LLO8200: Introduction to Data Science

Lacey Hartigan - Spring 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 (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 will be 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 will be 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 within one week of receiving feedback and earn back up to half of the points deducted.
    - **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).
    - Submissions missing one of the required files will immediately lose 20 points.
    - No late submissions 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, 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 will be 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.

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 0: January 6th, 2022 (optional R bootcamp session*)

- Review .Rmd file: 00_Getting Started
  - Go ahead and try to install the packages listed in this file before we meet for bootcamp. Come prepared with questions.

        - *If you choose not to attend bootcamp, you should review this file on your own and be sure everything runs/installs properly (I will record and post the bootcamp session if you have a scheduling conflict).

Week 1: January 19th, 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: January 26th, 2022

- LMS Module 4: Getting Data: Flat Files and "Tidy" Data
- Reading
    - Wickham:
        - Wrangle
            - Data import
            - Tidy data
    - Silver, Chapter 2
- Additional Resources
    - http://www.cookbook-r.com/Data_input_and_output/

Week 3: February 2nd, 2022

- Group project – In-class discussion of the project and group breakouts (we will use this time to help identify common interests and potential groupings)
- LMS Module 7: Webscraping
- Reading
    - Silver, Chapter 3
- Additional Resources
    - http://www.cookbook-r.com/Data_input_and_output/

Week 4: February 9th, 2022

- Assignment 1 due (LMS Modules 1, 4, & 7)
- LMS Module 2. 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: February 16th, 2022

- Group project – Progress Report 1 due
- LMS Module 3. 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: February 23rd, 2022

- Assignment 2 due (LMS Modules 2 & 3)
- LMS Module 5: Analyzing Data: Linear Regression
- Reading
    - Wickham:
        - Model
            - Introduction
            - Model basics
            - Model building
    - Silver, Chapter 6
- Additional Resources
    - http://www.cookbook-r.com/Statistical_analysis/

Week 7: March 2nd, 2022

- LMS Module 6: Scatterplots
- Reading
    - Wickham:
        - Model
            - Introduction
            - Model basics
            - Model building
        - Explore
            - Data visualization
    - Silver, Chapter 7

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

March 9th, 2022 – Spring Break (no class)

Week 8: March 16th, 2022

- Assignment 3 due (LMS Modules 5 & 6)
- 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: March 23rd, 2022

- 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
        - Model
            - Introduction
            - Model basics
            - Model building
    - Silver, Chapter 9

Week 10: March 30th, 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: April 6th, 2022

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

Week 12: April 13th, 2022

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

Week 13: April 20th, 2022

- GROUP PROJECT (DRAFT) PRESENTATIONS – Week 1 (3 to 4 groups will present)

Week 14: April 27th, 2022

- Assignment 5 due (LMS Modules 9 & 10)
- GROUP PROJECT (DRAFT) PRESENTATIONS – Week 2 (3 to 4 groups will present)

FINAL PRESENTATIONS and FINAL REPORTS DUE: Wednesday, May 4th, 2022 (by midnight, Pacific time)