Data Science Fundamentals

DACSS 601-02
University of Massachusetts Amherst
Summer 2022

Instructional Team

Instructor

Sean Conway
Tobin Hall 517
spconway@umass.edu

Teaching Assistant

Abhinav Kumar
abhinavkumar@umass.edu

*Course Time: Monday/Wednesday 10:00 - 11:30 AM ET (**OPTIONAL** synchronous meeting time)*
Course Location: Online (Zoom link, password: tidy)
Office Hours: DM the instruction team on Slack or email them.

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.

Course Objectives

- Equip students with the skills necessary to conduct statistical analyses in R, capable of understanding 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.

Textbook

There is one required text for this course, **freely available online:**

> Wickham, H., & Grolemund, G. (2016). R for data science: Visualize, model, transform, tidy, and import data. *OReilly Media*.

We **may** also reference the following texts and associated online content in this course. These are not required, but any interested students can also find them freely available online.

> Wickham, H. (2019). *Advanced R*. Chapman and Hall/CRC.
>
> Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational I and Graphical Statistics*, *19*(1), 3-28.

This syllabus outlines general areas of study throughout the course, as well as listing specific assignments on a daily basis. It is vital that you keep current with the assignments, as they will provide the basis for in-class lectures, discussions, and activities.

Course Website

All classroom material will be posted in the Google Classroom. You will be submitting assignments via RPubs as part of a course blog.

Grading

Grades are calculated as follows:

- Tutorials (45%)
- Homework (25%)
- Research Project (35%)

Tutorials Students are expected to complete eleven online R tutorials (9 regular, 2 statistics-based) that will walk them through data science tasks in R. These tutorials are FORMATIVE work - they are not intended to negatively affect your grade but to support your learning. Therefore, students who submit every assignment will receive at least a 90 (or 36 out of 40 points), regardless of their actual score.

Homework Various homework assignments will be distributed throughout the course. There are six significant assignments that are intended to help you engage with a specific dataset and related empirical question. Each homework builds on skills from the previous homework, and builds towards the final student project. The assignments will help you build your ability to capably and efficiently accomplish data science tasks in R and will establish a foundation for your subsequent methods courses. Feedback is provided to help individual students “stretch” and improve their data management and programming skills regardless of their background. All homework assignments will be due on **the Monday night of the associated due date, by midnight ET.**

Research Project More information regarding the research project will be made available on the course website. Students are expected to prepare a basic data analysis pipeline and resulting visualization at the end of the semester, and to draft a final discussion post that reflects on their approach, experience, and challenges in drafting the visualization, as well as what they would do differently next time. Students with no prior experience in R may want to consider one of the relatively clean datasets that we provide; while students with extensive programming experience should choose ambitious projects involving multiple datasets and more extensive data cleaning. Examples are provided of past projects, to provide a sense of the range of possible submissions. The goal is to increase individual skills and comfort with R and data management during the course, and we fully expect some students will submit projects considerably more complicated than others due to their prior experience.

Final letter grades are assigned using the University's Plus-Minus Grading Scale according to following rubric:

A (94-100%)

A- (90-93%)

B+ (86-89%)

B (81-85%)

B- (77-80%)

C+ (74-76%)

C (70-73%)

F (Below 70%)

Software

Students in this class will use R and RStudio. The software is free and available online; the course website includes a guide for installing both on your machine. The course assumes no familiarity with the R programming language. If you run into issues during R installation, we recommend that you create a free RStudio Cloud account to use while you consult with an expert to sort out any installation issues.

We will also make use of Google Classroom, GitHub and Slack. All of these are freely available for use online.

The instructor cannot provide support for installation or other general computing issues, and can provide only limited support for hands-on debugging (e.g., during class, during drop-in office hours, via slack.) We recommend that students work in small groups and support each other as much as possible during class. Additionally, DACSS maintains a list of tutors who can provide more specialized and intensive support for an hourly fee.

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.

Course Schedule

Detailed schedule information with all asynchronous materials, assignment instructions and due dates is provided on Google Classroom.

Week	Tutorials Due	Homework Due	Async Materials (to be completed ***PRIOR*** to synchronous discussion sessions)
5/23/2022	R Basics	Install R/RStudio	Intro to Data Science
5/30/2022	RMarkdown		Getting Started in R
6/6/2022	Statistics 1 & 2	HW1	Data Import
6/13/2022	Data Wrangling	HW2	Data Wrangling
6/20/2022	Transforming Data		Tidy Data 1
6/27/2022	Intro to Visualization	HW3	Intro to visualization
7/4/2022	Advanced Programming, Data Structures	HW4	Tidy Data 2
7/11/2022	Advanced Visualization		Visualizing statistics
7/18/2022	Modeling Basics	HW5	More visualization
7/25/2022	NONE		Best practices in visualization
8/1/2022	NONE		Programming with Functions
8/8/2022	NONE	HW6	More Data Science Topics
8/15/2022	NONE	FINAL PROJECT DUE 8/19	NONE (Final Project Work)