大數據資料處理實務

李水彬

2023-09-01

Chapter 04 資料連結、彙整、合併

課程內容

- 將不同來源的資料整併成一個檔案。
- 合併天氣與點餐紀錄: 研究天氣對營業狀況的影響

```
TaoW<-fread("桃園天氣2023.csv",header="auto") #天氣
Meal<-fread("Meal(Example).csv",header="auto") #點餐紀錄
#設定資料集名稱為 Meal
```

桃園天氣

```
str(TaoW)
```

```
## Classes 'data.table' and 'data.frame':   431 obs. of  8 variables:
##  $ date       : chr  "2023/2/1" "2023/2/1" "2023/2/2" "2023/2/2" ...
##  $ week       : int  3 3 4 4 4 5 5 6 6 6 ...
##  $ vacation   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ holiday    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ phase      : int  12 18 12 12 18 12 18 12 12 18 ...
##  $ temperature: num  21.2 17.9 15.6 15.6 15.3 18.4 16 18.5 18.5 16.8 ...
##  $ humidity   : num  0.44 0.65 0.9 0.9 0.87 0.74 0.89 0.8 0.8 0.98 ...
##  $ leave      : int  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

桃園天氣

```
head(TaoW[,c(1,5,6,7)])
```

date	phase	temperature	humidity
2023/2/1	12	21.2	0.44
2023/2/1	18	17.9	0.65
2023/2/2	12	15.6	0.90
2023/2/2	12	15.6	0.90
2023/2/2	18	15.3	0.87
2023/2/3	12	18.4	0.74

點餐紀錄

```
str(Meal)
```

```
## Classes 'data.table' and 'data.frame':   23 obs. of  8 variables:
##  $ 日期  : chr  "2023/2/1" "2023/2/1" "2023/2/1" "2023/5/11" ...
##  $ VIP_ID: chr  "YZ_10832" "YZ_15205" "YZ_17931" "YZ_11975" ...
##  $ 性別  : chr  "Male" "Male" "Female" "Female" ...
##  $ 時段  : chr  "中午" "中午" "中午" "晚上" ...
##  $ 主餐  : chr  "黃金脆皮雞腿" "香烤法式豬排" NA "香煎鮭魚排" ...
##  $ 飲料  : chr  "錫蘭紅茶" "焦糖奶茶" "美式咖啡" "珍珠奶茶" ...
##  $ 冷熱  : chr  "熱" "冷" "熱" "冷" ...
##  $ 實收  : int  320 405 40 495 450 390 320 40 465 320 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

點餐紀錄

```
head((head(Meal))
```

```
knitr::kable(head(Meal), "simple")
```

日期	VIP_ID	性別	時段	主餐	飲料	冷熱	實收
2023/2/1	YZ_10832	Male	中午	黃金脆皮雞腿	錫蘭紅茶	熱	320
2023/2/1	YZ_15205	Male	中午	香烤法式豬排	焦糖奶茶	冷	405
2023/2/1	YZ_17931	Female	中午	NA	美式咖啡	熱	40
2023/5/11	YZ_11975	Female	晚上	香煎鮭魚排	珍珠奶茶	冷	495
2023/3/23	YZ_10831	Female	中午	日式蒲燒鰻魚飯	可樂	冷	450
2023/5/15	YZ_11680	Male	晚上	香烤法式豬排	可樂	冷	390

變數名稱修改

- Meal 使用”日期”和”時段”中文變數名稱，TaoW使用 “date”和”phase”英文變數名稱，改成相同的變數名稱

```
oldNames<-colnames(TaoW) #取得變數名稱
oldNames[c(1,5)]<-c("日期","時段") #更改第五個變數名稱
colnames(TaoW)<-oldNames #置換更改後的變數名稱
colnames(TaoW)#顯示置換的變數名稱，確認完成變更
```

```
## [1] "日期"        "week"        "vacation"    "holiday"     "時段"
## [6] "temperature" "humidity"    "leave"
```

結構性資料合併

- 如何在點餐紀錄中加入天氣的資料?

結構性資料合併

- 先看簡單的例子
- 學號和姓名資料的合併

```
ID<-c("i10931001", "i1093xxx2","i1093xxx4","i1093xxx6","i1093xxx8","i1093xxx9",
      "i1093xxx0","i1093xxx1")
Names<-c("趙0賢","阮0勇","何00桃","陳00賢","黃00映","潘00竹","陳0義","阮00民")
```

- ID 學號，Names 姓名，分開在兩個變數

cbind() 函數

- cbind() 函數將兩個變數(資料筆數相同)存入同一個資料變數

```
x<-cbind(ID, Names)
dim(x)
```

```
## [1] 8 2
```

- x 每筆資料登入學生的學號和姓名
- x 的維度為8, 2

```
head((head(x))
```

ID	Names
i10931001	趙O賢
i1093xxx2	阮O勇
i1093xxx4	何OO桃
i1093xxx6	陳OO賢
i1093xxx8	黃OO映
i1093xxx9	潘OO竹
i1093xxx0	陳O義
i1093xxx1	阮OO民

paste() 以字串方是貼合資料

- paste() 函數將資料合併，產生新的資料變數

```
y<-paste(ID, Names);y<-data.frame(y)
dim(y)
```

```
## [1] 8 1
```

- 表示只有一個欄位變數，與x不同。

合併檔案

```
x1<-fread("IEI030甲1.csv",header="auto")
x2<-fread("IEI030甲2.csv",header="auto")
```

```
head(x1,5)
```

```
##                       科系 年       學號   姓名
## 1: 工管系(產學國四技)  4 i10931001 趙0賢
## 2: 工管系(產學國四技)  4 i10931002 阮0勇
## 3: 工管系(產學國四技)  4 i10931004 何0桃
## 4: 工管系(產學國四技)  4 i10931006 陳0賢
## 5: 工管系(產學國四技)  4 i10931008 黃0映
```

合併檔案

```
head(x2,5)
```

```
##                       科系 年       學號   姓名
## 1: 工管系(產學國四技)  4 i10931018 阮0越
## 2: 工管系(產學國四技)  4 i10931019 阮0瓊
## 3: 工管系(產學國四技)  4 i10931020 甲0瓊
## 4: 工管系(產學國四技)  4 i10931021 阮0河
## 5: 工管系(產學國四技)  4 i10931022 范0英
```

```
dim(x2)
```

```
## [1] 21  4
```

rbind() 函數

- rbind() : 堆疊欄位變數相同的資料

```
x<-rbind(x1,x2)
```

- 資料筆數

```
dim(x1)
```

```
## [1] 19  4
```

```
dim(x2)
```

```
## [1] 21  4
```

```
dim(x)
```

```
## [1] 40  4
```

找出重複的資料

- 使用迴圈逐筆比對

```
dm2<-dim(x)
du<-array(0,dm2[1])
for(k in 2: dm2[1]){
  for(h in 1:(k-1)){
    if(x[k,4]==x[h,4]){
      du[k]<-1 #若重複設為1
    }
  }
}
du
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [39] 0 0
```

查看哪幾筆重複

```
du.idx<-which(du==1)
du.idx
```

```
## [1] 20 21 22 23 24
```

duplicated() 函數

- 使用 duplicated() 函數: 把已經出現過的資料，它的row index 標為1。

```
duplicated(x)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [37] FALSE FALSE FALSE FALSE
```

移除重複的index

```
y<-x[-du.idx,]
dim(x) #移除前的維度
```

```
## [1] 40  4
```

```
dim(y) #移除後的維度
```

```
## [1] 35  4
```

合併點餐紀錄與天氣資訊

- 創立可以連接兩資料，具有唯一性的ID

日期	時段	temperature	humidity
2023/2/1	12	21.2	0.44
2023/2/1	18	17.9	0.65
2023/2/2	12	15.6	0.90
2023/2/2	12	15.6	0.90
2023/2/2	18	15.3	0.87

日期	VIP_ID	性別	時段	主餐	飲料	冷熱	實收
2023/2/1	YZ_10832	Male	中午	黃金脆皮雞腿	錫蘭紅茶	熱	320
2023/2/1	YZ_15205	Male	中午	香烤法式豬排	焦糖奶茶	冷	405
2023/2/1	YZ_17931	Female	中午	NA	美式咖啡	熱	40

合併點餐紀錄與天氣資訊

- 兩筆資料都有日期的資訊，但日期並不具有唯一性，
- TaoW 資料集顯示分12和18兩個時段，而且2023/2/2天氣紀錄有重複。Meal 2023/2/1 有三筆不同顧客點餐紀錄。
- TaoW 和 Meal 時段的紀錄不同。

直接合併的問題

- 使用 merge() 函數合併檔案

```
y<-merge(Meal, TaoW, all=T, by="日期")
head(y,10)
```

```
##          日期    VIP_ID   性別 時段.x          主餐      飲料 冷熱 實收 week
##  1:  2023/2/1 YZ_10832   Male    中午 黃金脆皮雞腿 錫蘭紅茶    熱  320    3
##  2:  2023/2/1 YZ_10832   Male    中午 黃金脆皮雞腿 錫蘭紅茶    熱  320    3
##  3:  2023/2/1 YZ_15205   Male    中午 香烤法式豬排 焦糖奶茶    冷  405    3
##  4:  2023/2/1 YZ_15205   Male    中午 香烤法式豬排 焦糖奶茶    冷  405    3
##  5:  2023/2/1 YZ_17931 Female    中午         <NA> 美式咖啡    熱   40    3
##  6:  2023/2/1 YZ_17931 Female    中午         <NA> 美式咖啡    熱   40    3
##  7:  2023/2/1 YZ_19839   Male    中午 香烤法式豬排 特調咖啡    熱  390    3
##  8:  2023/2/1 YZ_19839   Male    中午 香烤法式豬排 特調咖啡    熱  390    3
##  9: 2023/2/10     <NA>   <NA>    <NA>         <NA>     <NA> <NA>   NA    5
## 10: 2023/2/10     <NA>   <NA>    <NA>         <NA>     <NA> <NA>   NA    5
##     vacation holiday 時段.y temperature humidity leave
##  1:        1       0     12        21.2     0.44     0
##  2:        1       0     18        17.9     0.65     0
##  3:        1       0     12        21.2     0.44     0
##  4:        1       0     18        17.9     0.65     0
##  5:        1       0     12        21.2     0.44     0
```

直接合併的問題

- all=T, by="日期": 根據日期變數合併檔案，Meal 的每一筆都要連接 TaoW 上相同日期的資料。
- 2023/2/1 在Meal 資料集有3 筆，在 TaoW 有2筆，合併後 y 則會出現 $3 \times 2 = 6$ 筆
- 事實上，2023/2/1 在Meal 資料集的時段都是中午，因為連接時沒有考慮時段，會造成重複但是錯誤的連接。

將相同變數的記錄方式修改成具有一致性

在Meal 和 TaoW 的時段記錄方式不同，修改成以"中午"和"晚上"的紀錄方式

```
noon.idx<-which(TaoW$時段==12) #找出中午時段的row index
TaoW$時段[noon.idx]<-"中午" # 12 改成中午
TaoW$時段[-noon.idx]<-"晚上" # 18 改成晚上
head(TaoW,10)
```

```
##          日期 week vacation holiday 時段 temperature humidity leave
##  1: 2023/2/1    3        1       0 中午        21.2     0.44     0
##  2: 2023/2/1    3        1       0 晚上        17.9     0.65     0
##  3: 2023/2/2    4        1       0 中午        15.6     0.90     0
##  4: 2023/2/2    4        1       0 中午        15.6     0.90     0
##  5: 2023/2/2    4        1       0 晚上        15.3     0.87     0
##  6: 2023/2/3    5        1       0 中午        18.4     0.74     0
##  7: 2023/2/3    5        1       0 晚上        16.0     0.89     0
##  8: 2023/2/4    6        1       0 中午        18.5     0.80     0
##  9: 2023/2/4    6        1       0 中午        18.5     0.80     0
## 10: 2023/2/4    6        1       0 晚上        16.8     0.98     0
```

移除重複的資料

```
yes<-duplicated(TaoW) # 判斷是否重複
du.idx<-which(yes) # 重複的row index
du.idx
```

```
## [1] 4 9
```

移除重複的資料

```
TaoW<-TaoW[-du.idx]
TaoW
```

```
##          日期 week vacation holiday 時段 temperature humidity leave
##   1: 2023/2/1    3        1       0 中午        21.2     0.44     0
##   2: 2023/2/1    3        1       0 晚上        17.9     0.65     0
##   3: 2023/2/2    4        1       0 中午        15.6     0.90     0
##   4: 2023/2/2    4        1       0 晚上        15.3     0.87     0
##   5: 2023/2/3    5        1       0 中午        18.4     0.74     0
##  ---
## 425: 2023/9/1    5        1       0 晚上        29.1     0.76     0
## 426: 2023/9/2    6        1       0 中午        27.1     0.94     0
## 427: 2023/9/2    6        1       0 晚上        29.2     0.78     0
## 428: 2023/9/3    7        1       0 中午        27.9     0.83     0
## 429: 2023/9/3    7        1       0 晚上        27.0     0.75     1
```

創立唯一性變數欄位

- 將日期與時段變數合併

```
TaoW$DT<-paste(TaoW$日期, TaoW$時段, sep="-")
Meal$DT<-paste(Meal$日期, Meal$時段, sep="-")
```

```
head(TaoW)
```

```
##        日期 week vacation holiday 時段 temperature humidity leave           DT
## 1: 2023/2/1    3        1       0 中午        21.2     0.44     0 2023/2/1-中午
## 2: 2023/2/1    3        1       0 晚上        17.9     0.65     0 2023/2/1-晚上
## 3: 2023/2/2    4        1       0 中午        15.6     0.90     0 2023/2/2-中午
## 4: 2023/2/2    4        1       0 晚上        15.3     0.87     0 2023/2/2-晚上
## 5: 2023/2/3    5        1       0 中午        18.4     0.74     0 2023/2/3-中午
## 6: 2023/2/3    5        1       0 晚上        16.0     0.89     0 2023/2/3-晚上
```

創立唯一性變數欄位

```
head(Meal)
```

```
##          日期    VIP_ID    性別 時段          主餐       飲料 冷熱 實收
## 1:  2023/2/1 YZ_10832    Male 中午    黃金脆皮雞腿 錫蘭紅茶    熱  320
## 2:  2023/2/1 YZ_15205    Male 中午    香烤法式豬排 焦糖奶茶    冷  405
## 3:  2023/2/1 YZ_17931 Female 中午          <NA> 美式咖啡    熱   40
## 4: 2023/5/11 YZ_11975 Female 晚上      香煎鮭魚排 珍珠奶茶    冷  495
## 5: 2023/3/23 YZ_10831 Female 中午 日式蒲燒鰻魚飯       可樂    冷  450
## 6: 2023/5/15 YZ_11680    Male 晚上    香烤法式豬排       可樂    冷  390
##                 DT
## 1:  2023/2/1-中午
## 2:  2023/2/1-中午
## 3:  2023/2/1-中午
## 4: 2023/5/11-晚上
## 5: 2023/3/23-中午
## 6: 2023/5/15-晚上
```

- TaoW 和 Meal 都增加DT這個變數欄位

根據唯一欄位合併

- 設定 all=T
- 從Meal 和 TaoW 中，所有日期與時段找出點餐紀錄與天氣資訊。

```
MT<-merge(Meal,TaoW, all=T, by="DT") #雙向
head(MT[,c(1:5,15)],4)
```

```
##                DT   日期.x   VIP_ID   性別 時段.x temperature
## 1: 2023/2/1-中午 2023/2/1 YZ_10832   Male   中午        21.2
## 2: 2023/2/1-中午 2023/2/1 YZ_15205   Male   中午        21.2
## 3: 2023/2/1-中午 2023/2/1 YZ_17931 Female   中午        21.2
## 4: 2023/2/1-中午 2023/2/1 YZ_19839   Male   中午        21.2
```

根據唯一欄位合併

```
MT[5:10,c(1:5,15)]
```

```
##                 DT 日期.x VIP_ID 性別 時段.x temperature
## 1:  2023/2/1-晚上   <NA>   <NA> <NA>   <NA>        17.9
## 2: 2023/2/10-中午   <NA>   <NA> <NA>   <NA>        18.1
## 3: 2023/2/10-晚上   <NA>   <NA> <NA>   <NA>        16.5
## 4: 2023/2/11-中午   <NA>   <NA> <NA>   <NA>        14.2
## 5: 2023/2/11-晚上   <NA>   <NA> <NA>   <NA>        16.1
## 6: 2023/2/12-中午   <NA>   <NA> <NA>   <NA>        27.6
```

資料筆數

```
dim(TaoW)
```

```
## [1] 429   9
```

```
dim(Meal)
```

```
## [1] 23  9
```

```
dim(MT)
```

```
## [1] 433  17
```

將天氣併入TaoW資料中

- 設定 all.x=T
- 以 Meal 中每筆資料的日期和時段，從TaoW資料找出對應的天氣資訊。

```
MT<-merge(Meal,TaoW, all.x=T, by="DT")#第一個資料
dim(MT)
```

```
## [1] 23 17
```

```
head(MT[,c(1:5,15)],4)
```

```
##                DT   日期.x   VIP_ID   性別 時段.x temperature
## 1: 2023/2/1-中午 2023/2/1 YZ_10832   Male   中午        21.2
## 2: 2023/2/1-中午 2023/2/1 YZ_15205   Male   中午        21.2
## 3: 2023/2/1-中午 2023/2/1 YZ_17931 Female   中午        21.2
## 4: 2023/2/1-中午 2023/2/1 YZ_19839   Male   中午        21.2
```

資料筆數

```
dim(Meal)
```

```
## [1] 23  9
```

```
dim(MT)
```

```
## [1] 23 17
```

- 合併後的資料筆數與Meal相同。

根據唯一欄位合併

```
MT[5:10,c(1:5,15)]
```

```
##                 DT    日期.x   VIP_ID   性別 時段.x temperature
## 1: 2023/2/20-晚上 2023/2/20 YZ_11694   Male   晚上        14.1
## 2:  2023/2/6-晚上  2023/2/6 YZ_10831 Female   晚上        15.9
## 3: 2023/3/23-中午 2023/3/23 YZ_10831 Female   中午        29.5
## 4: 2023/3/23-中午 2023/3/23 YZ_15248 Female   中午        29.5
## 5: 2023/4/14-中午 2023/4/14 YZ_14756 Female   中午        27.6
## 6: 2023/4/24-中午 2023/4/24 YZ_12506 Female   中午          NA
```

課堂練習 6

- math.csv 選修管理數學學生的成績，stat.csv選修統計學學生的成績。請將兩個檔案合併。比較使用all=T, all.x=T和all.y=T合併的差異。