DATA 712: Advanced Analytics

Course Syllabus

Spring 2022
Joseph Cohen
Department of Sociology & Data Analytics
CUNY Queens College

Overview

This course will teach you to acquire, clean, describe, and model data for the purposes of decision-making support.

Learning Objectives. This class intends to train you in the following general classes of operations using the R platform. Over the course of this semester, you will be trained in the basics of:

- ***Using Analytics in Decision-Making Processes.*** Identify real world decisions that can benefit from data, and use data to improve decision-making.
- ***Data Acquisition and Wrangling.*** Getting data and preparing it for analysis.
- ***Descriptive Analysis.*** Describing how your metrics' distribution
- ***Visualization.*** Creating effective and attractive visual representations of data
- ***Regression Modeling.*** The bulk of this course focuses on methods to estimate linear functions that "explain" or "predict" events or outcomes using data.
- ***Multivariate Analysis.*** Methods to classify and group observations and variables.

These skills give you the foundations of skills necessary to apply advanced statistical analysis techniques in applied settings. They are relevant in jobs that involve data analysis, interpretation, and communication.

Prerequisites. You must have passed SOC/DATA 710 before starting this class.

Schedule

Seminars
Wednesdays, 6:30 PM – 8:15 PM
TBA

Office Hours
Wednesdays, 4:00 PM – 6:00 PM
PH 252 H

Note that there are no office hours in weeks in which there are no classes

Required Textbooks

- **Paul Teetor (2011) *R Cookbook.* O'Reilly.** Reference book.
- **Timothy Urdain (2010) *Statistics in Plain English.* Routledge.** Plain-language explanations.
- **Andrew Gelman, Jennifer Hill, and Aki Vehtari (2021) *Regression and Other Stories.* Cambridge**. Advanced applied statistics text.

Assignments

Assignment Information

- **DataCamp (20%).** *Complete assigned DataCamp modules. Student receive full credit for completing it on time, and partial credit for late submissions.*
- **Descriptives Assignment (5% x 6 = 30%).** Shoter assignments to practice skills.
- **Final Project (40%).** Final term paper with report.
- **Participation (10%).** Quality of contribution to class discussion and the success of the team.

Assessment

Grade Scale

Letter Grade	Raw Percentages to Letters	Number Equiv. of Letter	Degree of Implied Proficiency in / Grasp of Subject Matter
A+	97% - 100%	100%	Masterful
A	93% – 96.9%	95%	Excellent, Highly Professional
A-	90% - 92.9%	92%	Quality Early Career Submission
B+	87% - 89.9%	88%	Good Student Submission
B	83% - 86.9%	85%	Average Student Submission
B-	80% - 82.9%	82%	Acceptable Student Submission
C	70% - 79.9%	75%	Basic with Nontrivial Gaps
D	60% - 69.9%	60%	Serious Deficiencies
D-	50% - 59.9%	50%	Poor Submission
F	<60%	0%	Did Not Submit / Unacceptable

Schedule

Readings denoted with a (RES) are available on class reserves.

Week 1: Class Introduction

Feb 2. An introduction to the class

Topics

- ***Getting acquainted.*** Adjust to a new reality in which we study with other humans in the room
- ***Course overview.*** Course objectives. Requirements. Agenda. Assessment. Class policies.
- ***Study tips.*** Advice on succeeding in this class
- ***Equipment check.*** R, RStudio, LaTeX, DataCamp, Blackboard

Class Reading

- Syllabus
- Teetor: Chapter 1

Helpful Reads

- **Steps in a Data Analysis Job:** Will Hillier (2021) "A Step-by-Step Guide to the Data Analysis Process" Blog Post on *CareerFoundry.com* < https://careerfoundry.com/en/blog/data-analytics/the-data-analysis-process-step-by-step/>
- **An Introduction to R Markdown:** Garrett Grolemund (2014) "Introduction to R Markdown." July 16. Blog post. <https://rmarkdown.rstudio.com/articles_intro.html>
- **Free Online Textbook for R Markdown:** Yihui Xie, JJ Allaire, and Garrett Grolemund (2020) *R Markdown: The Definitive Guide.* Online textbook. <https://bookdown.org/yihui/rmarkdown/>
- **Markdown Cheat Sheet:** RStudio's R Markdown cheat sheet <https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>
- **To Enable Creation of PDF Reports on Windows Device:** Soren L. Kristiansen (2015) "Create PDF reports using R, R Markdown, LaTeX and knitr (on Windows 10)" Medium. October 28. <https://medium.com/@sorenlind/create-pdf-reports-using-r-r-markdown-latex-and-knitr-on-windows-10-952b0c48bfa9>

Homework

- Ensure that you have a working computer with working versions of R and R Studio
- Ensure that you have access to DataCamp and Blackboard
- Ensure that you can render PDF files from Markdown files
- Install Microsoft Teams and ensure that you are connected to the DATA 712 group

Week 2: The Task of Data Analysis

Feb 9. A look at the overall task of analyzing data

Topics

- **Science and Evidence-Based Decision-Making.** The underlying philosophies
- **What Do Data Analysts Do?** A description of the job. Coding and your long-term career plans.
- **Steps in an Analysis**: Steps to take in a data analysis job.
- **Importance of Communication**: Communication is at least as important as programming, communicating well, structure and style guide for written reports
- **Review of Basic R Operations**: Using Markdown, Working with Packages, Managing Memory, Setting Working Directories, Importing and Exporting Data, Working with Vectors and Data Frames, Basic Operations

Class Reading

- Rachel Schutt and Cathy O'Neill (2014) *Doing Data Science.* O'Reilly: Chapter 1 (ER)
- Urdain: Chapter 1
- Teetor: Chapter 2 - 4

Helpful Reads

- **Importing Data from Excel: *readxl*** package documentation <https://readxl.tidyverse.org/>
- **Simple Introduction to Hypothesis Testing:** Terence Shin (2020) "Hypothesis Testing Explained as Simply as Possible" *Towards Data Science.* February 28 <https://towardsdatascience.com/hypothesis-testing-explained-as-simply-as-possible-6e0a256293cf>
- **Engaging Introduction to P-Values**: Christie Ashwanden (2015) "Not Even Scientists Can Easily Explain P-Values" *FiveThirtyEight*, Nov 24 < https://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/>
- **Introduction to Environments in R:** "Environments" in Hadley Wickham (2019) *Advanced R, 2nd ed.* Chapman & Hall, Chapter 7 < http://adv-r.had.co.nz/Environments.html>
- **More on Working With Objects:** John Blischak, Daniel Chen, Harriet Dashnow, and Denis Haine (eds): "Software Carpentry: Programming with R." Web post. <https://swcarpentry.github.io/r-novice-inflammation/13-supp-data-structures/>

Homework

- Diagnostic Assignment Assigned
- DataCamp Course: *An Introduction to R* <https://learn.datacamp.com/courses/free-introduction-to-r>

Week 3: Describing Distributions

Feb 16. Describing data.

Topics

- **Describing Distributions**. What is a distribution? Data and cognitive processing. Data types.
- **Describing Univariate Distributions**. Describing continuous and discrete distributions.
- **Describing Bivariate Distributions**. Measuring association between two variables
- **Static Visualization using *ggplot2.*** A review of basic graphing operations using the ***ggplot2*** package.

Reading

- Urdain: Chs. 2 & 3

Homework

- DataCamp Course: Exploratory Data Analysis in R <https://learn.datacamp.com/courses/exploratory-data-analysis-in-r>
- DataCamp Module: Introduction to Data Visualization with ggplot2 < https://learn.datacamp.com/courses/introduction-to-data-visualization-with-ggplot2>
- Diagnostic Assignment Due
- Descriptives Assignment Assigned

Week 4: Data Wrangling

Feb 23. Acquiring and preparing data for analysis.

Topics

- **Importing and exporting data.** Bringing various forms of data into R, including APIs and HTML scraping.
- **Data cleaning and tidying.** Manipulating data into usable formats
- **Creating derivative variables and sets.** Creating new variables from your data.
- **Creating and applying functions and loops.** Tools for processing large amounts of data

Reading

- Teetor: Chapters 5, 6 & 7.1 – 7.7

Helpful Reads

- **What is Tidy Data?** Hadley Wickham "Tidy Data" *Journal of Statistical Software*
- **Wrangling Cheat Sheet.** R Studio Data Wrangling cheat sheet <https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf>
- **Online Textbook:** Claudia A. Engel (2019) *Data Wrangling with R*. Online text < https://cengel.github.io/R-data-wrangling/>
- **RStudio Data Wrangling Video:** From RStudio <https://rstudio.com/resources/webinars/data-wrangling-with-r-and-rstudio/>

Homework

- DataCamp: Cleaning Data in R <https://learn.datacamp.com/courses/cleaning-data-in-r>
- Descriptives Assignment Due
- Wrangling Assignment Assigned

Week 5: Basics of Statistical Inference

Mar 2. Using data to test ideas.

Topics

- ***Inferring Parameters from Samples.*** Populations and samples. Uncertainty. Inferences.
- ***Hypothesis Testing.*** What is a hypothesis test? Understanding significance and p-values. Univariate inferential statistics. Bivariate inferential statistics.
- ***Data Simulation.*** What is simulation? Uses of data simulation. Simulating variables and models. Bootstrapping.

Reading

- Urdain: Ch. 4 – 7, 9
- Gelman, Hill, and Vehtari: Ch 4 - 5
- Teetor: Ch. 8 - 9

Homework

- Wrangling Assignment Due
- Inferential Statistics Assignment Assigned

Week 6: An Introduction to Modeling

Mar 9. Creating and refining mathematical equations to describe the real world's workings.

Topics

- **An Introduction to Modeling.** What are statistical models? Approaches to empirical model development.
- **Inferring Causal Relationships from Empirically-Refined Models.** Regression as quasi-experiment. Inferring causation. Partial correlation and statistical control
- **Ordinary Least Squares Models.** Implementation of univariate and multivariate OLS for continuous outcomes.

Helpful Reads

- **General Introduction to Predictive Modeling:** Foster Provost & Tom Fawcett (2013) Chapters 3
- **How Linear Models are Fit:** Fox and Weisberg (2019): Chapter 4
- **Interpreting lm() Regulsts:** Felipe Rego (2015) "Quick Guide: Interpreting Simple Linear Model Output in R" Blog Post. https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R
- **Programming for Regression Analysis:** Horton and Kleinman (2015)*:* Chapter 6

Reading

- Urdain: Ch. 13
- Gelman, Hill, and Vehtari, Ch. 6 - 8
- Teetor: Ch. 11

Homework

- DataCamp Module: Correlation and Regression in R <https://learn.datacamp.com/courses/correlation-and-regression-in-r>
- Inferential Statistics Assignment Due
- Regression Assignment Assigned

Week 7: Better Regression Models

Topics

- **Transforming Predictors.** Cutting predictors. Log transformations. Top- and bottom-coding.
- **Model Respecification.** Adding and removing predictors. Likelihood-ratio tests.
- **Nonlinear and Interaction Effects.** Special conceptions of causal mechanisms.
- **Regression Diagnostics.** Tests to confirm OLS underlying assumptions hold

Reading

- Gelman, Hill, and Vehtari, Ch. 8

Further Reading

- Fox (1998): Chapters 11 – 13
- Fox and Weisberg (2019): Ch 8
- Horton and Kleinman (2015)*:* Chapter 6

Homework

- Regression Assignment Due
- Regression Diagnostics Assignment Assigned

Week 8: Basic Bayesian Models

Topics

- An introduction to the Bayesian approach to regression modeling
- Simple Bayesian models using ***rstanarm***
- Model diagnostics

Reading

- Gelman, Hill, and Vehtari, Ch. 9 - 12

Homework

- DataCamp Module: Bayeisan Regression Modeling with rstanarm < https://www.datacamp.com/courses/bayesian-regression-modeling-with-rstanarm>
- Regression Diagnostics Assignment Due

Week 9: Modeling Discrete Variables

Topics

- Logistic regression
- Ordered logit
- Poisson
- Beyond GLM

Reading

- Gelman, Hill, and Vehtari, Ch 13 - 15

Further Reading

- Fox (1997): Chapter 7, 15
- Fox and Weisberg (2019): Chapter 6
- Horton and Kleinman (2015)*:* Chapter

Homework

- DataCamp Module: Generalized Linear Models in R <https://learn.datacamp.com/courses/generalized-linear-models-in-r>
- Discrete Outcomes Regression Assigned

Week 10: Working with Survey Data

Topics

- Sampling and Weighting
- Survey Analysis in R

Reading

- Thomas Lumley (2010) *Complex Surveys: A Guide to Analysis Using R.* Wiley. Selections TBD.

Homework

- DataCamp Module: Analyzing Survey Data in R < https://learn.datacamp.com/courses/analyzing-survey-data-in-r>
- Final Project Assigned
- Discrete Outcomes Regression Due

Week 11: Missing Data / Causal Inference

Topics

- Missing data
- Data Imputation
- Topics in causal inference

Reading

- Gelman, Hill, and Vehtari, Ch 17 - 20

Helpful Reads

- **Amelia Documentation:** James Honaker, Gary King & Matthew Blackwell (2019) *Amelia II: Program for Missing Data*. < https://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf>

Week 12: Longitudinal Analysis

Topics

- TBD

Reading

- TBD
- Teetor: 7.8 – 7.14, Ch 14

Week 13: Multivariate Analysis

Topics

- Index Development
- Factor Analysis
- Similarity Measures
- Cluster Analysis

Reading

- Urdain: Ch. 15
- Everitt and Hothorn (2011): Chs. 5 & 7
- Teetor: 13.6, 13.9

Week 14: Final Project Presentations

Students will present the verbal presentation component of their final project.

Class Policies

Academic Honesty

The university defines the administrative processes and sanctions for violating its code of academic integrity in *The City University Policy on Academic Integrity.* You are advised to read and understand this policy immediately. Should you have any difficulty accessing or interpreting this policy, contact the instruction.

Students with Disabilities

Students who require accommodation based on special needs must contact the Office of Special Services for Students with Disabilities at 171 Kiely Hall, (718) 997-5870. The instructor cannot make accommodations for students with special needs unless this office works out arrangements with him.

Public Health or Safety Emergencies

This course may move online in the event of a public health or safety emergency. In the event that we move online, we will communicate via Zoom and exchange documents through Dropbox (instructions).

Access to Computer and Internet at Home

Many students do not have reliable access to a computer or Internet connection from home. This is far more common than students realize. Under normal circumstances, students without quality computer access can use the campus computer system. In the event of an emergency that cuts off access to campus computing resources, please notify the professor *before* you miss class or deadlines.

Late Grades

Students are encouraged to submit their assignments well in advance of the due date, so as to prevent them from falling victim to unanticipated technical or personal emergencies. Blackboard automatically designates assignments as having been submitted on-time or late.

For all assignments except *Final Project*:

- If an assignment is uploaded late, but is successfully uploaded to Blackboard within seven days of the due date, it will receive a half-alpha penalty. With this penalty, an "A" submission would be scored "A-", an "A-" submission scored a "B+" and so on.

- If an assignment is uploaded late, but is successfully uploaded to Blackboard within 21 days of the due date, it will receive a full alpha penalty, such that an "A" submission will be scored a "B", an "A-" scored as "B-", and so forth.

- If an assignment is uploaded beyond 21 days late, but before the due date of the final exam, then the professor may elect to assign it a grade of "C" if he judges its quality to be of at least a "B-" or better quality.

No other late submissions will be considered. The purpose of this assignment is to stir class discussion. If it is not done on time, the submission serves no purpose for the class.

Final Project. The professor will not consider late submissions of the final project. Students are encouraged to submit well in advance of the due date.

Emergency? The professor reserves the right to *prospectively* extend due dates if a student runs into an unanticipated personal emergency. In other words, if something disrupts your studies, please arrange a meeting with me to set up changes to due dates and/or assignments in advance. Extensions are at the sole discretion of the professor, and will not be considered on assignment due dates itself. It is important that you submit your materials in advance, here and in life in general. The whole purpose of this policy is to try to push last-minute people to learn to submit things in advance.

Classroom Conduct

Any student who takes this class agrees to act professionally in class. Any student who disrupts lectures, classes or tests will be asked to leave, and referred to the Dean's Office.

Communicating with the Instructor

Communicate with the instructor via e-mail. Under normal circumstances, you can expect a response within three business days. If your response is urgent and requires immediate attention, write "URGENT" in the subject line of your e-mail. The telephone is for use in case of emergencies – for example, if you are about to miss an exam. Do not leave messages on the instructors' voice mail – write an e-mail instead.

Grades

Grades will be available on Blackboard. Grades reflect careful consideration of student performance and will not be changed unless a recording or calculation error is revealed. It is your responsibly to monitor your grades in Blackboard and provide immediate feedback in the case of any discrepancy. Students must notify the instructor via e-mail of any problems with their grades within two weeks of their grades' posting on Blackboard. After two weeks from the grades posting on Blackboard, all grades are final and not subject to further discussion.

Office Hours

If you cannot attend any of the instructor's office hours due to conflict with your class schedule, you must notify the instructor and provide a copy of your class schedule within the second week of class.

Compliance with Public Health Instructions

Any student who does not comply with government or college public health directives represents a health threat to me and the class. If met with a situation in which a student refuses to comply with public health directives while in class, then I cannot responsibly expose you to other students or the College community at large. Should you wish to challenge this policy, I invite you to do so immediately so as not to disrupt other students with COVID-related conflict.