DATA 712: Advanced Analytics

Syllabus
Joseph N. Cohen
Spring 2024

Overview

My goal in this class is to train you to create information based on advanced data-analytical techniques. We will survey a range of advanced analytical operations using the R platform.

Instructor

Joseph Cohen
Associate Professor of Sociology
City University of New York, Queens College
www.josephnathancohen.info
joseph.cohen@qc.cuny.edu

Office: Powdermaker 252 H
Office Hours: Tuesdays, 1:30 PM – 3:30 PM on Microsoft Teams (click here for link)

Join on your computer, mobile app or room device
Click here to join the meeting
Meeting ID: 263 225 970 818
Passcode on Teams site
Download Teams | Join on the web

Or call in (audio only)
+1 315-314-2882,,656323638# United States, Syracuse
Phone Conference ID: 656 323 638#
Find a local number | Reset PIN

Schedule

This class meets on Wednesdays from 6:30 PM to 8PM in PH 351

Textbooks

There is no textbook in this class. Readings will be distributed via Blackboard.

Teams

Our class will communicate through Microsoft Teams. Click here to connect.

Assessment

Your work will be assessed as follows:

- **Class Participation (10%).** Students receive a letter grade related to the quality of their contribution to class discussions and the overall class learning experience. Student can perform well by participating thoughtfully in class discussions and exercises.
- **DataCamp (10%).** Students receive full credit for completing the required DataCamp modules by the deadline. Students can complete them later in the semester for reduced credit. Due throughout semester.
- **First Project (20%).** Student will create an original and useful information product as a paper or web page. Due February 28
- **Second Project (25%).** Student will create an original and useful information product as a paper or web page. Due April 4
- **Final Project (35%).** Student will create an original and useful information product as a paper or web page. Due May 15

Grade Scales

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%	Outstanding
A-	90% - 92.9%	92%	Excellent
B+	87% - 89.9%	88%	Good
B	83% - 86.9%	85%	Average
B-	80% - 82.9%	82%	Basic
C	70% - 79.9%	75%	Basic with Nontrivial Gaps
D	60% - 69.9%	65%	Serious Deficiencies
D-	50% - 59.9%	55%	Minimally Acceptable
F	<60%	0%	Unacceptable

Agenda

Readings in italics are optional.

Week 1: Class Introduction

January 31, 2024

Topics

- Class Introduction
- The Craft of Data Analysis
- Thinking Like a Social Scientist
- Generative AI

Readings

- Nate Silver (2012) *The Signal and the Noise.* Penguin Press. Chapters 2 - 3.
- *Joseph Cohen (2023) "Adapting to AI: How Will Generative AI Affect Work? How Should We Respond?" Discussion Paper. August 30.* https://jncohen.commons.gc.cuny.edu/wp-content/blogs.dir/15488/files/2023/12/Adapting-to-AI-101-Joseph-Cohen.pdf
- *Robert L. Bray (2023) "Lessons Learned When Teaching Data Analytics with ChatGPT to MBAs in Spring 2023" Discussion Paper. August 28.* https://www.kellogg.northwestern.edu/faculty/bray/doc/chatgpt/chatgpt.pdf

DataCamp

- Optional Course: Introduction to R. https://app.datacamp.com/learn/courses/free-introduction-to-r

Week 2: R Programming Basics

February 6, 2024

Topics

- Project Management Basics
- R Basics Reviewed
- R Markdown Reviewed
- Working with Environments
- Creating Functions
- Looping

Readings

- *Richard Cotton. 2013. Learning R. O'Reilly.* Ch. 6, 8, 9
- *R Studio Data Wrangling cheat sheet* https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

DataCamp

- **Required Course**: Intermediate R. (due Feb 16) https://app.datacamp.com/learn/courses/intermediate-r

Week 3: Describing & Visualizing Distributions

February 14, 2024

Topics

- Transitioning from Learner to Teacher
- Describing Distributions Reviewed
- Visualizing Distributions using ggplot2 Reviewed

Readings

- UC Business Analytics R Programming Guide. "An Introduction to ggplot2" https://uc-r.github.io/ggplot_intro
- *Noah Illinsky "On Beauty" Julia Steele and Noah Ilinsky (eds.) Beautiful Visualization. O'Reilly. Chapter 1*
- *Matthias Shapiro. "Once Upon a Stacked Time Series" Julia Steele and Noah Ilinsky (eds.) Beautiful Visualization. O'Reilly. Chapter 2*

DataCamp

- **Required Course**: Introduction to the Tidyverse. (due February 23) https://app.datacamp.com/learn/courses/introduction-to-the-tidyverse

Week 4: Data Wrangling

February 21, 2024. **This is an asynchronous class**.

DataCamp

- **Required Course:** Intermediate Importing Data in R. (due March 1) https://app.datacamp.com/learn/courses/intermediate-importing-data-in-r

Week 5: No Class This Week

February 28, 2024

On Wednesday, February 28, CUNY will follow a Monday schedule.

Week 6: Working with Survey Data

March 6, 2024

Topics

- Sample Representation
- Observation Weighting
- Working with Complex Survey Data

Readings

- Thomas Lumley (2010) Complex Surveys: A Guide to Analysis Using R. Wiley. Chapters 1 – 3, 5 - 6.

DataCamp

- Optional Course: Analyzing Survey Data in R. https://app.datacamp.com/learn/courses/analyzing-survey-data-in-r

Week 7: Data Reduction & Index Construction

March 13, 2024

Topics

- Cronbach's Alpha
- Factor Analysis
- Principal Components Analysis
- Index Construction

Readings

- Matteo Mazziotta and Adriano Pareto (2015) "Methods for Constructing Composite Indexes: Once for All or All for One? Rivista Italiana di Economia Demografia e Statistica. LXVIII (2)
- *Diana Reckien (2018) "What is in an index? Construction method, data metric, and weightingscheme determine the outcome of composite social vulnerability indicesin New York City" Regional Environmental Change.*
- *Timothy Urdan. 2010. Statistics in Plain English. Rutledge.* Chapter 15

DataCamp

- **Required Course:** Dimensionality Reduction in R. https://app.datacamp.com/learn/courses/dimensionality-reduction-in-r

Week 8: Regression Modeling

March 20, 2024

Topics

- Basics of inferential statistics
- OLS regression basics
- Regression diagnostics

Readings

- John Fox (1998) *Applied Regression Analysis, Linear Methods and Related Methods.* Sage. Chaters. 2, 3, 6, 11, and 13
- *Timothy Urdan. 2010. Statistics in Plain English. Rutledge.* Chapter 13
- *Foster Provost & Tom Fawcett (2013) Data Science for Business. O'Reilly. Chapters 3*

DataCamp

- **Required Course:** Introduction to Regression in R. (due March 29) https://app.datacamp.com/learn/courses/introduction-to-regression-in-r
- Optional Course: Intermediate Regression in R. https://app.datacamp.com/learn/courses/intermediate-regression-in-r

Week 9: Missing Data

March 27, 2024

Topics

- The effects of missing data
- Missing data handling methods
- Multiple Imputation with randomness

Readings

- Craig K. Enders *Applied Missing Data Analysis.* Guilford Press. Chapters 1 & 2
- Gary King *et al.* (2001) "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation" *American Political Sciences Review.* 95 (1)
- Matthew Blackwell, James Honaker, and Gary King. 2017. "A Unified Approach to Measurement Error and Missing Data: Overview and Applications." Sociological Methods and Research, 46, 3, Pp. 303-341.

DataCamp

- Optional Course: Dealing with Missing Data in R. https://app.datacamp.com/learn/courses/dealing-with-missing-data-in-r

Week 10: Classification

April 3, 2024

Topics

- Classification
- Similarity Metrics
- Multidimensional Scaling
- Cluster Analysis

Readings

- Brian Everitt and Thorsten Hothorn (2011) *Introduction to Applied Multivariate Analysis with R.* Springer. Ch. 1, 4, 5, and 6
- Mahmoud Harmouch (2021) "17 Types of Similarity and Dissimilarity Measures Used in Data Science. *Towards Data Science.* March 13. https://towardsdatascience.com/17-types-of-similarity-and-dissimilarity-measures-used-in-data-science-3eb914d2681

DataCamp

- Optional Course: Supervised Learning in R: Classification. https://app.datacamp.com/learn/courses/supervised-learning-in-r-classification

Week 11: Multilevel Models

April 10, 2024

Topics

- Introduction to multilevel modeling
- Implementing multilevel models in R

Readings

- TBD

DataCamp

- Optional. Hierarchical and Mixed Effects Models in R. https://app.datacamp.com/learn/courses/hierarchical-and-mixed-effects-models-in-r

Week 12: Geospatial Analysis

April 17, 2024

Topics

- Basic concepts in geospatial analysis
- Mapping data

Readings

- Guy Lansley and James Cheshire (2016) "An Introduction to Spatial Data Analysis and Visualization in R" https://www.spatialanalysisonline.com/An%20Introduction%20to%20Spatial%20Data%20Analysis%20in%20R.pdf

Week 13: No Class This Week

April 24, 2024

Spring Break: No classes this week

Week 14: Bayesian Analysis

May 1, 2024

Topics

- Frequentist vs. Bayesian Statistics
- Simple Bayesian Models
- Empirical Bayesian Regression with rstanarm

Readings

- TBD

DataCamp

- Fundamentals of Bayesian Data Analysis in R. https://app.datacamp.com/learn/courses/fundamentals-of-bayesian-data-analysis-in-r
- Bayesian Regression Modeling with rstanarm. https://app.datacamp.com/learn/courses/bayesian-regression-modeling-with-rstanarm

Week 15: Machine Learning

May 8, 2024

Topics

- Machine Learning
- Neural Networks
- Artificial Intelligence

Readings

- Qifang Bi etl al. "What is Machine Learning? A Primer for the Epidemiologist" *American Journal of Epidemiology.* 188 (12)
- *Joseph Cohen (2022) "Computers Can Make Art: Introducing "AI-Generated" Images. Blog Post. October 1. https://josephnathancohen.info/posts/computers-can-make-art-introducing-ai-generated-images/*

DataCamp

- Optional: Unsupervised Learning in R. https://app.datacamp.com/learn/courses/unsupervised-learning-in-r

Week 16: Final Presentations

May 15, 2024

Students will give a PowerPoint presentation that describes their Final Project, its findings, and these findings' implications.

Course Policies

Academic Integrity. 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.

For 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.

Policies on Late Submissions. All submissions should be uploaded to Blackboard by midnight Eastern time. Students can submit an assignment up to two weeks late for a one-alpha penalty.

You are Responsible for Teams, Blackboard and Computer Access. Students are expected to check Blackboard or their school emails regularly. Any announcements made on Blackboard are expected to have been read, acknowledged and clarified by students within four days of its posting. Contact the OCT Help Desk if you are having trouble with Blackboard. Students are also responsible for ensuring that they have access to a computer during class. It is a required class material. If you do not have a computer, you can borrow one from the library's Media Services desk.

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. 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.