Data Wrangling in R

Fall 2022, MATH8050: Homework 1
Your Name, Section XXX

Due September 7, 12:00 PM

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.

Commenting code Code should be commented. See the tidyverse style guide for questions regarding commenting or how to write code https://style.tidyverse.org/index.html. 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.

R Working Environment

Please load all the packages used in the following R chunk before the function `sessionInfo()`

```
# load packages


sessionInfo()
```

Working with data

Total points on assignment: 10 (reproducibility) + 22 (Q1) + 9 (Q2) + 10 (Q3) + 30 (Q4) + 16 (Q5) + 3 (Q6) = 100 points

Reproducibility component: 10 points.

1. (22 pts 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?

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.)

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

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

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

f. What does the following command do?

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

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

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. Explain why that histogram above cannot possibly be right.

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

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.

Data types

2. (9 pts total, 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]
```

Linear algebra

3. (10 pts total, equally weighted) Consider the linear system $AX = b$, where A is an $n \times n$ positive definite matrix and b is a n-dimensional vector, the unique solution is $X = A^{-1}b$. Please answer the following questions:

a. Write an R function called `my_solver()` such that given inputs A and b, the function `my_solver()` returns the solution of the linear system, i.e., `X <- my_solver(A, b)`.

b. Run the following code to get A and b.

```
n = 100
set.seed(123)
A = rWishart(1, 150, diag(n))[ , ,1]
b = rnorm(n,1)
```

Then use your function `my_solver()` to produce the answer and verify your solution. (hint: AX should be equal to b)

Working with ggplot2

4. (30 pts total, equally weighted) EPA monitors Air Quality data across the entire U.S. The file "AQS-data.csv" contains daily PM 2.5 concentrations and other information. Please make the following questions using the `ggplot()` function for plotting. In addition make sure that all the x-axis and y-axis labels have 14 font size.

a. Read the data file "AQSdata.csv" into R.

b. Generate density plots of PM2.5 concentrations grouped by `County` in one single panel, where each density should have its own color. What do you find from the figure?

c. Plot histograms of PM2.5 concentrations across different counties with one panel for one histogram.

d. Generate boxplots of PM2.5 concentrations by County. What would you say about the distributions?

e. Reorder the boxplots above by the median value of PM2.5 concentrations.

f. Converting the Site ID to a factor and plot the historgram grouped by `Site ID`.

g. Generate the time series plot for the monitoring Site ID 450190048.

h. Plot time series of PM2.5 concentrations for all monitoring sites in one panel, where each site has its own color

i. Plot time series of PM2.5 concentrations across all monitoring sites in multiple panels, where one panel only has one site, and each row only has two panels.

j. In the time series plot, there seems to be not enough space to hold the x-axis labels. One way to avoid this is to rotate the axis labels. Please rotate all the time labels 45 degree.

Working with dplyr

5. (16 pts total, equally weighted) Continuing working with the above PM 2.5 data.

a. Filter all the observations in the county Greenville. How many observations are there?

b. Filter all the observations in Greenville in August 2021

c. Filter all the observations in Greenville in August 2021 and select the variables PM2.5 concentrations, Date, latitude and longitude of sites

d. Generate scatter plot of PM2.5 against latitude and longitude in two different panels

6. (3 pts total, 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.