

# Introduction to Data Science

## Data Science Essentials

---

Mary van Valkenburg

Data Science Program Manager/Instructor

Nashville Software School

# Goals for today

- Review last session coding tasks
- A little more matplotlib
- Combining DataFrames (merge/concat)
- Feature engineering
- Choropleths

# Review last session coding tasks

**week3\_review** notebook

# More matplotlib – fig and ax

**week3\_review** notebook

# Combining DataFrames

## Concatenating two DataFrames:

***pd.concat***([<df1>, <df2>, <df3>])      pass a *list* of dataframes to concatenate

df1					Result				
	A	B	C	D		A	B	C	D
0	A0	B0	C0	D0	0	A0	B0	C0	D0
1	A1	B1	C1	D1	1	A1	B1	C1	D1
2	A2	B2	C2	D2	2	A2	B2	C2	D2
3	A3	B3	C3	D3	3	A3	B3	C3	D3
df2					4	A4	B4	C4	D4
	A	B	C	D	5	A5	B5	C5	D5
4	A4	B4	C4	D4	6	A6	B6	C6	D6
5	A5	B5	C5	D5	7	A7	B7	C7	D7
6	A6	B6	C6	D6	8	A8	B8	C8	D8
7	A7	B7	C7	D7	9	A9	B9	C9	D9
df3					10	A10	B10	C10	D10
	A	B	C	D	11	A11	B11	C11	D11
8	A8	B8	C8	D8					
9	A9	B9	C9	D9					
10	A10	B10	C10	D10					
11	A11	B11	C11	D11					

- Same columns
- Like pasting them together

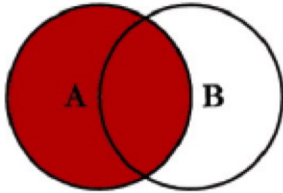
## Merging two DataFrames:

***pd.merge***(<df1>, <df2>, **on** = <col or list of cols to join on>, **how** = <join\_type>)

### pandas merge types

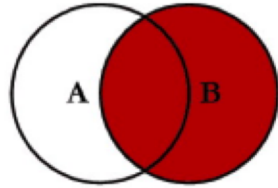
#### how = 'left'

Keeps all rows from the left table and only the matching rows from the right table.



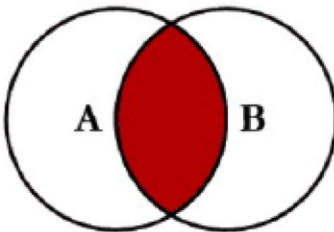
#### how = 'right'

Keeps all rows from the right table and only the matching rows from the left table.



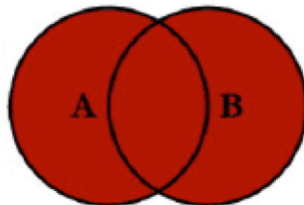
#### how = 'inner'

Keeps only rows that have a match in both tables.



#### how = 'outer'

Keeps all rows from both tables, whether they match on the specified key or not.



- One or more matching columns (keys)

# Feature Engineering

- **Create more meaningful features**
  - A statistic that compares the annual total cost of care by county to the county's average income (cost\_income\_ratio)
  - Others?
    - Average income per person (exemptions can be a proxy for person count)?



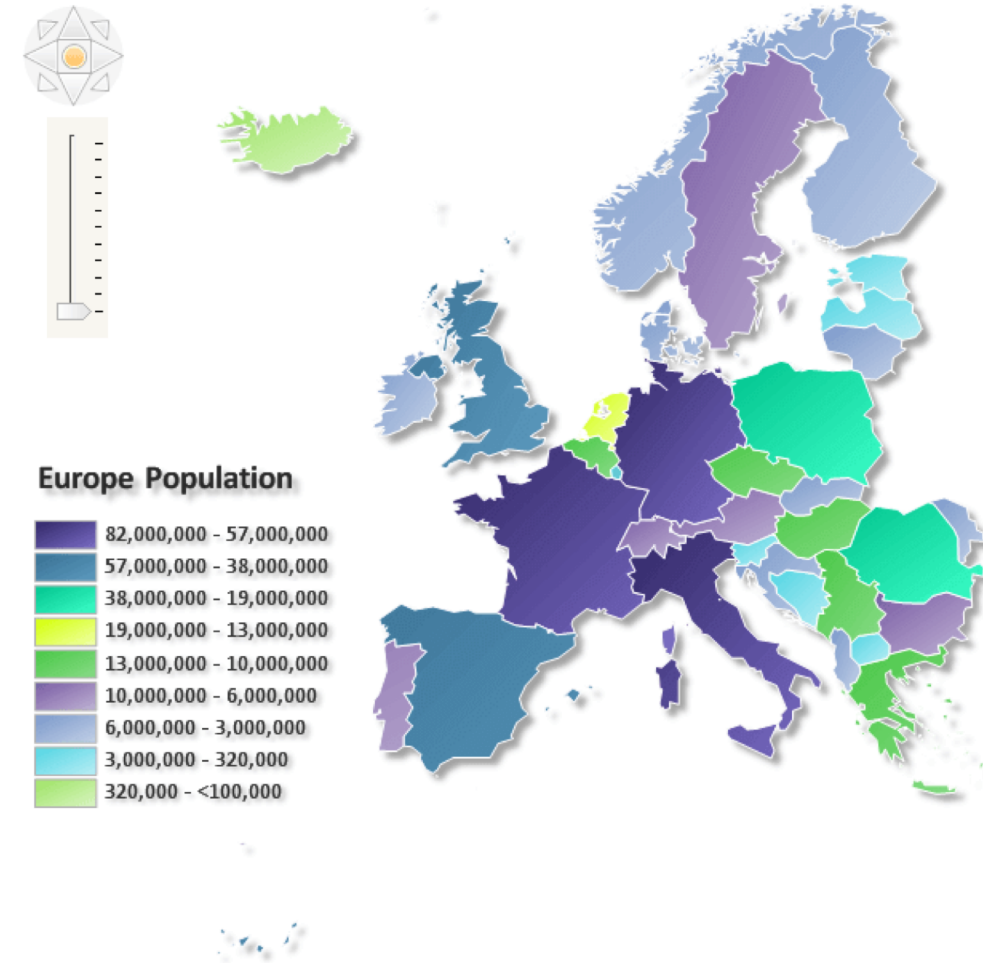
# Choropleths

A choropleth is a map where areas are colored or shaded according to the value of some aggregate statistic for that area (eg. average income, population density, unemployment rate, etc.)

We will create choropleths in Python by using the *geopandas* library, which you will most likely need to install.

To install geopandas, open the Terminal (Mac) or Anaconda Prompt (Windows) and type

\$ `conda install geopandas`



# Building a choropleth

**Choropleth\_Tutorial** notebook

**Questions?**