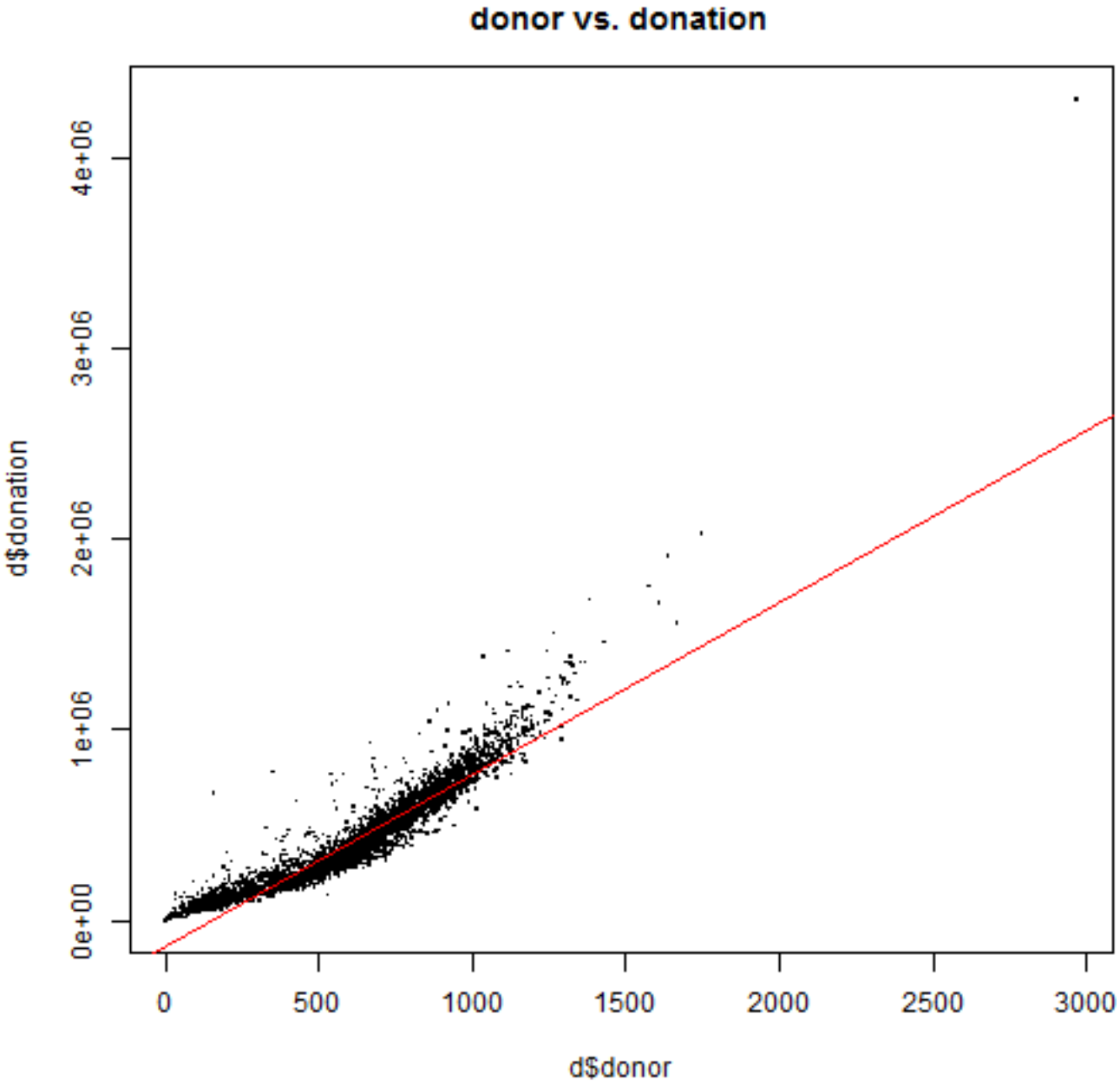
hist()

```
# use hist() to check donation distribution
> hist(d$donation, br = 100)
```


Histogram of d$donation
Frequency
0
50
100
150
200
250
300
0e+00
1e+06
2e+06
3e+06
4e+06
d$donation

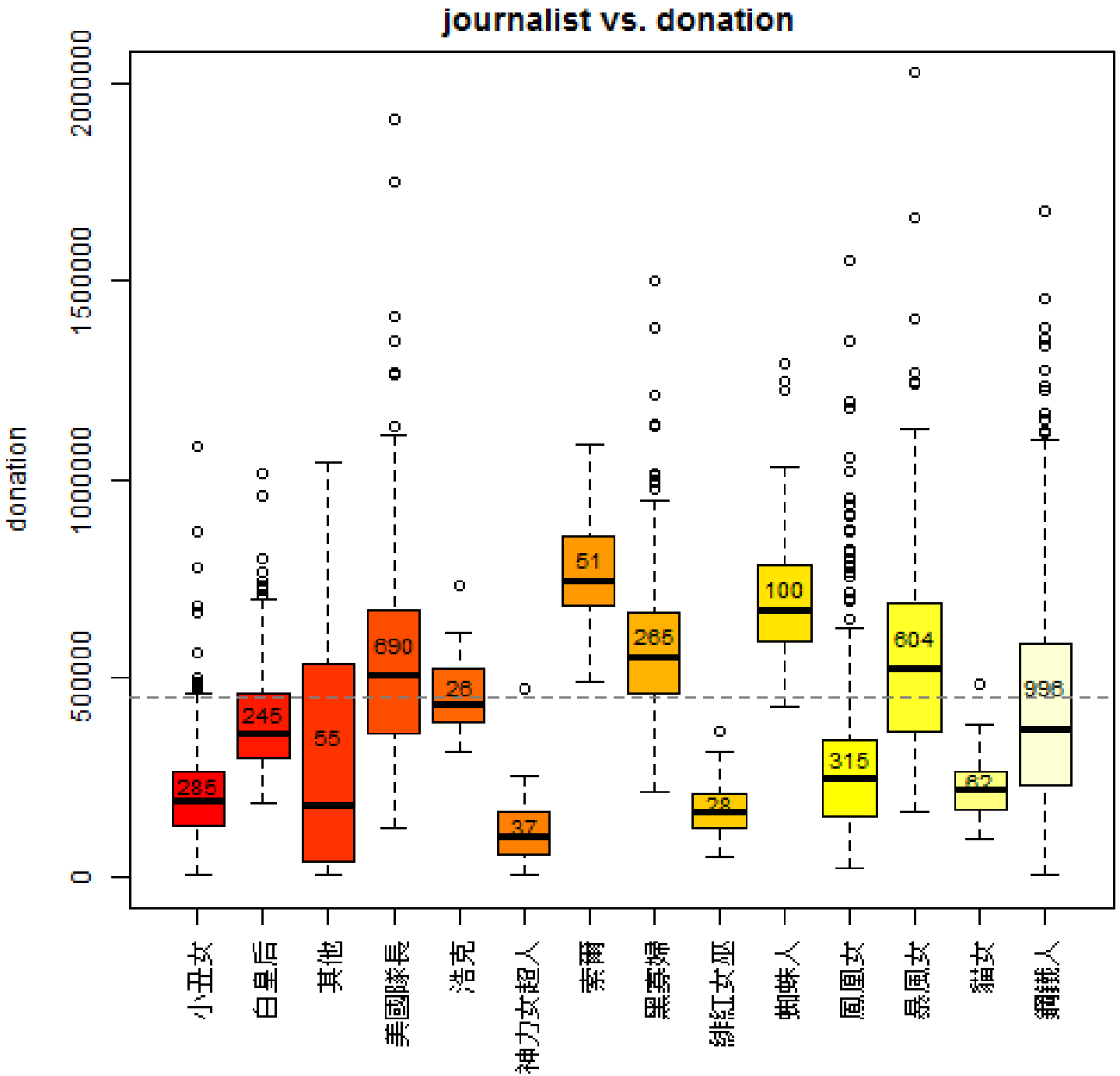
plot()

```
> plot(d$donor, d$donation, pch = '.', cex = 2)
> abline(lm(d$donation ~d$donor), col = 'red')
```


donor vs. donation
d$donation
0e+00
1e+06
2e+06
3e+06
4e+06
0
500
1000
1500
2000
2500
3000
d$donor

boxplot()

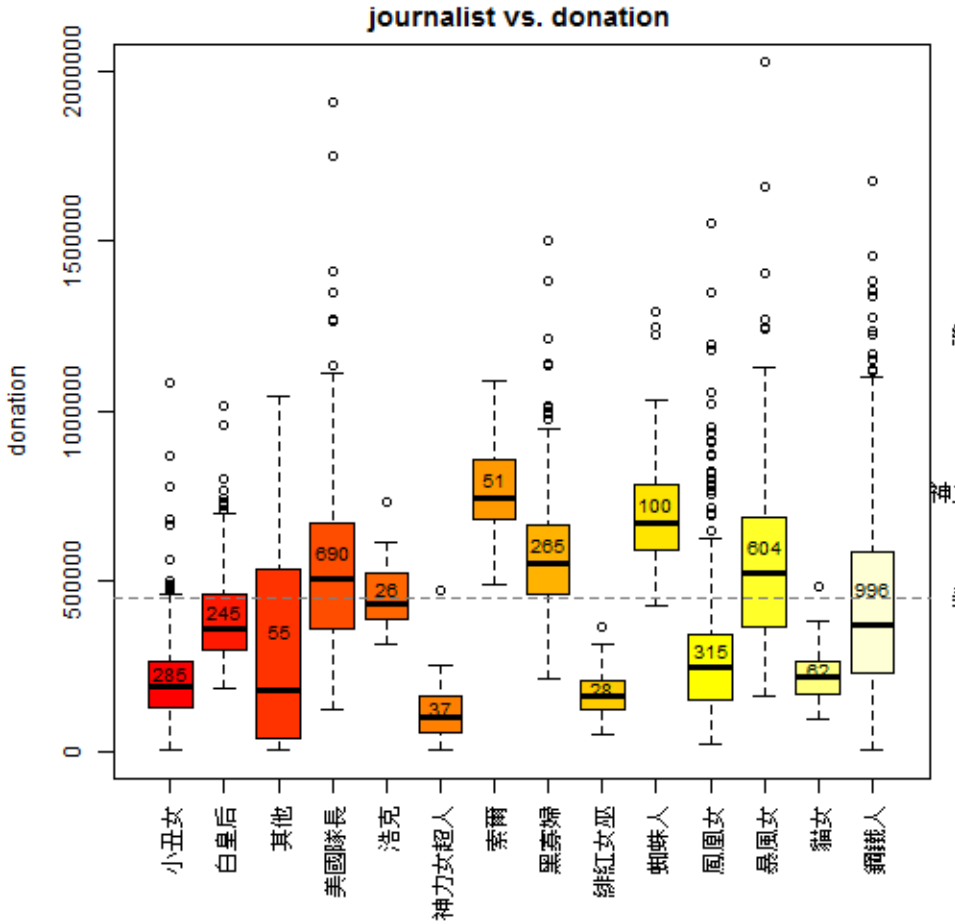
```
> n <- length(unique(d$journalist))
> b <- boxplot(d$donation ~ d$journalist, col =
heat.colors(n), las = 2, ylim = c(0,2e6))
> abline(h = mean(d$donation), lty = 2, cex = 2)
> text(1:n, (b$stats[3,]+b$stats[4,])/2, b$n, cex =
0.8)
```

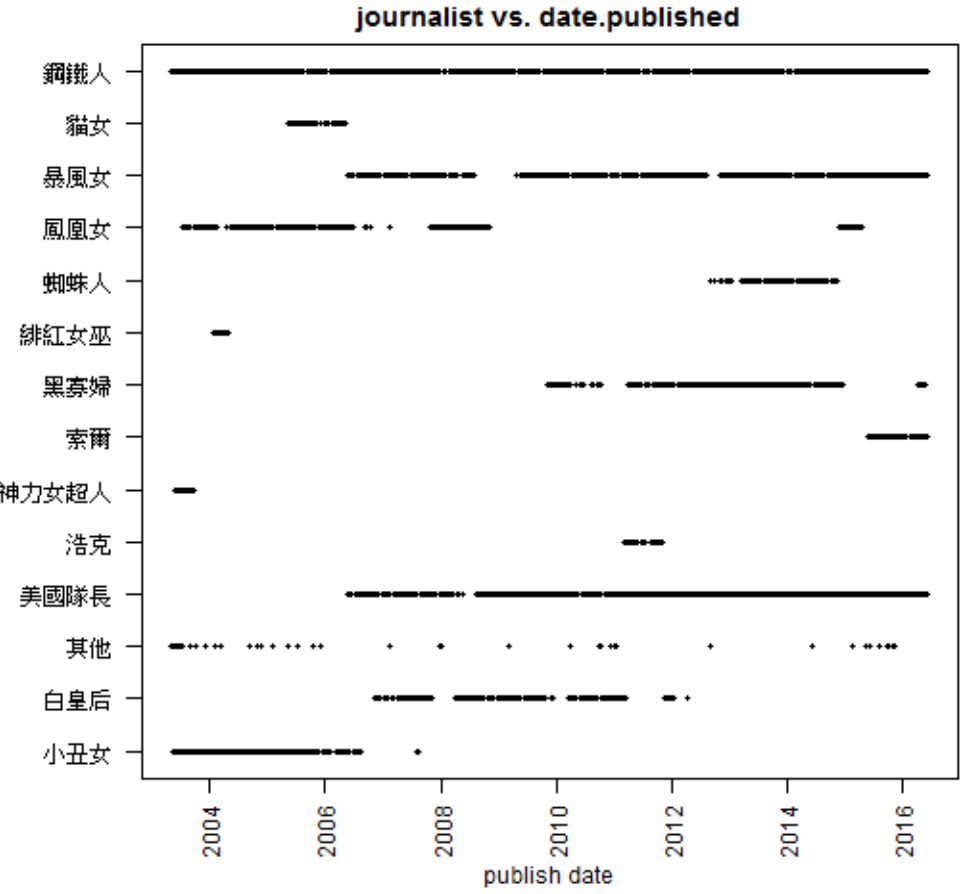

journalist vs. donation
donation
0
500000
1000000
1500000
2000000
285
245
55
690
26
37
51
265
28
100
315
604
62
996
小丑女
白皇后
其他
美國隊長
浩克
神力女超人
索爾
黑寡婦
緋紅女巫
蜘蛛人
鳳凰女
暴風女
貓女
鋼鐵人

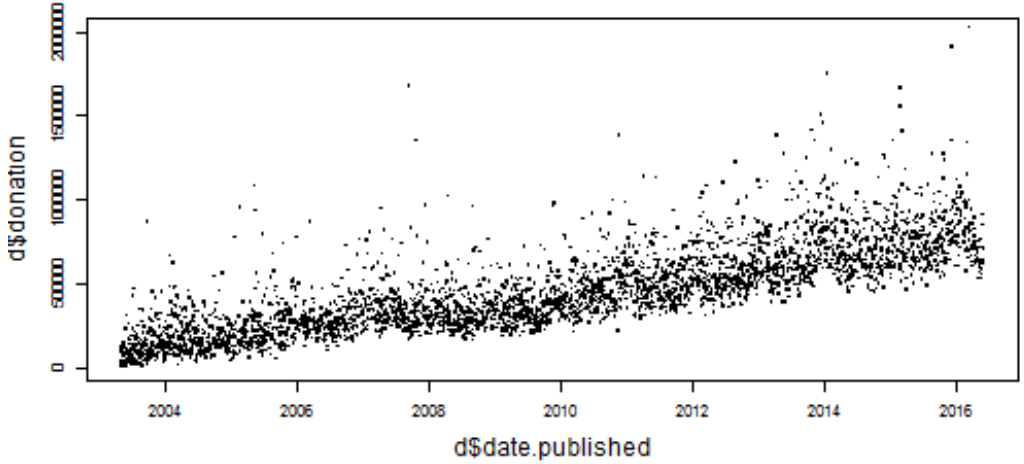
練習 B-01 (10 mins)

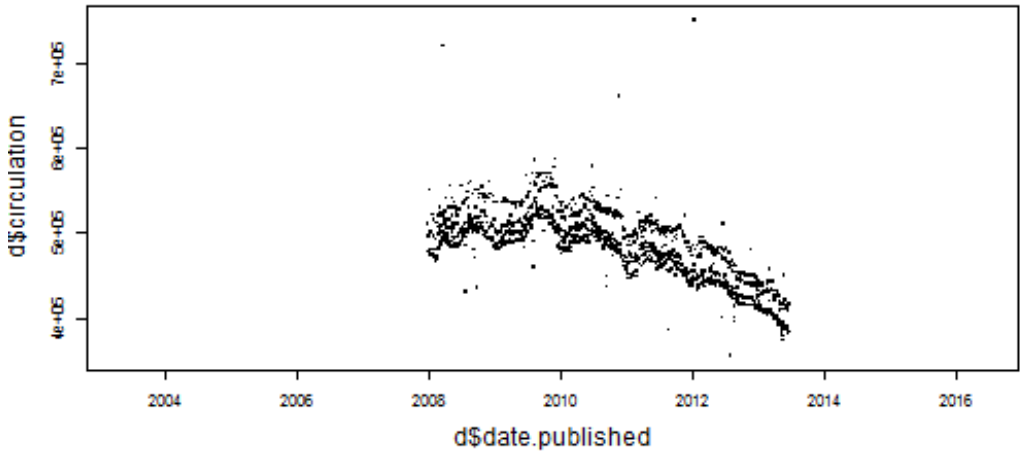
- 處理 NA 值
  - 將 n.word 的 NA 值補以平均數
  - 將 n.image 的 NA 值補以眾數
- 利用 hist() 觀察變數分佈
- 請找出 journalist 表現差異的原因
  - 試著運用 plot(), boxplot() 觀察變數間關係

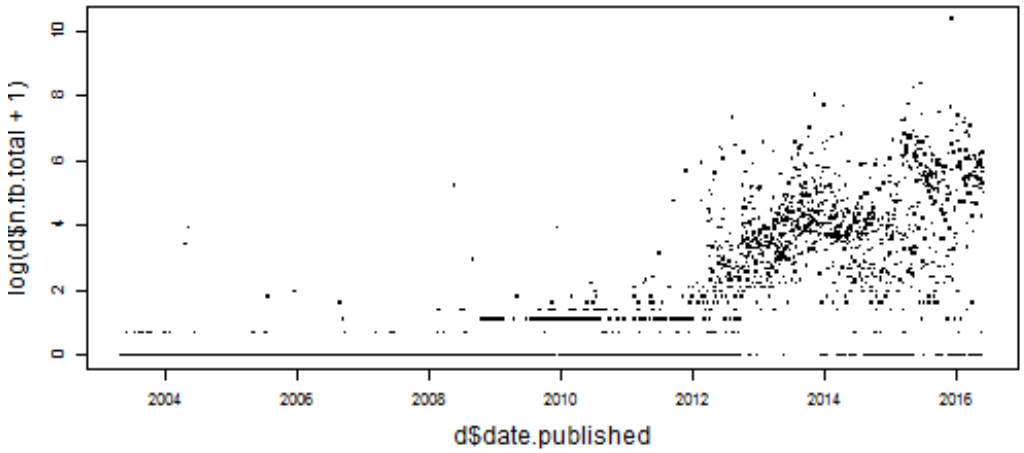
練習 B-01 解答


journalist vs. donation
donation
0
500000
1000000
1500000
2000000
285
245
55
690
26
37
51
265
28
100
315
604
82
996
小丑女
白皇后
其他
美國隊長
浩克
神力女超人
索爾
黑寡婦
緋紅女巫
蜘蛛人
鳳凰女
暴風女
貓女
鋼鐵人


journalist vs. date.published
鋼鐵人
貓女
暴風女
鳳凰女
蜘蛛人
緋紅女巫
黑寡婦
索爾
神力女超人
浩克
美國隊長
其他
白皇后
小丑女
2004
2006
2008
2010
2012
2014
2016
publish date


d$donation
0
500000
1000000
1500000
2000000
2004
2006
2008
2010
2012
2014
2016
d$date.published


d$circulation
4e+05
5e+05
6e+05
7e+05
2004
2006
2008
2010
2012
2014
2016
d$date.published

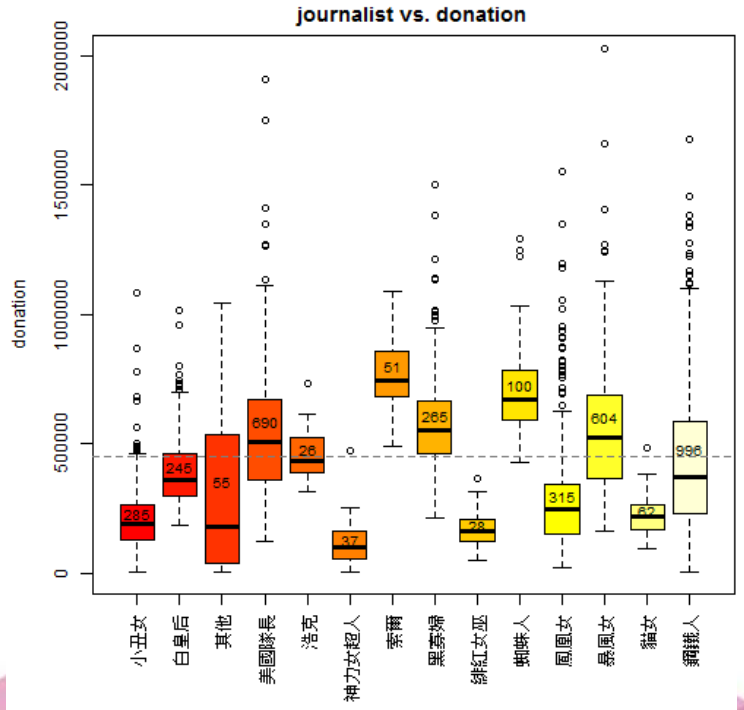

log(d$n.fb.total + 1)
0
2
4
6
8
10
2004
2006
2008
2010
2012
2014
2016
d$date.published

A FINDING!

- 蘋果日報可能被網路媒體影響
  - 捐款金額隨時間上升
  - 發行量隨時間下降
  - FB 總量隨時間上升

此時間趨勢可能造成...

- 誤解變數的影響力或遮蔽變數的效果
  - 以為索爾很厲害，寫出了高質量文章...
  - 但可能只是因為加入的時間比較好？



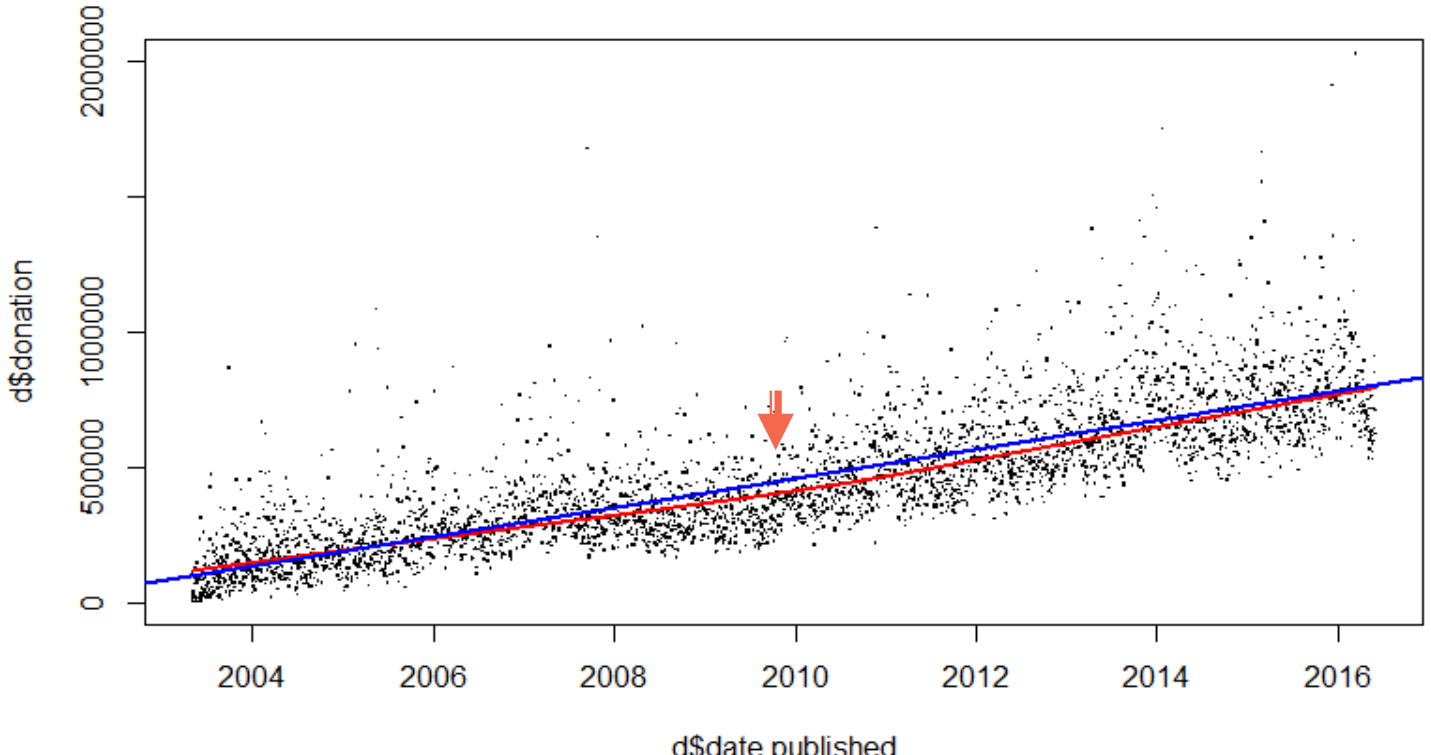
時間序列資料的 detrending

- 透過統計或數學的操作，移除時間上的趨勢，才能夠看清楚索爾的真正實力 (或許貓女比較強 ?)

- 基本的方法
  - 線性回歸
  - LOWESS 局部加權散點平滑法
    - Locally weighted scatterplot smoothing
    - 取一定比例的局部樣本做多項式回歸曲線
    - 查看二維變量之間關係的有力工具

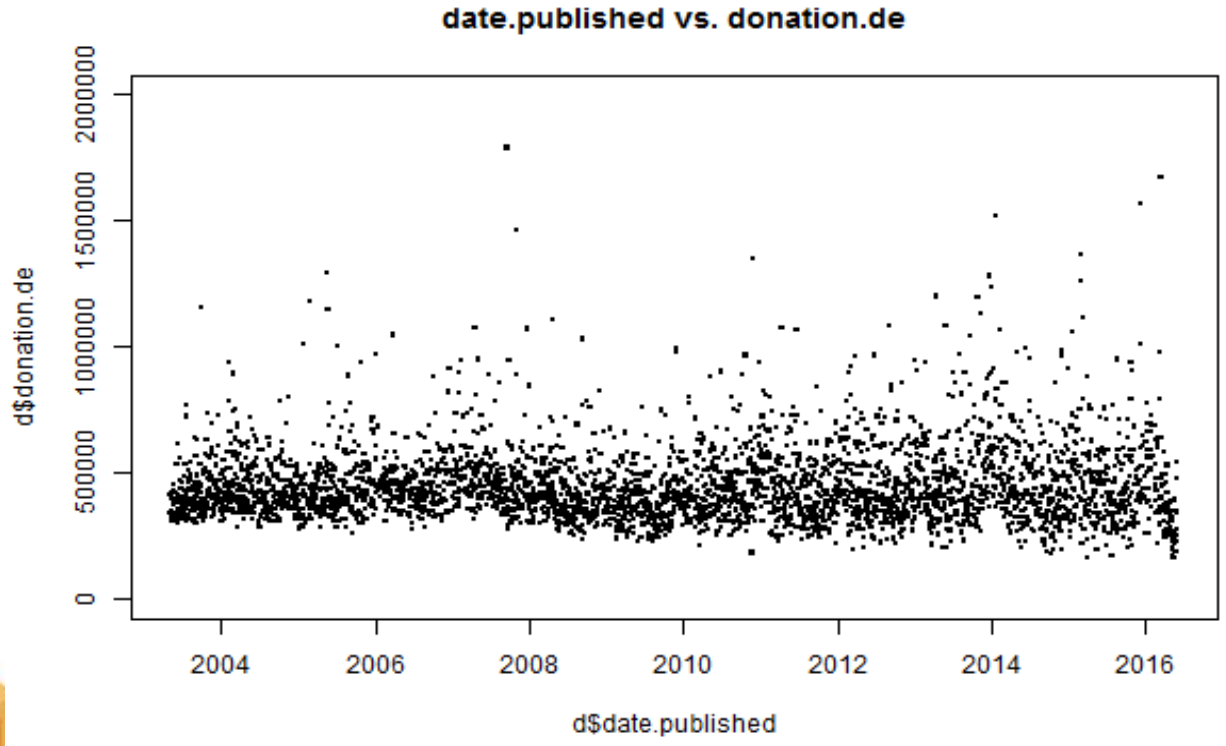
lm 與 lowess

```
# 用 lm() 與 lowess() 看 donation vs date.published
> plot(d$published, d$donation, pch = '.', cex = 3)
> lines(lowess(d$published, d$donation), col = 'red')
> abline(lm(d$published, d$donation), col = 'blue')
```



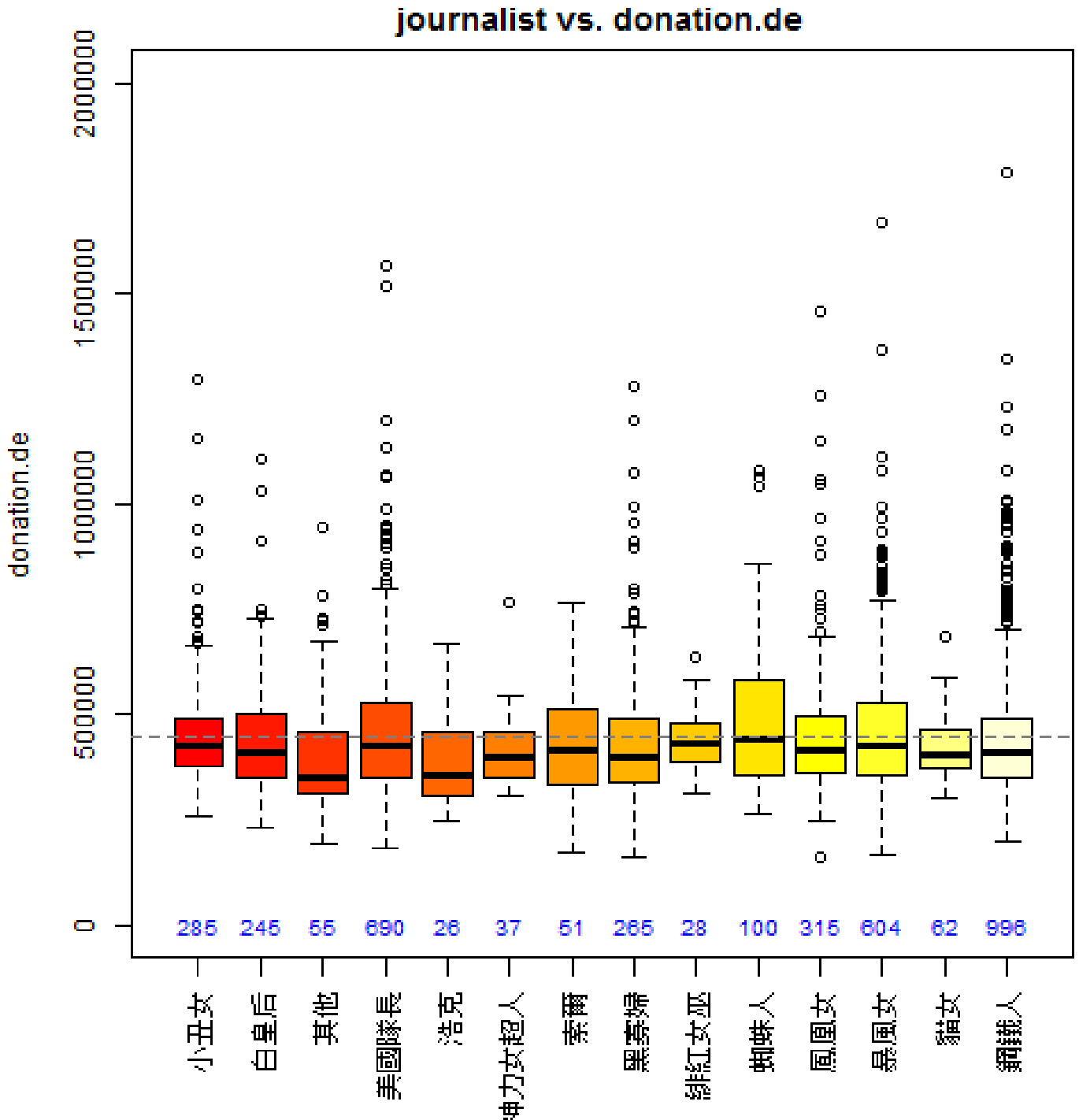
利用 lowess() 做 detrending

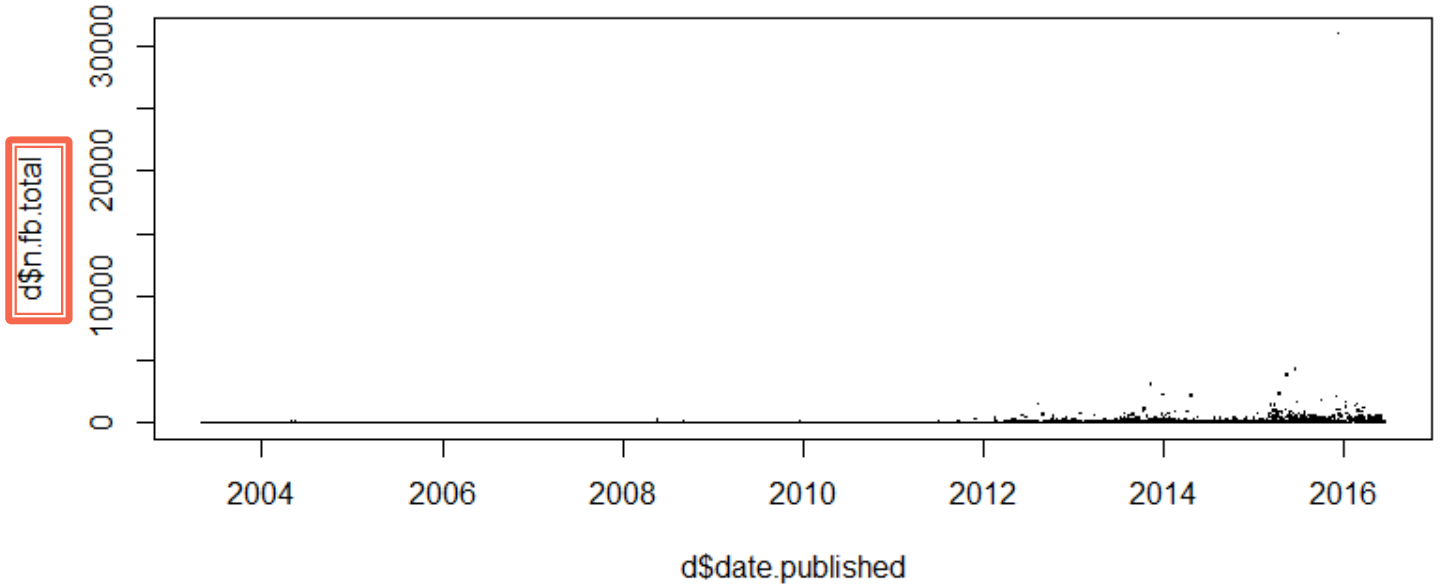
```
# 利用 R 內建的 lowess() 做 detrending
> l <- lowess(d$donation ~ d$date.published)
> d$donation.de <- d$donation - l$y + mean(l$y)
> plot(d$published, d$donation.de, pch = '.', cex = 3)
```

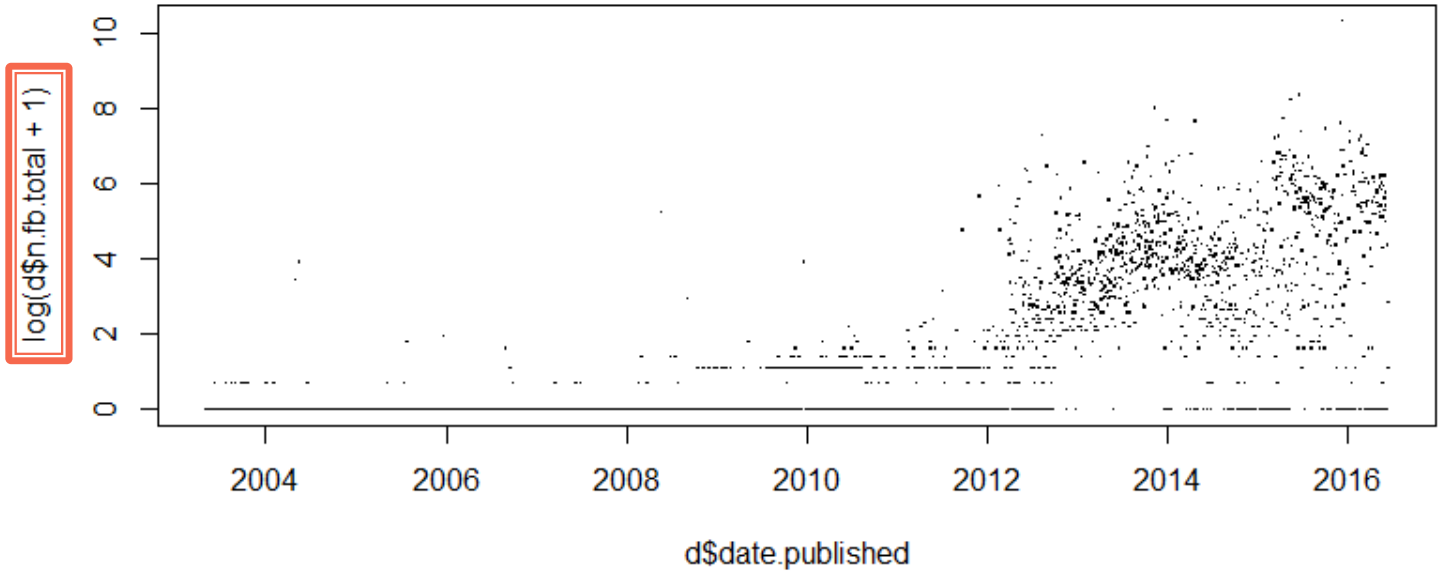


重新看一次 journalist 的表現

```
> n <- len(unique(d$journalist))
> b <- boxplot(d$donation.de ~ d$journalist, col =
    heat.colors(n), las = 3, ylim = c(0, 1e6))
> abline(h = mean(d$donation.de), lty = 2, cex = 2)
> title('journalist vs. donation.de', ylab =
'donation.de')
> text(1:n, 0, b$n, cex = 0.8, col = 'blue')
```


journalist vs. donation.de
donation.de
0
500000
1000000
1500000
2000000
285
245
55
690
26
37
51
265
28
100
315
604
62
996
小丑女
白皇后
其他
美國隊長
浩克
神力女超人
索爾
黑寡婦
绯红女巫
蜘蛛人
鳳凰女
暴風女
貓女
鋼鐵人


d$n.fb.total
0
10000
20000
30000
2004
2006
2008
2010
2012
2014
2016
d$date.published

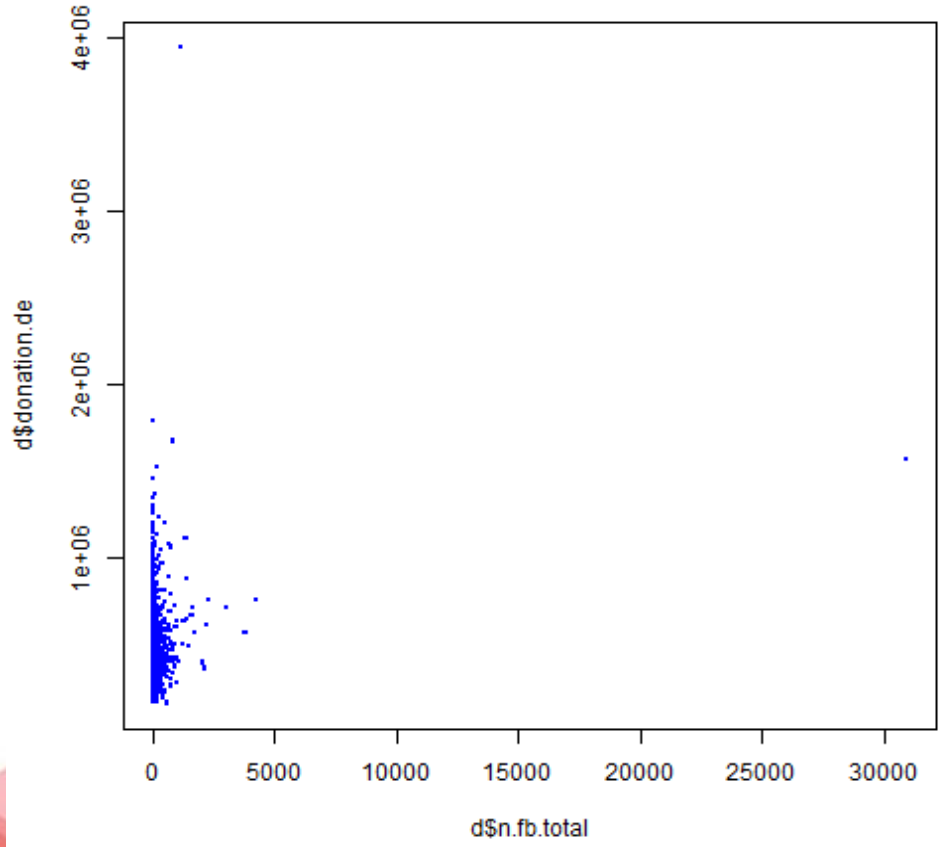

log(d$n.fb.total + 1)
0
2
4
6
8
10
2004
2006
2008
2010
2012
2014
2016
d$date.published

Log Transformation

- 目的
  - 更清楚地看出變數間的關聯性或增加解釋性
  - 使資料變得更線性
- 其他常用的變形
  - 1/x, x^2, sqrt(x) …

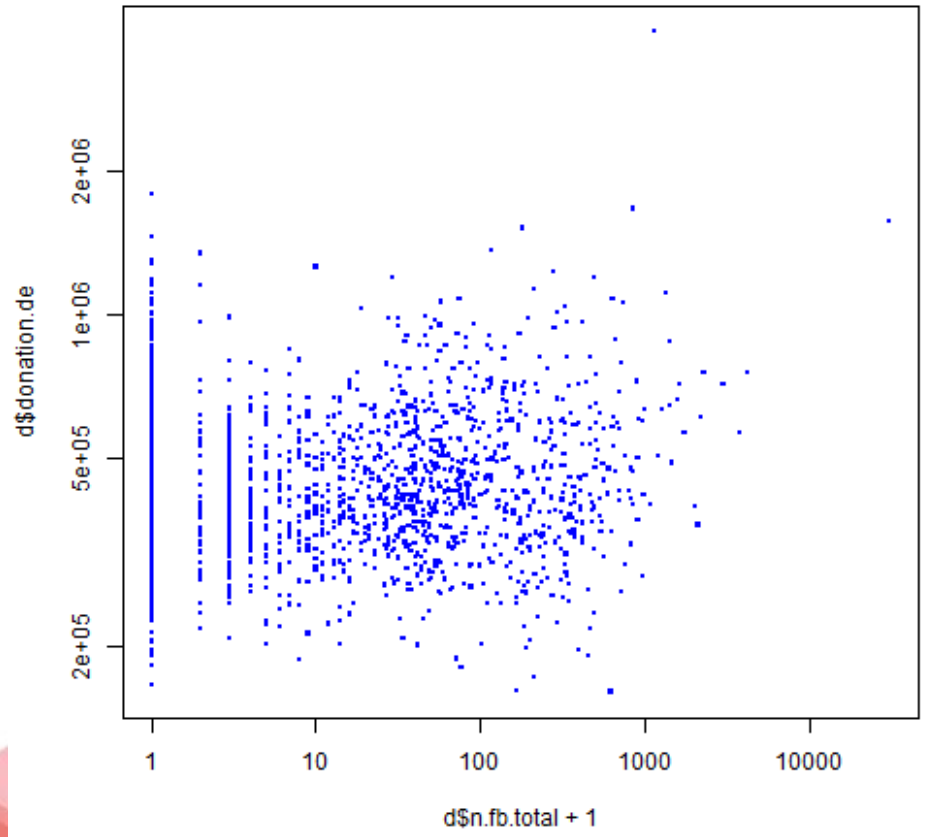
log()

```
# 未使用 log() 變形前
> plot(d$n.fb.total, d$donation.de, col = 'blue')
```



log()

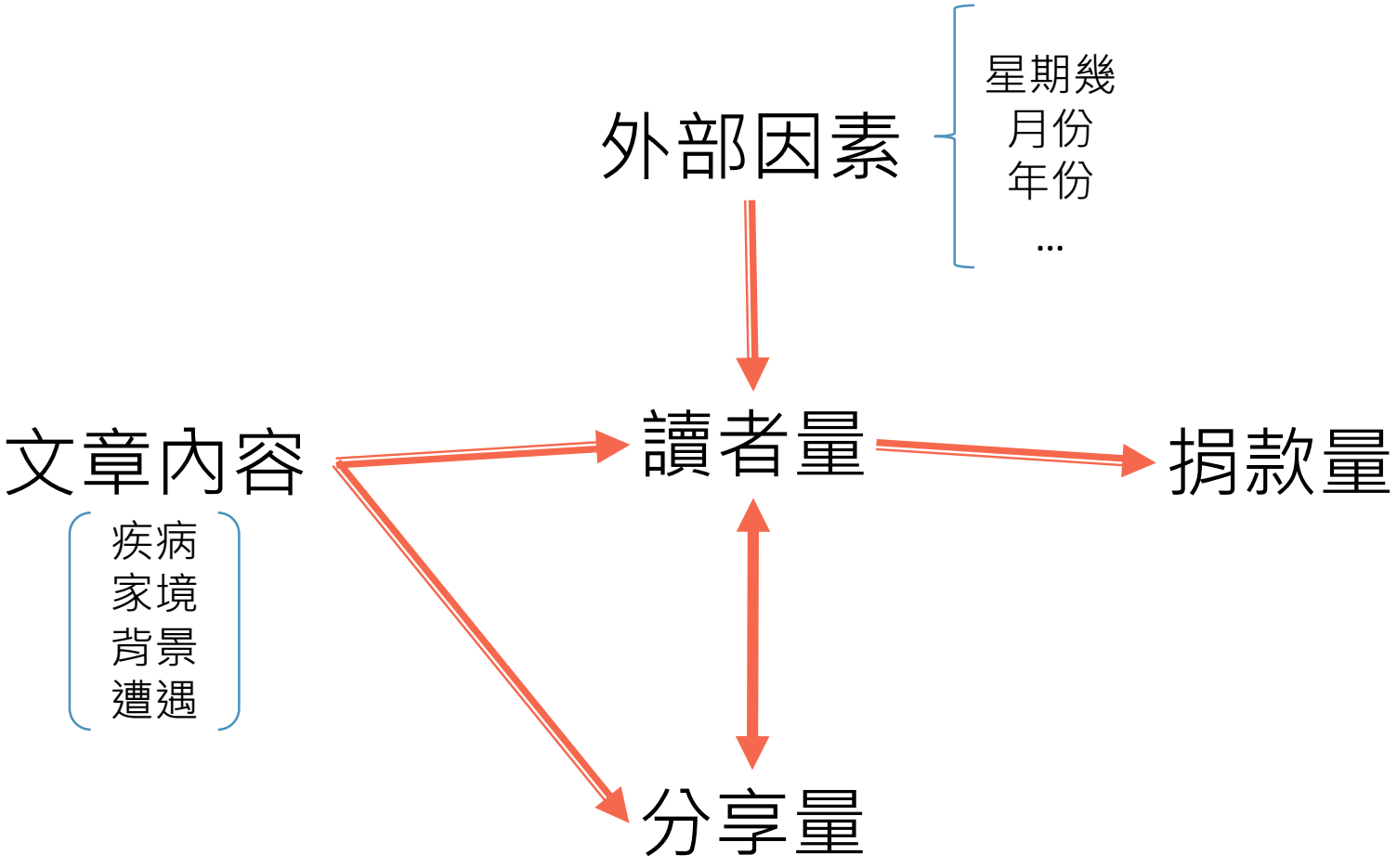
```
> plot(log(d$n.fb.total + 1), log(d$donation.de))
> plot(d$n.fb.total + 1, d$donation.de, log = 'xy')
```



變數看起來有點少...

- 爬新的資料
- 或 ...

想一想當初...

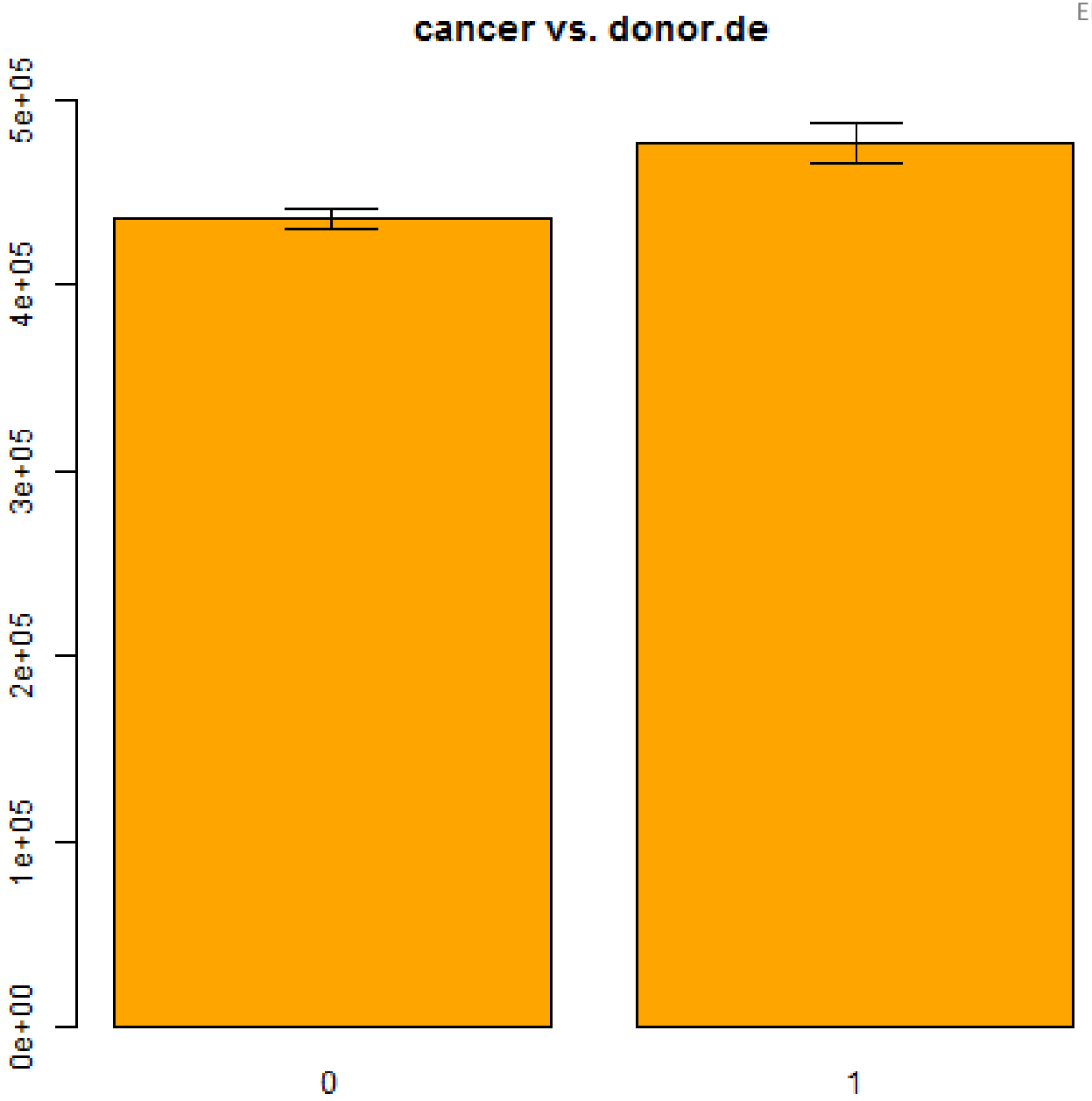


試著從文章標題創造新變數

```
> head(d$title)
[1] "泰雅災戶盼能重建家園"  "夫肺癌末期妻無力謀生"
[3] "中風殘胞無親人陪伴"    "阿嬤體弱孫女多重障礙"
[5] "單親啞母術後又罹病"    "跑船男烏腳病鋸雙腿"

> d$tle.n.word <- nchar(d$title)
```

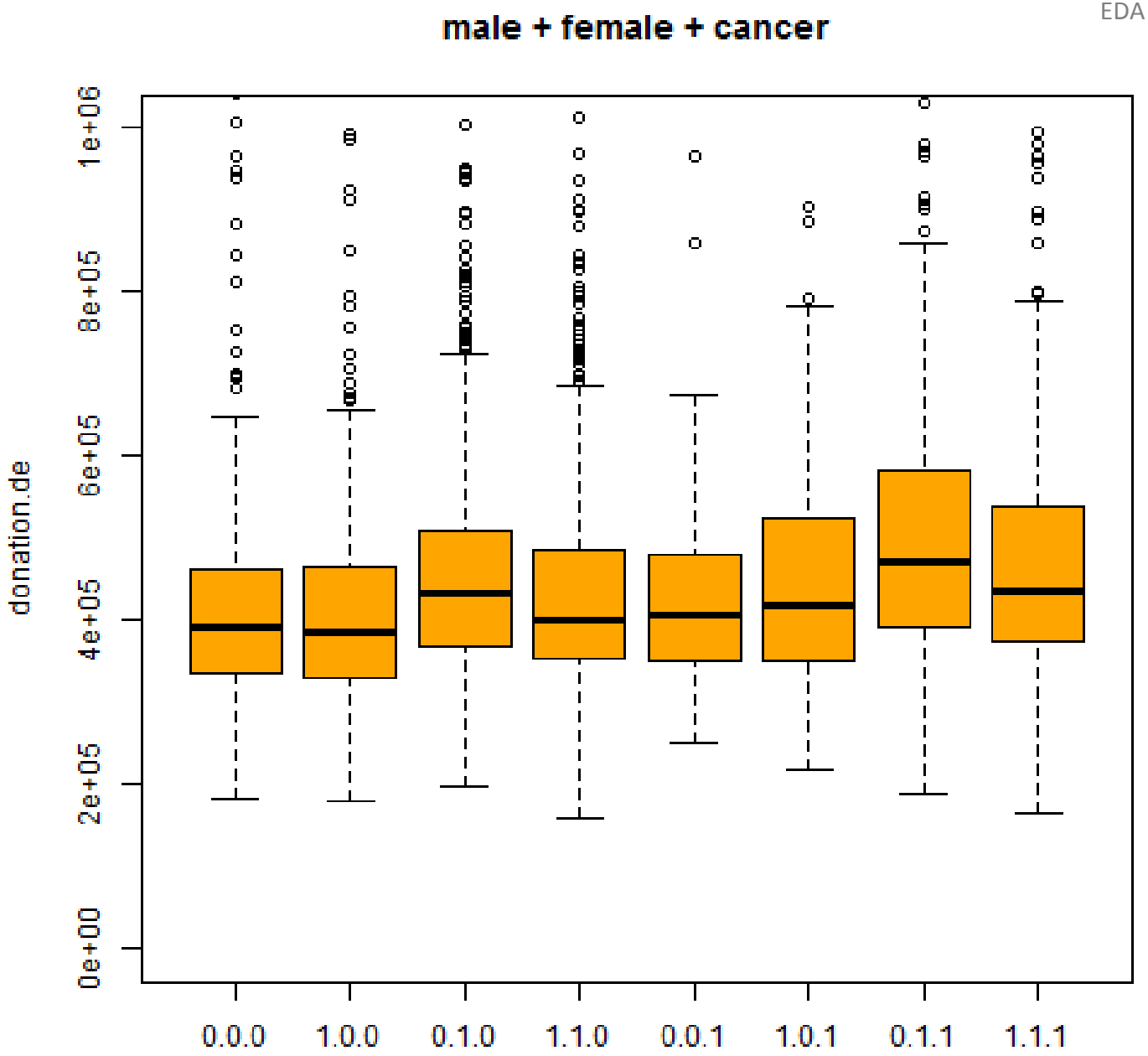
找出標題內的資訊

```
> i = grep("癌", d$title)
> d$tle.cancer <- 0
> d[i, ]$tle.cancer <- 1

> library(dplyr)
> tmp = group_by(d, tle.cancer)
> tmp = summarize(tmp, se = sd(donation.de) /
sqrt(n()), m = mean(donation.de))

> b <- barplot(tmp$m, names.arg =  tmp$tle.cancer, col
= 'orange', ylim = c(0, 5e5))
> arrows(b, tmp$m + tmp$se * 1.96, b, tmp$m - tmp$se *
1.96, angle = 90, code = 3)
```


cancer vs. donor.de
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
0
1

看男、女與癌的差異

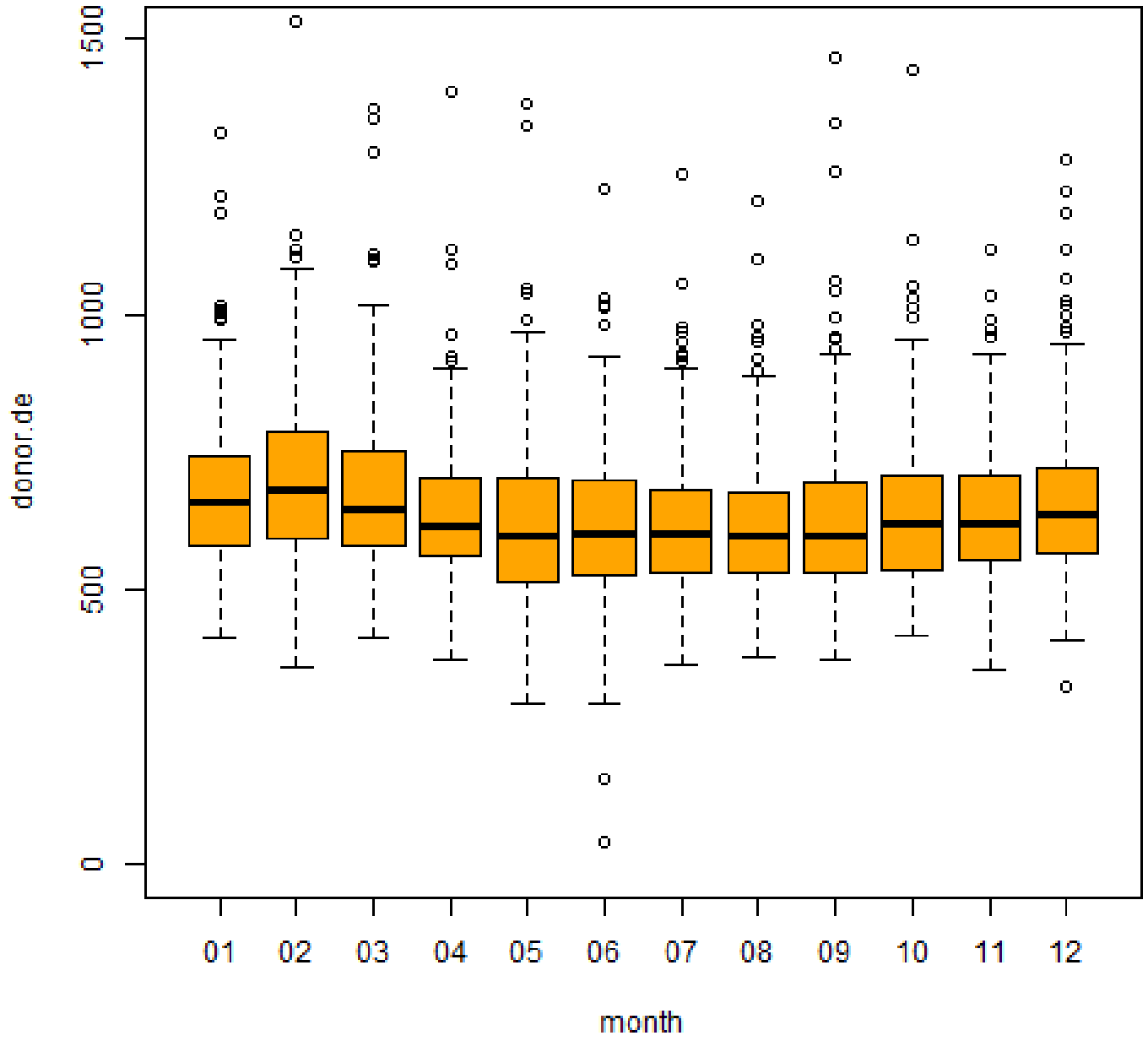
```
> i <- grep('夫|父|男|翁|公|爸|漢', d$title)
> d$tle.male <- 0
> d[i, ]$tle.male <- 1

> i <- grep('妻|母|女|婆|嬤|媽|婦', d$title)
> d$tle.female <- 0
> d[i, ]$tle.female <- 1

> boxplot(donation.de ~ tle.male + tle.female +
tle.cancer, data = d, col = ‘orange’)
> title(‘male + female + cancer’)
```


male + female + cancer
donation.de
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
0.0.0
1.0.0
0.1.0
1.1.0
0.0.1
1.0.1
0.1.1
1.1.1

找出時間上的差異

```
> class(d$date.published)
[1] "Date"

> d$month <- format(d$date.published, '%m')
> boxplot(d$donor.de ~ d$month, col = 'orange')
```

%y = year
%m = month
%w = week
%d = day


donor.de
0
500
1000
1500
01
02
03
04
05
06
07
08
09
10
11
12
month

各種新變數

- 平均每人捐款金額 = donation / donor
- 時間
  - 星期幾, 每月第幾周, 第幾個月, 哪一年.
- FB
  - # of TOTAL (likes, shares, comments) in logarithm
- 標題
  - 狀況 : 癌, 病, 火, 殘障, 男性, 女性, 男女性, 長輩, 孩童, 哀傷情緒, 視覺障礙, 單親, 刻苦, 死亡, 腎, …
  - 結構 : 「 」, 0-9, 歲數 …

練習 B-02, 03

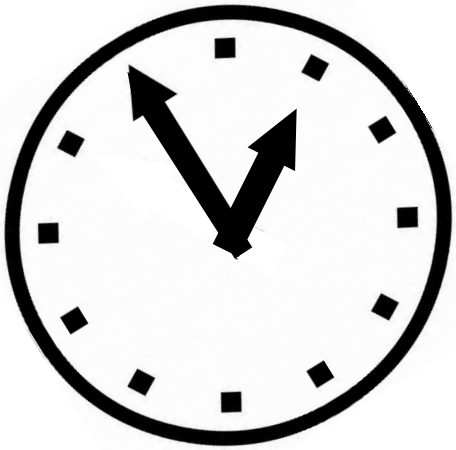
- Detrending
  - 畫出 date vs. donor 關係圖、用 lm, lowess 表現趨勢
  - 用 lowess 或 lm detrend donor
- 觀察資料特徵並創造新變數
  - 利用 donation & donor 創造新變數
  - 創造至少 10 個新的標題相關變數
  - 創造至少 3 個時間相關變數
- 找到彼此之間擁有最高相關性的變數組合
- 並且有合理的解釋

從文章標題看
捐款人對於被捐款者的傾向

- 負相關
  - ttl.male, ttl.heart, ttl.fire, ttl.brain, ttl.disable, ttl.liver
- 正相關
  - ttl.youth, ttl.death, ttl.female, ttl.cancer, ttl.have.age, ttl.single.p, ttl.blind

EDA 總結

- 基本函數運用與視覺化
- 時間序列的 detrending
- 資料變形 (data transformation)
- 創造新變數 (variable creation)



Stay Tuned…… We'll be back soon!!

Next session starts at **13:50**

90 分鐘的文字礦工
Text Mining in 90 minutes

Session C

A3887 3胞胎待哺 單親爸抗癌魔

2016年03月23日 傳送 讚 1,148 G+1 2 我要捐款 更多專欄文章

曾柏達突罹癌憂心家計，盼恢復健康看3胞胎長大。

40歲曾柏達去年此時離婚，成了帶3個1歲5個月大的三胞胎單親爸，忙跑計程車掙奶粉錢，怎料去年底他突罹肝癌，面對難料病況，他嘆：「一切都出乎預料。」

報導•攝影／向高彬

來到曾柏達家，40歲曾柏達剛手術不久，右腹仍掛著引流袋，他忙哄哭鬧娃兒，「乖，奶奶泡牛奶了。」67歲奶奶阿蘭婆將奶瓶塞進娃兒嘴裡，才安靜咕嚕喝奶。

單親癌爸曾柏達（左）右腹掛著引流袋，最擔憂「不知能陪孩子到哪天？」

「不知能陪多久」

曾柏達憂喜參半地說，身體病痛可忍，「心裡焦慮，靠孩子笑容安慰，只是我最擔憂的，也是他們，不知道能陪孩子到哪一天。」曾柏達去年底右腹疼痛，就醫查出12公分腫瘤、肝癌3期，醫師評估是B肝惡化，手術切除三分之二肝臟，現做化療療程。

曾柏達回述，人生起伏變化超乎他預料，早年曾開店生意失敗欠債，後來轉做汽車業務、但公司倒了，後來他受僱跑遊覽車接陸客觀光團時，在特產店認識現23歲的孩子生母，3年前先有後婚，「起先產檢是雙胞胎，覺得欣喜，4個月時才知是3胞胎，當時曾考慮減胎，但顧慮母嬰安全，仍決定生下，孩子7個月早產，住了2個月保溫箱。」在娃兒出生後4個月，兩人因經濟等問題爭執離異，3娃歸他撫養。曾柏達說，老大、老三是龍鳳胎，老二剛出生時最瘦小僅900克，早產兒狀況多，眼和肺都不好。

他又說，原本每天跑12小時，扣油錢車貸剩約2萬多元，加上每月2款低收補助共21700元，勉敷房貸家用及孩子尿布奶粉，無奈癌病襲來，當地公所獲悉轉介蘋果基金會關懷，基金會訪視後已先撥款紓困。

A2719 火燒厝 男寒夜睡破車

波及姪女4口「不知何時能重建」

2012年03月18日 傳送 讚 0 G+1 0 我要捐款 更多專欄文章

遭火災波及，阿菁的房子付之一炬。圖中孩童為阿菁2個子女。

初春乍暖還寒，55歲的單親爸爸阿水（周正水）窩在報廢廂型車裡準備就寢，雖已蓋上兩層被，還是冷得直發抖。阿水無奈地說，去年底住處鐵皮屋慘遭祝融，「現在我最擔心的，是住校讀高二兒子放假回來沒地方睡。」

報導・攝影／韓旭爾

多年來，單親爸爸阿水靠著山區林務零工掙錢撫養17歲的獨生子小傑，去年12月中，住處鐵皮屋被無名火燒毀，家當付之一炬，並波及隔壁姪女一家，「害姪女一家4口沒地方住，很對不起他們。」阿水說：「至今重建的錢不知到哪兒去找？」蘋果基金會獲悉後，已先撥款襄助阿水父子短期生活窘迫，並聯繫幫貧困家庭修屋的寶島行善義工團協力修繕，義工團評估，兩屋重建材料費用約需70萬元。當地村長說，目前已有縣府補助款共20萬元。

疑老舊電線走火

望著火場殘垣灰燼，阿火說，那天傍晚，他結束工作返家，一進門，便見屋後廚房濃煙不斷竄出，急忙打電話叫消防隊，住家位於偏遠山區，過了半個多小時消防車才趕到，一切都來不及了。當地村長說，鄰居趕緊拿出自家滅火器幫忙滅火，但當天風勢強助長火舌，用掉20多個滅火器卻沒用，「火災鑑定是老舊電線走火，沒人受傷已是不幸中的大幸。」

透過 Text Mining 了解文章的遣詞用字是如何影響人們的捐款行為

需要的材料？

- db_article_txt.rar

A3887 3胞胎待哺 單親爸抗癌魔

2016年03月23日

A2859 孝子罹癌 掛心憨弱母兄

退伍工作2月就發病 盼「重新開始」

2012年09月02日

A3629 病夫喚昏迷妻 快醒來

靠兒工讀扛家「不知怎辦」

2015年04月06日

A2719 火燒厝 男寒夜睡破車

波及姪女4口 「不知何時能重建」

2012年03月18日

初春乍暖還寒，55歲的單親爸爸阿水（周正水）窩在報廢廂型車裡準備就寢，雖已蓋上兩層被，還是冷得直發抖。阿水無奈地說，去年底住處鐵皮屋慘遭祝融，「現在我最擔心的，是住校讀高二兒子放假回來沒地方睡。」

報導．攝影／韓旭爾

多年來，單親爸爸阿水靠著山區林務零工掙錢撫養17歲的獨生子小傑，去年12月中，住處鐵皮屋被無名火燒毀，家當付之一炬，並波及隔壁姪女一家，「害姪女一家4口沒地方住，很對不起他們。」阿水說：「至今重建的錢不知到哪兒去找？」蘋果基金會獲悉後，已先撥款襄助阿水父子短期生活窘迫，並聯繫幫貧困家庭修屋的寶島行善義工團協力修繕，義工團評估，兩屋重建材料費用約需70萬元。當地村長說，目前已有縣府補助款共20萬元。

疑老舊電線走火

望著火場殘垣灰燼，阿火說，那天傍晚，他結束工作返家，一進門，便見屋後廚房濃煙不斷竄出，急忙打電話叫消防隊，住家位於偏遠山區，過了半個多小時消防車才趕到，一切都來不及了。當地村長說，鄰居趕緊拿出自家滅火器幫忙滅火，但當天風勢強助長火舌，用掉20多個滅火器卻沒用，「火災鑑定是老舊電線走火，沒人受傷已是不幸中的大幸。」

- df_article_after_eda.csv

	aid	date.published	donation	donor
1	A0001	2003-06-29	15900	22
2	A0002	2003-05-03	108415	37
3	A0004	2003-05-08	315	2
4	A0005	2003-05-04	33220	27
5	A0007	2003-05-07	26965	21
6	A0008	2003-05-22	143200	64
7	A0009	2003-05-06	38015	21
8	A0012	2003-05-20	65650	66
9	A0013	2003-05-17	86700	65
10	A0015	2003-05-23	70900	59
11	A0016	2003-07-21	52900	65
12	A0018	2003-05-14	18400	17
13	A0019	2003-05-08	17015	17
14	A0020	2003-05-18	6700	12
15	A0021	2003-05-05	6300	8
16	A0022	2003-05-10	35000	28
17	A0023	2003-05-15	123420	83
18	A0024	2003-05-16	67715	52
19	A0025	2003-06-02	18670	15
20	A0026	2003-05-29	116715	93

A2719 火燒厝 男寒夜睡破車

波及姪女4口 「不知何時能重建」

2012年03月18日 傳送 讚 0 G+1 0 我要捐款 更多專欄文章

遭火災波及，阿菁的房子付之一炬。圖中孩童為阿菁2個子女。

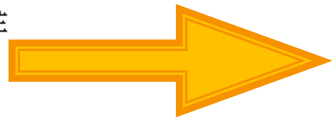
初春乍暖還寒，55歲的單親爸爸阿水（周正水）窩在報廢廂型車裡準備就寢，雖已蓋上兩層被，還是冷得直發抖。阿水無奈地說，去年底住處鐵皮屋慘遭祝融，「現在我最擔心的，是住校讀高二兒子放假回來沒地方睡。」

報導・攝影／韓旭爾

多年來，單親爸爸阿水靠著山區林務零工掙錢撫養17歲的獨生子小傑，去年12月中，住處鐵皮屋被無名火燒毀，家當付之一炬，並波及隔壁姪女一家，「害姪女一家4口沒地方住，很對不起他們。」阿水說：「至今重建的錢不知到哪兒去找？」蘋果基金會獲悉後，已先撥款襄助阿水父子短期生活窘迫，並聯繫幫貧困家庭修屋的寶島行善義工團協力修繕，義工團評估，兩屋重建材料費用約需70萬元。當地村長說，目前已有縣府補助款共20萬元。

疑老舊電線走火

望著火場殘垣灰燼，阿火說，那天傍晚，他結束工作返家，一進門，便見屋後廚房濃煙不斷竄出，急忙打電話叫消防隊，住家位於偏遠山區，過了半個多小時消防車才趕到，一切都來不及了。當地村長說，鄰居趕緊拿出自家滅火器幫忙滅火，但當天風勢強助長火舌，用掉20多個滅火器卻沒用，「火災鑑定是老舊電線走火，沒人受傷已是不幸中的大幸。」



```
$A2719
[1] "火燒厝 男寒夜睡破車初春乍暖還寒，55歲的單親爸爸阿水（周正水）窩在報廢廂型車裡準備就寢，雖已蓋上兩層被，還是冷得直發抖。阿水無奈地說，去年底住處鐵皮屋慘遭祝融，「現在我最擔心的，是住校讀高二兒子放假回來沒地方睡。」多年來，單親爸爸阿水靠著山區林務零工掙錢撫養17歲的獨生子小傑，去年12月中，住處鐵皮屋被無名火燒毀，家當付之一炬，並波及隔壁姪女一家，「害姪女一家4口沒地方住，很對不起他們。」阿水說：「至今重建的錢不知到哪兒去找？」蘋果基金會獲悉後，已先撥款襄助阿水父子短期生活窘迫，並聯繫幫貧困家庭修屋的寶島行善義工團協力修繕，義工團評估，兩屋重建材料費用約需70萬元。當地村長說，目前已有縣府補助款共20萬元。疑老舊電線走火望著火場殘垣灰燼，阿火說，那天傍晚，他結束工作返家，一進門，便見屋後廚房濃煙不斷竄出，急忙打電話叫消防隊，住家位於偏遠山區，過了半個多小時消防車才趕到，一切都來不及了。當地村長說，鄰居趕緊拿出自家滅火器幫忙滅火，但當天風勢強助長火舌，用掉20多個滅火器卻沒用，「火災鑑定是老舊電線走火，沒人受傷已是不幸中的大幸。」姪女家境也不好阿水33歲的姪女阿菁說，她與33歲的丈夫阿元育有6歲、4歲的子女，平時她在家帶孩子，家計由阿元打零工維持，日子只能勉強餬口。火災當時，「聽到外面有人叫『失火了』，我趕緊拉著兩個孩子往外衝。」阿菁說，目前一家暫住附近大伯家，屋子太小，只能在客廳打地鋪，「我也知道二伯阿水沒錢，我們手頭也不寬裕，不知何時房子才能重建。」阿水說，他現暫時睡在報廢廂型車中，並在車外搭上帆布充當煮飯、作息空間，向鄰居借廁所大小便、洗澡。「住在車裡，就當作野外露營，前幾波寒流來時，整晚冷到睡不著，春節也只能在哥哥家裡過。」阿水17歲的兒子小傑說：「車子空間小，我短期可借住同學家，久了不好意思。」"
```

讀入所有蘋果暖流文章

session_C_01_separate_words.R

```
# get all article names
> files = list.files('data/db_articles_txt/
db_articles_txt/', pattern = 'txt', full.names = T)
> file.name = gsub('.txt', '' ,basename(files))

# read in all articles
> article_txt = list()
> for(i in 1:len(files)) {
     a = readLines(files[i], encoding = 'UTF-8')
     article_txt[[file.name[i]]] =
          paste(a, collapse ='')
}

# check if all are correctly read in
> print(len(article_txt))
> fivenum(nchar(article_txt))
```

```
$A2719
[1] "火燒厝 男寒夜睡破車初春乍暖還寒，55歲的單親爸爸阿水（周正水）窩在報廢廂型車裡準備就寢，雖已蓋上兩層被，還是冷得直發抖。阿水無奈地說，去年底住處鐵皮屋慘遭祝融，「現在我最擔心的，是住校讀高二兒子放假回來沒地方睡。」多年來，單親爸爸阿水靠著山區林務零工掙錢撫養17歲的獨生子小傑，去年12月中，住處鐵皮屋被無名火燒毀，家當付之一炬，並波及隔壁姪女一家，「害姪女一家4口沒地方住，很對不起他們。」阿水說：「至今重建的錢不知到哪兒去找？」蘋果基金會獲悉後，已先撥款襄助阿水父子短期生活窘迫，並聯繫幫貧困家庭修屋的寶島行善義工團協力修繕，義工團評估，兩屋重建材料費用約需70萬元。當地村長說，目前已有縣府補助款共20萬元。疑老舊電線走火望著火場殘垣灰燼，阿火說，那天傍晚，他結束工作返家，一進門，便見屋後廚房濃煙不斷竄出，急忙打電話叫消防隊，住家位於偏遠山區，過了半個多小時消防車才趕到，一切都來不及了。當地村長說，鄰居趕緊拿出自家滅火器幫忙滅火，但當天風勢強助長火舌，用掉20多個滅火器卻沒用，「火災鑑定是老舊電線走火，沒人受傷已是不幸中的大幸。」姪女家境也不好阿水33歲的姪女阿菁說，她與33歲的丈夫阿元育有6歲、4歲的子女，平時她在家帶孩子，家計由阿元打零工維持，日子只能勉強餬口。火災當時，「聽到外面有人叫『失火了』，我趕緊拉著兩個孩子往外衝。」阿菁說，目前一家暫住附近大伯家，屋子太小，只能在客廳打地舖，「我也知道二伯阿水沒錢，我們手頭也不寬裕，不知何時房子才能重建。」阿水說，他現暫時睡在報廢廂型車中，並在車外搭上帆布充當煮飯、作息空間，向鄰居借廁所大小便、洗澡。「住在車裡，就當作野外露營，前幾波寒流來時，整晚冷到睡不著，春節也只能在哥哥家裡過。」阿水17歲的兒子小傑說：「車子空間小，我短期可借住同學家，久了不好意思。」"
```

```
[1] "火燒" "厝" "男" "寒夜" "睡" "破車" "初春" "乍暖還寒" "55" "歲" "的" "單親" "爸爸 "
[14] "阿水" "周正" "水" "窩" "在" "報廢" "廂型" "車裡" "準備" "就寢" "雖" "已" "蓋"
[27] "上" "兩層" "被" "還是" "冷得" "直發抖" "阿水" "無奈" "地說" "去年底" "住處" "鐵皮屋" "慘遭"
[40] "祝融" "現在" "我" "最" "擔心" "的" "是" "住校" "讀高二" "兒子" "放假" "回來" "沒"
[53] "地方" "睡" "多" "年來" "單親" "爸爸" "阿水靠" "著" "山區" "林務" "零工" "掙錢" "撫養"
[66] "17" "歲" "的" "獨生子" "小傑" "去年" "12" "月" "中" "住處" "鐵皮屋" "被" "無名"
[79] "火燒" "毀" "家當" "付之一炬" "並" "波及" "隔壁" "姪" "女" "一家" "害" "姪" "女"
[92] "一家" "4" "口" "沒" "地方" "住" "很" "對不起" "他們" "阿" "水" "說" "至今"
[105] "重建" "的" "錢" "不知" "到" "哪兒" "去找" "蘋果" "基金會" "獲悉" "後" "已先" "撥款"
[118] "襄助" "阿水" "父子" "短期" "生活" "窘迫" "並" "聯繫" "幫" "貧困家庭" "修屋" "的" "寶島"
[131] "行善" "義工" "團" "協力" "修繕" "義工" "團" "評估" "兩屋" "重建" "材料" "費用" "約"
[144] "需" "70" "萬元" "當地" "村長" "說" "目前" "已有" "縣府" "補助款" "共" "20" "萬元"
[157] "疑老舊" "電線走火" "望著" "火場" "殘垣" "灰燼" "阿火" "說" "那天" "傍晚" "他" "結束" "工作"
[170] "返家" "一" "進門" "便" "見" "屋後" "廚房" "濃煙" "不斷" "竄出" "急忙" "打電話" "叫"
[183] "消防隊" "住家" "位於" "偏遠" "山區" "過了" "半個" "多" "小時" "消防車" "才" "趕到" "一切"
[196] "都" "來不及" "了" "當地" "村長" "說" "鄰居" "趕緊" "拿出" "自家" "滅火器" "幫忙" "滅火"
```

文章斷詞

- jiebaR https://qinwenfeng.com/jiebaR/
  - 號稱最好的 Python 中文斷詞組件的 R 語言版本
  - 支持四種斷詞引擎
    - 最大概率法、隱式馬爾科夫模型、混和模型、索引模型
  - 可以標註詞性

- 中研院斷詞系統 http://ckipsvr.iis.sinica.edu.tw/
  - 號稱地表最強中文斷詞系統 (96% 精準度)
  - 自動標註詞性
  - 需要申請……

利用 jiebaR 斷詞

```
> library(jiebaR)
# initiate segmentation engine
> cutter = worker(bylines = T)
# cooler way to do segmentation
> article_words = lapply(article_txt, function(x)
    cutter <= x)
# traditional way
> article_words = lapply(article_txt, function(x)
    segment(x, cutter))
# check if all got segmented
> print(len(article_words))

# adjust to the format for text2vec::itoken
> article_words = lapply(article_words, '[[', 1)

> save(article_words, file =
      'data/list_article_words(jieba).RData')
```

詞的向量化

- text2vec
  - 作者 Dmitriy Selivanov 俄羅斯人
  - 支持 the state of art word embeddings (GloVe)
  - Count-based Model
  - https://cran.r-project.org/web/packages/text2vec/index.html
- word2vec
  - Tomas Mikolov 領軍的 Google Brain Team 研究團隊開發
  - Predictive Model

利用 text2vec 計算不重複的詞

session_C_02_create_glove.R

```
> load('data/list_article_words(jieba).RData')
> a = article_words

> library(text2vec)
> a.token <- itoken(a) # iterator
# to create unique word matrix
> a.vocab <- create_vocabulary(a.token,
    ngram=c(1, 1)) # 詞, 頻率, 文章佔比
```

計算 TCM

```
# vectorization of words
> a.token <- itoken(a)
> a.vectorizer <- vocab_vectorizer(a.vocab,
    grow_dtm = FALSE, skip_grams_window = 5)

# tcm = term co-occurrence matrix
> a.tcm <- create_tcm(a.token, a.vectorizer)
```

向量化詞

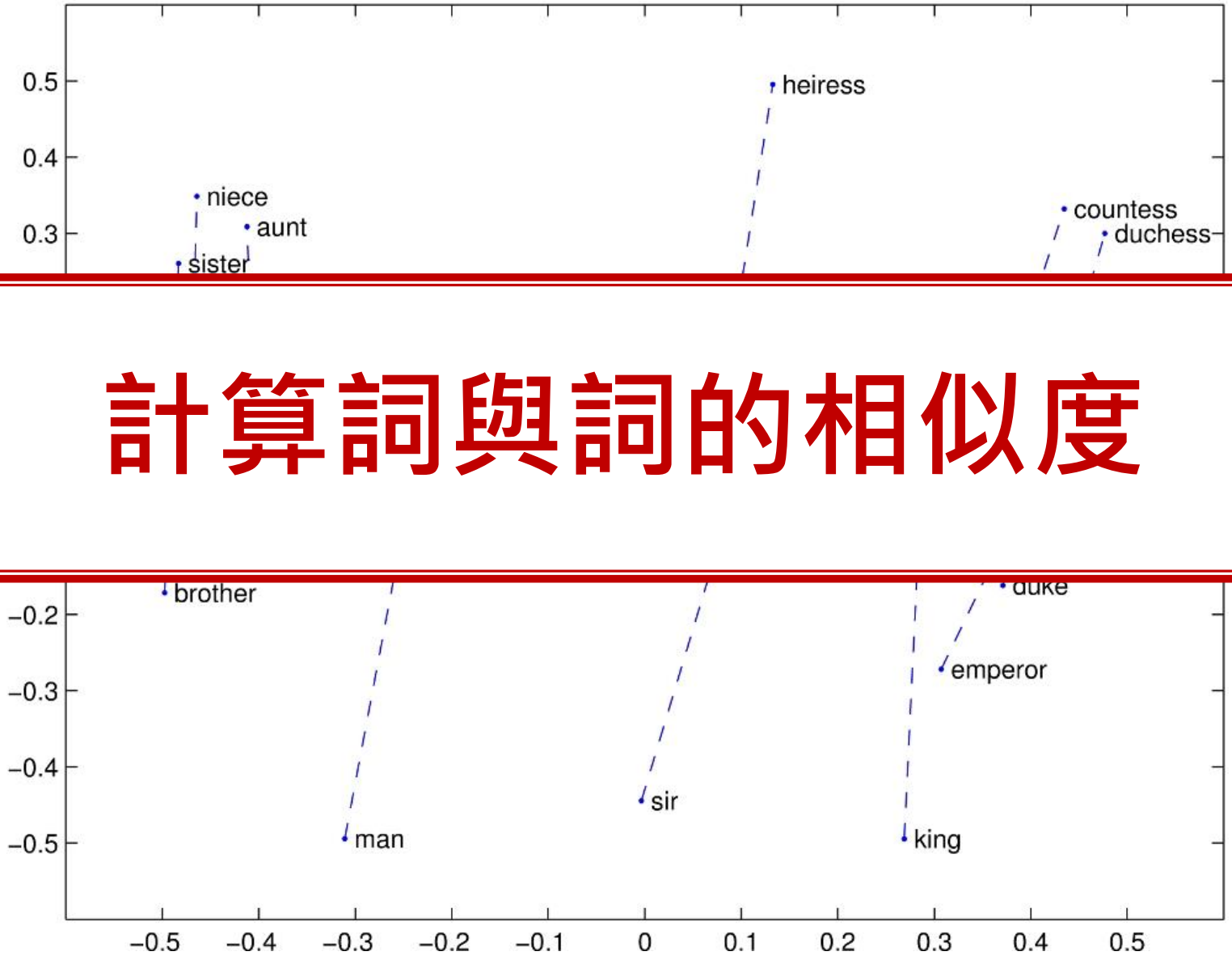
```
# glove fitting model
> fit <- glove(a.tcm, word_vectors_size = 100,
x_max =  10, learning_rate = 0.2, num_iters = 15)

# word_vectors$w_i = word vectors
# word_vectors$w_j = context vectors
> word.vec <- fit$word_vectors$w_i +
              fit$word_vectors$w_j
> rownames(word.vec) = rownames(awl.tcm)

> write.csv(word.vec, 'data/w2glv_100.csv')
```

練習 C-01, 02 (15 mins)

- 讀入文章
- jiebaR 斷詞
  - 用 jiebaR 斷詞並標註詞性 (格式：夫(n) 肺癌(n) 末期(f)...)
  - 嘗試用 jiebaR 直接讀入 6 篇文章並斷詞
- 用 text2vec 做出 2 種不同的詞向量結果
  - 調整不同向量長度
  - 做出只取 word vectors 和只取 context vectors 的兩種詞向量，給下一個部分使用

文詞向量化到底可以幹嘛!

From Stanford University Natural Language Processing Group GloVe Project



計算詞與詞的相似度

get_analogy

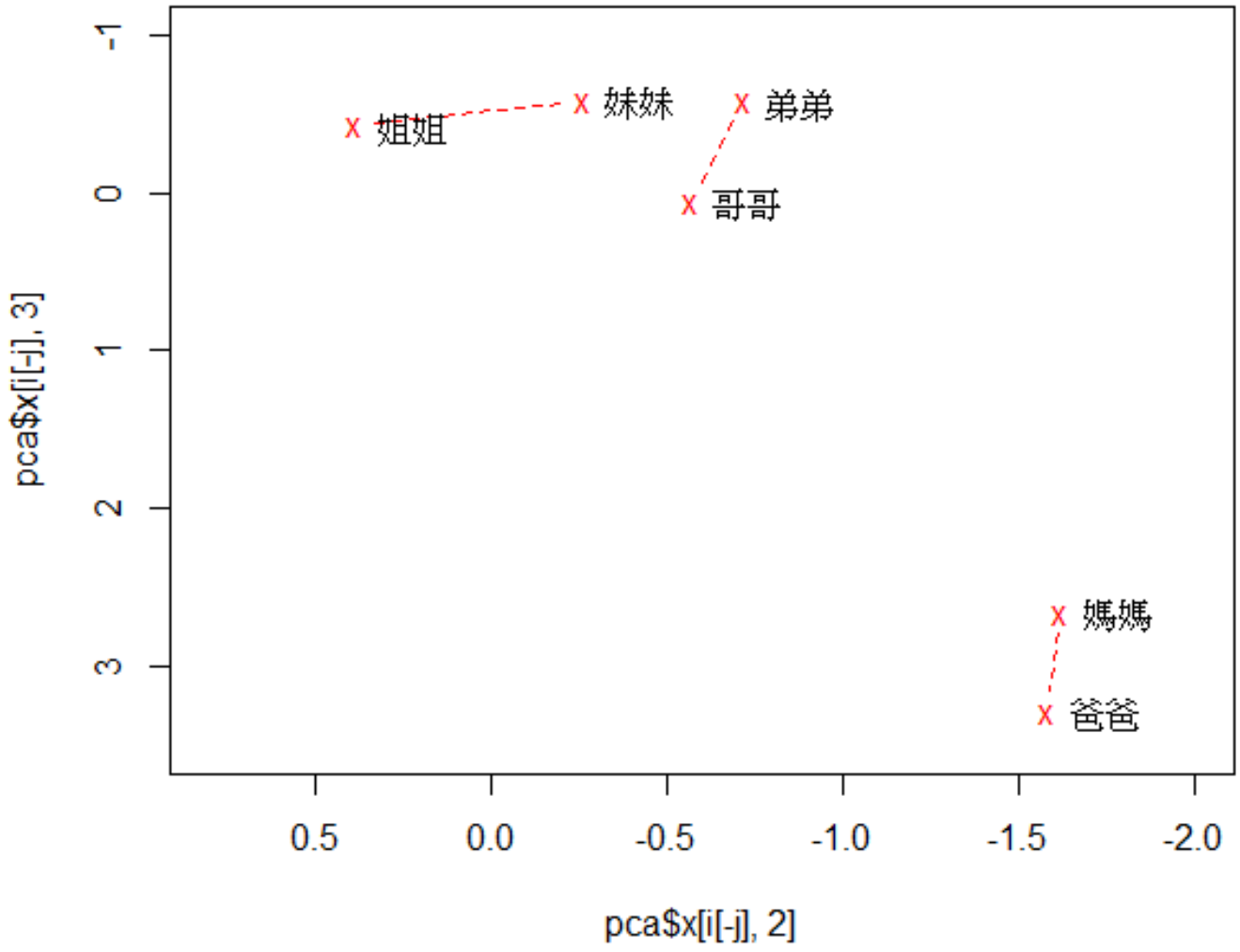
session_C_03_get_analogy.R

```
# read in word vector matrix
> word.vec = read.csv('data/w2glv_100.csv')
# calculate unit vector
> word.vec.norm = sqrt(rowSums(word.vec^2))
# build the function
> get_analogy = function(king, man, woman) {
    queen = word.vec[king, , drop = F] -
            word.vec[man, , drop = F] +
            word.vec[woman, , drop = F]
    cos.dist = text2vec:::cosine(queen,
            word.vec, word.vec.norm)
    head(sort(cos.dist[1, ], decreasing = T,10)
}
```

觀察高維度空間文詞的位置

```
> get_analogy('女兒', '媽媽', '爸爸')
    女兒       兒子       孫子       爸爸       兩個       弟弟
0.9157175 0.8439616 0.7206885 0.7206012 0.7134087 0.7012182
   小女兒       妹妹         她       孩子
0.6984936 0.6915945 0.6889442 0.6745889
> get_analogy('癌症', '爸爸', '媽媽')
    癌症       折磨       乳癌         是       復發     口腔癌
0.8955176 0.5950794 0.5772254 0.5744766 0.5619637 0.5523182
      得       發現     沒想到       肝癌
0.5453836 0.5374149 0.5351327 0.5248039
> get_analogy('肝癌', '爸爸','媽媽')
    肝癌       乳癌       肺癌       罹患     鼻咽癌     食道癌
0.8944141 0.6712295 0.6616412 0.6554880 0.6515760 0.6286956
   口腔癌       末期       發現         罹
0.6273844 0.6149330 0.5960807 0.5927049
> get_analogy('中風','左側','右側')
    中風       癱瘓       二度       右側         又       倒下
0.8708747 0.6337918 0.6243119 0.5605684 0.5543838 0.5480522
  行動不便       臥床       因車禍       罹癌
0.5471583 0.5304387 0.5289707 0.5146370
```

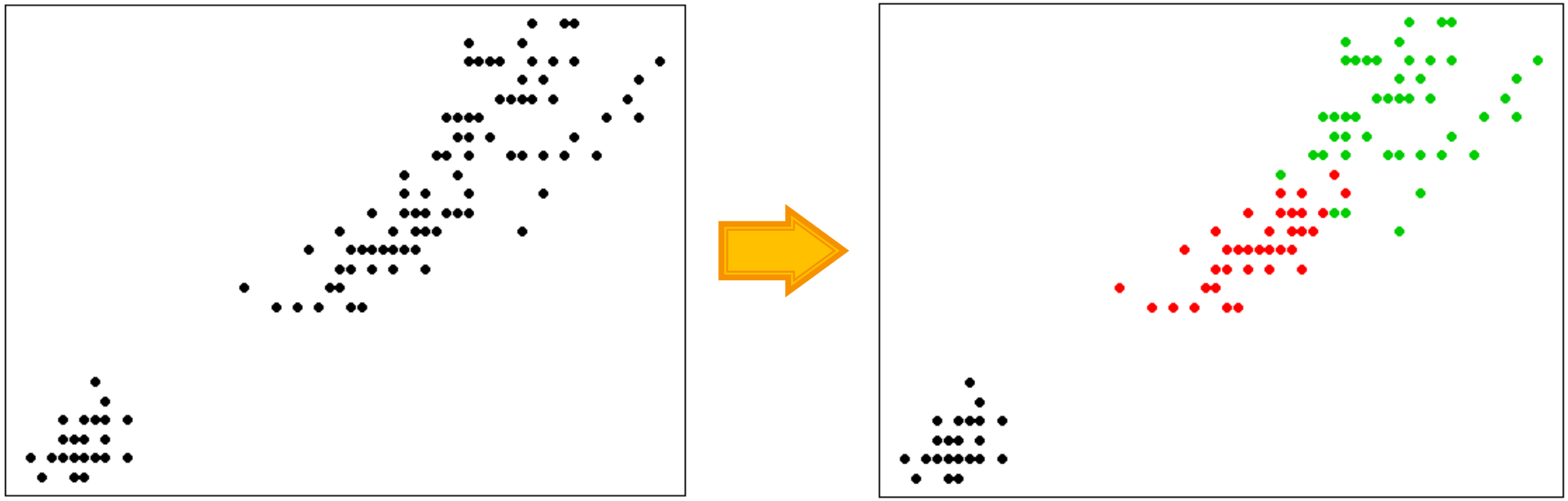
利用 PCA 降維看文詞的位置



詞在高維度的空間因為用字遣詞的順序，使較相似的詞較相近

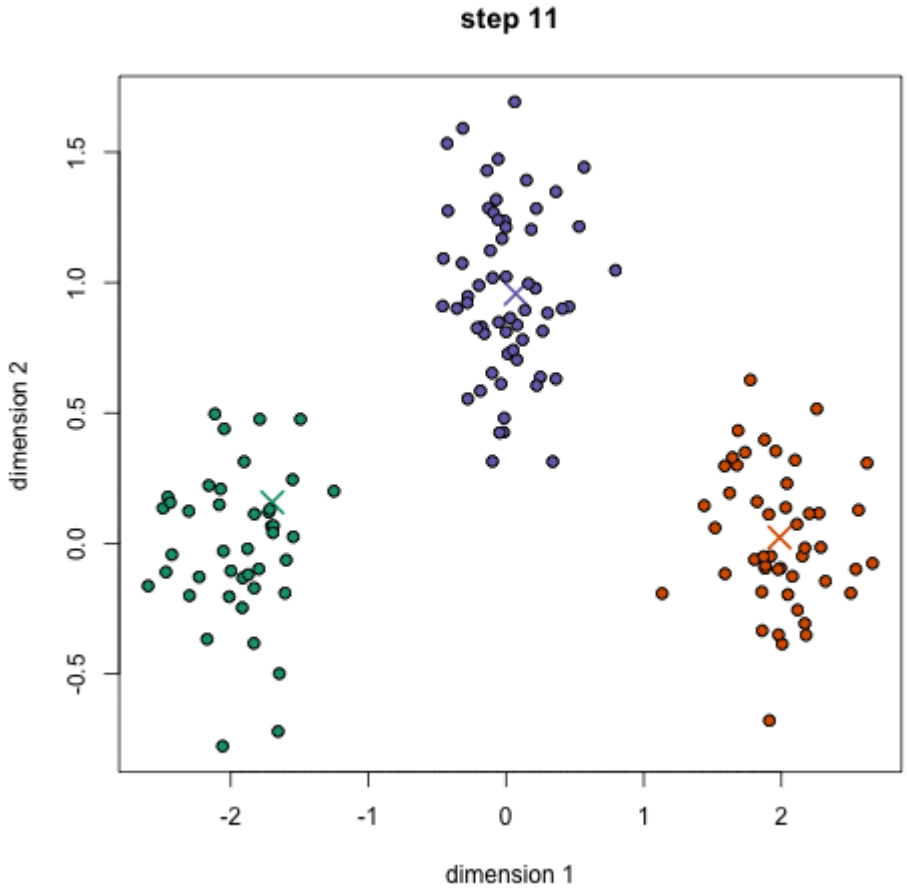
將相似的詞做分類 (Classification)

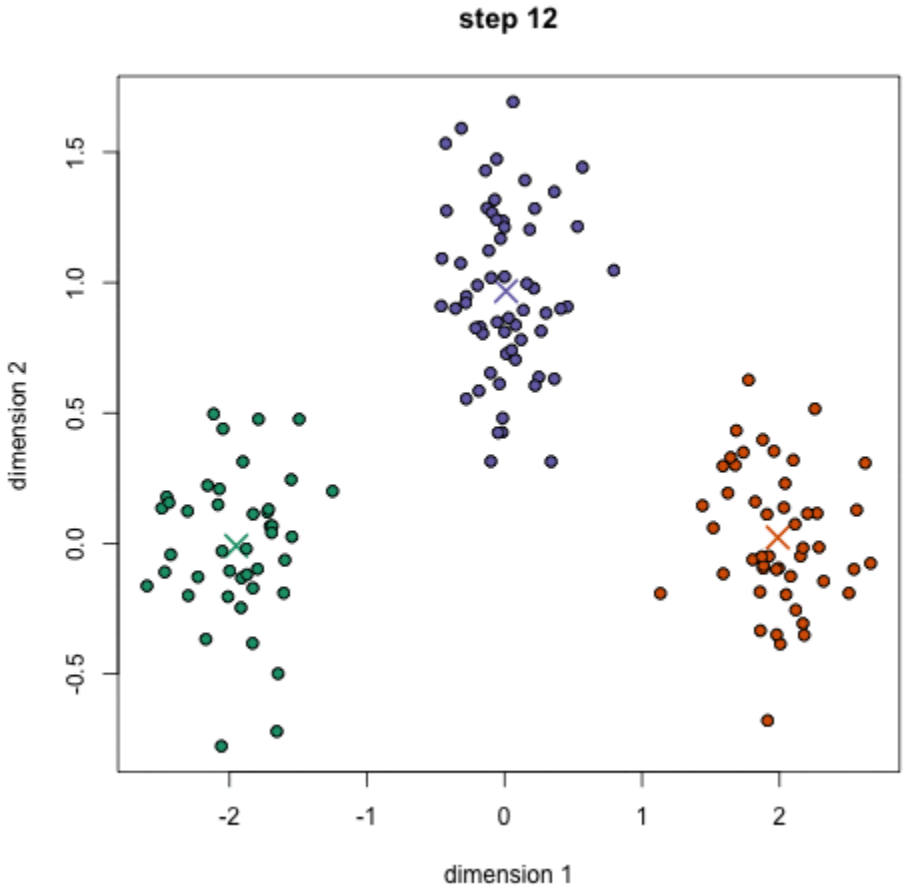
k-means clustering 做資料分類



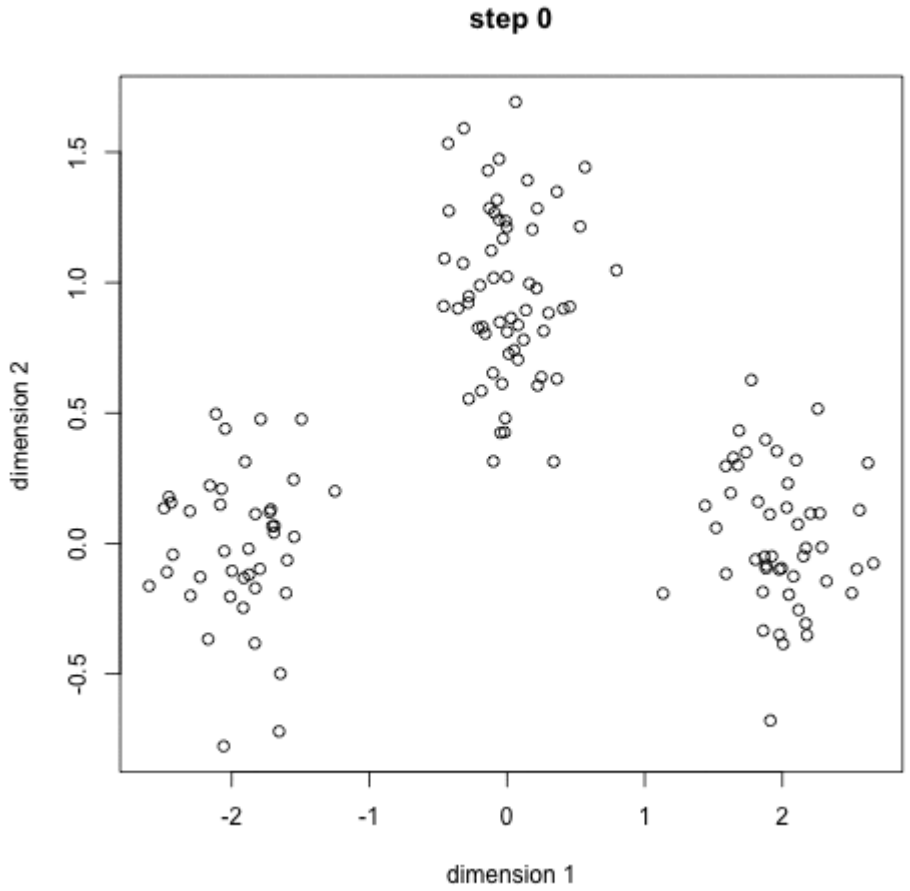
k-means clustering 步驟

- 隨機選取 k 個中心點 (cluster center)
- 計算所有點與中心點的距離，並分歸類到最近的中心
- 計算 k 個群的中心點，做為新的群中心點
- 重附上兩步驟，直到群中心點收斂到不改變





Source: http://simplystatistics.orgwp-content/uploads/2014/02/kmeans.gif



Source: http://simplystatistics.orgwp-content/uploads/2014/02/kmeans.gif

利用 k-means 將文字分群

session_C_04_glove_clustering.R

```
# do kmeans clustering
> k <- kmeans(w[,-i, with = F], centers = ?,
    iter.max = 100)
> w$k = k$cluster

> table(w$k)
```

計算文章的文字群比例

```
# calculate the ratio of word clusters per article
```

	aid	n.word	k1	k2	k3	k4	k5	k6	k7	k8	k9	k10
1	A0002	594	0.000000000	0.000000000	0.0000000000	0.000000000	0.000000000	0.001683502	0.000000000	0.000000000	0.0000000000	0.0000
2	A0004	566	0.000000000	0.000000000	0.0000000000	0.003533569	0.001766784	0.000000000	0.001766784	0.000000000	0.0000000000	0.0000
3	A0005	641	0.000000000	0.000000000	0.0000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.0015600624	0.0015
4	A0007	601	0.000000000	0.000000000	0.0000000000	0.001663894	0.000000000	0.001663894	0.003327787	0.000000000	0.0000000000	0.0000
5	A0008	616	0.000000000	0.000000000	0.0000000000	0.000000000	0.004870130	0.000000000	0.001623377	0.000000000	0.0000000000	0.0000
6	A0009	595	0.000000000	0.000000000	0.0000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.0000000000	0.0000
7	A0012	570	0.000000000	0.000000000	0.0000000000	0.001754386	0.000000000	0.001754386	0.000000000	0.001754386	0.0000000000	0.0017
8	A0013	707	0.000000000	0.000000000	0.0014144272	0.000000000	0.004243281	0.001414427	0.000000000	0.002828854	0.0000000000	0.0014
9	A0015	486	0.000000000	0.000000000	0.0000000000	0.000000000	0.002057613	0.002057613	0.000000000	0.002057613	0.0000000000	0.0020
10	A0016	649	0.000000000	0.000000000	0.0030816641	0.000000000	0.001540832	0.013867488	0.000000000	0.001540832	0.0000000000	0.0000
11	A0019	300	0.000000000	0.000000000	0.0000000000	0.000000000	0.006666667	0.000000000	0.000000000	0.003333333	0.0000000000	0.0033
12	A0020	320	0.000000000	0.000000000	0.0000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.0000000000	0.0000
13	A0021	698	0.000000000	0.000000000	0.0014326648	0.000000000	0.002865330	0.002865330	0.000000000	0.002865330	0.0000000000	0.0042
14	A0022	359	0.000000000	0.000000000	0.0000000000	0.000000000	0.002785515	0.000000000	0.000000000	0.000000000	0.0000000000	0.0000
15	A0023	528	0.000000000	0.000000000	0.0000000000	0.000000000	0.001893939	0.000000000	0.000000000	0.005681818	0.0000000000	0.0018
16	A0024	610	0.004918033	0.000000000	0.0016393443	0.000000000	0.000000000	0.001639344	0.001639344	0.000000000	0.0016393443	0.0000
17	A0025	463	0.000000000	0.000000000	0.0000000000	0.002159827	0.002159827	0.000000000	0.000000000	0.000000000	0.0000000000	0.0021

用相關性來觀察影響

```
# check the correlation between word clusters and
# the variables we care
> i <- grep('^k|donation|donor|log|n.fb|ttl',
    names(d))
> View(cor(d[,i]))
```

調整不同群的大小，抓出不同規模的詞

- k = 30
  - donor.de vs k11, cor = 0.1658

```
> w[w$k == 11,]$v1
 [1] "個(M)" "家(N)" "生活(N)" "兩(DET)" "兒子(N)" "照顧(Vt)" "無法(ADV)"
 [8] "丈夫(N)" "得(ADV)" "母親(N)" "三(DET)" "住(Vt)" "更(ADV)" "工作(Vi)"
[15] "父親(N)" "為了(P)" "中(N)" "經濟(N)" "靠(Vt)" "無(Vt)" "來(POST)"
[22] "阿嬤(N)" "生計(N)" "只好(ADV)" "孫子(N)" "困難(Vi)" "自(P)" "四(DET)"
[29] "壓力(N)" "之(T)" "孫女(N)" "智障(Vi)" "外(POST)" "除了(P)" "根本(ADV)"
[36] "期間(N)" "幾乎(ADV)" "需要(Vt)" "唯一(A)" "一家人(N)" "孫(N)" "五(DET)"
[43] "重擔(N)" "擔憂(Vt)" "幼子(N)" "公公(N)" "去世(Vi)" "老人家(N)" "往後(N)"
[50] "失業(Vi)" "四處(ADV)" "外出(Vi)" "患有(Vt)" "生活(Vi)" "養活(Vt)" "同時(N)"
[57] "年幼(Vi)" "不僅(C)" "兒女(N)" "年邁(Vi)" "自從(P)" "安養(Nv)" "只得(ADV)"
[64] "相依為命(Vi)" "起居(N)" "正在(ADV)" "依靠(Vt)" "獨自(ADV)" "過年(Vi)" "早已(ADV)"
[71] "儘管(C)" "如何(ADV)" "難以(ADV)" "相當(ADV)" "夫婦(N)" "屏東(N)" "方便(Vi) "
[1233] "松秋蘭(N)" "看上去(ADV)" "活兒(N)" "蔡謝蕊(N)" "霎時間(ADV)" "孝養(Vt)" "鬢髮(N)"
[1240] "託大(Vi)" "王素玲(N)" "自助餐廳(N)" "絕口不提(Vt)" "正光(N)" "陳清旗(N)" "止失(N)"
[1247] "孔金榮(N)" "林林總總(Vi)" "岔子(N)" "挽回不了(Vt)" "倍受(P)" "泥淖(N)" "殘貼(N)"
[1254] "安(ADV)" "鄉士文(N)" "廖文調(N)" "武國(N)" "莊景松(N)" "境地(N)" "微簿(N)"
[1261] "溫(Vt)" "許孟(N)" "桂(b)" "愁煩(N)" "揮別(Vt)" "隨侍在側(Vi)" "淒苦(Vi)"
[1268] "接連不斷(Vi)" "南化(Vt)" "司淑英(N)" "張和妹(N)" "對坐(Vi)" "長子家(N)" "何美麗(N)"
[1275] "成長期(N)" "子精障(N)" "晚輩(N)" "阿桂伯(N)" "朱光華(N)" "斷然(ADV)" "郭秀英(N)"
[1282] "牽繫(Nv)" "小高(N)" "夫婦倆人(N)" "人籌夫(N)" "烤雞(N)" "貼壁磚(N)" "憂慮體(N)"
[1289] "哀慟(Nv)" "無畏(Vt)" "陳張美(N)" "孫瑞龍(N)" "司家(N)" "錢金水(N)" "需要人(N)"
[1296] "東湊西湊(Vt)" "蘇耿毅(N)" "七千五百(DET)" "景松(N)" "李新助三姊(N)" "春花婆(N)" "住友屋(N)"
[1303] "王雅君(N)" "就醫費(N)" "智障兒女(N)" "次媳(N)" "曾雅慰(N)" "李丁貴(N)" "殷殷(Vi)"
[1310] "福田(N)" "劉余治(N)" "池上(N)" "阿治猝(N)" "奔走(Nv)" "水光(N)" "痴呆症(N)"
```

k = 100

- donor.de vs k13 (老人相關詞), cor = 0.2185

```
> w[w$k == 13,]$v1
[1] "兒子(N)" "阿嬤(N)" "孫子(N)" "過世(Vi)" "次子(N)" "孫女(N)" "留下(Vt)" "孫(N)"
[9] "媳婦(N)" "阿公(N)" "獨子(N)" "生前(N)" "兒女(N)" "年邁(Vi)" "獨力(ADV)" "相依為命(Vi)"
[17] "獨自(ADV)" "夫婦(N)" "么子(N)" "往生(Vi)" "祖孫(N)" "自幼(ADV)" "留(Vt)" "相依(Vi)"
[25] "成年(Vi)" "照(P)" "養育(Vt)" "相繼(ADV)" "清苦(Vi)" "拉拔(Vt)" "帶大(Vt)" "外孫(N)"
[33] "喪夫(Vi)" "秀蘭(N)" "智(N)" "三子(N)" "另外(ADV)" "長孫(N)" "姪子(N)" "倆(DET)"
[41] "靈堂(N)" "幼孫(N)" "孤苦(Vi)" "喪子(Vi)" "叔(N)" "秀英(N)" "孫子女(N)" "岳母(N)"
[49] "滿(ADV)" "媳(N)" "一手(ADV)" "喪父(Vi)" "弱智(N)" "丟給(Vt)" "養孫(N)" "守寡(Vi)"
[57] "託給(Vt)" "辛勞(Nv)" "阿武(N)" "阿弘(N)" "孫兒(N)" "長媳(N)" "嬸嬸(N)" "貧苦(Vi)"
[65] "老幼(N)" "明(N)" "雄(A)" "堂弟(N)" "祖父(N)" "伯伯(N)" "阿土伯(N)" "阿香嬤(N)"
[73] "祖父母(N)" "夭折(Vi)" "小萍(N)" "老奶奶(N)" "次(DET)" "挨餓(Vi)" "阿祖(N)" "阿妮(N)"
[81] "雙亡(Vi)" "曾孫(N)" "屏(N)" "後塵(N)" "雪花(N)" "秀麗(Vi)" "黑髮(N)" "小揚(N)"
[89] "楊春美(N)" "稚齡(N)" "清苦(Nv)" "春葉婆(N)" "黃玉(N)" "阿月嬤(N)" "月娥嬤(N)" "葬(Vt)"
[97] "同堂(Vi)" "李國材(N)" "阿枝嬤(N)" "涵妮(N)" "孤(Vi)" "喪生(Vi)" "墳(N)" "成材(Vi)"
[105] "三代(N)" "蓮妹(N)" "洪氷(N)" "恩惠(N)" "郭麗華(N)" "曾祖母(N)" "玉里(N)" "廖春嬤(N)"
[113] "金鶯嬤(N)" "古秋香(N)" "丫丫(N)" "阿淺嬤(N)" "徐玉安(N)" "員林(N)" "勤儉(Vi)" "處理完(Vt)"
[121] "雅萍(N)" "春蓮(N)" "阿錦婆(N)" "含笑(Vi)" "雅雅(N)" "雲英(N)" "阿枣(N)" "玉燕婆(N)"
[129] "余玉盛(N)" "阿治(N)" "李瑪美(N)" "阿素嬤(N)" "水田伯(N)" "黃貴妹(N)" "阿蜜嬤(N)" "童養媳(N)"
[137] "母逝(N)" "三孫(N)" "兩老(N)" "撫育(Vt)" "光輝(N)" "素玲(N)" "姪(N)" "行蹤(N)"
[145] "黃西(N)" "寬敏(N)" "蔡鎮(N)" "許永隆(N)" "劉玉秀(N)" "花玉婆(N)" "太魯閣(N)" "志盛(N)"
[153] "張秀玉(N)" "許文雄(N)" "堂(N)" "鄭寶桐(N)" "田銘昌(N)" "黃婿妮(N)" "池玉秀(N)" "黃貝(N) "
[161] "碧雲嬤(N)" "游瑛鳳(N)" "許麥(N)" "侯春蘭(N)" "感念(Vt)" "么孫(N)" "石松(N)" "分住(Vt)"
[169] "俊傑(N)" "粉妹(N)" "俊彥(N)" "偶(N)" "溫玉季(N)" "治喪(Vi)" "邱清課(N)" "寶菊(N)"
[177] "鍾月靜(N)" "孤女(N)" "麗梅(N)" "明女(N)" "文子(N)" "麗瓊嬤(N)" "阿嬌嬤(N)" "阿峰婆(N)"
[185] "胡梅芳(N)" "阿景(N)" "阿嬤家(N)" "阿參婆(N)" "葬身(Vt)" "徐玉蘭(N)" "宏文(N)" "依(Vt)"
[193] "樂(N)" "素麗(N)" "幼弱(Vi)" "世居(Vt)" "阿綿嬤(N)" "氷糖(N)" "耍(Vt)" "玉華嬤(N) "
```

k = 200

- donor.de vs k133 (死亡詞), cor = 0.203

```
> w[w$k == 133,]$v1
[1] "走(Vi)" "病逝(Vi)" "逝(Vi)" "留下(Vt)" "死(Vi)" "去世(Vi)" "生前(N)" "喪葬(N)" "喪葬費(N)" "往生(Vi)"
[11] "猝(ADV)" "交代(Vt)" "後事(N)" "留(Vt)" "驟(ADV)" "心肌(N)" "梗塞(Vi)" "來不及(ADV)" "身亡(Vi)" "悲傷(Vi)"
[21] "喪夫(Vi)" "喪事(N)" "遺照(N)" "送走(Vt)" "突然(Vi)" "出殯(Vi)" "辦妥(Vt)" "靈堂(N)" "猝死(Vi)" "喪子(Vi)"
[31] "程(N)" "國民(N)" "不敵(Vt)" "白髮人(N)" "黑髮人(N)" "臨終(Vi)" "病故(Vi)" "含淚(Vi)" "溺斃(Vi)" "辦好(Vt)"
[41] "撒手(Vi)" "遺體(N)" "辭世(Vi)" "老幼(N)" "人世(N)" "辦完(Vt)" "斷氣(Vi)" "殯儀館(N)" "冰櫃(N)" "悲慟(Vi)"
[51] "桂圓(N)" "奠儀(N)" "睡夢(N)" "繼(P)" "葬儀社(N)" "離世(Vi)" "紙錢(N)" "再見(Vi)" "黑髮(N)" "平復(Vt)"
[61] "小晟(N)" "喪禮(N)" "林美琴(N)" "遺言(N)" "棺木(N)" "葬(Vt)" "積勞成疾(Vi)" "遽逝(Vi)" "享年(Vt)" "陳芷涵(N)"
[71] "孤(Vi)" "安葬(Vt)" "永隔(N)" "杜氏(N)" "火化(Vt)" "薛(N)" "葉榮進(N)" "含悲(Vi)" "悲痛(Vi)" "遺書(N)"
[81] "處理完(Vt)" "周麗珠(N)" "江鎧安(N)" "莫惠萍(N)" "自焚(Vi)" "莊台(N)" "阿治(N)" "李瑪美(N)" "忠正(N)" "悲(Vi)"
```

- donor.de vs k180 (貧困食物), cor = 0.1008 (提升了!)

```
> w[w$k == 180,]$v1
[1] "吃(Vt)" "餐(M)" "飯(N)" "晚餐(N)" "煮(Vt)" "便當(N)" "青菜(N)" "餓(Vi)" "省錢(Vi)" "配(Vt)"
[11] "罐頭(N)" "飯菜(N)" "泡麵(N)" "早餐(N)" "稀飯(N)" "炒(Vt)" "熱(Vt)" "白飯(N)" "麵(N)" "肉(N)"
[21] "餓肚子(Vi)" "飽(Vi)" "包(M)" "鍋(M)" "碗(M)" "拌(Vt)" "果腹(Vi)" "粥(N)" "碗(N)" "吃飽(Vi)"
[31] "瓶(M)" "頓(M)" "好吃(Vi)" "吃完(Vt)" "一口(ADV)" "蛋(N)" "醬油(N)" "麵包(N)" "湯(N)" "加菜(Vi)"
[41] "餐桌(N)" "道(M)" "鍋(N)" "樣(M)" "麵線(N)" "煎(Vt)" "剩菜(N)" "饅頭(N)" "弄(Vt)" "麵條(N)"
[51] "開水(N)" "澆(Vt)" "鹹(Vi)" "填飽(Vt)" "煮好(Vt)" "鹽(N)" "肉鬆(N)" "炒飯(N)" "盤(M)" "蛋炒飯(N)"
[61] "加熱(Vi)" "荷包蛋(N)" "苦瓜(N)" "吃到(Vt)" "水餃(N)" "營養(Vi)" "滷肉(N)" "豆漿(N)" "菜色(N)" "年夜飯(N)"
[71] "放進(Vt)" "自助餐(N)" "省下來(Vi)" "扒(Vt)" "舀(Vt)" "吃剩(Vi)" "醬瓜(N)" "餓死(Vt)" "打發(Vt)" "蒸(Vt)"
[81] "充飢(Vi)" "燙(Vt)" "滷(Vt)" "菜肉(N)" "匙(M)" "攪拌(Vt)" "絲瓜(N)" "會兒(N)" "豆腐(N)" "配上(Vt)"
[91] "罐(N)" "飯桌(N)" "包子(N)" "白開水(N)" "津津有味(Vi)" "食量(N)" "吐司(N)" "熱騰騰(Vi)" "醬菜(N)" "用餐(Nv)"
[101] "填(Vt)" "湯汁(N)" "配飯(N)" "吃光(Vt)" "大鍋(N)" "菜葉(N)" "湯麵(N)" "午飯(N)" "食欲(N)" "菜湯(N)"
```

k = 500

- n.fb.share vs k233 (長輩名), cor = 0.1486

```
> w[w$k == 233,]$v1
[1] "老伴(N)" "嬸(N)" "阿英婆(N)" "阿雄伯(N)" "阿丹(N)" "阿義伯(N)" "阿玉婆(N)" "阿德伯(N)" "秋香(N)" "阿嬌婆(N)" "阿香嬤(
[13] "阿敏婆(N)" "阿燕婆(N)" "阿螢(N)" "阿金婆(N)" "聯喜伯(N)" "阿枝婆(N)" "金銀(N)" "阿蘭婆(N)" "信行伯(N)" "時(N)" "錦雲
[25] "阿益伯(N)" "阿苗(N)" "飛伯(N)" "阿樹伯(N)" "玉華(N)" "阿慢婆(N)" "金菊婆(N)" "阿嫃(N)" "銀達伯(N)" "阿泰伯(N)" "阿媛
[37] "阿錦婆(N)" "萬通伯(N)" "林文婆(N)" "玉枝(N)" "謝惠美(N)" "周明(N)" "阿容伯(N)" "阿票嬤(N)" "秀雄伯(N)" "李英圳(N)" "
[49] "阿葉嬤(N)" "阿容婆(N)" "明美婆(N)" "沈春(N)" "金系婆(N)" "吳金龍(N)" "阿格嬤(N)" "潘美玲(N)" "炳源伯(N)" "阿研婆(N)"
[61] "阿瑤嬤(N)" "月鈴嬤(N)" "秀妹婆(N)" "李金(N)" "嬤憂(Vi)" "義平伯(N)" "阿春婆(N)" "春蘭婆(N)" "水金婆(N)" "金美婆(N)"
[73] "振華伯(N)" "秀子(N)" "佩芳(N)" "佩芳婆(N)" "吳妹婆(N)"
```

- tll.cancer vs k63 (殘障相關詞), cor = -0.21

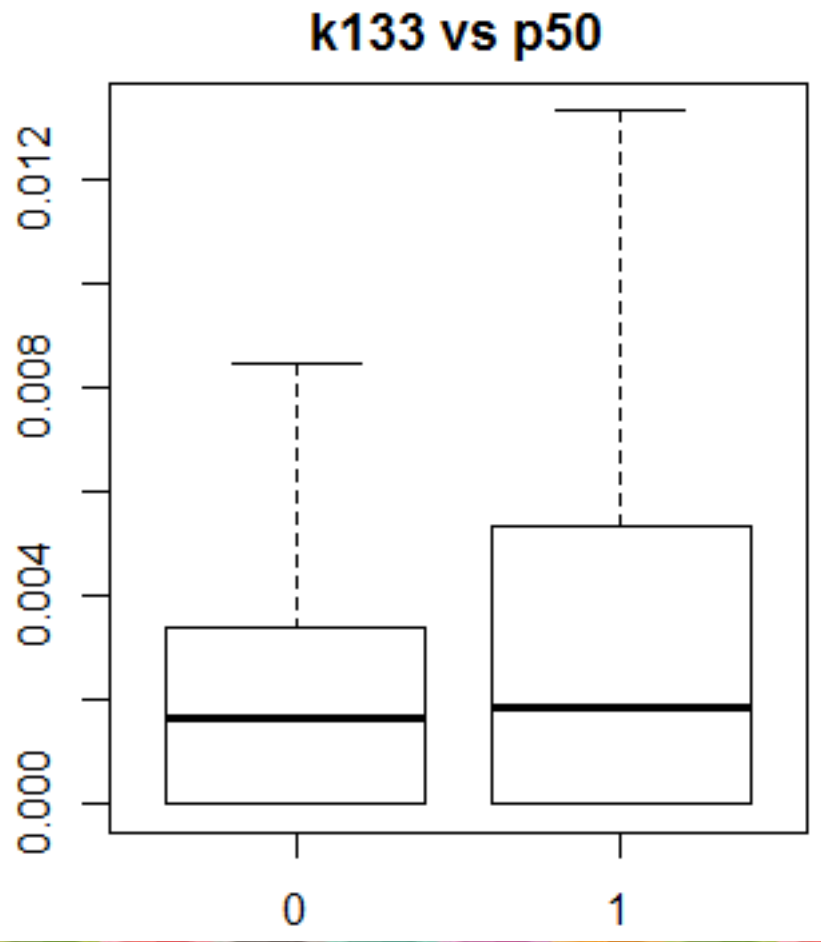
```
> w[w$k == 63,]$v1
[1] "中風(Vi)" "癱瘓(Vi)" "無力(Vi)" "手腳(N)" "癱(Vi)" "臥床(Vi)" "自理(Vt)" "半身(ADV)" "腦中風(N)" "四肢(N)"
[11] "左側(N)" "右側(N)" "麻痺(Vi)" "癱軟(Vi)" "輕微(Vi)" "癱臥(Vi)" "僵硬(Vi)" "傷腦(N)" "突發(Vi)" "行動(Vi)"
[21] "動彈(Vi)" "乏力(Vi)" "偏(ADV)" "右邊(N)" "寸步不離(Vi)" "腦梗塞(N)" "全癱(N)" "失語(N)" "帕金森(N)" "中風(N)"
[31] "麻木(Vi)" "阿農(N)" "神智不清(Vi)" "有礙(Vt)" "偏癱(N)" "一口口(DET)" "失語症(N)" "楊宇衡(N)" "華為(N)" "田英枝(N)"
[41] "緩步(Vi)" "王志發(N)" "語言區(N)" "致(N)"
```

找出有意義的字群

- w2v, k = 200
  - 觀察 donor.de
  - 負相關 (< -0.1): k84 (捐款詞), 101 (火災詞), 148 (公益活動), 125 (商務詞), 60 (宗教公益詞), 112 (學習詞), 194 (才藝與其他詞), 135 (公務詞)
  - 正相關 (> 0.1): k133 (死亡詞), 89 (隔代詞), 163 (小孩懂事詞), 46 (孤苦伶仃詞), 56 (小孩名), 180 (貧困食物詞)

用 boxplot 看前 50% 與後 50% donation.de 是否真的有差異?

```
> i <- which(d$donation.de > mean(d$donation.de))
> d$p50 = 0
> d[i,]$p50 = 1
> boxplot(d$k133 ~ d$p50,
    outline = F)
```



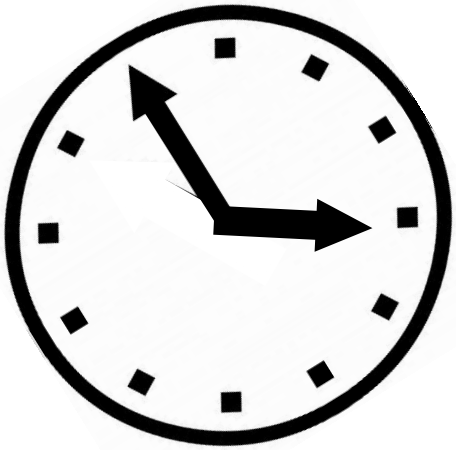
ANOVA 檢驗

```
> summary(aov(d$donation.de ~ d$p50))
```

```
              Df     Sum Sq   Mean Sq F value Pr(>F)
d$p50          1 2.438e+13 2.438e+13    1326 <2e-16 ***
Residuals   3748 6.892e+13 1.839e+10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

練習 C-03, 04

- 寫出 get_analogy function
- 利用 PCA 降維
  - 畫出兩種疾病與兩種災難與兩位種文章主詞的關係
- 運用 k-means 將文詞分群
  - 分出 10 群, 20 群
- 計算隨機 100 篇文章的文字群比例
  - 用 kmeans 的 10 群 / 20 群分 100 篇文章的比例
- 讀入已經寫好的檔案，觀察字群與變數的關係
  - 並用 ANOVA 檢驗



Stay Tuned…… We'll be back soon!!

Next session starts at **15:50**

建立預測模型
Model Building

Session D

手上有哪些材料？

- 原始資料：每篇文章的
  - 時間、捐款金額、捐款人數、FB 數
  - 標題、內文、字數、圖片、撰文者
- 萃取資訊
  - 數字轉化
    - mean, med, log……
  - 文本分析
    - 標題是否含有男女、單親、小孩、癌症、死亡……
    - 文章內文的用字、結構、語意

資料欄位分組

- 設定Response variables
  - **donor.de** (or detrended donation, donation.mean, etc.)
- 文章的 meta info
  - 時間、Fb 數量、文章標題資料
- 文章的詞向量
  - GloVe
  - Word2vec
- 文章的分群詞比例 (n 群)
  - “k-ratio”

資料欄位分組 (cont'd)

```
# read in the data
load('data/w2glv_name_clustering(k=200).Rdata')

# subset the column by names
ch <- which(sapply(d, is.factor))
ttl<- names(d)[grep("^ttl", colnames(d))]
fb <- names(d)[grep("fb", colnames(d))]
kg <- names(d)[grep('^g', names(d))]
time <- names(d)[grep('day|month', names(d))]
yns  <- names(d)[grep('^don', names(d))]
rest <- names(d)[!names(d)%in%c(ttl, fb, k, kg, yns,
time)]
```

練習 D-01 (8 mins)：資料修整

session_D_ex01.R

- 請讀取 “w2v_name_clustering(k=200).Rdata”
- 觀察極端值
  - 去除 outliers (保留約 95%資料)
- 取近一年資料
  - 從 2015-01-01 到 2016-06-03
- 設定 Response Variable
  - 捐款人數 donor.de ✓

練習 D-01 (解答)

- 請參考 session_D_ex01.R

建立預測模型

- 線性迴歸

```
lm(formula, data=data)
```

- 支持向量機

```
svm(formula, data=data)
```

- 決策樹

```
rpart(formula, data=data)
```

- 隨機森林

```
randomForest(formula, data=data)
```

建立預測模型

- 線性迴歸

```
lm(y ~ x1 + x2 + x3, data=data)
```

$$y = f(\text{WORLD}) + b$$

篩選資料

- 變數間相關性
  - Mutual-correlation check
- 共線性診斷
  - 變異數膨脹因素 VIF (Variance Inflation Factor)

篩選資料 (cont'd)

```
# pseudo code
# feature-feature correlation
data <- cor(data, use="complete")

# variance inflation fraction
source("func/vif.R")
vif(data, thresh=10)
```

練習 D-02 (3 mins)

session_D_ex02.R

- 練習檢查 mutual correlation 和 VIF

線性迴歸模型

```
# pseudo code
lm.fit <- lm(y ~ x1 + x2 + x3……, data=data)
```

```
Residual standard error: 87.33 on 275 degrees of freedom
Multiple R-squared:  0.5412,    Adjusted R-squared:  0.401
F-statistic: 3.861 on 84 and 275 DF,  p-value: < 2.2e-16
```

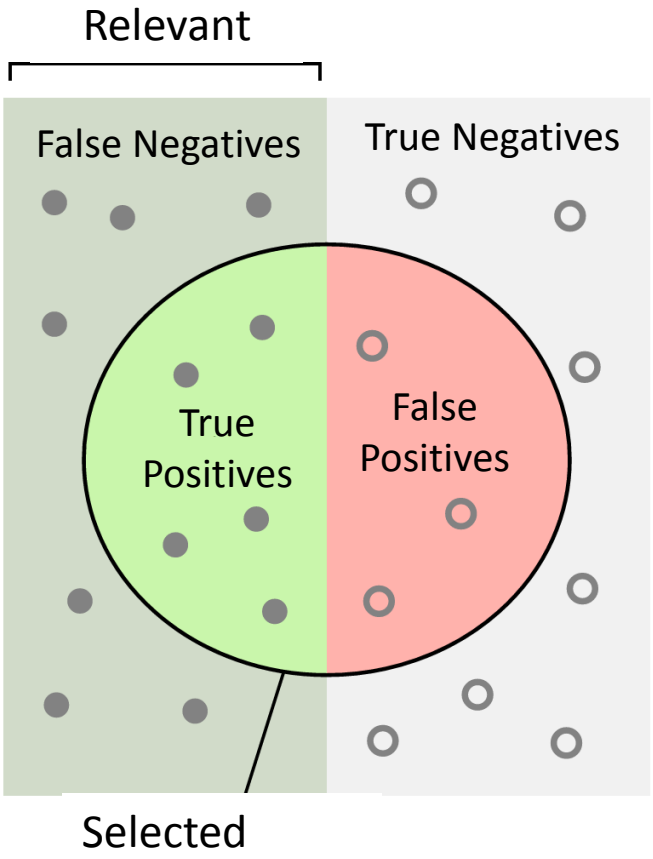
寫完考卷，來對答案……

- Continuous Variable
  - Pearson correlation coefficient $= \frac{cov(X,Y)}{\sigma_x \sigma_y}$ ✓
  - Coefficient of determination (R^2) $= 1 - \frac{SS_{reg}}{SS_{tot}}$ ✓
  - RMSE (Root mean square error) $= \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)}{n}}$
- Categorical Variable
  - Accuracy $= \frac{TP+TN}{P+N}$ ✓
  - F1-score $= 2 \cdot \frac{precision \cdot recall}{precision+recall} = \frac{2TP}{2TP+FP+FN}$ ✓

Diagnostic Testing

$$\text{Accuracy} = \frac{TP+TN}{P+N}$$

		預測值	
		+	-
實際值	T	TP	FN
	F	FP	TN



$$\text{F1-score} = \frac{2TP}{2TP+FP+FN}$$

		預測值	
		+	-
實際值	+	TP	FN
	-	FP	TN

Precision

Recall

線性迴歸模型

```
# pseudo code
lm.fit <- lm(y ~ x1 + x2 + x3……, data=data)
```

```
Residual standard error: 94.16 on 126 degrees of freedom
Multiple R-squared:  0.7556,    Adjusted R-squared:  0.3035
F-statistic: 1.672 on 233 and 126 DF,  p-value: 0.0007661
```

```
Residual standard error: 87.33 on 275 degrees of freedom
Multiple R-squared:  0.5412,    Adjusted R-squared:  0.401
F-statistic: 3.861 on 84 and 275 DF,  p-value: < 2.2e-16
```

Step Function

```
# pseudo code
lm.select <- step(lm.fit)
```

```
Residual standard error: 84.44 on 317 degrees of freedom
Multiple R-squared:  0.5055,    Adjusted R-squared:   0.44
F-statistic: 7.717 on 42 and 317 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = donor.de ~ ttl.n.words + ttl.male + ttl.youth +
    ttl.blind + ttl.dead + ttl.dig.if + n.fb.like + log.fb.share +
    log.fb.total + month + g2 + g4 + g13 + g19 + g21 + g25 +
    g36 + g38 + g41 + g46 + g47 + g49 + g50 + g54 + g58 + g64 +
    g66 + g69 + g91 + g96 + g101 + g107 + g121 + g131 + g141 +
    g155 + g157 + g158 + g161 + g166 + g169 + g200, data = d1)
```

線性迴歸模型

```
# pseudo code
lm.fit <- lm(y ~ x1 + x2 + x3……, data=data)
```

```
Residual standard error: 94.16 on 126 degrees of freedom
Multiple R-squared:  0.7556,    Adjusted R-squared:  0.3035
F-statistic: 1.672 on 233 and 126 DF,  p-value: 0.0007661
```

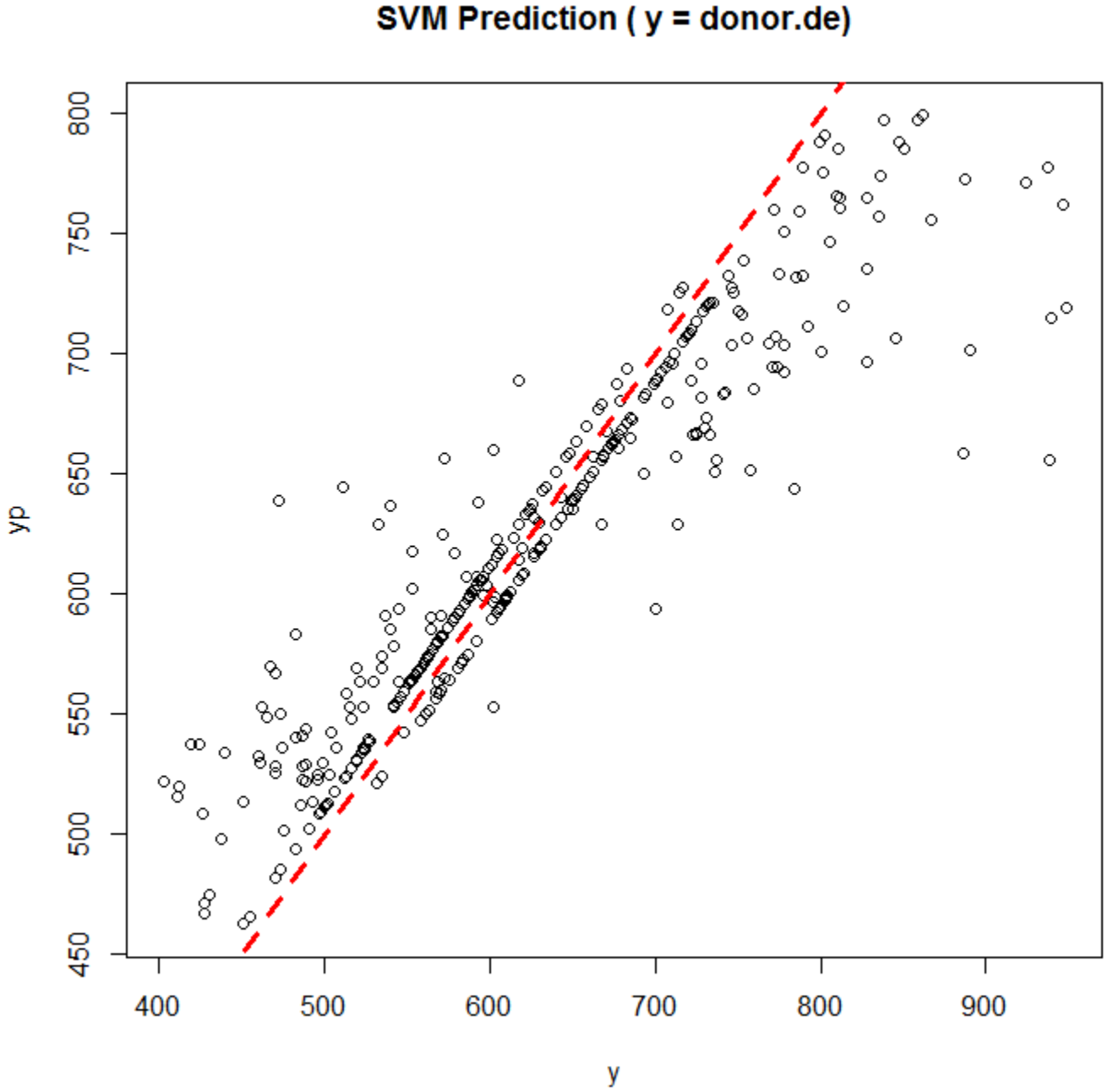
```
Residual standard error: 87.33 on 275 degrees of freedom
Multiple R-squared:  0.5412,    Adjusted R-squared:  0.401
F-statistic: 3.861 on 84 and 275 DF,  p-value: < 2.2e-16
```

```
Residual standard error: 84.44 on 317 degrees of freedom
Multiple R-squared:  0.5055,    Adjusted R-squared:   0.44
F-statistic: 7.717 on 42 and 317 DF,  p-value: < 2.2e-16
```

支持向量機 (pkg e1071)

```
# pseudo code
svm.fit <- svm(y ~ x1 + x2 + x3……, data=data)
```

```
> cor
[1] 0.9226098
> rsq
[1] 0.7850917
```


SVM Prediction (y = donor.de)
yp
y
450
500
550
600
650
700
750
800
400
500
600
700
800
900

Cross Validation

```
# leave-One-Out
i <- 1
testing <- d1[i, c(yn, x.ttl, x.fb, x.t, x.k, x.g)]
training <- d1[-i, c(yn, x.ttl, x.fb, x.t, x.k, x.g)]
svm.fit <- svm(form, data = training, kernel=kern,
                type=type)
p <- predict(fit, testing)
c(testing[, yn], p)
```

練習 D-03 (8 mins)

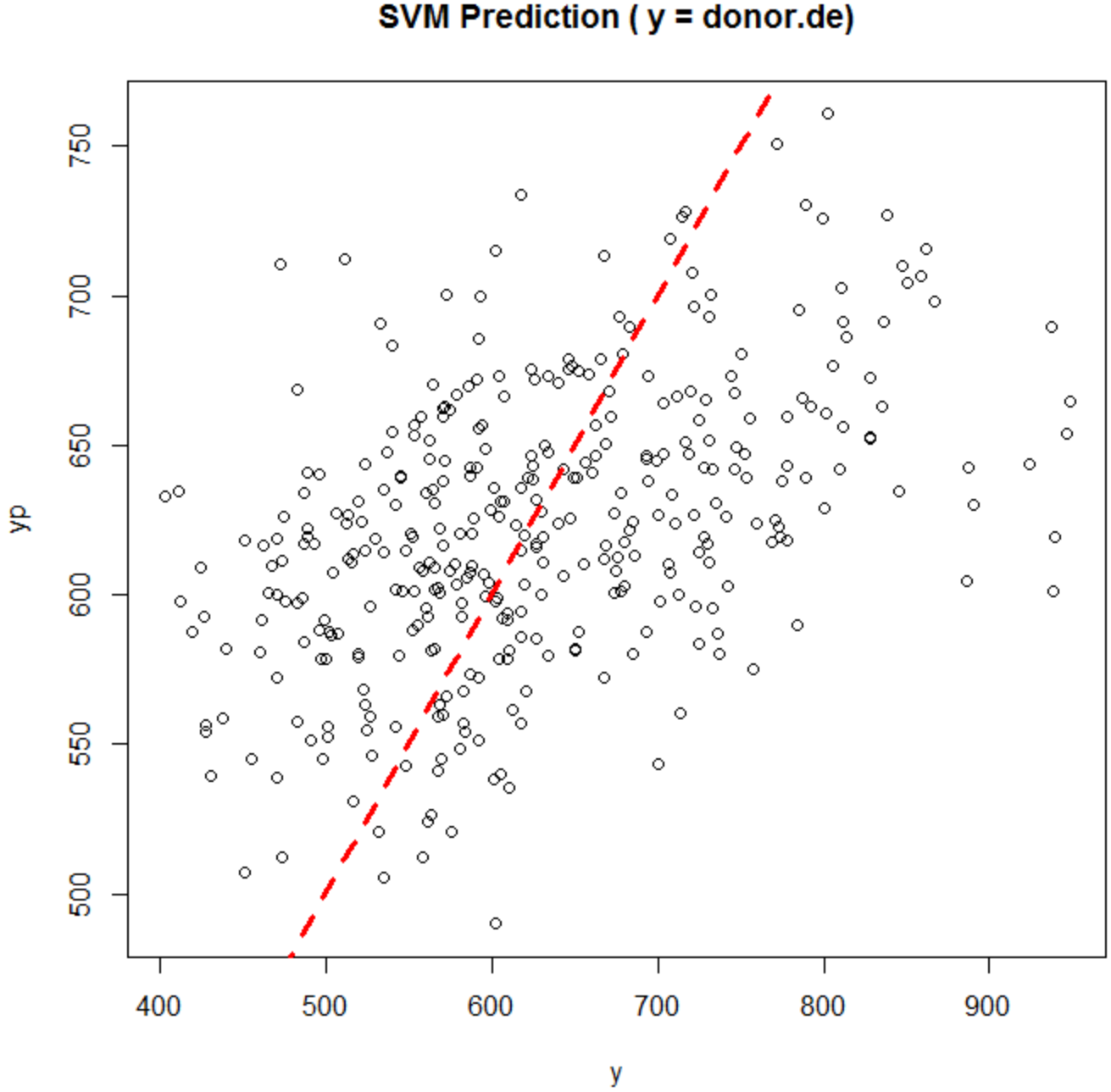
session_D_ex03.R

- 請以 leave-one-out 的方式做 SVM 的 cross validation
  - 計算 Pearson Correlation Coefficient
  - 計算 Coefficient Determination (R-squared)
  - 畫圖

練習 D-03 (解答)

- 請以 leave-one-out 的方式做 SVM 的 cross validation。
  - 計算 Pearson Correlation Coefficient
  - 計算 Coefficient Determination (R-squared)
  - 畫圖

```
> cor
[1] 0.458
> rsq
[1] 0.207
```


SVM Prediction (y = donor.de)
yp
y
500
550
600
650
700
750
400
500
600
700
800
900

練習 D-03 (進階)

session_D_ex03_bonus.R

- Bonus: 更換 SVM kernel 看結果有何不同。

Classification

- 將資料切為前 50% 與後 50% 兩類，做 SVM 的 classification。

donor.de	ttl.n.words	ttl.cancer
513.0973	12	0
597.7761	11	0
703.6155	11	0
629.4549	11	1
652.2943	11	0
489.1337	12	0
570.9731	11	0
539.4913	13	1
570.3307	12	0
535.1701	11	0
612.0095	11	0
777.8489	12	1
617.5277	11	0
719.3671	12	1
733.0459	11	0

donor.de	ttl.n.words	ttl.cancer
low	12	0
low	12	0
low	13	1
low	11	0
high	12	1
high	12	1
high	11	0
high	12	1
high	11	1
high	11	1
high	12	0
high	10	0
high	10	1
high	11	1
high	11	1

Classification (cont'd)

```
# pseudo code
# cut data in half
y <- ifelse(y > median(y), "high", "low")

# 做SVM並以leave-one-out檢查
# accuracy
acc <- mean(p$y==p$yp)
```

```
   yp
y     1   2
  1 123  57
  2  76 104
> acc
[1] 0.6305556
> fsc
[1] 0.6099707
```

```
# f1-score
fsc <- 2*TP / (2*TP+FP+FN)
```

練習 D-04 (5 mins)

- 將資料切為前 25% 與最後 25% 兩類，做 SVM 的 classification 後，計算 accuracy 與 F1-score。

練習 D-04 (解答)

session_D_ex04.R

- 將資料切為前 25% 與最後 25% 兩類，做 SVM 的 classification 後，計算 accuracy 與 F1-score。

```
# 切 upper / lower classes
quantile(d1[, yn], c(0.25, 0.75))[c("25%", "75%")]
```

```
> table(p)
   yp
y    1  2
  1 65 25
  2 22 68
> acc
[1] 0.7388889
> fsc
[1] 0.7431694
```

以文本資料建立預測模型

文本資料

- 我們有文章的資訊，機器是否可以判讀出這篇文章是哪一位 ~~Hero~~ 記者寫的呢？

aid	title	date.published	journalist	k1	g1
A2661	癌男綁便袋 盼打工撐家	2012-01-04	鋼鐵人	0.008333333	0.0000000000
A2662	翁棲陋屋 塞膠布擋雨	2012-01-05	美國隊長	0.015473888	0.0058027079
A2663	夫罹癌 婦養雞扛8口家	2012-01-06	美國隊長	0.028318584	0.0000000000
A2664	翁寒夜摘菜 扛4口吃穿	2012-01-08	美國隊長	0.027896996	0.0000000000
A2666	婦無懼癌魔 牽掛獨生女	2012-01-10	暴風女	0.003846154	0.0019230769
A2667	殘燭住陋屋 「怕家會垮」	2012-01-11	黑寡婦	0.007272727	0.0018181818
A2668	8歲童貼心 替癌父洗尿壺	2012-01-12	暴風女	0.008576329	0.0000000000
A2669	夫妻接連中風 4口拆散	2012-01-13	白皇后	0.018148820	0.0000000000
A2670	夫顧癌妻 棄場供品飽肚	2012-01-09	美國隊長	0.007797271	0.0000000000
A2671	殘夫嘆 無力養憨妻幼子	2012-01-16	美國隊長	0.016274864	0.0000000000

計算文章向量

- 上一堂課我們計算了每個文詞的詞向量，並且利用向量來看各個文詞之間的關係。
- 文章內的每個詞向量加總起來，可以用來代表每篇文章的屬性。


A3931子病女憨 寡母憂5口陷困

「我按怎嘸要緊，囝仔健康卡重要」

2016年05月16日

41歲阿育去年5月開始，因肝硬化病重無法工作，採訪時，骨瘦如柴的他，氣若游絲說，阿爸死得早，阿母不到30歲守寡做工撐養子女長大，如今他無奈因病重無法養育妻女、孝順母親。

報導•攝影／張嘉恬

阿育的61歲母親阿美嬸說，和已故先生育有2子和重度智能障礙女兒，30多年前先生過世時，排行老大的阿育11歲，老么小兒子才6歲，她當時靠著做手工、幫人洗月子衣物、打掃清潔共3份工撐養子女長大，如今兒子病重，她老來難享清福，她說：「我自己按怎嘸要緊，囝仔健康卡重要。」

阿美姨（右）準備替極重度智能障礙女兒洗淨沐浴。

兒等換肝現自費控制病情

目前阿美嬸和阿育夫妻、9歲孫女及37歲重度智能障礙女兒同住，之前生活主要靠阿育和太太，她說，自己和已逝先生雙方家族都有肝疾病史，阿育疑因遺傳，加上長期上大夜班作息顛倒，罹患肝硬化，至今2度肝昏迷住院治療，體弱氣虛難久站已無法工作，「醫生說要等換肝，現在我定期陪他到醫院打自費的白蛋白控制病情，1個月要1萬多元」。

阿育（中）目前等待換肝，因病重無法工作，右為女兒、左為母親。

文章

詞向量

撕心裂肺	0.250113823	0.288580771	-0.349648024
常叫	0.581631128	-0.168796440	-0.617956259
范玄珍說	0.059279153	-0.474639349	-0.236295720
駐台	-0.376782237	-0.797129303	0.651099000
銀枝則	-0.605435689	-0.069144042	0.080284280
麗莉	0.471099780	-0.551091980	-0.321278178
縱	0.244521769	-0.053195253	0.122502663
節儉	0.435786642	-0.150547734	-0.642815039

計算文章向量 (cont'd)

A3931子病女愁 寡母憂5口陷困

「我按怎嘸要緊，囝仔健康卡重要」

2016年05月16日

41歲阿育去年5月開始，因肝硬化病重無法工作，採訪時，骨瘦如柴的他，氣若游絲說，阿爸死得早，阿母不到30歲守寡做工撐養子女長大，如今他無奈因病重無法養育妻女、孝順母親。

報導•攝影／張嘉恬

阿育的61歲母親阿美嬌說，和已故先生育有2子和重度智能障礙女兒，30多年前先生過世時，排行老大的阿育11歲，老么小兒子才6歲，她當時靠著做手工、幫人洗月子衣物、打掃清潔共3份工撐養子女長大，如今兒子病重，她老來難享清福，她說：「我自己按怎嘸要緊，囝仔健康卡重要。」

阿美姨（右）準備替極重度智能障礙女兒洗淨沐浴。

阿育（中）目前等待換肝，因病重無法工作，右為女兒、左為母親。

兒等換肝現自費控制病情

目前阿美嬌和阿育夫妻、9歲孩女及37歲重度智能障礙女兒同住，之前生活主要靠阿育和太太，她說，自己和已逝先生雙方家族都有肝疾病史，阿育疑因遺傳，加上長期上大夜班作息顛倒，罹患肝硬化，至今2度肝昏迷住院治療，體弱氣虛難久站已無法工作，「醫生說要等換肝，現在我定期陪他到醫院打自費的白蛋白控制病情，1個月要1萬多元」。

文章

詞向量

撕心裂肺	0.250113823	0.288580771	-0.349648024
常叫	0.581631128	-0.168796440	-0.617956259
范玄珍說	0.059279153	-0.474639349	-0.236295720
駐台	-0.376782237	-0.797129303	0.651099000
銀枝則	-0.605435689	-0.069144042	0.080284280
麗莉	0.471099780	-0.551091980	-0.321278178
縱	0.244521769	-0.053195253	0.122502663
節儉	0.435786642	-0.150547734	-0.642815039

A3930	0.050570641	-0.1578603	0.06377170
A3931	0.075977638	-0.1898975	0.06608932
A3932	0.052882687	-0.1672281	0.09590967
A3933	0.046331780	-0.1613533	0.09096074
A3934	0.073454491	-0.1611603	0.11529204

練習 D-05 (8 mins)

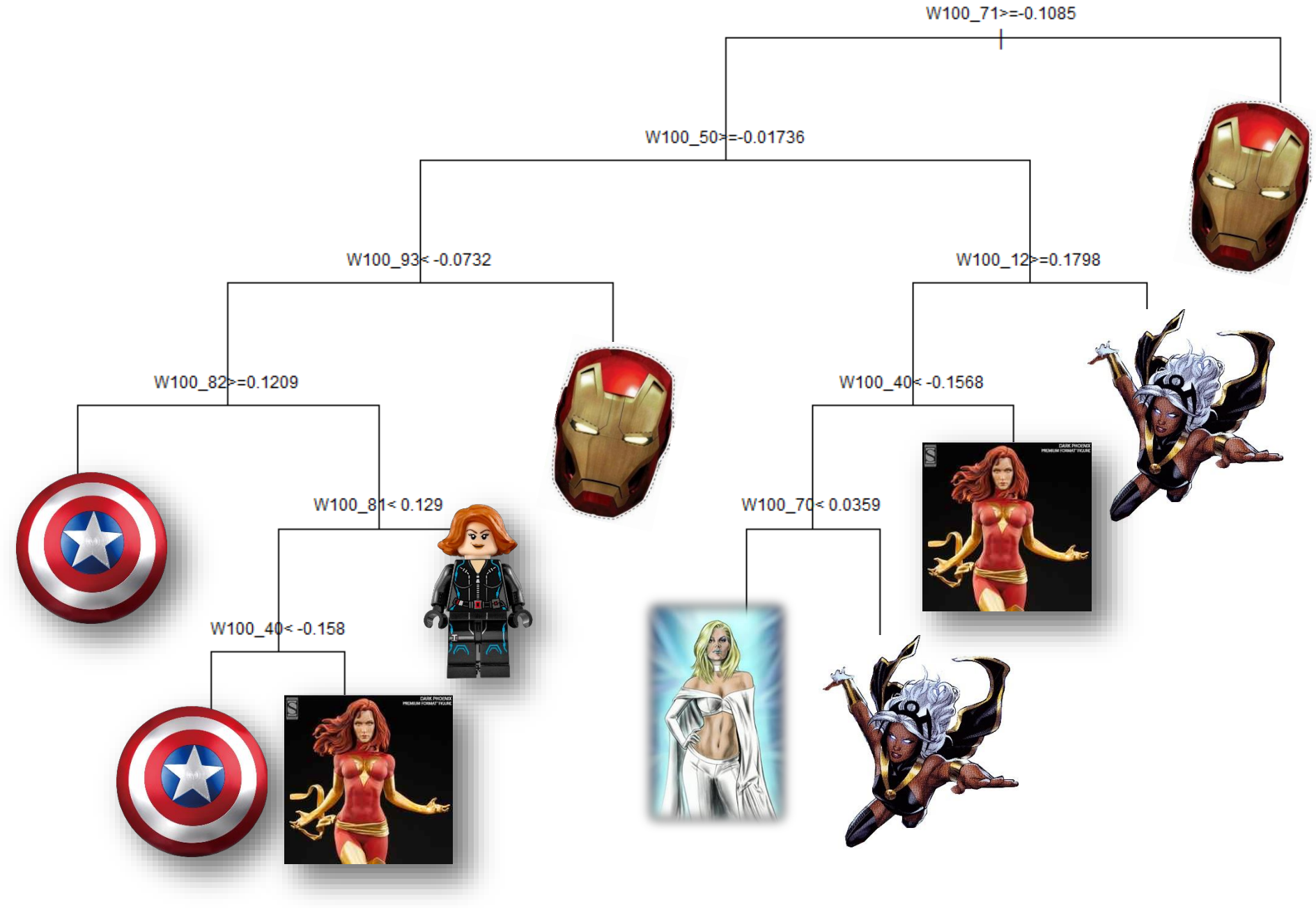
session_D_ex05.R

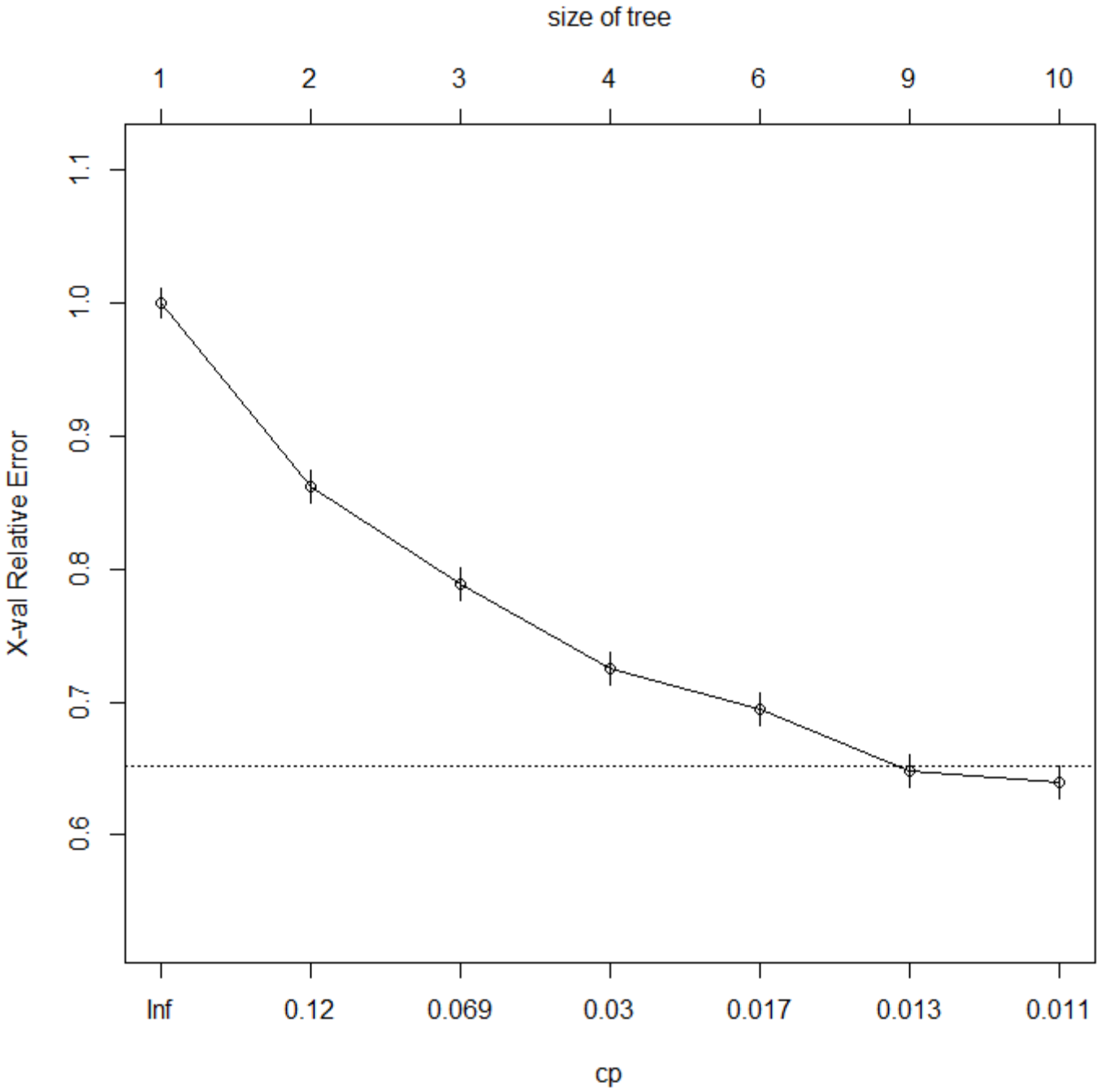
- 請將 Session C 產生的詞向量轉化成代表每篇文章的 “article_vector” (文章向量)。

Decision Tree

```
dt.fit <- rpart(form, data=training)
printcp(dt.fit)
plot(dt.fit, uniform=T)
text(dt.fit, use.n=T, cex=0.75)
```

```
Variables actually used in tree construction:
[1] W100_12 W100_40 W100_50 W100_70 W100_71 W100_81 W100_82 W100_93

Root node error: 2404/3400 = 0.70706

n= 3400
```


W100_71>=-0.1085
W100_50>=-0.01736
W100_93< -0.0732
W100_12>=0.1798
W100_82>=0.1209
W100_40< -0.1568
W100_81< 0.129
W100_70< 0.0359
W100_40< -0.158


size of tree
1
2
3
4
6
9
10
X-val Relative Error
1.1
1.0
0.9
0.8
0.7
0.6
Inf
0.12
0.069
0.03
0.017
0.013
0.011
cp

Random Forest

```
rf.fit <- randomForest(form, data=d4)
print(rf.fit)
plot(rf.fit)
```

```
Call:
 randomForest(formula = form, data = d4)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 20

        OOB estimate of  error rate: 26.85%
```

Random Forest

Confusion Matrix

	小丑女	白皇后	美國隊長	黑寡婦	鳳凰女	暴風女	鋼鐵人	class.error
小丑女	222	2	7	0	28	2	24	0.22105263
白皇后	4	78	82	0	5	56	20	0.68163265
美國隊長	3	0	639	2	12	12	22	0.07391304
黑寡婦	0	0	60	165	1	17	22	0.37735849
鳳凰女	22	0	33	1	217	7	35	0.31111111
暴風女	0	1	14	0	6	557	26	0.07781457
鋼鐵人	4	0	3	0	5	2	982	0.01405622

練習 D-06 (Homework)

- 請練習找出最好的 method 與 features 做預測模型，提昇準確率，並做 cross validation。

練習 D-06 (参考解答)

- gratio

```
> svm.acc(d4)
[1] 0.5386667
```

kratio

```
> svm.acc(d4)
[1] 0.816
```

- glv

```
> svm.acc(d4)
[1] 0.7226667
```

a2v

```
> svm.acc(d4)
[1] 0.912
```

THANK YOU ALL!!

Hard work pays off.