Introduction to R

Teaching Assistant: Parshan Pakiman Course: IDS 572

College of Business Administration ◇ University of Illinois at Chicago

If you find any issue in this document, please reach out to Parshan Pakiman at ppakim2@uic.edu.

1 Plan for the First Week

1.1 LinkedIn Learning

We use several videos on the LinkedIn Learning environment to learn and work with R in IDS 572. LinkedIn Learning is available to UIC students, and one can see the instructions available at [LINK] to setup up its account. Please make sure that your account works properly in the first week of classes.

1.2 R and RStudio

You need to install both R and RStudio, where the latter is an IDE for R. Feel free to use any online resources to learn how to install them on your machine. Below are some resources you might use:

- You can download R from [LINK].
- You can download RStudio from the products tab of the RStudio website [LINK].
- If you need more explanations on the installation of R and RStudio, you can use the LinkedIn Learning videos available at [LINK]. Specifically, see Section 2 of this video series titled "Getting Started."

1.3 Getting Started with RStudio

Once you installed R and RStudio, familiarize yourself with the RStudio environment. For example, try to navigate to the following parts of your RStudio: "console," "terminal," "files," "plots," "packages," "help," etc. Then, read Section 6 of the e-book titled "R for Data Science" available at [LINK]. It is useful to take a look at the exercises in Section 6.3 of this book. If needed, you can use the "Navigating the RStudio environment" video in Section 2 of the video series [LINK] to familiarize yourself with RStudio.

1.4 Managing Packages in R

Through the course, you might need to install R packages. If you need to learn how you can install new packages in R, you can see the "Packages for R" video in Section 2 of the video series [LINK]. Typically, we use the graphical environment of RStudio to see the list of installed R packages and install new ones. One can also use the following R script to view or install packages:

```
> install.packages("package_name")    # install the package named "package_name"
> remove.packages("package_name")    # uninstall the package named "package_name"
```

```
> installed.packages()    # list of all installed packages
```

1.5 Basics of R

Read Section 4 of R for Data Science. Ensure that you understand the coding basics (Section 4.1), naming of variables (Section 4.2), and calling of functions in R (Section 4.3). Also, read both Sections 2.2 and 3.1 of the book titled "R for Beginners" available at [LINK]. Below are some examples copied form these books:

```
> 1 / 200 * 30    # basic arithmetic
> (59 + 73 + 2) / 3    # use of parentheses in arithmetic
> sin(pi / 2)    # calling built-in function sin() and using constant pi
> # assigning value to object in R: object_name <- value
> # calling function in R: function_name(arg1 = val1, arg2 = val2, ...)
> seq(1, 10)    # calling function seq()
> x <- "hello world"    # defining an string object with the value of "hello world"
> x <- 1:30
> seq(length=9, from=1, to=5)
> c(1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5)
> rep(1, 30)
> gl(2, 6, label=c("Male", "Female"))
> data <- expand.grid(h=c(60,80), w=c(100, 300), sex=c("Male", "Female"))
> x = rnorm(10)
```

Solve Exercises 1 and 3 in Section 4.4 of R for Data Science. We need to also learn how to use R's help. Read section 2.3 of R for Beginners. The R's help is useful to search for built-in functions, their arguments, and their outputs. To get started with R's help, run the following codes:

```
> help.start()    # shows the graphical interface of R's help
> ?var    # pulls out the definition of function "var"
```

Next, we learn several data structures in R using the video series [LINK]. Please see following videos from this series:

1. Factor: watch the video titled "R data types: Factor".
2. Vectors: watch the videos titled "R data types: Vector", "Subsetting", and "Vector math". Then, I recommend reading Subsections 20.3.1, 20.3.2, 20.3.3, 20.4.5, 20.5.1, and 20.5.3 from R for Data Science.
3. Array: watch the video titled "R data types: Vector".
4. List: watch the videos titled "R data types: List" and "Brackets and double-brackets".
5. Array: watch the video titled "R data types: Array".

6. Sets: watch the videos titled "Sets: Union, intersect, and difference" and "Sets: Equal and in".

7. Matrix: watch the videos titled "R data types: Matrix" and "Matrix math review".

8. Dataframes: watch watch the videos titled "R data types: Data frame", "Data frames: Order and merge", "Data frames: rbind", and "Data frames: cbind".

1.6 Self-test Questions

In your RStudio, create the following 510×5 matrix:

```
> set.seed(1)
> data = matrix(rexp(50, rate=.1), ncol=5)
```

Now answer the following questions:

1. What is the minimum, mean, median, maximum, and the summation of each row of `data`? What are these values for each column of `data`?

2. What is the name of the columns of matrix `data`? Update the name of its columns to the following list `c('A','B','C','D','E')`. What happens if you run code `data$C`? Run code `data$C` where `df_data<-data.frame(data)`. Explain what is the type of `data` and `df_data`. Are they have the same type? Note: use code `class()` to see type of an object.

3. What does code `data[1:10,3]` do? Compute inner product of the first and the third columns of this matrix. Note: the inner product between two vectors, say u and v, can be computed in R using code $u\%*\%v$.

4. Transpose `data` using code `dataT = t(data)`. Change the column name of $dataT$ and remove its current row names.

5. Replace the 5th row and the third column of `data` to 3.75. Replace the 5th row and the third column of `df_data` to 3.75.

6. Add the column `F` to `data` using code `df_data['F'] <- 0`. Add a new column to `data` (name it `RATIO`) which gives the ratio of two existing columns. What happens if you divide column `df_data['A']` by `df_data['F']`?

7. Define column `df_data['G'] <- NA`. Use function `summary()` to compute minimum, mean, median, maximum, and other statistic of every column of `df_data`. What do you see in the output of `summary()` for column `df_data['G']`.