First Notebook War

PyData DC 2018

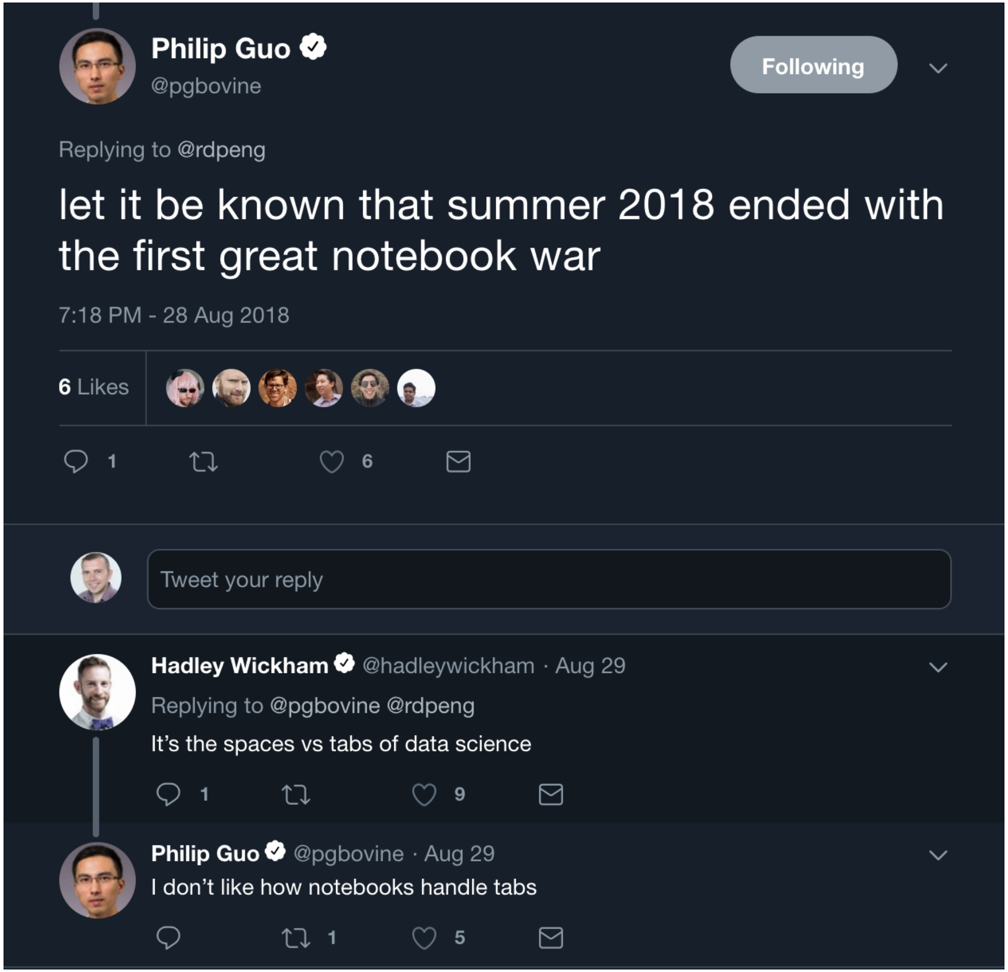
Martin Skarzynski

2018-11-17

About me

- Cancer Prevention Fellow
  - at National Cancer Institute
- Co-Chair, Bioinformatics & Data Science
  - at Foundation for Advanced Education in the Sciences
- Website: https://marskar.github.io
- Twitter: @marskar

Title inspired by a tweet by Philip Guo



Philip Guo
@pgbovine

Replying to @rdpeng

let it be known that summer 2018 ended with the first great notebook war

7:18 PM - 28 Aug 2018

6 Likes

1 6

Tweet your reply

Hadley Wickham @hadleywickham · Aug 29
Replying to @pgbovine @rdpeng
It's the spaces vs tabs of data science

1 9

Philip Guo @pgbovine · Aug 29
I don't like how notebooks handle tabs

1 5

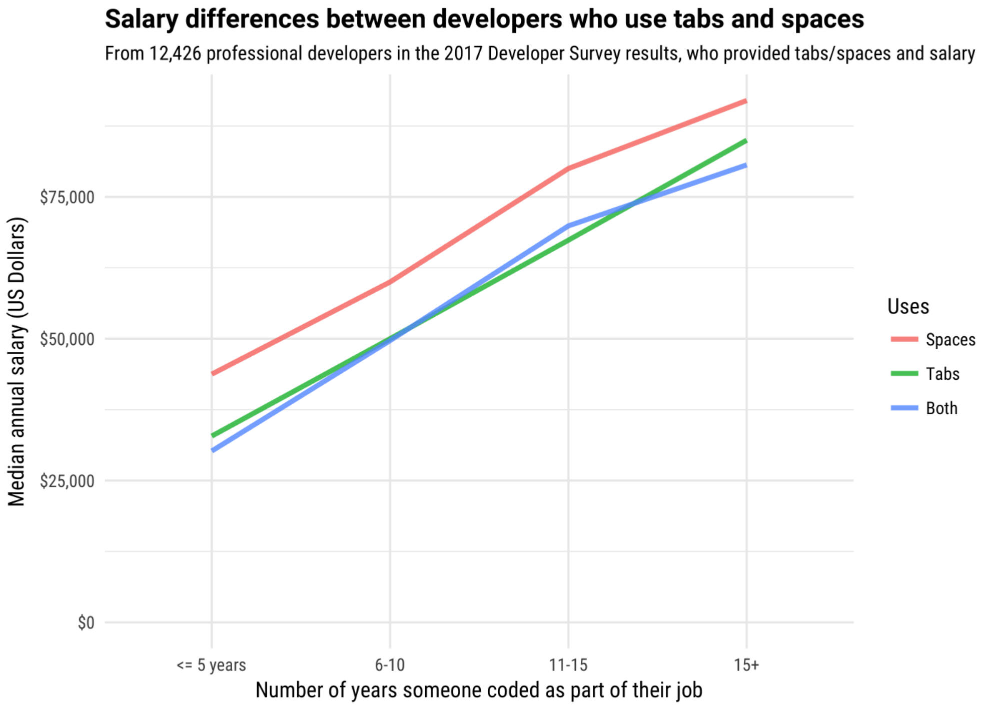
Previous conflicts

- Spaces versus Tabs
- Emacs versus Vim

- Source: https://goo.gl/images/U2KpcG

Spaces versus Tabs

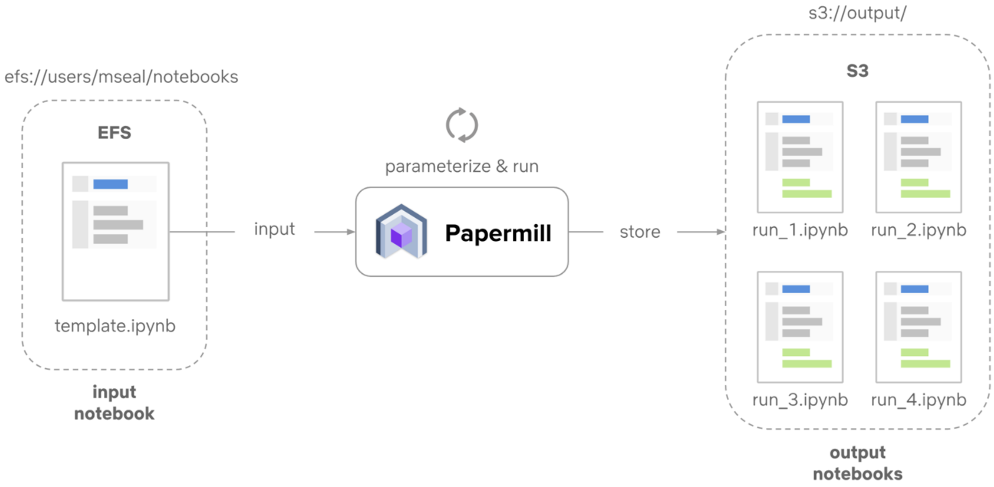
- Code editor setup: Tab = 4 spaces
- GNU Make requires tabs!
- Use spaces, get paid more!
  - according to blog post by David Robinson (@drob)'s



Notebooks

Jupyter notebooks

- are data science tools
- built on IPython by Fernando Perez (@fperez)
- combine Markdown text, code, and output
- help data scientists communicate goals, methods, and results
- used in academia, Amazon, Netflix, and PayPal

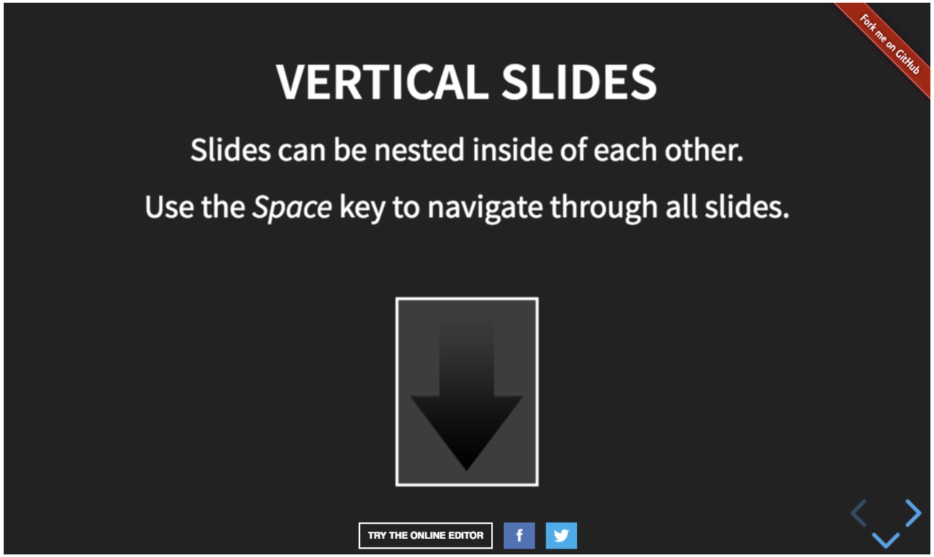


Joel doesn't like notebooks

- "I Don't Like Notebooks" by Joel Grus (@joelgrus) at JupyterCon 2018
- Slides
- Video
- Modularity and Reusability

Joel also doesn't like vertical slides

PLEASE COME TO MY NEXT TALK:
"I DON'T LIKE THAT SLIDESHOW PROGRAM WHERE SOMETIMES THE NEXT SLIDE IS RIGHT AND SOMETIMES THE NEXT SLIDE IS DOWN"



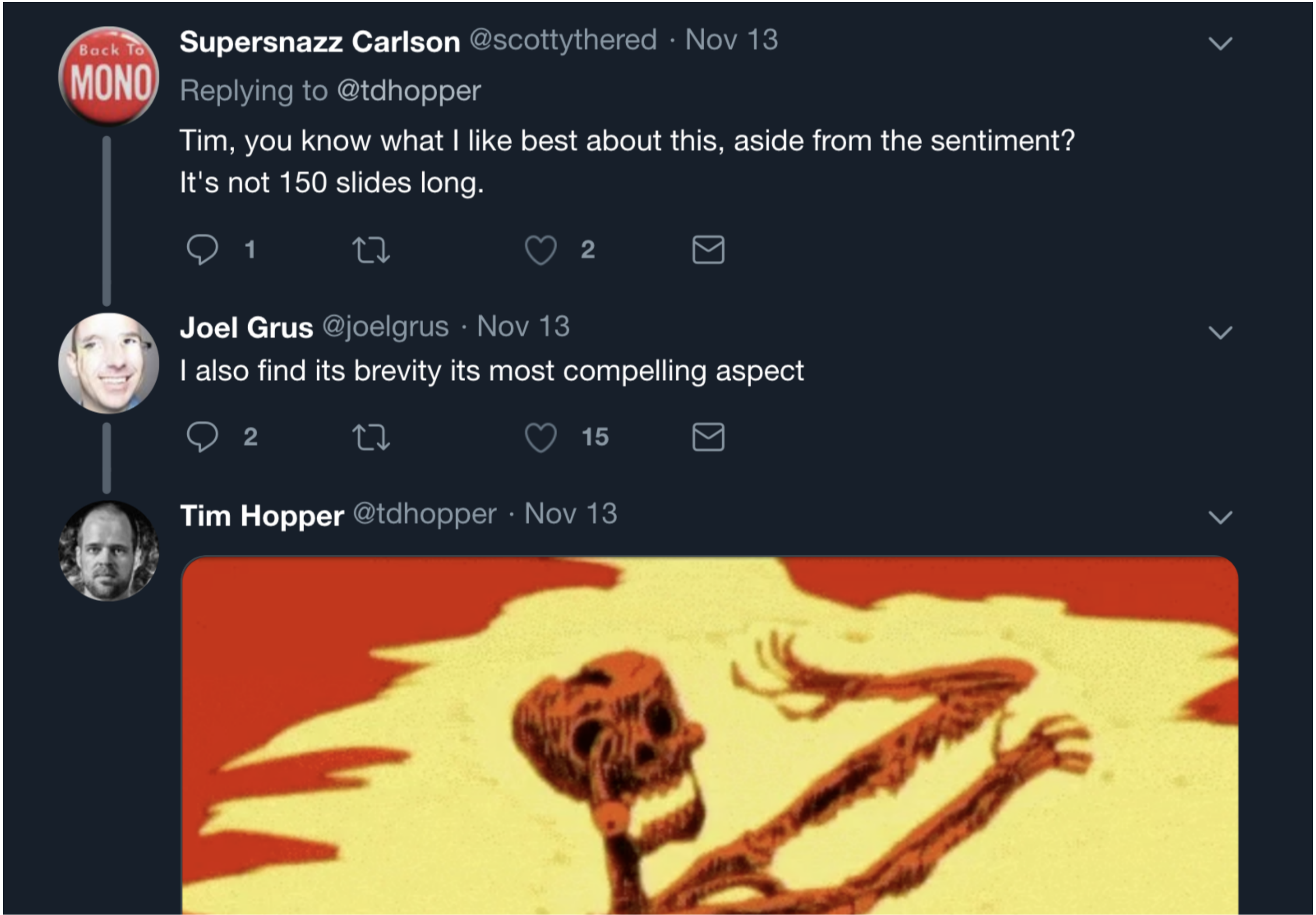
VERTICAL SLIDES

Slides can be nested inside of each other.

Use the *Space* key to navigate through all slides.

@joelgrus #jupytercon

Tim Hopper (@tdhopper) likes notebooks



Supersnazz Carlson @scottythered · Nov 13

Replying to @tdhopper

Tim, you know what I like best about this, aside from the sentiment? It's not 150 slides long.

1 2

Joel Grus @joelgrus · Nov 13

I also find its brevity its most compelling aspect

2 15

Tim Hopper @tdhopper · Nov 13

DataFramed

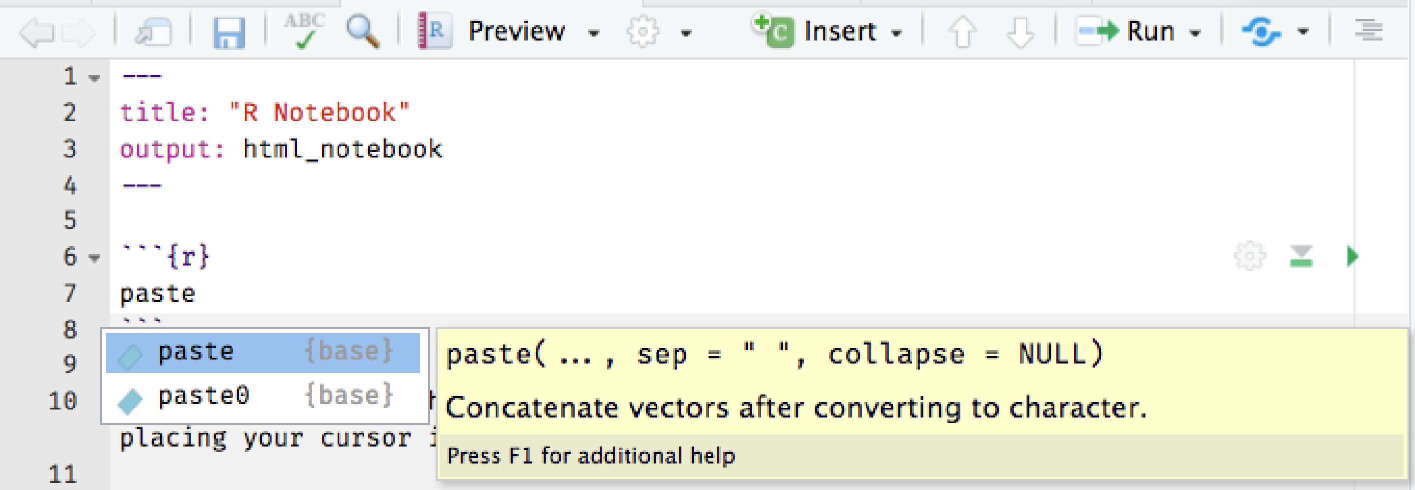
- Podcast by Hugo Bowne-Anderson (@hugobowne)
- Episode 44 with Brian Granger (@ellisonbg)
- JupyterLab

« We are entering an era where large, complex organizations need to scale interactive computing with data to their entire organization in a manner that is collaborative, secure, and human centered. »

BRIAN GRANGER

Yihui Xie (@xieyihui)'s Blog post

- I used R markdown to make these slides!
- No problems with version control
- R notebooks are just another output format!



~~~
---
title: "R Notebook"
output: html_notebook
---

```{r}
paste
```

placing your cursor i
~~~

```
paste       {base}
paste0      {base}
```

paste(... , sep = " ", collapse = NULL)

Concatenate vectors after converting to character.

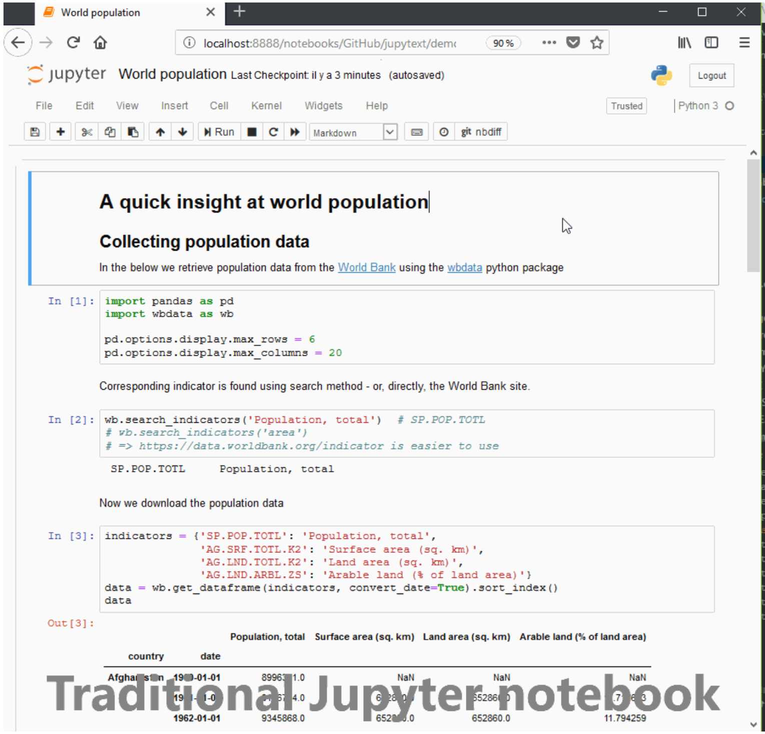
Press F1 for additional help

Why use notebooks

- Literate programming
- Rendered by GitHub and nbviewer
- Google colab
- Kaggle kernels
- Binder

Notebook tools

1. version control tool for notebooks - nbdime
2. work with Jupyter notebooks and scripts in parallel using JupyText
3. configure notebooks to run on markdown (md) files with notedown
4. create and run Jupyter notebooks from scripts and md files with nbless



jupyter World population Last Checkpoint: il y a 3 minutes (autosaved)

A quick insight at world population

Collecting population data

In the below we retrieve population data from the World Bank using the wbdata python package

```
In [1]: import pandas as pd
        import wbdata as wb

        pd.options.display.max_rows = 6
        pd.options.display.max_columns = 20
```

Corresponding indicator is found using search method - or, directly, the World Bank site.

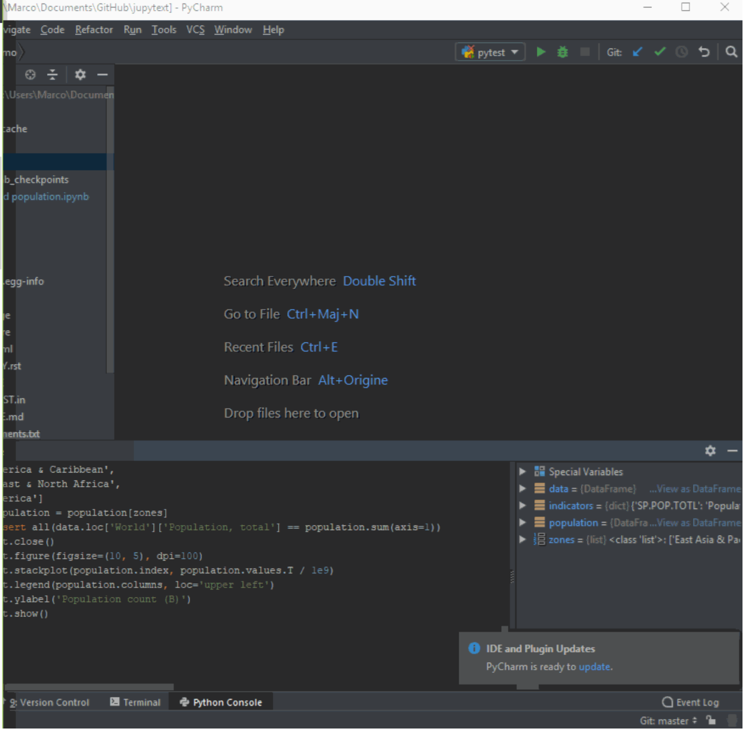
```
In [2]: wb.search_indicators('Population, total')  # SP.POP.TOTL
        # wb.search_indicators('area')
        # => https://data.worldbank.org/indicator is easier to use

         SP.POP.TOTL     Population, total
```

Now we download the population data

```
In [3]: indicators = {'SP.POP.TOTL': 'Population, total',
                      'AG.SRF.TOTL.K2': 'Surface area (sq. km)',
                      'AG.LND.TOTL.K2': 'Land area (sq. km)',
                      'AG.LND.ARBL.ZS': 'Arable land (% of land area)'}
        data = wb.get_dataframe(indicators, convert_date=True).sort_index()
        data
Out[3]:
```

		Population, total	Surface area (sq. km)	Land area (sq. km)	Arable land (% of land area)
country	date				
Afghanistan	[illegible]	[illegible]	NaN	NaN	NaN
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	1962-01-01	9345868.0	[illegible]	652860.0	11.794259

Traditional Jupyter notebook



Search Everywhere Double Shift

Go to File Ctrl+Maj+N

Recent Files Ctrl+E

Navigation Bar Alt+Origine

Drop files here to open

```
erica & Caribbean',
ast & North Africa',
erica']
pulation = population[zones]
sert all(data.loc['World']['Population, total'] == population.sum(axis=1))
t.close()
t.figure(figsize=(10, 5), dpi=100)
t.stackplot(population.index, population.values.T / 1e9)
t.legend(population.columns, loc='upper left')
t.ylabel('Population count (B)')
t.show()
```

Special Variables
data = {DataFrame} ...View as DataFrame
indicators = {dict} {'SP.POP.TOTL': 'Popula
population = {DataFra ...View as DataFrame
zones = {list} <class 'list'>: ['East Asia & Pa

IDE and Plugin Updates
PyCharm is ready to update.

Write modules!

- Imports
    1. Standard Library
    2. Third Party
    3. User Defined
- Definitions
    - Classes
    - Functions (for more check out Steven Lott's PyData DC tutorial)
- Type Hints
- Docstrings (with examples!)
- `if __name__ == '__main__':`
- Function call(s), e.g. `doctest.testmod(verbose=True)`
- doctest: docstring examples -> test suite (with `unittest` API)
- run test suite with `unittest` or `pytest`
- use cookiecutter for project structure
- deploy projects/packages to PyPI

Thanks for listening!