R Programming Week 2: Assignment 1

Initialisation

Configure evironment

Hide

```
rm(list = ls())
setwd("~/OneDrive/datasciencecoursera/R_Programming/week2")
```

The function ‘pollutantmean’ calculates the mean of a pollutant (sulfate or nitrate) across a specified list of monitors. The function ‘pollutantmean’ takes three arguments: ‘directory’, ‘pollutant’, and ‘id’. Given a vector monitor ID numbers, ‘pollutantmean’ reads that monitors’ particulate matter data from the directory specified in the ‘directory’ argument and returns the mean of the pollutant across all of the monitors, ignoring any missing values coded as NA.

Hide

```
pollutantmean <- function(directory, pollutant, id = 1:332) {
    ## 'directory' is a character vector of length 1 indicating
    ## the location of the CSV files

    ## 'pollutant' is a character vector of length 1 indicating
    ## the name of  the pollutant for which we will calcultate the
    ## mean; either "sulfate" or "nitrate"

    ## 'id' is an integer vector indicating the monitor ID numbers
    ## to be used

    ## Return the mean of the pollutant across all monitors list
    ## in the 'id' vector (ignoring NA values)
    ## NOTE: Do not round the result
    means <- c()

    for(monitor in id){
        path <- paste(getwd(), "/", directory, "/", sprintf("%03d", monitor), ".csv", sep = "")
        monitor_data <- read.csv(path)
        interested_data <- monitor_data[pollutant]
        means <- c(means, interested_data[!is.na(interested_data)])
    }

    mean(means)
}
```

The function ‘complete’ reads a directory full of files and reports the number of completely observed cases in each data file. The function should return a data frame where the first column is the name of the file and the second column is the number of complete cases.

Hide

```r
complete <- function(directory, id = 1:332){
    ## 'director' is a character vector of length 1 indicating
    ## the location of the CSV files

    ## 'id' is an integer vector indicating the monitor ID numbers
    ## to be used

    ## Return a data frame of the from:
    ## id nobs
    ## 1  117
    ## 2  1041
    ## ...
    ## where 'id' is the monitor ID number and 'nobs' is the
    ## number of complete cases
    results <- data.frame(id=numeric(0), nobs=numeric(0))
    for(monitor in id){
        path <- paste(getwd(), "/", directory, "/", sprintf("%03d", monitor), ".csv", sep = "")
        monitor_data <- read.csv(path)
        interested_data <- monitor_data[(!is.na(monitor_data$sulfate)), ]
        interested_data <- interested_data[(!is.na(interested_data$nitrate)), ]
        nobs <- nrow(interested_data)
        results <- rbind(results, data.frame(id=monitor, nobs=nobs))
    }
    results
}
```

The function ‘corr’ takes a directory of data files and a threshold for complete cases and calculates the correlation between sulfate and nitrate for monitor locations where the number of completely observed cases (on all variables) is greater than the threshold. The function should return a vector of correlations for the monitors that meet the threshold requirement. If no monitors meet the threshold requirement, then the function should return a numeric vector of length 0.

Hide

```
corr <- function(directory, threshold = 0){
    ## 'directory' is a character vector of length 1 indicating
    ## the location of the CSV files

    ## 'threshold' is a numeric vector of length 1 indicating the
    ## number of completely observed observations (on all
    ## variables) requi?red to compute the correlation between
    ## nitrate and sulfate; the default is 0

    ## Return a numeric vector of correlations
    ## NOTE: Do not round the result!
    cor_results <- numeric(0)

    complete_cases <- complete(directory)
    complete_cases <- complete_cases[complete_cases$nobs>=threshold, ]
    #print(complete_cases["id"])
    #print(unlist(complete_cases["id"]))
    #print(complete_cases$id)

    if(nrow(complete_cases)>0){
        for(monitor in complete_cases$id){
            path <- paste(getwd(), "/", directory, "/", sprintf("%03d", monitor), ".csv", sep =
"")
            #print(path)
            monitor_data <- read.csv(path)
            #print(monitor_data)
            interested_data <- monitor_data[(!is.na(monitor_data$sulfate)), ]
            interested_data <- interested_data[(!is.na(interested_data$nitrate)), ]
            sulfate_data <- interested_data["sulfate"]
            nitrate_data <- interested_data["nitrate"]
            cor_results <- c(cor_results, cor(sulfate_data, nitrate_data))
        }
    }
    cor_results
}
```

Quiz

Q1. What value is returned by the following call to pollutantmean()? You should round your output to 3 digits.

Hide

```
pollutantmean("specdata", "sulfate", 1:10)
```

```
[1] 4.064128
```

Q2. What value is returned by the following call to pollutantmean()? You should round your output to 3 digits.

Hide

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

```
[1] 1.706047
```

Q3. What value is returned by the following call to pollutantmean()? You should round your output to 3 digits.

Hide

```
pollutantmean("specdata", "sulfate", 34)
```

```
[1] 1.477143
```

Q4. What value is returned by the following call to pollutantmean()? You should round your output to 3 digits.

Hide

```
pollutantmean("specdata", "nitrate")
```

```
[1] 1.702932
```

Q5. What value is printed at end of the following code?

Hide

```
cc <- complete("specdata", c(6, 10, 20, 34, 100, 200, 310))
print(cc$nobs)
```

```
[1] 228 148 124 165 104 460 232
```

Q6. What value is printed at end of the following code?

Hide

```
cc <- complete("specdata", 54)
print(cc$nobs)
```

```
[1] 219
```

Q7. What value is printed at end of the following code?

Hide

```
set.seed(42)
cc <- complete("specdata", 332:1)
use <- sample(332, 10)
print(cc[use, "nobs"])
```

```
 [1] 711 135  74 445 178  73  49   0 687 237
```

Q8. What value is printed at end of the following code?

Hide

```
cr <- corr("specdata")
cr <- sort(cr)
set.seed(868)
out <- round(cr[sample(length(cr), 5)], 4)
print(out)
```

```
[1]  0.2688  0.1127 -0.0085  0.4586  0.0447
```

Q9. What value is printed at end of the following code?

Hide

```
cr <- corr("specdata", 129)
cr <- sort(cr)
n <- length(cr)
set.seed(197)
out <- c(n, round(cr[sample(n, 5)], 4))
print(out)
```

```
[1] 243.0000   0.2540   0.0504  -0.1462  -0.1680   0.5969
```

Q10. What value is printed at end of the following code?

Hide

```
cr <- corr("specdata", 2000)
n <- length(cr)
cr <- corr("specdata", 1000)
cr <- sort(cr)
print(c(n, round(cr, 4)))
```

```
[1]  0.0000 -0.0190  0.0419  0.1901
```