Professor Howard T. Everson
MS Program in Data Analysis & Visualization
365 Fifth Avenue, Suite 3300
New York, NY 10016-4309
gc.cuny.edu

DATA 71200: Advanced Data Analysis
Course Syllabus Spring '26

Course Description[1]

This survey course is designed for students who want to extend their data analytic skills beyond a basic knowledge of multiple regression analysis, and who want to communicate their findings clearly to audiences of business groups, researchers, and other practitioners. The course will introduce students to advanced data analytic methods and toolkits, including machine learning methods using the sci-kit learning library, that will equip them with the ability to perform analyses of complex data from business, industry, and the arts and sciences. To succeed in the course, students should have a working knowledge of Statistics (through Regression methods) and at least a basic knowledge of the programming language *Python* to do the data analytic course work.

The course will begin with an overview of regression analyses, including logistic regression, and continues with latent variable analysis and related exploratory data analytic methods. This course will cover machine learning approaches: including both supervised methods (e.g., k-Nearest neighbors, naïve Bayes classifiers, decision trees, and support vector machines) and unsupervised methods (e.g., principal component analysis, non-negative matrix factorization, and k-means clustering). The supervised methods will focus primarily on "classic" machine learning techniques where features are designed rather than learned, although we briefly look at recent deep learning models with neural networks

The course will offer conceptual explanations of these statistical techniques and will provide opportunities for students to implement and practice these techniques using real data with the goal of helping students develop a sense of when machine learning is an appropriate tool versus other statistical modeling methods. Students will be encouraged to use Python and sci-kit learning tools to produce readable and sensible code that will enable others to replicate and extend their analyses.

Learning Objectives

The course is designed to provide students with a sound conceptual understanding of the collection of advanced statistical techniques and models and an introduction to the programming tools needed to implement these advanced statistical methods. At the end of this course, students should have:

- developed understanding of the statistical basis for machine learning methods
- articulate the importance of data preprocessing approaches

[1] Prof. Everson is a faculty member in the Data Visualization and Analysis Program, and a Visiting Scholar in the Educational Psychology Program at the Graduate Center CUNY. Dr. Everson also consults widely on statistical and psychometric issues with a variety of start-ups, schools, and other educational and governmental agencies throughout the U.S.

- articulate the main assumptions underlying machine learning approaches
- evaluate machine learning algorithms
- articulate the difference between supervised and unsupervised learning
- apply a range of supervised and unsupervised learning techniques

Approach to the Course

The emphasis throughout will be on the development of statistical thinking, i.e., thinking like a data scientist—by promoting a model-based understanding of statistics, not merely memorization of statistical formula or programmed procedures. Students will be encouraged to be curious—to make sense of the course materials and tools, and to ask questions. Much of the content of the course will be conveyed in various readings, textbooks, and interactive online textbooks**,** many of which include exercises and assessment questions. To succeed in this course, I expect all students to prepare by reading all the reading assignments, complete all the data analysis exercises, and submit and share work for review. Most importantly, I will encourage all students to take the time necessary to make sense of the readings and the data analyses assigned throughout the course. The "big" idea is to help students develop and hone their statistical reasoning skills.

This course will be taught online via Zoom. The weekly lectures & discussion sessions will focus on working with a variety of datasets with the goal of deepening students' understanding of statistical concepts and the connections among those ideas. In addition, the online sessions will address topics, concepts and methods students find challenging. Be sure you have your laptop at the ready so that you can follow along. **I** want to help you succeed in this course. If you are feeling confused or having difficulty keeping pace with the course, please schedule at time for us to talk. Because of the online nature of the course, I do not keep regular office hours. If you wish to speak with me, send me an email (howard.everson@gmail.com or Heverson@gc.cuny.edu) and I will find a time to meet with you.

Required Texts

The following books are required for this course: (1) *Regression Analysis* by J. Frost (e-Book); (2) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd Ed.) by T. Hastie, R. Tibshirani, & J. Freidman, 2017, Springer; (3) *Machine Learning for Social and Behavioral Research* by Jacobucci, Grimmm, and Zhang, 2023; and (4) *Introduction to Machine Learning with Python: A Guide for Data Scientists* by A. C. Muller & S. Guido (2017, O'Reilly Media). PDF versions of the *Regression Analysis* e-book and *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* book are in the class DropBox folder.

In addition, throughout the semester supplemental readings will be assigned—most, if not all, will be available as PDFs and accessible in the class DropBox Folder. (To claim your Dropbox account, please navigate to https://dropbox.cuny.edu/ and log in with your CUNY Login account Username: Firstname.Lastname##@login.cuny.edu). Please use the same email address for all communications related to the course. If you aren't sure which email address you have filed with the GC Registrar, you can log into CUNYFIRST and check the course list to find out. If you prefer to use a different email address, please send your alternate email address to Heverson@gc.cuny.edu or howard.everson@gmail.com.

Lectures

Lectures are not intended to provide new information, but to give you a chance to think deeply about and ask questions about - what you are learning in the online book. All lectures will be recorded via Zoom and the weekly recordings will be placed in the course DropBox folder. They recordings can be accessed on the day after the lecture.

Lectures and online activities will assume you have read the relevant texts, thus allowing for rich and lively class discussions. Material in the assigned readings will not always be repeated in class though they may be necessary for understanding the material presented in class. Students will be responsible for completing all reading before the class session for which they are assigned. If, however, there are topics or assigned readings that remain unclear, please let me know and we will discuss them in class. Lectures are designed to supplement the homework, not substitute for it. Lectures are not intended to provide new information, but to give you a chance to think deeply about—and ask questions about—what you are learning from reading the required books and the supplemental readings.

WEEKLY LECTURE TOPICS*

Class	Date	Topic	Readings
1	1/29	Introduction to Advanced Data Analysis	Review of Syllabus
2	2/5	Exploratory Data Analysis	Articles by J. Behrens (EDA) & D. Donoho (History of Data Science
3	2/12	PRESIDENTS' DAY NO CLASS	
4	2/19	Linear Regression & Correlation	Regression Analysis e-Book
5	2/26	Modeling Multivariate Relationships	Regression Analysis e-Book
6	3/5	Prediction & Explanation with Regression	Elements of Statistical Learning
7	3/12	PROFESSIONAL MEETING NO CLASS	
8	3/19	Bayesian Statistics Introduction	Supplemental Readings
9	3/26	Analysis of Categorical Data	Logistic Regression Readings
N/A	4/1-9	NO CLASS SPRING BREAK	
10	4/16	Introduction to Machine Learning	Mυller & Guido Book
11	4/21	How Machine Learning Works	ML Books & Readings
12	4/23	Supervised Learning	Mυller & Guido Book
13	4/30	Unsupervised Learning	Mυller & Guido Book
14	5/7	Classification Methods	ESL & ML Books
15	5/14	Machine Learning Practices	Jacobucci et al. Book
16	5/21	Course Review	ML Books & Readings
17	5/28	Final Exams Due	

* Note: *This is a tentative schedule of the topics and lectures for the Spring '26 semester. Weekly topics and dates are likely to be revised as we work through the course and the data analyses. Readings will be assigned during class each week, and most readings will be available in the course's DropBox file.*

Statistical Programs & Software, GitHub & Jupyter Notebooks

Data scientists and statisticians are increasingly using advanced software and tools, e.g., Python programming language, Jupyter notebooks, GitHub, and other A.I. tools. If you have an interest in accessing these tools, please bring up the question in class and I will arrange for you to learn more by using the *CourseKata* materials developed at UCLA.

Course Requirements: Course participation and homework assignments will comprise roughly 60% of the final grade, and the final exam will be worth 40% of the final grade. The course will not be graded on a curve. Each student's grade will be based solely on his or her own performance, not on the performance of others in the class. With that in mind, I encourage you to study with classmates and friends, if you choose to, you're all likely to learn more and improve your grades.

Homework: All homework assignments will be sent and submitted by email. Take-home exercises and readings will be assigned on a weekly and/or bi-weekly schedule. **You should expect the homework, including readings to take 6-8 hours (or more) per week, so start early in the week.** Homework will be graded for completion and correctness.

As you work through the readings and the exercises, be sure to write down questions about the ideas or methods you don't understand. **Bring these questions to lecture or office hours or send me an email.** It is **your** responsibility to make sure you're grasping the statistical ideas and methods of analysis you have encountered during the course.

Take-Home Exercises and Final Exam: The final exam will be distributed on or around May 18th and will be due on May 30th. All the take-home exercises and the final exam will be given online. These exercises and the final exam will include a mix problem sets and analytic results requiring tables, graphs, and narrative explanations of the data. The take home exercises are cumulative and can cover a range of topics—from the beginning of the course up through those lecture topics addressed in the course.

Incomplete Grades: If you cannot complete the semester's work and final exam, we will need to discuss a plan for ensuring you can complete the course requirements by the end of the Fall '23 semester. Per the GC's grading rules, INC grades turn into an F after two semesters. Awarding an INC grade will not be done without agreeing on a plan for completing the course requirements.

Special Accommodations: Every effort will be made to provide reasonable accommodations for students in need of them. The Graduate Center serves the needs of a growing number of students with special needs. For information on how to request these services, contact the Graduate Center's Office of Student Services. Such requests should be made during the first week of class, or as early as possible.

A Note on Plagiarism: Please consult the Graduate Center's Student Handbook, in particular the section on the CUNY Policy on Academic Integrity at www.gc.cuny.edu/current students/handbook/acadPol.htm. There is a helpful discussion on how to handle properly the work of others in your own writing in *Writing with Sources: A Guide for Students*, by Gordon Harvey, 1998 (ISBN: 0-877220-434-0). If you have any questions about what constitutes proper use of the work of others, please see me.

Office Hours & Available Statistical Consulting: Office hours will be held via Zoom by appointment only. If you wish to make an appointment, please e-mail me at howard.everson@gmail.com or Heverson@gc.cuny.edu or call or text me @ 917.520.6648. The Graduate Center also offers statistical consulting through the *Quantitative Research Consulting Center*. To arrange for consulting contact the QRCC at https://qrcc.commons.gc.cuny.edu/.