Applied Data Analytics

Pandas basics

(Multi)Indices

Hans-Martin von Gaudecker and Aapo Stenhammar

Why talk about indices?

- Only way to label data in a Series, used it implicitly for DataFrames
- Many operations are aligned by index
- Using a meaningful index makes this safer
- Index should be unique and not contain floats

Example: Gapminder data

	country	continent	year	lifeExp
0	Cuba	Americas	2002	77.158
1	Cuba	Americas	2007	78.273
2	Spain	Europe	2002	79.78
3	Spain	Europe	2007	80.941

Setting and resetting the index

```
[1] df.index
... RangeIndex(start=0, stop=4, step=1)

[2] df_clean = df.set_index(["country", "year"])

[3] df_clean.index
... MultiIndex([( 'Cuba', 2002),
                ( 'Cuba', 2007),
                ('Spain', 2002),
                ('Spain', 2007)],
               names=['country', 'year'])

[4] df_round_tripped = df_clean.reset_index()
[5] df_round_tripped.index
... RangeIndex(start=0, stop=4, step=1)
```

- **`set_index`** and **`reset_index`** are inverse functions
- **`set_index`** can take any column or list of columns
- Optional argument **`drop=True`** or **`drop=False`** determines what happens with the old index in **`set_index`** / **`reset_index`**

MultiIndex

- When your data has multiple dimensions
  - Firm ID × date × stock price
  - Country × age group × variables
  - Individual ID × time × variables
- Obtain by passing a list with multiple columns to `set_index`

MultiIndex: Selecting data

```
[6] df_clean.loc["Cuba"]
...      continent  lifeExp
    year
    2002  Americas   77.158
    2007  Americas   78.273

[7] df_clean.loc["Cuba", "lifeExp"]
... year
    2002    77.158
    2007    78.273
    Name: lifeExp, dtype: float64

[7] df_clean.loc[("Cuba", 2002), "lifeExp"]
... 77.158
```

- `.loc[.]` with a single argument selects all rows with that index at the first level
- `.loc[., .]` selects all rows with that index at the first level and the column
- `.loc[(., .), .]` selects the value at the intersection of the two indices

MultiIndex: Set the order carefully

- Saw it is very easy to do things with the first level of the index
- Works for higher level(s), too — but more tricky