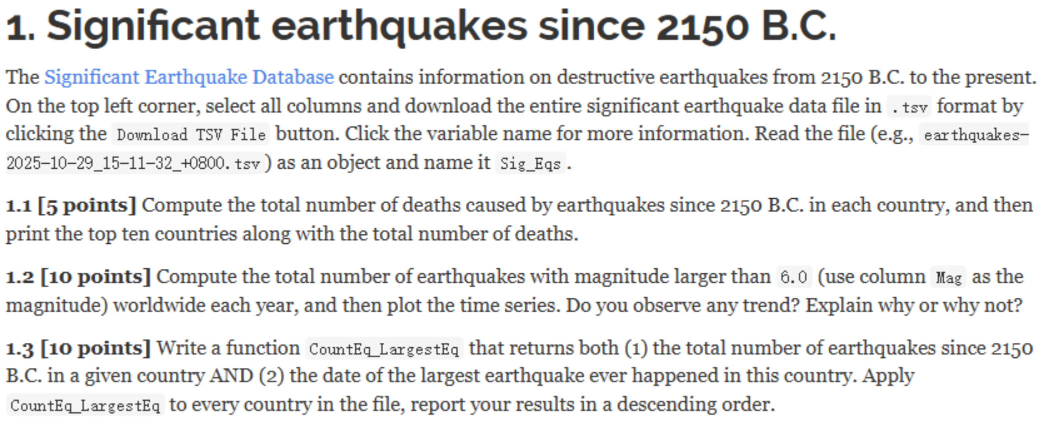
HOMEWORK#2



1. Significant earthquakes since 2150 B.C.

The Significant Earthquake Database contains information on destructive earthquakes from 2150 B.C. to the present. On the top left corner, select all columns and download the entire significant earthquake data file in `.tsv` format by clicking the `Download TSV File` button. Click the variable name for more information. Read the file (e.g., `earthquakes-2025-10-29_15-11-32_+0800.tsv`) as an object and name it `Sig_Eqs`.

1.1 [5 points] Compute the total number of deaths caused by earthquakes since 2150 B.C. in each country, and then print the top ten countries along with the total number of deaths.

1.2 [10 points] Compute the total number of earthquakes with magnitude larger than `6.0` (use column `Mag` as the magnitude) worldwide each year, and then plot the time series. Do you observe any trend? Explain why or why not?

1.3 [10 points] Write a function `CountEq_LargestEq` that returns both (1) the total number of earthquakes since 2150 B.C. in a given country AND (2) the date of the largest earthquake ever happened in this country. Apply `CountEq_LargestEq` to every country in the file, report your results in a descending order.

1.1:

思路：该题主要使用数据清洗+分组聚合的方式处理数据，然后根据统计结果过滤出想要的数据。首先使用 pandas 的 dropna 函数过滤删除无国家、无震级的无效记录，再将死亡人数缺失值填充为 0，确保数据有效性；然后按“国家”列分组，对“死亡人数”列求和得到各国总死亡人数；最后按死亡人数降序排序，筛选前 10 名国家。

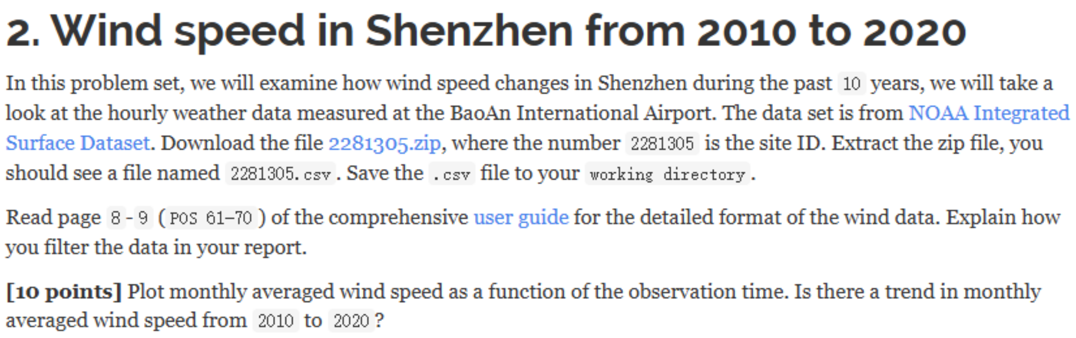
1.2:

思路：先筛选“震级>6.0”且“年份在合理范围”的记录，确保数据符合分析需求；再提取年份信息，生成完整年份序列，通过分组计数得到每年地震数量，补全无地震年份的计数，构建连续时间序列；接着计算 50 年移动平均线以突出长期趋势，用折线图绘制年度地震数与移动平均线。

1.3:

思路：首先创建`CountEq_LargestEq`函数，通过条件判断筛选指定国家的地震记录：若无记录则返回“无记录”，若有记录则统计地震总数，按震级降序找到最大地震，提取其日期与震级；再遍历数据中所有唯一国家，批量调用该函数，收集每个国家的“地震总数、最大地震震级、最大地震日期”；最后将结果整理为 DataFrame，按地震总数降序排序。

代码来源：学计算机的朋友的指导。



2. Wind speed in Shenzhen from 2010 to 2020

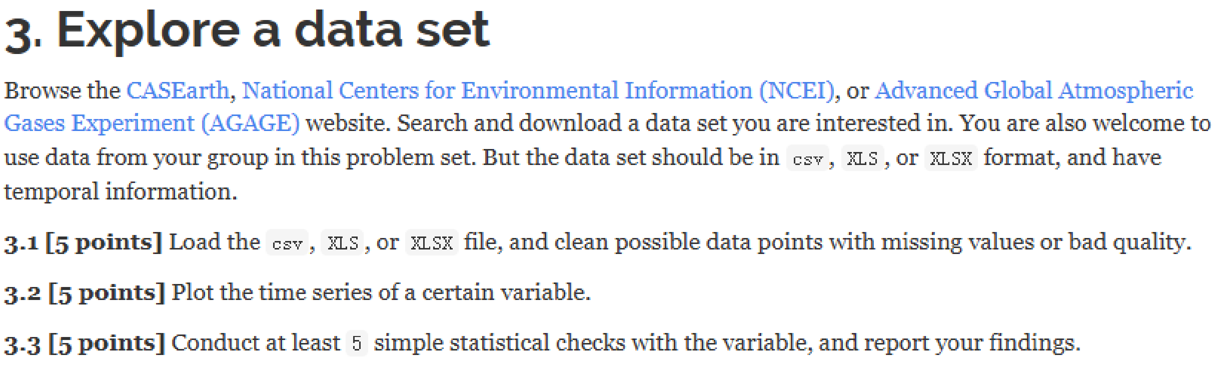
In this problem set, we will examine how wind speed changes in Shenzhen during the past `10` years, we will take a look at the hourly weather data measured at the BaoAn International Airport. The data set is from NOAA Integrated Surface Dataset. Download the file 2281305.zip, where the number `2281305` is the site ID. Extract the zip file, you should see a file named `2281305.csv`. Save the `.csv` file to your `working directory`.

Read page `8-9` (`POS 61-70`) of the comprehensive user guide for the detailed format of the wind data. Explain how you filter the data in your report.

[10 points] Plot monthly averaged wind speed as a function of the observation time. Is there a trend in monthly averaged wind speed from `2010` to `2020`?

思路：首先读取 CSV 文件，拆分出风向、风速等信息，再格式化数据单位，确保数据格式正确；再将“DATE”列转换为 datetime 类型，筛选 2010-2020 年的记录，过滤风速异常值；接着按“年月”分组，计算每月平均风速，构建月度平均风速数据集；最后用折线图绘制月度平均风速随时间的变化。

代码来源：学计算机的朋友的指导。



3. Explore a data set

Browse the CASEarth, National Centers for Environmental Information (NCEI), or Advanced Global Atmospheric Gases Experiment (AGAGE) website. Search and download a data set you are interested in. You are also welcome to use data from your group in this problem set. But the data set should be in `csv`, `XLS`, or `XLSX` format, and have temporal information.

3.1 [5 points] Load the `csv`, `XLS`, or `XLSX` file, and clean possible data points with missing values or bad quality.

3.2 [5 points] Plot the time series of a certain variable.

3.3 [5 points] Conduct at least `5` simple statistical checks with the variable, and report your findings.

3.1

思路：先从网址下载 CSV 文件，根据 CSV 文件来看要跳过前 40 行说明文字，定义核心列（年份、月份、CO_2浓度等）；再将数据中表示缺失值的“-9.99, -0.99”替换为 NaN，通过`dropna`删除 CO_2浓度缺失的行；最后合并“年份”与“月份”列，创建“date”列并设为索引，统一时间格式。

3.2

思路：先确认目标变量为“CO_2浓度”，提取“date”列（时间）与“co2”列（浓度）、“deseasonalized”列；再用折线图绘制两条曲线，分别代表月度 CO_2浓度原始值与去季节化趋势值；最后添加图表标题、坐标轴标签与图例，通过图表观察 CO_2浓度的长期变化与季节波动。

3.3

思路：该题针对 CO_2浓度变量开展 5 项统计：1.计算均值、中位数、标准差

等描述性统计量，最后发现基于810条CO_2浓度统计数据，CO_2浓度呈现春夏高、秋冬低的特征，与植物光合作用及人类活动的季节性变化相符。

代码来源：学计算机的朋友的指导。