Math 446, Data Science with Python, Fall 2024

Exterior Course Website: http://www.stevenheilman.org/~heilman/446f24.html
Prerequisite: MATH 408 and 1 from (MATH 225 or MATH 235) and 1 from (ITP 115 or ITP 116)
Course Content: Python implementations of: data collection, data wrangling, exploratory data analysis, dimensionality reduction, unsupervised / supervised learning, clustering, classification, common predictive statistical / machine learning algorithms, model validation.

Lecture Meeting Time/Location: Mondays, Wednesdays, and Fridays, 1PM-150PM VHE 206
Instructor: Steven Heilman, stevenmheilman@gmail.com
Office Hours: TBD
TA: TBD, tbd@usc.edu
TA Office Hours: Kap 263 (the Math Center)
Discussion Session Meeting Time/Location:

- Thursdays, 4PM-450PM, GFS 222

Textbook: There is no required textbook. The first course resource is a freely available book: Python for Data Analysis, 3E by Wes McKinney, available online at: https://wesmckinney.com/book/.
Some other textbooks that might be helpful are:
A Hands-On Introduction to Data Science, by Chirag Shah
An Introduction to Statistical Learning, with Applications in Python by James, Witten, Hastie and Tibshirani. (Available online at: this link)

Software: Python is a freely available software. You should download and install this software on your personal computer. Specifically, dowload Anaconda (a popular Python distribution platform) from here: https://www.anaconda.com/download. Instructions for downloading and installing this software can be found: here. It might be helpful to bring a laptop to class with Python installed. Students who do not own a laptop may consider the USC Laptop Loaner Program: https://itservices.usc.edu/spaces/laptoploaner/.

We will most commonly be using the Jupyter notebook, within Anaconda.

Final Project Guidelines: The final project is an opportunity to work with a data set of your choice, apply some of the techniques we have discussed in class, and perhaps learn some new things we did not cover in class. A project could begin with an interesting question or a well-known problem, and perhaps lead to investigating or implementing various algorithms, conducting an empirical analysis, etc.

Along the way, you will review relevant literature, identify appropriate data sources, select appropriate means of evaluation, and either develop novel methodology for your problem or deploy and comprehensively evaluate existing methodology for your new application.

The goal is to say something interesting about a problem in data science, broadly construed. You could perhaps develop new methodology for an existing problem or application that has no fully satisfactory solution. You could alternatively tackle a new problem or application with existing

methodology; in this case, you should identify one or more questions without satisfactory answers in your chosen domain and explore how the methodology can help you answer those questions. You may draw inspiration from particular data sets, but your focus should rest not on the data itself but rather on the questions about the world that you can answer with that data.

While a substantial theoretical component is not required for this project, it could be beneficial if your project is supported by some theoretical results.

You may work alone or in a group of two; the standards for a group project will be twice as high. In certain cases I might approve a group of three, but this is unlikely.

We strongly encourage you to come to office hours to discuss your project ideas, progress, and difficulties.

Final Project Milestones: [Submission format TBD, but will probably be LaTeX]

I: Project Proposal [due Sep 26, 4PM]. By this first milestone, you should have selected a question or problem of interest, identified relevant data sources, begun exploring the literature surrounding the question, and discussed your ideas with the course staff. Your project proposal deliverable is a 1/2 - 1 page report (single spaced) describing the question or problem you intend to tackle, why this question is important or interesting, prior work on this problem, what data you intend to use in your analyses, and the principal challenges that you anticipate.

If you would like to receive feedback about particular aspects of your proposal, please indicate this in your submission.

I can try to help in problem selection. Ideally, the problem should be something you are very interested in. As such, it might be helpful to first tell me about your interests (maybe after class or in office hours), and we can try to think of something to work on. Selecting problems to work on is a difficult skill that takes years to develop, so it would be nice if you find a project idea on your own, but I expect everyone will need at least a little help in their choice. I know some things about some fields but I don't know everything about every field, so I might not be so helpful with certain projects outside my own background, but I can learn a bit myself to help you along if your interests are outside my knowledge.

II: Progress Report [due Oct 31, 4PM]. By this second milestone, you should have some initial results to share; for example, you may have implemented and evaluated the performance of existing algorithms on your dataset and task of interest, or you may have conducted an initial study with simulated data to better understand the properties of certain methods, or you were able to prove some preliminary result about some question of interest, etc.

Your progress report deliverable is a write-up of no more than 2 pages (single spaced) (not including references) describing what you have accomplished so far and, briefly, what you intend to do in the remainder of the term. You should be able to reuse at least part of the text of this milestone in your final report.

III: Pre-recorded presentation [due date TBD]. You will present your work in a pre-recorded video (it is easy to record a presentation using zoom, but I guess you don't have to use zoom). Depending on enrollment numbers, we might watch videos in class. The length of the presentation will vary

according to course enrollment, but each person should expect to speak for about 5-10 minutes. Since the talk will be short, you should consider practicing (and timing) your talk before recording it. You can practice part of it with me in office hours if you want. If we view the talks in class, expect around 3-5 minutes of questions from myself and your fellow students. Unlike other times in the class, attendance is mandatory during the presentations, and I will be taking attendance during them. Once I set the time limits they will be strict. Going over time will result in severe penalties.

You will be graded on your presentation skills, e.g. voice volume, screen/board usage, pacing of material, choice of material, etc. Minor technical problems will not be penalized, but major technical problems will be penalized. If you want to give a short version of your presentation in office hours before the actual presentation, and then have me give feedback that might be a good idea.

IV: Final Report [due Dec 18, 11AM]. Your final project report (not including acknowledgements and references) should be around 5-8 pages in length (single spaced) (using at most 12 point font and maximum 1 inch margins) and should follow a typical scientific style (with abstract, introduction, etc.). The write-up should clearly define your problem or question of interest, review relevant past work, and introduce and detail your approach. A comprehensive empirical evaluation should follow, along with an interpretation of your results. Any elucidation of the theoretical properties of an empirical method under consideration is also welcome.

If this work was done in collaboration with someone outside of the class (e.g., a professor), please describe their contributions in an acknowledgements section.

The final report PDF file should be submitted on brightspace. No hardcopy is needed.

Extra Credit Project: TBD

Email Policy:

- My email address for this course is stevenmheilman@gmail.com.
- It is your responsibility to make sure you are receiving emails from stevenmheilman@gmail.com, and they are not being sent to your spam folder.
- Do NOT email me with questions that can be answered from this document.

Exam Procedures: Students must bring their USCID cards to the midterms and to the final exam. Phones must be turned off. Cheating on an exam results in a score of zero on that exam. Exams can be regraded at most 15 days after the date of the exam. This policy extends to homeworks as well. All students are expected to be familiar with the USC Student Conduct Code. (See also here.)

Student Conduct: Discrimination, sexual assault, and harassment are not tolerated by the university. You are encouraged to report any incidents to the Office of Equity and Diversity http://equity.usc.edu/or to the Department of Public Safety http://capsnet.usc.edu/department/department-public-safety/online-forms/contact-us. This is important for the safety whole USC community. Another member of the university community - such as a friend, classmate, advisor, or faculty

member - can help initiate the report, or can initiate the report on behalf of another person. The Center for Women and Men http://www.usc.edu/student-affairs/cwm/ provides 24/7 confidential support, and the sexual assault resource center webpage sarc@usc.edu describes reporting options and other resources.

Accessibility Services: If you are registered with accessibility services, I would be happy to discuss this at the beginning of the course. Any student requesting accommodations based on a disability is required to register with Accessibility Services and Programs (OSAS) each semester. A letter of verification for approved accommodations can be obtained from OSAS. Please be sure the letter is delivered to me as early in the semester as possible. OSAS is located in 301 STU and is open 8:30am-5:00pm, Monday through Friday.
https://osas.usc.edu
213-740-0776 (phone)
213-740-6948 (TDD only)
213-740-8216 (fax)
OSASFrontDesk@usc.edu

Other Resources: An introduction to mathematical arguments

Homework Policy:

- Homeworks are due roughly every week, at **4 PM Thursdays**, i.e. at the beginning of the first discussion session on Thursdays.
- Homeworks are submitted in brightspace, under the "Assignments" tab. You are allowed unlimited submission "attempts" for an assignment, but only the last submission will be graded. To avoid internet issues, I recommend making your first submission of an assignment well in advance of the deadline. (Note that phone tethering can also give you an internet connection to a computer.)
- Homeworks should be submitted as single PDF documents. One way to create a PDF document from paper homework assignments is the freely available Adode Scan App.
- Late homework is not accepted.
- If you still want to turn in late homework, then the number of minutes late, divided by ten, will be deducted from the score. (The time estimate is not guaranteed to be accurate.)
- **Do not submit homework via email.**
- The **two lowest** homework scores will be dropped. This policy is meant to account for illnesses, emergencies, dropped internet connections, etc.
- You may not use the internet to try to find answers to homework problems.
- A random subset of the homework problems will be graded each week. However, it is strongly recommended that you try to complete the entire homework assignment.
- All homework assignments must be **written by you**, i.e. you cannot copy someone else's solution verbatim. However, collaboration on homeworks is allowed and encouraged.

- Homework solutions will be posted a few days after the homework is turned in.

Grading Policy:

- The final course grade is weighted as the larger of the following two schemes:
- Scheme 1: class participation (5%), homework (25%), the first midterm (15%), the second midterm (15%), project abstract (3%), project progress report (7%), final presentation (10%), final project report (20%).
- Scheme 2: class participation (5%), homework (25%), largest midterm grade (20%), project abstract (3%), project progress report (7%), final presentation (15%), final project report (25%).
- The grade for the semester will be curved. However, I do not "curve down" since anyone who exceeds my expectations in the class by showing A-level performance on the exams and homeworks will receive an A for the class.
- If you cannot attend one of the exams, you must notify me within the first two weeks of the start of the quarter. Later requests for rescheduling will most likely be denied.
- Class participation is not the same as attendance. I will never explicitly take attendance, but I will notice if someone is frequently absent. Things that increase your class participation grade include: asking good questions, paying attention in class, showing up on time or early to class, etc. Things that decrease your class participation grade include: excessive talking or disruptions during class, frequent absences, excessive texting/smartphone usage in class, frequent tardiness, etc.
- You must take the final exam to pass the course.

Tentative Schedule: (This schedule may change slightly during the course.)

Week	Monday	Tuesday	Wednesday	Thursday	Friday
1	Aug 26: Intro to Jupyter Notebook	Aug 27	Aug 28: Review of Python	Aug 29:	Aug 30: Review of Python
2	Sep 2: No class	Sep 3	Sep 4: Intro to Numpy	Sep 5: Homework 1 due	Sep 6: Numpy and Floating Point Arithmetic
3	Sep 9: Estimation and Numpy	Sep 10	Sep 11: Review of Linear Algebra	Sep 12: Homework 2 due	Sep 13: Review of Linear Algebra
4	Sep 16: Least Squares Minimization	Sep 17	Sep 18: Singular Value Decomposition	Sep 19: Homework 3 due	Sep 20: Principal Component Analysis
5	Sep 23: k-means Clustering, Scikit-learn, Matplotlib	Sep 24	Sep 25: Dimension Reduction, Johnson-Lindenstrauss	Sep 26: Project proposal (abstract) due	Sep 27: Intro to Pandas
6	Sep 30: Intro to Pandas	Oct 1	Oct 2: Exam 1	Oct 3: No homework due	Oct 4: Data Loading, Storage, File Formats
7	Oct 7: JSON, CSV	Oct 8	Oct 9: Web Scraping	Oct 10: Homework 4 due. No class	Oct 11: No class
8	Oct 14: Data Cleaning, Preparation	Oct 15	Oct 16: Data Cleaning, Preparation	Oct 17: Homework 5 due	Oct 18: Data Wrangling
9	Oct 21: Data Wrangling	Oct 22	Oct 23: Exploratory Data Analysis	Oct 24: Homework 6 due	Oct 25: Classification
10	Oct 28: Classification	Oct 29	Oct 30: Regression	Oct 31: Progress report due	Nov 1: Regression
11	Nov 4: 8.4, Multiclass classification	Nov 5	Nov 6: Multiclass classification	Nov 7: No homework due	Nov 8: Exam 2
12	Nov 11: No class	Nov 12	Nov 13: Deep Learning, keras	Nov 14: Homework 7 due	Nov 15: No class
13	Nov 18: Deep Learning, keras	Nov 19	Nov 20: Deep Learning, keras	Nov 21: Homework 8	Nov 22: Deep Learning, keras
14	Nov 25: Leeway	Nov 26	Nov 27: No class	Nov 28: No class	Nov 29: No class
15	Dec 2: Final Presentations	Dec 3	Dec 4: Final Presentations	Dec 5: Homework 9 due	Dec 6: Final Presentations [Dec 18: Final Report Due]

Advice on succeeding in a math class:

- Review the relevant course material **before** you come to lecture. Consider reviewing course material a week or two before the semester starts.
- When reading mathematics, use a pencil and paper to sketch the calculations that are performed by the author.

- Come to class with questions, so you can get more out of the lecture. Also, finish your homework at least **two days** before it is due, to alleviate deadline stress.
- Write a rough draft and a separate final draft for your homework. This procedure will help you catch mistakes. Also, I would very much recommend typesetting your homework. Learning LaTeX is a very important skill to have for doing mathematics. Here is a template .tex file if you want to get started typesetting.
- If you are having difficulty with the material or a particular homework problem, review Polya's Problem Solving Strategies, and come to office hours.