

Economia, Finanza e Analisi dei Dati

Laurea Magistrale

Data Analysis for Economics

Data from the Web

Marco Nieddu

Fall 2026

Who Are We Building On?

- This lecture builds on the lecture slides from Jared Hutchins: https://github.com/jphutch/ACE-592-SAE/.
- And we are going to surf two websites:
  - the MIMIT page with the daily fuel prices we used in the previous classes, carburanti-archivio-prezzi, through a copy we created to avoid putting pressure on the MIMIT servers: https://garaugarau.github.io/DAEc/;
  - Wikipedia: https://it.wikipedia.org/wiki/Comuni_della_Sardegna.

Surfing the Web with Python

- A vital skill for data analysis: talking to websites from Python.
- **Why?**
  - Data collection becomes much easier: no tedious clicking. The MIMIT archives are 94 files; Wikipedia has one page for each of the 377 municipalities of Sardinia.
  - Tasks that usually take hours can be automated, and repeated when new data come out.
  - Many datasets used in economics today were never published as datasets: online prices, job ads, rents, local information (e.g., Cavallo and Rigobon, 2016).

Today

Notebook	Page	What we collect
01 Web scraping	MIMIT archive page (our copy)	94 quarterly archives: daily prices and stations, 2015–2026
02 Wikipedia	The list of the municipalities of Sardinia, then the page of each	The local holiday of each municipality

Same tools in both: `requests` to ask for a page, `BeautifulSoup` to read it, loops to repeat.

What Is "HTTP"?

- Hypertext Transfer Protocol (HTTP) is the way clients (e.g., your computer) talk to servers (e.g., websites).
- The **user agent** (usually a web browser) sends a **request** for a resource, and the server sends back a **response**: a page in HTML, or another resource (a file, an image).
- The browser is a high-level user agent; we can use Python as our user agent instead.

There is way more to this than we need to know...

What's the Upshot?

- With Python as the user agent, we can send requests to websites and get back responses without opening a browser window.
- **GET**: the common way to "request data". It is the only request we need: `requests.get(url)`.
- Others exist: **POST** sends data to a server (e.g., a form), **HEAD** asks only for the header of the response, ...
- Every response comes with a **status code**: a number that says whether the request worked.

The Golden Rule of Status Codes

- If it starts with a 2? **GOOD**: 200 OK.
- If it starts with a 4 or 5? **BAD**
  - 403 Forbidden: we are not allowed (e.g., Wikipedia, if we do not say who we are);
  - 404 Not Found: the page does not exist (e.g., a typo in the address);
  - 500 Internal Server Error: the server has a problem.
- Something else? **Look it up**: 3xx are redirections (the resource is somewhere else), 1xx are "intermediate" responses.

HTML in One Slide

- A page is text in HTML, made of **tags** that open and close: `<td> ... </td>`.
- A table body is a `<tbody>`, each row a `<tr>`, each cell a `<td>`; a link is an `<a>`, with the address in its `href`.

```
<tbody>
  <tr>
    <td>Terzo trimestre 2026</td>
    <td><a href="https://github.com/.../2026_3_tr.tar.gz">Anagrafica 3-2026</a></td>
    <td><a href="https://github.com/.../p_2026_3_tr.tar.gz">Prezzo 3-2026</a></td>
  </tr>
  ...
</tbody>
```

In the browser: right-click on the page, *Inspect*.

Reading HTML with BeautifulSoup

- `soup.find('tbody')`: the first tag of a kind; `find_all('tr')`: all of them, in a list.
- `rows[0]`: one element of the list, as in any list.
- `cell.get_text(strip=True)`: the text of a tag, without the tags.
- `cell.find('a')['href']`: the address of a link.
- Searching by class or by text: `find('table', class_='infobox sinottico')`, `find('a', string='Giorno festivo')`; and climbing to the row that contains a tag: `find_parent('tr')`.

The Plan, in Both Notebooks

1. **One by hand**: one page, one row, one cell, one link, one download.
2. **All with a loop**: the same steps for every row, collected in a DataFrame.
3. **Structure**: the downloads in folders (one per quarter, then one per year), the results in a file in `processed/`.

Local Saints and Holidays

- The goal of notebook 02 is to build a dataset with the local (religious) holiday of each municipality in Sardinia.
- Steps:
  1. Identify a data source that includes this information (Wikipedia): see the *Giorno festivo* of Abbasanta, https://it.wikipedia.org/wiki/Abbasanta.
  2. Write a script that extracts the information from a single page.
  3. Find a list of all Sardinian municipalities (with the link to their Wikipedia page) and store the links in a DataFrame.
  4. Write a script that goes through each link in the DataFrame and collects the local holiday.

Why?

- These data can be used (and have actually been used) to address **economic** questions.
  - The impact of religious celebrations on economic development: Montero and Yang (2022).
  - Mexican municipalities whose patron-saint festival falls in the planting or harvest season are poorer today.

Culture and Identity in Economics

- More broadly, a growing literature emphasizes the role of **culture, identity, and geography** in shaping long-run economic and political outcomes:
  - **Squicciarini (2020)** shows that in nineteenth-century France, regions with stronger *religiosity* invested less in secular education and industrialization, slowing economic progress.
  - **Bazzi and Gudgeon (2021)** show that in Indonesia, redrawing the *boundaries* of districts changed their ethnic composition, and with it the risk of conflict.
  - **McGuirk and Nunn (2023)** show that *climate change and rainfall variability* increased conflict in regions dependent on transhumant pastoralism in Africa.

Religious Holidays in Mexico

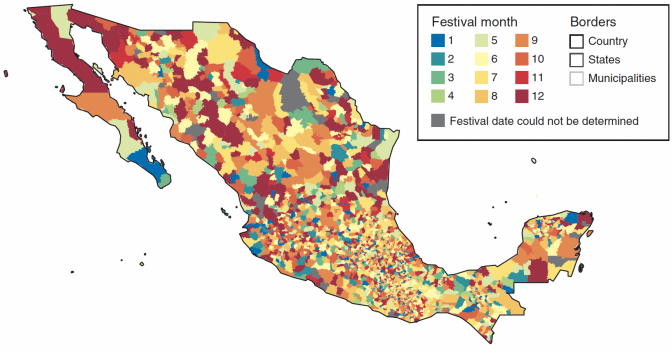


FIGURE 1. SAINT DAY FESTIVAL MONTHS: ALL OF MEXICO

Notes: The map presents the month that each municipality in Mexico celebrates its respective Catholic patron saint day festival. Municipalities where we were unable to determine the festival date are shaded in dark gray. See online Appendix A.1 for more information on the construction of the festival date dataset.

When Religious Holidays Coincide with the Harvesting Season

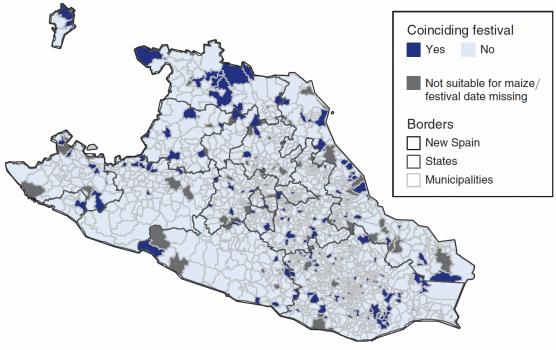


FIGURE 3. AGRICULTURALLY COINCIDING FESTIVALS: NEW SPAIN REGION OF MEXICO

Notes: Coinciding festival is equal to "yes" if the saint day festival in a municipality occurs either 0 to 30 days prior to the optimal maize planting date or 0 to 30 days after the optimal maize harvest date for a municipality using FAO GAEZ data, and "No" otherwise for each municipality in the New Spain region of Mexico. Municipalities where we were unable to determine the festival date or are unsuitable for maize are shaded in dark gray.

Holidays and Economic Development

TABLE 2—IMPACT OF AGRICULTURALLY COINCIDING FESTIVALS ON DEVELOPMENT OUTCOMES

	(1)	(2)	(3)	(4)	(5)
Panel A. Log household income					
Festival coincides with maize planting or harvest	−0.275	−0.204	−0.251	−0.255	−0.206
	(0.099)	(0.080)	(0.071)	(0.070)	(0.077)
	[0.107]	[0.080]	[0.082]	[0.079]	[0.075]
State fixed effects	N	Y	Y	Y	Y
Geography controls	N	N	Y	Y	Y
Colonial controls	N	N	N	Y	Y
Planting-month fixed effects	N	N	N	N	Y
Harvest-month fixed effects	N	N	N	N	Y
Festival-week fixed effects	N	N	N	N	Y
Observations	1,593	1,593	1,593	1,593	1,593
Adjusted R^2	0.004	0.347	0.538	0.543	0.567
Mean dependent variable	3.234	3.234	3.234	3.234	3.234
Standard deviation dependent variable	1.330	1.330	1.330	1.330	1.330
Panel B. Index of economic development					
Festival coincides with maize planting or harvest	−0.695	−0.422	−0.593	−0.613	−0.537
	(0.300)	(0.240)	(0.209)	(0.208)	(0.224)
	[0.342]	[0.252]	[0.226]	[0.220]	[0.190]
State fixed effects	N	Y	Y	Y	Y
Geography controls	N	N	Y	Y	Y
Colonial controls	N	N	N	Y	Y
Planting-month fixed effects	N	N	N	N	Y
Harvest-month fixed effects	N	N	N	N	Y
Festival-week fixed effects	N	N	N	N	Y
Observations	1,593	1,593	1,593	1,593	1,593
Adjusted R^2	0.002	0.348	0.566	0.572	0.597
Mean dependent variable	−0.589	−0.589	−0.589	−0.589	−0.589
Standard deviation dependent variable	4.039	4.039	4.039	4.039	4.039

Good Manners

- **Say who you are**: a *User-Agent* with the name of the program and a contact. Wikipedia refuses requests without it (403).
- **Pause between requests** (`time.sleep(1)`): a loop without pauses can look like an attack.
- **Read `robots.txt`**, at the root of the site: the parts the owner does not want programs to visit.
- **Use a download page or an API when there is one**; scrape only what is not offered as data.
- **Keep what you download**: download again only when you need newer data.
- **Check the terms of use** before publishing anything built on the data (e.g., cite the source).

References

- Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using online prices for measurement and research. *Journal of Economic Perspectives*, 30(2), 151–178.
- Montero, E., & Yang, D. (2022). Religious festivals and economic development: Evidence from the timing of Mexican saint day festivals. *American Economic Review*, 112(10), 3176–3214.
- Hutchins, J. Lecture slides, ACE 592 SAE: https://github.com/jphutch/ACE-592-SAE/.