Python Loop through Excel sheets, place into one df

Asked 3 years, 8 months ago Active 1 year, 5 months ago Viewed 35k times

13

6

I have an excel file `foo.xlsx` with about 40 sheets `sh1`, `sh2`, etc. Each sheet has the format:

```
area      cnt   name\nparty1   name\nparty2
blah      9         5              5
word      3         7              5
```

In each sheet I want to rename the vars with the format `name\nparty` to only have the `party` as a label. Example output:

```
area      cnt    party1    party2     sheet
bacon     9         5         5        sh1
spam      3         7         5        sh1
eggs      2         18        4        sh2
```

I am reading in the file with:

```
book = pd.ExcelFile(path)
```

And then wondering if I need to do:

```
for f in filelist:
    df = pd.ExcelFile.parse(book,sheetname=??)
    'more operations here'
    # only change column names 2 and 3
     i, col in enumerate(df):
     if i>=2 and i<=3:
        new_col_name = col.split("\n")[-1]
        df[new_col_name] =
```

Or something like that?

python excel pandas

Share Improve this question

Follow

edited Jun 14 '17 at 16:46

hpaulj

169k 12 155 264

asked Jun 14 '17 at 15:36

Yolo_chicken

733 2 6 19

1 Answer

Active | Oldest | Votes

27

UPDATE as of 2019-09-09:

use `sheet_name` for v0.25.1 instead of `sheetname`

The `read_excel` method of `pandas` lets you read all sheets in at once if you set the keyword parameter `sheetname=None`. This returns a dictionary - the keys are the sheet names, and the values are the sheets as dataframes.

Using this, we can simply loop through the dictionary and:

1. Add an extra column to the dataframes containing the relevant sheetname
2. Use the `rename` method to rename our columns - by using a `lambda`, we simply take the final entry of the list obtained by splitting each column name any time there is a new line. If there is no new line, the column name is unchanged.
3. Append to the "full table"

Once this is done, we reset the index and all should be well. Note: if you have parties present on one sheet but not others, this will still work but will fill any missing columns for each sheet with `NaN`.

```
import pandas as pd

sheets_dict = pd.read_excel('Book1.xlsx', sheetname=None)

full_table = pd.DataFrame()
for name, sheet in sheets_dict.items():
    sheet['sheet'] = name
    sheet = sheet.rename(columns=lambda x: x.split('\n')[-1])
    full_table = full_table.append(sheet)

full_table.reset_index(inplace=True, drop=True)

print full_table
```

Prints:

```
    area  cnt  party1  party2   sheet
0  bacon    9       5       5  Sheet1
1   spam    3       7       5  Sheet1
2   eggs    2      18       4  Sheet2
```

Share Improve this answer

Follow

edited Sep 9 '19 at 1:48

Kim Stacks

10.6k 29 125 237

answered Jun 14 '17 at 15:46

asongtoruin

8,121 2 25 37

I'm sorry I was unclear. The name in `name\nparty` changes each sheet. It's electoral results and I don't want the candidates name just their party. Is there some kind of wild card or string split to only keep everything after the `\n` ? – Yolo_chicken Jun 14 '17 at 16:01

2 I think .split('\n')[-1] only keeps parts of a string after the \n. For example, "Frank Underwood\nFictional Democrat".split('\n')[-1] returns 'Fictional Democrat' – DalekSec Jun 14 '17 at 16:09

1 @DalekSec was just editing this in! The correct approach for sure. – asongtoruin Jun 14 '17 at 16:13

1 for version 0.25.1 it should be `sheet_name` not `sheetname` . I didn't check if `sheetname` works in previous versions. – Kim Stacks Sep 9 '19 at 1:47