BIG3003, Spring 2023 Homework 01

Due: PM 1:00 on April 12, 2023

Your responses contain textual explanation and the code you use to produce the results.

1 Rainfall data

The data set `rnf6080.dat` records hourly rainfall at a certain location in Canada, every day from 1960 to 1980. Load the data set into a data frame called rain.df using the following command:

```
rain.df <- read.table("rnf6080.dat", header=FALSE)
```

Use the `help` function to learn what arguments this function takes. Once you have the necessary input, load the data set into R and make it a data frame called `rain.df`. Here we use the option `header=FALSE` because there are no column names in the given file.

(a) How many rows are there in `rain.df`? How many columns? How do you know?

(b) What are the names of the columns of `rain.df`?

(c) What command would you use to get the value at row 1, column 3? What is the value?

(d) Display the first row of `rain.df` in its entirety.

(e) What does the following command?

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

(f) Create a new column called daily which is the sum of the 24 hourly columns. What command did you use?

2 Data types

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

```
x <- c("5", "15", "250", "2023")
max(x)
sort(x)
sum(x)
is.factor(x)
```

(b) For the following commands, either explain the results or why they produce errors.

```
y <- c("5",83,2023)
y[2] + y[3]
length(y)
dim(y)
is.factor(y)
```

3 Working with functions and operators

(a) The colon operator will create a sequence of integers in order. It is a special case of the function `seq()` which you saw earlier in this assignment. Using the help command `?seq` to learn about the function, design an expression that will give you the sequence of numbers from 1 to 10000 in increments of 372. Design another that will give you a sequence between 1 and 10000 that is exactly 50 numbers in length.

(b) The function `rep()` repeats a vector some number of times. Explain the difference between `rep(1:3, times = 3)` and `rep(1 : 3, each = 3)`.

4 Working with `airquality` dataset

Load the R dataset `airquality` using `data(airquality)`. Learn about the dataset using `help(airquality)`. If you do not have access to R, you may choose to work with other software.

(a) Using `summary` take a look at some summary statistics for the data frame. Note that there are some missing data and that all of the variables are numeric.

(b) Using `pairs`, construct of a scatterplot matrix including all of the variables in the dataset. These will all be joint (bivariate) distributions. Describe the relationships between each pair of variables. Are there nay associations among variables?

(c) Using `boxplot`, construct separate side-by-side boxplots for ozone concentrations conditioning on month and ozone concentrations conditioning on day. Does the ozone distribution vary by month of the year? In what way? What about by day?

(d) Construct a three-dimensional scatterplot with ozone concentrations as the response and temperature and wind speed as predictors. This will be a joint distribution. For plotting, you may use plotting methods are provided in the library `plot3D` and in library `scatterplot3d`. For really fancy plotting, have a look at the library `ggplot2`. What patterns can you make out?

(e) Construct an indicator variable for missing data for the variable `Ozone`. For example, use

```
is.na(airquality$Ozone)
```

Applying `table`, cross-tabulate the indicator against month. What do you learn about the pattern of missing data? How might your earlier analyses using the conditioning plot be affected? (If you want to percentage the table, `prop.table` is a good way.)