The pandas Library

Ivan Corneillet

Data Scientist



Learning Objectives

After this lesson, you should be able to:

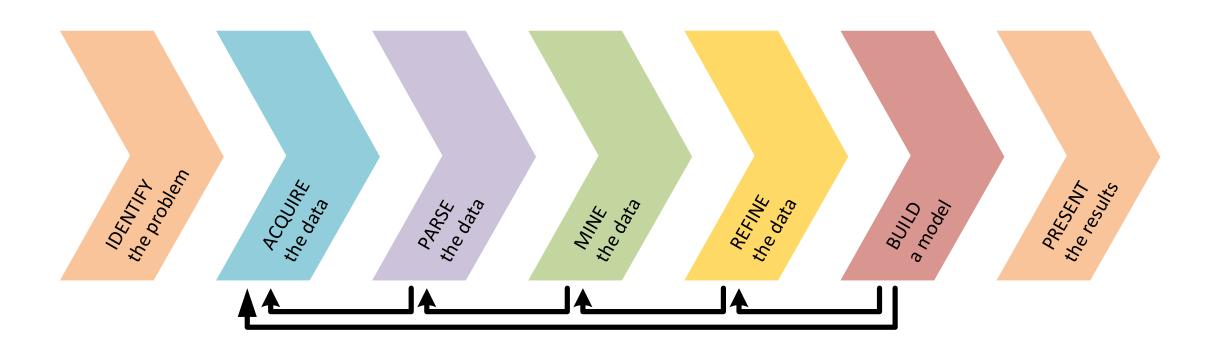
• Write an Jupyter notebook to import, format, and clean data using the *pandas* library

The San Francisco housing dataset: a dataset we will use throughout this course



- Recently Sold Homes (Source: Zillow)
 - 1,000 homes sold in San Francisco between 11/10/2015
 and 2/12/2016

The Data Science Workflow



When reaching the 6 BUILD a model step, our data needs to be tidy and in the form of a feature matrix X (i.e., the stimuli, e.g., "ring bell") and a response vector y (i.e., the response, e.g., "dog salivates")

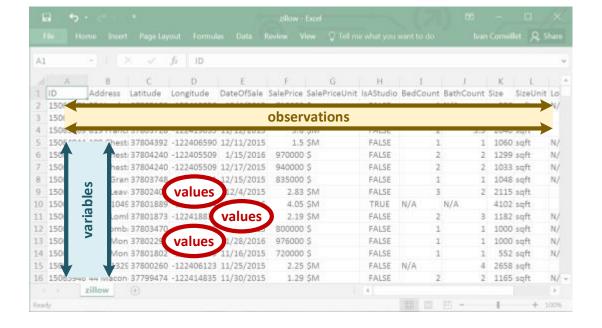
Feature Matrix *X*

Response Vector *y*

	col0	col1	col2	col3		col
row0					row0	
row1					row1	
row2					row2	
row3					row3	

What is tidy data?

- Your data is tidy if you follow these three rules:
 - Each **observation** (or **sample**) in the dataset is placed in its own **row**
 - Each variable (or feature) is placed
 in its own column
 - Each value is placed in its own cell



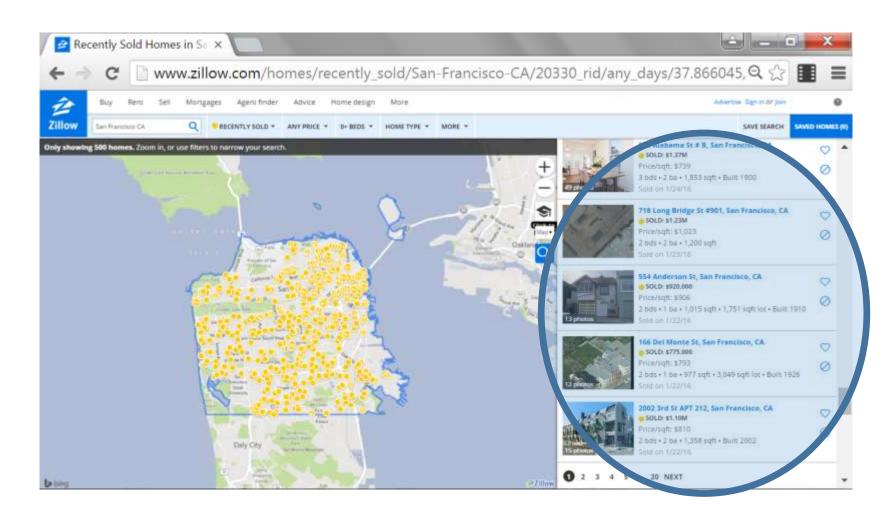
Unfortunately, the output of step **②** ACQUIRE the data will usually be raw, i.e., unstructured...



<div class="property-info"</pre> id="yui 3 18 1 1 1456167242885 71870"><strong id="yui_3_18_1_1_1456167242885_71869"><dt class="property-address" id="yui_3_18_1_1_1456167242885_71868">149 Shipley St, San Francisco, CA</dt><dt class="listing-type zsgcontent_collapsed" id="yui_3_18_1_1_1456167242885_71875"><span</pre> class="zsg-icon-recently-sold type-icon">Sold: \$1.18M</dt><dt class="zsg-fineprint" id="yui 3 18 1 1 1456167242885 71877">Price/sqft: \$1,116</dt><dt class="property-data" id="yui_3_18_1_1_1456167242885_71880">3 bds • 2 ba • 1,057 sqft<span</pre> class="built-year" id="yui_3_18_1_1_1456167242885_71879"> • Built 1992</dt><dt class="sold-date zsg-fineprint" id="yui_3_18_1_1_1456167242885_71975">Sold on 2/22/16</dt></div>

(E.g., unstructured data scrapped from websites)





... and/or messy...



Trouble tickets inspect and maintain manholes in New Year
 City

* "Service box," a common piece of infrastructure, had at least 38 variants, including SB, S, S/B, S.B, S?B, S.B., SBX, S/BX, SB/X, S/XB, /SBX, S.BX, S &BX, S?BX, S BX, S/B/X, S BOX, SVBX, SERV BX, SERV-BOX, SERV/BOX, and SERVICE BOX

Source: Big Data: A Revolution That Will Transform How We Live, Work, and Think

Question: what tool can we use to wrangle raw data into tidy data?

- Answer: pandas
 - pandas is a Python library that provides the ability to index, retrieve, tidy, reshape, combine, slice, tabular and other multidimensional datasets
 - pandas also provides facilities to perform statistical and mathematical analysis which will come handy for exploratory data Analysis
- Wrangling data is the most fruitful skill you can learn as a data scientist. It will save you hours of time and make your data much easier to visualize, manipulate, and model
- Today, we will use pandas to explore and manipulate the San Francisco housing dataset

pandas.DataFrame and pandas.Series (cont.)

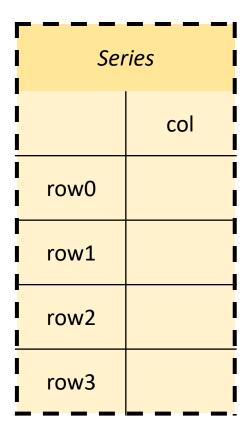
DataFrame						
		col0	col1	col2	col3	
	row0					
	row1					
	row2					Series
	row3					
		DataFrame			Series	

When reaching the 6 BUILD a model step, our feature matrix X will be modeled as a DataFrame and the response vector y as a Series

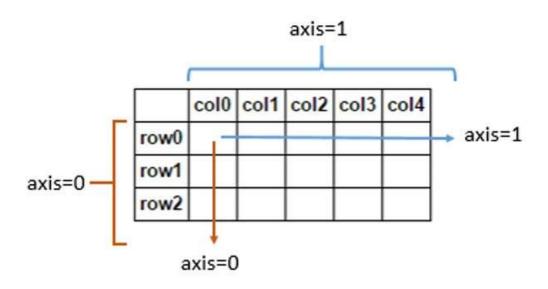
Feature Matrix *X*

DataFrame							
i I	col0	col1	col2	col3			
row0							
row1							
row2							
row3				I			

Response Vector *y*



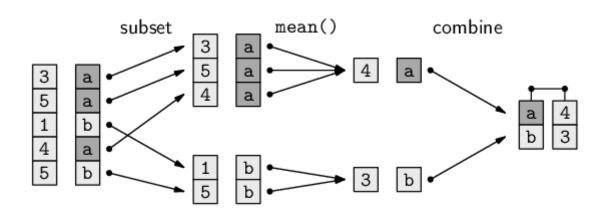
(a note on *pandas* axes)



Courses Steels Overflow

- Axes are defined for arrays with more than one dimension, e.g., a
 DataFrame
- A DataFrame has two corresponding
 axes: the first running vertically
 downwards across rows (axis o),
 and the second running horizontally
 across columns (axis 1)

(another note of the split-apply-combine performed by .groupby())



Source: Python for Social Scientists

Slides © 2016 Ivan Corneillet Where Applicable Do Not Reproduce Without Permission