Data Wrangling in R

STA 360/602: Assignment 1, Spring 2019

1. Rainfall Dataset

(a) Load the data set into R and make it a data frame called rain.df.

```r
rain.df <- read.table("data/rnf6080.dat")
```

(b) How many rows and columns does rain.df have? How do you know?

There are 5070 rows and 27 columns. The number of rows and columns are indicated by command dim().

```r
dim(rain.df)
```

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

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

```r
names(rain.df)
```

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

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

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

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

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

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

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

(f) What does the following command do?

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

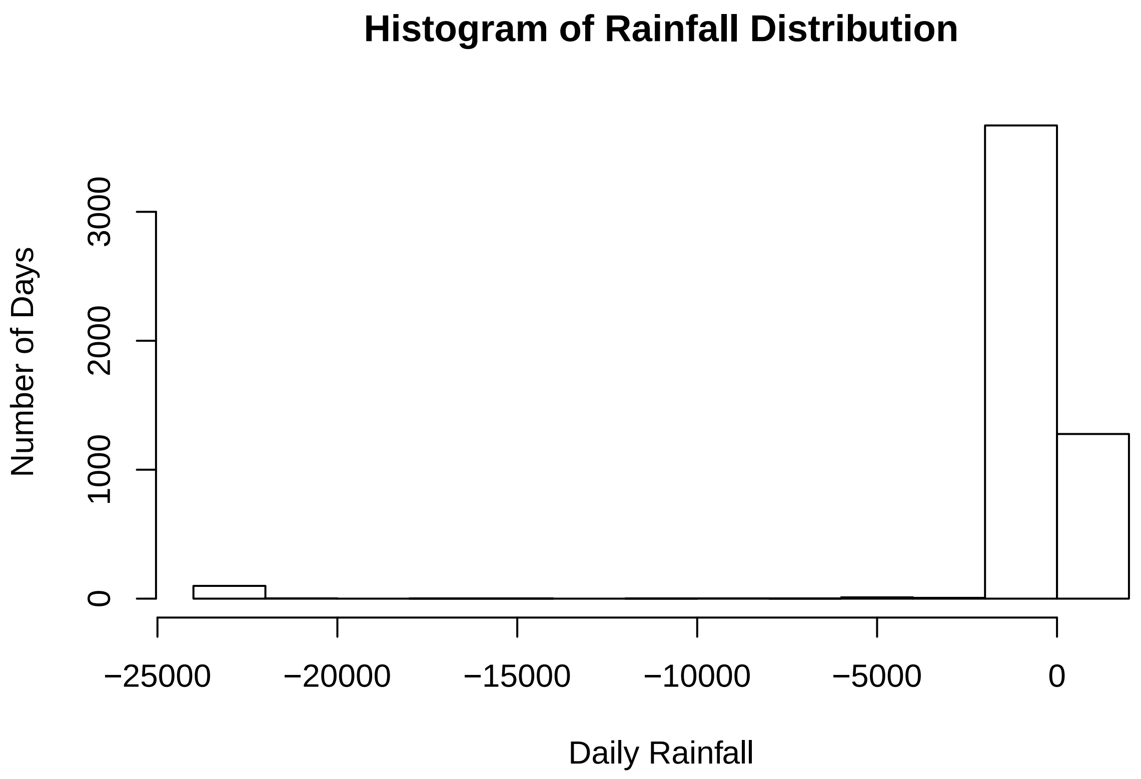
The command assigns new column names to rain.df. The first three columns are year, month, and day. The rest (sequence from 0 to 23) represents rainfall amount at specific hour (0 - 23).

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

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

(h) Give the command you would use to create a histogram of the daily rainfall amounts. Submit this histogram as a separate PDF file, named with your Duke ID, the assignment number, and Fig1.

```r
#pdf("rcs46-hw01-Fig1.pdf")
hist(rain.df$daily,main = "Histogram of Rainfall Distribution",
     xlab = "Daily Rainfall",ylab = "Number of Days")
```



```
#dev.off()
```

(i) Explain why that histogram cannot possibly be right.

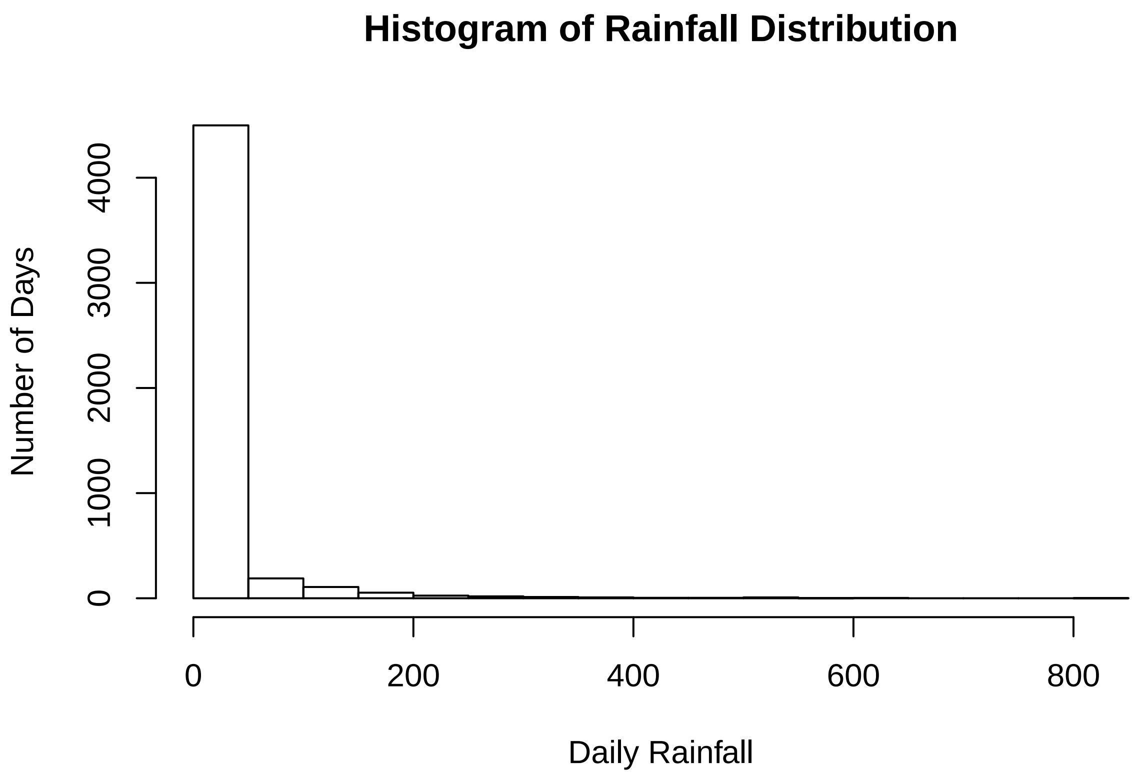
Since hourly rainful amount should be recorded as either zero or a positive number, the histogram cannot possibly be right as the data frame contains several negative numbers. Specifically, the number -999 occurs several times in the data set.

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

```
rain.df$daily[rain.df$daily < 0] <- NA
```

(k) Create a corrected histogram and submit a PDF file of it, named as before but with Fig2 instead of Fig1. Explain why it is more reasonable than the previous histogram.

```
#pdf("rcs46-hw01-Fig2.pdf")
hist(rain.df$daily,main = "Histogram of Rainfall Distribution",
     xlab = "Daily Rainfall",ylab = "Number of Days")
```



```
#dev.off()
```

This histogram is more reasonable because it removes negative numbers that does not make sense in terms of rainfall. As a result, the histogram is more appropriate.

2. Data Types

(a) For each of the following commands, either explain why they should be errors, or explain the non-erroneous result.

(1) Non-erroneous result. This command combines three characters into a vector.

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

(2) Erroneous result. Because the vector has character data type, max does not return the right numeric result.

```
max(x)
```

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

(3) Erroneous result. Character data type also leads to incorrect result for sort().

```
sort(x)
```

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

(4) Error: sum(x). Because the sum() operation will return a value if and only if the types are numeric, complex, or logical, summing up characters will produce an error.

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

(1) Non-erroneous result. This command combines one character and two integers into a vector. However, because the output type is determined from the highest type of the components in the hierarchy (integer < character), the second and third elements are converted from integer into character. As a result, the y vector contains three character type elements.

```r
y <- c("5",7,12)
```

(2) Error: y[2] + y[3]. This command adds the second element of the vector (character) to the third element of the vector (character). Because the addition operation does not support character type, adding up two characters will produce an error.

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

(1) Non-erroneous result. This command creates data frame called z, in which contains one character and two numeric values. The first element (first row, first column) is a character; the second (first row, second column) and third (first row, third column) elements are integers.

```r
z <- data.frame(z1="5",z2=7,z3=12)
```

(2) Non-erroneous result. This command adds the second (first row, second column) element to the third (first row, third column) element. Because both elements are of integer data type, the addition operation returns 19.

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

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