Data Wrangling in R

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.

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") #Convert .dat to dataframe
```

I used the `read.table` command to convert the dataset into a data frame.

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. The number is rows is indicated by the number of observations stated in the environment panel, and the number of columns is indicated by the number of variables.

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

I would use the `colnames` command to get the column names in rain.df

```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"
```

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

I would use the `rain.df[4,2]` command to get the value, which is 4

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

```
## [1] 4
```

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

I would use the `rain.df[2,]` command to get the value, the content is below:

```
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 above command renames the first three columns to “year”, “month”, and “day”, and renames the following columns sequentially from 0 to 23

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

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

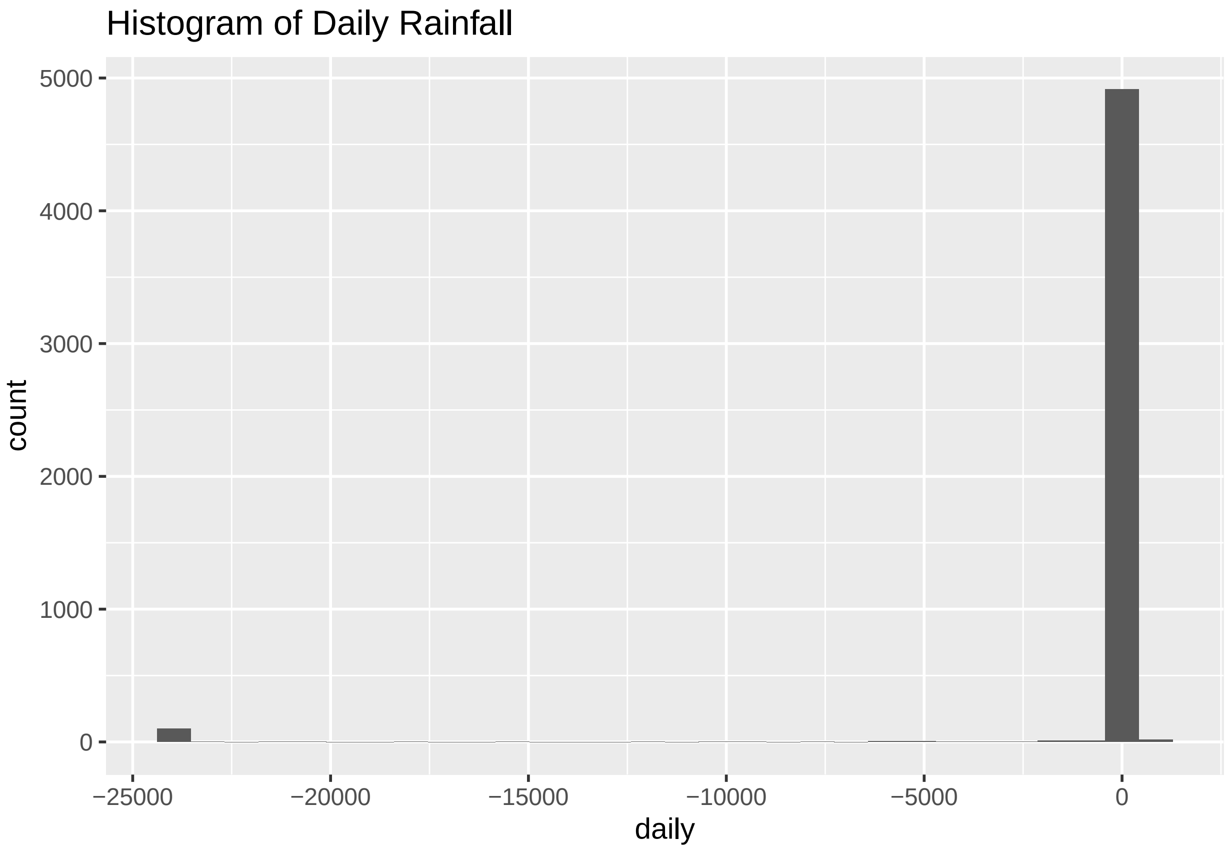
```
rain.df <- mutate(rain.df, daily = rowSums(rain.df[4:27]))
```

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.

I would use the \texttt{ggplot(rain.df, aes(x = daily)) + geom_histogram()} command to create a histogram of the daily rainfall amounts.

```
ggplot(rain.df, aes(x = daily)) +
  geom_histogram() +
  labs(title = "Histogram of Daily Rainfall")
```

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



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

The histogram cannot be correct because it is impossible for there to be a negative amount of rainfall in a day.

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

I would use the `rain.df[rain.df < 0] <- NA` command to fix the data frame, because the negative values likely reflect recording errors.

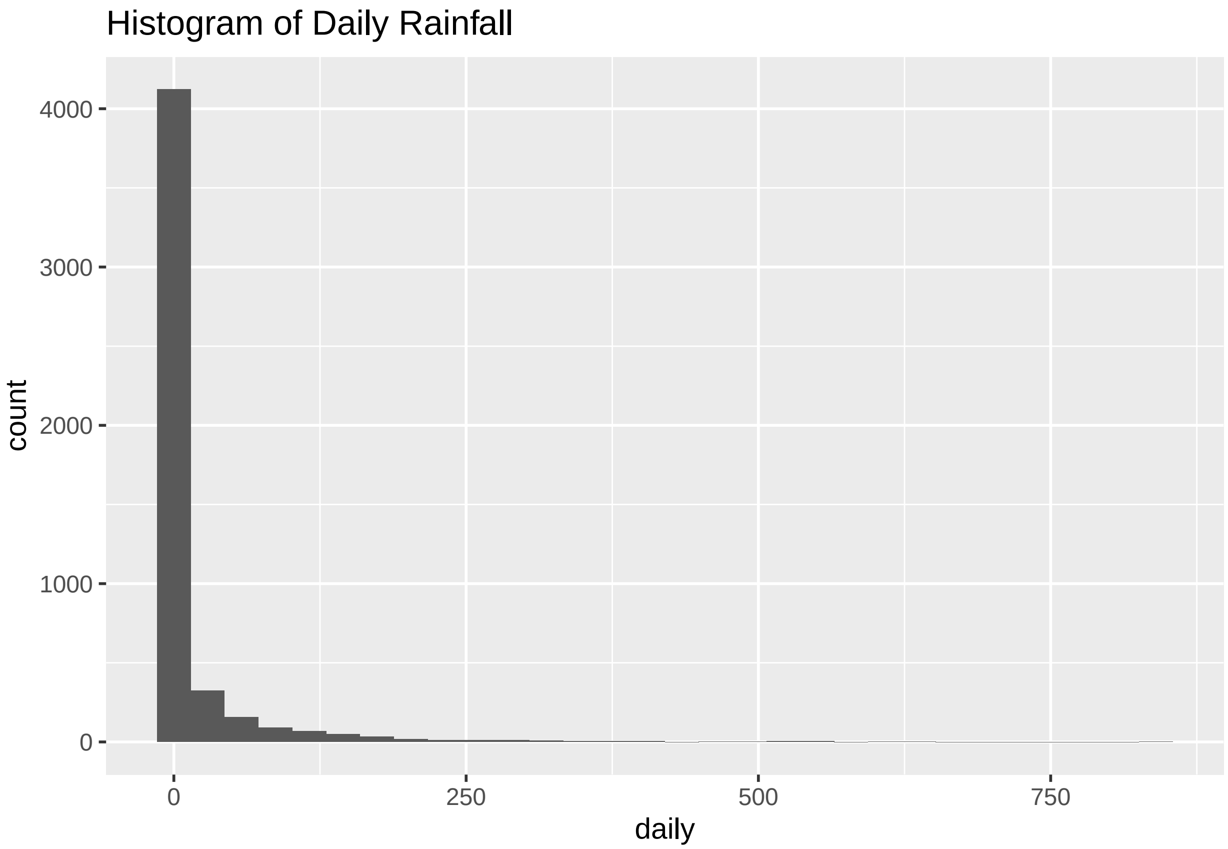
```
rain.df[rain.df < 0] <- NA
```

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.

```
ggplot(rain.df, aes(x = daily)) +
  geom_histogram() +
  labs(title = "Histogram of Daily Rainfall")
```

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

```
## Warning: Removed 139 rows containing non-finite values (stat_bin).
```



The above histogram is more reasonable than the previous one becuase it only contains positive values for rainfall. ***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")
```

Sets x to a character vector containing the strings “5”, “12”, and “7”

```
max(x)
```

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

Gets “7” as 7 has a higher ascii value than 5 or 1, the first characters of the other strings

```
sort(x)
```

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

Sorts the strings by their first characters, giving the order “12”, “5”, “7”.

```
sum(x)
```

Throws an error, because it is not possible to add non-numerical values.

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

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

Since one value is a character, all of the values are stored as strings in the vector for consistency. As such y is ("5", "7", "12")

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

Throws an error, because the values of y are strings, which cannot be added.

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)
```

The above command converts the information in the parenthesis to a dataframe, with the characters behind the '=' signs becoming column headers, and the information after becoming the data.

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

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

The above command sums the value at the first row and second column and the value at the first row and third column. This gives 19 (7+12)

3. (3 pts, equally weighted).

a.) What is the point of reproducible code?

The purpose of reproducible code is for others to be able to readily use it when collaborating.

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

If

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.

```
4
```