DSI Data Science: Introduction to R
Assessment 2

Anjali Silva, PhD

Fall, 2022

Due: Email TA before 9.00 pm EST on 29 November 2022.

Introduction and Goals

Read all instructions carefully. There are 5 pages in this document.

During the course we had a chance to navigate RStudio environment, download the tidyverse package (Wickham, H. et al, 2019) and explore functions for data manipulation, wrangling and visualization. Further we learned about writing custom functions, explored functional programming tools and methods for simulating data. Further, we touched on concepts of reproducibility, bias, diversity, inclusion, ethical considerations, equity concepts, data security and best coding practices. The aim of this assessment is to ensure that learners are able to:

- Manipulate tabular data with *dplyr*: A Grammar of Data Manipulation.
- Use of manipulation and wrangling techniques for reshaping data.
- Generate intuitive visualizations appropriate for data type with *ggplot2*: A Grammar of Graphics.
- Design custom functions that can take data as input, perform simple analyses, and generate output.
- Recognize equity, diversity & inclusion (EDI) practices and their importance in data sciences.

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. [15 marks] For this set of problems, we will use the built-in dataset `iris` from R *base* package *datasets*. To access this dataset, you can simply type `iris` on your script or console.

   (a) [7 marks] Using *tidyverse* functions and other base R functions as necessary, write a custom function to perform the following. The custom function you are to write should be called *speciesFilter*. The *speciesFilter* function needs to have two arguments called *dataset* and *irisSpecies*. Within the function body, use `dplyr::filter` to filter the user provided *dataset* for the species that the user would specify in argument *irisSpecies*. Return the resulting output to user within the function. Provide all your code for full marks.

   (b) [3 mark] Run the *speciesFilter* function you wrote using the built-in dataset `iris`. Set the *dataset* argument to `iris` and *irisSpecies* argument to "setosa". Save the resulting output from function *speciesFilter* into an object called *outputSpeciesFilter*. 'What is the size of the resulting data frame, i.e., the dimensions of *outputSpeciesFilter*? Show all your code for full marks.

   (c) [5 marks] Using *outputSpeciesFilter* object from above, modify it to contain a new column using a *dplyr* function, that shows the ratio of Sepal.Length to Petal.Length from `iris` dataset. This can be achieved by dividing Sepal.Length from Petal.Length as shown below. It would be up to you as to what you would name the new column. Below, results are saved as *RatioSLPL*. Show all your code for full marks.

```
# RatioSLPL shows the ratio of Sepal.Length to Petal.Length
RatioSLPL = Sepal.Length / Petal.Length
```

2. [15 marks] Select a public dataset of your choice and save a copy into your data subdirectory of Rproject. Read the dataset into RStudio using the appropriate function. Preprocess the dataset if needed (i.e., remove missing values, check and fix column labels, remove special characters, subset the dataset, etc.). Initialize and plot the dataset using *ggplot2::ggplot* function. Customize plots with at least **4 layers** from *ggplot2* package to ensure that the plot contains an informative title, clear labels, axes, color, size, and overall look. Show all your code (including code for reading of dataset into RStudio). Provide the plot created along with a detailed description of what the plot shows (3-4 sentences) for full marks. Be sure to provide full citation for dataset used, including the URL. You may attach the plot separately.

   Potential data sources:

   - Open Data | Open Government, Government of Canada: `https://open.canada.ca/en/open-data`
   - Open Data - City of Toronto: `https://www.toronto.ca/city-government/data-research-maps/open-data/`
   - University of Toronto Dataverse: `https://borealisdata.ca/dataverse/toronto`
   - Find Open Datasets - Kaggle: `https://www.kaggle.com/datasets?`

   Other helpful websites:

   - University of Toronto Data Visualization Guide:`https://mdl.library.utoronto.ca/dataviz/getting-started`
   - Colorbrewer2: `https://colorbrewer2.org/`

3. [10 marks] How can you ensure equity, diversity inclusion (EDI) practices are respected in your work related to data sciences? Give examples. What are some challenges or what maybe beyond your control? Explain. Use half to 3/4ths of this page.

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_A2.PDF. E.g., SILVA_A_A2.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.

Extra Readings

- Zook, M., et al. (2017) "Ten simple rules for responsible big data research". *PLoS Computational Biology*. 13. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5373508/`.

References

- Wickham, H. (2016). "ggplot2: Elegant Graphics for Data Analysis". Springer-Verlag New York. URL: `https://ggplot2.tidyverse.org`.
- 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. et al. (2019). "Welcome to the tidyverse". *Journal of Open Source Software*, 4(43). URL: `https://joss.theoj.org/papers/10.21105/joss.01686`.
- Xie, Y. (2022). tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents. R package version 0.41.