Syllabus : Programming for Data Analysis

Summer 2020

Instructor: Fred LaPolla, MLS

Email: fred.lapolla@nyulangone.org
401-480-4917

Duration: Mondays and Thursdays July 6 - August 13

10:00am-11:30am

Note: Dates and topics subject to change.

Overview:

An introductory class for beginners in R programming for data analysis. This Zoom-based class will address getting up and running with R and R Studio, data types, data structures, the use of functions and their arguments, data cleaning/transformations, and data visualization, hypothesis testing in R and basic regression.

Objectives:

Students will be able to read and program in R syntax including:

- Loading data into R
- Creating and naming variables
- Cleaning data
- Conducting hypothesis testing data analysis in R
- Presenting data discoveries in R

Expectations and Grading

Students will be graded for completing a final presentation (40%), submitting homework (40%); and class attendance/participation (20%). Each class session will include R coding based homework questions meant to reinforce that sessions lessons, and each week will include a brief written synopsis of progress towards the final project.

Final Project

Students will be expected to identify a set of biological or health sciences data that they can access and analyze. This can include publicly available data or data that you have access and permission to use for this project. We will discuss some sources but options can include data hosted in repositories, such as https://phenome.jax.org/, or publicly available datasets found here: https://datacatalog.med.nyu.edu/. Students will be expected to identify a data source early in the class, and provide regular updates in the form of written synopses on their progress. In the final two sessions, students will create a brief presentation using R Markdown and explain how they analyzed their data to their peers. Students will also be expected to briefly explain the scientific premise of their work.

General Policies

- Late/missed work: You will lose 10% for each day of lateness for up to one week.
- I will make announcements throughout the semester by e-mail. Make sure that your email address is updated; otherwise you may miss important emails from me.
- Always back up your work in a safe place (electronic file with a backup is recommended) and make a hard copy. Do not wait for the last minute to do your work. Allow time for deadlines.
- Plagiarism, the presentation of someone else's words or ideas as your own, is a serious offence and will not be tolerated in this class. The first time you plagiarize someone else's work, you will receive a zero for that assignment. The second time you plagiarize, you will fail the course with a notation of academic dishonesty on your official record.
- If you need extra time for health, familial or other personal reasons please speak with the instructor **before** missing any work or deadlines.

Pre-requisites:

Students are not expected to have any prior knowledge of R, but are expected to have worked with tabular data (i.e. spreadsheets) and have a basic familiarity with descriptive and inferential statistics.

Software needed:

Please first install R, and then install R Studio. A guide to doing this can be found at: https://www.datacamp.com/community/tutorials/installing-R-windows-mac-ubuntu

Schedule

Monday July 6 10:00-11:30
Class introduction; Syllabus Review; Introduction to R and R studio; Discussion of Final Project; Discussion of Data Resources; Beginning of Pulling Data into R

Thursday July 9 10:00-11:30
Troubleshooting; R Projects, R & GitHub; Data Types in R

End of Week 1: Identify a data source and write up briefly why you are interested in analyzing it. Provide general information: how many variables and observations it contains, potential areas of exploration or research, potential difficulties, why it is relevant. Also provide a brief explanation of what types of variables it contains.

Monday July 13 10:00-11:30
Data Structures in R

Thursday July 16 10:00-11:30
Functions, Apply, For Loops and If Statements

End of Week 2: Identify variables of interest in your dataset for analysis and presentation. Explain why they may be of interest or use in your research. What changes might need to be made to make them easier to analyze (e.g. dichotomization, normalizing).

Monday July 20 10:00-11:30
Exploratory Data Visualization

Thursday July 23 10:00-11:30
Bioinformatics Visualizations: Heatmaps

End of Week 3: Provide brief information based on summary statistics and basic visualizations about nature of your data: is it normally distributed or skewed, is it continuous or categorical. Focus on variables of interest to your final presentation.

Monday July 27 10:00-11:30
Publication-ready graphics with GGPlot2

Thursday July 30 10:00-11:30
Data Cleaning: Dealing with missing values, using DPlyr

End of Week 4: Describe any challenges you perceive or necessary transformations to your data. Outline your analysis plan

Monday August 3 10:00-11:30
Hypothesis Tests

Thursday August 6 10:00-11:30
Linear Regression

Monday August 10 10:00-11:30
Final Presentations Day I

Thursday August 13 10:00-11:30
Final Presentations Day II

Communication: Please post assignments to Brightspace and communicate by email.