Data Wrangling

Presenter: Steve Baskauf
steve.baskauf@vanderbilt.edu

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

CodeGraf landing page

- vanderbi.lt/codegraf

Data "wrangling"

- a.k.a. data "munging"
- Can involve cleaning, reformatting, summarizing, and changing the data organization to make it more fit for some use like visualization.
- A very large topic – we are only scratching the surface.
- See chapters 5, 7, and 8 in Python for Data Analysis

Basic DataFrame manipulation

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

Column vs. index label

integer index

regular column

	State	Commercial	Electric Power	Residential	Industrial	Transportation	Total
0	Alabama	2.22	55.25	1.87	21.06	34.69	115.09
1	Alaska	2.03	2.75	1.50	16.78	11.85	34.91
2	Arizona	2.87	44.28	2.19	4.59	33.08	87.01
3	Arkansas	2.94	30.22	1.66	8.21	19.38	62.41
4	California	18.87	36.57	24.11	68.84	212.95	361.35

index label

State	Commercial	Electric Power	Residential	Industrial	Transportation	Total
Alabama	2.22	55.25	1.87	21.06	34.69	115.09
Alaska	2.03	2.75	1.50	16.78	11.85	34.91
Arizona	2.87	44.28	2.19	4.59	33.08	87.01
Arkansas	2.94	30.22	1.66	8.21	19.38	62.41
California	18.87	36.57	24.11	68.84	212.95	361.35

Ways to make changes

- Assign to a named view

```
sorted_view = state_co2_fuel.sort_values(by='Total mmt')
```

- Assign to a named copy

```
sorted_copy = state_co2_fuel.copy().sort_values(by='Total mmt')
```

- Perform operation "inplace"

```
state_co2_fuel.sort_values(by='Total mmt', inplace=True)
```

no assignment

Removing rows and columns

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

"Axes" of a data frame

axis 1
column axis

axis names

axis 0
row axis

Sector	Commercial	Electric Power	Industrial	Residential	Transportation
State					
Alabama	43.996634	1463.430990	503.697916	64.751214	678.018321
Alaska	50.408486	60.772091	406.923918	37.226044	303.227119
Arizona	42.886448	902.591302	96.147374	42.705125	650.674600
Arkansas	41.095011	534.948110	213.471009	52.310787	407.280788
California	329.118829	926.954976	1540.460128	602.205615	4449.921497

Handling missing data

- Pandas has a method for broadly replacing missing data: **`.fillna()`**
- Selection is also possible using **`.isnull()`** and **`.notnull()`** to generate boolean array to be used for selection indexing.

Sorting rows

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

Slicing columns and rows

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

Recall:

- use `.loc[]` for label indices.
- use `.iloc[]` for integer indices.
- only the first index is required to slice rows
- to slice columns, specify the row as `:` , then the column range.
- by default, slices are only views of the data, not copies.

Selecting data

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

How selecting works

- A boolean operation is done on a column. Any common operation (**==**, **<**, **>**, etc.) is possible.
- That generates a series of boolean (**`True`** or **`False`**) the same length as the number of table rows.
- If the series item corresponding to the row is **`True`**, the row is included. If the series item for that row is **`False`**, the row is excluded.
- The resulting DataFrame maintains the indices of the original DataFrame.

Selection indexing process

`organism_info`

index	group (0)	number legs (1)
'lizard' (0)	'reptile'	4
'spider' (1)	'arachnid'	8
'worm' (2)	'annelid'	0
'bee' (3)	'insect'	6

Insert this sequence as the index (in the square brackets).

organism_info['number legs'] > 5
False
True
False
True

index	group (0)	number legs (1)
'spider' (1)	'arachnid'	8
'bee' (3)	'insect'	6

```
organism_info[organism_info['number legs'] > 5]
```

Remote Support for Teaching and Research Needs

- Access to digital collections 24/7
- Skype consultations with your subject librarian
- Ask a Librarian: an easy way to submit a question via email
- Live chat available from the Library home page

Jean & Alexander Heard
LIBRARIES

NEED HELP? ASK A LIBRARIAN!

https://www.library.vanderbilt.edu/ask-librarian.php

Jean & Alexander Heard
LIBRARIES