Memorial University of Newfoundland
Department of Mathematics and Statistics

DSCI 6607 – Programmatic Data Analysis Using R and Python
Fall 2024

Lectures:	Monday and Wednesday (12:00 – 13:15)
Location:	**CS–1009**
Instructor:	Dr. Armin Hatefi **E-mail**: ahatefi@mun.ca **Office hours**: Monday and Wednesday (9:00 – 10:00) - **HH-2025** or by appointment.
Course website:	Available through https://online.mun.ca/d2l/loginh/ (D2L)

Course description

Programmatic data analysis is an essential part of modern sciences. Data Scientists must not just be able to run existing programs, but to understand the principles on which they work. They must also be able to read, modify, and write scripts, so that they can assemble the computational tools needed to solve their data analysis problems. This class is an introduction to programmatic data analysis, targeted at data science students without assuming extensive programming background. Students will learn the core of ideas of programming—data structures, functions, iteration, input and output, simulations and models through writing code to assist in statistical analyses. Students will learn how to write maintainable code, as well as test code for correctness. They will learn how to set up and run stochastic simulations, how to fit basic statistical models and assess the results, and how to work with and filter data sets and data bases. Since code is an important form of communication among scientists, students will also learn how to comment and organize code. *Welcome to Data Science Club*!

Course Preparation & Prerequisites

There are several different programming languages for data science. By far the two most popular are R and Python. In this course we will be extensively using R and Python which are freely available for different OS. Students in this course are from various backgrounds; hence, we start with an introduction to programming; However, prior exposure to statistical thinking and to basic probability concepts is essential. Previous programming experience is not assumed.

Recommended Textbooks

1. Scientific programming and simulations using R, by Jones, Maillardet and Robinson, 2008.

2. R for Data Science by Wickham and Grolemund, 2016.

3. Python data science handbook: essential tools for working with data, by J Vanderplas, 2007.

4. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd ed., by Wes McKinney, O'Reilly Media, 2017.

Evaluation:

	Weight	Date	Time	Location
Assignments	50%*	See the following Table		
Projects	20%	See the following Table		
Final Exam	30%*	TBA	TBA	TBA

*__Note:__ If you perform better on the final exam than the assignments then the final exam weight will increase to 40% and the weight of assignments will decrease to 40%. Final exam will be cumulative and will cover all contents covered during the semester.
**__Note:__ More information on due date of projects will be discussed in class.

Projects	Weight	Due Date
Data Analysis Project with R	10%	TBA
Data Analysis Project with Python	10%	TBA

Assignments	Weight	Due Date
Assignment 1	10%	September 27
Assignment 2	10%	October 11
Assignment 3	10%	October 25
Assignment 4	10%	November 8
Assignment 5	10%	November 29

1. "Within the University community there is a collective responsibility to maintain a high level of scholarly integrity. A student is expected to adhere to those principles which constitute proper academic conduct." For more information on Memorial University's regulation on academic misconduct, see http://www.mun.ca/regoff/calendar/sectionNo=REGS-0748

2. "Memorial University of Newfoundland is committed to ensuring an environment of understanding and respect for the dignity and worth of each student and also to supporting inclusive education based on the principles of equity, accessibility and collaboration." For more information on Memorial University's commitment to accommodation of students with disabilities, see http://www.mun.ca/policy/site/policy.php?id=239. Students needing traditional exam accommodations should contact the Blundon Centre early in the term.

3. According to the University Calendar, the last day to drop courses without academic prejudice is **October 30, 2024**.

Copyright

The lectures and displays (and all materials) delivered or provided in DSCI 6607, including any visual or audio recording thereof, are subject to copyright owned by Dr. Armin Hatefi. It is prohibited to record, copy or share by any means, in any format, openly or surreptitiously, in whole or in part during or from this course, in the absence of express written permission from the instructor.

Course website and Communication

The course website is available through D2L portal and will be regularly updated with lecture notes, practice problems, assignments and readings. The course website will also be used for announcements and your grades. Students are encouraged to attend lectures, instructor office hours. Note that we follow **one email address policy. According to the policy, students must use**

only their MUN email account for communication and write a proper email including your name and student number; otherwise, your email will not be replied.

The tentative Course Contents

This course will be focusing on computational/numerical aspects of data analysis and statistical methods. Hence, you will need to make extensive use of the computer throughout the course. The general topics include: R basics; data wrangling and visualization using tidyverse; basic programming: branching, looping, vector based programming, and program flow; working with libraries in R: download, usage, extracting the results; programming with functions; Introduction to statistical models, simulations and numerical optimizations; Python language basics and environments; built-in constructs: data structures, functions, file I/O; NumPy basics: multidimensional array, array-oriented programming; pandas basics: data structures, functionality, mapping, ranking, descriptive statistics, working with formats (CSV, JSON, XML), fundamentals of SQL databases (SQLite), data wrangling and data aggregation; plotting and visualization: matplotlib, pandas and seaborn; time series.

The tentative Schedule

- *Waiter! R menu, please!* – `R & RStudio Basics`.
- *Waiter! I'm not comfortable with my data frame!* – `tibble`.
- *Waiter! How can I present my analysis!* – `RMarkdown`.
- *Waiter! There is a bug in my code!* – `Functional Programming`.
- *Waiter! Here is too hot?!* – `An Introduction to Quarto`.
- *Waiter! Why is your bar graph too ugly?!* – `An Introduction to ggplot2`.
- *Waiter! What's chef's special for tonight?!* – `Data Wrangling using tidyverse`.
- *Waiter! Why is restaurant so quiet today?!* – `Monte Carlo simulation and numerical optimization methods`.
- *Waiter! My analysis does not make sense?!* – `Statistical Models & Inference`.
- *Waiter! Python menu, please?!* – `Python Basics`.
- *Waiter! I don't like this environment?!* – `IPython, Anaconda, Jupyter & Quarto`.
- *Waiter! I can't handle my array?!* – `Array-oriented programming using NumPy`.
- *Waiter! I think there is a wrong order here?!* – `Data manipulation using Pandas`.
- *Waiter! Visualize my data?!* – `Visualization with matplotlib & seaborn`.
- *Waiter! Is there a Costco nearby?!* – `SQL databases with Python`.
- *Waiter! What time do you close?!* – `Time series with Python`.

And if time permits,

- *Waiter! Surprise me!* – `Resampling techniques for analysis: Bootstrap.`
- *Waiter! May I see your supervisor?!* – `Validation methods for analysis.`