HW1

2023-01-12

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6      ✓ purrr   0.3.5
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

Reproducibility component: 10 points.

1. (22 points total, equally weighted) The data set rnf6080.dat records hourly rainfall at a certain location in Canada, every day from 1960 to 1980.

a. Load the data set into R and make it a data frame called rain.df. What command did you use?

```
rain.df <- read.table("rnf6080.dat.txt") #read in the data
```

I used the read.table function to read in the .txt file.

b. How many rows and columns does rain.df have? How do you know? (If there are not 5070 rows and 27 columns, you did something wrong in the first part of the problem.)

```
dim(rain.df) #check dimensions of data
```

```
## [1] 5070   27
```

There are 5070 rows and 27 columns.

c. What command would you use to get the names of the columns of rain.df? What are those names?

```
names(rain.df) #get column names with names()
```

```
##  [1] "V1"  "V2"  "V3"  "V4"  "V5"  "V6"  "V7"  "V8"  "V9"  "V10" "V11" "V12"
## [13] "V13" "V14" "V15" "V16" "V17" "V18" "V19" "V20" "V21" "V22" "V23" "V24"
## [25] "V25" "V26" "V27"
```

```
colnames(rain.df) #get column names with colnames()
```

```
##  [1] "V1"  "V2"  "V3"  "V4"  "V5"  "V6"  "V7"  "V8"  "V9"  "V10" "V11" "V12"
## [13] "V13" "V14" "V15" "V16" "V17" "V18" "V19" "V20" "V21" "V22" "V23" "V24"
## [25] "V25" "V26" "V27"
```

d. What command would you use to get the value at row 2, column 4? What is the value?

```
rain.df[2,4] #get row 2, column 4 value
```

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

Using base R, the value at row 2, column 4 can be found with rain.df[2,4].

e. What command would you use to display the whole second row? What is the content of that row?

```
rain.df[2,] #look at second row
```

```
##   V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21
## 2 60  4  2  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0
##   V22 V23 V24 V25 V26 V27
## 2   0   0   0   0   0   0
```

rain.df[2,] prints the entire row, as shown above.

f. What does the following command do?

```
names(rain.df) <- c("year","month","day",seq(0,23))
names(rain.df)
```

```
##  [1] "year"  "month" "day"   "0"     "1"     "2"     "3"     "4"     "5"
## [10] "6"     "7"     "8"     "9"     "10"    "11"    "12"    "13"    "14"
## [19] "15"    "16"    "17"    "18"    "19"    "20"    "21"    "22"    "23"
```

This command changes the column names to year, month, day, 0, 1, 2, …, 23.

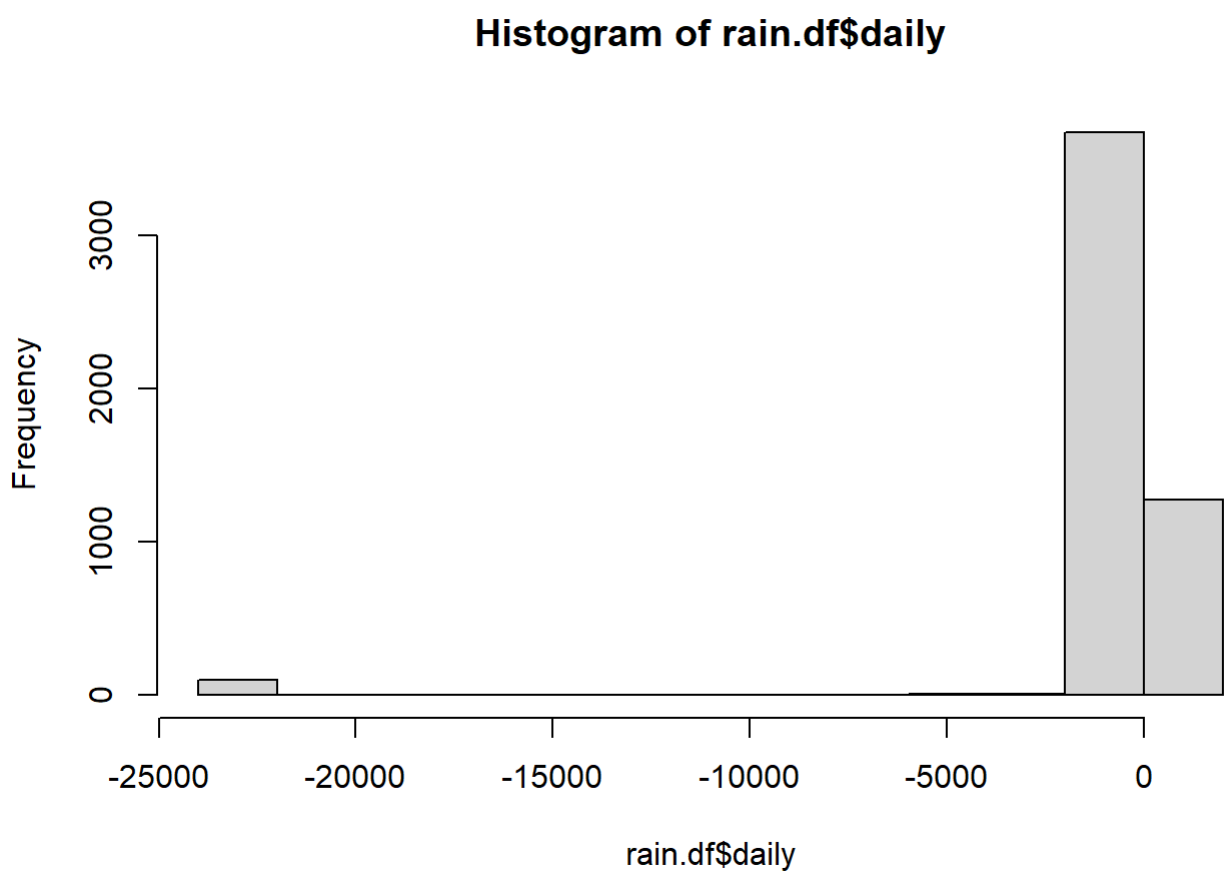
g. Create a new column called daily, which is the sum of the 24 hourly columns.

```
rain.df$daily <- rowSums(rain.df[,4:27]) #create new column called dailey with sum of columns 4-
27
summary(rain.df$daily)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -23976.0      0.0      0.0   -494.4      2.0    840.0
```

h. Give the command you would use to create a histogram of the daily rainfall amounts. Please make sure to attach your figures in your .pdf report.

```
hist(rain.df$daily) #plot a histogram of the daily column
```



i. Explain why that histogram above cannot possibly be right. The histogram has negative values, and it is not possible to have negative rainfall.

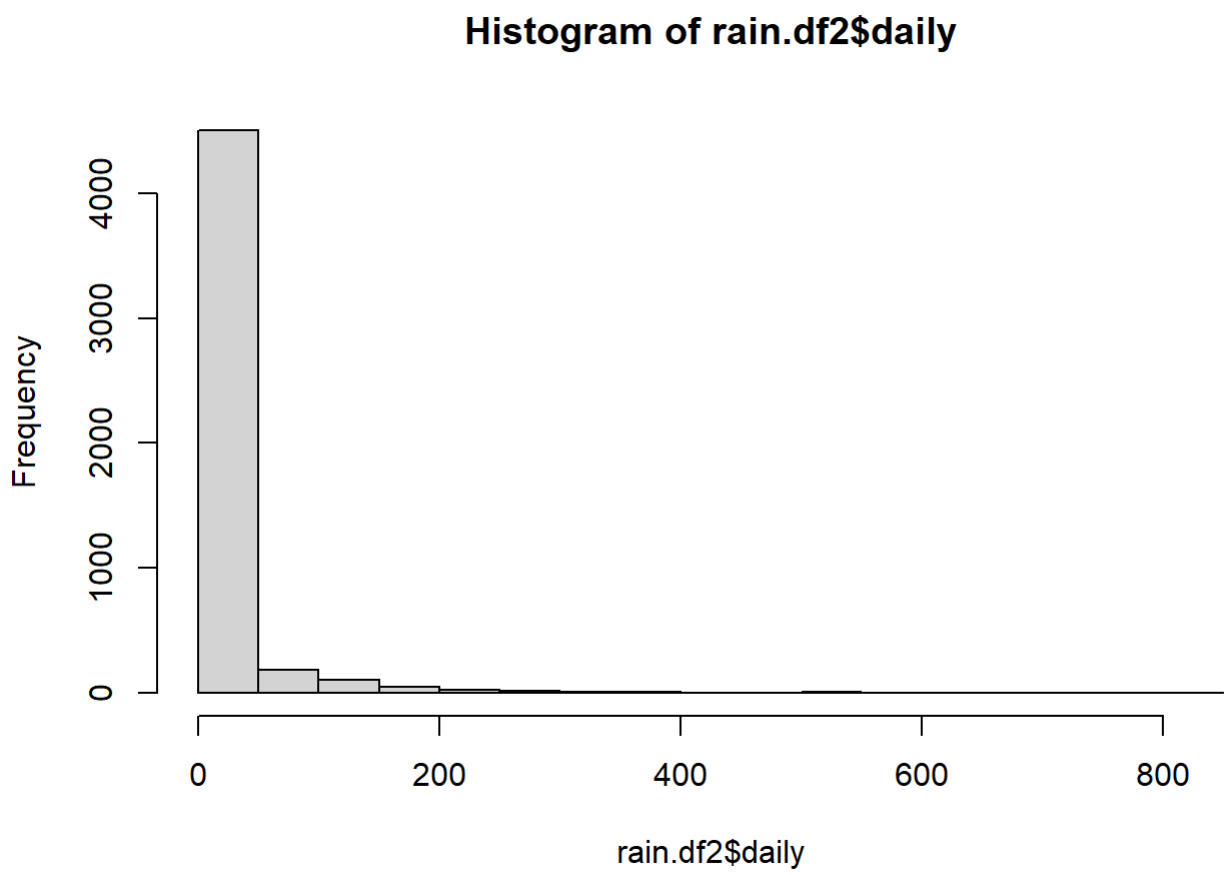
j. Give the command you would use to fix the data frame.

```
rain.df2 <- rain.df %>% filter(daily >= 0) #keep rows where daily >= 0
```

I would use the filter command in tidyverse to remove rows with negative rainfall.

k. Create a corrected histogram and again include it as part of your submitted report. Explain why it is more reasonable than the previous histogram.

```
hist(rain.df2$daily)
```



This histogram is more reasonable because all values for amount of rainfall are greater than or equal to 0.

Data types

2. (9 points, equally weighted) Make sure your answers to different parts of this problem are compatible with each other.

a. For each of the following commands, either explain why they should be errors, or explain the nonerroneous result.

```
x <- c("5","12","7")
max(x)
```

```
## [1] "7"
```

Because the values in x are characters, the max function returns the value with the maximum position (the last value).

```
sort(x)
```

```
## [1] "12" "5"  "7"
```

The sort function sorts character values in alpha-numeric order. This means the value starting with 1 goes before 5 and 7.

```
# sum(x)
```

R cannot perform mathematical operations on character values, which is why the sum function produces an error.

b. For the next two commands, either explain their results, or why they should produce errors.

```
y <- c("5",7,12)
# y[2] + y[3]
```

```
typeof(y[2])
```

```
## [1] "character"
```

```
typeof(y[3])
```

```
## [1] "character"
```

Because the first argument in y is a character, the remaining elements were read in as characters as well. Similar to the last question, R cannot perform mathematical operations on character values.

c. For the next two commands, either explain their results, or why they should produce errors.

```
z <- data.frame(z1="5",z2=7,z3=12) #create data frame with three columns
z[1,2] + z[1,3]
```

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

The first line creates a data frame with three columns, z1, z2, and z3. The second line performs 7+12. Because this was made as a data frame, the values in the first column can be different type than the values in the other columns, which is why the numeric values work here and not in the previous example.

3. (3 pts, equally weighted). a.) What is the point of reproducible code? Reproducible code is crucial in science to allow for repeated analyses and ease of interpretation.

b.) Given an example of why making your code reproducible is important for you to know in this class and moving forward. Making reproducible code is important to help the TA and professor read and interpret my assignments and give partial credit where possible. Moving forward, I will use it to help me work more efficiently and effectively on all coding projects.

c.) On a scale of 1 (easy) – 10 (hard), how hard was this assignment. If this assignment was hard (> 5), please state in one sentence what you struggled with. This assignment was a 1.