Data Wrangling in R

STA 360/602: Assignment 1, Spring 2019

Frank Xu

Due Tuesday, 15 January 2019, 10 AM, Sakai

R Markdown Test

Passed

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 and make it a data frame called `rain.df`. What command did you use?

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

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.)

```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
colnames(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"
```

These are default names by R because the data set does not have names assigned for each columns.

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
```

The command would give a new df with only 1 row and 27 columns, and the content of that row is every value from the rain.df with a row number of 2.

f. What does the following command do?

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

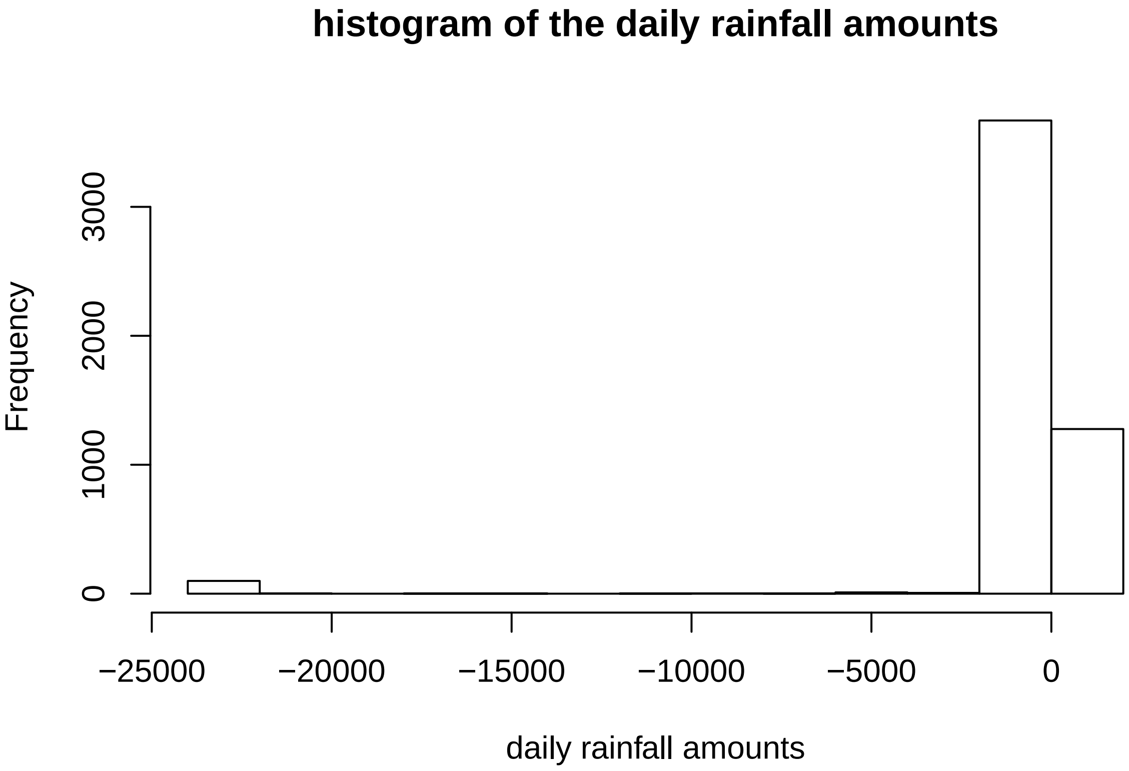
It changes the column names of rain.df respectively, into the sequence of “year”, “month”, “day”, then a sequence from 0 to 23.

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

```r
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.

```r
hist(rain.df[,28], main = "histogram of the daily rainfall amounts", xlab = "daily rainfall amounts")
```



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

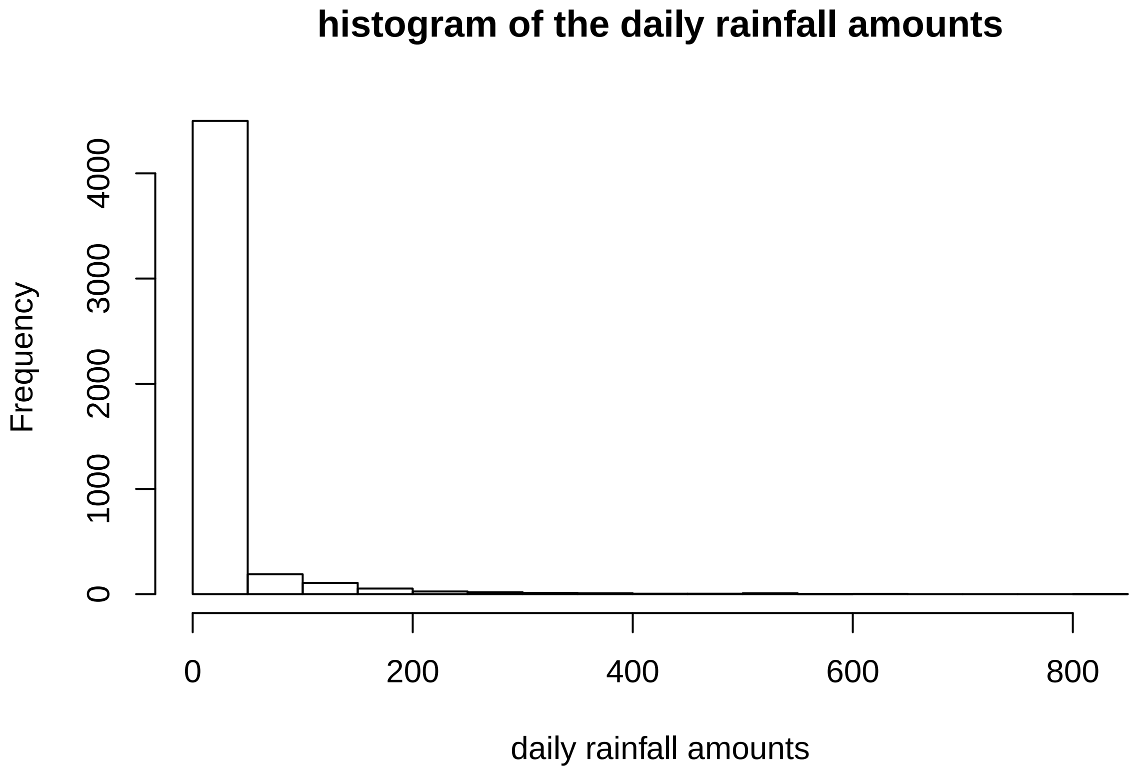
Because the precipitation cannot be a negative number.

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

```r
rain.df[rain.df < 0] <- NA
```

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.

```r
hist(rain.df[,28], main = "histogram of the daily rainfall amounts", xlab = "daily rainfall amounts")
```



Clearly it has a more reasonable range of x, at least no negative values. Also, it shows a decreasing trend as the daily rainfall amounts increased, which could be related to that most days have 0 or very little rainfall, which is natural in certain areas and thus seems more reasonable.

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")

# No error. It creates x as an vector of three characters/factors, "5", "12", "7"

max(x)

# No error. "7" will be returned because "7" is the maximum in these factors

sort(x)

# No error. "12" "5" "7" will be returned because the factors are sorted that way

sum(x)

# An error will be returned because you cannot sum characters
```

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

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

```
# No error. It creates y as an vector of three characters/factors, "5", "7", "12", because
a vector can only have 1 type of elements, and it will automatically change 7 and 12 into
characters
```

```
y[2] + y[3]
```

```
# An error will be returned because "+" do not apply on characters
```

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)
```

```
# No error. It creates z as a df with column z1 has the element "5", column z2 has
the numeric element 7, z3 has 12
```

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

```
# No error. Numeric 19 will be returned because 7 + 12 = 19
```

3. (3 pts, equally weighted).

a.) What is the point of reproducible code?

Load related packages in the beginning. Create reproducible objects. Comment.

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

If the TAs going to check my assignment answers, they can just easily copy the rmd file in their directory and run it instantly without modifying the major part of the code.

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.

1. Easy, but long.