Applied Data Science I

Kyle Scot Shank

2021-09-12

Contents

Welcome

This is the website for the **"Applied Data Science I"** course series at College of the Atlantic. It will be part book, part course page, and part collaborative learning effort between student and faculty contributors. The focus of this document will be to teach practical and applied skills in "data science" using the R programming language. No formal knowledge of computer science or R programming is required and the mathematical pre-requisites will be minimal. You'll learn how to explore and visualize data from both a practical and theoretical perspective, with a special focus on the real-world implications and ethical considerations often ignored when teaching introductory data science skills and techniques. We will focus on learning a lot and having fun while doing it.

Basic Information

Instructor: Kyle Scot Shank, '14

COA Email: ksshank@coa.edu (I will always try to get back to you within 24 hours, but there may be ocassional periods when I am a bit slower.)

Pronouns: he/him/his

Class Meeting: CHE 102 (Center for Human Ecology, Flexible Classroom), Tuesdays and Fridays, 1:00pm to 2:25pm

Course Description

This course will be both a broad overview of how "data science" is done in the "real world". We will have a specific focus on learning the *craft* of data science and analytics, working to build a suite of various practical data skills that can extend across a variety of different knowledge domains.

This would be a great class for those interested in an introductory exploration of data science as a topic, as well as those looking to add a degree of analytical expertise to pre-existing work and interests. We'll be focusing on four major areas of application: (1) properly building data-driven questions and hypotheses;

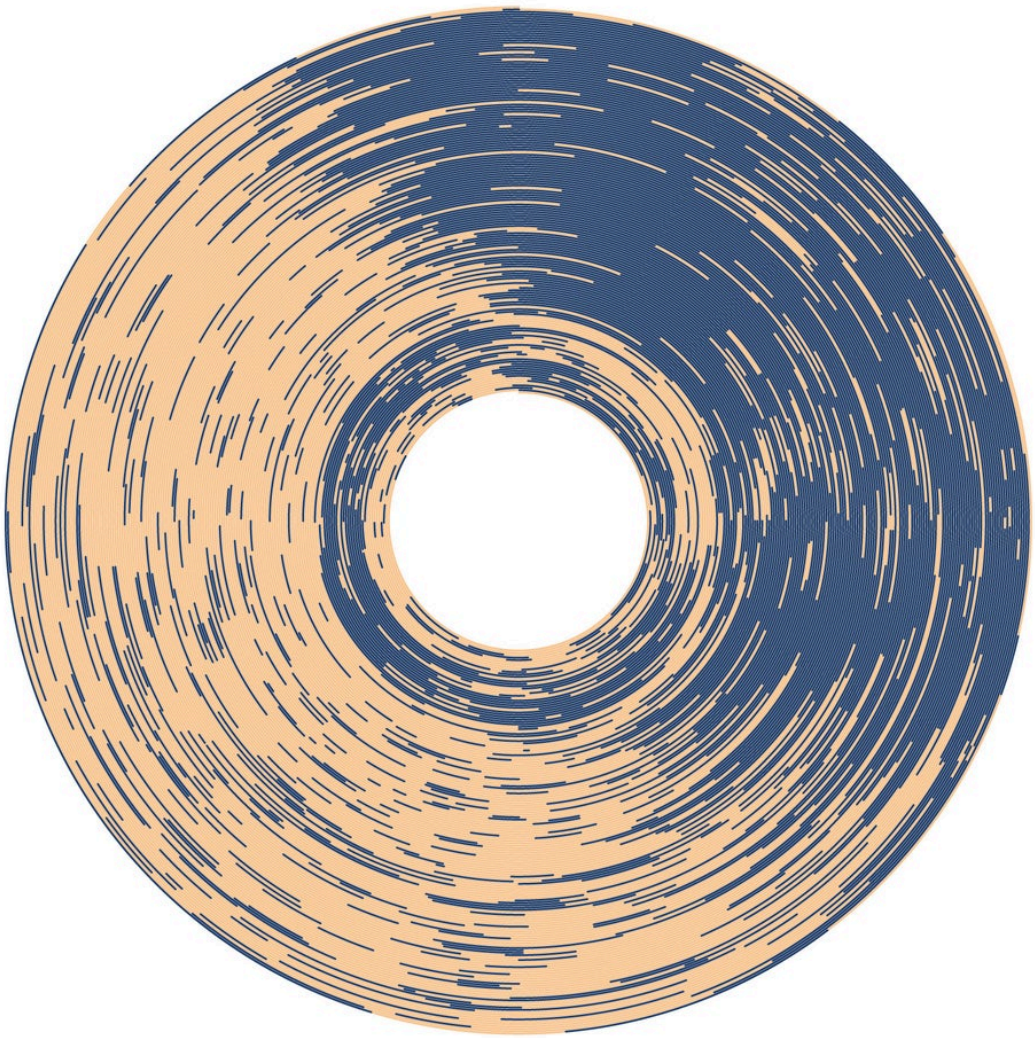

Figure 1: A data visualization from Redditor `andrew_elliott`

(2) loading, cleaning, organizing, and transforming data; (3) exploratory data analysis; and (4) data visualization and communication. Leaving this class, students will be able to immediately apply these skills to a broad array of interests. For example: students should be able to generate hypotheses from disparate data sources (such as minutes from ACM), obtain and transform data from websites (like accessing data from the U.S. Census Bureau), explore and analyze data to discover patterns (such as through spatial biological data sets), and visualize such data for easy communication to peers and laypersons. We will also intersperse throughout these topics various readings and discussions related to ethics and integrity in data science and analytics to help ground our work in a more holistic, human ecological point of view.

Classes will be taught as a mix of live coding exercises (bring your laptops!), lectures, and group discussions. No prior programming experience is required - we're going to be learning to use the R programming language in this course! - but a familiarity with computers and data will be helpful. Students will need to use either their personal laptop or a COA loaner laptop for class and programming exercises.

Course Objectives

This course has four (more or less) intertwined objectives. After completing the course, students will have both intellectual understanding and practical experience in the following domains:

- **Fundamentals of data wrangling**: Students will gain proficiency performing basic data importing, cleaning, and aggregating tasks using the `R` programming language.
- **Fundamentals of exploratory data analysis**: Plan, implement, and present a research project focusing on using exploratory data analysis techniques to discuss a research question.
- **Fundamentals of data visualization**: Create quality plots and data visualizations using `R` (generally) and the `tidyverse` (specifically).
- **Fundamentals of data literacy and ethics**: Students will practice reading, writing, and communicating data in context, as well as be able to identify the nature of data sources and constructs within an analysis. Students will also gain experience in highlighting the strengths and weaknesses of a given approach, as well tying analytical decisions back to the proposed hypothesis to check for rigor and bias. We will also pay special attention to the numerous ways in which ethical considerations can inform (or be ignored by) an a given analysis.

Course Schedule & Flow

You'll notice that this course changes "feels" several different times: the first few weeks will feel the most "computer sciency"-y, followed by a few weeks that will feel "statistics"-y, then a few weeks that will very much be reminiscent of a graphic design course, followed by a holistic capstone period. This is by design and reflects the inherent interdisciplinary nature of data science today.

- **Week 1**: Discuss syllabus, class focus, and get ourselves setup from a technical perspective.
- **Week 2**: Introduction to R - a crash course in computer science!, as well as practice obtaining, loading, and cleaning data.
- **Week 3**: Introduction to R - analyzing data sets
  - (**First Written Assignment Due: 2021-10-01**)
- **Week 4**: Some introductory statistics - understanding the "shape" of your data
- **Week 5**: Some more introductory statistics - testing hypothesis with your data.
  - (**Second Written Assignment Due: 2021-10-15**)
- **Week 6**: Data visualization - what does it mean to make a *good* visualization
- **Week 7**: Data visualization - conveying information with visualization
  - (**Third Written Assignment Due: 2021-10-29**)
- **Week 8**: Building a data science workflow (load, examine, visualize)
- **Week 9**: Working on our final projects, miscellaneous topics
  - (**Final Project Written Component Due: 2021-11-12**)
- **Week 10**: In-class presentations of final projects, course wrap-up
  - (**Final Project Presentations: 2021-11-16**)

Note: all dates and assignments subject to change.

Evaluation

There will be no quota of A's, B's, etc. You may take this class according to any grading structure you prefer (letter grades, pass/fail, etc.) - please feel free to reach out early so we can discuss the best plan for you.

Evaluation will be through class participation and discussion, several data investigation exercises, and a final project. The data investigations will take the form of written analyses of several well-known data sets as well as investigations of synthetic ones created specifically for the course. The final project will take the form of an oral presentation of an analysis. This can be either done in a group or as an individual and may be of any topic of sufficient interest to the student(s) involved.

In general, the breakdown of course credit will be as follows:

- **Class Participation**: 20%

- **Written Analyses**: 40%
- **Final Project**: 40%

Let's define what these mean!

Class participation

I define class participation as a balance between *presence*, *attention*, and *preparation*. Being *present*, in my perspective, is attending class and being actively engaged in the learning process. *Attention* can take a variety of different forms: some students may be more comfortable asking questions in class (either of me or of your peers), others may be more inclined to take notes and digest and absorb information on their own time, etc. *Preparation* is having completed the previous readings and assignments to the best of your ability and being ready to discuss any problems (or insights!) you may have had in doing so.

Written Analyses

Over the course of the term, we will be working on several written analyses of data sets where you'll be applying the skills you've learned to answer specific questions *or* hypotheses on what new kinds of questions could be asked. These assignments will be **submitted** individually, but you may work together in groups if you like so long as you make note of who all worked together on your submission.

Final Project

There will be a capstone project (either individual or group) focused on an in-depth analysis of a specific data set. We will discuss the final form of this project during the semester.

Diversity and Inclusion

It is my intent that students from all backgrounds and perspectives be well served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength, and benefit. It is my intent to present materials and activities that are respectful of diversity: gender, sexuality, disability, age, religion, socioeconomic status, ethnicity, race, and culture.

Learning about diverse perspectives and identities is an ongoing process. I am always looking to learn more about power and privilege and the harmful effects of racism, sexism, homophobia, classism, and other forms of discrimination and oppression. Your suggestions are encouraged and appreciated. Please let me know ways to improve the effectiveness of the course for you personally, or for other students or student groups. If something was said or done in class (by

anyone, including me) that made you feel uncomfortable, please talk to me about it. You may also reach out to the Provost or Associate Deans for further information or discussion.

Late Work

Try your best to turn in work when it is due. That said - the world is pretty crazy right now. If you're going to need some extra time, please just let me know as far in advance as possible and we'll find a way to work things out.

COVID-19 & Remote Instruction

The goal for this term is to be **in-person, together** as much as possible. With that said - it's still a pandemic, so who knows what might happen. If we're going to be switching to remote instruction at any point I will make sure to post in Google Classroom as well as directly e-mail all of you as far ahead of time as I can. If we need to be remote, we'll be using Zoom.

Textbook

You're already looking at it!

In all seriousness - one of the meta-goals of the Applied Data Science course series is to generate a textbook to meet the diverse needs of individuals interesting in using data science skills but who not have had exposure to the standardized math and computer science skills presupposed in introductory data science courses.

We will be using the following texts. Please note that **all** of these texts are available online for free and purchase is only necessary if you'd like to have your own physical copy for reference.

- Grolemund and Wickham - **R For Data Science**
- Peng and Matsui - **The Art of Data Science**.
- Barr, Ceinkaya-Rundel, Diez, and OpenIntro - **OpenIntro Statistics**
  - Note: you'll need to download this from Leanpub. You can set the pay scale to $0 to purchase this PDF for free - but you'll need to provide an e-mail.

Hardcopy Syllabus

If you would like to keep a copy of the syllabus, there is a download button above in the top toolboar that you can use to obtain one for your records. This may be particularly useful for students who plan to continue their graduate education after COA.

Standard Disclaimers

150 hours of academic engagement Our accreditation requires that we communicate that students should expect 150 hours of academic engagement for a one-credit COA course This total includes weekly meetings, fieldtrips, office hours, film screenings, readings and other assignments, service-learning, practicum, or other course requirements. You should therefore expect to spend a minimum of 150 academically engaged hours associated with this one-credit course. These 150 hours will be spent roughly as follows: 3 hr/wk "in" class, 4 hr/wk reading, 8 hr/wk on homework.

Plagiarism By enrolling in an academic institution, a student is subscribing to common standards of academic honesty. Any cheating, plagiarism, falsifying or fabricating of data is a breach of such standards. A student must make it his or her responsibility to not use words or works of others without proper acknowledgment. Plagiarism is unacceptable and evidence of such activity is reported to the academic dean or his/her designee. Two violations of academic integrity are grounds for dismissal from the college. Students should request in-class discussions of such questions when complex issues of ethical scholarship arise.

Library Resources Thorndike Library offers many resources and services that can assist you in your academic endeavors, including individualized research support and access to resources beyond COA. Study spaces are also available. The library is open 7 days/week. Remote access to the research databases is available 24/7. Contact library@coa.edu or visit the library website for details.

License

Part I

Getting Started

Chapter 1

Introductions

Goal for Today

- Introductions, course overview

Readings for Today

Required

- None!

Optional

- None!

Important Links and Files

- Lecture slides - these are also able to be found in the Google Classroom Drive.
- The class website - you should probably bookmark this!

Chapter 2

Gettting Started

Goal for Today

- Getting R and RStudio installed and functioning for everybody
- Discussing why we're using R vs. other programming languages (like Python)
- Some example analyses

Readings for Today

Required

- Peng, R Programming for Data Science - Chapter 2, History and Overview of R
- OpenIntro to Statistics, Sections 1.1 and 1.2
- Ziemann et al., Gene name errors are widespread in the scientific literature, *Genome Biology* (2016) 17:177 DOI 10.1186/s13059-016-1044-7

Optional

- Hugo Bowne-Anderson, What Data Scientists Really Do, According to 35 Data Scientists, *Harvard Business Review*. 15 Aug. 2018
- Knuth, Donald E., Computer programming as an art, *ACM Turing award lectures*. 2007.
- Thomas Herndon, Michael Ash and Robert Pollin, Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff, *Cambridge Journal of Economics* 2013, 1 of 23 (Sections 1 and 3)
- Cassidy, John, The Reinhart and Rogoff Controversy: A Summing Up, *New Yorker*, 2013 (Note: this may be useful background reading for the paper above)

Important Links and Files

- R: The R Project for Statistical Computing
- The Comprehensive R Archive Network (CRAN)
- R Studio
- Install R on Windows - YouTube Instructions
- Install R on Mac - YouTube Instructions

Part II

A Crash Course in R (pt. I)

Part III

A Crash Course in R (pt. II)

Part IV

Introducing: Statistics!

Part V

Even More Statistics!

Part VI

Data Visualization Basics

Part VII

Data Viz: Good, Bag, and Ugly

Part VIII

Building Workflows