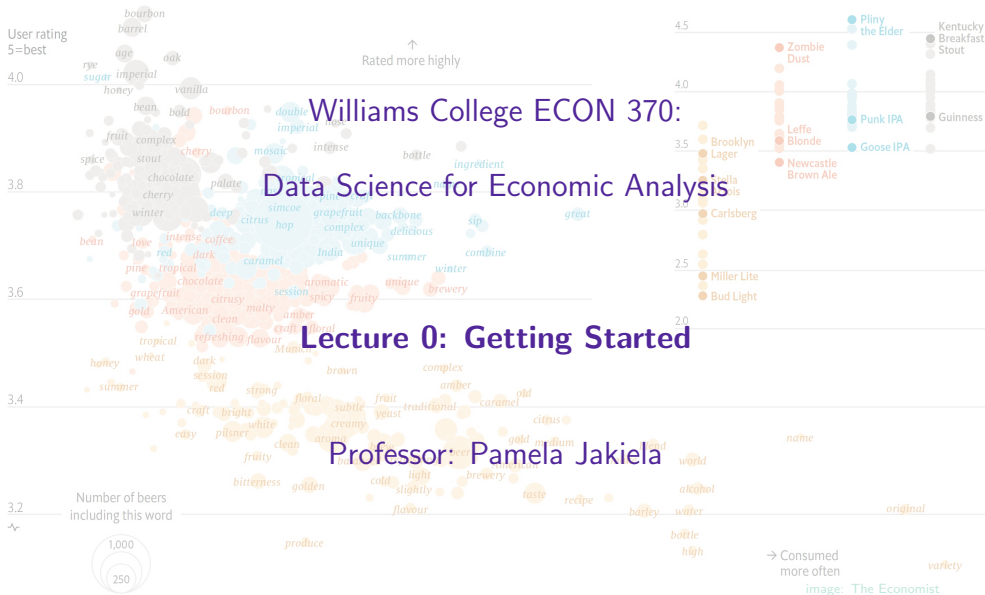


Lecture 0: Getting Started

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 and Thursdays from 3:00 to 4:00 in 339 Schapiro
- Teaching Assistant: Christopher Hughes (cph4@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/1390275> (submitting assignments)

Learning Objectives

- Developing advanced data skills, with a particular focus on identifying new data 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 economists:
 - ▶ 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, clustering

What This Course Is NOT About

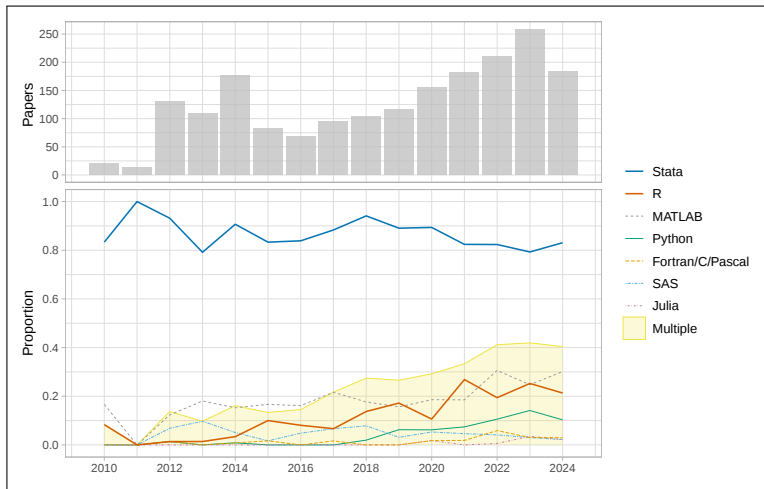
- Big data
- Computation
- Database management
- Data science methods that are focused exclusively on prediction (see STAT/CSCI classes)
- Traditional econometric approaches to causal inference (see ECON 379 and ECON 474)
- Machine learning (or other statistical) techniques that are not used by economists

Stata vs. R vs. Python

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

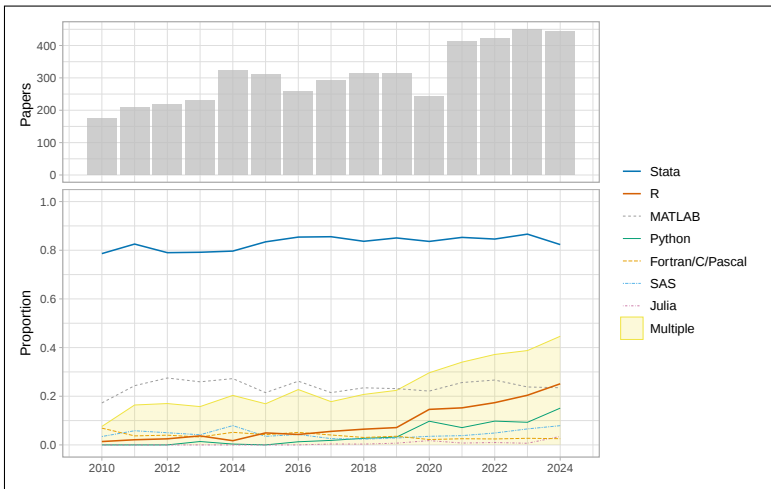
- Many of the techniques we'll cover in this class can't be done in Stata (at present)
- Economics is shifting away from Stata, albeit slowly; not yet clear where we are heading
- Being able to get yourself up-to-speed in a new language is a (data science) skill in itself
- Familiarity with R or Python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

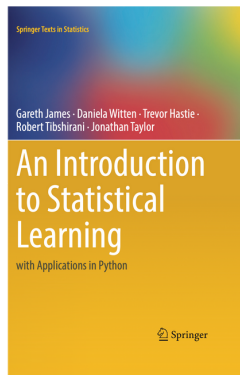
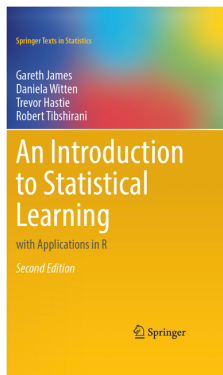
I plan to teach the course primarily in R, to the extent that I focus on a particular language

- 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

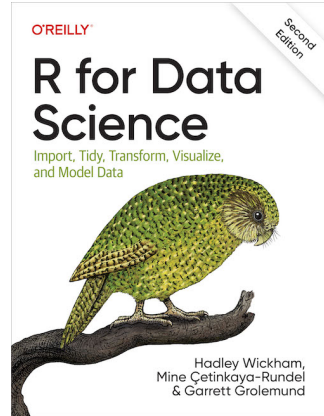
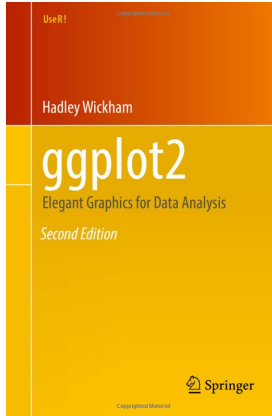
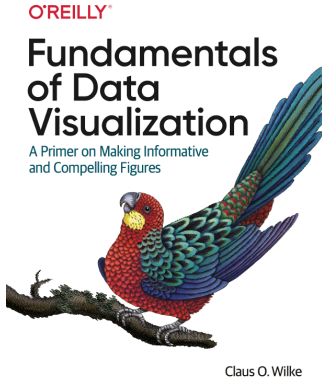
You should decide whether you want to focus your energy on R or Python, either is fine

- 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 10 can be done in R or Python, with some exceptions

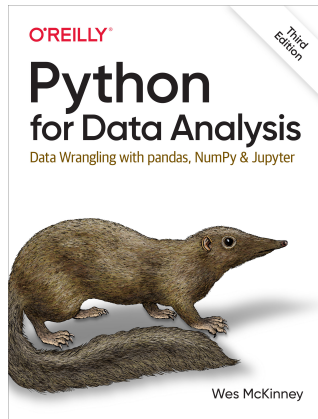
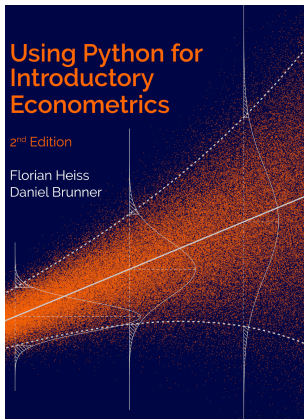
Readings: Download This



Readings: Bookmark These for R

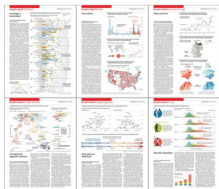


Readings: Bookmark These for Python



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

- An **Exploratory Data Analysis Project** where you seek to explore cross-country data.
- A **Treatment Effect Heterogeneity Project** where you replicate and extend a randomized control trial, testing for differential impacts of a program or policy.
- A **Text as Data Project** where you build and analyze a text data source.

Assignments and Grading

Class Participation	10 points
In-Class Worksheets (3)	10 points
Lab Assignments	10 points
Exploratory Data Analysis Project	10 points
Treatment Effect Heterogeneity Project	20 points
Text as Data Project	20 points
Final Exam	20 points

Presentations and Other Important Dates

Important dates are listed in the syllabus.

Generative AI

- 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 (optional)
3. Create a gradescope account
4. If you are new to R and/or Python, complete the Getting Started Activities

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

The screenshot displays the RStudio environment with a script editor on the left, a console at the bottom left, and a viewer pane on the right showing the output of the `biplot` function.

Script Editor:

```
1 # ECON 370: GETTING STARTED IN R
2
3
4 # preliminaries
5
6 ## libraries
7
8 ## if you have not already done so, install these packages
9 #install.packages("tidyverse")
10 #install.packages("haven") # to load data in Stata's .dta form
11
12 library(tidyverse)
13 library(haven)
14
15 ## file path
16 username <- Sys.getenv("USERNAME")
17 mypath <- paste0("C:/Users/", username, "/Dropbox/ECON-370/getting-started/R/")
18
19 ## load data
20 urlfile <- "https://raw.githubusercontent.com/barrolee/barroleeDataset/master/BLData/BL2013_MF1599_v2.2.dta"
21 bl <- read_dta(urlfile)
22
23 urlfile <- "https://github.com/barrolee/barroleeDataset/raw/refs/heads/master/BLData/BL2013_MF1599_v2.2.csv"
24 blcsv <- read_csv(urlfile)
```

Console:

```
> ## load data
> urlfile <- "https://raw.githubusercontent.com/barrolee/barroleeDataset/master/BLData/BL2013_MF1599_v2.2.dta"
> bl <- read_dta(urlfile)
> x <- $
> x
[1] 5
> ## load data
> urlfile <- "https://raw.githubusercontent.com/barrolee/barroleeDataset/master/BLData/BL2013_MF1599_v2.2.dta"
> bl <- read_dta(urlfile)
> bl
# A tibble: 1,898 × 20
  BLcode country year sex agefrom ageto lu lp lpc ls lsc lh lhc yr_sch yr_sch_pri
  <dbl> <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 1 Algeria 1950 MF 15 999 80.7 17.6 3.75 1.45 0.460 0.298 0.165 0.846 0.744
2 1 Algeria 1955 MF 15 999 81.1 17.0 3.46 1.64 0.495 0.259 0.142 0.835 0.728
3 1 Algeria 1960 MF 15 999 82.6 14.3 3.07 2.75 1.05 0.323 0.173 0.880 0.706
4 1 Algeria 1965 MF 15 999 80.9 14.4 4.01 4.21 1.79 0.426 0.227 1.10 0.831
5 1 Algeria 1970 MF 15 999 73.6 19.2 5.23 6.69 3.26 0.345 0.179 1.55 1.16
6 1 Algeria 1975 MF 15 999 64.4 25.2 4.26 9.57 5.10 0.738 0.375 2.11 1.50
7 1 Algeria 1980 MF 15 999 55.8 29.5 4.59 13.2 7.56 1.52 0.767 2.61 1.91
8 1 Algeria 1985 MF 15 999 45.8 33.1 13.0 18.6 11.5 2.54 1.34 3.99 2.65
9 1 Algeria 1990 MF 15 999 38.9 35.2 14.7 22.2 14.4 3.70 1.97 4.74 3.05
10 1 Algeria 1995 MF 15 999 32.9 38.0 19.3 24.4 16.4 4.71 2.53 5.41 3.47
# 1,888 more rows
# 15 more variables: yr_sch_sec <dbl>, yr_sch_ter <dbl>, pop <dbl>, wcode <chr>, region_code <chr>
# Use `print(n = ...)` to see more rows
```

Viewer Pane:

Biplot of Multivariate Data

Description

Plot a biplot on the current graphics device.

Usage

```
biplot(x, ...)
```

Default S3 method:

```
biplot(x, y, var.axes = TRUE, col, cex = rep(par("cex"), 2),
       xlabs = NULL, ylabs = NULL, expand = 1,
       xlim = NULL, ylim = NULL, arrow.len = 0.1,
       main = NULL, sub = NULL, xlab = NULL, ylab = NULL, ...)
```

Arguments

- x** The biplot, a fitted object. For `biplot.default`, the first set of points (a two-column matrix), usually associated with observations.
- y** The second set of points (a two-column matrix), usually associated with variables.
- var.axes** If TRUE the second set of points have arrows representing them as (unscaled) axes.
- col** A vector of length 2 giving the colours for the first and second set of points respectively (and the corresponding axes). If a single colour is specified it will be used for both sets. If missing the default colour is looked for in the [palette](#): if there it and the next colour as used, otherwise the first two colours of the palette are used.
- cex** The character expansion factor used for labelling the points. The labels can be of different sizes for the two sets by supplying a vector of length two.
- xlabs** A vector of character strings to label the first set of points: the default is to use the row dimname of `x`, or 1:n if the dimname is NULL.

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

The screenshot displays the Spyder Python IDE interface. The left pane shows a Jupyter Notebook with a linear regression model. The right pane shows the Variable Explorer with a DataFrame containing 1898 rows and 20 columns. Below the Variable Explorer, the IPython console shows the output of the model fit.

```
1
2 ## ECON 523: PROGRAM EVALUATION FOR DEVELOPMENT
3 ## PROFESSOR PAPELA JAKIELA
4 ## EMPIRICAL EXERCISE 11: IN-CLASS ACTIVITY
5
6 # preliminaries -----
7
8 ## libraries
9 import os
10 import numpy as np
11 import pandas as pd
12 import statsmodels.formula.api as smf
13
14 # setup -----
15
16 np.random.seed(24601)
17 numclusters = 100
18 obspercluster = 1
19 effect = 0
20
21 loopmax = 100
22
23 # empty vector to store results
24
25 pvals = np.full(loopmax, np.nan)
26
27 # create clustered data -----
28
29 for i in range(1, loopmax + 1):
30     print("loop iteration: ", i)
31     j = 1 - i
32     data = pd.DataFrame({'clusterid': np.arange(1, numclusters+1, 1),
33                          'clusterEffect': np.random.randn(numclusters)})
34     data['treatment'] = np.where(data['clusterid'] >= numclusters/2, 1, 0)
35     data = data.loc[np.repeat(data.index.values, obspercluster)].reset_index(drop=True)
36     data['y'] = data['clusterEffect'] + effect*data['treatment'] + np.random.randn(numclusters*obspercluster)
37     ols = smf.ols('y ~ treatment', data = data).fit()
38     pvals[j] = ols.pvalues.iloc[0]
39
40 significant = np.where(pvals < 0.05, 1, 0)
41 print(significant.mean())
```

Variable Explorer

Name	Type	Size
b1	DataFrame (1898, 20)	Column names: B1code, country, year, sex, agefrom, agefo, lu, lp, ipc, ...
urifile	str	95
x	int	7

IPython console

```
In [8]: b1
Out[8]:
B1code country year ... pop hCode region_code
0 1.0 Algeria 1950.0 ... 5241.0 DZA Middle East and North Africa
1 1.0 Algeria 1955.0 ... 5699.0 DZA Middle East and North Africa
2 1.0 Algeria 1960.0 ... 6071.0 DZA Middle East and North Africa
3 1.0 Algeria 1965.0 ... 6374.0 DZA Middle East and North Africa
4 1.0 Algeria 1970.0 ... 7100.0 DZA Middle East and North Africa
... ..
1893 358.0 Ukraine 1990.0 ... 40770.0 UKR Europe and Central Asia
1894 358.0 Ukraine 1995.0 ... 40889.0 UKR Europe and Central Asia
1895 358.0 Ukraine 2000.0 ... 40332.0 UKR Europe and Central Asia
1896 358.0 Ukraine 2005.0 ... 39573.0 UKR Europe and Central Asia
1897 358.0 Ukraine 2010.0 ... 38220.0 UKR Europe and Central Asia

[1898 rows x 20 columns]

In [9]: b1.dtypes
Out[9]:
B1code float64
country object
year float64
sex object
agefrom float64
agefo float64
lu float32
lp float32
```

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 Saturday 9/19 at 11:00 PM
3. Friday: lecture on dimension reduction and unsupervised learning