Econ 370: Data Science for Economic Analysis

Department of Economics
Williams College
Fall 2024

1 Contact Information and Course Logistics

1.1 Professor: Dr. Pamela Jakiela

Email: pj5@williams.edu

1.2 Course Meetings

Wednesdays, Fridays from 8:30 to 9:45 AM in 129 Schapiro

1.3 Course Websites

https://pjakiela.github.io/ECON370/

Course materials are available on the public website for the course: https://pjakiela.github.io/ECON370/. That is where I will post lecture slides, information about assignments, class schedules, and other course information.

https://www.gradescope.com/courses/854937

All assignments will be submitted through gradescope. You'll receive an email inviting you to join the course, and that will allow you to set up a gradescope account linked to your Williams email address.

1.4 Office Hours

My drop-in office hours are on Mondays from 1:30 to 3:30 in 339 Schapiro.

If you have a course that conflicts with my scheduled office hours, you can email me for the link to my one-on-one appointment schedule.

1.5 Communication

Email is the best way to contact me (pj5@williams.edu). I will try to respond to course-related emails withing two working days of receipt.[1] I am sometimes overwhelmed by the volume of email I receive; if I have not responded to you within two working days, please feel free to send me an email reminder.

[1] So, if you send me an email on Tuesday at noon, I will try to respond by the end of the day on Thursday. I do not check my email on weekends. If you send me an email on Friday afternoon, I will respond to it by the end of the day on Tuesday.

2 Course Description

2.1 From the Course Catalog

This course provides a hands-on introduction to data science tools most relevant for economic analysis including data visualization, machine learning, and text analysis. Economists and other social scientists tend to use these data science tools differently than many researchers in statistics and computer science - conducting empirical analysis that is explicitly grounded in economic theory, and focusing on causal inference rather than prediction. Through a combination of lectures, hands-on labs, and group projects, students will develop the theoretical and practical skills needed to analyze economic data using modern data science techniques in both R and Python.

2.2 Learning Objectives

I have three main objectives in teaching ECON 370:

1. To help students develop advanced data skills, with a particular focus on identifying new sources of data, data cleaning/wrangling, and visualization;
2. To help students hone their ability to master new empirical tools (e.g. programming languages, statistical methods) independently to keep their skills current; and
3. To introduce students to new statistical tools (e.g. machine learning, text analysis) which are increasingly being incorporated into the empirical economist's toolbox.

2.3 Who Should Take ECON 370?

ECON 370 is an advanced elective intended for economics majors who have completed the intermediate level courses. ECON 255 (or its equivalent in the statistics department, STAT 346) is a required prerequisite.

3 Tentative Class Schedule and Important Dates

Student presentations will take place on Wednesday September 27, Wednesday October 23, Friday October 25, Wednesday December 4, and Friday December 6. **If you have a foreseeable conflict with any of those dates, you should not take the course.**

Date	Description
9/6	Introduction
9/11	Data
9/13	Exploratory Data Analysis
9/18	Data Visualization
9/20	Guest Speaker: Bilal Zia, Head of Data Science & Analytics, Duolingo
9/25	*Data Visualization Group Meetings*
9/27	**Data Visualization Project Presentations**
10/2	Numeric Approaches to OLS
10/4	Cross-Validation
10/9	Subset Selection, Regularization, and Lasso
10/11	Regression Trees and Random Forests
10/16	*Predicting Infant Mortality Group Meetings*
10/18	*Mountain Day (possibly earlier)*
10/23	**Predicting Infant Mortality Presentations**
10/25	**Predicting Infant Mortality Presentations**
10/30	Causal Forests
11/1	k-Means Clustering
11/6	Intro to Text as Data, Regular Expressions
11/8	Web Scraping
11/13	Word Frequencies, Document-Term Matrices
11/15	Clustering Documents
11/20	Topic Models
11/22	*Final Project Group Meetings*
11/27	*Thanksgiving Break*
11/29	*Thanksgiving Break*
12/4	**Final Project Presentations**
12/6	**Final Project Presentations**

Italics indicates no regularly schedule class meeting.

Bold indicates student presentations during class meeting time.

4 Readings

Data science is a new and rapidly evolving field, and there is no single textbook that is appropriate for the course. We will make use of several reference texts, all of which are freely available online. An up-to-date list of useful references is available at `https://pjakiela.github.io/ECON370/references.html`. Links to recommended readings associated with each course module will be posted on the course website.

You should download a pdf copy of *An Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani. I recommend downloading the second edition of the R version, though students with prior Python experience are welcome to use the Python version instead (the two versions are identical except for the labs, which we will not be using). Both versions are available at `https://www.statlearning.com/`.

You should also bookmark the web pages for the following reference books:

- *Fundamentals of Data Visualization* by Claus O. Wilke, available at `https://clauswilke.com/dataviz/`
- *ggplot2: Elegant Graphics for Data Analysis (3e)* by Hadley Wickham, available at `https://ggplot2-book.org/`
- *R for Data Science (2e)* by Hadley Wickham, Mine Cetinkaya-Rundel, and Garrett Grolemund, available at `https://r4ds.hadley.nz/`

5 Assignments and Grading

Grades are calculated as follows:

Lab Assignments	45 points
Data Visualization Project	12 points
DHS Prediction Project	12 points
Final Empirical Project	21 points
Class Participation	9 points
Getting-to-Know-You Survey	1 point

5.1 Labs

Each class will include a lab component that you will complete in R or Python (or, in one case, both). You will submit your code and data files via gradescope, and you may also be asked to answer some questions (in gradescope) about your results.

Each lab is worth three percent of your final grade. To receive full credit for a lab, your code **must** run from start to finish and generate correct results (though I will allow each student one file path and/or package related failure). Scripts and programs that do not run will not receive (any) credit. Your code should be clear, concise, and explained through detailed comments. Try to avoid obscure packages that may not be maintained.

5.2 Group Projects

Over the course of the semester, you will complete three empirical projects and present your results to the class. For each of these projects, you will need to submit detailed replication files that allow me reproduce all the empirical results included in your presentation. Your replication code needs to run from start to finish; it should be clear, concise, and explained through detailed comments.

5.2.1 Data Visualization Project

Early in the semester, you will work in groups to complete a data visualization project. Your project will help a local small business better understand their transaction data. The data visualization project accounts for 12 percent of your final grade. Your grade will depend on your attendance at a meeting with me to discuss your project (2 points) and the overall execution of the project in terms of originality, the quality of your visualizations, your class presentation, and the code and other supporting materials underlying your presentation (all of which together account for 10 percent of your final grade).

5.2.2 Predicting Infant Mortality Project

Later in the semester, you will work in pairs to complete an empirical project using machine learning to predict infant mortality in a low- or middle-income country. Your grade will depend on your attendance at a meeting with me to discuss your project (2 points) and the overall execution of the project, your class presentation, and the code and other supporting materials underlying your presentation (all of which together account for 10 percent of your final grade).

5.2.3 Final Empirical Project

In the second half of the semester, you will complete a group project on a topic of your choosing that you will present in the last week of class. Your grade will depend on the quality of your research proposal (4 points), your attendance at a meeting with me to discuss your project (2 points), your presentation (10 points), and the code and other supporting materials underlying your presentation (5 points).

5.3 Class Participation

Active, constructive participation in class meetings is a critical part of the course. During each class, I will be assigning you a participation score on a scale of 0 to 3, as follows:

- 3: Present in class, paying attention, fully participating, and making constructive comments and/or asking thoughtful questions
- 2: Present in class, fully engaged, limited contributions to class/lab discussion
- 1: Physically present but not fully engaged
- 0: Not present

I do not expect students to have perfect attendance: if you need to miss class once or twice during the semester, you do not need to seek approval from me in advance (though you are welcome to alert me if you wish). At the end of the semester, I will drop your two lowest class participation scores. If you expect to miss more than two classes, or you encounter challenges (such as illness) that prevent you from attending class consistently, you should discuss these issues with me in office hours or over email. Regular absences will result in a loss of class participation credit unless they are discussed with and approved by me in a timely manner.

5.4 Getting-to-Know-You Survey

After the first class meeting, you will be invited to take a brief survey (through a google form) to provide me with more information about you.

5.5 Late Assignments

Unless otherwise stated, all (unexcused) late assignments will be penalized: the maximum grade will be lowered by 10 percent for every day late for the first five days (including weekends). Assignments can only be submitted more than five days late with permission from the instructor.

6 Honor Code

The English language content of your labs and presentation slides should reflect your own work (or the work of the members of your group, in the case of group projects). To avoid violating the honor code, any English text written or prepared by someone else must be identified as such and cited appropriately. Any assignment containing either an image/figure/graph or more than five consecutive words taken from another source (e.g. a published paper or a website) without attribution will automatically receive a zero.

You will also receive no credit for graphs or tables that are identical or nearly identical to existing work (e.g. graphs from published papers or those posted on blogs), even if you submit code that generates the table or figure from the relevant raw data files. This means that you will not receive credit if you submit someone else's replication file(s) as part of an empirical project or lab. If I find a table or figure that you submit posted elsewhere, you will automatically receive a zero for the assignment with no opportunity for renegotiation.

6.1 ChatGPT, etc.

You **are** allowed to use ChatGPT and other generative AI to assist with your coding. Generative AI is a powerful tool, and you should learn how to use it. On each assignment, I will ask you to describe your use of generative AI, but your use of such tools has no direct impact on your grade.

To avoid inadvertently duplicating others' work, I suggest that you ask chat to translate or debug small sections of code (based on your own pseudocode). You are always responsible for submitting code that is correct, well-commented, etc., and for making sure that your

work is original. In addition, I reserve the right to ask you to explain any code that you submit; if you are unable to do so, you will not receive credit for the assignment.

7 Health and Accessibility Resources

Students with disabilities or disabling conditions who experience barriers in this course are encouraged to contact me to discuss options for access and full course participation. The Office of Accessible Education is also available to facilitate the removal of barriers and to ensure access and reasonable accommodations. Students with documented disabilities or disabling conditions of any kind who may need accommodations for this course or who have questions about appropriate resources are encouraged to contact the Office of Accessible Education at `oaestaff@williams.edu`.

8 Classroom Culture

The Williams community embraces diversity of age, background, beliefs, ethnicity, gender, gender identity, gender expression, national origin, religious affiliation, sexual orientation, and other visible and non visible categories. I welcome all students in this course and expect that all students contribute to a respectful, welcoming and inclusive environment. If you have any concerns about classroom climate, please come to me to share your concern.

9 Concluding Remarks

Congratulations on making it to the end of the syllabus. Do professors enjoy writing syllabi? Not really. Do students enjoy reading them? Probably not. Hopefully, this document will provide us with a shared set of expectations for the semester, making the course more constructive and enjoyable for everyone. Also, if you send me an email containing a "bad" data visualization together with a sentence explaining why it is bad before September 13 (subject line: ECON 370 bad data viz), you will earn a point of extra credit.