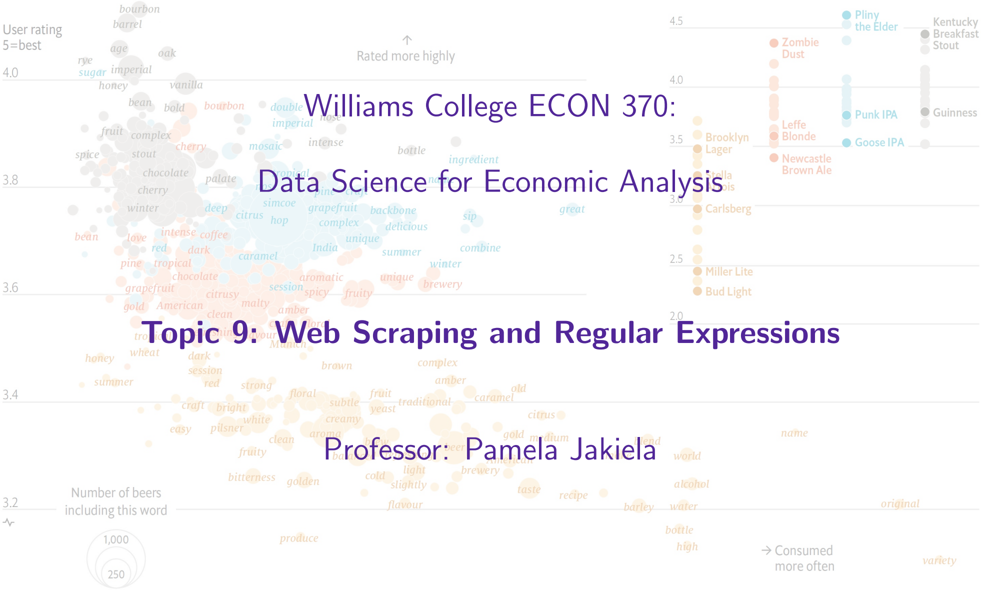
Williams College ECON 370:

Data Science for Economic Analysis

Topic 9: Web Scraping and Regular Expressions

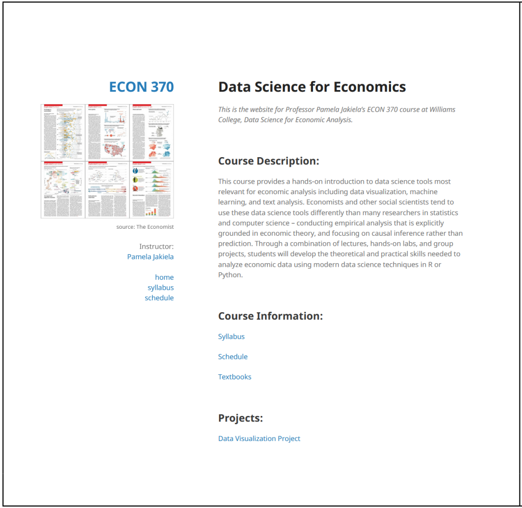
Professor: Pamela Jakiela



Outline

- Web scraping basics
- Regular expression basics
- Scraping the Williams website

Web Pages Are Built in **H**yper**T**ext **M**arkup **L**anguage (HTML)



ECON 370

source: The Economist

Instructor:
Pamela Jakiela

home
syllabus
schedule

Data Science for Economics

This is the website for Professor Pamela Jakiela's ECON 370 course at Williams College, Data Science for Economic Analysis.

Course Description:

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 R or Python.

Course Information:

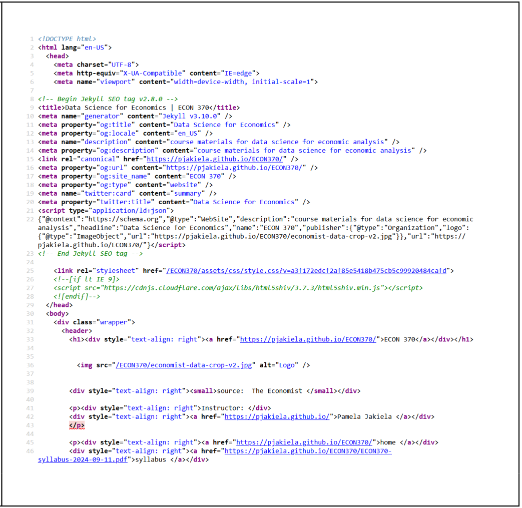
Syllabus

Schedule

Textbooks

Projects:

Data Visualization Project



```html
<!DOCTYPE html>
<html lang="en-US">
  <head>
    <meta charset="UTF-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1">

<!-- Begin Jekyll SEO tag v2.8.0 -->
<title>Data Science for Economics | ECON 370</title>
<meta name="generator" content="Jekyll v3.10.0" />
<meta property="og:title" content="Data Science for Economics" />
<meta property="og:locale" content="en_US" />
<meta name="description" content="course materials for data science for economic analysis" />
<meta property="og:description" content="course materials for data science for economic analysis" />
<link rel="canonical" href="https://pjakiela.github.io/ECON370/" />
<meta property="og:url" content="https://pjakiela.github.io/ECON370/" />
<meta property="og:site_name" content="ECON 370" />
<meta property="og:type" content="website" />
<meta name="twitter:card" content="summary" />
<meta property="twitter:title" content="Data Science for Economics" />
<script type="application/ld+json">
{"@context":"https://schema.org","@type":"WebSite","description":"course materials for data science for economic
analysis","headline":"Data Science for Economics","name":"ECON 370","publisher":{"@type":"Organization","logo":
{"@type":"ImageObject","url":"https://pjakiela.github.io/ECON370/economist-data-crop-v2.jpg"}},"url":"https://
pjakiela.github.io/ECON370/"}</script>
<!-- End Jekyll SEO tag -->

    <link rel="stylesheet" href="/ECON370/assets/css/style.css?v=a3f172edcf2af85e5418b475cb5c99920484cafd">
    <!--[if lt IE 9]>
    <script src="https://cdnjs.cloudflare.com/ajax/libs/html5shiv/3.7.3/html5shiv.min.js"></script>
    <![endif]-->
  </head>
  <body>
    <div class="wrapper">
      <header>
        <h1><div style="text-align: right"><a href="https://pjakiela.github.io/ECON370/">ECON 370</a></div></h1>


          <img src="/ECON370/economist-data-crop-v2.jpg" alt="Logo" />


        <div style="text-align: right"><small>source:  The Economist </small></div>

        <p><div style="text-align: right">Instructor: </div>
        <div style="text-align: right"><a href="https://pjakiela.github.io/">Pamela Jakiela </a></div>
        </p>

        <p><div style="text-align: right"><a href="https://pjakiela.github.io/ECON370/">home </a></div>
        <div style="text-align: right"><a href="https://pjakiela.github.io/ECON370/ECON370-
syllabus-2024-09-11.pdf">syllabus </a></div>
```

Elements and Attributes

A web page is made up of **elements** that are arranged in a hierarchical structure

- Element start and end with a **tag**: for example, `<title>ECON 370</title>`
  - Every html page contains an html element (`<html>...</html>`)
  - The html element contains the elements (children) **head** and **body**
- Many elements appear multiple times within a page: for example, `<p>a paragraph</p>`
- Tags can contain **attributes** that encode element-specific information, for example:
  - `<h1 id="data-science-for-economics">Data Science for Economics</h1>`
  - `<p><a href="https://pjakiela.github.io/syllabus.pdf">Syllabus</a></p>`

When to Scrape

Simple web scraping tools are useful when:

- A page contains many repetitions of the same structure/sequence of elements, and you want to combine those elements into a data frame for analysis

  Example:
  ```
  <h4><span class="course_code">ECON 105</span><span
  class="course_terms">(F)</span> <span class="course_code">SEM</span> <span
  class="course_title">Gender in the Global Economy</span></h4><h4><span
  class="course_code">ECON 107</span><span class="course_terms_blank"></span>
  <span class="course_code">SEM</span> <span class="course_title">Inequality in a
  Classless Society:  The Soviet Experiment and its Aftermath</span></h4>
  ```

- A sequence of (ideally numerically indexed) pages use the same structure and elements, and you want to combine the information from multiple pages to build a data set
  - **Example:** building a data set of all recent NBER working papers
  - **Example:** extracting information on undergraduate institution from faculty profile pages

When Not to Scrape

Many websites cannot be scraped easily with simple tools (can you do it?)

- Policy advice: don't try to scrape websites that don't want to be scraped

Web scraping raises a range of ethical issues (should you do it?)

- Some personally-identifiable information posted on the web should not be used for research
    - In general, work products and things shared under creative comments licenses are in-bounds, personal (non-professional) content posted with some expectation of privacy is out-of bounds
    - When collecting personally identifiable information for research, check with your IRB
    - Collecting personally-identifiable information is often unnecessary
- Also: don't violate a websites terms of service (if you are bound by them) or copyright law
- Finally, be polite: to avoid over-burdening a server, always build in pauses between queries

How to Scrape: Get the HTML, Extract Text and Attributes

1. Do the actual scraping using `rvest` in R or `requests` + `BeautifulSoup` in Python
2. Extract elements and the text and attributes they contain:
   - Extract elements with tag "a": `html_elements(mypage, "a")` or `mypage.select("a")`
   - Extract elements with `class` equal to a with `(".a")`
   - Extract elements with `is` equal to a with `("#a")`
   - Extraction tasks are often sequenced (first extract parents, then specific children)
   - Final step is to extract text (printed on web page) or the value of an attribute

String Variables: Provisions Data

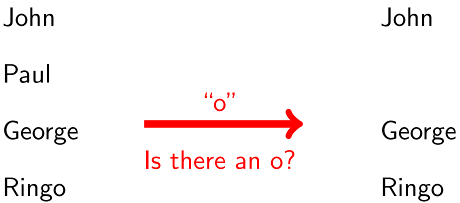
amt	item_description
17	Snow Buddies (Regular) Nettle Meadow Farm, Kunik Mini Organic (Regular)
33	Czechs Out Pilsner (Regular) Shire Space Junk, IPA 4 Pack (Regular)
23	Mitica Almonds Marcona Prepacked (Regular), Sangria Mix (Regular)
9	Spicy Maple Almonds (Regular) Red Jacket Orchards, Juice Strawberry (Regular)
15	Sourdough Loaf (Regular) Nettle Meadow Farm, Kunik Mini Organic (Regular)
16	Equator Coffees Equator Blend Roasted, Whole Bean Coffee (Regular)
16	Equator Coffees Equator Blend Roasted, Whole Bean Coffee (Regular)
23	Finback Brewery IPA Something Blanc, 4pk (Regular)
20	Ommegang Quadrupel 3 Philosophers, 4pk (Regular)
15	Sourdough Loaf (Regular) Nettle Meadow Farm, Kunik Mini Organic (Regular)

String Variables: REPEC Rankings of U.S. Economics Departments

id	department	school
40	Dept. of Economics	Washington University in St. LouisSt. Louis
41	Dept. of Economics	University of ColoradoBoulder
42	Economics Dept.	University of California-Santa Cruz (UCSC)Santa Cruz
43	Dept. of Economics	W.P. Carey School of Business, Arizona State UniversityTempe
44	Dept. of Economics	George Washington UniversityWashington
45	Dept. of Economics	University of PittsburghPittsburgh
46	Dept. of Economics	Andrew Young School of Policy Studies, Georgia State UniversityAtlanta
47	Dept. of Economics	University of WashingtonSeattle
48	Dept. of Economics	Tufts UniversityMedford
49	Economics Dept.	Williams CollegeWilliamstown
50	Economics Dept.	Eller College of Management, University of ArizonaTucson

Regular Expressions

Regular expressions (**regex**) are strings of characters used for pattern matching in strings



John

Paul

George

Ringo

“o”

Is there an o?

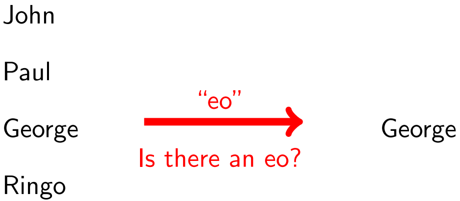
John

George

Ringo

Regular Expressions

Regular expressions (**regex**) are strings of characters used for pattern matching in strings



John

Paul

George

Ringo

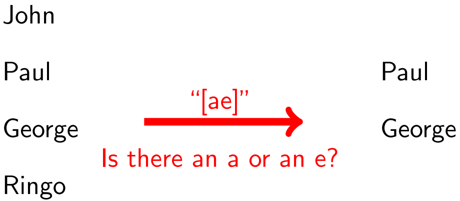
“eo”

Is there an eo?

George

Regular Expressions

Regular expressions (regex) are strings of characters used for pattern matching in strings



John

Paul

George

Ringo

“[ae]”

Is there an a or an e?

Paul

George

String Manipulation with Regular Expressions

1. Find a match (to create an indicator or filter/subset data)
   1.1 Starts with, ends with, any letter, any number, or, a year
   1.2 Escape characters
2. Count the number of matches
3. Find and replace one sequence of characters with another sequence of characters
4. Split a string into two parts using a delimiter (for example, split words/names using " ")

Lab #9

Objective: scrape departmental websites to collect data on Williams College faculty

1. `ECON370-lab9` provides a template that collects data from the English department
   1.1 Scrapes the page listing names, titles, and profile pages for faculty and staff
   1.2 Extracts names, titles, and profile page urls
   1.3 Restricts the sample to tenure-line teaching faculty
   1.4 Guesses gender identity based on first name using social security data
   1.5 Extracts information on education (schools attended) from faculty profile pages
2. Each of you will replicate and extend the template for a different departments