Revised: September 8, 2021

- Addition of "group" option to the final project
- Minor grammatical corrections.
- Reference to GitHub site for schedule

LLO8200: Introduction to Data Science

Rafael Garcia – Fall 2021
rafael.garcia@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 two areas: weekly problem sets and a 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 3 live session.
    - 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. Your lowest grade will be dropped.
    - **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).
    - The grading standards will be as follows:
        1. 50 = turned in problem set, did not attempt most of the problems
        2. 75 = turned in problem set, attempted most of the problems
        3. 100 = turned in problem set, attempted all of the problems
- 35% - Final Project: During the semester, you will work on a final 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. Two components are required:
        1. Four progress reports (17.5% of final grade)
        2. A final paper produced using .Rmd (and knit to either html or pdf) (17.5% of final grade).
    - The final project can be worked on in small groups (please notify the instructor prior to the first Progress Report due date), but each student will be required to submit an individual report. The analyses with the group will, likely, be identical, but the interpretations and explanations of your results need to be done individually.

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.

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 on the final project. 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.

Course schedule can be found on the course website in the Organizational Materials section