intro_python

September 15, 2021

1 Getting started with data analysis in Python

Bartosz Telenczuk, 2021 https://github.com/btel

Some of the examples were taken from "Plotting and Programming in Python" by The Carpentries, licensed under CC BY 4.0

1.1 Intro to Python for (data) scientists

1.1.1 JupyterLab

- starting
- creating jupyter notebook
- keyboard shorcuts: `Enter` (to enter edito mode), `Shit-Enter` (Run), `Esc` (enter command mode), `M` (markdown, in command mode), `X` (remove cell, in command mode)

1.1.2 Variables

- defining strings and integers
- variables stay defined even if you remove a cell
- indexing with integers and slices
- zero-based indexing!
- type

Exercise (types) Test the following operations in your notebook. Which output do the produce? What is the type?

```
first_name = 'Adam'
age = 100

variable_1 = 'hello' + first_name
variable_2 = age + 1
variable_3 = 5.1
variable_4 = first_name + 1
```

1.1.3 Built-in functions, methods and and help

- builtin functions

- positional arguments
- string methods
- official Python docs: https://docs.python.org/3/
- types have methods

Exercise (comparing strings) What will the following program show:

```
rich = "gold"
poor = "tin"
print(max(rich, poor))
```

1.2 Data analysis with pandas

1.2.1 Working with data

- openning files in jupyter lab
- importing extra function libraries (pandas)
- importing csv data with read_csv
- keyword arguments
- showing dataframe

Try `pd.read_<Tab>` to find other formats (or look them up in docs)

1.2.2 Plotting

- line and dot plots
- histograms
- scatter plots

Exercise (plotting styles) Plot the relation between age and BMI using different ploting styles (such as ‘o’, ‘:’, ‘-.’, ‘ro’, ‘bo’)

1.2.3 Indexing data frame

- extract column
- iloc vs loc
- dataframe index
- two-dimensional indexing
- using empty slice

Exercise (automatic alignment) Normalize all variables in the data frame (subtract mean and divide by standard deviation)

1.3 Linear regression with sklearn

- split data into train/test set
- plotting with matplotlib
- fitting scikit learn linear regression on train set
- predicting on test set

Question Why do we have 3 different coefficients?