LLO8200: Introduction to Data Science

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: getting data, analyzing data to make predictions, and presenting the results of analysis. For each area, the subtopics are as follows:

Getting Data Topics

- Tools of the trade: R and RStudio
- Working with pre-processed data and flat files
- Getting data from the web: webscraping, using forms, 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 Analysis 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 two areas: weekly problem sets and the final project.

- 65% - Problem sets: Each week students will be assigned a problem set to complete. The problem sets will be due 24 hours prior to the following week's live session. For example, the Week 1 problem set will be due 24 hours prior to the Week 2 live session. No late problem sets will be accepted without prior permission.
    - Each problem set is worth 100 points.
    - **All Problem Set Submissions must be in "knitted" format: html, doc, or pdf. You will upload two files**
        1. Your .Rmd code file
        2. One "knit" document (in the format of your choosing).
    - Note that your grade on problem sets does not depend entirely on being correct on all problems but making a serious attempt to answer all problems (you will earn partial credit for all serious attempts).
- 35% - Final Project: During the course of the semester, you will work on a final assignment utilizing your skills as a data analyst.
    - Progress reports 17.5%: 100 points each
    - Final Product 17.5%: 100 points

Texts

We will have two texts for the course.

- Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. San Francisco, CA: O'Reilly Media, Inc.
- 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_spring2021/

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 a maximum of three 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 may work in groups (maximum of four people to a group) 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: January 12, 2021

- LMS Module 1. Welcome to Data Science: Tools of the Trade
- Reading
    - Wickham:
        - Welcome: Introduction
        - Explore
            - Introduction
            - Workflow: basics
            - Workflow: projects
    - Silver, Chapters 1–4

Week 2: January 19, 2021

- LMS Module 2. Analyzing Data: Conditional Means
- Reading
    - Wickham:
        - Explore: Data transformation
    - Silver, Chapters 5–9
- Assignment 1 due (upload BEFORE class begins)

Week 3: January 26, 2021

- LMS Module 3. Presenting Data: Descriptive Plots
- Reading
    - Wickham:

        - Explore
            - Data visualization
            - Data transformation
- Additional Resources
    - http://www.cookbook-r.com/Graphs/Bar_and_line_graphs_(ggplot2)/
    - http://www.cookbook-r.com/Graphs/Plotting_distributions_(ggplot2)/
- Assignment 2 due (upload BEFORE class begins)

Week 4: February 2, 2021

- (continuation of) LMS Module 3. Presenting Data: Descriptive Plots
    - Supplemental Ntiles code (uploaded to course website)
- Reading
    - Wickham:
        - Explore
            - Data visualization
            - Data transformation
- Additional Resources
    - http://www.cookbook-r.com/Graphs/Bar_and_line_graphs_(ggplot2)/
    - http://www.cookbook-r.com/Graphs/Plotting_distributions_(ggplot2)/

Week 5: February 9, 2021

- LMS Module 4: Getting Data: Flat Files and "Tidy" Data
- Reading
    - Wickham:
        - Wrangle
            - Data import
            - Tidy data
    - Silver, Chapter 10 (catch up on previous chapters!)
- Additional Resources
    - http://www.cookbook-r.com/Data_input_and_output/
- Assignment 3 due (upload BEFORE class begins)

Week 6: February 16, 2021 **** CLASS CANCELED DUE TO WINTER STORM URI ****

Week 7: February 23, 2021

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

            - Model basics
            - Model building
        - Silver, Chapter 11 (catch up on previous chapters!)
- Additional Resources
    - http://www.cookbook-r.com/Statistical_analysis/
- GROUP PROJECT: Progress Report 1 - Research question(s) and potential dataset(s)
    - Please have one group representative email this to Lacey.
- Assignment 4 due (upload BEFORE class begins)

Week 8: March 2, 2021

- (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 12 (catch up on previous chapters!)
- Additional Resources
    - http://www.cookbook-r.com/Statistical_analysis/

Week 9: March 9, 2021

- LMS Module 7: Webscraping
    - Optional additional code: Twitter API (saved to course website)
- Reading
    - Silver (catch up on previous chapters!)
- 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.
- Assignment 5 due (upload BEFORE class begins)

Week 10: March 16, 2021

- LMS Module 8: Analyzing Data: Classification
- 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

Week 11: March 23, 2021

- LMS Module 9: Presenting Data: Plots and Tables for Classification
- 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
- Group Project – progress report due (one group member, please email to Lacey before class begins)

Week 12: March 30, 2021

- LMS Module 10: Cross Validation
- Reading
    - Wickham
        - Model
            - Many Models
- Assignment 6 due (upload BEFORE class begins)

Week 13: April 6, 2021

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

Week 14: April 13, 2021

- GROUP PROJECT PRESENTATIONS (3 to 4 groups will present)
- Assignment 7 due (upload BEFORE class begins)

FINAL PRESENTATIONS and FINAL REPORTS DUE: April 21, 2021 (by midnight, Pacific time)