updated 5/11/2022

LLO8200: Introduction to Data Science

Lacey Hartigan - Summer 2022
lacey.hartigan@vanderbilt.edu

Introduction

We have entered a time in which vast amounts of data are more widely available than ever before. At the same time, a new set of tools has been developed to analyze this data and provide decision makers with information to help them accomplish their goals. Those who engage with data and interpret it for organizational leaders have taken to calling themselves data scientists, and their craft data science. Other terms that have come into vogue are *big data*, *predictive analytics*, and *data mining*. These can seem to be mysterious domains. The point of this class is to demystify much of this endeavor for individuals who will be organizational leaders.

The class is structured around developing students' skills in three areas: obtaining and organizing data, analyzing data to make predictions that answer key research questions, and presenting the results of analyses. For each area, the subtopics are as follows:

Obtaining and organizing data

- Tools of the trade: R and RStudio
- Working with pre-processed data and flat files
- Tidying/cleaning data
- Getting data from the web: webscraping, using application programming interfaces (APIs)

Analyzing Data Topics

- Descriptives and conditional means
- Regression
- Supervised learning: classification
- Cross validation

Presenting Data Analyses Topics

- Descriptives: histograms, density plots, bar plots, dot plots
- Scatterplots
- Lattice graphics and small multiples
- Interactive graphics
- Communicating results effectively

Evaluation

Students are evaluated based on three areas: problem sets, a group project that culminates in a presentation and a paper, and class attendance/participation.

- 45% - Problem sets: Students are assigned a problem set to complete approximately every other week (see the syllabus and course website for due dates). Problem sets are due *any time before* the live session meets for that week.
    - Each problem set is worth 100 points. While each problem set question has a right answer, you will earn partial credit for most serious attempts. Additionally, you can submit corrections.
        - Corrections must be submitted within one week of receiving feedback.
        - They must be completed alone (no group work is allowed for corrections).
        - Corrected work will be awarded half of the points originally deducted (partial credit will be awarded when appropriate as well).
    - **All Problem Set Submissions must be in "knit" format: pdf or html. You must upload two files to receive full credit.**
        1. Your .Rmd code file,
        2. One "knit" document (pdf preferred; html accepted – WORD docs NOT accepted).
    - Submissions missing one of the required files will immediately lose 20 points.
    - No late submissions will be accepted without prior approval.
- 45% - Final Project: During the semester, you will work on a group project focused on using predictive modeling to answer a research question of your choosing, utilizing your skills as a data analyst.
    - Each component of the final project is worth 100 points. The following components are required:
        1. Three progress reports (worth 30, 30, and 40 points),
        2. A presentation to be delivered in class (in PowerPoint or another preferred presentation platform),
        3. A "knit" paper produced using .Rmd (in pdf or html).
- 10% - Participation: Attendance at and active/constructive participation in synchronous live sessions. It is expected that you make every attempt to attend each session, have your camera on, and your attention focused on the topic at hand. Missing more than one class will result in lost participation points (unless you

have a doctor's note, deployment paperwork, religious exemption, etc.).

Texts

- Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. San Francisco, CA: O'Reilly Media, Inc.
    - This is available for free online. You may also choose to purchase a hard copy.
- Silver, N. (2012). *The signal and the noise: Why so many predictions fail—but some don't*. New York, NY: Penguin.
- Articles available through the Vanderbilt library system (see week-by-week schedule below).

Software

We will use only free, open-source software in this course. We will use R, an open-source data analytic platform for all analysis. R appears to be the most widely used data analysis software in data science. We will utilize RStudio as our integrated development environment (IDE) for R.

Course Webpage

All files (weekly .Rmd files, course syllabus, datafiles, etc.) will be maintained on the course webpage. Because we are always working to improve the course (and because code is not static, it is always evolving and improving), updates are housed on the course webpage (**NOT the LMS**). You are expected to check this webpage frequently.

https://lhartigan15.github.io/LLO8200_updated/

Honor Code Statement

All assignments for this class, including weekly problem sets and the final project, are to be conducted under the obligations set out in Vanderbilt's Honor Code, found in the student handbook: https://www.vanderbilt.edu/student_handbook/the-honor-system/. If you have any doubts about how the Honor Code applies to your work in this class, please ask me for clarification. Uncertainty about application of the Honor Code does not excuse a violation.

Problem sets. You may collaborate with other classmates on your problem sets; however, all code must be your own (i.e., you are not allowed to email each other code files). The only copy/pasted code in your files should be from class .Rmd files (async and live session) or from the internet. Copying/pasting other students' code verbatim is an honor code violation.

Final Project. You will work in groups of three-to-four people for the final project; however, I expect that every group member will make a meaningful contribution to the products. We will talk more about the final project in the first few weeks' class sessions.

Weekly Schedule

**Items listed under each given week are DUE before live session on that date*

Week 0: May 5th, 2022 (optional R bootcamp session*)

- Review .Rmd file: 00_Getting Started
    - If you did not attend bootcamp, you should review the .Rmd file and the video recording and be sure everything runs smoothly.

Week 1: May 11th, 2022

- LMS Module 1. Welcome to Data Science: Tools of the Trade
- Reading
    - Wickham:
        - Welcome: Introduction
        - Explore
            - Introduction
            - Workflow: basics
            - Workflow: projects
    - Silver, Chapter 1

Week 2: May 18th, 2022

- LMS Module 2. Getting Data: Flat Files and "Tidy" Data
- Assignment: Post on the LMS wall some potential areas of interest for the group project.
    - The project will consist of analyzing data to build a predictive model for a particular phenomenon.
    - Groups will consist of 3-5 people (no more, no less).
    - I encourage making connections based on substantive interests rather than friendships, but as long as you have a group, I won't push you on this.
- Reading
    - Wickham:
        - Wrangle
            - Data import
            - Tidy data
    - Silver, Chapter 2
- Additional Resources
    - http://www.cookbook-r.com/Data_input_and_output/

Week 3: May 25th, 2022

- Group project – In-class discussion of the project (including norms surrounding communication and expectations) and group breakouts (we will use this time to finalize groups – if someone doesn't have a group at this point, we will find that person a group).

    - I expect that every group (if not at the max number of 5 people) will be open to additional group members if needed.
- LMS Module 3. Getting Data: Web Sources
- Reading
    - Silver, Chapter 3
- Additional Resources
    - http://www.cookbook-r.com/Data_input_and_output/

Week 4: June 1st, 2022

- Assignment 1 due (LMS Modules 1, 2, & 3)
- LMS Module 4. Analyzing Data: Conditional Means (supplemental Ntiles code included in .Rmd but not in async videos)
- Reading
    - Wickham:
        - Explore: Data transformation
    - Silver, Chapter 4

Week 5: June 8th, 2022

- Group project – Progress Report 1 due
- LMS Module 5. Presenting Data: Descriptive Plots
- Reading
    - Wickham:
        - Explore
            - Data visualization
            - Data transformation
    - Silver, Chapter 5
- Additional Resources
    - http://www.cookbook-r.com/Graphs/Bar_and_line_graphs_(ggplot2)/
    - http://www.cookbook-r.com/Graphs/Plotting_distributions_(ggplot2)/

Week 6: June 15th, 2022

- Assignment 2 due (LMS Modules 4 & 5)
- LMS Module 6. Analyzing Data: Linear Regression (we will split this unit across 1.5 weeks)
- Reading
    - Wickham:
        - Model
            - Introduction
            - Model basics
            - Model building
    - Silver, Chapter 6

- Additional Resources
    - http://www.cookbook-r.com/Statistical_analysis/

Week 7: June 22nd, 2022

- (the rest of LMS Module 6) and then LMS Module 7. Presenting data: Scatterplots
- Reading
    - Wickham:
        - Model
            - Introduction
            - Model basics
            - Model building
        - Explore
            - Data visualization
    - Silver, Chapter 7
- Additional Resources
    - http://www.cookbook-r.com/Statistical_analysis/

Week 8: June 29th, 2022

- Assignment 3 due (LMS Modules 6 & 7)
- Final Project – in-class group work – we will be using breakout rooms to have group meetings this week. Come prepared to do group work and have a short check-in with Lacey.
- Reading
    - Silver, Chapter 8

Week 9: July 6th, 2022

- Group project – Progress Report 2 due
- LMS Module 8. Analyzing Data: Classification (part 1 – we are splitting this module across two weeks)
- Reading
    - Article: Althoff, T., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2014). How to ask for a favor: A case study on the success of altruistic requests. In ICWSM. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8106/8101
    - Wickham
        - Model
            - Introduction
            - Model basics
            - Model building
    - Silver, Chapter 9

Week 10: July 13th, 2022

- LMS Module 8. Analyzing Data: Classification (part 2)
- Reading
    - Article: Althoff, T., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2014). How to ask for a favor: A case study on the success of altruistic requests. In ICWSM. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8106/8101
    - Wickham
        - Model
            - Introduction
            - Model basics
            - Model building
    - Silver, Chapter 10

Week 11: July 20th, 2022

- Assignment 4 due (LMS Module 8)
- LMS Module 9. Presenting Data: Plots and Tables for Classification
- Reading
    - Silver, Chapter 11

Week 12: July 27th, 2022

- Group project – Progress Report 3 due
- LMS Module 10. Cross Validation
- In-class code review and group project work (time permitting).
- Reading
    - Wickham
        - Model
            - Many Models
    - Silver, Chapter 12

Week 13: August 3rd, 2022

- GROUP PROJECT (DRAFT) PRESENTATIONS – Week 1

Week 14: August 10th, 2022

- Assignment 5 due (LMS Modules 9 & 10)
- GROUP PROJECT (DRAFT) PRESENTATIONS – Week 2

FINAL PRESENTATIONS and FINAL REPORTS DUE: Friday, August 12th, 2022 (by midnight, Pacific time)

Classroom Accommodations

Vanderbilt is committed to equal opportunity for students with disabilities. If you need course

accommodations due to a disability, please contact VU Student Access Services to initiate the process. After SAS has notified me of relevant accommodations, we will discuss how these accommodations may best be approached in this class, and I will facilitate the accommodations.

Mental Health & Wellness

If you are experiencing undue stress that may be interfering with your ability to perform academically, Vanderbilt's Student Care Network offers a range of support services. The Office of Student Care Coordination (OSCC) is the central and first point of contact to help you navigate and connect to appropriate resources. You can schedule an appointment with the OSCC at https://www.vanderbilt.edu/carecoordination/ or call 615-343-WELL. You can find a calendar of services at https://www.vanderbilt.edu/studentcarenetwork/satellite-services/.

If you or someone you know needs to speak with a professional counselor immediately, the University Counseling Center offers Urgent Care Counseling. Students should call the UCC at (615) 322-2571 during office hours to speak with an urgent care clinician. You can also reach an on-call counselor after hours or on the weekends by calling (615) 322-2571 and pressing option 2 at any time. You can find additional information at https://www.vanderbilt.edu//ucc/.

Mandatory Reporter Obligations

All University faculty and administrators are mandatory reporters. What this means is that all faculty, including me, must report allegations of sexual misconduct and intimate partner violence to the Title IX Coordinator. In addition, all faculty are obligated to report any allegations of discrimination. I am willing to discuss with you such incidents but can only do so in the context of us both understanding my reporting obligations. If you want to talk with someone in confidence, officials in the Student Health Center, the University Counseling Center, and the Office of the Chaplain and Religious Life (when acting as clergy) can maintain confidentiality. In addition, officials in the Project Safe Center have limited confidentiality, in that they must report the incidents but can do so without providing identifying information. The Project Safe Center serves as the central resource for those impacted by sexual misconduct and intimate partner violence and can assist with navigating all facets of the University's resource and support network and other processes.

Names and Pronouns

If you would like to use a different name or pronouns than those provided through YES, please let me know at any time prior to or during the semester. Information is available through the LGBTQI Life offices about how to change either or both of these in YES.