DATA 333: Data Management, Processing, and Visualization

Syllabus
Joseph N. Cohen
Spring 2024

Overview

This course is part of a larger curriculum on the field of *data analysis*, the task of gathering and analyzing data to create insights. This particular course examines the practical process of creating these insights, from project conception to audience communications. Topics include gathering and storing data, making data usable, the general insights that can be gleaned from data and the practicalities of doing so, and the task of communicating the knowledge that you create to users. We will discuss data wrangling, visualization, statistical analysis, regression modeling, classification, and dimensionality reduction. This hands-on course develops these skills using R, the world's leading statistical analytical platform.

Baseball Statistics. This semester, DATA 333 will introduce these topics through its application to tasks of evaluation, analysis, and prediction in baseball. The course will not require readings on baseball, but will use the sport as a vehicle to impart class concepts.

Meeting Times

Wednesdays, 1:40 PM – 4:30 PM in PH 155.

Instructor

Joseph Cohen
Associate Professor of Sociology, City University of New York, Queens College
Web: www.josephnathancohen.info; Email: 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

Teams

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

Textbooks

Richard Cotton. 2013. *Learning R*. O'Reilly.

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 deadlines. Students can complete them later in the semester for reduced credit. Due throughout semester.
- **Homework (20%).** Students will be graded on homework exercises that demonstrate their grasp of class work. Due throughout the semester.
- **Midterm Projects (20%).** Students will write a paper or produce a web page that conveys original, rigorously-derived information with real world implications. Due March 27.
- **Final Project (40%).** Students will write a paper or produce a web page that conveys original, rigorously-derived information with real world implications. Due on 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
- Using Chat-GPT
- Craft of Data Analysis
- Thinking Like a Social Scientist
- *Baseball as a Vehicle for Learning Data Analysis*
- *An Introduction to the Game of Baseball*

Readings

Data Analysis

- 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*

Baseball

- *Neil Weinberg (2015) "Getting Started" FanGraphs. January 15. <https://library.fangraphs.com/getting-started/>*
- *Anon. (n.d.) "Baseball for Beginners" PBS. Web page <https://www.pbs.org/kenburns/baseball/baseball-for-beginners>*
- *wikiHow (2020) "How to Play Baseball" YouTube Video.*

Week 2: An Introduction to R

February 7, 2024

Topics

- Running Analytics Projects
- An Introduction to R, RStudio, and R Markdown
- *Assessing hitters: Traditional metrics*

DataCamp

- **Required Course**: Introduction to R. https://app.datacamp.com/learn/courses/free-introduction-to-r (due February 16)

Readings

Data Analysis

- DataCamp (2022) "RStudio Tutorial" September. https://www.datacamp.com/tutorial/r-studio-tutorial
- RStudio. "R Markdown from RStudio" https://rmarkdown.rstudio.com/lesson-1.html
- Richard Cotton (2013) *Learning R.* O'Reilly. Chs. 1 - 3

Baseball

- *Neil Weinberg (2014) "The Beginner's Guide to Using Statistics Properly" FanGraphs. September 15. https://library.fangraphs.com/the-beginners-guide-to-using-statistics-properly/*
- *Anon. (n.d.) "Baseball for Beginners" PBS. Web page <https://www.pbs.org/kenburns/baseball/baseball-for-beginners>*

Week 3: R Basics

February 14, 2024

Topics

- R Basics
- *Assessing hitters: Advanced metrics*

DataCamp

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

Readings

Data Analysis

- Richard Cotton (2013) *Learning R.* O'Reilly. Chs. 4, 5, 12

Baseball

- *Neil Weinberg (2014) "How to evaluate a hitter, sabermetrically" SBNation. May 26. https://www.beyondtheboxscore.com/2014/5/26/5743956/sabermetrics-stats-offense-learn-sabermetrics*
- *Neil Weinberg (2014) "Complete List (Offense)" FanGraphs. October 30. https://library.fangraphs.com/offense/offensive-statistics-list/*

Week 4: Project Workshop

February 21, 2024

Workshop in which class will consolidate skills and develop Midterm Project.

DataCamp

- **Required Course**: Introduction to the Tidyverse. (due March 1) https://app.datacamp.com/learn/courses/introduction-to-the-tidyverse

Week 5: No Class This Week

February 28, 2024

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

Week 6: Data Wrangling Basics

March 6, 2024

Topics

- Data Wrangling
- *Assessing pitchers*

DataCamp

- Optional Course: Cleaning Data in R. https://app.datacamp.com/learn/courses/cleaning-data-in-r

Readings

Data Analysis

- Richard Cotton. Learning with R. Chapters 6, 7, 8, 10, 13
- Hadley Wickham (2014) "Tidy Data" *Journal of Statistical Software*, 59 (10): 1 – 23. (preprint version)
- Hadley Wickham (2023) *R for Data Science*. O'Reilly. Ch. 5 (https://r4ds.hadley.nz/data-tidy), Chapter 19 (https://r4ds.hadley.nz/joins)
- *Erin Sovansky Winter. A Language, not a Letter; Learning Statistics in R. Chapters 3 (https://ademos.people.uic.edu/Chapter3.html) and Chapter 4 (https://ademos.people.uic.edu/Chapter4.html)*
- *Timothy Carsel "Meling and Casting" A Language, not a Letter; Learning Statistics in R. Chapter 8 (https://ademos.people.uic.edu/Chapter8.html)*

Baseball

- *David (2006) "Evaluating Pitcher Talent" U.S.S. Mariner. August 29.* *http://www.ussmariner.com/2006/08/29/evaluating-pitcher-talent/*
- *Neil Weinberg (2014) "Complete List (Pitching)" FanGraphs. December 18.* *https://library.fangraphs.com/pitching/complete-list-pitching/*
- *Owan McGrattan (2023) "Stuff+, Location+, and Pitching+ Primer" FanGraphs. March 10.* *https://library.fangraphs.com/pitching/stuff-location-and-pitching-primer/*

Week 7: Getting Data from the Internet

March 13, 2024

Topics

- APIs
- Web Scraping
- Data Governance
- *Generic pitch types*
- *Assessing pitches*
- *Pitch Repertoire*

DataCamp

- **Required Course:** Intermediate Importing Data in R. (due March 29) https://app.datacamp.com/learn/courses/intermediate-importing-data-in-r
- Optional Course**:** Web Scraping in R. https://app.datacamp.com/learn/courses/web-scraping-in-r

Readings

Data Analysis

- Hadley Wickham (2023) "Web Scraping" *R for Data Science.* O'Reilly. https://r4ds.hadley.nz/webscraping
- Richie Cotton (2023) "Data Governance Fundamentals Cheat Sheet" *DataCamp.* January. https://www.datacamp.com/cheat-sheet/data-governance-fundamentals-cheatsheet
- Joseph Nathan Cohen (2023) "Download Baseball Data with baseballr" Post. January 10. https://josephnathancohen.info/posts/download-baseball-data-into-r/
- *Christian Pascual (2020) "R API Tutorial: Getting Started with APIs in R" DataQuest. February 13.* *https://www.dataquest.io/blog/r-api-tutorial/*

Baseball

- *Cameron Key (2024) "Using Statcast Pitch Data" baseballr. January 15. https://billpetti.github.io/baseballr/articles/using_statcast_pitch_data.html*
- *David Adler (2019) "Identifying Pitch Types: A Fan's Guide" MLB. June 1. https://www.mlb.com/news/identifying-pitch-types-a-fan-s-guide*
- *Ethan Moore (2020) "Measuring Pitch Quality" Something Tangible, Medium. https://medium.com/something-tangible/measuring-pitch-quality-590043713e74*

Week 8: Describing Distributions

March 20, 2024

Topics

- Describing univariate distributions
- Describing relationships
- Basic inferential statistics
- *Game Theory Basics*
- *Win Probabilities and Game Strategy*

DataCamp

- **Required Course:** Introduction to Statistics in R. https://app.datacamp.com/learn/courses/introduction-to-statistics-in-r

Readings

Data Analysis

- Timothy Urdain (2010) *Statistics in Plain English.* Chapters 2, 3, 8

Baseball

- *Matt Swartz (2012) "Game Theory and Baseball, Part 1: Concepts" Hardball Times. December 17. https://tht.fangraphs.com/game-theory-and-baseball-part-1-concepts/*
- *Tom M. Tango, Mitchel G. Lichtman, and Andrew E. Dolphin (2006) The Book: Playing the Percentages in Baseball. TMA Press. Chapter 1. (BB)*
- *Piper Slowinski (2010) "Regression toward the Mean" FanGraphs. February 16. <https://library.fangraphs.com/principles/regression/>*
- *Dylan Drummey (2023) "The Pitcher's Dilemma: Applying Game Theory to Pitching Decisions." Pitcher List. March 7. https://pitcherlist.com/the-pitchers-dilemma-applying-game-theory-to-pitching-decisions/*

Week 9: Visualization Basics

March 27, 2024.
Note: Midterm Paper is due today

Topics

- Basics of Visualization
- Visualization using ggplot2
- *Fielding Metrics*

DataCamp

- **Required Course:** Introduction to Data Visualization with ggplot2. (due April 12) https://app.datacamp.com/learn/courses/introduction-to-data-visualization-with-ggplot2

Readings

Data Analysis

- 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*

Baseball

- *Jeff Zimmerman and Dan Basco (2010) "Mesauring Defense: Entering the Zones of Fielding Statistics" Baseball Research Journal. https://sabr.org/journal/article/measuring-defense-entering-the-zones-of-fielding-statistics/*

Week 10: Data Reduction & Index Construction

April 3, 2024

Topics

- Dimensionality Reduction
- Feature Selection and Exaction
- Cronbach's Alpha
- Factor Analysis
- Principal Components Analysis
- Composite Index Development

DataCamp

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

Readings

Data Analysis

- 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

Week 11: Classification

April 10, 2024

Topics

- Similarity metrics
- Multidimensional scaling
- Simple clustering methods
- Hierarchical clustering methods
- *Player Similarities*

DataCamp

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

Readings

Data Analysis

- 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

Baseball

- *John Moore (2021) "Identifying Overlooked Position Players with K-Means Clustering" BaseballCloudBlog. May 25. https://baseballcloud.blog/2021/05/25/identifying-overlooked-position-players-with-k-means-clustering/*
- *Jonah Simon. "Grouping Major League Hitters with Hierarchical Methods" Analytics Vidhya. https://medium.com/analytics-vidhya/grouping-major-league-hitters-with-hierarchical-methods-e7dc35b7f665*

Week 12: Regression Modeling

April 17, 2024

Topics

- Basics of inferential statistics
- OLS regression basics
- Regression diagnostics
- *Performance Forecasting*
- *Player Valuation*

DataCamp

- **Required Course:** Introduction to Regression in R. (due May 2)
  https://app.datacamp.com/learn/courses/introduction-to-regression-in-r

Readings

Data Analysis

- *Timothy Urdan. 2010. Statistics in Plain English. Rutledge.* Chapter 13
- *Foster Provost & Tom Fawcett (2013) Data Science for Business. O'Reilly. Chapters 3*

Baseball

- *Nate R. "Using Regression to Predict Baseball Salaries" LinkedIn.*
  https://www.linkedin.com/pulse/using-regression-predict-baseball-salaries-nate-reed/
- *Vince Gennaro (2009) "Estimating the Dollar Value of Players" Baseball Research Journal. Summer.*

Week 13: No Classes This Week

April 24, 2024

Spring Break: No classes this week

Week 14: Dynamic Visualization

May 1, 2024

Topics

- Basic Principles Dynamic Visualization
- Introduction to Tableau
- *Small Sample Problem*
- *Regression to the mean*

DataCamp

- Optional Course: Introduction to Tableau.
  https://app.datacamp.com/learn/courses/introduction-to-tableau

Readings

Data Analysis

- TBD

Baseball

- *David Cameron (2015) "The Meaning of Small Sample Data" April 13. FanGraphs. https://blogs.fangraphs.com/the-meaning-of-small-sample-data/*
- *Russell A. Carleton (2012) "Baseball Therapy: It's a Small Sample Size After All" Baseball Prospectus. July 16. https://www.baseballprospectus.com/news/article/17659/baseball-therapy-its-a-small-sample-size-after-all/*

Week 15: Further Topics

May 8, 2024

Topics

- Ethics, Privacy, and Security
- Machine Learning

Readings

- Qifang Bi etl al. "What is Machine Learning? A Primer for the Epidemiologist" *American Journal of Epidemiology.* 188 (12)

Week 16: Final Project Presentations

May 15, 2024
Note: Final Project Due
Students will present their final projects

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, A student can submit up to two weeks late for a one-alpha penalty.

You are Responsible for Blackboard and Computer Access. Students are expected to check Blackboard or their campus email 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.