3.0 - Pandas sources

Pandas is compatible with lots of different data sources, converting from a txt into a multitude of other sources (vice-verse) has never been easier. For instance, if one wishes to transform the speed_measurements file used for the numpy chapter then it is easily done so with the following code snippet.

On Canvas as *txt_load_sources.py*

```
import os
import pandas as pd

CURR_DIR_PATH = os.path.dirname(os.path.realpath(__file__))

speed_data_path = CURR_DIR_PATH + "/speed_measurements"

speed_data = pd.read_csv(
  speed_data_path + ".txt",
  sep=",",
  header=None,
  encoding="unicode_escape"
)

speed_data.columns = ["speed", "lisence_plate", "color", "time"]

speed_data.to_csv(speed_data_path + ".csv", index=False)
speed_data.to_json(speed_data_path + ".json")
speed_data.to_html(speed_data_path + ".html")
```

If you run the code, assuming it is in the same directory as the speed_measurements.txt file, you should observe 3 new files being created with the same data in different formats.

On Canvas you will find a second file called txt_load_sources_sql.py which also contains a sql loading process using postgres and sqlite with sqlalchemy. Consider it a bunos treat (not expected to fully understand it, yet) of which you can chew on after dinner, but do feel free to ask about it.

3.1 - Pandas ETL

Time to extract, transform and load.

The data that you will work with is compiled by 5 different teachers in 5 different subjects. Naturally they all use different formats for noting down student attendance, however, the headmaster wants a final report of all students' attendances for the last 30 days.

A. Download the file called *attendance_data.zip* and unpack it

B. Put all of the files into a folder called “data” and write your script in the folder above where the data exist

```
__init__.py ← write script here
data
  biology_00.txt
  biology_01.txt
  ….
  physics_030.csv
```

C. In the script file you can reference data by their dictionary name

Assuming you have a file called biology_06.csv in the data folder then you can access it with the code below.

```
import os
import pandas as pd

CURR_DIR_PATH = os.path.dirname(os.path.realpath(__file__))

biology_path = CURR_DIR_PATH = "/data/biology_06.csv"

biology_df = pd.read_csv(biology_path)
```

3.1.1 - Extract

D. Start with inspecting the various dataframes and reflect about their dissimilarities.

E. Write a script that merges all of the data from various days of the same subject, that is,
biology_00.txt through biology_30.txt becomes biology_data.csv
math_00.csv through math_30.csv becomes math_data.csv
…

   a. It is prefered if the name of the file is according to previous month, e.g. 06_biology or june_biology to describe data collected during June.

3.1.2 - Transform

At this point you should have 5 files, one for each subject.

F. Join the first name and surname columns in the dataframes where the name was split.

G. Convert all the dataframes with the column “attendance” into “late” (observe that the difference is how much someone attended to how many minutes they were late), you may assume that a full lecture is 60 minutes.

3.1.3 - Load

H. Compile **one** new csv with the names and late duration of all the students that have had any absence during the last 30 days from all subjects.

   The file's name should be absence_june.csv
   Also preferable if the table is sorted according to absence