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. Each problem set will be graded on a 100-point scale.
    - 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: your .Rmd code file and one "knit" document (in the format of your choosing).**
    - Note that your grade on problem sets does not depend on your being correct on all problems but making a serious attempt to answer all problems.
- 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

- LMS Module 4: Getting Data: Flat Files and "Tidy" Data
- Reading
    - Wickham:
        - Wrangle
            - Data import
            - Tidy data
- Assignment 3 due (upload BEFORE class begins)