Exercise 1.1

One popular Python library (module) for data preprocessing is sklearn.impute (scikit-learn). Check out the classes SimpleImputer and KNNImputer it contains.

Make the following data (- use the value np.nan for the missing values):

A	B	C	D
1	2	6	4
4		6	4
7	10	9	5
10	11	12	
	14	15	5
16	17		2
18	12	12	5
20		23	2

Then preprocess it as follows:

1. For column B use the KNNImputer with 2 neighbors to replace the missing values.
2. For column C use the KNNImputer with 3 neighbors to replace the missing values.
3. For column A use the SimpleImputer to replace the missing values with the average value rounded to integer.
4. For column D use the SimpleImputer to replace the missing values with the most frequent value.

After handling missing values what is the average of all the values in data? Give the answer rounded to three decimals.

Exercise 1.2

Many machine learning algorithms use the concept of similarity. One way to measure similarity is to calculate the distance between vectors. Learn about the concept of Manhattan distance and Euclidean distance.

Make a Python function that returns the distance between two vectors as desired. The function is given the vectors and the distance calculation method ('euclidean' or 'manhattan') as parameters. The function returns the calculated distance as its value.

Make another Python function given a matrix of m vectors each containing n values ($m \times n$) and a distance calculation method ('euclidean' or 'manhattan'). The function returns the matrix ($m \times m$) including the distances calculated in pairs of rows.

> Note
>
> The calculated distance values for different vectors appears twice in the output matrix. For ith and jth vector in the input there is two cells in output matrix: cell [i, j] and [j, i].

Calculate Manhattan distances for vectors (rows) in matrix:

```
array([[ 1,  2,  3,  5,  3],
       [ 3,  1,  5,  7, -1],
       [ 2,  7,  1,  8, -1],
       [ 4,  6,  1, -2,  0],
       [ 3,  0, -1,  2,  2],
       [ 0,  0,  0,  0,  0]])
```

What is the longest distance between vectors? Give the answer rounded to two decimals.

Exercise 1.3

In supervised learning the data is typically split to two or three subsets: train and test data or train, validation and test data.

Make a splitting to three subsets for the data in the following source: https://raw.githubusercontent.com/haniemi/deeplearning/main/data/airbnb.csv. This file has column 'id' as the first column, and it can be used as an index column.

This data contains information about airbnb rentals in New York, USA (2019). This public dataset is part of Airbnb, and the original source can be found on here https://insideairbnb.com/get-the-data.

Let's assume that the column room_type is the target for classification.

Split the data in three subsets using the function `train_test_split` from module `sklearn.model_selection`. Make the splitting in the following way:

- test data has 75 % of the data, validation data 5 % and test data 20 %.
    - First split the data in train and temp subsets.
    - Then split the temp subset to validation and test subsets.
- all subsets should have the same distribution in the column room_type, i.e. each set contains approximately the same percentage of samples of each target (room_type) class as the complete set.
    - Read the documentation of the `train_test_function`: sklearn.model_selection.train_test_split — scikit-learn 1.4.2 documentation
- use the seed (random_state) value 125 (- so that everyone gets the same splitting).

How many samples in the test data have the room type 'Private room'?