Introduction to Data Science with RStudio

2022 Young Research Fellow Program

Schedule: 09:00–12:00, April 9, 2022.

Instructor: Nith Kosal. You can reach out me at `nithkosal@futureforum.asia`.

Course Objective: The core content of the seminar focuses on data acquisition and wrangling, exploratory data analysis, data visualization, statistics descriptive and inferential sample econometrics, and effective communication of results. A heavy emphasis is placed on a consistent syntax (with tools from the tidyverse), reproducibility (with R Markdown), and version control and collaboration (with Git and GitHub). In addition, out-of-class learning is supplemented with interactive tutorials. The goal of the course is to bring you from zero to being able to work in a team on a fully reproducible data science project analysing a dataset of their choice and answering questions they care about.

Requirement Before Seminar:

- Downloaded and Installed R and RStudio into your laptop.
- Understood the Integrated development environment (IDE) of the RStudio.
- Understood how to create a working directory and used *help* function.
- You should able to use a sample calculator through arithmetic operators: addition, subtraction, division, exponentiation, and modulo.
- Installed the important packages such as *tidyverse, tidyr, dplyr, ggplot2, ggraph, tidyquant, shiny, caret, kernlab, plotly, xml, readxl, readr, tidyxl, foreign, data.table, knitr,* and so on to your laptop.
- You should able to create and calculate vectors, matrices, factors, data frames, and lists.
- Understood how to reading data from files: *.csv, .xls, .xlsx, .dta, .sav, .mat, .wf1, xml, .html,* and *.txt*.
- Understood uncertain programming with data like logical instructions, loops, repetitions and functions, as well as, I have excepted you are probably can write your own function.
- You should able to use R packages functions for data wrangling and transforming.
- Understood the concept how to visualize data into *bar chart, pie chart, tree map, histogram, kernel density plot, scatterplot, line plot, box plots, violin plots, ridgeline plots, strip plots, beeswarm plots, cleveland dot charts, dot density maps, time series plot, area charts, correlation plots, survival plots, mosaic plots, biplots, diagrams, heatmaps, radar charts, waterfall charts,* and *word clouds*. It is seen many things to know, even with limited time—I would not recommend you look at it at all. it is good if you can well-know how to use five or seven types of data visualization.
- You should understand the basic statistical functions, statistics, and basic econometrics.

Seminar Material: As we have limited time for the seminar, I will not give a presentation on a theory or concept of how to use the R program, statistics, or econometric. To be clear, this is a RStuido practice class with the exercise in the real world datasets from household surveys and R Database. Everyone require to read and take the time to listen in advance to the materials provided prior to the seminar. The following textbooks are helpful for background reading:

- **R Programming Fundamentals**, Susan Holmes. A recorded training videos from the Stanford Center for Professional Development. `shorturl.at/eoyU3`
- **An Introduction to R**, W. N. Venables, D. M. Smith and the R Core Team, 2022. Complete introduction to base R. `https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf`
- **R for Data Science**, Garrett Grolemund and Hadley Wickham, 2016. Introduction to data analysis using R, focused on the *tidyverse* packages. `http://r4ds.had.co.nz/`
- **Advanced R**, Hadley Wickha, 2019. In-depth discussion of programming in R. Read later, if you want to become a good R programmer. `https://adv-r.hadley.nz`
- **Data Visualization: A Practical Introduction**, Kieran Healy, 2018. Textbook on data visualization, using *ggplot2*. `https://socviz.co`
- ***ggplot2*: Elegant Graphics for Data Analysis**, Hadley Wickham, 2015. In depth discussion of R-package for data vizualization. `https://ggplot2-book.org`
- **An Economist's Guide to Visualizing Data**, Jonathan A. Schwabish, 2014. Guidelines for good visualizations (not R-specific). `https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.28.1.209`
- **A Layered Grammar of Graphics**, Hadley Wickham, 2010. The theory behind *ggplot2*. `https://byrneslab.net/classes/biol607/readings/wickham_layered-grammar.pdf`

Starting today, you will have a week for self-learning and exploration with the R program and statistics or econometrics. If you have any questions, please feel free to ask me.

Exercise Before Seminar: In order to have a comprehensive knowledge prior to a practical session, you should complete the following exercises and submit them to me for review on April 8, 2022. Your exercise should be submitted in the R or Markdown file. Before the seminar starts on Saturday, I will provide the answer and comments in general and then we can start the practice. All exercises are connected. To better understand the basic program language in R, you should do it. You can do it in groups, and each group should have 3 or 4 members. At the time, you should use a GitHub for sharing this exercise project with your teammate. By doing so, you should create a GitHub account and a sharing repository on the exercise project with your teammate. All exercises are interconnected. You can do it in groups, and each group should have three or four members. At this point, you should use a GitHub to share this exercise project with your teammate. In doing so, you should create a GitHub account and a sharing repository on the exercise project with your teammate.