DATA 200 Data Science Fluency

Linda E. Clark, Ph.D.
Spring 2020

Preferred Contact Email: linda_clark@brown.edu
Office Hours: TBA

Office Location: Room 325 164 Angell St.

This course will use Canvas
Class Hours T 9:00 – 10:20 (live, in person/remote); asynchronous
Class Room TBD

If you cannot make my office hours, please email me to arrange an alternative time

Course (Catalogue) Description

As data science becomes more visible, are you becoming more curious about its unique amalgamation of computer programming, statistics, and visualizing or storytelling? Have you ever wondered how these areas fit together and what a data scientist does? This course offers all students regardless of background the opportunity for hands-on data science experience, following a data science process from an initial research question, through data analysis, to the storytelling of the data. Along the way, you will learn about the ethical considerations of working with data through ethics spotting, and become more aware of societal impacts of data science.

Prerequisites

CSCI 0111 CSCI 0150 CSCI 0170 CSCI 0190
DATA 200 assumes foundational knowledge from CS 0111. Specifically, DATA 200 leverages students' skills in understanding the impacts of data organization on programming, the importance of understanding initial data collection, and the ability to work from a requirement to programming subtasks.
Students will use these skills in DATA 200 to manipulate aspects such as data aggregation, joins, filters and conduct exploratory data analysis.

Students enrolled in the CS Concentration will not receive concentration credit for this course

Required Textbook

Title: We will be using **Introduction to Data Science: *Data Analysis and Prediction Algorithms with R*** by *Rafael A. Irizarry*
This test is open source.
Despite the title, we will be using PYTHON, not R!

Cost of Textbook: This is an open source textbook, there is no cost.
Expenses and Financial Concerns

> Brown University undergraduates with concerns about the non-tuition cost(s) of a course at Brown, including this course, may apply to the Dean of the College Academic Emergency Fund to determine options for financing these costs, while ensuring their privacy. The fund can be found in the Emergency Funds, Curricular & Co-curricular Gap (E-Gap) Funds in UFunds. Information and procedures are available at this link: http://brown.edu/go/egap.

Other Electronic Resources/Readings

Throughout the course I will supplement our textbook with other readings. These will be provided to you by posting them electronically on CANVAS. If you need alternative formats of the materials, please contact me at the start of the course to arrange for alternative delivery methods to meet your needs.

The course uses google co-lab for programming in Python. There is no need for any additional software or resources.

Course Goals

Data is all around us, taking many forms. From numbers, text, and patterns, the challenge is no longer getting data, our challenge is managing, manipulating, and interpreting data so that we can continue to make informed decisions.

As opposed to being a fully formed professional field, data science is an emerging area, often associated with data driven business decision making. However, the practice of data science applies to just about any other area. Interested in analyzing the text of all prior episodes of "The Office" for specific phrases? Data science can help. What about looking for patterns within classical pieces art? Data science can help.

As data science becomes more visible, do you become more curious it's unique amalgamation of computer programming, statistics, and visualizing or storytelling? Have you ever wondered how these areas fit together and what a data scientist does? This course offers non-CS concentrators the opportunity for hands-on data science experience, following a data science process from an initial research question, through data analysis, to the storytelling of the data. Along the way, you will learn about the ethical considerations of working with data through ethics spotting, and become more aware of societal impacts of data science.

Learning Objectives:

1. Define data science as an emerging interdisciplinary field (programming, statistics, and visualization) to consider their own personal space within data science
2. Identify instances where ethical issues arise around data and apply strategies weighing alternatives to such instances to insure proper data governance and inclusive practices are followed
3. Create valid usable databases for future analysis by iteratively adjusting initial data import definitions based on data exploration
4. Use Python to successfully import, export, and manipulate data to connect to the technical requirements for cloud computing
5. Match appropriate statistical techniques to appropriate data types, ensuring conclusions from data are valid and understandable for public consumption and making informed decisions that can impact society moving forward.
6. Produce ongoing meaningful visual summaries of data to enable more in-depth understanding of the data for iterative analysis and to understand the phenomena for a public audience
7. Develop self-awareness of the impact of work on other parts of interdisciplinary processes to ensure clear communicate the requirements and products of their work.

Evaluation/Required Work (5 categories)

This course will be graded either A,B,C/No Credit or Satisfactory/No Credit (S/NC)

Initial self-assessment (graded as zero or 1 for submission only) 1%

Content: level of python knowledge; level of statistical experience; software tools (awareness use)

Weekly warm-ups (graded for completion) 15%

Most weeks students will be responsible for completing a short simple coding task before lecture.

Python Method Presentation 4%

Each student will be responsible for one brief video-taped presentation of a Python method.

Final Summative Project (45%)

Students will be divided into teams based on topical interest. Groups will find a dataset, query the dataset, generate a research question and complete the data science work as described below.

Scaffolded Activities

1. Identify a research question and data. Write initial python code to import and describe data;

   Products: Requirements for the project (research question and data source)
   Python Colab code for importing the data
   Written summary of the data (variables, observations) (aka data dictionary)

2. Data wrangling and exploratory data analysis.

   Python code ensuring the research question can be addressed with the final dataset (unit of analysis is appropriately aggregated, null values are appropriately handled, and any extreme values are identified.
   Conduct univariate and bivariate descriptive statistics, noting similarities and discrepancies.
   Products: Refined Python Code
   Univariate and bivariate summaries

3. Conduct the appropriate statistical technique to for the problem

   Identify the appropriate statistical technique and make any necessary data transformations. Identify and test any assumptions. Report and interpret findings from the technique
   Products: Python Code for the statistical analysis
   Rational for the statistical technique
   Test of the statistical assumptions
   Findings and interpretation of the statistical findings

4. Use data visualizations to summarize the findings and visualize the data to convey the story or making meaning from the data

   Products: Data visualization summarizing the most main conclusions from the analysis

Reflections (total of three (after activities 1,2, and 3)) (30%)

For each scaffolded activity, students will write a reflection (no more than 4 pages) including 3 sections (their own work and the products from the previous step and ethics spotting)

Section 1: Reflections on their own work:

What was the most significant challenge of this activity?

What knowledge would have been helpful before starting the step?

Section 2: (NOT REQUIRED FOR REFLECTION #1)

Reflection on how the previous activity impacted the current activity:

What aspect of the product from the previous activity helped the most in the current activity

What was missing or unclear in the previous activity that would have been helpful for the current activity?

Section 3: Ethics Spotting: Identify an actual or potential ethical issue encountered in working with the data?

Contributions to an Ethics in Data Science Blog. 5%Each student submits at least one entry

The Data Science Initiative will host an Ethics in Data Science Blog to promote ethical awareness in working with data. Students in Data Science 0200 will be required to contribute one actual or potential ethical issue. This issue can be copied from one of the reflection papers. Issues can be published anonymously but must be submitted to the course instructor with names.

Course Expectations

- It is my expectation that the students in this class and I will create a productive, caring learning environment. Part of creating this learning environment is being present in class and participating. Much of your learning will be facilitated by class discussions and activities. Multiple absences will impede your opportunities to learn this material and will adversely impact your performance on the required assignments. If you miss more than 1 class, you will need to see me to determine if you can complete the course satisfactorily.
- Because much of our class will involve discussions, I ask each student to follow basic rules to respect everyone in the room. This includes

  Basic conversational decorum (not interrupting, talking over, or side conversations)

  Respect for the opinions and views of each member of our class

  Owning your own opinions

  If at any point in the semester you feel uncomfortable in our classroom discussions, please see me so that we may work together to ensure everyone feels welcome and free to share their opinions and thoughts
- As a general rule, late assignments are not accepted. If, for any reason you anticipate being unable to meet a deadline, please see me in advance of the deadline to make other arrangements.
- If you will miss class due to a University sponsored event, religious holiday, or other absence according to University policy, please communicate with me in advance to ensure we can decide for you to complete any in-class activities.
- This class will have a mix of demonstration, discussion, and hands-on activities. Unless specifically specified, technology in class is to be avoided. Students will be notified in advance if technology will be required for any class meeting.

Semester Hours

The time estimate for this course is a total of 184 hours

Class Time	39 Hours
Reading (9 pg/hour)	20 Hours
Reflections (2hr/page)	48 Hours
Scaffolded Project	72 Hours
Blog Development and Post	5 Hours

Class Schedule

Unit	Topic	Assignment
1	Introduction to data science and the ethics of data science	Initial Assessment (first day of class attendance)
2	Planning and managing data science projects	
3	Python programming in data sciences, libraries and manipulating data	
4	Exploring data for data science	Scaffolded Activity 1
5	Inferential statistics (sampling, variability, randomness, probabilities)	Reflection 1
	Spring Recess (no classes)	
6	Hypothesis testing, prediction, and classification	Scaffolded Activity 2 Reflection 2
7	Overview of machine learning	
8	Overview of data visualizations with Tableau	Scaffolded Activity 3
9	Presenting data	Reflection 3
	Final Exam Period (No Final)	Scaffolded Activity 4

Accessibility and Accommodations Statement

Brown University is committed to full inclusion of all students. Students who, by nature of a documented disability, require academic accommodations should contact the professor during office hours. Students may also speak with Student and Employee Accessibility Services at 401-863-9588 to discuss the process for requesting accommodations.

Inclusion

This course is designed to support an inclusive learning environment where diverse perspectives are recognized, respected and seen as a source of strength. It is our intent to provide materials and activities that are respectful of various levels of diversity: mathematical background, previous computing skills, gender, sexuality, disability, age, socioeconomic status, ethnicity, race, and culture.

I would like to create a learning environment for my students that supports a diversity of thoughts, perspectives and experiences, and honors your identities (including race, gender, class, sexuality, religion, ability, etc.) To help accomplish this:

- If you have a name and/or set of pronouns that differ from those that appear in your official Brown records, please let me know!
- If you feel like your performance in the class is being impacted by your experiences outside of class, please don't hesitate to come and talk with me. I want to be a resource for you. If you prefer to speak with someone outside of the course, Dean Bhattacharyya, Associate Dean of the College for Diversity Programs, is an excellent resource.
- I (like many people) am still in the process of learning about diverse perspectives and identities. If something was said in class (by anyone) that made you feel uncomfortable, please talk to me about it.

Multilingual Learners

Brown University welcomes students from around the world, and the unique perspectives international students bring enrich the campus community. To empower students whose first language is not English, an array of ELL support is available on campus including language and culture workshops and individual appointments. For more information about English Language Learning at Brown, contact the ELL Specialists at ellwriting@brown.edu.

Academic Integrity

A student’s name on any exercise (e.g., a theme, report, notebook, performance, computer program, course paper, quiz, or examination) is regarded as assurance that the exercise is the result of the student’s own thoughts and study, stated in their own words, and produced without assistance, except as quotation marks, references, and footnotes acknowledge the use of printed sources or other outside help.

If you need assistance in defining or recognizing plagiarism, please see me and we can work to further your understanding of plagiarism and strategies to avoid plagiarizing.