DACSS 601 Data Science Fundamentals

University of Massachusetts Amherst
Meeting Time: 5:30–6:30 pm on MWF
Instruction Mode: Asynchronous
Meeting Venue: Slack
Summer 2023

Instructor

Dr. Hui Zhou
Data Analytics and Computational Social Science (DACSS)
College of Social & Behavioral Sciences
University of Massachusetts Amherst
Email: huizhou@umass.edu
Office hours: by appointment
Time Zone: EDT

Tutor

Larrisa Miller
Ph.D. student, Department of Communication
Office hours: make appointments on calendly.
Email: larrisamille@umass.edu

Course Description

This 3 credit course provides students with an introduction to the R programming language that will be used in all core courses and many of the technical electives. There is a growing demand for students with a background in generalist data science languages such as R, as opposed to more limited software such as Excel or statistics packages such as SPSS or Stata. The course will also provide students with a solid grounding in general data management and data wrangling skills that are required in all advanced quantitative and data analysis courses.

This course does not assume that students have any prior knowledge of statistical programming, although some experience with R or other programming languages can help to some extent. However, students without programming experience should not feel disadvantaged, as I assume everyone is learning R from scratch. This course is offered in an asynchronous manner. An *optional* one-hour online discussion section is

scheduled between 5:30 pm and 6:30 pm on Mondays, Wednesdays and Fridays. Before each section, I will make and share a 45–60 minutes video with the class one day earlier at noon. Students are supposed to read the required book chapters, watch the online video, and attend the discussion section *if available*. The online video is designed to help students understand reading materials and acquire coding skills. The discussion section will enable students to ask questions and interact with me and classmates.

Course Objectives

- Equip students with the skills necessary to conduct statistical analyses in R and help them understand implementing data science research designs across a variety of settings.
- Provide students with the tools to design and complete basic data science tasks of their own and in group collaborations.
- Demonstrate the importance of technological and statistical literacy for purposes of analysis, argument, and understanding, with students capable of critically engaging research and identifying both the strengths and weaknesses of increasingly common arguments based on empirical evidence.
- Enable students to communicate clearly and appropriately in both oral and written format the results or shortcomings of data-centered research.

Statistical Software

`R` is an open-source statistical software developed by statisticians. It is one of the most popular statistical tools in both academia and business. There are platforms that make `R` easier to use. Those platforms are called Integrated Developing Environment (IDE). IDEs have a battery of features important to developers, including coding style, package management, debugging, etc. This course will adopt RStudio, which is probably the most popular IDE for `R`. Recently, the RStudio company changed its name to posit, but this change does not have any impact on our course. We will still refer to it as RStudio for straightforwardness. Just keep in mind $Posit = RStudio$.

Although a desktop version of RStudio is readily available for different operating systems, we will be using RStudio Cloud, or Posit Cloud, which allows all of us to use RStudio online for free. There are three advantages of utilizing the RStudio Cloud relative to downloading a desktop version for local usage.

- There is no need to install R and RStudio. All we need is an RStudio Cloud account that enables us to log in to use the R language.

- When we utilize RStudio Cloud, everyone is on the same page regardless of the differences in their computer operating systems (Windows vs. Mac OS). Additionally, everyone will have the same version of R and RStudio. This will make it much easier for you to follow my instructions.
- RStudio Cloud not only serves as online statistical software but it also enables us to store files on the cloud. That means you can always work on your projects remotely. Just grab a digital device such as a laptop, tablet or even cell phone; log into your RStudio Cloud account; start working on your projects.

The only downside to the RStudio Cloud is that you must have an Internet connection, as it relies on cloud services. RStudio Cloud can be free if you are a light user ($\leq 25\ hours\ per\ month$) and do not need teamwork. However, in this class we will use RStudio Cloud heavily and collaboratively (e.g., I will grade your homework in RStudio Cloud). Thus, each student must purchase the Cloud Plus plan at the price of $5 per month, which is quite affordable given that all texts are publicly available for free.

If we compare RStudio to a smartphone, there are numerous packages—equivalent to smartphone apps—that make RStudio even more powerful and useful. One of those packages is called `RMarkdown`. `RMarkdown` can incorporate `R` code, outputs, and texts in a single file, thus avoiding repeatedly copying and pasting R code and outputs from R to other editors such as Microsoft Word. It can also be a good tool for document formatting. All our lecture notes, homework assignments, online modules, and the final project must be written in `RMarkdown`. We will learn it in the first learning unit.

Course Management

Blackboard

We will use Blackboard to manage course materials, including the syllabus, homework assignments and grades. Importantly, homework assignments must be completed in RMarkdown via RStudio Cloud, compiled as a MS Word document, and submitted to Blackboard via a Turn-It-In link. Grades will also be managed through Blackboard.

RStudio Cloud

We adopt RStudio Cloud to write and execute R code. The basic workflow for each learning unit is that I will post lecture notes (i.e., lab handouts) on RStudio Cloud. All lecture notes will be saved in my RStudio Cloud workspace entitled *DACSS 601*, which will be shared with the entire class so that students can access these files. As noted earlier, students must purchase the *Cloud Plus* plan before they can accept the invitation

and become a member of the workspace. This membership costs $5 per month.

Slack

Although the Blackboard and the RStudio Cloud are fascinating tools, we cannot use them to communicate with each other synchronously. Neither are they suitable for managing large files such as a one-hour video. Thus, we need to use Slack, a free and convenient tool for teamwork. As with the RStudio Cloud, I also created a workspace named *DACSS 601* on Slack. That will be our virtual classroom. Course videos and Zoom links for discussion sections will be available on Slack. Additionally, we will have several chatting channels on Slack. Feel free to ask questions and share thoughts!

Text

This course adopts the book *R for Data Science: Visualize, Model, Transform, Tidy, and Import Data*, dubbed *R4DS*. This is a popular entry-level text for data science. While the second edition is about to come out, we will make use of the first edition published in 2017. The reason for this choice is twofold. First, the second edition does not change dramatically in the most important chapters. In fact, the major changes come from some newly introduced chapters that focus on additional topics. Second, the authors have not yet announced the completion of their second edition. We do not need to risk the chaos brought about by their updates to the most-up-to-date version. When necessary, we will reference their second edition, dubbed *R4DS2E*.

In addition, we will draw on several chapters from the four books listed below. All books are publicly available and free of charge. Please click the following hyperlinks to either view them online or download a copy.

- (Required) Wickham, H., & Grolemund, G. (2017). *R for Data Science: Visualize, Model, Transform, Tidy and Import Data*. O'Reilly Media. [dubbed **R4DS**]
- (Chapters Selected) Wickham, H., Cetinkaya-Rundel, M., & Grolemund, G. (2023). *R for Data Science: Import, Tidy, Transform, Visualize and Model Data*. O'Reilly Media. [dubbed **R4DS2E**]
- (Chapters Selected) Dalgaard, P. (2008). *Introductory Statistics with R*. Springer Publication. [dubbed **ISR**]
- (Chapters Selected) James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: Springer. [dubbed **ISL**]
- (Chapters Selected) Wickham, H. (2021). *Mastering Shiny: Build Interactive Apps, Reports & Dashboards Powered by R*. O'Reilly Media. [dubbed **MS**]

Assignments and Grading Policy

1. Online Modules (45%). There are nine online modules for students to complete on RStudio Cloud. These online modules are closely based on the reading/video/discussion materials covered in the past two days. Online modules mostly consist of coding questions, although there might be short-answer questions as well. Students have two days to complete a module. They earn points as long as they demonstrate the ability to apply R in a reasonable way, and they lose points if they don't even give it a try. **Online modules should be completed and left on RStudio Cloud. It does not need to be submitted to Blackboard**.

2. Homework (30%). Three homework assignments account for 30% of the final grade. A homework assignment will be posted on RStudio Cloud after three learning units have been completed (noncumulative). Students shoud complete homework in RMarkdown on RStudio Cloud. While they can engage in group study, I have a zero-tolerance policy for academic dishonesty, including but not limited to plagiarism and AI-assisted work. To be clear, this course forbids the use of ChatGPT and other AI tools in assignments. **I require that students must compile homework into a Word file and submit it to Blackboard via a Turn-It-In link**.

3. Final Project (15%). Additionally, students must complete a final project, which could be a research paper, web app or scraping project.

    - If a student intends to write a research paper, she or he must have a clear question and conduct quantitative analysis in R to answer the question. The analysis could be descriptive or inferential, depending on the nature of the question and availability of data. When citing literature, use the APA style and make sure the citation style is consistent throughout the paper.
    - In the case of a web app, students can develop an online dashboard using RShiny, which will be taught in the Learning Unit 13 on August 14.
    - Students might choose to develop a scraping project by drawing on the skills they acquire from Learning Unit 10. In this scenario, they must make sure the target website stores data on at least five web pages, and their scraping script should yield a clean dataset with no less than five columns/variables and one hundred rows/observations.

4. Participation (10%). It is imperative that students actively participate in class discussion on Slack. Students are expected to participate regularly, and participation should reflect careful consideration of the topic. Participation does not need to reflect expertise; rather, students should try to ask and answer questions regularly.

Final letter grades are assigned using the University's Plus-Minus Grading Scale. Final grade percentages ending in a decimal of .5 or greater will be rounded up to the next whole number. Below is the grading scale.

A	$\in$	$[94, 100]$	
A-	$\in$	$[90, 93]$	Excellent
B+	$\in$	$[86, 89]$	
B	$\in$	$[81, 85]$	Good
B-	$\in$	$[77, 80]$	
C+	$\in$	$[74, 76]$	
C	$\in$	$[70, 73]$	Fair
F	$\in$	$[0, 70)$	Fail

Deadlines

Both online modules and homework appear in the form of `.Rmd` files (i.e., RMarkdown files). Online modules are posted on the RStudio Cloud (labeled as assignment). Students should complete the modules by inserting code, executing code, typing text to answer questions. I will grade and make comments on the online modules once the due date comes. Homework is posted on Blackboard. Students should upload a word document to Blackboard after completing the homework. The word document can be produced by knitting a `.Rmd` file. Below are the assignment deadlines.

- Jul 16: I will post **Online Module 1** at noon. It will be due **at noon, Jul 18**.
- Jul 18: I will post **Online Module 2** at noon. It will be due **at noon, Jul 20**.
- Jul 20: I will post **Online Module 3** at noon. It will be due **at noon, Jul 22**.
- Jul 23: I will post **Homework 1** at noon. It will be due **at noon, Jul 30**.
- Jul 23: I will post **Online Module 4** at noon. It will be due **at noon, Jul 25**.
- Jul 25: I will post **Online Module 5** at noon. It will be due **at noon, Jul 27**.
- Jul 27: I will post **Online Module 6** at noon. It will be due **at noon, Jul 29**.
- Jul 30: I will post **Homework 2** at noon. It will be due **at noon, Aug 6**.
- Jul 30: I will post **Online Module 7** at noon. It will be due **at noon, Aug 1**.
- Aug 1: I will post **Online Module 8** at noon. It will be due **at noon, Aug 3**.
- Aug 3: I will post **Online Module 9** at noon. It will be due **at noon, Aug 5**.
- Aug 6: I will post **Homework 3** at noon. It will be due **at noon, Aug 13**.

- Aug 14: Students should discuss the **Final Project** with me if they have no clue about what to do at this point. The Final Project will be due **at noon, Aug 24**. The final grade will be posted by Aug 28.

Academic Honesty

Since the integrity of the academic enterprise of any institution of higher education requires honesty in scholarship and research, academic honesty is required of all students at the University of Massachusetts Amherst.

Academic dishonesty is prohibited in all programs of the University. Academic dishonesty includes but is not limited to: cheating, fabrication, plagiarism, and facilitating dishonesty. Appropriate sanctions may be imposed on any student who has committed an act of academic dishonesty. Instructors should take reasonable steps to address academic misconduct. Any person who has reason to believe that a student has committed academic dishonesty should bring such information to the attention of the appropriate course instructor as soon as possible. Instances of academic dishonesty not related to a specific course should be brought to the attention of the appropriate department Head or Chair. The procedures outlined below are intended to provide an efficient and orderly process by which action may be taken if it appears that academic dishonesty has occurred and by which students may appeal such actions. Since students are expected to be familiar with this policy and the commonly accepted standards of academic integrity, ignorance of such standards is not normally sufficient evidence of lack of intent.

For more information about what constitutes academic dishonesty, please see the Dean of Students' website: http://umass.edu/dean_students/codeofconduct/acadhonesty/.

Statement on Disabilities

The University of Massachusetts Amherst is committed to making reasonable, effective and appropriate accommodations to meet the needs of students with disabilities and help create a barrier-free campus.

If you are in need of accommodation for a documented disability, register with Disability Services to have an accommodation letter sent to your faculty. It is your responsibility to initiate these services and to communicate with faculty ahead of time to manage accommodations in a timely manner. For more information, consult the Disability Services website at http://www.umass.edu/disability/.

Course Schedule

Learning Unit 1: Getting Started with RStudio (Posit) Cloud and RMarkdown

1. Meeting schedule: July 17 (Monday)
2. Readings: Chapters 1 & 27, R4DS
3. Skills to be acquired:
    - the scope of data science
    - cloud account set up
    - package installlation
    - help documentation
    - RMarkdown syntax and workflow

Learning Unit 2: Dealing with Variables in Base R

1. Meeting schedule: July 19 (Wednesday)
2. Readings: Chapter 1.2.1–1.2.9, ISR
3. Skills to be acquired:
    - work directory
    - data import
    - data structure: scalar, vector, matrix, list, dataframe
    - type of variables: logical, integer, double, character, date-time, factor
    - functions for different types of variables
    - missing values
    - indexing

Learning Unit 3: Dealing with Dataframes in Base R

1. Meeting schedule: July 21 (Friday)
2. Readings: Chapter 1.2.10–1.2.16, ISR
3. Skills to be acquired:
    - data generation (random and customized data)
    - R data

- indexing in a dataframe
- naming
- recoding
- reshaping
- subsetting
- sorting
- ordering
- merging
- appending

Learning Unit 4: Data Visualization in Base R

1. Meeting schedule: July 24 (Monday)
2. Readings: Chapter 4, ISR
3. Skills to be acquired:
   - histograms
   - bar plots
   - time series plots
   - pie charts
   - scatter plots
   - par() function
   - saving a figure

Learning Unit 5: Data Wrangling with Tidyverse: Part I

1. Meeting schedule: July 26 (Wednesday)
2. Readings: Chapters 10, 11, 12, 14 & 15, R4DS
   - Additional reading: Differences between the base R and magrittr pipes by Hadley Wickham
3. Skills to be acquired:
   - pipe operator
   - tibble

- readr
- readxl
- tidyr
- stringr

Learning Unit 6: Data Wrangling with Tidyverse: Part II

1. Meeting schedule: July 28 (Friday)
2. Readings: Chapters 5, 13, 15, 16 & 18, R4DS
3. Skills to be acquired:
    - dplyr
    - forcats
    - lubridate

Learning Unit 7: Data Visualization with ggplot2

1. Meeting schedule: July 31 (Monday)
2. Readings: Chapter 3, R4DS.
3. Skills to be acquired:
    - geom elements
    - aesthetics
    - annotations
    - facet
    - coordinates
    - theme setting

Learning Unit 8: R Programming: Part I

1. Meeting schedule: August 2 (Wednesday)
2. Readings: Chapters 19.2, 19.4, 19.5, 19.6, 21.1, 21.2, 21.3, 21.4, R4DS
3. Skills to be acquired:
    - conditions
    - functions

- iterations (i.e., for and while loops)
- other keywords: switch, repeat, next, break

Learning Unit 9: R Programming: Part II

1. Meeting schedule: August 4 (Friday)
2. Readings: Chapters 21.5, 21.6, 21.7, 21.8, R4DS
3. Skills to be acquired:
    - the apply family in base R
    - the map family in purrr package
    - package development

Learning Unit 10: Web Scraping with rvest

1. Meeting schedule: August 7 (Monday)
2. Readings: Chapters 16 & 25, R4DS2E
3. Skills to be acquired:
    - scraping html source code
    - xpath
    - regular expressions
    - automatic execution

Learning Unit 11: Statistical Modeling

1. Meeting schedule: August 9 (Wednesday)
2. Readings: Chapter 3, ISL
3. Skills to be acquired:
    - linear models
    - generalized linear models
    - interactive models
    - effects interpretation
    - effects visualization
    - model goodness-of-fit

Learning Unit 12: Machine Learning

1. Meeting schedule: August 11 (Friday)
2. Readings: Chapter 2, ISL
   - Additional reading: Evaluation Metrics for Classification Models—How to measure performance of machine learning models? by Selva Prabhakaran
3. Skills to be acquired:
   - supervised and unsupervised learning
   - variance-bias trade-off
   - overfitting
   - confusion matrix
   - cross validation
   - bootstrap
   - caret

Learning Unit 13: Dashboard Development with RShiny

1. Meeting schedule: August 14 (Monday)
2. Readings: Chapters 1, 2 & 3, MS
3. Skills to be acquired:
   - UI setting
   - server setting
   - publishing RShiny apps

Learning Unit 14: Introduction to Python and Google Colab (and Jupyter Notebook)

1. Meeting schedule: August 16 (Wednesday)
2. Readings: None
3. Skills to be acquired:
   - pandas
   - numpy
   - Python programming
   - scikit-learn