Wicaksa Munajat

CST383 CSUMB Summer ‘22 Data Science

05/16/2022

In this lab, we'll get some practice loading data. Note: I often use this code at the top of my Python files when doing data science.

import pandas as pd

import seaborn as sns

from matplotlib import rcParams

# allow output to span multiple output lines in the console

pd.set\_option('display.max\_columns', 500)

# switch to seaborn default stylistic parameters

# see the useful https://seaborn.pydata.org/tutorial/aesthetics.html

sns.set()

sns.set\_context('paper') # 'talk' for slightly larger

# change default plot size

rcParams['figure.figsize'] = 9,7

1. Go to data.gov, and look around a little. See if you can find a data set of interest to you.

<https://catalog.data.gov/dataset/major-sport-venues>. This data set looks at all of the venues for all of the professional leagues in the USA.

Graphical user interface, text, application, email

Description automatically generated

2. Look at the icons explaining the format of the data. Note that often the data is in HTML format, meaning a web page is available. Normally such a web page will let a person who is not a data scientist explore the data easily. We want the raw data!

While that previous data set is only available for private use, I can see that others have different formats such as RDF, JSON, and XML.

Text

Description automatically generated with medium confidence

3. Go to the following page, which shows various data sets for counties around the US:

https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data.aspx (Links to an external site.)

4. Download the .xls file for unemployment and median household income for the US, states, and counties.

Graphical user interface, text, application

Description automatically generated

5. Load the data as a pandas data frame.

If you have Excel on your laptop, load the file into Excel, and export it as a CSV file. This is a bit awkward because the Excel file contains multiple sheets. Then use pandas.read\_csv() to read the data. You will want to use the 'skiprows' option to skip the first rows of the exported CSV. This is much better than manually editing the CSV, as it is repeatable.

If you don't have Excel, load the data directly from the Excel data file using pandas read\_excel(). Note the 'header' option. Look at the data frame you load and make sure it looks right. You should have 3275 rows and 52 columns (it may be 88 now -- more than when I loaded the data). The third column should be 'Area\_name'.

6. There is a (slightly older) CSV file for this data on the following github page.

Try loading the data set using this URL. Use pandas read\_csv():

<https://raw.githubusercontent.com/grbruns/cst383/master/unemployment.csv>

(Links to an external site.)

I ended up using this old file for the data.

Table

Description automatically generated

7. How does the work you’re doing with this data to get it into pandas data frame relate to the need to have your results be replicable? How would you document the source of the data set, and what you did to get it into a format that Python can use?

Documenting the source by creating metadata elements such as title, creator, date, description so that anyone who uses the data in the future can easily interpret and use the data.

8. Data sets often come with associated data dictionaries that explain where the data comes from, what the attributes mean, etc. Where is the documentation for this data set?

The documentation can be found in this url: <https://www.ers.usda.gov/data-products/county-level-data-sets/documentation/>

A picture containing application

Description automatically generated

9. If you still have time, try exploring the data with Python.