CBA

Congresso Brasileiro de Automática

João Pessoa - PB

Introduction to Data Structure in Python: a Pandas crash course

João Pessoa - PB
09 September 2018

60 ANOS UFRN

UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

IMD INSTITUTO METRÓPOLE DIGITAL

CIÊNCIAS & TECNOLOGIA UFRN

list of anything

numpy (ndarrays)

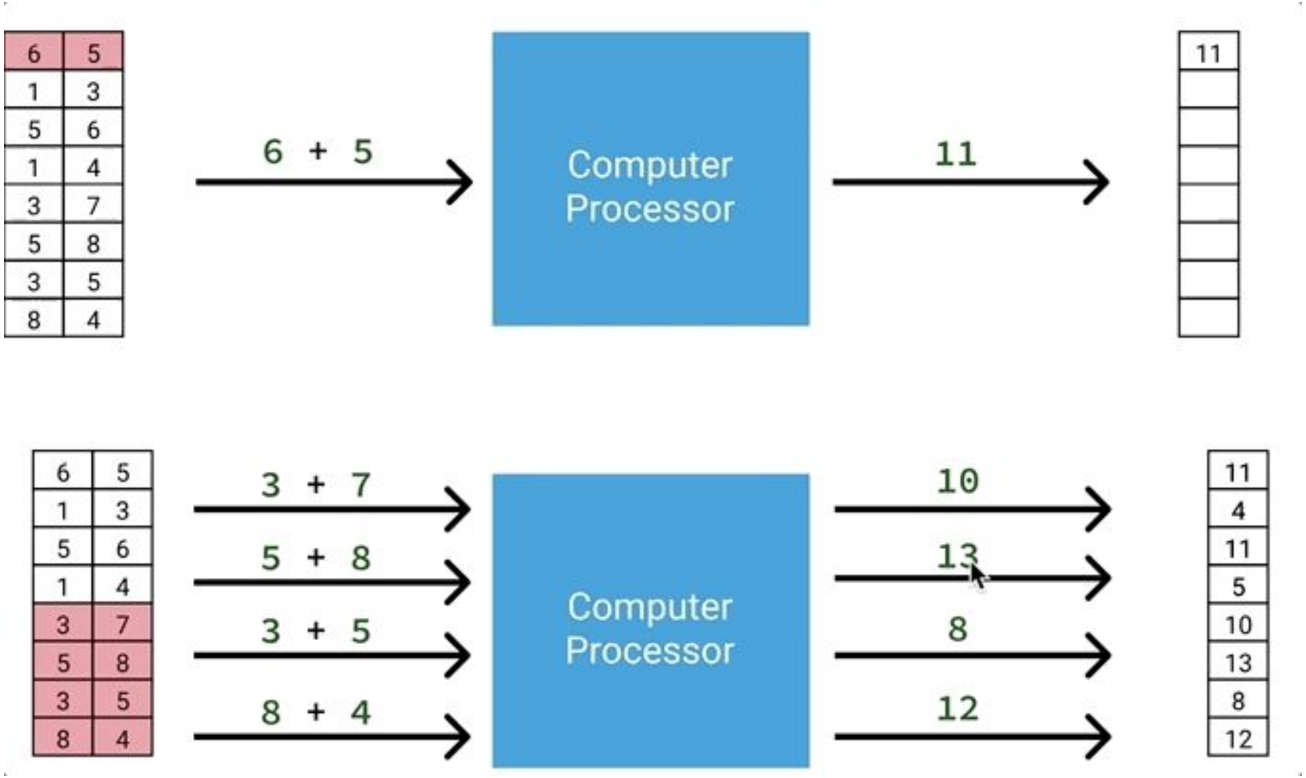
pandas (dataframe,series)

How Vectorization Makes Code Faster

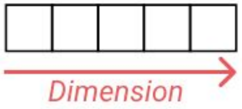
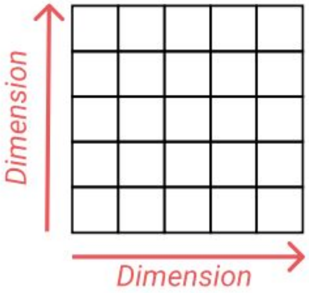
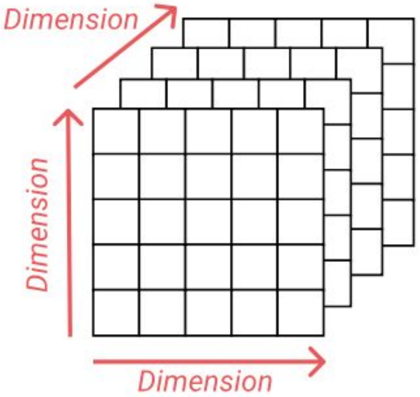
Single Instruction Multiple Data (SIMD)

List of Lists

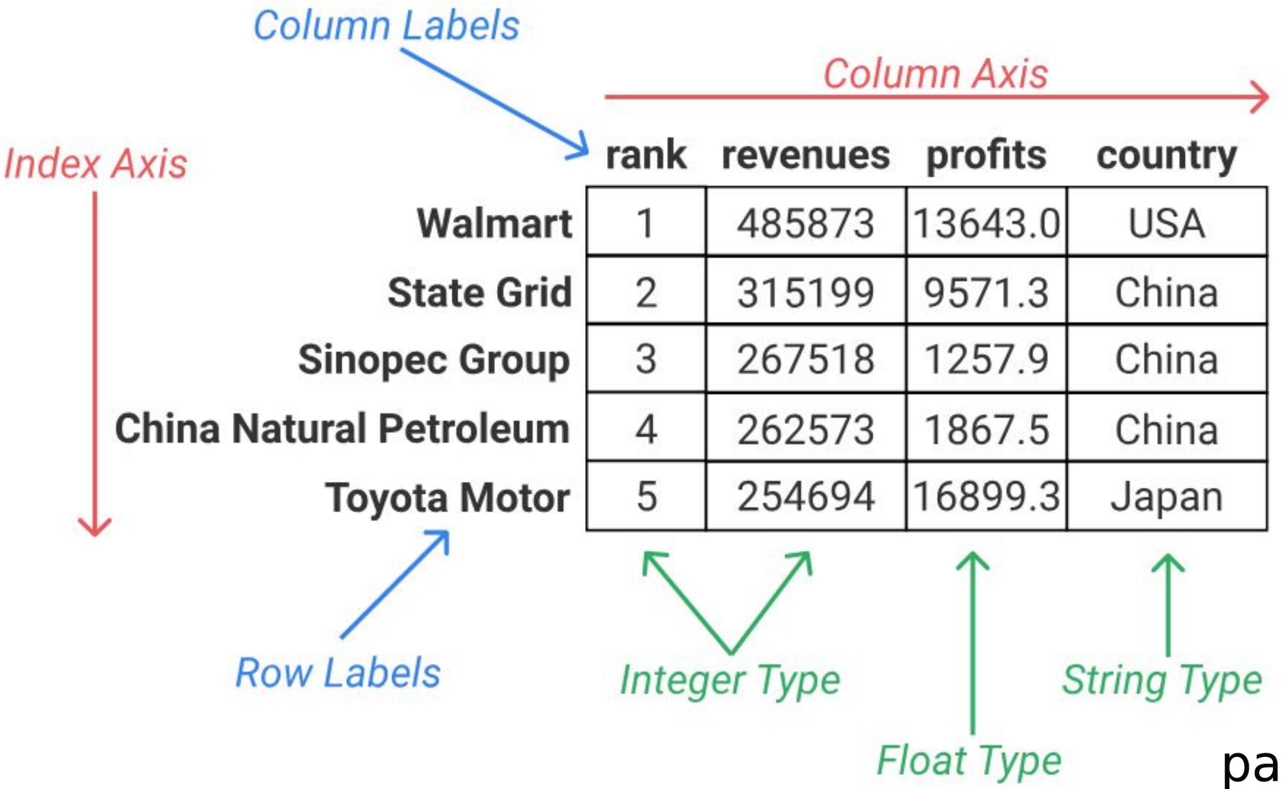
Numpy (ndarray)



Understanding Numpy ndarray

	Number of Dimensions	Known As
	One	One-dimensional array, array, list, vector, sequence
	Two	Two-dimensional array, matrix, table, list of lists, spreadsheet
	Three	Three-dimensional array, multi-dimensional array, panel

Dataframe vs Series

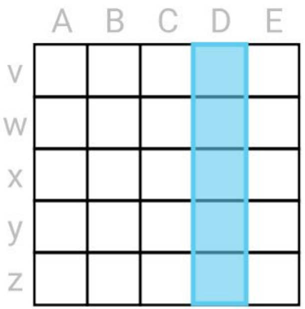


	rank	revenues	profits	country
Walmart	1	485873	13643.0	USA
State Grid	2	315199	9571.3	China
Sinopec Group	3	267518	1257.9	China
China Natural Petroleum	4	262573	1867.5	China
Toyota Motor	5	254694	16899.3	Japan

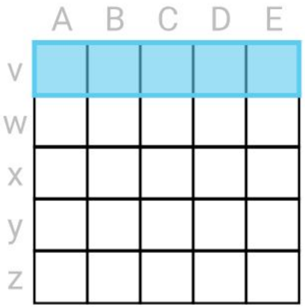
Dataframe vs Series

Original Dataframe

Code



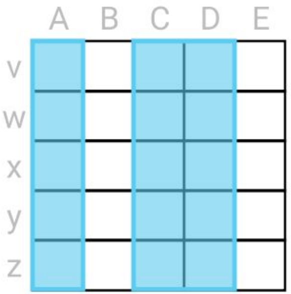
```
single_col = df["D"]
```



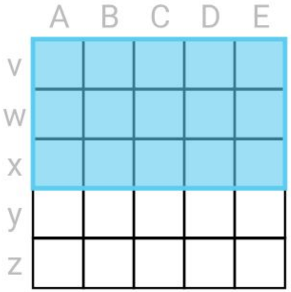
```
single_row = df.head(1)
```

Original Dataframe

Code

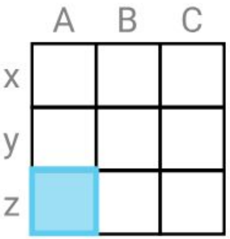


```
multi_cols = df[["A", "C", "D"]]
```



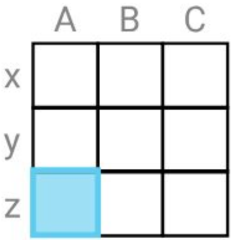
```
multi_rows = df.head(3)
```

Selecting elements

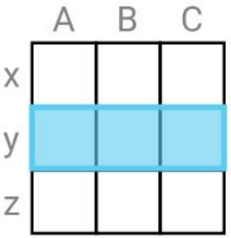


```
df.loc["z","A"]
```

located at row with label **z**, *column with label* **A**

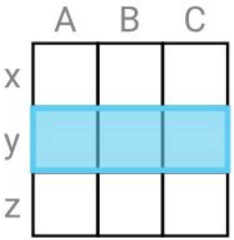


```
df.iloc[2,0]
```



```
df.loc["y"]
```

located at row with label **y**

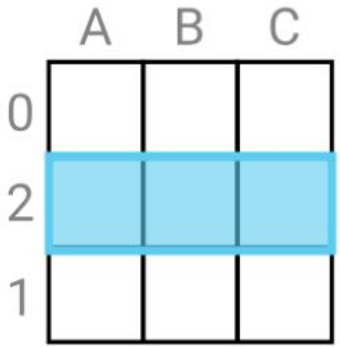


```
df.iloc[1]
```

Loc() vs iLoc()

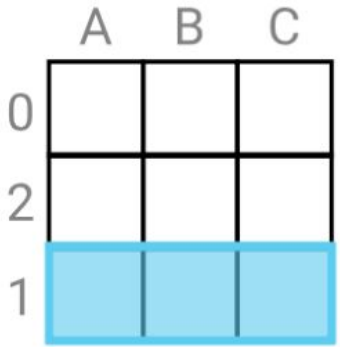
```
df.iloc[1]
```

`iloc[1]` *uses the integer position of the row to select the second row*



```
df.loc[1]
```

`loc[1]` *uses the label of the row to select the row with an axis label of* `1`*.*



Hands on

Global Power Plant Database

- **country** - 3 character country code.
- **country_long** - longer form of the country designation.
- **name** - name of the powerplant.
- **gppd_idnr** - 10 or 12 character identifier for the power plant.
- **capacity_mw** - electrical generating capacity in megawatts.
- **latitude** - geolocation of power plant.
- **longitude** - geolocation of power plant.
- **fuel1** - energy source used in electricity generation.
- **fuel2** - energy source used in electricity generation.
- **fuel3** - energy source used in electricity generation.
- **fuel4** - energy source used in electricity generation.
- **comissioning_year** - year of plant operation
- **owner** - majority shareholder of the power plant
- **source** - entity reporting the data; could be an organization, report, or document
- **ulr** - web document corresponding to the `source` field
- **geolocation_source** - attribution for geolocation information
- **year_of_capacity_data** - year the capacity information was reported
- **generation_gwh_2013** - electricity generation in gigawatt-hours reported for the year 2013.
- **generation_gwh_2014** - electricity generation in gigawatt-hours reported for the year 2014.
- **generation_gwh_2015** - electricity generation in gigawatt-hours reported for the year 2015.
- **generation_gwh_2016** - electricity generation in gigawatt-hours reported for the year 2016.
- **estimated_generation_gwh** - estimated annual electricity generation in gigawatt-hours for the year 2014