STAT 4700/5030/7700 (Data Analytics and Statistical Computing)

Fall 2022 Syllabus

Time and Location:

- Mon/Wed 1:45 - 3:14 pm
- 8/30/2022 – 12/12/2022, JMHH F50 (MBA students are not required to attend class on days that they are on break according to the MBA calendar)

General Content:

- This course provides an introduction to programming for non-programmers in the widely used R language.
- The course also presents the basics of data visualization, data wrangling, stochastic simulation, statistical inference based on simulation (bootstrap), text analysis, and numerical optimization using R.

Prerequisites:

- STAT 111/112 or STAT 101/102 or STAT 431 or STAT 613 or ECON 103/104.
- No programming experience is required. This may not be the best course for experienced programmers since they can easily pick up R using online resources.

Organization:

- Instructor: Rommel G. Regis
  E-mail: rgregis@wharton.upenn.edu
  Office Hours: Tue/Thu, 5 – 6 pm or by appointment, Zoom
  https://upenn.zoom.us/j/95241061647
- There are two TA/Graders for the course. Their office hours will be announced later.
- Canvas will be used to post announcements, lecture slides, R scripts, R markdown files and html renderings, data sets, homework, online quizzes, and take-home exams. The materials will be organized in weekly modules.
- Poll Everywhere will be used for class participation.
- Ed Discussion will be used for online discussions.

COVID Policies:

- We will follow all university guidelines regarding safety and precautions against COVID.
- Masks are not required in our classroom, but anyone who wishes to wear a mask is welcome to do so.

- The university currently recommends masks in indoor public settings, including classrooms, for individuals with underlying medical conditions. Hence, please be considerate of your classmates who prefer to wear a mask since they (or someone they live with) might have underlying medical conditions.

Class Materials:

- Textbook: None
- Lecture Slides (to be posted on Canvas)
- The open-source R statistical software and the R studio IDE (Integrated Development Environment) for R (https://www.rstudio.com/products/rstudio/download/)
- MAC Users: If RStudio does not work for you, you may have to download the raw version of R from https://cran.r-project.org
- R scripts, R markdown files and html renderings, and data sets (to be posted on Canvas)

Course Requirements and Grading:

- 15% Midterm Exam (take-home)
- 20% Final Exam (take-home)
- 65% Homework, Quizzes and Class Participation

Quiz Policies:

- Quizzes are to be taken in class unless otherwise specified.
- To be completed individually, but you are free you use any of the course materials or online sources.
- Make-up quizzes will only be given in special circumstances (e.g., illness or family emergency) at the discretion of the instructor. Additionally, you may be asked to provide some documentation to avail of a make-up. Please notify me by e-mail (rgregis@wharton.upenn.edu) to request a make-up, if you have a reasonable excuse for missing a quiz.

Homework Policies and Collaboration:

- You are expected to turn in your assignments on time. Points will be deducted for late work and a homework that is three days late will no longer be accepted unless there is a special circumstance (e.g., illness or family emergency) that prevented you from submitting your assignment on time. Please notify me by e-mail (rgregis@wharton.upenn.edu) to request an extension without penalty, if you have a reasonable excuse. Again, you may be asked to provide some documentation to avail of an extension.
- You may discuss the homework with other students. However, you are expected to prepare your homework by yourself (e.g., write your own code). In addition, you are expected to acknowledge any collaboration by including a sentence of the following

form at the beginning of your assignment: “Help on this class requirement was received from __________.” Verbatim copying of another student’s code or failure to acknowledge collaboration may result in a zero for the assignment.

Class Participation:

- Class participation is part of your grade.
- You are expected to participate in class by answering or asking questions. There is no penalty for a wrong answer or for a question whose answer may be obvious to many of your classmates. It is better to participate and give a wrong answer than never to participate at all. Think of making a mistake as a learning experience.
- We will be using Poll Everywhere (pollev.com/rommelregis328) regularly in class. You are expected to answer the poll questions in class, but again, there is no grade penalty for a mistake.
- We will also be using Ed Discussion for class discussions. This is accessible through Canvas. You are highly encouraged to use Ed Discussion to ask your questions about the course material or homework.
- Your level of participation in Poll Everywhere or Ed Discussion will determine your class participation grade.

Classroom rules of conduct:

- Be respectful and helpful to fellow students. Helping your classmates debug their code will also help you become a more proficient programmer.
- Please use your laptops or tablets only for purposes related to our class. No email, texting, and social media during class time.

Students with Disabilities:

- If you have a documented disability, please contact me as soon as you can, preferably within the first two weeks of class.

Course Content: We will cover the following topics and demonstrate how to perform various computing and data analysis tasks using R:

I. Introduction to R Programming
   a. Syntax and atomic data types
   b. Data structures (vectors, matrices, arrays, data frames, lists, factors)
   c. Numerical and graphical summaries of data, empirical distributions
   d. Writing your own functions
   e. Flow control and iteration (conditionals and loops)

II. Introduction to Simulation
   a. Random number generation
   b. Probability simulations and Central Limit Theorem demo

c. Monte Carlo integration

III. Data Visualization

a. Grammar of graphics of the ggplot2 package

b. Choropleth maps and heat maps

c. Time series plots

d. Visualization of multivariate data

IV. Data Wrangling

a. Grammar of data wrangling using the dplyr package

b. Using the pipe operator as an alternative to function composition

V. Review of Basic Statistical Inference

VI. Review of Simple and Multiple Linear Regression Models

VII. More on Stochastic Simulation

a. Monte Carlo methods in inference

b. The bootstrap

c. Permutation tests

VIII. Introduction to Text Mining

a. The tidy text format

b. Word clouds

c. Introduction to Sentiment Analysis

IX. Introduction to Numerical Optimization Using R

a. Overview of optimization in Statistics and Machine Learning

b. R packages for Linear Programming and Nonlinear Optimization

c. Application to Numerical Maximum Likelihood Estimation and Simulation-Based Optimization