大數據資料處理實務

李水彬

2023-09-01

Chapter 02

資料讀取

資料讀取分成在地機內資料讀取和網路公開資料

- 在地機內資料: 檔案名稱
  - 讀取檔案咖啡簡餐的營業資訊 MealRecord2023(Example).csv
- 網路公開資料:
  - Covid 19 data

https://covid.ourworldindata.org/data/owid-covid-data.csv

- 新竹縣生育率

網站 https://data.gov.tw/dataset/165188 資料 https://ws.hsinchu.gov.tw/001/Upload/1/opendata/8774/2170/941b2959-4902-4c15-bd2c-b919af376066.csv

載入套件

- 在使用fread() 載入資料前，必須先以library()載入data.table套件。
- 換言之，fread() 是套件 data.table 內的一個讀取資料的函數。
- 函數 function 是一組指令集，可以完成一個特定的功能。
- 套件 package 是包含許多函數的函數庫。

```
library(data.table)
```

根據R使用手冊說明，data.table這個套件提供快速且高效的記憶體，具備以下這些功能：檔案讀取和寫入、聚合、更新、等值、非等值、滾動、範圍和間隔連接，採用簡短而靈活的語法，以實現更快的開發。

本機內資料讀取

- 使用fread()函數讀取 .csv 檔案

fread(檔案名稱,,header=“auto”)

- CSV（Comma-Separated Values）是一種逗號分隔值文字檔案，它使用逗號作為數值間格。

```
setwd("E:/資料視覺化IE0260/") # 資料的路徑
Meal<-fread("MealRecord2023(Example).csv",header="auto")
```

- setwd() 設定被讀取資料**MealRecord2023(Example).csv**的路徑，你必須根據練習時資料放置的路徑，修改上列範例的路徑**E:/資料視覺化IE0260/**。
- header 告知資料首列是否為欄位變數名稱，若不確定就採本範例 header=“auto” 不用修改。
- **header=TRUE** 首列為欄位變數名稱，**header=FALSE** 首列不是為欄位變數名稱

網路公開資料

fread(網址,header="auto")

```
# 用url取代檔案名稱即可
Covid19<-fread("https://covid.ourworldindata.org/data/owid-covid-data.csv",header="auto")
```

資料內容

- 大數據資料的資料龐大，使得無法同時查看全部資料，只能看**部分資料**。
- 從部分資料的內容，亦可以了解資料的類型和重要資訊。常查看的方式:
    - 前幾筆資料
    - 後幾筆資料
    - 特定資料
    - 特定變數
- 查看資訊
    - 資料筆數、欄位名稱(變數)和變數類型
    - 資料維度 (正規型資料)
    - 值的範圍
    - 有無遺漏值
    - 不合理的數據

資料內容

- 前幾筆資料

```
head(Meal,4) # 顯示Meal 前4筆資料
```

```
##          日期    VIP_ID    性別 星期 寒暑假 特殊假日 時段           主餐       飲料
## 1: 2023/2/1 YZ_10832    Male    3      1         0 中午 黃金脆皮雞腿 錫蘭紅茶
## 2: 2023/2/1 YZ_15205    Male    3      1         0 中午 香烤法式豬排 焦糖奶茶
## 3: 2023/2/1 YZ_17931  Female    3      1         0 中午         <NA> 美式咖啡
## 4: 2023/2/1 YZ_18925    Male    3      1         0 中午   雲南椒麻雞 美式咖啡
##    氣溫 濕度 冷熱 實收
## 1: 21.2 0.44   熱  320
## 2: 21.2 0.44   冷  405
## 3: 21.2 0.44   熱   40
## 4: 21.2 0.44   冷  320
```

資料內容

- 後幾筆資料

```
tail(Meal,4) # 顯示Meal 後4筆資料
```

```
##          日期    VIP_ID    性別 星期 寒暑假 特殊假日 時段          主餐       飲料
## 1: 2023/9/3 YZ_12364    Male    7      1         0 中午 黃金脆皮雞腿 百香綠茶
## 2: 2023/9/3 YZ_13277 Female    7      1         0 中午    香煎鮭魚排 焦糖奶茶
## 3: 2023/9/3 YZ_17472    Male    7      1         0 中午 泰式椰香雞肉 梅子綠茶
## 4: 2023/9/3 YZ_15596    Male    7      1         0 中午 香烤法式豬排 錫蘭紅茶
##    氣溫 濕度 冷熱 實收
## 1: 27.9 0.83   冷  320
## 2: 27.9 0.83   冷  465
## 3: 27.9 0.83   冷  260
## 4: 27.9 0.83   熱  390
```

資料內容

- 第10筆資料的所有變數(欄位)的資料

```
Meal[5,]
```

```
##         日期    VIP_ID    性別 星期 寒暑假 特殊假日 時段         主餐       飲料 氣溫
## 1: 2023/2/1 YZ_16780 Female    3      1         0 中午 雲南椒麻雞 珍珠奶茶 21.2
##    濕度 冷熱 實收
## 1: 0.44   冷  365
```

- 第5筆資料之第7個變數(欄位)的資料

```
Meal[5,7]
```

```
##    時段
## 1: 中午
```

資料內容

- 第5筆資料之第2和第7個變數(欄位)的資料

```
Meal[5,c(2,7)] #
```

```
##      VIP_ID 時段
## 1: YZ_16780 中午
```

- 第3和第5筆資料之第2和第7個變數(欄位)的資料

```
Meal[c(3,5),c(2,7)] #
```

```
##      VIP_ID 時段
## 1: YZ_17931 中午
## 2: YZ_16780 中午
```

資料內容

- 第3~5筆資料之第2和第7個變數(欄位)的資料

```
Meal[3:5,c(2,7)] #
```

```
##      VIP_ID 時段
## 1: YZ_17931 中午
## 2: YZ_18925 中午
## 3: YZ_16780 中午
```

- 指令10:12 為10到12 連續號，例如

```
1:10 #1到10的連續號
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

資料內容

- str(): 資料摘要資訊提供資料筆數、欄位名稱(變數)和變數類型等資訊。
- MealRecord2023(Example) 資料的筆數

```
str(Meal)
```

```
## Classes 'data.table' and 'data.frame':   21504 obs. of  13 variables:
##  $ 日期    : chr  "2023/2/1" "2023/2/1" "2023/2/1" "2023/2/1" ...
##  $ VIP_ID  : chr  "YZ_10832" "YZ_15205" "YZ_17931" "YZ_18925" ...
##  $ 性別    : chr  "Male" "Male" "Female" "Male" ...
##  $ 星期    : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ 寒暑假  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ 特殊假日: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 時段    : chr  "中午" "中午" "中午" "中午" ...
##  $ 主餐    : chr  "黃金脆皮雞腿" "香烤法式豬排" NA "雲南椒麻雞" ...
##  $ 飲料    : chr  "錫蘭紅茶" "焦糖奶茶" "美式咖啡" "美式咖啡" ...
##  $ 氣溫    : num  21.2 21.2 21.2 21.2 21.2 21.2 21.2 21.2 21.2 21.2 ...
##  $ 濕度    : num  0.44 0.44 0.44 0.44 0.44 0.44 0.44 0.44 0.44 0.44 ...
##  $ 冷熱    : chr  "熱" "冷" "熱" "冷" ...
##  $ 實收    : int  320 405 40 320 365 390 335 320 280 340 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

資料內容

- str() 函數輸出的的第一段: Classes ‘data.table’ and ‘data.frame’: 21504 obs. of 13 variables。它說明這是一個資料有 21504 筆客戶用餐紀錄，每一筆紀錄有有 13 變數。
  - str() 函數除了提供資料筆數與變數個數外，它還有變數型態(variable type) 的資訊
    - chr: character
    - num: numerical
    - int: integer

資料內容

- dim(): dimension 資料維度。

```
tmp<-dim(Meal) #求算資料Meal資料的維度，暫存在變數tmp。
tmp # 顯示變數tmp內的數值
```

```
## [1] 21504    13
```

- str() 函數輸出的的第一段為資料維度的資訊，但還有變數屬性的資訊。
- dim() 僅提供正規型資料在各維度的大小。

課堂練習 1

- Covid 19 資料的筆數和維度。
- Covid 19 哪些欄位屬 character，哪些屬 numerical。

課堂練習 2

- 使用fread() 函數讀取檔案 MealRecord2023(Assignment).csv，
- 查詢前5筆資料和第20~23筆資料。
- 資料筆數和變數個數分別為多少。
- 哪些變數的屬性為chr?

值的範圍

- 數值型(num)資料

```
summary(Meal[,13])
```

```
##       實收
##  Min.   : 40.0
##  1st Qu.:320.0
##  Median :340.0
##  Mean   :344.2
##  3rd Qu.:435.0
##  Max.   :515.0
```

這摘要資訊顯示客單價(實收)

- 最小值(Min) 為 40.0 ，
- 第一個四分位數(1st Qu.) 為 320.0 ，
- 中位數(Median) 為 340.0 ，
- 平均值(Mean) 為 344.2 ，
- 第三個四分位數(3rd Qu.) 為 435.0 和
- 最大值(Max) 為 515.0 。
- 資料範圍為 40.0 ~ 515.0 。

值的範圍

- 類別型(chr)資料

```
##      時段
## 1: 中午
## 2: 晚上
```

- unique() 找出變數內所有項目。

- table() 次數分配表。

```
##
##   中午   晚上
## 12900  8604
```

課堂練習3

- 列出所有主餐名稱。
- 統計每種飲料的銷售次數。

遺漏值

- 遺漏值常以 空格、 NA或其他符號註記(-,_...)
- 當資料內有遺漏值可能會造成計算錯誤或發生終止計算的情形。
- summary() 對數值變數時，會統計遺漏值的個數。

遺漏值

```
##       濕度
##  Min.   :0.2900
##  1st Qu.:0.6000
##  Median :0.6900
##  Mean   :0.7032
##  3rd Qu.:0.7800
##  Max.   :0.9900
##  NA's   :794
```

- 這摘要資訊顯示氣溫變數共有 794 遺漏值，

遺漏值

```
##                 飲料
##  1:     錫蘭紅茶
##  2:     焦糖奶茶
##  3:     美式咖啡
##  4:     珍珠奶茶
##  5:     特調咖啡
##  6:     百香綠茶
##  7: 蒟蒻檸檬綠
##  8:           可樂
##  9:          <NA>
## 10:     椰香綠茶
## 11:     梅子綠茶
## 12: 曼特寧咖啡
```

- unique() 函數視NA為項目名稱，故會被標示出來。
- 這個例子的NA不能視為遺漏(沒有紀錄)，NA在這為客戶點餐時沒有點主餐，而不是點了沒有記錄到。
- 為了避免NA被視為類別變數的一個選項，若客戶沒有點餐最好用**沒有點餐**或其他名稱註記。

不合理的數據

- 存在的數據(非遺漏值)，其值與經驗或其他資訊不匹配。

```
##       氣溫
##  Min.   : 10.40
##  1st Qu.: 23.20
##  Median : 28.60
##  Mean   : 26.67
##  3rd Qu.: 30.90
##  Max.   :126.90
##  NA's   :794
```

- 這摘要資訊顯示氣溫的最大值為 126.90 ，顯然很不合理，(之後我們再講如何處理這樣的問題)

作業練習4

- MealRecord2023(Assignment).csv 的主餐有幾種?
- 主餐變數的遺漏值有幾筆?
- 客單價的平均值和最大值為多少?