Project: Programming with R

Tony Yao-Jen Kuo

Project Description

Project source

- Assignment from Programming with R

Write 3 functions to interact with data

- `pollutantmean(directory, pollutant, id = 1:332)`

Write 3 functions to interact with data

- `pollutantmean(directory, pollutant, id = 1:332)`

- `complete(directory, id = 1:332)`

Write 3 functions to interact with data

- `pollutantmean(directory, pollutant, id = 1:332)`
- `complete(directory, id = 1:332)`
- `corr(directory, threshold = 0)`

Getting data

specdata.zip

How to download, unzip data with R?

- `download.file()` for downloading
- `unzip()` for unzipping

About data

- 332 CSV files after unzipping
- Each CSV file has 4 variables

Function 1

Try to calculate the mean value of certain pollutant from different stations

```
pollutantmean(directory, pollutant, id = 1:332)
```

Hints for function 1

- Set `na.rm = TRUE` in `mean()` if there are NAs

Sample outputs

```
my_dir <- "/Users/kuoyaojen/Downloads/specdata"
pollutantmean(my_dir, "sulfate", 1:10)
```

```
## [1] 4.064128
```

```
pollutantmean(my_dir, "nitrate", 70:72)
```

```
## [1] 1.706047
```

```
pollutantmean(my_dir, "nitrate", 23)
```

```
## [1] 1.280833
```

Function 2

Try to calculate how many complete rows are in different CSV files

```
complete(directory, id = 1:332)
```

Hints for function 2

- Use `complete.cases()` to get complete rows from a data frame

Sample output 1

```r
my_dir <- "/Users/kuoyaojen/Downloads/specdata"
complete(my_dir, 1)
```

```
##   id nobs
## 1  1  117
```

```r
complete(my_dir, c(2, 4, 8, 10, 12))
```

```
##   id nobs
## 1  2 1041
## 2  4  474
## 3  8  192
## 4 10  148
## 5 12   96
```

Sample output 2

```r
complete(my_dir, 30:25)
```

```
##   id nobs
## 1 30  932
## 2 29  711
## 3 28  475
## 4 27  338
## 5 26  586
## 6 25  463
```

```r
complete(my_dir, 3)
```

```
##   id nobs
## 1  3  243
```

Function 3

Try to calculate the correlation coefficient for CSV files, which have complete observations over `threshold`

```
corr(directory, threshold = 0)
```

Hints for function 3

- Use `cor(x, y, use = "pairwise.complete.obs")` function for correlation coefficient

Sample output 1

Sample output 2

```r
my_dir <- "/Users/kuoyaojen/Downloads/specdata"
cr <- corr(my_dir, 150)
head(cr)
```

```
## [1] -0.01895754 -0.14051254 -0.04389737 -0.06815956 -0.
```

```r
summary(cr)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -0.21057 -0.05147  0.09333  0.12401  0.26836  0.76313
```

Sample output 3

```
cr <- corr(my_dir, 400)
head(cr)
```

```
## [1] -0.01895754 -0.04389737 -0.06815956 -0.07588814  0.7
```

```
summary(cr)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.17623 -0.03109  0.10021  0.13969  0.26849  0.76313
```

Sample output 4

```
cr <- corr(my_dir, 5000)
summary(cr)
```

```
## Length  Class   Mode
##      0   NULL   NULL
```

```
length(cr)
```

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

Sample output 5

```
cr <- corr(my_dir)
summary(cr)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.00000 -0.05282  0.10718  0.13684  0.27831  1.00000
```

```
length(cr)
```

```
## [1] 323
```