

Statistics and Data Analysis - Lab-01

Barbara Żogała-Siudem

2026/2027

1 Software

Mostly: `python` with packages `numpy`, `pandas`, `scipy.stats`, `matplotlib`

1.1 Where can you learn python for data analysis?

There are many tutorials online. There is also a free book:

<https://datawranglingpy.gagolewski.com/datawranglingpy.pdf>

Check also file `python_numpy_pandas.py` for brief reminder.

1.2 Test basics of numpy and pandas skills

1.2.1 Excercise

Create a random 1-dimensional vector

```
import numpy as np
np.random.seed(13)
x = np.random.normal(0, 2, 20).round(2)
```

- (a) Write to console all numbers from vector `x` that consist in $[-2, -1] \cup [1, 2]$.
- (b) Write to console the number and fraction of all non-negative values.
- (c) Find the mean of all positive values.
- (d) Find the value closest to 0 (remember to keep the sign of this value).
- (e) Divide `x` into two vectors `x1` – consisting of all values greater than the mean value of `x` and `x2` – consisting of values smaller or equal to the mean.
- (f) Standardize vector `x`, i.e. transform its values so that the mean of a new vector is equal 0 and standard deviation is equal to 1.

1.2.2 Exercise

Create a random matrix `x` with n rows and k columns, for $n = 4$, $k = 5$.

- (a) Standardize the columns of `x`.
- (b) Standardize the rows of `x`.
- (c) Find the maximal value in each row of `x`.
- (d) Apply the softmax function to each row of an array, i.e.,

$$x_{i,j} \rightarrow \frac{\exp(x_{i,j})}{\sum_{l=1}^k \exp(x_{i,l})}.$$

- (e) Then one-hot decode the values in each row, i.e., find the column number with the value most close to 1. Return a vector of size n .

1.2.3 Exercise

Assume that an $n \times d$ matrix X represents n points in R^d . Write a function that determines the pairwise distances between all the points in X and a given $y \in R^d$. Return a vector $d \in R^n$ with $d_i = \|x_{i,\cdot} - y\|_2$

1.2.4 Exercise

Load `grades.csv` file into `pandas.DataFrame`.

- (a) Select only records for which a column `school` has value `SP1`.
- (b) Find the average grade for `math` and `eng` in the whole school.
- (c) Find the average grades for each `class` for each school subject.
- (d) Sort the frame obtained in (c) by the average `math` grade.
- (e) Add a new column to the initial data frame indicating if a student's math grade is above the average grade in school or not.