**[Weather Data Analysis](https://ku.blackboard.com/webapps/assignment/uploadAssignment?content_id=_233785_1&course_id=_28169_1&group_id=&mode=view)**

This is a graded assignment. **Deadline  1 December 2019 Sunday 23:30**

You are expected to do 2 basic data analysis using a given data set starting from 1853, the dataset given in Exercises folder named “Oxford\_data\_weather.txt” alternatively the data is published on web its URL is given in “Oxford\_data URL” link in the same folder.

You are asked to analyze:

1- Minimum temperature during the total data period\* see below for explanations.

2-Total sunshine hours during the total data period\* see below for explanations.

The data consists of:

* Mean daily maximum temperature (tmax)
* Mean daily minimum temperature (tmin)
* Days of air frost (af)
* Total rainfall (rain)
* Total sunshine duration (sun)

The monthly mean temperature is calculated from the average of the mean daily maximum and mean daily minimum temperature i.e. (tmax+tmin)/2.

Missing data (more than two days missing in a month) have been indicated by "---".

Estimated data (where available) have been indicated by "\*" after the value.

Station data files are updated on a rolling monthly basis, around 10 days after the end of the month. Data are indicated as provisional until the full network quality control has been carried out. After this, data are final.

Source: https://www.metoffice.gov.uk/public/weather/climate-historic/#?tab=climateHistoric

 To do the analysis :

**ONCE** Using nested Lists

**SECOND TIME** Using numpy arrays or pandas data frames

For each part:

1. First open the file and have a look at it yourselves see the explanations in the heading (first 5 lines)
2. Then**read** the monthly time series data from file into memory using data structures of Part 1 and Part 2 . Be aware that the starting year of data for max temp. and sunshine hours differ, sunshine hours started to be recorded after 1929. Hence you will make use of different parts (columns) of the file and you will have different number of rows for each analysis requested.
3. Do some data cleansing by:

* You may discard data records which are not finalized, marked with a label  of  “Provisional”. Those consider them will get a bonus of 10 pts.
* removing “\*” after a value (for example see data records of November 1985, several months in 1996),
* discarding records which has no sunny hours data for sunny hours analysis.
* or any other you think is needed, pls. comment them clearly

When constructing lists the first index in the nested list corresponds to year and the second index corresponds to the month number. More precisely, the double index [i][j] corresponds to year 1929 + i and month 1 + j (January being month number 1).

The task is to define this nested list in a program and do the following data analysis.

For **analysis (1) following are requested**:

**a)** Compute the average min temperature for each month during the total data period (1853-2018 if  provisional labelled records discarded). Output shall be an average calculated for each month for the whole period in the following form:

Jan.  Feb.  March April   May  June   July  August  Sept.  Oct.  Nov.  December

-----   ------  --------  ------  ------  ------   ------  ---------  -------   -----  ------  --------------

avg1 avg2 avg3    avg4 avg5 avg6… …………………………………….avg12

**b)** Which month has the best weather according to the means found in (a) ?

**c)** compute the average of mean daily minimum temperature (value of tmin column) for each month in each decade. Output shall be in the following form:

       - 1929 : avg\_decade20

1930-1939 : avg\_decade30

1940-1949 : avg\_decade40

1950-1959 : avg\_decade50

1960-1969 : avg\_decade60

…

2010-2018 : avg\_decade10 (including 2019 data is optional/bonus)

For **analysis (2) following are requested:**

**a)** Compute the average number of sunshine hours for each month during the total data period (1929-2018 if  provisional labelled records discarded). Output shall be an average calculated for each month for the whole period in the following form:

Jan.  Feb.  March April   May  June   July  August  Sept.  Oct.  Nov.  December

-----   ------  --------  ------  ------  ------   ------  ---------  -------   -----  ------  --------------

avg1 avg2 avg3    avg4 avg5 avg6… avg12

**b)** Which month has the best weather according to the means found in (a) ?

**c)** compute the average of sunshine hours (value of sun hour column) for each month in each decade. Output shall be in the following form:

       - 1929 : avg\_decade20

1930-1939 : avg\_decade30

1940-1949 : avg\_decade40

1950-1959 : avg\_decade50

1960-1969 : avg\_decade60

…

2010-2018 : avg\_decade10 (including 2019 data is optional/bonus)

Please consider the coding  template provided in Exercises folder named "Standalone\_module\_template.py"and document appropriately

At the end please write a paragraph:

1. explaining why you choose numpy arrays or pandas dataframe
2. comparing and interpreting the differences and your comments on using nested lists **and u**sing numpy arrays or pandas data frames (whichever one you had choosen).