Assignment 5: Attendance

General instructions:

- Use the file “assignment_5_template.py” to do the assignment. Change the name of the file to “assignment_5.py”.
- Write your code after the comment: “#Your code here:” (do not modify the code before this comment)
- Office hours information are available in the office hours channel on Teams.
- Authorized libraries (you cannot import other library than the authorized ones):
    - pandas
    - csv
    - datetime
    - numpy

Deadline:

- 03/11/2022 at 23:59:59
- Delayed submissions are not allowed

Coding instruction:

- Your objective is to clean and print the dataframe of attendance to the ICT class.

1st step:

- Find all missing values. Some are already recognized by pandas as NaN values. Two other names have been used for missing values, find them, and transform them into NaN values.

2nd step:

- 3 data are under the wrong format, find them and replace them with NaN values.

3rd step:

- There are some outliners (values too high or too low). Find them and replace them by NaN values.

4th step:

- Delete all lines with NaN values (27 lines)

5th step:

- Delete duplicated rows (17 lines)

6th step:

- Reset the index of the dataframe, sort it by:
    - First priority: Name of student
    - Then: date
    - Then: Begin hour
    - Then: Week

- Here you can see an example of the final required dataframe:

	KBTU_ID	NAME_STUDENT	TYPE	DATE	BEGIN_HOUR	WEEK	COUNT
0	21B030804	Abdigali Almagul	LECTURE	2022-09-14	11.0	2.0	2.0
1	21B030804	Abdigali Almagul	PRACTICE	2022-09-16	12.0	2.0	1.0
2	21B030804	Abdigali Almagul	LECTURE	2022-09-21	11.0	3.0	2.0
3	21B030804	Abdigali Almagul	PRACTICE	2022-09-23	12.0	3.0	1.0
4	21B030804	Abdigali Almagul	LECTURE	2022-09-28	11.0	4.0	2.0
...	...	...	...	...	...	...	...
3665	21B031270	Zurgambayev Bakhredin	PRACTICE	2022-09-23	10.0	3.0	1.0
3666	21B031270	Zurgambayev Bakhredin	LECTURE	2022-09-27	16.0	4.0	2.0
3667	21B031270	Zurgambayev Bakhredin	PRACTICE	2022-09-30	10.0	4.0	1.0
3668	21B031270	Zurgambayev Bakhredin	LECTURE	2022-10-04	16.0	5.0	2.0
3669	21B031270	Zurgambayev Bakhredin	LECTURE	2022-10-11	16.0	6.0	2.0

[3670 rows x 7 columns]

- Do not use any function to print all dataframe or just specific parts or line. Just create a dataframe and print it.
- If the displayed dataframe is not exactly the same, it can be due to the python interpreter (VSCode, Jupyter, …). Try to upload anyway to check.
- When you upload your code, the path of the csv file should be only “attendance_to_clean.csv” (it should not be something like “D:/User…../attendance_to_clean.csv”)
- Do the steps in the same order than indicated otherwise it may not work.
- The types of columns are:
    - String for ‘KBTU_ID’, ‘NAME_STUDENT’ and ‘TYPE’
    - Datetime object for ‘DATE’
    - Float for ‘BEGIN_HOUR’, ‘WEEK’ and ‘COUNT’
- Value 8 for the weeks data is not an outliner.

Tests Information:

- There will be no input, just print the expected dataframe.

Upload instructions:

- When your code is finish and work. Upload it to codepost.io (Tutorial here)
- You can reupload your code as many times as you want until the deadline
- The last upload is taken into account, so if you success all tests but reupload another file which is not working, you assignment will be failed.
- The expected name for your file is “assignment_5.py”. If the name is different, your code will fail.