Data Analysis with Python

Session-4

pandas

Data Frames

Table of Contents

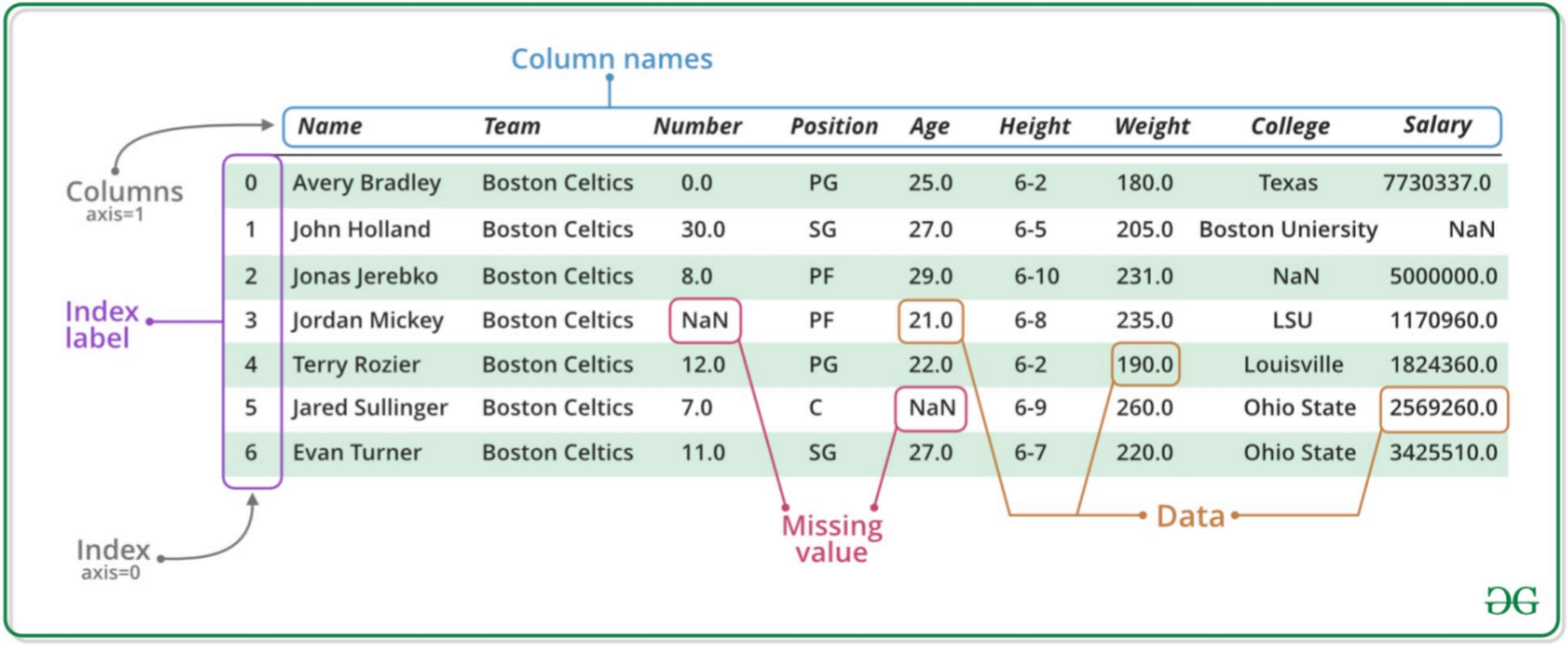
Pandas DataFrames

What is Pandas DataFrames?

- DataFrames are the **workhorse** of Pandas.
- We can think of a DataFrame as a **bunch of Series** objects put together to **share the same index**.
- A DataFrame is a **two-dimensional** data structure where data is aligned **in rows and columns**.
- Three principal components; the **data, rows, and columns** form the Pandas DataFrame.

Pandas DataFrames

What is Pandas DataFrames?



	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

Pandas DataFrames

Creating Pandas DataFrames

```
pandas.DataFrame(data=None, index=None, columns=None,
                 dtype=None, copy=None)
```

- “data” parameter can be;
  - NumPy Array
  - List
  - Dictionary
  - Scalar value

Pandas DataFrames

Creating Pandas DataFrames

Row Oriented

Dictionary

```
sales = [{'account': 'Jones LLC', 'Jan': 150, 'Feb': 200, 'Mar': 140},
         {'account': 'Alpha Co', 'Jan': 200, 'Feb': 210, 'Mar': 215},
         {'account': 'Blue Inc', 'Jan': 50, 'Feb': 90, 'Mar': 95 }]
df = pd.DataFrame(sales)
```

default

List

```
sales = [('Jones LLC', 150, 200, 50),
         ('Alpha Co', 200, 210, 90),
         ('Blue Inc', 140, 215, 95)]
labels = ['account', 'Jan', 'Feb', 'Mar']
df = pd.DataFrame.from_records(sales, columns=labels)
```

from_records

	account	Jan	Feb	Mar
0	Jones LLC	150	200	140
1	Alpha Co	200	210	215
2	Blue Inc	50	90	95

Column Oriented

```
sales = {'account': ['Jones LLC', 'Alpha Co', 'Blue Inc'],
         'Jan': [150, 200, 50],
         'Feb': [200, 210, 90],
         'Mar': [140, 215, 95]}
df = pd.DataFrame.from_dict(sales)
```

from_dict

```
sales = [('account', ['Jones LLC', 'Alpha Co', 'Blue Inc']),
         ('Jan', [150, 200, 50]),
         ('Feb', [200, 210, 90]),
         ('Mar', [140, 215, 95]) ]
df = pd.DataFrame.from_items(sales)
```

from_items

Pandas DataFrames

Basic Methods & Attributes

- .dtype
- .size
- .ndim
- .head
- .tail
- .shape
- .sample
- .sort_index()
- .sort_values()
- .isin
- .index
- .keys()
- .values
- .items()

Pandas DataFrames

Basic Methods & Attributes

- `.columns`
- `.reset_index`
- `.set_index()`
- `.iloc[]`
- `.loc[]`
- `.rename()`
- `.info()`
- `.describe()`
- `.value_counts()`
- `.unique()`
- `.nunique()`
- `.drop()`

Draw lines to match the attributes/methods to their definitions:

`df.values`

`df.shape`

`df.drop`

Return a Numpy representation of the DataFrame.

Return an int representing the number of axes / array dimensions.

Return an int representing the number of elements in this object.

Return a tuple representing the dimensionality of the DataFrame.

Return the first n rows.

Drop specified labels from rows or columns.

`df.head`

`df.ndim`

`df.size`

Students, draw anywhere on this slide!