03 | *pandas*

Ivan Corneillet

Data Scientist

GENERAL ASSEMBLY

Learning Objectives

After this lesson, you should be able to:

- Define a data science problem
- Write a Jupyter notebook to import, format, and clean data using *pandas*

Business Understanding

By asking a good question and setting a clear aim:

- You set yourself up for success
  - "A problem well stated is half solved" – Charles Kettering
- You help other data scientists learn from and reproduce your work
  - You establish the basis for making your analysis reproducible
- You also help them expand on your work in the future

The SMART Goals Framework for Data Science

(https://en.wikipedia.org/wiki/SMART_criteria)

SPECIFIC	The dataset and key variables are clearly defined
MEASURABLE	The type of analysis and major assumptions are articulated
ATTAINABLE	The question you are asking is feasible for your dataset and is not likely to be biased
REPRODUCIBLE	Another person (or you in 6 months!) can read your state and understand exactly how your analysis is performed
TIME-BOUND	You clearly state the time period and population for which this analysis will pertain

Trends often change over time and vary by the population of source of your data. It is important to clearly define who/what you included in your analysis as well as the time period for the analysis

Models, Feature Matrix X, Response Vector y, and Tidy Data

Before **modeling**, our data needs to be **tidy** and in the form of a **feature matrix $\boldsymbol{X}$** (i.e., the stimuli, e.g., *“ring bell”*) and a **response vector $\boldsymbol{y}$** (i.e., the response, e.g., *“dog salivates”*)

Feature Matrix X

	col0	col1	col2	col3
row0				
row1				
row2				
row3				

Response Vector y

	col
row0	
row1	
row2	
row3	

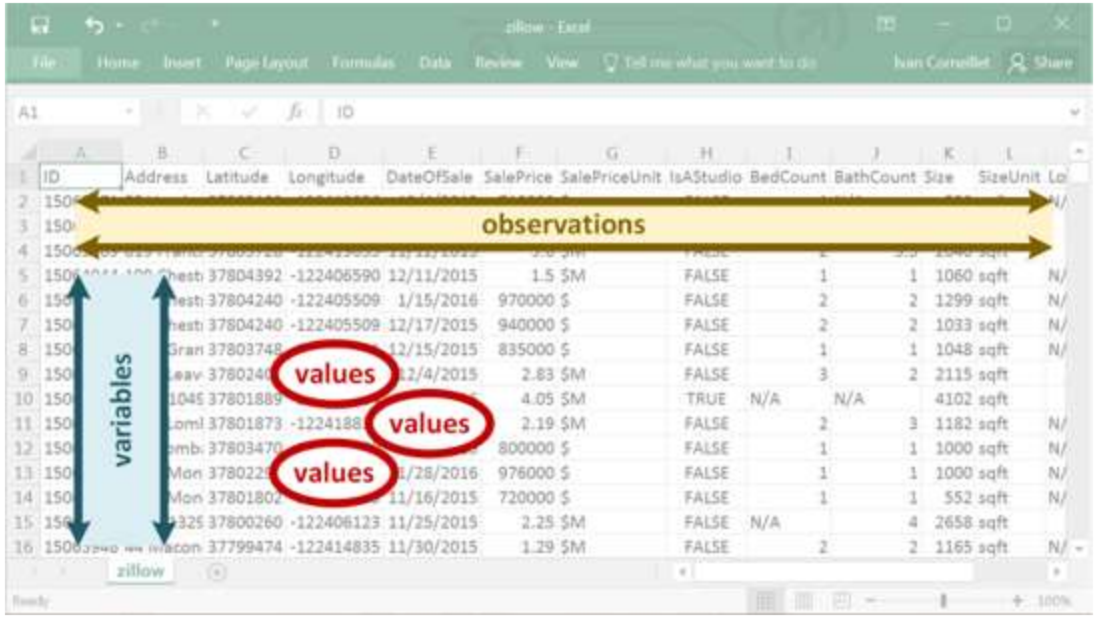
San Francisco Dataset: a dataset we will use throughout this course

CASE STUDY

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

What is Tidy Data?

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



	A	B	C	D	E	F	G	H	I	J	K	L	
1	ID	Address	Latitude	Longitude	DateOfSale	SalePrice	SalePriceUnit	IsAStudio	BedCount	BathCount	Size	SizeUnit	Lo
2	150	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	N/
3	150												
4	150	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
5	150	Chest	37804392	-122406590	12/11/2015	1.5	$M	FALSE	1	1	1060	sqft	N/
6	150	est	37804240	-122405509	1/15/2016	970000	$	FALSE	2	2	1299	sqft	N/
7	150	hest	37804240	-122405509	12/17/2015	940000	$	FALSE	2	2	1033	sqft	N/
8	150	Gran	37803748	[illegible]	12/15/2015	835000	$	FALSE	1	1	1048	sqft	N/
9	150	Leav	3780240	[illegible]	2/4/2015	2.83	$M	FALSE	3	2	2115	sqft	
10	150	1045	37801889	[illegible]	[illegible]	4.05	$M	TRUE	N/A	N/A	4102	sqft	
11	150	Loml	37801873	-1224188	[illegible]	2.19	$M	FALSE	2	3	1182	sqft	N/
12	150	mb	37803470	[illegible]	[illegible]	800000	$	FALSE	1	1	1000	sqft	N/
13	150	Mon	3780225	[illegible]	1/28/2016	976000	$	FALSE	1	1	1000	sqft	N/
14	150	Mon	37801802	[illegible]	11/16/2015	720000	$	FALSE	1	1	552	sqft	N/
15	150	325	37800260	-122406123	11/25/2015	2.25	$M	FALSE	N/A	4	2658	sqft	
16	150	Macon	37799474	-122414835	11/30/2015	1.29	$M	FALSE	2	2	1165	sqft	N/

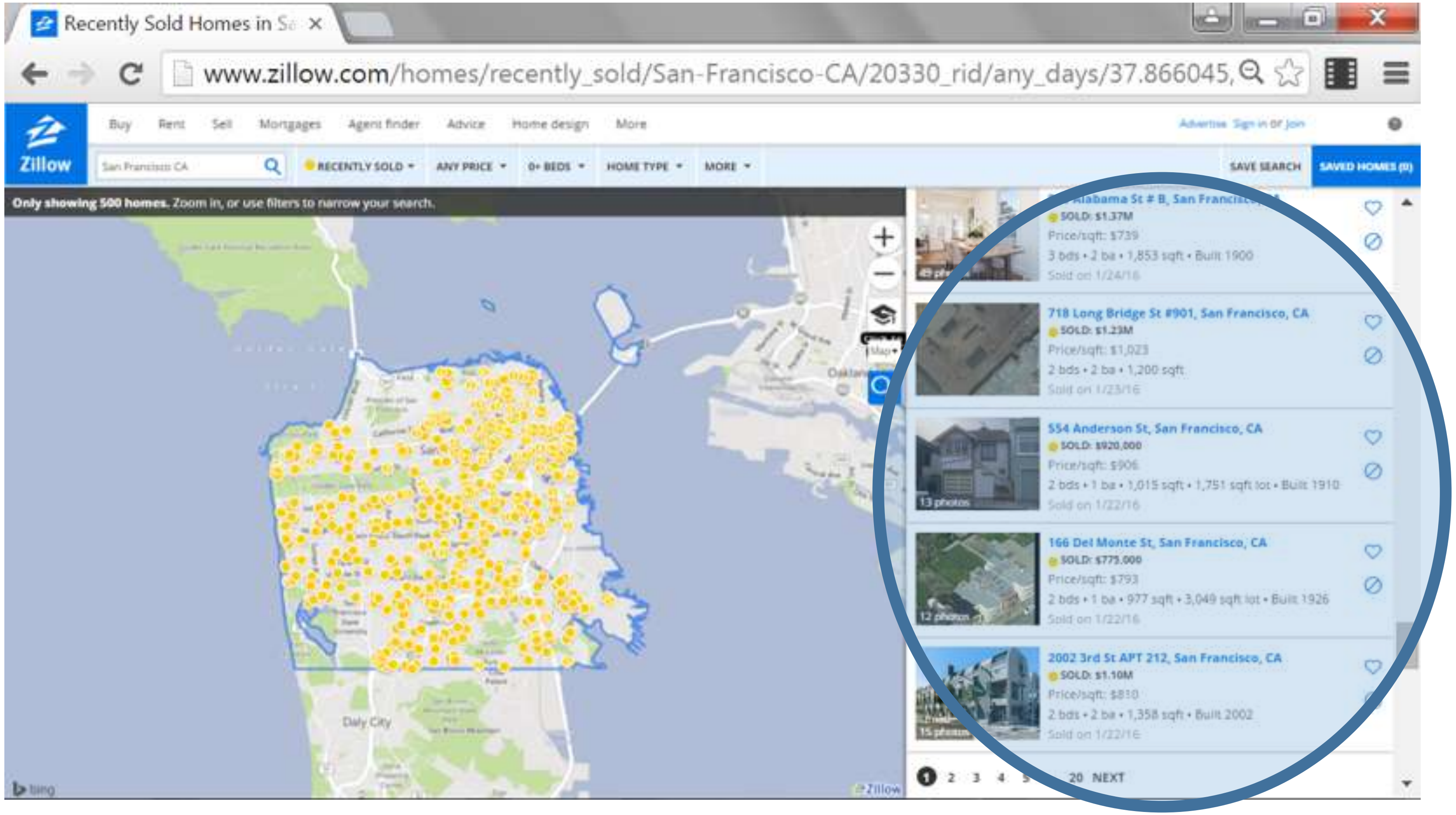
Unfortunately, data will usually come to you **raw**, i.e., **unstructured…**

CASE STUDY

```
<div class="property-info"
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"><a href="/homedetails/149-
Shipley-St-San-Francisco-CA-94107/15147894_zpid/" class="hdp-link
routable" title="149 Shipley St, San Francisco, CA Real Estate"
id="yui_3_18_1_1_1456167242885_71873">149 Shipley St, San
Francisco, CA</a></dt></strong><dt class="listing-type zsg-
content_collapsed" id="yui_3_18_1_1_1456167242885_71875"><span
class="zsg-icon-recently-sold type-icon"></span>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"><span
class="beds-baths-sqft">3 bds • 2 ba • 1,057 sqft</span><span
class="built-year" id="yui_3_18_1_1_1456167242885_71879"> • Built
1992</span></dt><dt class="sold-date zsg-fineprint"
id="yui_3_18_1_1_1456167242885_71975">Sold on 2/22/16</dt></div>
```

(E.g., raw/unstructured scrapped data)

CASE STUDY



www.zillow.com/homes/recently_sold/San-Francisco-CA/20330_rid/any_days/37.866045,

Only showing 500 homes. Zoom in, or use filters to narrow your search.

Alabama St # B, San Francisc
SOLD: $1.37M
Price/sqft: $739
3 bds • 2 ba • 1,853 sqft • Built 1900
Sold on 1/24/16

718 Long Bridge St #901, San Francisco, CA
SOLD: $1.23M
Price/sqft: $1,023
2 bds • 2 ba • 1,200 sqft
Sold on 1/23/16

554 Anderson St, San Francisco, CA
SOLD: $920,000
Price/sqft: $906
2 bds • 1 ba • 1,015 sqft • 1,751 sqft lot • Built 1910
Sold on 1/22/16

166 Del Monte St, San Francisco, CA
SOLD: $775,000
Price/sqft: $793
2 bds • 1 ba • 977 sqft • 3,049 sqft lot • Built 1926
Sold on 1/22/16

2002 3rd St APT 212, San Francisco, CA
SOLD: $1.10M
Price/sqft: $810
2 bds • 2 ba • 1,358 sqft • Built 2002
Sold on 1/22/16

… and/or **messy**…

EXAMPLE

- 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

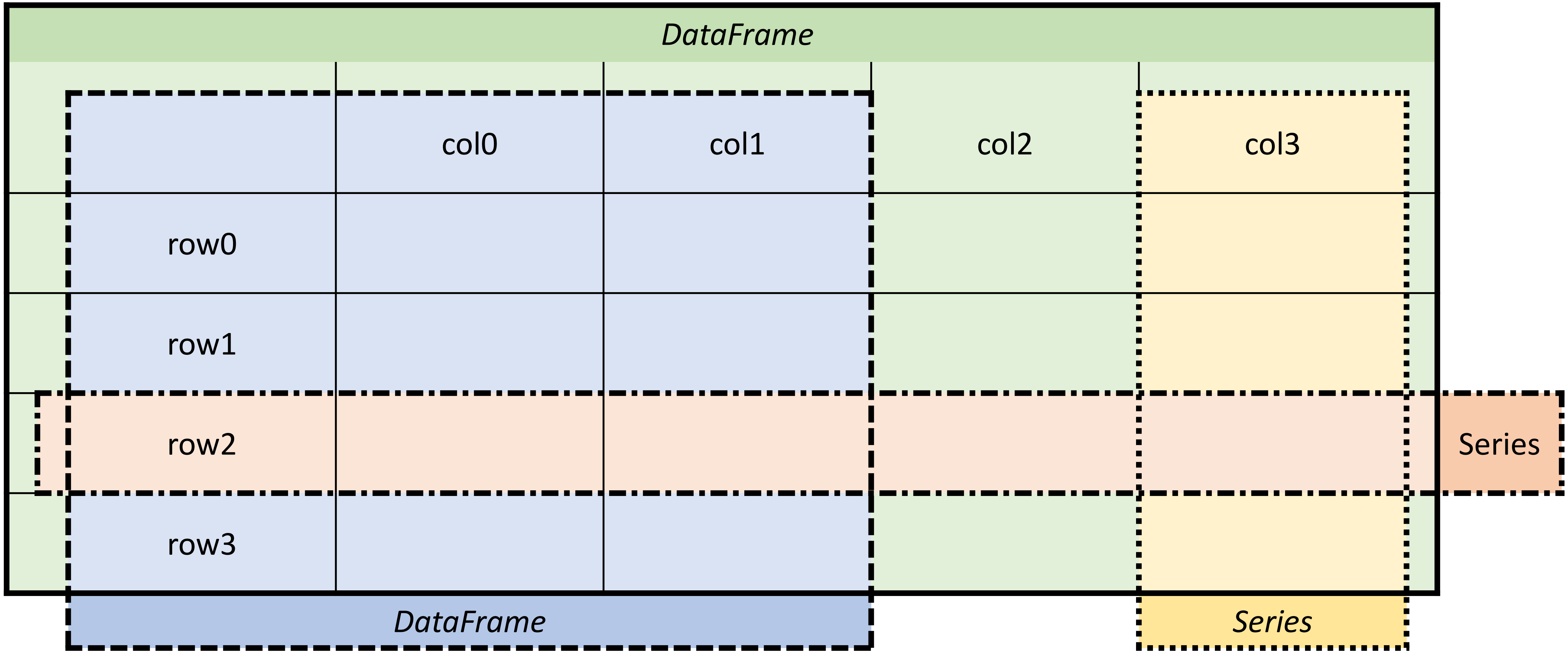
Question: What tool can we use to wrangle **raw data** into **tidy data**?

- 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

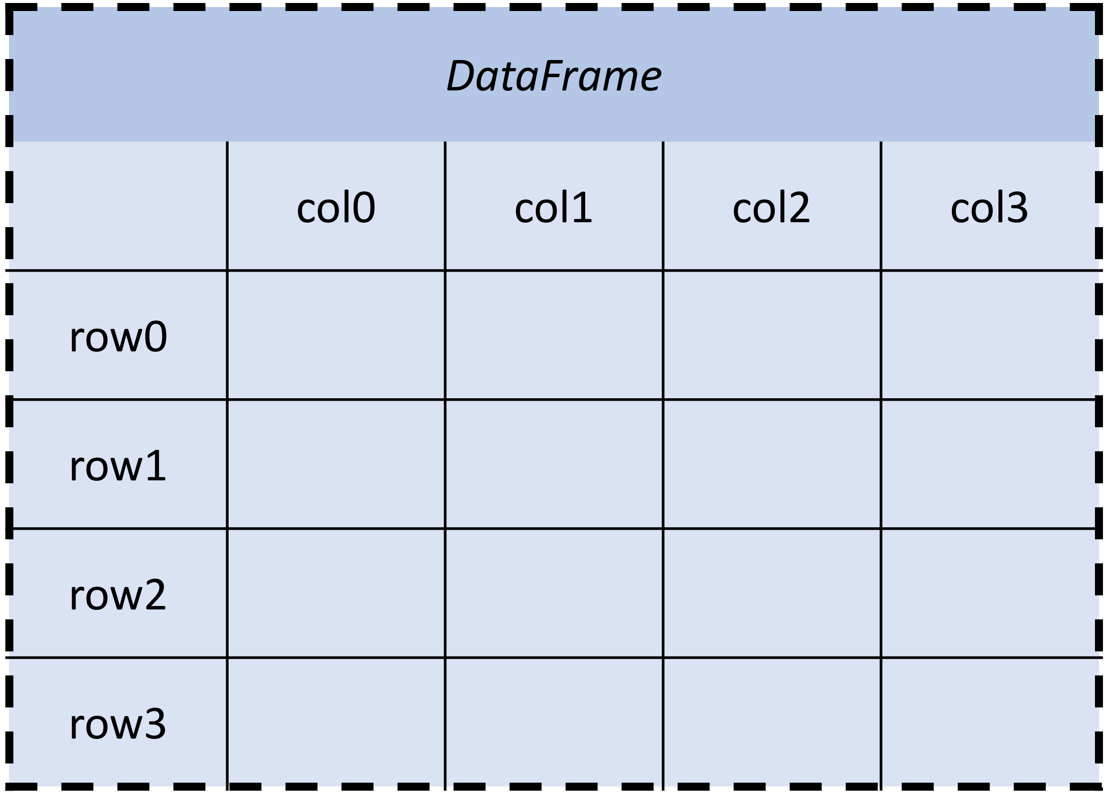
pandas

pandas.DataFrame and *pandas.Series* (cont.)



When reaching the **modeling** step, our **feature matrix $\boldsymbol{X}$** will be modeled as a **DataFrame** and the **response vector $\boldsymbol{y}$** as a **Series**

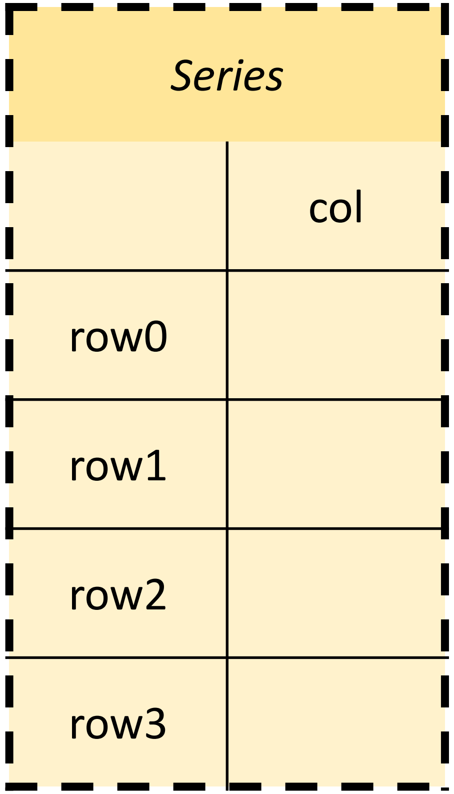
Feature Matrix X



DataFrame

	col0	col1	col2	col3
row0				
row1				
row2				
row3				

Response Vector y



Series

	col
row0	
row1	
row2	
row3	