Adriane Fresh
Assistant Research Professor
Office Hours: TBD

IDS 590
Fall 2025

Practical Data Science I

1 Course Description

Practical Data Science I is a flipped-classroom, exercise and project-focused course. It requires zero prior experience with programming and begins with an introduction to Python, computational thinking, and the principles of good programming using the 7 Steps method. The class focus then shifts to data analysis with an emphasis on the type of analyses of interest to social scientists and public policy students. The course provides students with experience manipulating and analyzing real (often messy, error-ridden, and poorly documented) data using the full range of Python data science tools (like the command line, git, VS Code, numpy, pandas, matplotlib, statsmodels, and more).

2 For Whom Is This Course Meant?

This class is for anyone interested in answering questions about the world using data with limited programming experience. Social scientists, public policy students, biostatisticians, statisticians, you name it!

The ideal student is someone who has taken some courses in statistical inference, and has some experience working with data in a language like R or Stata, but wants to dramatically improve their ability to manipulate data and employ more sophisticated data science tools to answer questions that matter to them.

2.1 For Whom Would This Course Be Inappropriate?

If you were a computer science major as an undergraduate or worked in a job that made intense use of Python for Data Science applications, please speak to me after class, as the first portion of this course might be somewhat boring for you. With that said, even students who have taken computer science courses may find that this class offers a very different perspective on familiar tools. CS programs tend to be oriented towards a style of programming best suited for software development which can differ substantially from the tools and style used in data science.

3 What Do You Mean By Data Science?

There are, broadly speaking, two branches of what is often referred to as Data Science, which I will term *Software Development Data Science* and *Data Analysis Data Science*.

In *Software Development Data Science*, programmers write programs that gets bundled up in software and distributed widely, or gets run on the cloud for millions of people. For example, software development data scientists wrote the recommendation engine that lets Netflix tell you what movies you might enjoy, or what people might be your friends on Facebook. As a result, they generally write *generalizable* code that is designed to run on data with a known structure.

In *Data Analysis Data Science*, the data scientist is generally employed to answer a single, specific question. For example, a Data Analysis Data Scientist may be hired to figure out how to reduce anti-biotic resistant infections in a hospital, or to identify what campaign promises are most likely to convince voters to support a politician. As a result, Data Analysis Data Scientists are generally writing code that is only meant to be used for their specific project. Moreover, Data Analysis Data Scientists don't generally have the luxury of working with data with a known structure – where a Netflix Data Scientist may get data from a company database that's clean and well organized, a Data Analysis Data Scientist may have to work with data that has come from lots of different sources and which no one has cleaned and organized (e.g. notes from nurses, or voting data from different states compiled by hand by minimum wage government employees).

To be clear, these branches are not completely distinct. Most data scientists do things that fall into both categories (for example, even a Software Developer will likely do some *ad hoc* analyses before developing a fully deployable tool). However these two types of data science do emphasize different skills. Software Development Data Scientists, for example, are well served by traditional computer science curricula and need a much deeper understanding of concepts like object-oriented programming, and software deployment. By contrast, Data Analysis Data Scientists need to be comfortable working with data in different formats and to understand how to clean and fit together datasets that were never actually built to be integrated.

The focus of this course will be on the skills of Data Analysis Data Science: cleaning and merging data, data exploration, and designing projects to answer very specific questions. If you're interested in policy analysis, or health-sector analysis, or applied empirical research, this course is for you; if you're interested in developing programs you can deploy in an iPhone app to improve recommendations, then while there will be material that will be of use to you (the Python data science stack, working at the command line, git and github), the emphasis of the material won't quite be what you're looking for.

4 Python

In this class we will primarily be working with Python.

Why Python? Because it's currently one of the two most-used programs in data science (the other being R, which you'll be working with in other classes), which means there is a good chance you'll be called upon to use it when working in teams.

It is worth emphasizing that we're not learning Python because it is necessarily the "the best" language. The reality is that there are *lots* of tools for statistical programming, and each has its own strengths and weaknesses (e.g. R, Stata, SPSS, Python, Julia, Matlab, etc., etc.). People often develop strong opinions about which language is *best*, and sometimes pass judgement on people who use other languages. Every programming language has its strengths and weaknesses, and what is "best" depends on your use-case (the types of things you are using the language to

do). This is true not only because languages themselves have strengths and weaknesses, but also because the tools and packages that have been created for use in different languages differ (e.g. people just haven't made a good package for doing geo-spatial work in Julia yet, for example). And if you're working on teams, you'll also have to make decisions based on the backgrounds of your tool sets. All of which is to say: there is no single *best* language for all purposes. But Python is a very popular, strong, general purpose language, so will serve as a great starting point.

As a result, over the course of your career you may find yourself gravitating to one tool or another as required by your research. But in providing you with a firm foundation in a very popular language like Python, you will not only be learning a tool that will allow you to do most everything you'll want to do in graduate school, but you will also be providing yourself with a solid foundation in *generalizable* skills that you will find useful if you later change platforms.

5 Class Organization

Data science is an applied discipline, and so this will be an intensely applied class with *lots* of hands-on exercises.

To make it possible for us to work through problems together as they arise, we will dedicate most of our class time to completing these exercises in small groups. That means that students will be required to read instructional material *before every class* so they will be ready to do these exercises. This is what is referred to as "flipping the classroom."

In order to make this class organization work, it will be ***critically*** important that students do their assigned readings before *every* class, and as discussed below, this will be reflected in how grades are assigned in this class. Students who do not complete their assigned readings and tutorials before each class should not expect to receive good grades, regardless of performance on project assignments.

This class is organized around having two (synchronous) class sessions every week. While the plan is for most of these will be in person, some classes will inevitably end up needing to be held online. **Synchronous attendance, whether classes online or in person, is required unless you are unable to participate synchronously due to extenuating circumstances (such as an internet connection that will not support synchronous participation (for online classes) or illness (for in person classes)).**

With that said, everyone's health and safety is of course our first priority, so while it is very important you attend class whenever possible, you should **never** hesitate to stay home if you're not feeling well. If you are not feeling well and need to miss class – or need to miss class for covid related reasons (e.g. quarantine) – please reach out to me **before class** so that we can make a plan to make sure you're fully supported!

6 Assignments & Grading

6.1 Participation (20% of Grade)

Note that a major component of good participation is good *preparation*. Because we will mostly reserve class time for hands-on exercises, it is absolutely critical that students do their assigned

readings before *every* class. Students who do not work through the instructional materials they have been assigned before class will not only get very little out of the hands-on exercises designed to reinforce the assigned materials, but they will also undermine the learning of the students they are asked to work with. With that in mind, students who do not complete their assigned readings before every class should expect to see this reflected in their participation grades.

Participation will be graded as follows:[1]

A range. You are fully *and consistently* engaged in class discussion and exercises. You both listen and contribute actively. You are well-prepared for class. Having done more than merely read the material, you have spent time thinking *carefully and deeply* about the material's relationship to other materials and ideas presented in previous classes. You are not only able to answer questions about the material, but also come to class with thoughtful questions. When working in teams, you work *with* your partner. If your partner is struggling with an exercise, you help them understand the material rather than just completing the material on your own. If you are struggling with material, you ask for help (both from the instructor and your fellow students) and do not simply lean on your partner to complete the exercise.

B range. You are engaged in class discussion and exercises. You listen and contribute regularly. You come well-prepared to class having read the material and your contributions show your familiarity, but your level of engagement lacks the depth accumulated through extra time spent thinking about the material. When working in teams, you work *with* your partner when they have a similar level of understanding, but do not always invest in helping a struggling partner to understand the material. You often ask for help when you are struggling, but other times you let your partner just complete the exercise.

C range. You have met the minimum requirements of participation. You are usually, but not always prepared. You participate sometimes, but not regularly. The comments that you offer show a basic familiarity with the materials, but do not help to build a coherent or productive discussion. When working in teams, you only sometimes work *with* your partner. When your partner is struggling, you often just do the exercise yourself. If you are struggling, you often do not ask for help and allow your partner to take over the exercise.

D range. You have not met the minimum requirements of participation. You are unprepared for class. You have not read with the material with sufficient engagement to know even the most basic elements. When working in teams, you do not attempt to work *with* your partner. When your partner is struggling, you just do the exercise yourself. If you are struggling, you do not ask for help and allow your partner to take over the exercise.

As should be clear from this rubric, above all it is important to emphasize that participation is evaluated on the basis of *quality* and *consistently*, *not* quantity. Moreover, when completing in-class exercises, good participation is not about finishing first or without ever asking for help; good participation in in-class exercises is about helping your partner understand the material, and asking for help when you need it.

[1] This rubric is adapted from that of Duke Political Science Professor Adriane Fresh.

6.2 Quizzes (20%)

To ensure students are doing their readings in advance of class, from time to time we will start class with short quizzes. These quizzes are designed to be relatively straightforward if you did the readings—they won't be full of gotcha questions—but will require you to have done the readings.

6.3 Exercises (35% of Grade)

Over the course of the semester, students will be asked to complete a number of small assignments as homework. These assignments will, in total, be worth 35% of student grades. **Note:** because of the way students get autograder feedback before final submission and because the lowest exercise grade gets dropped, grades on exercises tend to converge towards 100%. In light of that, I will generally adjust exercise scores down 5 percentage points when doing grade calculations so that getting all available points on an exercise (reported as 100% on gradescope) maps to a 95% solid A (but not a 100%).

6.4 Team Data Science Project (25% of Grade)

Around mid-semester, students will be assigned a large team Data Science Project. The goal and general framework for this team project will be provided to students, but the project will require students to complete the analysis component of a full data science project, including gathering data, cleaning and merging that data, analyzing the data, and presenting results.

6.5 Late Assignments, Make Up Exams and Extra Credit

Grading

All assignments will be given a numerical score on a 0-1 scale. These scores will be multiplied by the value of the assignment (see above) and the following scale will be used to assign a final letter grade.

Late Assignment

All late assignments will be penalized 10% per day the assignment is late, up to a maximum penalty of 50%.

The final deadline for accepting assignments that are more than one week late is at the discretion of the instructor and may vary by assignment.

Exceptions to these late penalties may be made for students dealing with exceptional circumstances (illness for themselves or family, etc.) — if you are dealing with a difficult situation, please feel free to contact me to discuss your situation.

Dropping Lowest Scores

To accommodate the fact that life happens, at the end of the semester, I will drop each student's lowest Quiz *and* lowest Exercise **that have been completed.** Essentially, this is a free pass for one exercise and one quiz you totally whiff, submit very late. But it is *not* a free pass to *skip* an Exercise or Quiz — uncompleted Exercises or Quizzes are not eligible for being dropped.

Quizzes and Absentees

As detailed above, attendance in class is always required barring injury, illness, or other significant conflict.

To ensure fairness:

- students who are not present in class will not be allowed to make up quizzes
- except if the student has reached out to me **before** class to alert me to their absence and the reason for it. If the reason is it deemed to be grounds for an excused absence, the student will be able to take the quiz at a different time.
- students who are not physically in class are also **NOT** allowed to take quizzes remotely without express permission. If you are found to have taken a quiz when not in class, not only will you receive a zero on that quiz, but it may be considered an honor code violation. Note that because classes are recorded (for students with excused absences) and gradescope tracks IP addresses, this is not a hard thing to figure out.

7 Texts

We will be working almost entirely with the material at practicaldatascience.org. So you don't have to buy anything!

8 Honor Policy

Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and nonacademic endeavors, and to protect and promote a culture of integrity.

Remember the Duke Community Standard that you have agreed to abide by:

- I will not lie, cheat, or steal in my academic endeavors;
- I will conduct myself honorably in all my endeavors; and
- I will act if the Standard is compromised.

Cheating on exams or plagiarism on homework assignments, lying about an illness or absence and other forms of academic dishonesty are a breach of trust with classmates and faculty, violate the Duke Community Standard, and will not be tolerated. Such incidences will result in a 0 grade for all parties involved. Additionally, there may be penalties to your final class grade along with being reported to the MIDS program directors.

9 Disability Statement

In an effort to prevent students with disabilities from having to explain and justify their condition separately to each of their various instructors, Duke has centralized disability management in the Student Disabilities Access Office. If you think there is a possibility you may need an accommodation during this course, please reach out to their office as soon as possible (processing can take a little time).

Medical information shared with the SDAO are strictly confidential, and if SDAO determines an accommodation is appropriate, faculty members will simply be informed of the accommodation they are required to provide, not the underlying medical reason for the accommodation.

If you have any problems with SDAO, please let me know as soon as possible.

10 Student Signature

I, the undersigned, confirm I have read and understand the expectations of this class.

Name: _____

Signature: _____

Date: _____

I, the undersigned, confirm I have also read and understand the chatGPT and You reading.

Name: _____

Signature: _____

Date: _____