大數據資料處理實務

李水彬

2023-09-01

講解讀取.csv檔案進行資料分析。

Chapter 06-4 統計量-檔案資料分析

R code

- 讀取檔案 fread()

```
library(data.table) #引用data.table的函數
x<-fread("桃園天氣2023.csv",header="auto")
```

- 顯示資料欄位

```
colnames(x)#顯示資料的欄位名稱(變數名稱)
```

```
## [1] "date"        "week"        "vacation"    "holiday"     "phase"
## [6] "temperature" "humidity"    "leave"
```

```
dim(x) #資料維度
```

```
## [1] 431   8
```

重複數據

```
du.idx<-which(duplicated(x)) # 檢查是否重複
du.idx
```

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

```
x<-x[-du.idx,] #移除重複
dim(x)
```

```
## [1] 429   8
```

- 減少兩筆重複的數據

遺漏值

- 計算temperature平均值&中位數

```
mean(x$temperature) # 平均值mean
```

```
## [1] NA
```

```
median(x$temperture) #中位數 median
```

```
## NULL
```

- 變數內有遺漏值(NA)，無法得出統計量。

排除遺漏值

- 設定 na.rm=TRUE 排除遺漏值

```
# 平均值
tmean<-mean(x$temperature,na.rm=TRUE) #排除NA
# 中位數
tmedian<-median(x$temperature,na.rm=TRUE) #排除NA
# 標準差
tsd<-sd(x$temperature,na.rm=TRUE) #排除NA
tmean;tmedian;tsd
```

```
## [1] 25.25266
```

```
## [1] 27
```

```
## [1] 5.933553
```

課堂練習 18

計算 humidity 的中位數、平均數、四分位距和標準差。

部分資料的統計量

```
str(x)
```

```
## Classes 'data.table' and 'data.frame':   429 obs. of  8 variables:
##  $ date       : chr  "2023/2/1" "2023/2/1" "2023/2/2" "2023/2/2" ...
##  $ week       : int  3 3 4 4 5 5 6 6 7 7 ...
##  $ 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 18 12 18 12 18 12 18 ...
##  $ temperature: num  21.2 17.9 15.6 15.3 18.4 16 18.5 16.8 17.2 14.6 ...
##  $ humidity   : num  0.44 0.65 0.9 0.87 0.74 0.89 0.8 0.98 0.99 0.99 ...
##  $ leave      : int  0 0 0 0 0 0 0 0 0 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

- 有兩個phase (12:中午, 18:傍晚) 的資料

篩選資料

- subset(資料, 列的條件, 欄的條件)

```
x1<-subset(x,phase==12) # 選變數phase=12的資料
x2<-subset(x,phase==18) # 選變數phase=18的資料
dim(x1);dim(x2)
```

```
## [1] 214   8
```

```
## [1] 215   8
```

篩選資料

```
y1<-subset(x,phase==12,c("date","humidity"))
#選變數phase=12，保留日期(date)與濕度(humidity)兩個欄位
dim(y1)
```

```
## [1] 214   2
```

- [1]放假日[2]晚上的溫、濕度含日期的資料。

```
y2<-subset(x,phase==18&week==7,c("date","temperature","humidity"))
#選變數phase=12，保留日期(date)，溫度(temperature)與濕度(humidity)兩個欄位
dim(y2)
```

```
## [1] 31  3
```

篩選資料

```
head(y2,5)
```

```
##         date temperature humidity
## 1:  2023/2/5        14.6     0.99
## 2: 2023/2/12        20.9     0.82
## 3: 2023/2/19        15.4     0.83
## 4: 2023/2/26        12.6     0.63
## 5:  2023/3/5        17.1     0.45
```

課堂練習 19

- 取出星期日中午的溫、濕度資料。
- 取出中午溫度大於20的資料。

篩選單一月份的資料

```
x$month<-month(x$date) #將date轉成月份，產生新變數month
```

- 二月份中午的氣溫

```
i<-2
tmp<-subset(x,month==i&phase==12,"temperature") #只選temperature資料
```

```
##  [1] 21.2 15.6 18.4 18.5 17.2 14.4 19.2 20.3 14.9 18.1 14.2 27.6 24.9 10.4 10.5
## [16] 16.5 18.0 22.9 17.8 14.4 13.8 17.6 16.0 14.7 12.7 12.5 18.1 20.8
```

計算篩選後的平均值

```
mean(tmp$temperature,na.rm=TRUE) # na.rm 排除NA
```

```
## [1] 17.18571
```

使用迴圈計算每個月的平均氣溫

```
tm<-array(0,8) # 宣告一個8x1的陣列, 初始值為0
for(i in 2:9){
  tmp<-subset(x,month==i & phase==12,select="temperature") #只選temperature資料
  tm[i-1]<-mean(tmp$temperature,na.rm=TRUE) # na.rm 排除NA
}
names(tm)<-2:9
tm
```

```
##        2        3        4        5        6        7        8        9
## 17.18571 20.97097 24.29091 26.94516 30.04138 31.91613 31.01613 28.90000
```

For迴圈 說明

```
for(x in groups){
……
statements
……
}
```

- if for all x in groups, the statements in blankets will run.

For迴圈範例

- 計算 $\sum_{i=1}^{10} i$

```
ss<- 0# set the initial value of variable ss to zero.
for(i in 1:10){
ss<-ss+i
}
ss
```

```
## [1] 55
```

課堂練習 20

- 計算每個月傍晚的平均濕度
- 計算每個月中午的平均濕度