NYU | ROBERT F. WAGNER GRADUATE SCHOOL OF PUBLIC SERVICE

PADM-GP 2505

Advanced Data Analytics and Evidence Building
Spring 2024

Instructor Information

- Benjamin Feder
- Email: benjamin.feder@nyu.edu
- Office Hours (remote): Tuesday 5:30 PM - 6:30 PM; Wednesday 5:30 PM - 6:30 PM
    - Subject to change pending student availability

Teaching Assistant Information

- Xiangyu Ren
- Email: xiangyu.ren@nyu.edu

Course Information

- Lecture: Friday, 10:00 AM - 11:40 AM
- Lab: Friday, 11:50 AM - 12:50 PM
- Location: 60 Fifth Avenue, Room 125

Course Description

This course aims to develop key data analytics skill sets necessary to inform evidence-based policy. Its design offers hands-on training in how to make sense of and use large-scale heterogeneous datasets in the context of addressing real-world problems. The main learning objectives are to develop a better understanding of how to analyze issues using data from a variety of different sources and how to communicate the resulting insights through a public policy lens. It is designed for graduate students who are seeking a stronger foundation in data analytics and scoping questions that can be answered with available data. Students are expected to apply core concepts to an individual, class-long data project focused on a critical public policy issue. The associated textbook provides more information on the content covered in this class.

Course Structure

The course will be structured in weekly sessions. The sessions will consist of interactive lectures with some time also devoted to class projects and will be supplemented with a weekly lab directly following each class session. The lectures will cover both general theory and specific applications, and the labs will expand on the lectures by discussing relevant coding techniques

to implement these methods on the class data in Python. Labs will typically focus on the coding materials based on concepts introduced in the previous week's lecture.

The class schedule is tentative and is subject to change. Readings and interactive coding materials are expected to be completed before class on the day of the assignment. Additional resources can be found on NYU Brightspace.

Readings

The syllabus provides the relevant chapters from the textbook and additional readings for each week. Students may be pointed to more readings during lectures. This course's textbook is available at the NYU Library through the hyperlink below.

Big Data and Social Science: A practical guide to models and tools, 2nd edition, Taylor Francis 2020, Ian Foster, Rayid Ghani, Ron Jarmin, Frauke Kreuter and Julia Lane.

Housekeeping

- The NYU Brightspace site for this course will contain the lecture slides, additional reading materials, and assignments. Since instruction will take place in person, lectures will not be recorded. Slides will be available on Brightspace after class for those who wish to review content. Notifications and updates will be sent out through NYU Brightspace regularly.
- Students are expected to attend classes in person and use the lab time to apply the class concepts to their projects. **Students should plan to devote 1-2 hours a week outside of class to developing their class project.**
- Students are not permitted to take notes via laptop or phone. Laptops and phones may only be used for participating in class polls.
- Punctuality is **very important**. Unforeseen circumstances may arise, but all students are generally expected to be on time. Students should notify Professor Feder in advance if they will be late or unable to attend.
- Students are expected to be prepared for class discussions and to keep up with prior class content. The open exchange of ideas will be respected by all. Respectful and inclusive discussion is required.
- Grades are non-negotiable.
- Late submissions are accepted but are counted as late and will be penalized as specified in the Evaluation section. Students can always turn in an assignment early to avoid penalties. There are no make-up assignments.

Evaluation

Project work
Each student will work on their individual research project throughout the semester. The goal of the research project is for students to independently develop and apply the techniques taught in the class to the provided dataset, and supplemental resources—a project template, notebooks, and lecture notes—will be provided to help guide the work. Questions or concerns about individual projects can be discussed via email and office hours. All students are required to schedule two separate graded, 10-minute informal sessions to discuss their project with Professor Feder before the midterm and final presentations.

There will be a midterm presentation and a final presentation of the project results. At the end of the semester, students are required to submit a short research paper documenting their project work. Submitting the final research paper up to 24 hours late will result in a 25% grade reduction; between 25 and 48 hours late, a 50% reduction; beyond 48 hours late, no credit.

Class preparation and participation
Preparation and class participation will constitute 25% of the final grade (examples: participation in class discussions, responses in class polls, engagement in peer midterm and final presentations, posting on NYU Classes forum, responding to questions when asked, helping classmates by sharing code snippets and helping them debug code, sharing information that might be interesting for classmates). Class participation and concise, clear communication is essential for this class. Attendance is mandatory, and unexcused absences will be noted in the participation grade.

The breakdown of the evaluation activities is summarized in the table below:

Activity	Percentage of Grade
Project work	**75%**
1-on-1 midterm discussion	10%
Midterm presentation	15%
Final presentation	20%
1-on-1 final discussion	15%
Final paper (10 pages)	15%
Class preparation and participation	**25%**
Polling and class participation	15%
Peer feedback on the midterm and final presentations	10%

Project Guidance

All prepared coding materials—in the form of interactive Jupyter Notebooks—will be written in Python. Students should repurpose as much code as possible from the notebooks for their class projects. Stack Overflow and generative AI tools such as ChatGPT (more on this in the Plagiarism section) can also be useful coding resources. Students are permitted to use additional software tools and programming languages to complete their projects, but no support will be provided for non-Python tools and languages.

A project template has been developed to serve as both a general benchmarking tool and a decision tracker. Keeping the project template up to date will make it much easier to develop final presentations and papers. Students are not required to submit the project template at the end of the semester.

Plagiarism

All students must produce original work. Outside sources are to be properly referenced and/or quoted. Lifting copy from websites or other sources and trying to pass it off as original words constitutes plagiarism. Such cases can lead to academic dismissal from the university, as taking credit for other’s writing is a violation of NYU’s Academic Integrity policy.

The use of generative AI (ChatGPT and related tools) is allowed in this class as a coding support tool. Since part of the individual project evaluation for each activity will focus on creative and critical thinking, using generative AI will likely be counterproductive. In particular, copying generative AI output and submitting it as a final report is prohibited. Plagiarism of any form—including copying generative AI output—will not be tolerated. Students in this class are expected to report violations to Professor Feder.

Academic Integrity

Academic integrity is a vital component of Wagner and NYU. All students enrolled in this class are required to read and abide by Wagner's Academic Code. All Wagner students have already read and signed the Wagner Academic Oath. Students unsure about expectations and how to abide by the academic code should consult with Professor Feder immediately.

Henry and Lucy Moses Center for Students with Disabilities at NYU

Academic accommodations are available for students with disabilities. Please visit the Moses Center for Students with Disabilities (CSD) website and click the "Get Started" button to apply for accommodation. Students can also contact CSD directly (212-998-4980 or mosescsd@nyu.edu) for information. Students requesting academic accommodations are strongly advised to reach out to the Moses Center as early as possible in the semester for assistance.

NYU's Calendar Policy on Religious Holidays

NYU's Calendar Policy on Religious Holidays states that members of any religious group may, without penalty, absent themselves from classes when required in compliance with their religious obligations. Please notify Professor Feder in advance of religious holidays that might coincide with presentations to schedule mutually acceptable alternatives.

NYU's Wellness Exchange

NYU's Wellness Exchange has extensive student health and mental health resources. A private hotline (212-443-9999) is available 24/7 to connect students with a professional who can help them address day-to-day challenges as well as other health-related concerns.

Projected Class Schedule and Assignments

The schedule listed below is subject to change and should only be used as a reference. The most up-to-date information will be available on NYU Brightspace.

Session 1: Introduction to class work, structure, and research projects

- Date: 01/26/2024
- Lecture:
    - Organizational details for class/housekeeping
    - Developing a research question
    - Project scoping

    - Basics of Theory of Change and Evaluation
- Readings:
    - Textbook Chapter 1
    - Project template
- Coding Application: Environment set up

Session 2: Developing an empirical research question

- Date: 02/02/2024
- Lecture:
    - The Challenge of Evidence-based Policymaking
    - Scoping a question
    - Finding data sources
    - Rescoping
- Readings:
    - Data Science and Public Policy Lab at Carnegie Mellon University. (2021, November 3). Data Science Project Scoping Guide.
    - Lane, J. (2010). Let's make science metrics more scientific. Nature, 464(7288), 488-489.
    - Lane, Julia. "A Vision for Democratizing Government Data." Issues in Science and Technology 39, no. 1 (Fall 2022): 84–88.
- Coding Application: Starter notebook

Session 3: Data exploration and management

- Date: 02/09/2024
- Lecture:
    - Data discovery and exploratory data analysis techniques
    - Understanding the data-generating process
    - Different data collection methods (traditional surveys, APIs, and web scraping)
    - Data types and structures
    - Relational databases and schemas
- Readings:
    - Behrens, J. T., & Yu, C. H. (2003). Exploratory data analysis. Handbook of psychology, 2, 33-64.
    - Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). Exploratory data analysis. In R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. essay, O'Reilly. Chapter 10.
    - Textbook Chapters 2 + 4
- Coding Application: Data Exploration Notebook

Session 4: Record Linkage

- Date: 02/16/2024
- Lecture:
    - Understanding the Problem
    - The conceptual framework
    - Deterministic approaches
    - Probabilistic approaches
    - Bias and ethics issues
- Readings:
    - Textbook Chapter 3

    - Chang, W.-Y., Garner, M., Basner, J., Weinberg, B., & Owen-Smith, J. (2022). A Linked Data Mosaic for Policy-Relevant Research on Science and Innovation: Value, Transparency, Rigor, and Community. Harvard Data Science Review, 4(2).

Session 5: Measurement

- Date: 02/23/2024
- Lecture:
    - Defining analytical datasets
    - Defining input measures
    - Defining output measures
- Readings:
    - Chapters 2 and 3 of Democratizing our Data
    - Bergstrom, C. T. (2017, June 12). Case study: NIH's rule of 21 - Part 1. Calling Bullshit.
    - Bergstrom, C. T. (2017, June 23). Case study: NIH's rule of 21 - Part 2. Calling Bullshit.
- Coding Application: Record Linkage Notebook

Session 6: Visualization

- Date: 03/01/2024
- Lecture:
    - Basics of visualization
    - Examples of successful visualizations
    - Applications (two notable uses for visualization: data exploration, and presentation)
- Readings:
    - Textbook Chapter 6

Session 7: Communicating Results

- Date: 03/08/2024
- Lecture:
    - Papers
    - Memos
    - Presentations
    - Guidance, tips, and tricks for communication
- Reading:
    - Review one of the presentations in Session 1 or 2 the Value of Science conference hosted by the National Center for Science and Engineering Statistics. You will be asked to comment on what you think was effective and what was not effective.
    - Review the Theory of Change in Appendix I - Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource. (2023, January). National Artificial Intelligence Research Resource Task Force. Retrieved from.
- Coding Application: Visualization Notebook

Session 8: Midterm project presentations

- Date: 03/15/2024
    - Students present current stage of their project
    - Students provide feedback on projects

    - **Note: No lab**
- Reading: N/A

03/22/2024 Spring Break: No classes

Session 9: Text Analysis
- Date: 03/29/2024
- Lecture:
    - Conceptual framework
    - Introduction to text analysis: Information retrieval, clustering and text categorization, text summarization
    - Learn how to implement topic modeling
    - Application to scientific fields
    - Evaluation
- Reading:
    - Chapter 8 of textbook

Session 10: Machine Learning Models I
- Date: 04/05/2024
- Lecture:
    - Formulate research questions in a machine learning framework: from transformation of raw data to feeding them into a model
    - How to build, evaluate, compare, and select models
    - How to reasonably and accurately interpret models
- Readings:
    - Chapter 7 of the textbook
    - Hao, J., & Ho, T. K. (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. Journal of Educational and Behavioral Statistics, 44(3), 348–361.
- Coding Application: Text Analysis Notebook

Session 11: Machine Learning Models II
- Date: 04/12/2024
- Lecture:
    - Supervised Machine Learning
    - Assessing model fit
- Readings:
    - Chapter 7 of the textbook
    - Athey, S. (2015, August). Machine learning and causal inference for policy evaluation. In Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining (pp. 5-6).
    - Agrawal, A., Gans, J., Goldfarb, A., & Athey, S. (2019). The Impact of Machine Learning on Economics. In The Economics of Artificial Intelligence: An agenda (pp. 507–547). essay, The University of Chicago Press.
    - Coyle, D. (2020, September 15). The tensions between explainable AI and good public policy. Brookings.
    - Hurley, D. (2018, January 2). Can an algorithm tell when kids are in danger? The New York Times Magazine.

Session 12: Biases, Fairness, and Inference
- Date: 04/19/2024

- Lecture:
    - Address biases in machine learning techniques and their consequences for public policy
    - How to deal with inference and the errors associated with big data
    - Problems of Big data and the errors resulting from it
- Readings:
    - Implicit bias test
    - Chapters 10 and 11 of the textbook
    - Ross, M.B., Glennon, B.M., Murciano-Goroff, R. et al. Women are credited less in science than men. Nature **608**, 135–145 (2022).
    - Dealing with Bias and Fairness in Data Science Systems: A Practical Hands-on Tutorial
- Coding Application: Machine Learning Notebook

Session 13: Privacy, Confidentiality, and Ethics in Research

- Date: 04/26/2024
- Lecture
    - Recognize where and understand why ethical and confidentiality issues can arise when applying analytics to policy problems
    - Plan, execute, and evaluate a research project along privacy concerns and ethical obligations
- Readings:
    - Chapter 12 of the textbook
    - Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (2014). Privacy, big data and the public good: Frameworks for engagement. Cambridge University Press.

Session 14: Final Project Presentations

- Date: 05/03/2024
- Lecture: N/A, students present final projects
- **Note: No lab**