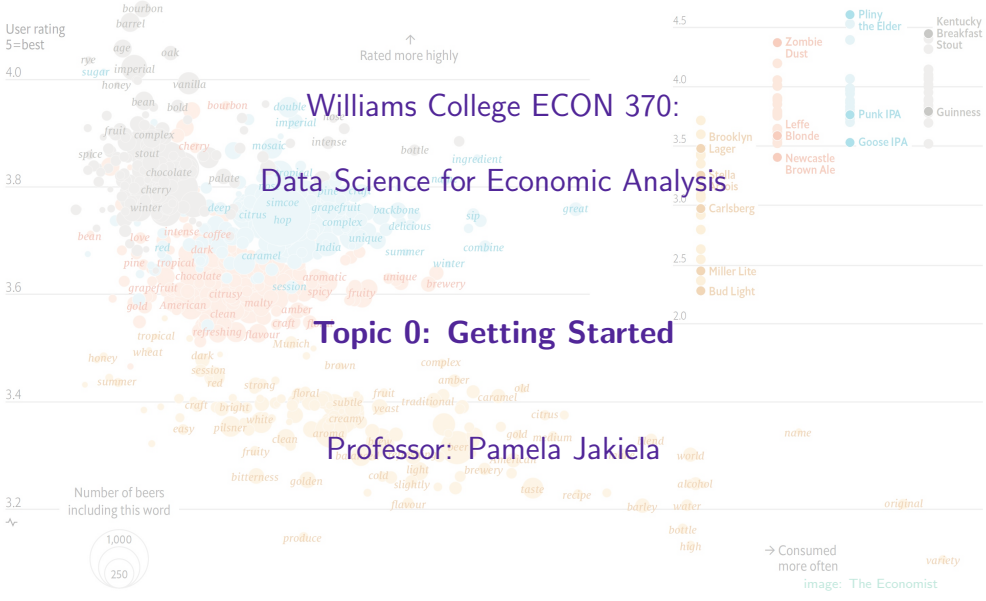


Williams College ECON 370:

Data Science for Economic Analysis

Topic 0: Getting Started

Professor: Pamela Jakiela

Outline

- Introductions
- Syllabus
- Next steps

ECON 370: Data Science for Economic Analysis

- Class meetings: MWF from 11:00 AM to 12:15 PM in 129 Schapiro
- Instructor: Professor Pamela Jakiela (`pj5@williams.edu`
  - Office hours: Mondays from 2:00 to 3:00, Thursdays from 3:00 to 4:00 in 339 Schapiro
- Teaching Assistant: Lewis Schrock (`lps3@williams.edu`), office hours TBD
- Prerequisites: ECON 255 or STAT 346, ideally one other economics elective using data
- Course websites:
  - `https://pjakiela.github.io/ECON370/` (course materials)
  - `https://www.gradescope.com/courses/1105402` (submitting assignments)

Learning Objectives

- Developing advanced data skills, with a particular focus on identifying data new sources, data cleaning and wrangling, exploratory data analysis, and data visualization
- Learning to implement new statistical methods (e.g. machine learning, text analysis) which are increasingly being incorporated into the empirical economist's toolbox, and learning how these tools are used by economists for prediction and causal inference
- Honing the ability to master new empirical tools such as programming languages (e.g. R, Python) and statistical methods (e.g. machine learning, text analysis) as they emerge

TLDR: how to be a super RA and/or be ready for grad school in quantitative social science

What This Course Is About

- Working with data: finding data, cleaning and wrangling data, data visualization, exploratory data analysis, dimension reduction and unsupervised learning techniques
- The machine/statistical learning tools used (the most) by economics:
  - Shrinkage methods, particularly lasso, and the ML approach to covariate selection
  - Using random forests and other ML techniques to capture treatment effect heterogeneity
- Text analysis: working with text, (simple) web scraping, word frequencies, topic models

What This Course Is NOT About

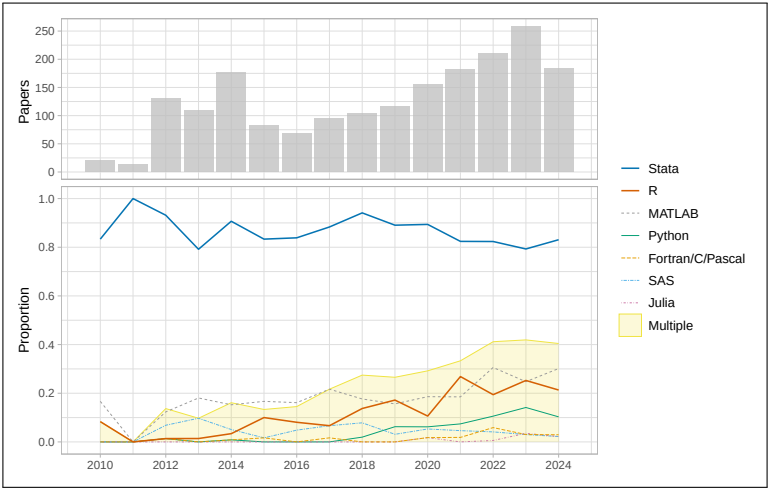
- Big data
- Computation
- Database management
- Data science methods that are focused exclusively on prediction (see STAT/CSCI classes)
- Traditional econometric approaches to causal inference (see ECON 379 and ECON 474)
- Machine learning (or other statistical) techniques that are not used by economists

Stata vs. R vs. Python

ECON 255 is taught in Stata, and Stata is still by far the most widely used in by economists

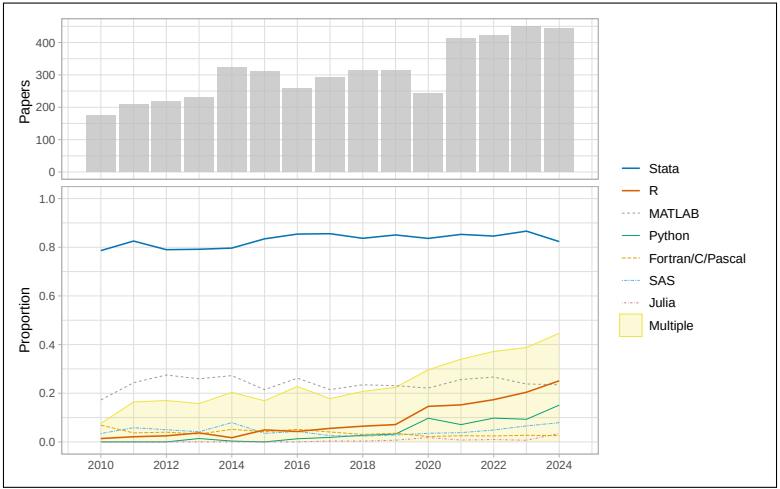
- Many of the techniques we'll cover in this class can't be done in Stata (at present)
- Economics is shifting away from Stata, albeit slowly; not yet clear where we are heading
- Being able to get yourself up-to-speed in a new language is a (data science) skill in itself
- Familiarity with R or Python is not a prerequisite (as long as you know Stata)

Stata vs. R vs. Python: Data from QJE and REStat



Source: Upton, Cai, Jakiela, Ozier, and Raman (2025)

Stata vs. R vs. Python: Data from AEA Journals



Source: Upton, Cai, Jakiela, Ozier, and Raman (2025)

R vs. Python

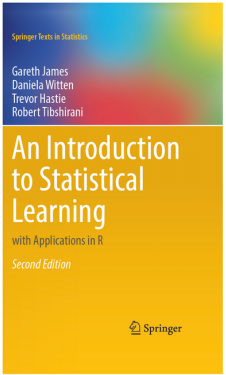
I plan to teach the course primarily in R, to the extent that I focus on a particular language

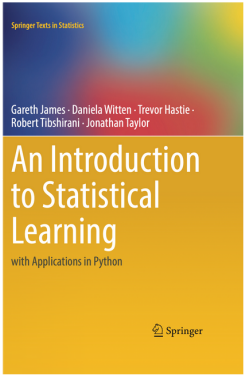
- R is (at present) the most widely used alternative to Stata in economics
- R is the dominant approach to statistical computing in statistics and political science
- R has better tools for data visualization (ggplot2 vs. matplotlib)
- Some economics-specific machine learning tools are only available in R
- But... Python has many, many uses beyond statistical computing

You should decide whether you want to focus your energy on R or Python, either is fine

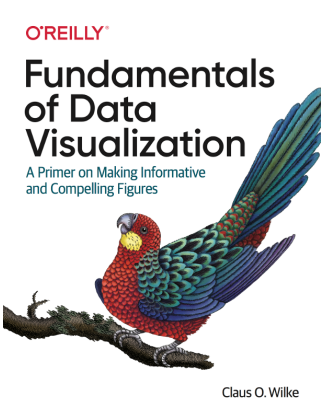
- I would like you to end the course (somewhat) familiar with both languages
- Lab 1 is in both R and Python (you will submit two versions)
- Labs 2 through 9 can be done in R or Python, with some exceptions

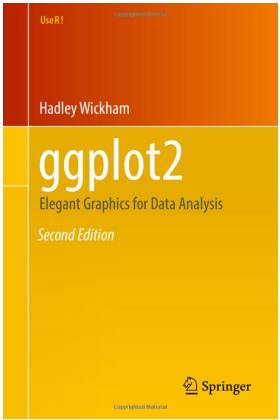
Readings: Download This

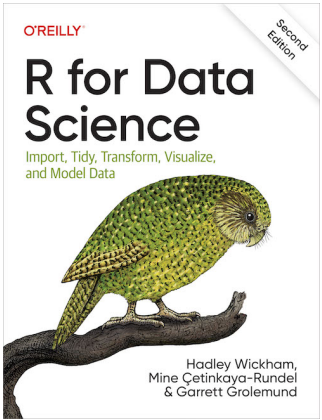




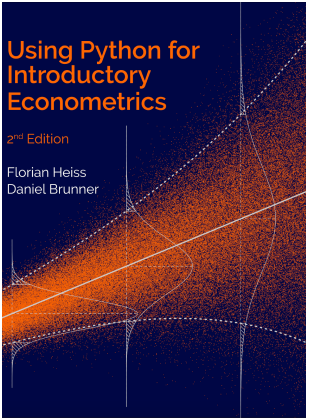
Readings: Bookmark These for R

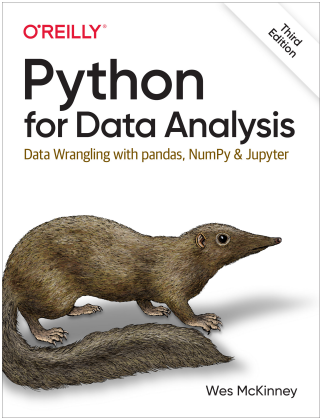






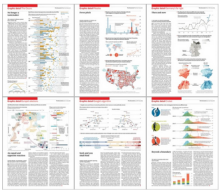
Readings: Bookmark These for Python





Topic-Specific Readings Will Be Posted Online

ECON 370



source: The Economist

Instructor:
Pamela Jakiela

home
syllabus
schedule

1 Data

Readings

These are useful references as opposed to required readings.

R for Data Science: 7, 5, and 19

Intro to Data Science: 2, 4, 6

An Intro to Stats with Python: 2.1, 2.2, and 3.1

Lecture

Slides from Lecture 1

Lab

Assignments: Labs

```
## ECON 370 LAB 1:  DATA CLEANING AND PREP
## NAME:
## DATE:


# preliminaries ------------------------------------------------------------

## install packages / load libraries as needed and load libraries
## you'll need at least tidyverse for R, numpy and pandas for Python

## specify your file path by defining mypath as the path to your working directory for this lab



## R suggestion: mypath <- "[YOUR FILE PATH IN QUOTES]"
## Python suggestion: mypath = "[YOUR FILE PATH IN QUOTES]"



# get the barro-lee data on educational attainment -------------------------

## load barro-lee educational attainment data set from a csv file
## the barro lee website is here:  http://barrolee.com/
## download the csv for the data set on Education Attainment for Population Aged 15 and Over (Total Population)
## the file should be called:  BL2013_MF1599_v2.2.csv
## load the data as a tibble or pandas dataframe called bldata

## R suggestion: read_csv()
## Python suggestion: mypath = pd.read_csv



## how many observations are in the data set?  how many variables?
## which variables are string (or character) variables?

## R suggestions:  dim(), spec(), head(), summary()
## Python suggestions: type(), df.info, df.shape, df.columns, df.dtypes



## define a new tibble called bl2010 which only contains BL data from 2010
## keep only the columns: BLcode, country, yr_sch, WBcode, region_code
## rename yr_sch as mean_edu, WBcode as isocode, and region_code as wb_region

## R suggestion:  use the pipe and the filter, select, and rename functions
## Python suggestion:  df[df['x1'] == VALUE] selects a subset of the rows of dataframe df
## Python suggestion:  df[['x1', 'x2', 'x3']] selects a subset of the columns of dataframe df
## Python suggestion:  rename columns with
## df = df.rename(columns={
##     'old_var_name': 'new_var_name',
## })



## make a table/tibble summarizing the average rate of educational attainment by region

## R suggestion:  use group_by() and pipe to summarize()
## Python suggestion:  use df.groupby().agg()
```

Assignments: Projects/Presentations

- Exploratory Data Analysis (EDA) project
- Treatment Effect Heterogeneity in Randomized Control Trials project
- Final project on a topic of your choice (text? unsupervised learning? new data?)

Assignments and Grading

Lab Assignments	45 points
In-Class Worksheets + Optional Final Exam	10 points
Exploratory Data Analysis Project	7 points
Treatment Effect Heterogeneity in RCTs Project	12 points
Final Empirical Project	21 points
Class Participation in Lecture and Lab	4 points
Getting-to-Know-You Survey	1 point

Presentations and Other Important Dates

Important dates are listed in the syllabus.

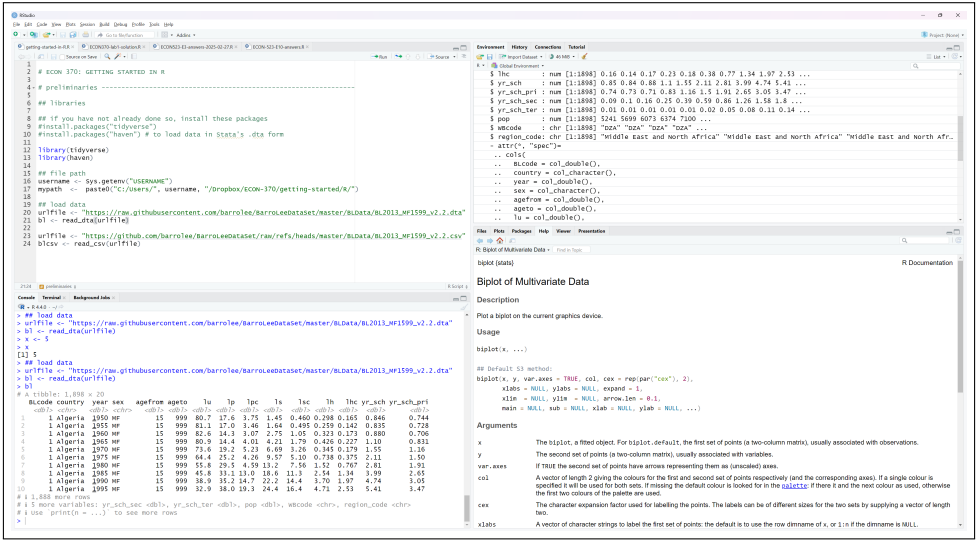
ChatGPT

- The words, analysis tables, and figures/images you produce should be your own
  - If the English words/sentences, tables, or figures that you submit can be found in published articles, on the internet, or anywhere else, you do not receive credit for that lab/project (as if you submitted an article to a journal but that article was already published elsewhere)
  - You must always submit replication code that generates your tables and figures
- Generative AI is a powerful coding tool, and you should learn to use it
  - I recommend handing it small coding tasks (e.g. turning a few lines of R code into Python)
  - Make sure you understand each step in your final script (I may ask)
  - It is **always** your responsibility to make sure that your code is correct, and that it runs

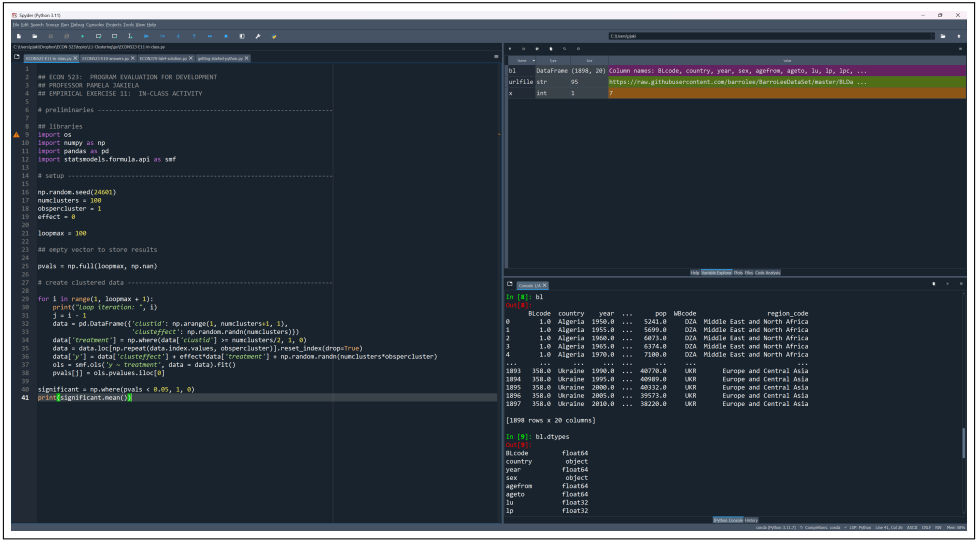
Next Steps

1. Install R and RStudio
2. Install Python
3. Create a gradescope account
4. Complete the getting-to-know-you survey
5. If you are new to R and/or Python, complete the Getting Started Activities

RStudio: https://posit.co/download/rstudio-desktop/



Python/Spyder: https://www.anaconda.com/download/success



Next Week

1. Monday: lecture on data cleaning and exploratory data analysis
2. Wednesday: exploratory data analysis lab
    2.1 Lab 1 due in R and Python by Thursday at 11:00 PM
3. Friday: lecture on dimension reduction and unsupervised learning