Data Wrangling in R

Jerry Xin STA 360: Homework 1

Due Friday August 27, 5 PM EDT

Today's agenda: Manipulating data objects; using the built-in functions, doing numerical calculations, and basic plots; reinforcing core probabilistic ideas.

General instructions for homeworks: Please follow the uploading file instructions according to the syllabus. You will give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Each answer must be supported by written statements as well as any code used. Your code must be completely reproducible and must compile.

Advice: Start early on the homeworks and it is advised that you not wait until the day of. While the professor and the TA's check emails, they will be answered in the order they are received and last minute help will not be given unless we happen to be free.

Commenting code Code should be commented. See the Google style guide for questions regarding commenting or how to write code https://google.github.io/styleguide/Rguide.xml. No late homework's will be accepted.

R Markdown Test

0. Open a new R Markdown file; set the output to HTML mode and "Knit". This should produce a web page with the knitting procedure executing your code blocks. You can edit this new file to produce your homework submission.

Working with data

Total points on assignment: 10 (reproducibility) + 22 (Q1) + 9 (Q2) + 3 (Q3) = 44 points

Reproducibility component: 10 points.

```r
library(tidyverse)
library(knitr)
library(broom)
```

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
rain.df <- read.table("data/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.)

rain.df has 5070 rows and 27 columns, because it has 5070 observations of 27 variables. The 5070 observations of 27 variables is shown in the environment tab.

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" "V12"
## [13] "V13" "V14" "V15" "V16" "V17" "V18" "V19" "V20" "V21" "V22" "V23" "V24"
## [25] "V25" "V26" "V27"
```

Use the command colnames(), to get the names of the columns. The names of the columns are: "V1" "V2" "V3" "V4" "V5" "V6" "V7" "V8" "V9" "V10" "V11" "V12" "V13" "V14" "V15" "V16" "V17" "V18" "V19" "V20" "V21" "V22" "V23" "V24" "V25" "V26" "V27".

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

```
view(rain.df[2,"V4"])
```

Use view(rain.df[2,"V4"]). The value is 0.

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

```
view(rain.df[2, ])
```

Use view(rain.df[2,]). The content of that row is: 60 4 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

f. What does the following command do?

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

The command relabels the columns to year, month, day, and then the hours from 0-23, because rainfall is collected hourly.

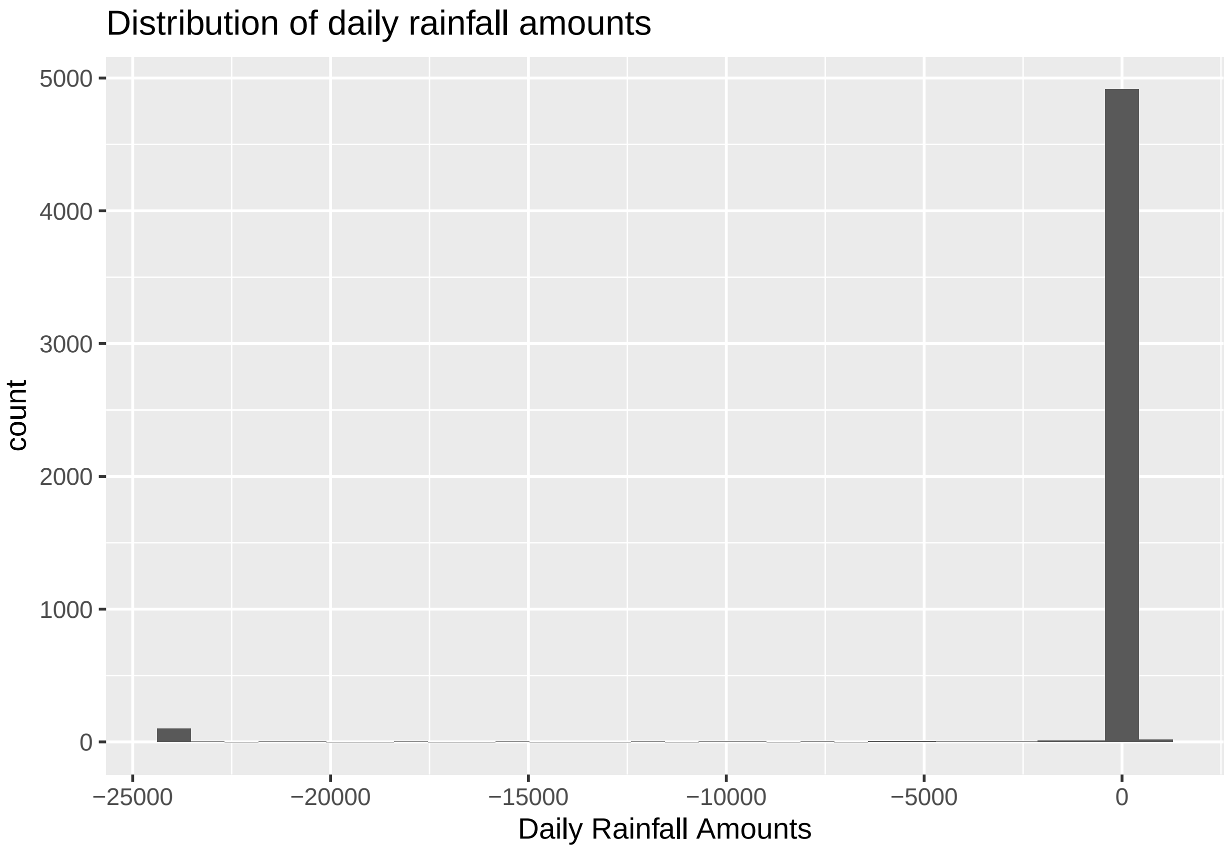
g. Create a new column called `daily`, which is the sum of the 24 hourly columns.

```
rain.df <- rain.df%>%
  mutate(daily = rowSums(.[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.

```
ggplot(data = rain.df, aes(x = daily)) +
  geom_histogram() +
  labs(x = "Daily Rainfall Amounts",
       title = "Distribution of daily rainfall amounts")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

It cannot possibly be right because there seems to be 1 daily rainfall amount that is around -24000. It doesn't make sense for rain to fall at a negative amount.

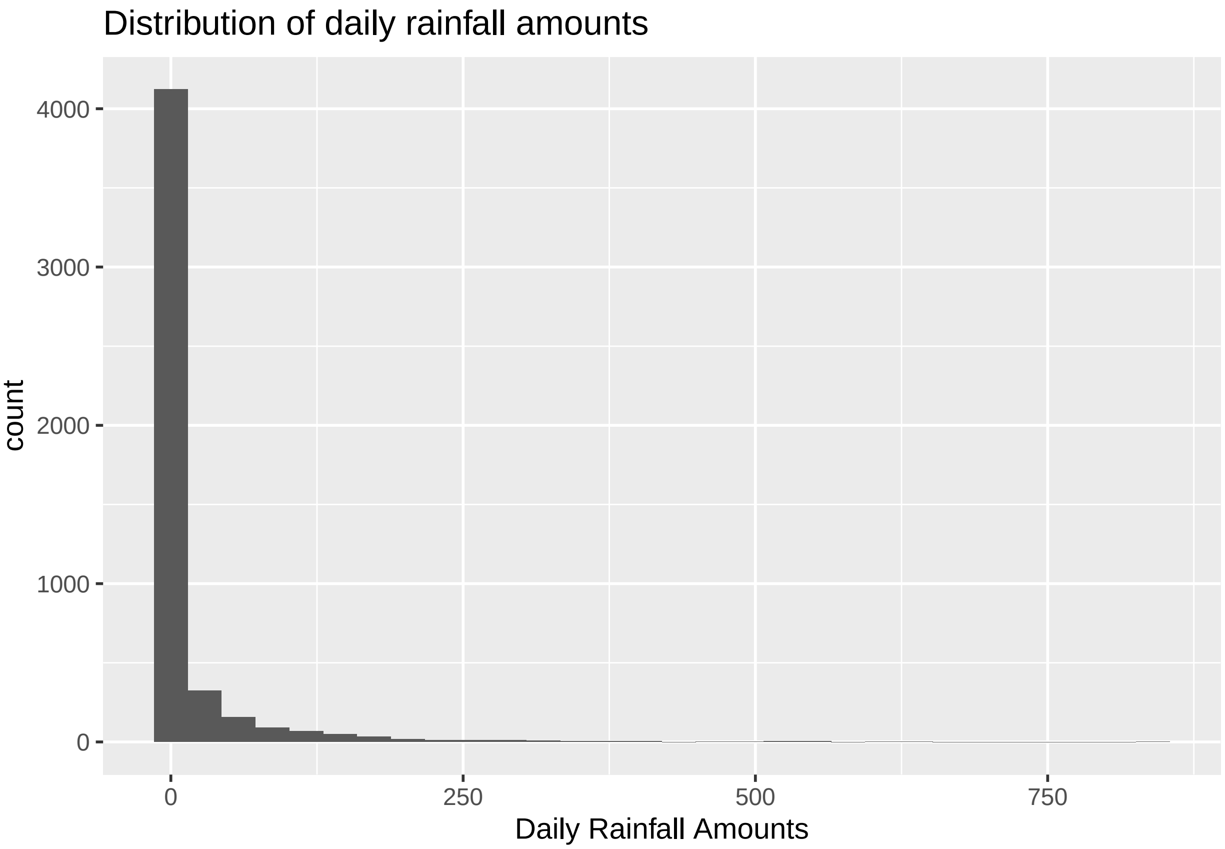
j. Give the command you would use to fix the data frame.

```
rain.df <- rain.df%>%
  filter(daily >= 0)
```

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(data = rain.df, aes(x = daily)) +
  geom_histogram() +
  labs(x = "Daily Rainfall Amounts",
       title = "Distribution of daily rainfall amounts")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This is more reasonable because the lowest amount of daily rainfall is reasonably set at 0(on days which it doesn't rain). The histogram also doesn't have any extreme outliers.

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)
```

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]
```

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]
```

3. (3 pts, equally weighted).

a.) What is the point of reproducible code?

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

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.