Data Wrangling in R

Leona

Due Friday August 27, 5 PM EDT

Working with data

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 data frame called

```
rain.df <- read.table( header = FALSE,
  file = "~/Desktop/STA 360/Homework/homework-1/data/rnf6080.dat")
```

b. How many rows and columns does rain.df have?

```
row_num <- nrow(rain.df)
col_num <- ncol(rain.df)
```

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?

```
col_names <-  colnames(rain.df)
col_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"
```

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

```
rain.df[2,4]
```

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

The value at row 2, column 4 is 0.

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

```
rain.df[2,]
```

```
##   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
```

f. What does the following command do?

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

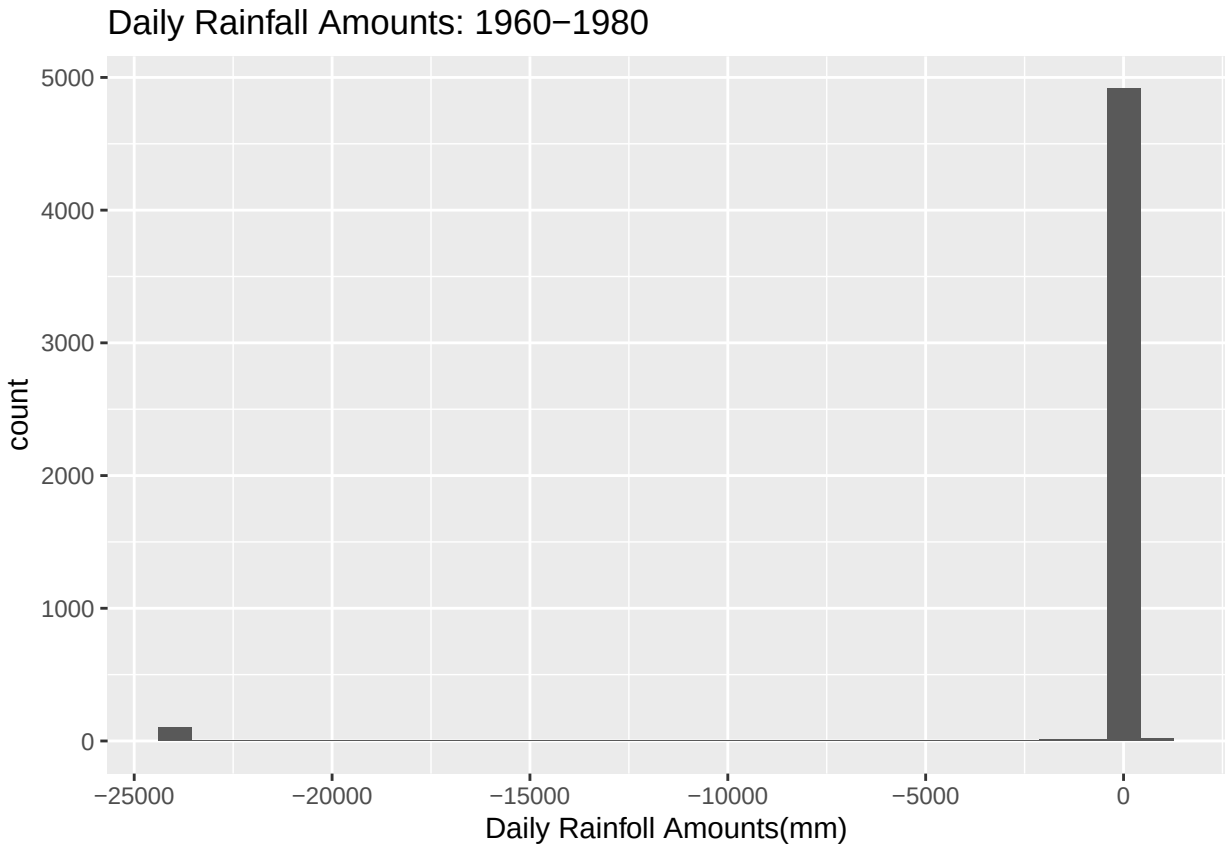
The following command set the column names of the dataframe.

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

```
rain.df$daily <- rowSums(rain.df[,4:27])
```

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.

```
library(tidyverse)
ggplot(rain.df, aes(x=daily)) +
  geom_histogram() +
  labs(title = "Daily Rainfall Amounts: 1960-1980",
       x = "Daily Rainfoll Amounts(mm)")
```



i. Explain why that histogram above cannot possibly be right.

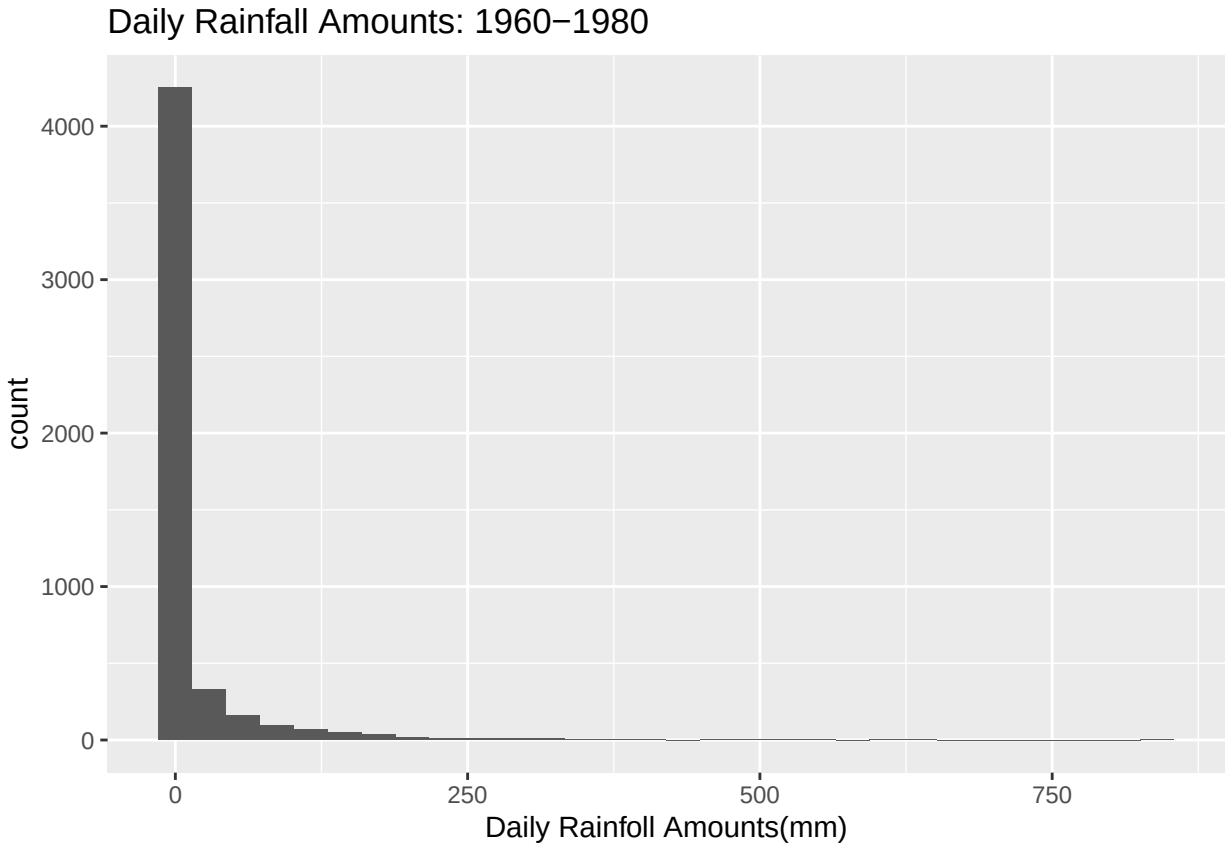
Histogram cannot be correct, because daily rainfall amount cannot be below zero.

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

```
rain.df[rain.df < 0 ] <- 0
rain.df$daily <- rowSums(rain.df[,4:27])
```

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.

```
ggplot(rain.df, aes(x=daily)) +
  geom_histogram() +
  labs(title = "Daily Rainfall Amounts: 1960-1980",
       x = "Daily Rainfoll Amounts(mm)")
```



This histogram is more reasonable because the daily rainfall amounts are all above zero. Most common daily rainfall amounts are below 200, which is normal for a location in Canada.

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 non-erroneous result.

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

Because values in x are all characters. Max and Sort function can operate on characters, but they only operate on the first character of x. Therefore, "7" is the max while "12" (1) in the min.Function sum() should produce an error, because sum can only operate on numeric type.

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]
```

- is an binary operator. However the type of y is "character" due to the character "5". Binary operator requires numeric argument while y is not.

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)
z[1,2] + z[1,3]
```

When creating dataframe, value is assigned based on the order of the command. As a result, z[1,2] will extract value from first line second column in dataframe z, which is 7.z[1,3] will extract value from first line third column in dataframe z, which is 12.

3. (3 pts, equally weighted).

a.) What is the point of reproducible code?

Reproducible code ensures that other can understand your code easily to review/ validate your method and result, increasing the reliability of your result. It can also allow people to replicate your method and create their own outputs.

b.) Given an example of why making your code reproducible is important for you to know in this class and moving forward.

Computational reproducibility will be helpful for when conducting statistical research (maybe honor thesis) moving forward. Such practice will increase the reliability for the result. Making code reproducible also makes your result less influenced by randomness for repeated testing/peer review is made possible.

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.

The assignment is around 1. Data types was a little bit challenging, but it is nice to refresh my knowledge of R fundamentals.