DSI Data Science: Introduction to R
Assessment 1

Anjali Silva, PhD

Fall, 2022

Due: Email TA before 9.00 pm EST on 20 November 2022.

Introduction and Goals

Read all instructions carefully. There are 6 pages in this document.

During the course we had a chance to navigate the RStudio environment, learn its anatomy, use R syntax, identify how to get help, and apply built-in functions. Further, we learned how to install R packages (CRAN, Bioconductor, GitHub). We explored R data types and structures. Then we downloaded tidyverse package (Wickham, H. et al, 2019) and explored some of its functions. The aim of this assessment is to ensure that learners are able to:

- Identify RStudio anatomy and be able to navigate the RStudio environment.
- Understand differences in R data types and structures.
- Apply R coercion rules.
- Detect and work with missing values.
- Work with built-in functions and functions from downloaded R packages.

Tasks

Answer all the questions, in order. You may use this document with a PDF editor, the R markdown template provided or any other platform/software of your choice to generate the PDF document containing the questions and answers. Alternatively, you may open your R script using a text editor and export it as a PDF document. **In your submission to TA, you must provide both the question number AND the question, in addition to your answers.** This is done to ensure that you do not skip any questions and to ensure all sub-questions within each question are answered. You may ignore the formatting (e.g., *italicizing*) in questions when copying and pasting questions.

1. [1 mark] In RStudio, the _____ tab lists the objects/variables and functions present in the current R session.

2. [1 mark] In RStudio, the _____ is where a user can give R commands and where R will show the results of a command.

3. [4 marks] In the below vector called 'testVector', list the type (character, logical, integer, double, raw, complex) for each element.

```
# a vector containing different element types
testVector <- c("a", 1L, 1.5, TRUE)
```

"a": _____
1L: _____
1.5: _____
TRUE: _____

4. [1 mark] What will be the resulting type of vector for 'testVector' from above question?

_____

5. [1 mark] What type of coercion occurs when creating the vector'testVector'; implicit or explicit?

_____

6. [2 marks] Explain why this type of vector results for 'testVector'? Be sure to explain by outlining R coercion rules. (Hint: see 'Details' section of help documentation for `c()` function, which is for concatenation. To access help documentation, do `?c` on console.)

_____
_____
_____
_____

7. [1 mark] Look up help for R base function `mean()` using `?mean` on console. Create an integer vector called 'integerVector'. Demonstrate an application of `mean` function using 'integerVector', i.e., provide an example of using `mean` function with 'integerVector'. Provide all your code.

8. [1 mark] Create the following vector called 'missingValueVector'. This vector will be built using reserved term `letters`, which contain letters of the alphabet in R. Using the `c()` function, which is for concatenation, the `letters` have been combined with 3 missing values indicated by `NA`. The resulting vector is called 'missingValueVector' as shown below. How would you detect missing values in this vector called 'missingValueVector'? Provide all your code.

```
# vector containing missing values
missingValueVector <- c(letters, NA, NA, NA)

# autoprint the object
missingValueVector
```

9. [2 marks] How would you identify which positions contain missing values in 'missingValueVector'? Provide all your code. (Hint: There could be multiple ways to perform this. One method is to use `which()` and `is.NA()` functions and make use of logical values).

10. [2 marks] Can you mix positive and negative integers simultaneously when subsetting a vector? Yes or No? _____ Justify your answer by showing one example. You may show this using'missingValueVector'. Provide your code for the example.

11. [2 marks] R comes with its own datasets, which you can access using command `data()` on the console. This will open up a new tab showing all the datasets. Pick a dataset of interest to you. Take a look at the dataset, its help documentation, and ensure it has tabular structure. How would you convert this dataset to a tibble? Provide all your code.

12. [2 marks] A package is a way to organize code for the purpose of shareability. A package may include code, functions, documentation, tests to check package/functions and, sometimes, datasets. A repository is a place where packages are located, so users can locate, download and use them. Three of the most popular repositories for R packages are The Comprehensive R Archive Network (CRAN), Bioconductor and GitHub.

    Package *covid19.analytics* is a recently developed package from Toronto, that is publicly available via both CRAN and GitHub. Note, the package may have been updated since the development of this document (finalized October, 2022). Packages are removed from CRAN if the packages violate *CRAN policies*.

    For CRAN availability of this package, see: `https://cran.r-project.org/web/packages/covid19.analytics/index.html`. Visit and read this page. This page will be referred to as the CRAN page for *covid19.analytics* package.

    Give the commands for downloading this package from CRAN to your computer via R. Also include command to attaching the package to your current R session.

13. [2 marks] Let's work with some string manipulations. The *tidyverse* package contain *stringr* package that contains many functions for string manipulation. Use the following code to download R package *protr* and create the object `P00750` as shown. Which function from *stringr* R package can you use to count the number of 'G' letter occurings in `P00750`? How many 'G' letters are present? Provide all your code.

```
# download and attach protr R package
install.packages("protr")
library("protr")
# attach tidyverse R package
library("tidyverse")
# attach stringr R package
library("stringr")

# to learn more about this function
?readFASTA

# copy the example
P00750 <- readFASTA(system.file("protseq/P00750.fasta", package = "protr"))

# view object P00750
P00750

# Which function from stringr package to find all occurrences of G in P00750?
# How many 'G' letters are present?
```

Grading Scheme

The mark assigned for each question is indicated with the question. Use that to guide the answers. There will be marks assigned for submitting the Assessment in correct format.

Submission Instructions

[1 mark] Remember: Submissions should only be in PDF format. When emailing TA, name PDF using format: LASTNAME_FirstInitial_A1.PDF. E.g., SILVA_A_A1.PDF.

[2 marks] **In your submission to TA, you must provide both the question number AND the question, in addition to your answers.**

Answer all the questions, in order.

1 To Use R Markdown To Create A PDF With Answers

R Markdown is a format for creating reproducible, dynamic reports with R. R Markdown reports permit to embed R code and results into slideshows, pdfs, html files, Word documents, etc. If you prefer to use R markdown to generate your PDF files with solutions, you will have to spent sometime to learn R markdown first. To make this easier, a template is provided for you and you may use this. To locate the R Markdown template provided to you, called '02_Assessment1_template.Rmd', visit Assessments folder

within GitHub:
`https://github.com/anjalisilva/IntroductionToR/tree/main/Assessments` and download the file. Open this file using RStudio.

To create PDF documents from R Markdown, you will need *rmarkdown* R package and a LaTeX distribution installed. There are several options for LaTeX including MiKTeX, MacTeX, TeX Live, and TinyTeX. You are welcome to download any option of LaTeX that may work. Here, I will show the use of TinyTeX. You can install TinyTeX with the *tinytex* R package.

```
# install the rmarkdown package
install.packages("rmarkdown")
library("rmarkdown")

# install the tinytex package
install.packages("tinytex")
library("tinytex")
# to install TinyTeX
tinytex::install_tinytex()
```

The template has all question numbers and questions. You will only need to insert your answers. Before starting to enter answers, do the following. First, with '02_Assessment1_template.Rmd' opened in RStudio, press 'Knit' icon and then select 'Knit to PDF'. See Figures 1 and 2 to locate the Knit icon on RStudio. Alternatively, from the above menu of RStudio, you can select 'File', then 'Knit Document' to convert '02_Assessment1_template.Rmd' to a PDF called '02_Assessment1_template.PDF'. Another option is to use the console. On console, type the following:

```
library(rmarkdown)
rmarkdown::render("02_Assessment1_template.Rmd", output_format = "pdf_document")
```

You are recommended to add small chunks of code at a time and 'Knit' the document. For more information including basics, see `https://rmarkdown.rstudio.com/lesson-1.html` or seek help for TA early.

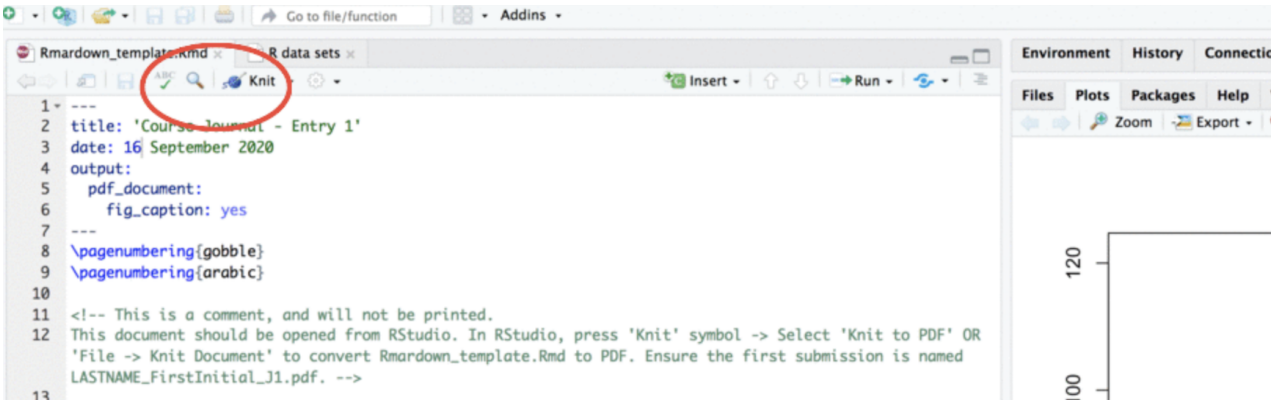

Figure 1: Locating 'Knit' icon within RStudio. See red circle.

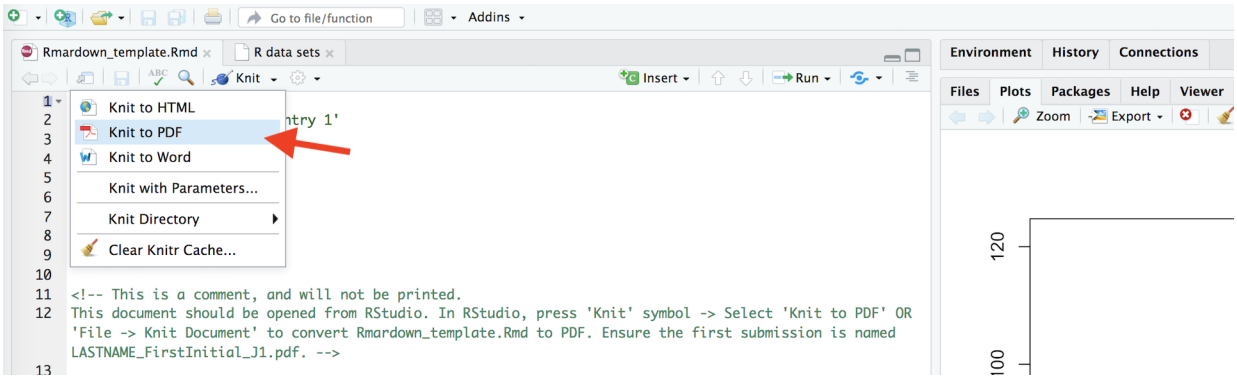

Figure 2: Use the down arrowhead immediately to the right of 'Knit'. Select 'Knit to PDF'. See red arrow.

Extra Readings

- Ihaka,R. and Gentleman,R. (1996) "R: a language for data analysis and graphics". *Journal of Computational and Graphical Statistics*, 5, 299–314. `https://www.stat.auckland.ac.nz/~ihaka/downloads/R-paper.pdf`.
- Chambers, J. M. (2009) "The R Journal: Facets of R". *The R Journal*, 1, 5-8. `https://journal.r-project.org/dev/articles/RJ-2009-008`.
- (2017) "The worlds most valuable resource is no longer oil but data". *The Economist*. `https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data`.
- Silge, J., Nash, J. C., and Graves, S. (2018) "Navigating the R Package Universe". *The R Journal*, 10. URL: `https://journal.r-project.org/archive/2018/RJ-2018-058/index.html`.
- Wickham, H. et al. (2019). "Welcome to the tidyverse". *Journal of Open Source Software*, 4(43). URL: `https://joss.theoj.org/papers/10.21105/joss.01686`.

References

- Wickham, H. et al. (2019). "Welcome to the tidyverse". *Journal of Open Source Software*, 4(43). URL: `https://joss.theoj.org/papers/10.21105/joss.01686`.
- Ponce, M. et al. (2021). "covid19.analytics: An R Package to Obtain, Analyze and Visualize Data from the Coronavirus Disease Pandemic". *Journal of Open Source Software*, 6(59). URL: `https://doi.org/10.21105/joss.02995`.
- Allaire, J. et al. (2022). rmarkdown: Dynamic Documents for R. R package version 2.16. URL: `https://rmarkdown.rstudio.com`.
- R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`.
- RStudio Team (2022). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL: `http://www.rstudio.com/`.
- Wickham, H. (2022). "stringr: Simple, Consistent Wrappers for Common String Operations". R package version 1.4.1, URL: `https://CRAN.R-project.org/package=stringr`.
- Xie, Y. (2022). tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents. R package version 0.41.