Introduction to Data Science

Vanderbilt University

Human and Organizational Development

Course Number HOD 3200

Fall 2023

Mark J. Chin

Office Hours: On Zoom, Tuesdays 11-1 (Calendly Link) or by appointment

Email: mark.chin@vanderbilt.edu

Twitter: @MemeMedianMode

Course github: Link

Introduction

Note: The materials for this course have been heavily adapted from Will Doyle's HOD 3200 Course at Vanderbilt University

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, git and Github
- 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
- Evaluating multiple models/Cross Validation

Presenting Data Analysis Topics

- Descriptives: histograms, density plots, bar plots, dot plots
- Scatterplots
- Plots for Classification
- Interactive Graphics

Evaluation

Students will be evaluated based in two areas: weekly assignments and the final project.

- Problem sets 65%: Almost every week I will assign a problem set for students to complete focused on practicing the topic of the week. These problem sets will generally be due one week after introduction of the topic, at 11:59:59 pm. So if we introduce a new topic on Tuesday, the assignment will be due Monday 11:59:59 pm the following week. Thursday introductions mean Wednesday deadlines. No late assignments will be accepted. Each assignment will be graded on a 100 point scale. Your lowest grade will be dropped.

- Final Project 35%: During the course of the semester you will work **with a partner** on a final assignment utilizing your skills as a data analyst. We will discuss this assignment and my expectations in detail during the course of the semester. There will be four progress reports due for the final project, each of which will be worth 12.5% of the final grade for the project. No late progress reports will be accepted. The final product will account for the remaining 50%. No late final products will be accepted.

Texts

Required Texts

We will have two texts for the course. The first is Hadley Wickham's book, R for Data Science. Wickham is generously making this book available for free. However, I strongly encourage you to buy this book from O'Reilly.

Amazon

The other text is Nate Silver's *Signal and the Noise*.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't.* New York: Penguin.

Amazon

Your local bookseller

Reserve

I've placed three books by Edward Tufte on reserve for you. These are masterpieces in the area of visualizing quantitative information. You should take a look at these for ideas and inspiration–I've noted the sections that are most helpful in various parts of the syllabus.

Tufte, E. R. (1990). Envisioning information. Cheshire, CT: Graphics Press.

Tufte, E. R. (1997). Visual explanations. Cheshire, CT: Graphics press.

Tufte, E. R. (2001). The visual display of quantitative information (2nd Edition). Cheshire, CT: Graphics press.

Lecture Notes

My lecture notes include both code and notes for the week. They will be available in github.

Web Resources

When appropriate for each week, web resources are linked directly from the syllabus. You will find a wealth of resources online, including other versions of this class offered as Massive Online Open Courses. I encourage you to take full advantage of the wealth of online materials that are available. Stack Overflow is your friend, but search carefully for your question. It is VERY likely that your question has already been asked.

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.

We will also use git, a distributed version control program, and Github, an online hosting platform. Github Desktop will serve as our Graphical User Interface to git and GitHub. RStudio is fully integrated with git and Github, making it an ideal IDE for these purposes. Up to date class assignments will be distributed through GitHub and will be collected and graded through Brightspace.

Communication

You can book my office hours at: https://calendly.com/mark-chin/hod3200-office-hours If my office hours don't work for you, please make an appointment over email. Student communications, including emails are my priority. However, due to the volume of email I receive, I may miss your message. To help with this problem, please place the phrase "HOD 3200' ' in your subject line. I will search for these messages every time I access my email. You can also use Brightspace's email function, which will automatically do this for you.

The Tedious Stuff

This class will be impossible if you don't show up. It's reasonable to contact me if for some reason you can't make it for a given class session.

We must use laptops in this class, despite the substantial body of evidence that laptop use in classrooms hinders learning. To mitigate this problem, the following standards will ALWAYS apply in class. You may have RStudio open. You may also have a web browser open to a web page that is relevant to course content. You MUST turn off all notifications and messaging programs. If your web browser is open to Facebook, Instagram or other purely social sites, or if you are responding to messaging apps, I will ask you to leave class for the day. If you're joining us remotely the same rules apply.

Mobile phone use is never appropriate in class. I will ask you to leave if you are using your mobile phone at any time. Exceptions are to be arranged BEFORE class, not when I observe you using your mobile phone.

Honor Code Statement

All assignments for this class, including weekly assignments and the final project, are to be conducted under the obligations set out in Vanderbilt's Honor Code. Please click here to review the honor code.

There will be two quite different standards for completing the assignments and the final project.

Assignments You may collaborate with anyone and you may utilize any resource you wish to complete these assignments.

Final Project All of the work on the final assignment must be the work of you and your partner alone. Anyone's work that you reference should be cited, as usual. All data that you do not personally collect must be cited, as with any other resource.

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

Schedule

Thursday, August 24, Introduction

Tuesday, August 29, Getting Data: Tools of the Trade

Resources

Wickham: Introduction, Explore: Introduction, Workflow: basics, Workflow: projects

Silver, Chapters 1-4

R Intro and Resources

Download R

Download Rstudio You want the "Desktop" version, free license

Download git

Download GitHub Desktop

Github Intro and Resources

Lecture Notes

Introduction: `01-intro.Rmd`.

Thursday, August 31, Tools of the Trade, continued

Subtopics: "verbs" of data wrangling, file types, working with git and GitHub.

Lab Practical

R Basics, "verbs" of data wrangling

Tuesday, September 5, Analyzing Data: Conditional Means

Resources

Wickham: Data transformation

Silver, Chapters 5-9, 12-13

Lecture Notes

Conditional Means: `02-conditional_means.Rmd`.

Assignments

Assignment 1 Due 11:59 PM, Monday, September 4

Thursday, September 7, Conditional Means, continued

Standing Meetings

Lab Practical

Conditional Means

Tuesday, September 12, Presenting Data: Descriptives

Subtopics: bar plot, density plot, dot plots, histograms

Resources

Wickham: Data visualization, Exploratory Data Analysis

Cookbook for R:Bar and Line Graphs

Cookbook for R: Plotting Distributions

Lecture Notes

Plotting Distributions and Conditional Means: `03-plot_means.Rmd`.

Assignments

Assignment 2 Due 11:59 PM, Monday, September 11

Thursday, September 14, Descriptives, continued

Standing Meeting

Lab Practical

Presenting results in graphical format: barplots, density plots, dot plots, histograms

Assignments

Progress Report 1 Due before class

Tuesday, September 19, Getting Data: Flat Files, Basic Concepts of "Tidy Data"

Resources

Wickham: Data import, Tidy data

Lecture Notes

Flat Data: `04-flat_data.Rmd`

Assignments

Assignment 3 Due 11:59 PM, Monday, September 18

Thursday, September 21, Flat Files, Basic Concepts of “Tidy Data”, continued

Standing Meeting

Lab Practical

Working with various data formats

Tuesday, September 26, Analyzing Data: Linear Regression

Resources

Wickham: Model: Introduction, Model Basics, Model Building

Lecture Notes

Linear Regression: `05-regression.Rmd`.

Thursday, September 28, Linear Regression, continued

Subtopics: training and testing models

Assignments

Assignment 4 Due 11:59 PM, Wednesday, September 27

Tuesday, October 3, Linear Regression, continued

Standing Meeting

Lab Practical

Kaggle competition

Assignments

Progress Report 2 Due before class

Thursday, October 5, Presenting Data: Scatterplots

Resources

Wickham: Data Visualization, Graphics for Communication

Tufte, Visual Display chapters 4 and 5.

Tufte, Envisioning Information, chapter 2

Assignments

Assignment 5 Due 11:59 PM, Wednesday, October 4

Lecture Notes

Scatterplots: `06-scatterplots.Rmd`.

Tuesday, October 10, Scatterplots, continued

Standing Meetings

Lab Practical

Presenting Data via Scatterplots

Thursday, October 12, Getting Data: Scraping Data from the Web, APIs

Resources

Rvest Vignette: https://cran.r-project.org/web/packages/rvest/vignettes/rvest.html

Lecture Notes

Web Scraping and APIs: `07-webscraping.Rmd`.

Assignments

Assignment 6 Due 11:59 PM, Wednesday, October 11

Tuesday, October 17, Web Data, continued

Standing Meetings

Lab Practical

Web Scraping and APIs

Thursday, October 19, No Class (Fall Break)

Tuesday, October 24, Analyzing Data: Classification

Resources

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An introduction to statistical learning. 2nd Edition with R. Chapter 4, Chapter 4 Lab R Code

Althoff, T., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2014, May). How to Ask for a Favor: A Case Study on the Success of Altruistic Requests. In ICWSM. (Available Here)

$$< https://arxiv.org/abs/1405.3282 >$$

Lecture Notes

Classification: `08-classification.Rmd`.

Assignments

Assignment 7 Due 11:59 PM, Monday, October 23

Thursday, October 26, Classification, continued

Standing Meetings

Lab Practical

Classifying behavior via text analysis: random acts of pizza.

Tuesday, October 31, Presenting Data: Plots for Classification

Resources

Lecture Notes

Plots for Classification: `09-plots_classification.Rmd`.

Assignments

Assignment 8 Due 11:59 PM, Monday, October 30

Thursday, November 2, Plots for Classification, continued

Standing Meetings

Lab Practical

Plots for understanding classification

Tuesday, November 7, Analyzing Data: Cross Validation

Resources

Wickham: Many Models

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An introduction to statistical learning. 2nd Edition with R. Chapter 5

Lecture Notes

Cross Validation: `10-cross_validation.Rmd`.

Assignments

Progress Report 3 Due before class

Thursday, November 9, Cross Validation, continued

Standing Meeting

Lab Practical

Cross Validation

Tuesday, November 14, Analyzing Data: Unsupervised Learning

Resources

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An introduction to statistical learning. 2nd Edition with R. Chapter 12, Chapter 12 Lab R Code

Lecture Notes

Unsupervised Learning: `12-unsupervised.Rmd`.

Assignments

Assignment 9 Due 11:59 PM, Monday, November 12

Thursday, November 16, Unsupervised Learning, continued

Standing Meeting

Lab Practical

Unsupervised Learning

Tuesday, November 21, No Class (Thanksgiving)

Thursday, November 23, No Class (Thanksgiving)

Tuesday, November 28, Special Topics: Databases

Lecture Notes

Databases: `11-databases.Rmd`.

Assignments

Assignment 10 Due 11:59 PM, Monday, November 26

Thursday, November 30, Special Topics: Interactive Graphics

Lecture Notes

Interactive Graphics: `13-interactive_graphics.Rmd`.

Assignments

Progress Report 4 Due before class

Tuesday, December 5, Class Presentations

Group 1

Thursday, December 7, Class Presentations

Group 2

Final Projects Due Thursday, December 14, 11:59 PM CT