LLO8200: Introduction to Data Science

Lacey Hartigan - Fall 2021
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
- Using databases

Analyzing Data Topics

- Descriptives and conditional means
- Regression
- Supervised learning: classification
- Unsupervised learning: *K*-means and nearest neighbors clustering
- 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 will be evaluated based on three areas: weekly problem sets, a group project that culminates in a presentation and paper, and class attendance/participation.

- 45% - Problem sets: Students will be assigned a problem set to complete 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 all serious attempts. Additionally, you can submit corrections within one week of receiving feedback and earn back up to half of the points originally deducted.
    - **All Problem Set Submissions must be in "knitted" format: html or pdf. You must upload two files to receive full credit.**
        1. Your .Rmd code file
        2. One "knit" document (in the format of your choosing).
    - Each student is allowed one late submission (submitted within one week of the original due date) without penalty per term. After that, each late submission is docked -10 points for the first day and -20 points for every day thereafter.
- 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 final presentation to be delivered in class (in PowerPoint or another preferred presentation platform),
        3. A final paper produced using .Rmd (and knit to either html or pdf).
- 10% - Participation: Attendance at and active 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.

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.

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 will be housed on the course webpage. You are expected to check it frequently.

https://lhartigan15.github.io/LLO8200_fall2021/

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.

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 considered 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.

If you have any questions at all about the Honor Code or how it will be applied, ask me right away.

Schedule

Week 1: August 25, 2021

- 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: September 1, 2021

- LMS Module 2. Analyzing Data: Conditional Means
    - Supplemental Ntiles code (included in .Rmd on course website)
- Reading
    - Wickham:
        - Explore: Data transformation
    - Silver, Chapter 2

Week 3: September 8, 2021

- LMS Module 3. Presenting Data: Descriptive Plots
- Reading
    - Wickham:
        - Explore
            - Data visualization
            - Data transformation
    - Silver, Chapter 3
- Additional Resources
    - http://www.cookbook-r.com/Graphs/Bar_and_line_graphs_(ggplot2)/
    - http://www.cookbook-r.com/Graphs/Plotting_distributions_(ggplot2)/

Week 4: September 15, 2021

- Group project – In-class discussion of the project and group breakouts
- LMS Module 4: Getting Data: Flat Files and "Tidy" Data
- Reading
    - Wickham:
        - Wrangle
            - Data import
            - Tidy data
    - Silver, Chapter 4
- Additional Resources
    - http://www.cookbook-r.com/Data_input_and_output/

Week 5: September 22, 2021

- LMS Module 5: Analyzing Data: Linear Regression
- Reading
    - Wickham:
        - Model
            - Introduction
            - Model basics

            - Model building
        - Silver, Chapter 5
- Additional Resources
    - http://www.cookbook-r.com/Statistical_analysis/

Week 6: September 29, 2021

- Group project – Progress Report 1
- (continuation of) LMS Module 5: Analyzing Data: Linear Regression and LMS Module 6: Scatterplots
- Reading
    - Wickham:
        - Model
            - Introduction
            - Model basics
            - Model building
        - Explore
            - Data visualization
    - Silver, Chapter 6
- Additional Resources
    - http://www.cookbook-r.com/Statistical_analysis/

Week 7: October 6, 2021

- LMS Module 7: Webscraping
    - Optional additional code: Twitter API (saved to course website)
- Reading
    - Silver, Chapter 7
- Additional Resources
    - http://www.cookbook-r.com/Statistical_analysis/

Week 8: October 13, 2021

- 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: October 20, 2021

- Group project – Progress Report 2 due
- LMS Module 8: Analyzing Data: Classification (part 1)
- 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 (if you did not read this previously, now would be a good time to do so)
        - Model
            - Introduction
            - Model basics
            - Model building
    - Silver, Chapter 9

Week 10: October 27, 2021

- LMS Module 8: Analyzing Data: Classification (part 2)
- Reading (same as last week)
    - 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 (if you did not read this previously, now would be a good time to do so)
        - Model
            - Introduction
            - Model basics
            - Model building
    - Silver, Chapter 10

Week 11: November 3, 2021

- Group project – Progress Report 3 due
- LMS Module 9: Presenting Data: Plots and Tables for Classification
- Reading
    - Silver, Chapter 11

Week 12: November 10, 2021

- LMS Module 10: Cross Validation
- In-class code review and group project work.
- Reading
    - Wickham
        - Model
            - Many Models
    - Silver, Chapter 12

Week 13: December 1, 2021

- GROUP PROJECT PRESENTATIONS (3 to 4 groups will present)

Week 14: December 8, 2021

- GROUP PROJECT PRESENTATIONS (3 to 4 groups will present)

FINAL PRESENTATIONS and FINAL REPORTS DUE: TBD, 2021 (by midnight, Pacific time)