Data Science Fundamentals

DACSS 601
University of Massachusetts Amherst
SS2 2021

Contact Information

Dr. Curtis Atkisson
School of Public Policy
E-mail: `catkisson@umass.edu`

Course Time: Mon 4-5; Wed 7-8; Fri 7-8
Course Location: Online
Office Hours: Tues 11-1; Fri 4-6; or by appointment

Course Description

This three-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 document

This is a living and breathing document. I believe that students learn best when they take control of their education. You will have plenty of opportunities to do so, and this document may change to reflect that.

Course Goal

There is one overriding goal in this course: to get you set up and exposed to the tools used in modern data analysis. For most of you, this will set you up to succeed in your other courses that you take in the program. In short, I don't care how you get to a basic understanding of these tools—we will get you there by hook or by crook. And that is the entirety of what you will be graded on. That's the whole thing.

Course Objectives

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

Text

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

Wickham, Hadley and Garrett Grolemund. *R for Data Science: Visualize, Model, Transform, Tidy, and Import Data*. O'Reilly. [link]

We will also reference the following texts and associated online content in this course. Interested students might consider acquiring either or both of the following texts, but again, this is not required:

Imai, Kosuke. *Quantitative Social Science: An Introduction*. Princeton University Press.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics.

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 presentations, discussions, and activities.

Grading

Grades are calculated as follows:

Modules (30%)

Homework (25%)

Midterm Examination (10%)

Research Project (25%)

Participation (10%)

Modules Students will be required to complete online R modules that will walk them through data science tasks in R. Students are expected to complete each and every module; modules are graded

for completeness rather than having the perfect answer. For that reason, failing to complete modules will carry the penalty of forfeiting the entirety of the points available for the module. Students seeking accommodations *must notify the professor **BEFORE** the modules are due.*

Homework Eight homework assignments will be distributed during the course of the semester. The assignments are intended to further engage students with a particular empirical question, to familiarize students with data science, and to build student ability to capably and efficiently accomplish data science tasks in `R`, establishing a foundation for subsequent courses. Each homework assignment will include web-based data science tutorials, and a section that requires students to apply the lessons from the tutorial with their own data. Details on each assignment will be distributed in class and through the course website. Homework assignments will be submitted online on the dates indicated in the syllabus. Late assignments are assessed a one letter grade penalty per calendar day late. The homeworks all build towards your final project, so by doing them you will accomplish your final project.

Midterm Exam More information regarding the midterm exam will be made available in class and online. Students will be required to complete an open-notes online exam that includes a series of short-answer questions related to elementary data science tasks in `R`, then a more elaborate task of getting data into `R` and prepared for analysis. Students are expected to complete the exam within the scheduled time period. Failing to complete the exam will carry the penalty of forfeiting the entirety of the points available for the exam. Students seeking accommodations *must notify the professor **BEFORE** the exam.* The goal of the midterm is to show me what you know and what we need to go over again. This is what is called a formative assessment.

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 seeking accommodations *must notify the professor **BEFORE** the due date.*

Participation It is imperative that students actively participate in class discussion wherever it may be: Slack, Piazza, Google Classroom, course blog, etc. 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 seek to both ask and answer questions regularly and in equal proportion.

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

A (90-100%)

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`, `RStudio`, and `git`. 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.

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

Note that reading assignments are listed according to the day on which the subject matter will be discussed; they should therefore be read prior to that date. This course officially has 3 meeting days per week, for a total of 17 sessions. 13 sessions will have content that we will work through as a class and 4 sessions will be available for general questions and as working sessions.

Day 1: Introduction to Data Science

Before the first class, students should view the introductory course video posted to Classroom, which provides an overview of the syllabus, background material, `R`, and `RStudio`.

Videos: Intro to Data Science Fundamentals, & The `tidyverse`
Tutorials: Tutorial 1 & Tutorial 2.
Reading: Chapter 1, RDS.
Due: Install `R`, `RStudio`, and `git`.

Day 2: Getting Started in `R`

Data types, vectors, data wrangling, coding basics.

Videos: Coding basics, Functions, Writing scripts, Github, RMarkdown basics, & RMarkdown Text and Code
Tutorials: Tutorial 2B
Reading: Chapters 4, 6, 27, RDS.
Due:

Day 3: Programming in `R` and `tidyverse`

Command interface, directory structures.

Videos: Working directories and data input, Data import, Tibbbles & Measurement and coding
Tutorials: Tutorial 5
Reading: Chapters 10 & 11, RDS.
Due: Homework 6

Day 4: Tidy data and data wrangling

Principles of tidy data, basic dplyr verbs.

Videos: Tidy data, Filters, & Arrange and select
Tutorials: Tutorial 6
Reading: Chapter 5 & 12, RDS.
Due: Homework 1

Day 5: Transforming Data

Filtering, arranging, subsetting, and grouping data.

Videos: Pipes
Tutorials: Statistics Tutorial 1
Reading: Chapter 18, 29, & 30, RDS.
Due: Homework 2

Day 6: Project day

Due: Homework 3

Day 7: Introduction to Visualization

What is data viz and ggplot2.

Videos: ggplot intro
Tutorials: Tutorial 3
Reading: Chapter 3, RDS.
Due:

Day 8: More visualization

Univariate vs bivariate and make some pretty pictures.

Videos: Central tendency and dispersion, Covariation and missing data, & Univariate and Bivariate Visualizations
Tutorials: Statistics Tutorial 2
Reading:
Due:

Day 9: Advanced visualizations

Grammar of graphics and variable conversion.

Videos: Adding dimensions
Tutorials: Tutorial 4
Reading: Chapters 7 & 15, RDS.
Due: Homework 4: Univariate

Day 10: Best practices in visualization

Facets and more ggplot.

Videos: Facets & Best practices for visualization
Tutorials: Tutorial 7
Reading: Chapter 28, RDS.
Due: Homework 5: Data Visualization

Day 11: Project day

Due: ***Midterm Exam***

Day 12: Vectors and Functions

Functions and vectors.

Videos: Vectors & Functions
Tutorials: Tutorial 8
Reading: Chapter19 & 20, RDS.
Due:

Day 13: Basics of Modeling

Intro to modeling.

Videos: Model basics & Model fitting
Tutorials:
Reading: Chapter 22 & 23, RDS.
Due: Homework 6: Bivariate

Day 14: Advanced modeling

Linear model and predictions.

Videos: Model families, Predictions, Conditional relationships, & Cross validation
Tutorials:
Reading: Chapter 25 & 25, RDS.
Due:

Day 15: Project day

Day 16: Fun stuff!

Here are a few specialized applications of tools we have learned this term.

Videos:Text-as-data, Machine learning, & Network Analysis

Day 17: Project day
