Applied Data Analytics

Pandas basics

Shifting and differencing rows

Hans-Martin von Gaudecker and Aapo Stenhammar

Example: Gapminder data

MultiIndex: `country`, `year`

country	year	continent	lifeExp
Cuba	1997	Americas	76.151
Cuba	2002	Americas	77.158
Cuba	2007	Americas	78.273
Spain	1997	Europe	78.77
Spain	2002	Europe	79.78
Spain	2007	Europe	80.941

Shifting rows

```
[1] df.shift(1)
```

- Index remains the same
- Data is shifted by the number of rows specified in the argument

country	year	continent	lifeExp
Cuba	1997	nan	nan
Cuba	2002	Americas	76.151
Cuba	2007	Americas	77.158
Spain	1997	Americas	78.273
Spain	2002	Europe	78.77
Spain	2007	Europe	79.78

Including a column with lags

```
[2] df["lag_lifeExp"] = (
        df.shift(1)["lifeExp"]
    )
    df[["lifeExp", "lag_lifeExp"]]
```

country	year	lifeExp	lag_lifeExp
Cuba	1997	76.151	nan
Cuba	2002	77.158	76.151
Cuba	2007	78.273	77.158
Spain	1997	78.77	78.273
Spain	2002	79.78	78.77
Spain	2007	80.941	79.78

Combining groupby and shift

```
[3] df.groupby("country").shift(1)
    df[["continent", "lifeExp"]]
```

- Essentially always necessary when your data has a MultiIndex
- Or should have one

country	year	continent	lifeExp
Cuba	1997	nan	nan
Cuba	2002	Americas	76.151
Cuba	2007	Americas	77.158
Spain	1997	nan	nan
Spain	2002	Europe	78.77
Spain	2007	Europe	79.78

Differencing rows

```
[4] df.groupby("country").diff(1)
    ---------------------------------------------------------------------------
    TypeError                                 Traceback (most recent call last)

    [clipped]

    TypeError: unsupported operand type(s) for -: 'str' and 'str'
```

- Only defined for numerical data

Differencing rows

```
[5]  df.groupby("country")[["lifeExp"]].diff(1)
```

country	year	lifeExp
Cuba	1997	nan
Cuba	2002	1.007
Cuba	2007	1.115
Spain	1997	nan
Spain	2002	1.01
Spain	2007	1.161