For a PDF version of this document, please click here (https://mfiorina.github.io/sais_r_course/spring_2023/syllabus/r_course_syllabus.pdf).

Professional Skills Course

Syllabus – Spring 2023

Programming for Professional Research Using R

Marc-Andrea Fiorina

Contact: mfiorina@worldbank.org (mailto:mfiorina@worldbank.org)

Schedule

Weekday	Date	Room	Start Time	End Time
Thursday	March 30, 2023	BOB 616	6:00pm	8:00pm
Thursday	April 6, 2023	BOB 616	6:00pm	8:00pm
Thursday	April 13, 2023	BOB 616	6:00pm	8:00pm

Course Description

Entry-level research positions in universities, government offices, think tanks, and multilateral institutions are increasingly expected to perform basic quantitative tasks using statistical software such as Stata, R, or Python. As data work has become near-ubiquitous in the policy world, so have basic tasks like aggregating, analyzing, summarizing, and visualizing data.

This course approaches the use of a statistical software – R – from the perspective of a research assistant. It will introduce you to the methods needed for quantitative work in such a position. This course will also aim to provide you with the resources to continue developing your knowledge and experience of R beyond its duration.

Course Objectives

By the end of this course, students will be familiar with the instruments they are expected to use in entry-level professional research positions. They will also have resources to further their education in R past the course.

In more detail, students will be introduced to the use of:

- `dplyr`'s `select()` and `mutate()` functions to explore and modify datasets;
- `dplyr`'s `group_by()` and `summarize()` functions to generate summary statistics;
- `tidyr`'s `pivot_longer()` and `pivot_wider()` functions to reshape data for easier analysis;
- The `gt` and `stargazer` packages to generate HTML, PNG, or LateX summary tables;
- The `ggplot2` package to generate descriptive scatter plots of data

Students will also be taught how to clean raw data by:

- Identifying duplicates
- Finding outliers
- Looking for data inconsistencies
- Codifying missing values
- Encoding categorical character variables

Course Outline

Session	Description
I – Learning R through the Tidyverse	- Basic R Use – Introduction to R and RStudio, How to import/export data - Using Basic Tidyverse Functions – how to subset, mutate, summarize, and reshape data - Using R in a Collaborative Setting – Introduction to downloading and publishing data using GitHub
II – Cleaning and Constructing Data Using R	- Raw Data Cleaning – How to identify and fix data errors prior to its use - Data Construction – How to build and reshape datasets for analysis
III – Producing Research Output Using R	- Exporting Summary/Regression Tables – How to produce (i) regression tables and (ii) descriptive summary tables for academic or policy audiences - Introduction to R Data Visualization – How to produce beautiful and informative visualizations: scatter plots, density plots, and more

Note – The course slides, syllabus, and other resources will be shared at the following link: https://mfiorina.github.io/sais_r_course/ (https://mfiorina.github.io/sais_r_course/).

Course Structure

Each two-hour session will be split into two halves. The first half (approx. one hour) will consist of an interactive lecture using slides and live coding. The second half (approx. one hour) will consist of practical exercises that the students will accomplish with my support.

The last two sessions will begin with multiple-choice questionnaires on the topic of the previous week's content. At the end of the course, there will be an open-ended assignment in which the students will have the option to create a script, which I will then review and provide feedback.

Course Readings/Resources

Students should download data and code templates from this linked Dropbox folder (https://www.dropbox.com/sh/hx6r5hcqvndorgu/AAA2LbKrtMXmJSaPUsMpN2_2a?dl=0).

As there will be no time for this in class, **YOU NEED TO DO THE FOLLOWING BEFORE THE FIRST SESSION:**

- Download R at this link: https://cran.r-project.org/ (https://cran.r-project.org/)
    - Windows users, click on 'Download R for Windows' and then download the 'base' version. The file should be called `R-4.2.3-win.exe`.
    - Mac users, click on 'Download R for MacOS' and then download the file called:
        - `R-4.2.3-arm64.pkg` if you have a more recent MacBook with an M1 chip.
        - `R-4.2.3.pkg` if you have an older MacBook with an Intel chip or run an older iOs.
- Download RStudio at this link: https://www.rstudio.com/products/rstudio/ (https://www.rstudio.com/products/rstudio/). Download the 'RStudio Desktop: Open Source Edition' version.

Note – Readings and resources below are optional and are provided for context and use after the course is finished. The session slides will cover everything needed for the course.

Overall Readings and Resources

- Hadley Wickham & Garrett Grolemund, *R for Data Science* (https://r4ds.had.co.nz/). This is the foundational textbook for use of the "Tidyverse" package suite in R.
- RStudio, RStudio Cheatsheets (https://www.rstudio.com/resources/cheatsheets/). Cheatsheets to help perform basic data tasks in R.
- Thomas Mock's The Mockup Blog (https://themockup.blog/) has a great array of tutorials for all levels. You'll see posts from there below.
- The World Bank DIME Wiki (https://dimewiki.worldbank.org/Main_Page). A wiki with open-source articles on how to be a research assistant with the World Bank. Great insights into collaborative data work, reproducibility, and the responsibilities of an entry-level data researcher.

Session 1 – Learning R through the Tidyverse

- For those who are fully new to R, I strongly recommend: Thomas Mock, "A Gentle Introduction to Tidy Statistics in R" – blog post (https://themockup.blog/posts/2018-12-10-a-gentle-guide-to-tidy-statistics-in-r/) and video (https://www.rstudio.com/resources/webinars/a-gentle-introduction-to-tidy-statistics-in-r/). We won't have much time to review the basics of R programming during the first session.
- Dominic Royé, "A very short introduction to Tidyverse" (https://dominicroye.github.io/en/2020/a-very-short-introduction-to-tidyverse/). Blog post covering the basics of Tidyverse use in R.
- tidyr, "Pivoting" (https://tidyr.tidyverse.org/articles/pivot.html). Vignette explaining how to reshape datasets using `pivot_longer` and `pivot_wider`.
- Hadley Wickham, "dplyr 1.0.0: working across columns" (https://www.tidyverse.org/blog/2020/04/dplyr-1-0-0-colwise/). Explains the basics for flexible column-wise operations using `across` in R.

Session 2 – Cleaning and Constructing Data Using R

- DIME, "Data Cleaning" (https://dimewiki.worldbank.org/Data_Cleaning). Instructions on how to clean raw household data for use in a development setting.
- Wickham and Grolemund, *R for Data Science* Chapter 12 – Tidy Data (https://r4ds.had.co.nz/tidy-data.html). How to structure ("tidy") your dataset for flexible use in data analysis.

Session 3 – Producing Research Output Using R

Tables

- Marek Hlavac, "stargazer: beautiful LATEX, HTML and ASCII tables from R statistical output" (https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf). Vignette for the `stargazer` package, main tool to export regression tables to LateX
- Thomas Mock, "gt - a (G)rammar of (T)ables" (https://themockup.blog/posts/2020-05-16-gt-a-grammer-of-tables/). Introduction to the `gt` package, a more flexible instrument to export tables in PNG, PDF, or HTML formats.

Plots

- Alicia Horsch, “A quick introduction to ggplot2” (https://towardsdatascience.com/a-quick-introduction-to-ggplot2-d406f83bb9c9). Introduction to the `ggplot2` package, the main instrument for plot creation in R.

Further Resources

- We won’t have time to explore more advanced data manipulation techniques. Here are some resources for those who are interested in learning on their own:
    - Iterative coding (using loops for repetitive code) – `purrr`’s `map` function is your friend. I recommend Thomas Mock, “Functional programming in R with Purrr” (https://themockup.blog/posts/2018-12-11-functional-progamming-in-r-with-purrr/) to get you started.
    - User-made functions in R – At some point, it will become time-effective to create your own functions to apply to your work. Hadley Wickham, *Advanced R* Chapter 6 – Functions (https://adv-r.hadley.nz/functions.html).

- For those interested in learning more about publishing your work using RMarkdown:
    - RStudio, “Introduction to RMarkdown” (https://rmarkdown.rstudio.com/lesson-1.html). Summarizes the uses and utility of the RMarkdown framework.
    - Yihui Xie, “xaringan Presentations” – book chapter (https://bookdown.org/yihui/rmarkdown/xaringan.html) and presentation (https://slides.yihui.org/xaringan/). Introduction to `xaringan`, a package that allows you to create slide decks using R. Also explore the `xaringanExtra` package (https://pkg.garrickadenbuie.com/xaringanExtra/#/).
    - With RMarkdown, create books using `bookdown` (https://bookdown.org/) or a blog using `blogdown` (https://bookdown.org/yihui/blogdown/)

- For those interested in data visualization:
    - The R community organizes “Tidy Tuesday” (https://www.tidytuesday.com/). This is a weekly challenge where users are provided a dataset and participants then swap graphs and scripts used to create their visualizations.
    - David Robinson’s Tidy Tuesday live screencasts (https://www.youtube.com/user/safe4democracy) on YouTube. The

perfect resource to follow along and try to replicate a professional coder's scripts.
    - Yan Holtz and Conor Healy, "From Data to Viz" (https://www.data-to-viz.com/). An amazing repository of methods to create different data visualizations using R.

- For those interested in geospatial work and visualizations using maps:
    - Robin Lovelace, Jakub Nowosad, and Jannes Muenchow, *Geocomputation with R* (https://geocompr.robinlovelace.net/index.html). A great introduction to manipulating geospatial data (shapefiles and rasters) in R.
    - Edzer Pebesma, "Simple Features for R" (https://r-spatial.github.io/sf/articles/sf1.html). An introduction to the `sf` package, commonly used for geospatial work in R.
    - Edzer Pebesma, "Plotting Simple Features" (https://r-spatial.github.io/sf/articles/sf5.html). How to use `sf` and `ggplot2` to visualize data using maps.

- For those interested in conducting data work in the development world: Kristoffer Bjarkefur, Luiza Cardoso de Andrade, Benjamin Daniels, and Maria Ruth Jones, *Development Research in Practice – The DIME Analytics Data Handbook* (https://worldbank.github.io/dime-data-handbook/). A comprehensive account of tools and instruments to conduct quantitative development research.
- For those looking for more hands-on, real-world data work: Ben Baldwin, "A beginner's guide to nflfastR" (https://www.nflfastr.com/articles/beginners_guide.html). How to download and explore NFL play-by-play data. This is how I learnt how to use R. Further tutorials using this data can be found at the "Open Source Football" blog (https://www.opensourcefootball.com/).

Instructor Biography

My name is Marc-Andrea Fiorina (MA – SAIS Bologna 2018, DC 2019), and I am a research analyst with the Development Impact Evaluation unit (DIME) at the World Bank. Prior to studying at SAIS, I received my BA (Hons) in Philosophy, Politics, and Economics from the University of Oxford.

I started at DIME as an intern in the Spring 2019 Semester and worked as a research assistant there from October 2019 to August 2021. I now work on a project that analyzes the gendered impacts of providing cash-for-work programs to women in low- and middle-income countries.

As a research assistant, I learnt how to work with data in a collaborative space and how to improve my coding language learning through continuous use and good practices. I hope to share those practices and resources with you through this course.