R-Ladies Taipei

資料預處理 Part I

Michelle/Pepper

R-Ladies Taipei

資料預處理

- 清理資料
  - 遺漏值
  - 離群值
- 資料轉換
  - 資料型態 + 格式簡介、轉換
  - 資料維度轉換
  - 資料數值轉換

R-Ladies Taipei

清理資料-遺漏值

- 缺失比率
  - 比率一定比例(如20%)可考慮將變數、觀察資料排除
- 補值方法
  - 統計量：連續變數的預設補值方法為平均數，類別變數預設的補值方法為眾數
  - 分布：依母體資料分布的均數、利用建模補值，如用Y及單一變數建立線性迴歸模型，透過模型補值
  - 自訂：連續數值常設為0、99999，類別數值常設為N/A
  - 不處理：忽略遺漏值

R-Ladies Taipei

清理資料-缺失比率

- IRIS

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

- DATA-實務資料

```
> head(data)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5           NA         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3           NA          NA          1.3         0.2    <NA>
4           NA          NA           NA         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

R-Ladies Taipei

如何知道缺失值的比例

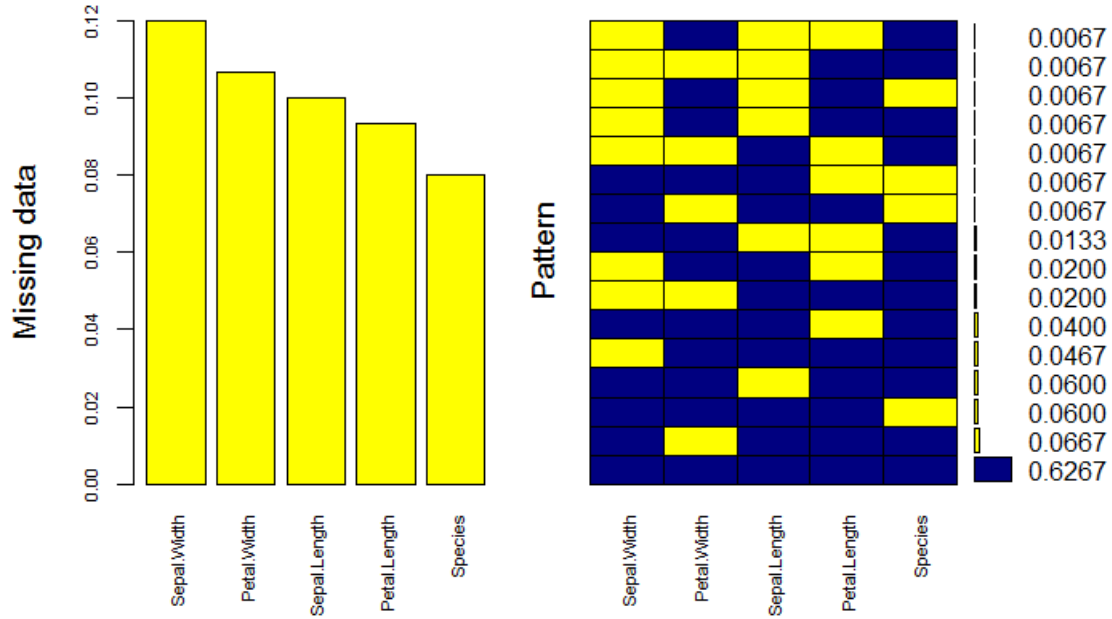
```
> summary(data)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :47
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:48
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :43
 Mean   :5.858   Mean   :3.063   Mean   :3.759   Mean   :1.184   NA's      :12
 3rd Qu.:6.400   3rd Qu.:3.325   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.700   Max.   :2.500
 NA's   :15      NA's   :18      NA's   :14      NA's   :16
```

- library(mice)
- md.pattern(data)

```
> md.pattern(data)
    Species Petal.Length Sepal.Length Petal.Width Sepal.Width
94        1            1            1           1           1  0
 9        1            1            0           1           1  1
 7        1            1            1           1           0  1
 6        1            0            1           1           1  1
10        1            1            1           0           1  1
 9        0            1            1           1           1  1
 1        1            1            0           1           0  2
 2        1            0            0           1           1  2
 3        1            0            1           1           0  2
 3        1            1            1           0           0  2
 1        0            0            1           1           1  2
 1        0            1            1           0           1  2
 1        1            0            0           1           0  3
 1        1            1            0           0           0  3
 1        1            0            1           0           0  3
 1        0            1            0           1           0  3
         12           14           15          16          18 75
```

清理資料-缺失比率(圖示)

- library(VIM)
- mice_plot <- aggr(data, col=c('navyblue','yellow'),
  numbers=TRUE, sortVars=TRUE,
  labels=names(data), cex.axis=.7,
  gap=3, ylab=c("Missing data","Pattern"))



R-Ladies Taipei

清理資料-統計量補值

- 連續變數：花萼長度、花萼寬度、花瓣長度、花瓣寬度
- 統計量：用平均數來填補遺漏值
  - mean.data <- data
  - mean.1 <- mean(mean.data[, 1], na.rm = T)

```
> mean.1
[1] 5.857778
```

  - na.rows <- is.na(mean.data[, 1])

```
> na.rows
  [1] FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE
 [19] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
 [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
 [55] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [73] FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
 [91] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
[127] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[145] FALSE FALSE FALSE FALSE FALSE
```

  - mean.data[na.rows, 1] <- mean.1

R-Ladies Taipei

清

```
> head(data)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5           NA         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3           NA          NA          1.3         0.2    <NA>
4           NA          NA           NA         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

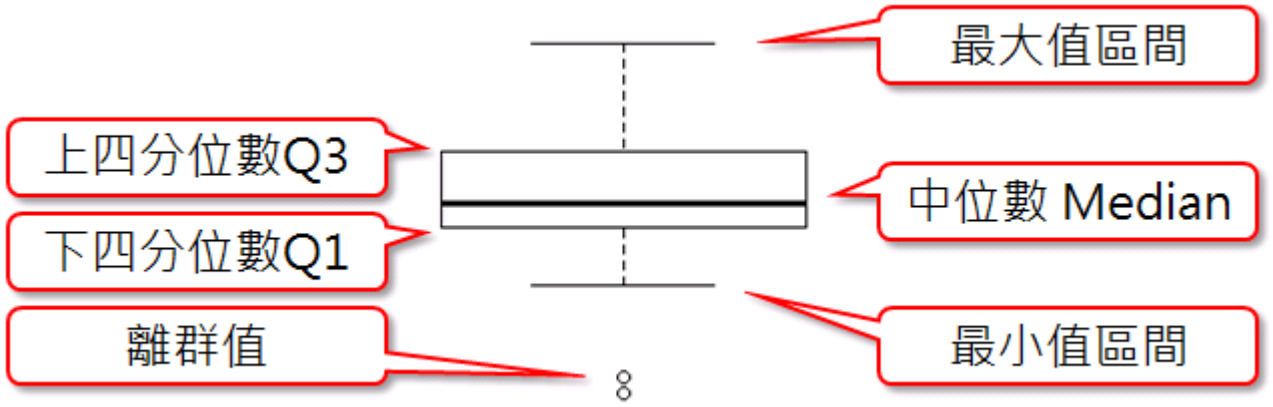
- 分布：KNN
  - 概念：找和自己很像的K個鄰居，然後從他們身上複製自己所沒有的東西
  - 語法：
    - library(DMwR)
    - imputeData <- knnImputation(data)
    - head(imputeData)

```
> head(imputeData)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1     5.100000    3.500000     1.521534         0.2  setosa
2     4.900000    3.000000     1.400000         0.2  setosa
3     4.893551    3.376589     1.300000         0.2  setosa
4     4.940000    3.420000     1.450000         0.2  setosa
5     5.000000    3.600000     1.400000         0.2  setosa
6     5.400000    3.900000     1.700000         0.4  setosa
```

R-Ladies Taipei

清理資料-離群值

- 盒型圖(box plot)

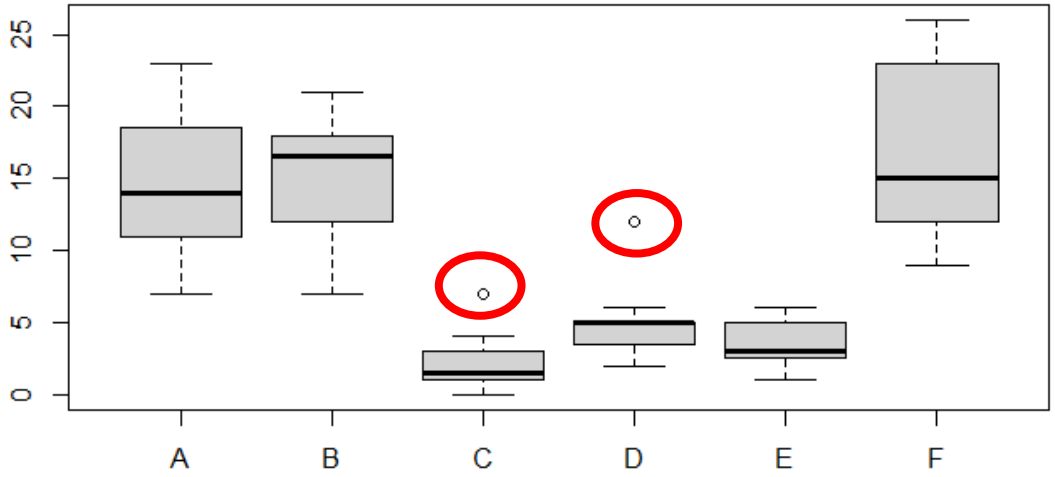


  - 最大值區間：Q3+1.5 *(Q3-Q1)
  - 最小值區間：Q1 –1.5 *(Q3-Q1)
  - 離群值：超過最大值區間或最小值區間的樣本

R-Ladies Taipei

清理資料-離群值

- 語法：
  - 盒型圖：boxplot(count ~ spray, data=InsectSprays, col="lightgray")



R-Ladies Taipei

資料轉換

- 資料型態 + 格式簡介
- 資料型態 + 格式轉換
- 資料維度轉換
- 資料數值轉換

R-Ladies Taipei

假設有一筆資料長這樣...

訂單編號	訂購單位	訂單金額
201705081102	A廠	1500
201705081331	B廠	2000
201705081612	C廠	1700

訂單編號	訂購單位	訂單金額
2.0171E+11	A廠	1500
2.0171E+11	B廠	2000
2.0171E+11	C廠	1700

R-Ladies Taipei

R的資料型態

- numeric(數值) : integer , double
- character(字串)
- logical(邏輯)
- matrix(矩陣)
- array(陣列)
- data.frame /data.table (資料表)
- list (列表)

可以參照過去RBasic的
R數據結構和格式

R-Ladies Taipei

看各自的資料型態

- typeof()
  - 與mode很相似，但更精細，是比較新的方法
  - 物件在記憶體中怎樣被記錄
- class()
  - 物件型態，沿用物件導向的概念
- attributes()
  - 物件的屬性

R-Ladies Taipei

用iris來作範例

- head(iris)
- typeof(iris)
  - list
- class(iris)
  - data.frame
- typeof(iris$Sepal.Length) #花萼的長度
  - double
- class(iris$Sepal.Length)
  - numeric

R-Ladies Taipei

用ggplot來作範例

- library(ggplot2)
- is <- ggplot(iris)
- typeof(is)
  - list
- class(is)
  - gg,ggplot

R-Ladies Taipei

- attributes(is)

```
> attributes(is)
$names
[1] "data"        "layers"      "scales"      "mapping"
[5] "theme"       "coordinates" "facet"       "plot_env"
[9] "labels"

$class
[1] "gg"     "ggplot"
```

R-Ladies Taipei

檢查讀進來的資料型態

- str()
  - 描述每個col的資料型態
- summary()
  - 資料分布情況
- Hmisc:: describe()
  - 也是資料分布，只是更為精細
  - 缺點是程式要比較久

R-Ladies Taipei

檢查讀進來的資料型態

- str(iris)

```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1
 1 1 1 1 1 1 ...
```

R-Ladies Taipei

- summary(iris)

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length
 Min.   :4.300   Min.   :2.000   Min.   :1.000
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600
 Median :5.800   Median :3.000   Median :4.350
 Mean   :5.843   Mean   :3.057   Mean   :3.758
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100
 Max.   :7.900   Max.   :4.400   Max.   :6.900
  Petal.Width          Species
 Min.   :0.100   setosa    :50
 1st Qu.:0.300   versicolor:50
 Median :1.300   virginica :50
 Mean   :1.199
 3rd Qu.:1.800
 Max.   :2.500
```

R-Ladies Taipei

- library(Hmisc)
- describe(iris)

```
> Hmisc::describe(iris)
iris

 5  Variables      150  Observations
---------------------------------------------------------------
Sepal.Length
       n  missing distinct     Info     Mean      Gmd
     150        0       35    0.998    5.843   0.9462
     .05      .10      .25      .50      .75      .90
   4.600    4.800    5.100    5.800    6.400    6.900
     .95
   7.255

lowest : 4.3 4.4 4.5 4.6 4.7, highest: 7.3 7.4 7.6 7.7 7.9
---------------------------------------------------------------
```

R-Ladies Taipei

資料轉換

- 資料型態 + 格式簡介
- 資料型態 + 格式轉換
- 資料維度轉換
- 資料數值轉換

R-Ladies Taipei

修改資料格式

資料型態	轉換函式
numeric	as.numeric()
character	as.character()
logical	as.logical()
factor	as.factor()
data.frame	as.data.frame()
matrix	as.matrix()
list	as.list()

R-Ladies Taipei

小心Factor換成數字！

- Factor(因子/因素)是R語言的資料型態，可以想像成是分類，R同時會給每一個分類打上一個編號
- m <- factor(month.abb) #月份縮寫
- as.numeric(m)
  - 5 4 8 1…
- as.numeric(month.abb)
  - NA NA NA…
  - Warning message:… 字串不能直接轉數字!

- t <- factor(c(100, 200, 300))
  - 100 200 300
  - Levels: 100 200 300
- as.numeric(t)
  - 1 2 3
- 要先轉換成字串，再換成數字
- as.numeric(as.character(t))
  - 100 200 300
- 建議盡量在讀資料時就避免這個問題
- read.csv(..., stringsAsFactors = FALSE)
- data.table::fread()

R-Ladies Taipei

資料轉換

- 資料型態 + 格式簡介
- 資料型態 + 格式轉換
- 資料維度轉換
- 資料數值轉換

R-Ladies Taipei

資料維度轉換

- 轉置(行轉成列，列轉成行)
- as.data.frame(t(x))
- data.table::transpose(x)

R-Ladies Taipei

- library(reshape2)
- dcast()
  - 長的數據拉開變寬
- melt()
  - 寬的數據濃縮變長

可以參照過去RBasic的
R套件教學

R-Ladies Taipei

JSON轉data.frame

- 政府開放資料 :https://data.gov.tw/dataset/25940

- library(rjson)
- cars <- fromJSON(file = '路外停車資訊.json')
- attributes(cars)
- cars$parkingLots[1]
- library(dplyr)
- cars_df <- bind_rows(cars$parkingLots)
- View(cars_df)

R-Ladies Taipei

資料轉換

- 資料型態 + 格式簡介
- 資料型態 + 格式轉換
- 資料維度轉換
- 資料數值轉換

R-Ladies Taipei

資料數值轉換

- 有時候資料需要作修正
  - 數字很極端，想要正規化.....
- 取決於你的模型和資料作評估

R-Ladies Taipei

資料型態	轉換函式
log	log(x, base = exp(1))
	log10(x)
Zscore	scale(x)
	(x-mean(x))/sd(x)
開根號	sqrt(x)
平方/次方	a ^ b

R-Ladies Taipei

參考資料

- http://rpubs.com/skydome20/R-Note10-Missing_Value
- https://rpubs.com/davoodastaraky/lubridate
- https://stackoverflow.com/questions/35445112/what-is-the-difference-between-mode-and-class-in-r
- http://www.cnblogs.com/nxld/p/6083731.html