R Programming for Data Analysis

### Instructor: Fred LaPolla, MLS

[fred.lapolla@nyulangone.org](mailto:fred.lapolla@nyulangone.org)

212.263.8535

577 First Ave, Room 209,

New York, NY 10016

### Teaching assistants: Linda Procell and Dania Annuar

[Linda.procell@nyulangone.org](mailto:Linda.procell@nyulangone.org)

[Dania.annuar@nuyulangone.org](mailto:Dania.annuar@nuyulangone.org)

## Dates: Mondays and Thursday 10:00-11:30am, July 8 - August 15, 2024

## Location: <https://nyumc.webex.com/nyumc/j.php?MTID=m969e50981d7af7618b8d1f4f0b08d697> and Science Building Room 920 (seats 16)

## Office Hours: By appointment

Course Site: [fredlapolla/VilcekSummerDataAnalysis2024: Course Materials for Vilcek Programming for Data Analysis Summer 2024 (github.com)](https://github.com/fredlapolla/VilcekSummerDataAnalysis2024)

## Overview:

Understanding code-based approaches to data analysis provide researchers with the ability to conduct reproducible, rigorous research without having to invest in expensive software. R is a widely used language for data science and statistical analysis and provides researchers with tools to unlock insights from their data. This class will introduce students to R using the RStudio development environment, walk them through features and functions of the program to complete original analysis on a dataset of their choosing.

## Objectives:

Students will be able to:

* Upload, clean and transform data to make it suitable for statistical analysis and visualization.
* Utilize R syntax to analyze their data.
* Troubleshoot R challenges to enable them future success in R coding and analysis.
* Use R Markdown to generate a reproducible report suitable for presenting their findings to others.

## Class Policies:

If you have any reason for needing accommodation of these policies, please inform the instructor or have the appropriate department contact him to inform of arrangements.

Homework must be submitted on time or you will lose points on your assignment. If there is a valid reason it must be late, you need to discuss the possibility of delay **before** the assignment is due to see if an extension can be made.

Homework should be submitted on GitHub Classes to the class session’s repository.

You are expected to do your own work and site sources for materials referenced/used.

Generative AI: This class aims to teach fundamentals of R programming, including syntax and structure. The homework are designed to reinforce these lessons, and I encourage you to try them without AI so as to cement an understanding of the basics of R. Working with generative AI in a coding environment requires a high degree of trouble-shooting, which requires understanding of the fundamentals. Any code you submit must actually run and accomplish its intended purpose. If you do decide to use generative AI on your final or other work, please be transparent and cite it.

### Schedule (topics and dates subject to change)

|  |  |  |
| --- | --- | --- |
| Date | Topic | Progress toward final project |
| Monday, July 8, 2024 | Syllabus review, discussion of final project, discussion of how to identify data sources | Explanation of final project and data finding |
| Thursday, July 11, 2024 | Introduction to R and the RStudio Environment, troubleshooting, packages |  |
| Monday, July 15, 2024 | Indexing and Data Types and Structures in R | Exploration of a dataset |
| Thursday, July 18, 2024 | Functions and Packages | Students should have identified a data set to work with and explore |
| Monday, July 22, 2024 | Apply, For Loops and If statements |  |
| Thursday, July 25, 2024 | Data Cleaning and Dealing with Missing Values | Students should begin to  identify a research question of interest to address from their dataset |
| Monday, July 29, 2024 | Advanced data cleaning with the tidyverse |  |
| Thursday, August 1, 2024 | Data Visualization for exploratory data analysis |  |
| Monday, August 5, 2022 | Data Visualization for publication | Students should be identifying variables to use to answer their research question |
| Thursday, August 8, 2022 | Hypothesis testing and analysis in R |  |
| Monday, August 12, 2022 | Hypothesis tests continued, linear regression | Students should have their data cleaned and usable and be ready to run analyses and write up results. |
| Thursday, August 15, 2022 | Reproducibility with R and R Markdown |  |

## Grading

Homework completion: 65%

Final Project: 35%

Final Project Guidelines: Students will find and choose a dataset from the biomedical life-sciences field. Using what we learn in class, they will identify an original research question to answer based on the data and use R to clean the data, transform it into a usable format and analyze it. Students will submit a 10-20 slide html presentation created using an RMD file on their results and methodology.

## Suggested Further Reading:

This class will not use a textbook, and learning will primarily be done by doing coding assignments following each lecture.

1. R for Data Science by Garrett Grolemund & Hadley Wickham (available [here](http://r4ds.had.co.nz/))

2. Discovering Statistics Using R by Andy Field, Jeremey Miles and Zoe Field: <https://us.sagepub.com/en-us/nam/discovering-statistics-using-r/book236067>