Discussion Forums

Week 4

SUBFORUMS

All

Assignment: Visualizing Two Datasets

← Week 4

CF

How to download data from a website into a CSV file?

claudio freitas Week 4 · a year ago

Can anyone provide an example?

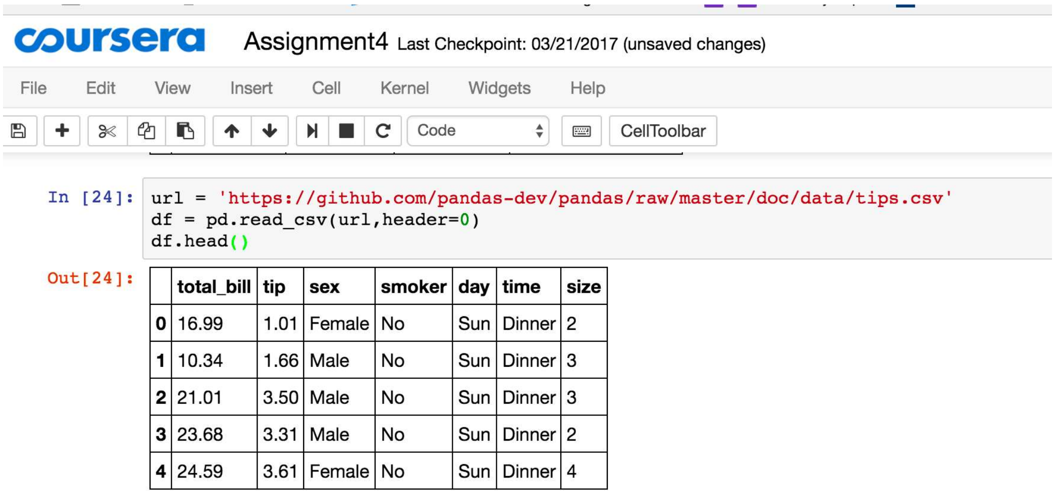
⇧ 2 Upvotes Reply Follow this discussion

HIGHLIGHTED POST

SG Sophie Greene Teaching Staff · a year ago · Edited

If you have a link to the CSV file, pandas read_csv can be used exactly the same way, instead of the file name, you pass the url

here is an example



```
In [24]: url = 'https://github.com/pandas-dev/pandas/raw/master/doc/data/tips.csv'
         df = pd.read_csv(url,header=0)
         df.head()
```

Out[24]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

If you have a web page containing a table and you want to read that table you can use read_html. You will need to load the missing modules like lxml in the online platform, that said, the changes are not permanent, i.e. the modules you install will be deleted when you log out /restart the Jupyter server

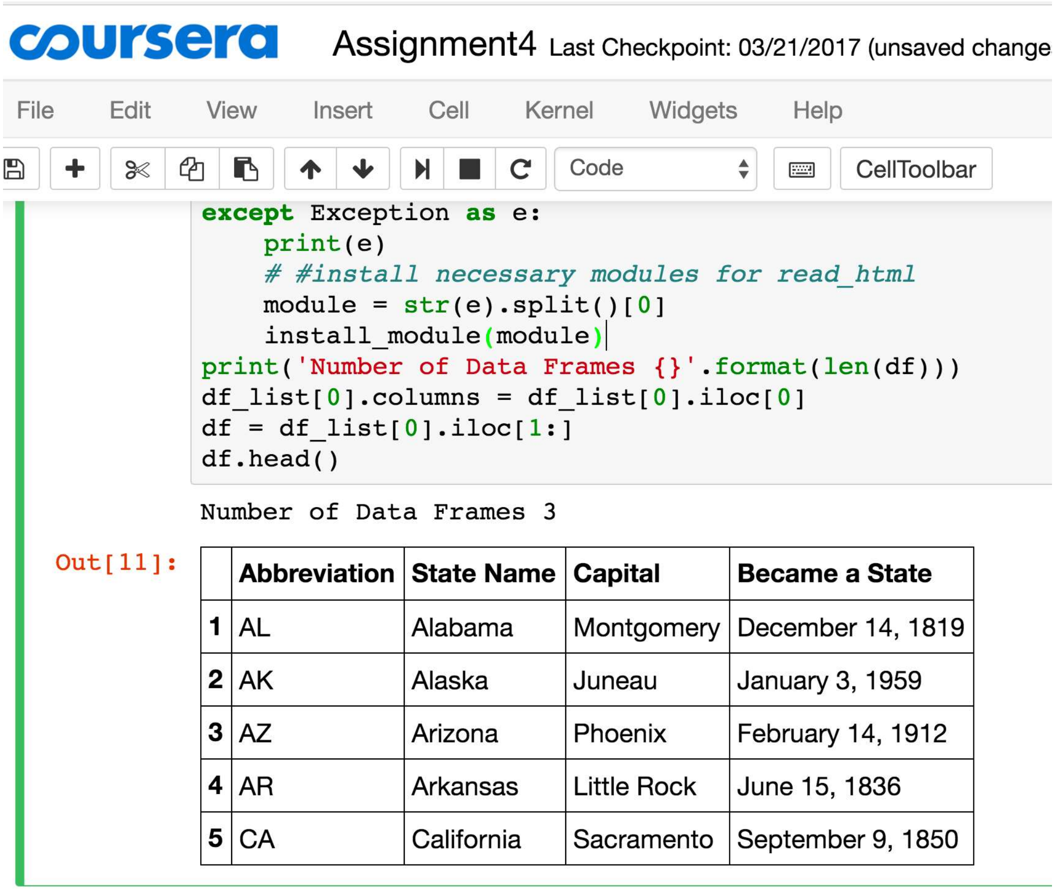
I've written a script that can help load the needed modules automatically

```
import pandas as pd
from IPython.display import display, HTML


def install_module(module):
    ! conda install "$module" -y
    js_cmd = ['IPython.notebook.kernel.restart();',
              'IPython.notebook.select(1);',
              'IPython.notebook.execute_cell();'
              ]
    js = "<script>{0}</script>".format(' '.join(js_cmd))
    display(HTML(js))

url = 'https://simple.wikipedia.org/wiki/List_of_U.S._states'
try:
    df_list = pd.read_html(url)
except Exception as e:
    print(e)
    # #install necessary modules for read_html
    module = str(e).split()[0]
    install_module(module)
print('Number of Data Frames {}'.format(len(df_list)))
df_list[0].columns = df_list[0].iloc[0]
df = df_list[0].iloc[1:]
df.head()
```

the script will automatically restart the kernel to propagate the changes and will return the list tables available in the input url page



```
except Exception as e:
    print(e)
    # #install necessary modules for read_html
    module = str(e).split()[0]
    install_module(module)
print('Number of Data Frames {}'.format(len(df)))
df_list[0].columns = df_list[0].iloc[0]
df = df_list[0].iloc[1:]
df.head()
```

```
Number of Data Frames 3
```

Out[11]:

	Abbreviation	State Name	Capital	Became a State
1	AL	Alabama	Montgomery	December 14, 1819
2	AK	Alaska	Juneau	January 3, 1959
3	AZ	Arizona	Phoenix	February 14, 1912
4	AR	Arkansas	Little Rock	June 15, 1836
5	CA	California	Sacramento	September 9, 1850

This way your peer reviewers will be able to run the code in the online platform without issues

If you want to save the html table you read using read-html to datafile you can use

```
#csv
df.to_csv('filename.csv')
#or excel might need module xlwt installed
df.to_excel('file.xls')

```

⇧ 7 Upvotes Jump to post

Earliest **Top** **Most Recent**

Pierre Masson · a year ago

Hi Sophie,

I have tried to use the script. As a result, I receive the following message

lxml not found, please install it