Applied Data Analytics

Pandas

Introduction to Series

Hans-Martin von Gaudecker and Aapo Stenhammar

What is a Series?

- A vector (one-dimensional array) of data with associated labels
- All elements have the same data type

Example

Ship	Capacity
Laura Mærsk	1,926
CMA CGM San Francisco	6,956

How to create pd.Series

```
[1] pd.Series([2_100, 6_956])
[1] 0    2100
    1    6956
    dtype: int64
```

```
[2] pd.Series(
        [2_100, 6_956],
        name="capacity"
    )
[2] 0    2100
    1    6956
    Name: capacity, dtype: int64
```

- Call **`pd.Series()`** with values in square brackets, separated by commas.

  Result: Unnamed Series with default index (consecutive integers starting at 0).

- Set the **`name`** parameter to communicate the contents of the Series.

How to create pd.Series (cont'd)

```
[3] pd.Series(
        [2_100, 6_956],
        index=["Laura Mærsk", "CMA CGM San Francisco"]
        name="capacity"
    )
[3] Laura Mærsk              2100
    CMA CGM San Francisco    6956
    Name: capacity, dtype: int64
```

```
[4] capacity = pd.Series(
        [2_100, 6_956],
        index=["Laura Mærsk", "CMA CGM San Francisco"]
    )
```

- Set a meaningful **`index`** with values in square brackets, separated by commas. 0 The number of elements must be the same as in the data.
- To continue working with the Series, assign it to a variable

Accessing elements

```
[5] capacity.loc["Laura Mærsk"]
[5] 2100

[6] capacity.iloc[1]
[6] 6956

[7] capacity.loc[1]
[7] [very long message clipped]

    KeyError: 1

[8] capacity.iloc[2]
[8] [very long message clipped]

    IndexError: single positional indexer
                is out-of-bounds
```

- Use loc + index value (label) in square brackets
- Use iloc + index position (integer) in square brackets

  Indexing starts at 0 in Python!!!

- Indices must exist, else you get errors

  Error messages are scary at first.
  Reading the last line is often enough.

Attributes and methods

```
[9] capacity.dtype
[9] dtype('int64')

[10] capacity.value_counts
[10] <bound method IndexOpsMixin.value_cou
     CMA CGM San Francisco    6956
     Name: capacity, dtype: int64>

[11] capacity.value_counts()
[11] 2100    1
     6956    1
     Name: capacity, dtype: int64
```

- `[pd.Series].[xyz]` lets you do lots of useful stuff
- Attributes are static characteristics of the Series
- Methods are functions that operate on the Series
- Need to call them with parentheses