DATA HANDLING IN PYTHON MODULE - 4

SINCE WE NEED TO HANDLE HUGE AMOUNTS OF DATA, WE WILL BE IMPLEMENTING DATA HANDLING TECHNIQUES WITH PANDAS LIBRARY. AND WE WILL EXPLORE THE DIFFERENT MISCELLANEOUS FUNCTIONS OF PANDAS LIBRARY IN DETAIL.

SOURCE CODE

- Numpy
  - https://colab.research.google.com/drive/1ZYlWEmkbv1OfnIzRe2JaCFl42FlUd7Af?usp=sharing
- Pandas
  - https://colab.research.google.com/drive/1nh4fONray99z9vSbb8_5jakYAjHV6zYl?usp=sharing

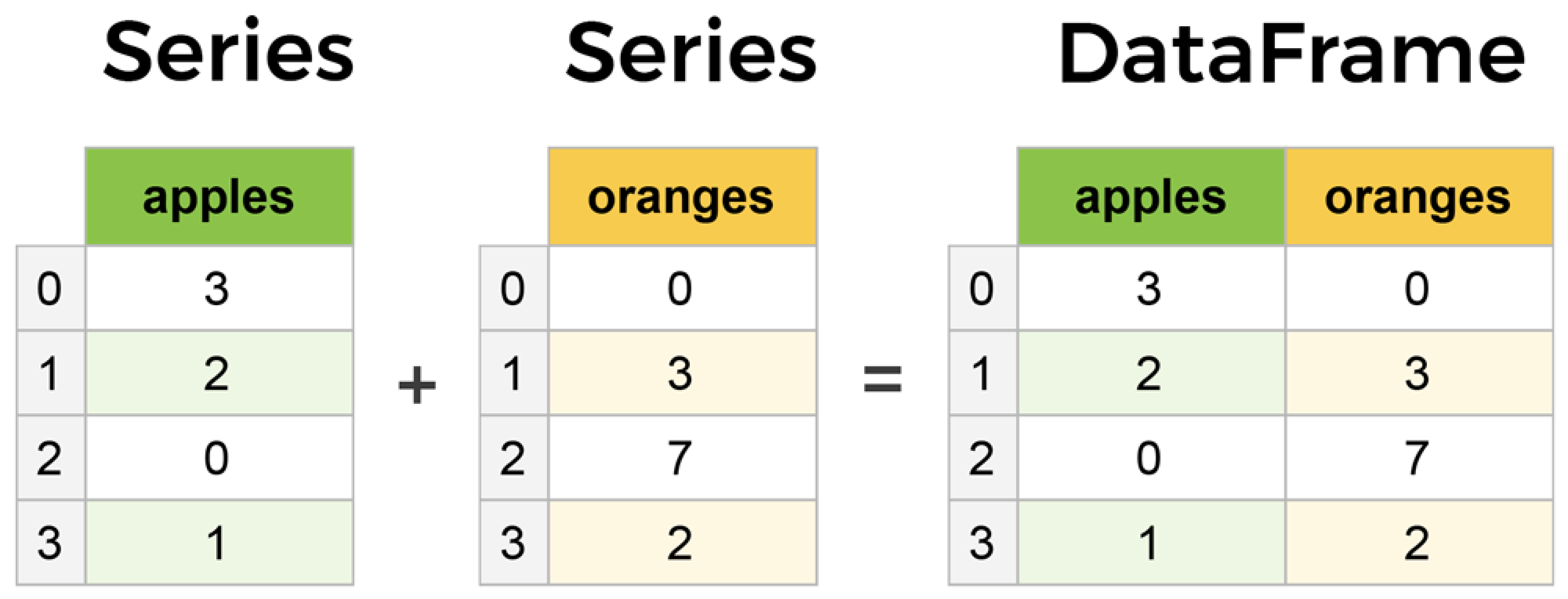
THE PANDAS PACKAGE

- is the most important tool at the disposal of Data Scientists and Analysts
- backbone of most data projects
- makes you get acquainted with your data by cleaning, transforming, and analyzing it.
- For example, say you want to explore a dataset stored in a CSV on your computer. Pandas will extract the data from that CSV into a DataFrame — a table, basically — then let you do things like:
  - Calculate statistics and answer questions about the data, like
    - What's the average, median, max, or min of each column?
    - Does column A correlate with column B?
    - What does the distribution of data in column C look like?
  - Clean the data by doing things like removing missing values and filtering rows or columns by some criteria
  - Visualize the data with help from Matplotlib. Plot bars, lines, histograms, bubbles, and more.
  - Store the cleaned, transformed data back into a CSV, other file or database

INSTALL AND IMPORT PANDAS

CORE COMPONENTS OF PANDAS

- The primary two components of pandas are the Series and DataFrame.
- A Series is essentially a column, and a DataFrame is a multi-dimensional table made up of a collection of Series.



Series

	apples
0	3
1	2
2	0
3	1

+

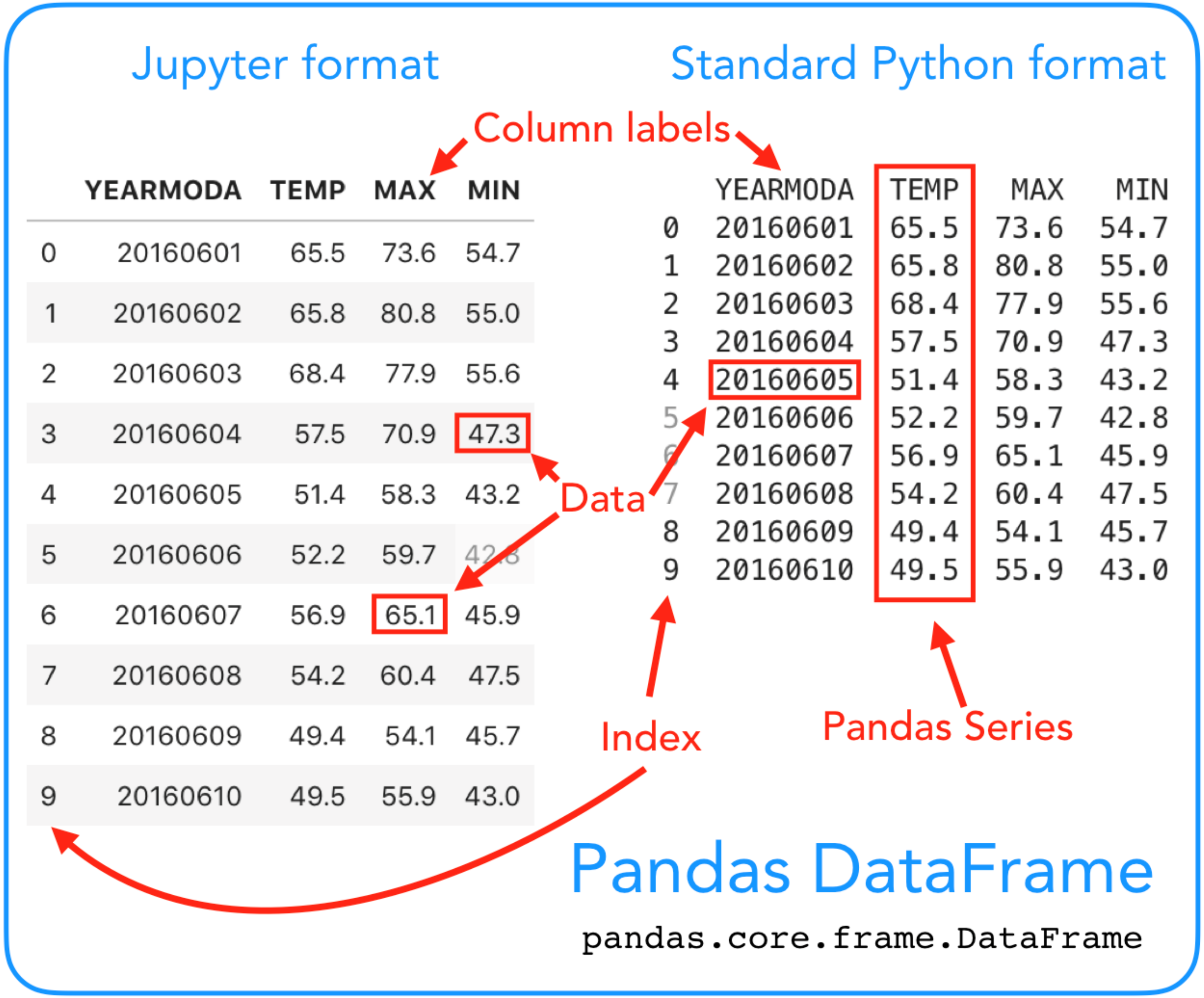
Series

	oranges
0	0
1	3
2	7
3	2

=

DataFrame

	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2



Jupyter format

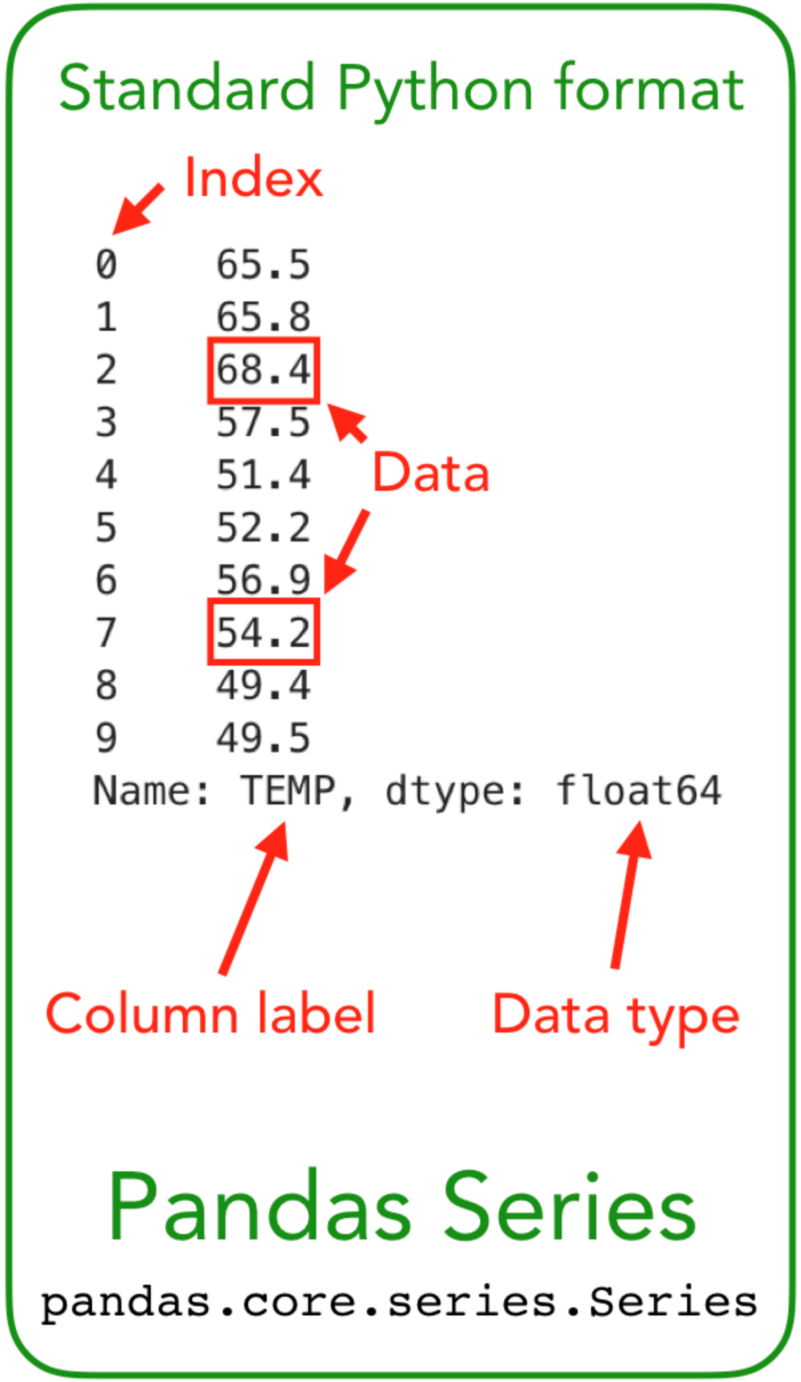
	YEARMODA	TEMP	MAX	MIN
0	20160601	65.5	73.6	54.7
1	20160602	65.8	80.8	55.0
2	20160603	68.4	77.9	55.6
3	20160604	57.5	70.9	47.3
4	20160605	51.4	58.3	43.2
5	20160606	52.2	59.7	42.8
6	20160607	56.9	65.1	45.9
7	20160608	54.2	60.4	47.5
8	20160609	49.4	54.1	45.7
9	20160610	49.5	55.9	43.0

Standard Python format

```
   YEARMODA  TEMP  MAX   MIN
0  20160601  65.5  73.6  54.7
1  20160602  65.8  80.8  55.0
2  20160603  68.4  77.9  55.6
3  20160604  57.5  70.9  47.3
4  20160605  51.4  58.3  43.2
5  20160606  52.2  59.7  42.8
6  20160607  56.9  65.1  45.9
7  20160608  54.2  60.4  47.5
8  20160609  49.4  54.1  45.7
9  20160610  49.5  55.9  43.0
```

Pandas DataFrame

`pandas.core.frame.DataFrame`



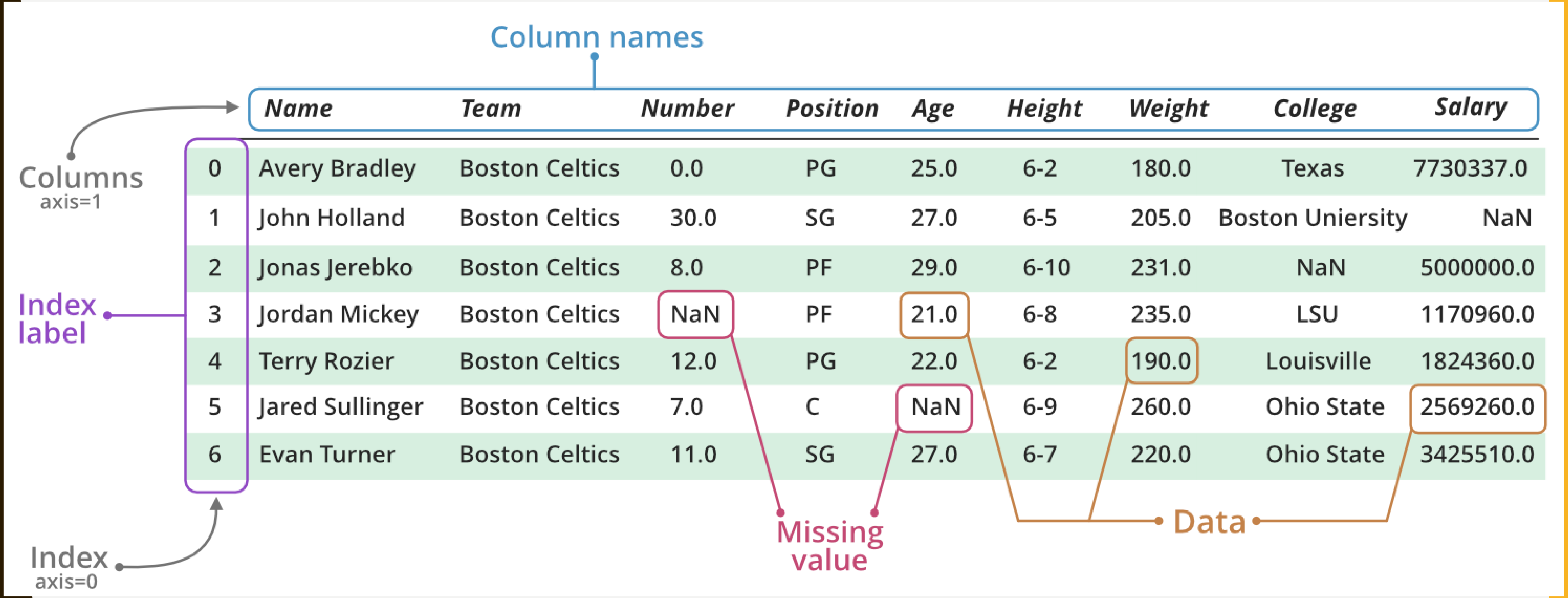
Standard Python format

```
0     65.5
1     65.8
2     68.4
3     57.5
4     51.4
5     52.2
6     56.9
7     54.2
8     49.4
9     49.5
Name: TEMP, dtype: float64
```

Pandas Series

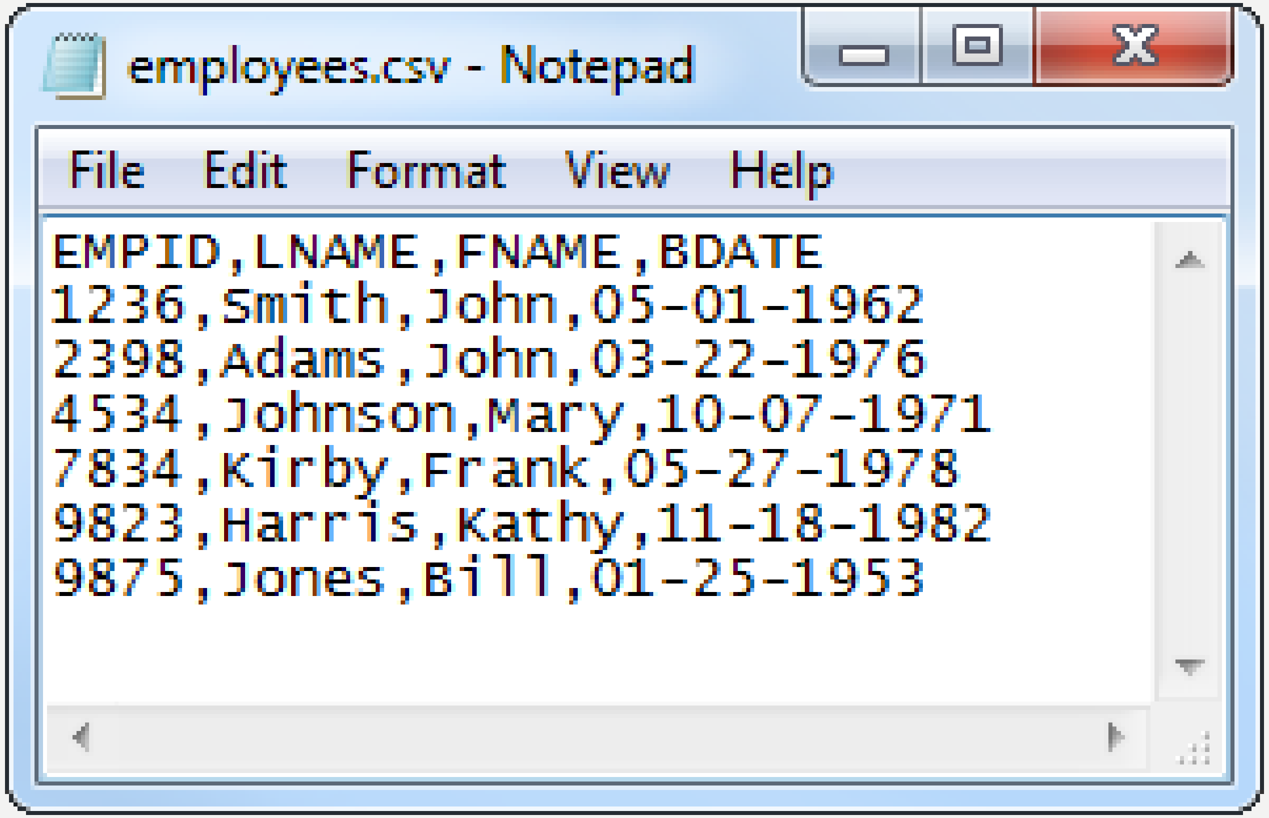
`pandas.core.series.Series`

MORE ABOUT A DATAFRAME



	Name	*Team*	*Number*	*Position*	*Age*	*Height*	*Weight*	*College*	*Salary*
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston Uniersity	NaN
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0	6-8	235.0	LSU	1170960.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0	6-2	190.0	Louisville	1824360.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN	6-9	260.0	Ohio State	2569260.0
6	Evan Turner	Boston Celtics	11.0	SG	27.0	6-7	220.0	Ohio State	3425510.0

SAMPLE CSV FILE



```
EMPID,LNAME,FNAME,BDATE
1236,Smith,John,05-01-1962
2398,Adams,John,03-22-1976
4534,Johnson,Mary,10-07-1971
7834,Kirby,Frank,05-27-1978
9823,Harris,Kathy,11-18-1982
9875,Jones,Bill,01-25-1953
```