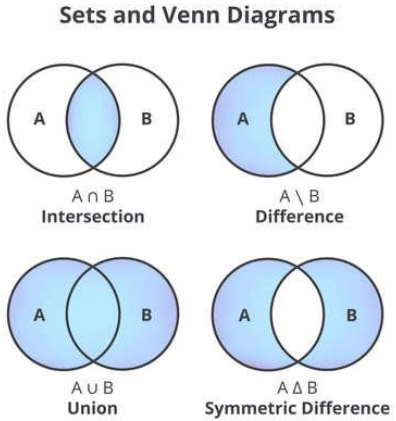
Pandas

Agenda

- Pandas
  - objects
  - data types
  - API

Pandas - objects

- Series
  - a one-dimensional array of indexed data.
  - has an explicitly defined index associated with the values - need not be an integer
  - `pd.Series(data, index=index)`
  - data - list, np.array, scalar (single value), dictionary
  - can add values just like in a dictionary
- Dataframe
  - two-dimensional - like a sequence of Series objects with the same (row) index.
  - flexible row indices and column names



- Index
  - Immutable array
  - Ordered set
    - union
    - intersection
    - differences
  - See ***loc*** and ***iloc***

Pandas - data types

Two types of data: continuous and discrete/categorical.
numpy and pandas allow missing values for most, but not all data types.

Allows null & NaN

- Int64 (pandas)
- ‘category’ (pandas)
- ‘boolean’ (pandas)
- datetime64 (numpy)
- float (numpy)

Can**no**t be null or NaN

- int (numpy)
- bool (numpy)
  - None = False
  - np.nan = True

The use of NaN allows us to perform arithmetic operations, but the result will be NaN.

```
pd.Series([1, np.nan, 2, None]) + 1
```

```
0    2.0
1    NaN
2    3.0
3    NaN
```

Pandas selections

Series - Like a dictionary: collection of keys and values.

`['key']`	
`.loc`	explicit
`.iloc`	implicit
`'key' in data`	
`.keys()`	
`['key1' : 'key5']`	slicing, explicit
`[0:4]`	slicing, implicit
`[list_of_keys]`	fancy indexing
`data > a_val`	masking (range)

Dataframe - multidimensional.

`["column_name"]`

`.column_name` attribute style, works only with strings that don't conflict with DataFrame method names.

`[:, 'column_name']`

`[:, 1] - first column`

Pandas - API

- lt, gt, eq
- div, mul, mod
- apply(func)
- count
- drop
- rename
- join
- merge
- transpose

- loc
- iloc
- matplotlib - plot
  - line (default)
  - area
  - bar
  - box
  - scatter

- to_
  - csv
  - sql
  - dict
  - json
  - html
  - markdown
  - pickle

https://pandas.pydata.org/docs/reference/frame.html

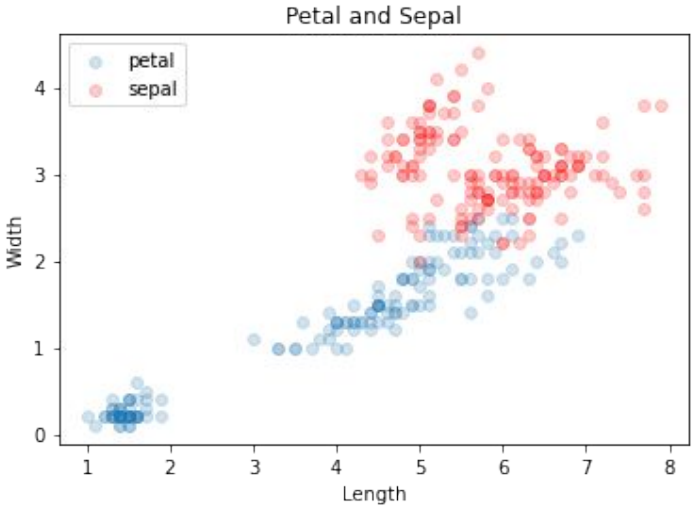
EDA and Pandas

Plots with pandas

RESTORY...

Exploratory Data Analysis - examples

If dataset is very large - take a subset (see splits) for faster manipulations.



Exercise

Iris dataset - load into dataframe

For all plots: Annotate with appropriate title and labels.

A) Scatterplot that shows the petal length and width and sepal length and width all in one plot. Let petals be blue dots and sepals be red.

B) Scatterplot of sepal length and width, color by target.

C) Two scatterplots, Petal and Sepal, colored by target.

Correlation and Covariance

Measure the **relationship** and **dependency** between two variables.

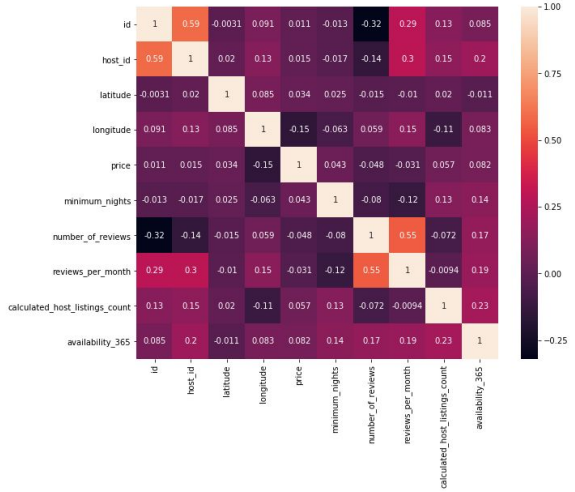
Covariance - direction, non-standardised

Correlation - direction and strength, standardised: values are between -1 (negative association) and +1 (positive association).

The closer to zero the weaker the correlation.

Negative association - one variable increases as the other decreases and vice versa

Positive association - move together, whether increase or decrease.



Correlation helps us investigate and establish relationships between variables.

```
from sklearn.preprocessing import
StandardScaler
iris_std = StandardScaler().fit_transform(iris)
```

RESTORY...

Länkar

- PythonDataScienceHandbook
- EDA basics
- EDA demographics ex
- Data visualisation how to
- Handling missing Data (handbook)
- RedEye - Sleep Cycle
- RedEye - Univrses
- pandas user guide
- plots examples with matplotlib & pyplot
- frekvens

Exercise

- Google dev