R Programming Assignment 1: Air Pollution

Mihai M. Craiu

29/6/2020

Introduction

For this first programming assignment you will write three functions that are meant to interact with dataset that accompanies this assignment. The dataset is contained in a zip file specdata.zip that you can download from the Coursera web site.

Data

The zip file containing the data can be downloaded here:

specdata.zip

The zip file contains 332 comma-separated-value (CSV) files containing pollution monitoring data for fine particulate matter (PM) air pollution at 332 locations in the United States. Each file contains data from a single monitor and the ID number for each monitor is contained in the file name. For example, data for monitor 200 is contained in the file "200.csv". Each file contains three variables:

- Date: the date of the observation in YYYY-MM-DD format (year-month-day)
- sulfate: the level of sulfate PM in the air on that date (measured in micrograms per cubic meter)
- nitrate: the level of nitrate PM in the air on that date (measured in micrograms per cubic meter)

For this programming assignment you will need to unzip this file and create the directory 'specdata'. Once you have unzipped the zip file, do not make any modifications to the files in the 'specdata' directory. In each file you'll notice that there are many days where either sulfate or nitrate (or both) are missing (coded as NA). This is common with air pollution monitoring data in the United States.

Solution

First step

Directory have to be "C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/"

getwd()

setwd("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/")

Now we have to create a variable whit .csv data

```r
getwd()
```

```
## [1] "C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2"
```

```r
setwd("specdata")

getwd()
```

```
## [1] "C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata"
```

```r
data <- read.csv("001.csv", header = T)
```

Now we are going to create 3 functions

Part 1

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.

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 }

```r
setwd("specdata")

pollutantmean <- function(directory, pollutant, id=1:332){
  mylist <- list.files(path=directory, pattern=".csv")
  x <- numeric()
  for(i in id){
    mydata <- read.csv(mylist[i])
    x <- c(x, mydata[[pollutant]])
  }
  mean(x, na.rm=T)
}
#directory = "C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/"
```

You can see some example output from this function below. The function that you write should be able to match this output. Please save your code to a file named pollutantmean.R.

```r
setwd("specdata")

pollutantmean("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
              "sulfate", 1:10)
```

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

```r
pollutantmean("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
              "nitrate", 70:72)
```

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

```r
pollutantmean("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
              "nitrate", 23)
```

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

Part 2

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.

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 }

```r
setwd("specdata")
complete <- function(directory, id=1:332){
  mylist <- list.files(path=directory, pattern=".csv")
  nobs <- numeric()
  for(i in id){
    mydata <- read.csv(mylist[i])
    mysum <- sum(complete.cases(mydata))
    nobs <- c(nobs, mysum)
  }
  data.frame(id, nobs)
}

#directory = "C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/"
```

You can see some example output from this function below. The function that you write should be able to match this output. Please save your code to a file named complete.R. To run the submit script for this part, make sure your working directory has the file complete.R in it.

```r
setwd("specdata")

complete("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/", 1)
```

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

```
complete("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
         c(2, 4, 8, 10, 12))
```

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

```
complete("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
         30:25)
```

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

```
complete("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/", 3)
```

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

Part 3

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.

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! }

```
setwd("specdata")

corr <- function(directory, threshold = 0){
  mylist <- list.files(path=directory, pattern=".csv")
  df <- complete(directory)
  ids <- df[df["nobs"] > threshold, ]$id
  corrr <- numeric()
  for(i in ids){
    mydata <- read.csv(mylist[i])
    dff <- mydata[complete.cases(mydata), ]
    corrr <- c(corrr, cor(dff$sulfate, dff$nitrate))
  }
  return(corrr)
}

#directory = "C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/"
```

For this function you will need to use the 'cor' function in R which calculates the correlation between two vectors. Please read the help page for this function via '?cor' and make sure that you know how to use it.

You can see some example output from this function below. The function that you write should be able to approximately match this output. Note that because of how R rounds and presents floating point numbers, the output you generate may differ slightly from the example output. Please save your code to a file named corr.R. To run the submit script for this part, make sure your working directory has the file corr.R in it.

```
setwd("specdata")

cr <- corr("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
           150)
head(cr)
```

```
## [1] -0.01895754 -0.14051254 -0.04389737 -0.06815956 -0.12350667 -0.07588814
```

```
summary(cr)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.21057 -0.04999  0.09463  0.12525  0.26844  0.76313
```

```
cr_1<- corr("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
            400)
head(cr_1)
```

```
## [1] -0.01895754 -0.04389737 -0.06815956 -0.07588814  0.76312884 -0.15782860
```

```
summary(cr_1)
```

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

```
cr_2 <- corr("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
             5000)
summary(cr_2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

```
length(cr)
```

```
## [1] 234
```

```r
cr_all <- corr("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/")

summary(cr_all)
```

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

```r
length(cr_all)
```

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

Quiz Solutions

Questions

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

```r
setwd("specdata")
pollutantmean("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/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.

```r
setwd("specdata")
pollutantmean("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/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.

```r
setwd("specdata")

pollutantmean("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/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.

```
setwd("specdata")

pollutantmean("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
              "nitrate")
```

```
## [1] 1.702932
```

```
round(pollutantmean("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
                    "nitrate"),3)
```

```
## [1] 1.703
```

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

```
setwd("specdata")

cc <- complete("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/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?

```
setwd("specdata")

cc_1<- complete("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
                54)
print(cc$nobs)
```

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

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

```
setwd("specdata")

set.seed(42)
cc_2<-complete("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
               331:1)
use<-sample(332,10)
print(cc_2[use,"nobs"])
```

```
##  [1]  90 443 551 444 385 432 311  NA 125 711
```

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

```
setwd("specdata")

cr<- corr("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/")

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

```
## [1] -0.0331  0.5509  0.2621  0.1624  0.1433
```

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

```
setwd("specdata")

cr <- corr("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/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   0.1384   0.2996  -0.0648  -0.1063  -0.1405
```

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

```
setwd("specdata")

cr <- corr("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
           2000)
n <- length(cr)
cr <- corr("C:/Users/Mihai/Desktop/Data_Science_JHU_Coursera/Rprogramming/Week_2/specdata/",
           1000)
cr <- sort(cr)
print(c(n, round(cr, 4)))
```

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