# Exploring Data with Pandas

Chapter 23

# What can Pandas do for us?

- Python Data Analysis Library


- Organizing data
- Calculate simple statistics
- Store data in formats that facilitate analysis
- And so much more

# Working with Data

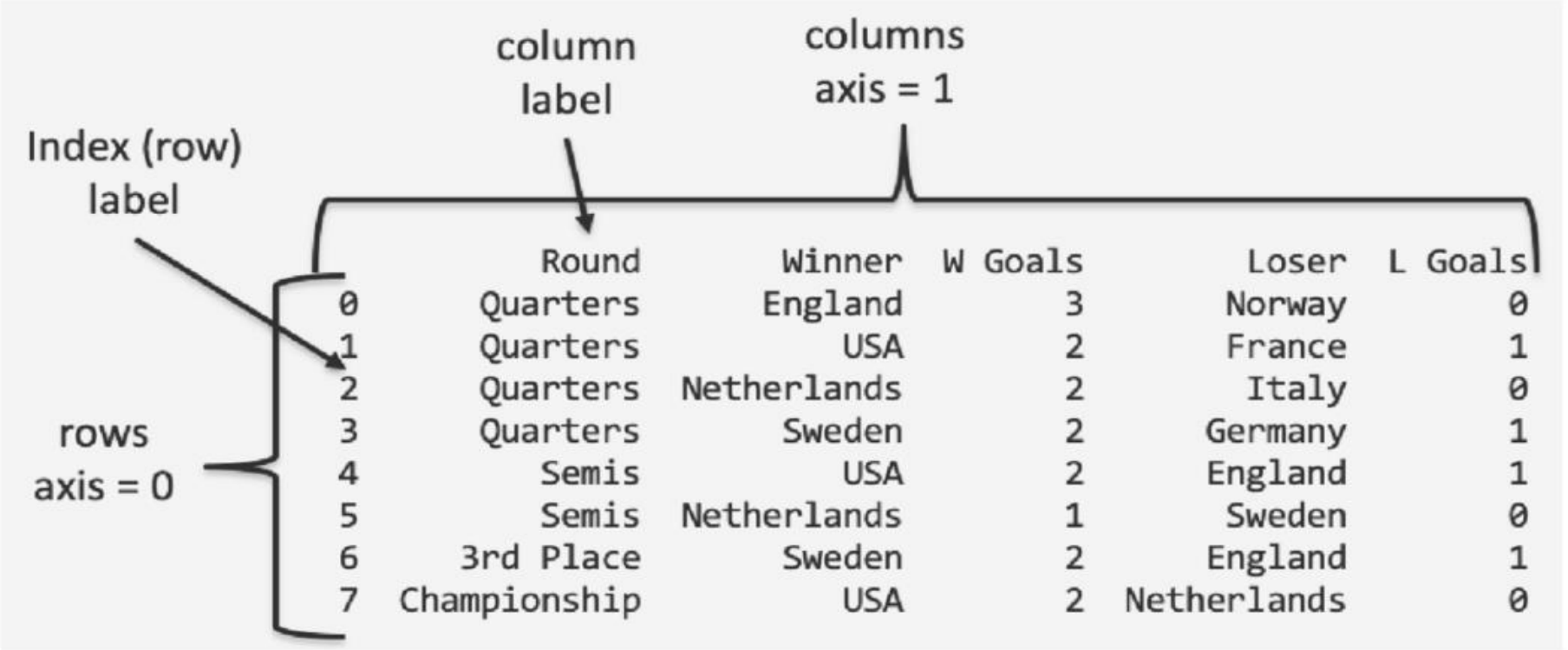
- Represents mixed-type tabular data in rows & columns

- Final Rounds of the 2019 Women's World Cup

```
Round,Winner,W Goals,Loser,L Goals
Quarters,England,3,Norway,0
Quarters,USA,2,France,1
Quarters,Netherlands,2,Italy,0
Quarters,Sweden,2,Germany,1
Semis,USA,2,England,1
Semis,Netherlands,1,Sweden,0
3rd Place,Sweden,2,England,1
Championship,USA,2,Netherlands,0
```

# Pandas DataFrame

```
wwc = pd.read_csv('wwc2019_q-f.csv')
print(wwc.to_string())
```




| | Round | Winner | W Goals | Loser | L Goals |
|---|---|---|---|---|---|
| 0 | Quarters | England | 3 | Norway | 0 |
| 1 | Quarters | USA | 2 | France | 1 |
| 2 | Quarters | Netherlands | 2 | Italy | 0 |
| 3 | Quarters | Sweden | 2 | Germany | 1 |
| 4 | Semis | USA | 2 | England | 1 |
| 5 | Semis | Netherlands | 1 | Sweden | 0 |
| 6 | 3rd Place | Sweden | 2 | England | 1 |
| 7 | Championship | USA | 2 | Netherlands | 0 |

Figure 23-1 A sample Pandas DataFrame bound to the variable wwc

# Types and Shapes

```
print(type(wwc))
print(type(wwc.index))
print(type(wwc.columns))
print(type(wwc.values))
print(wwc.values.shape)
print(wwc.shape)
```

```
<class 'pandas.core.frame.DataFrame'>
<class 'pandas.core.indexes.range.RangeIndex'>
<class 'pandas.core.indexes.base.Index'>
<class 'numpy.ndarray'>
(8, 5)
(8, 5)
```

# Building a DataFrame

- “Usually” will load from a file or query result
  - CSV, Excel
  - Fixed width text
  - JSON, HTML, XML
  - SQL, Google Query
  - And more ..
- Why build “by hand”
  - Data from an API
  - Data from several sources

# Selecting Data

- Select a column (Series)
  - `wwc['Winner']`
- Select a cell
  - `wwc['Winner'][2]`
- Select a set of columns (DataFrame)
  - `wwc[['Winner','Loser']]`

# Locating data with loc and iloc

- Optimized for row selection
- Index is the primary parameter
  - `wwc.loc[0]` (Series)
  - `wwc.loc[0:3]` (DataFrame)
  - `wwc.loc[0:3:2]`
  - Notice range is inclusive
- Index is really a label not a numerical index
- Use `iloc` to use numerical indices

# Other selections

- Select by group
  - DataFrameGroupBy
  - Similar to SQL GROUP BY
- Select by context
  - SQL WHERE
  - DataFrame