PS2 Report

Name:左小幸　SID:12132243

PS2_1: Significant earthquakes since 2150 B.C.

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.

Code:

```
import pandas as pd

df = pd.read_csv("Sig_Eqs.tsv",sep='\t')
print(df[["Country","Deaths"]].groupby("Country").sum().sort_values(by="Deaths",ascending=False).head(10))
```

Result:

```
In [1]: runfile('C:/repo/ESE5023_Assignments_12132243/PS2/PS2_1.1.py', wdir='C:/repo/
ESE5023_Assignments_12132243/PS2')
               Deaths
Country
CHINA       2074900.0
TURKEY      1074769.0
IRAN        1011437.0
SYRIA        439224.0
ITALY        434863.0
HAITI        323472.0
AZERBAIJAN   317219.0
JAPAN        278138.0
ARMENIA      191890.0
PAKISTAN     148783.0
```

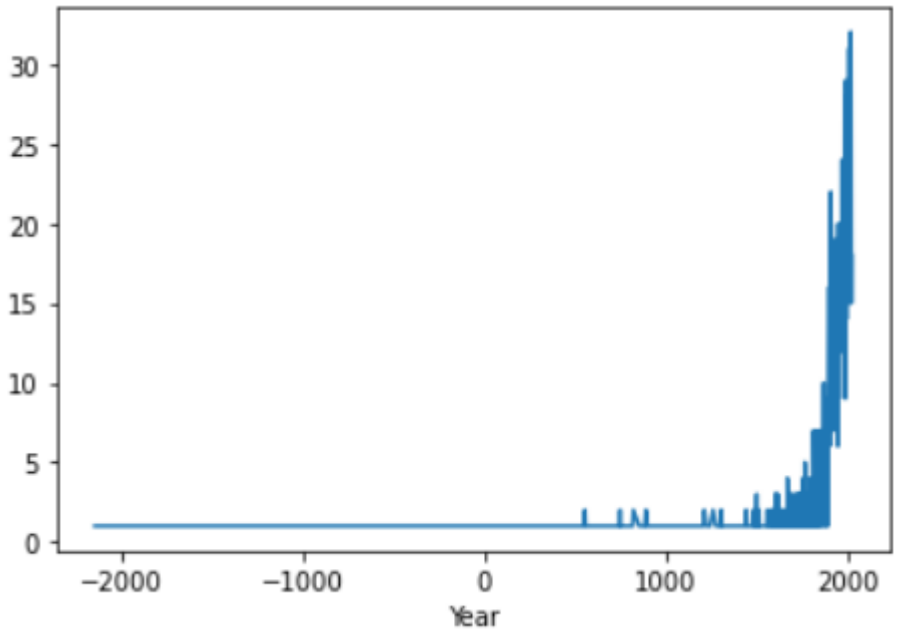
1.2 [10 points] Compute the total number of earthquakes with magnitude larger than 6.0 (use column Mag as the magnitude) worldwide each year, and then plot the time series. Do you observe any trend? Explain why or why not?

Code:

```
import pandas as pd

df = pd.read_csv("Sig_Eqs.tsv",sep='\t')
df[df["Mag"]>6][["Year","Mag"]].groupby("Year")["Year"].count().plot()
```

Result:



The trend I have observed is that the number of earthquakes greater than magnitude 6 globally has increased significantly over the last 500 years. I think the reason for this trend may not be that the number of high-intensity earthquakes has really increased, but that with the development of technology, people are observing and recording a lot of earthquakes that were not observed and recorded before.

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.

Code:

```
8   import pandas as pd
9
10  df = pd.read_csv("Sig_Eqs.tsv",sep='\t')
11
12  countrylist = df['Country'].unique()
13  result = df[['Country','Mag','Year','Mo','Dy']].head(0)
14
15  def CountEq_LargestEq(a):
16      df1 = df[df['Country'] == str(a)]
17      df2 = df1[df1['Mag']==df1['Mag'].max()][['Country','Mag','Year','Mo','Dy']]
18      df2['total_number'] = df[df['Country'] == str(a)]['Country'].count()
19      global result
20      result = result.append(df2)
21
22
23  for i in countrylist:
24      CountEq_LargestEq(i)
25
26  result = result.sort_values('total_number',ascending=False,ignore_index=True)
27  print(result)
```

Result:

```
In [3]: runfile('C:/repo/ESE5023_Assignments_12132243/PS2/PS2_1.3.py', wdir='C:/repo/
ESE5023_Assignments_12132243/PS2')
              Country  Mag    Year    Mo    Dy  total_number
0               CHINA  8.5  1668.0   7.0  25.0         610.0
1               JAPAN  9.1  2011.0   3.0  11.0         409.0
2           INDONESIA  9.1  2004.0  12.0  26.0         401.0
3                IRAN  7.9   856.0  12.0  22.0         380.0
4              TURKEY  7.8  1912.0   8.0   9.0         330.0
..                ...  ...     ...   ...   ...           ...
162             PALAU  7.6  1914.0  10.0  23.0           1.0
163            NORWAY  5.8  1819.0   8.0  31.0           1.0
164  FRENCH POLYNESIA  6.5  1848.0   7.0  12.0           1.0
165          KIRIBATI  7.6  1905.0   6.0  30.0           1.0
166           COMOROS  5.9  2018.0   5.0  15.0           1.0

[167 rows x 6 columns]
```

The way I think about this problem is. Store **(1)**(the total number**)** and **(2)**(the date of the maximum earthquake) in the same Data frame and print it.

PS2: Wind speed in Shenzhen during the past 10 years

[10 points] Plot monthly averaged wind speed as a function of the observation time. Is there a trend in monthly averaged wind speed within the past 10 years?

How to filter data:

WIND-OBSERVATION speed quality code
The code that denotes a quality status of a reported WIND-OBSERVATION speed rate.
DOM: A specific domain comprised of the characters in the ASCII character set.

- 0 = Passed gross limits check
- 1 = Passed all quality control checks
- 2 = Suspect
- 3 = Erroneous
- 4 = Passed gross limits check, data originate from an NCEI data source
- 5 = Passed all quality control checks, data originate from an NCEI data source
- 6 = Suspect, data originate from an NCEI data source
- 7 = Erroneous, data originate from an NCEI data source
- 9 = Passed gross limits check if element is present

I filtered the data according to the description of wind speed data quality on the user guide. I selected the data which **speed quality code** = 1 as valid data.

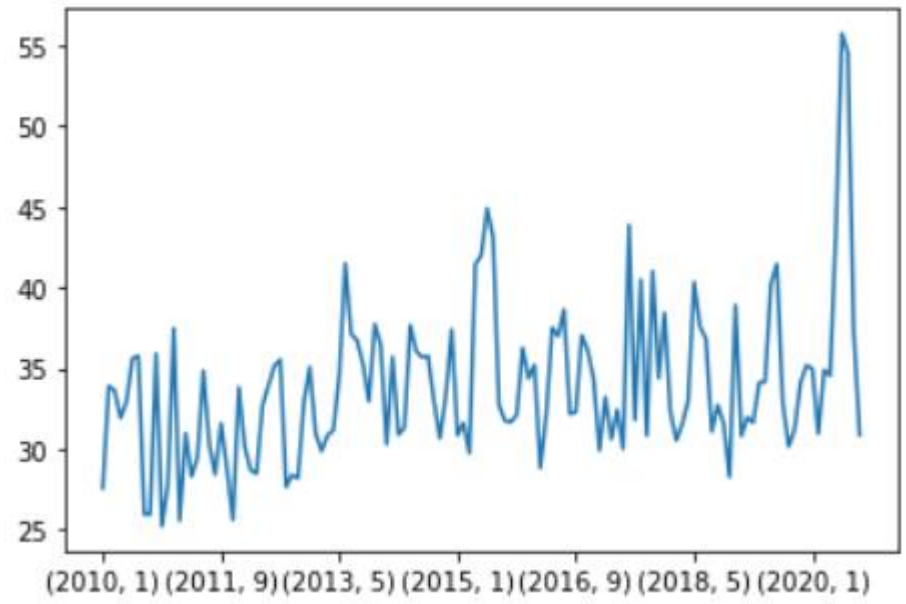
Code:

```
import pandas as pd

noaa = pd.read_csv('2281305.csv')
wind = noaa.loc[:,('DATE','WND')]
wind[['DA','DQC','TC','SR','SQC']] = wind['WND'].str.split(',',5,expand = True)

paqccwind = wind[wind["SQC"].astype('int') == 1]
paqccwind['DATE'] = pd.to_datetime(paqccwind['DATE'])
paqccwind['SR1'] = paqccwind['SR'].astype('int')
paqccwind.groupby([paqccwind['DATE'].dt.year,paqccwind['DATE'].dt.month])['SR1'].mean().plot()
```

Result:



I found that the monthly average wind speed of Bao An International Airport has been fluctuating for nearly ten years, and there seems to be no obvious trend. If one had to give a trend, it would be a slight increase in average wind speeds over the decade.

PS3: Explore a data set

Browse the CASEarth, NOAA Land-Based Datasets and Products, 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.

I chose a piece of data from my research, which contains information on more than 6,000 DAMS around the world. GRanD_dams_v1_3.xlsx

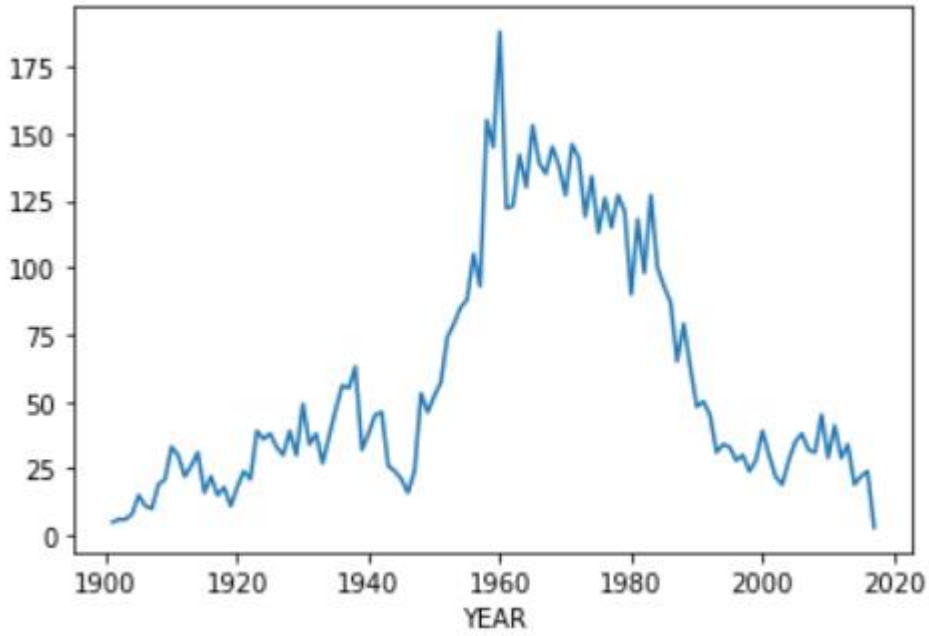
I wanted to count DAMS built since the 19th century, so I ran the following filter

```
#clean data
dam1 = dam[dam['YEAR']>1900]
```

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

I plotted the number of DAMS built each year

```
#the number of dams built since 1900
dam1.groupby(dam['YEAR'])['YEAR'].count().plot()
```



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

1. #Top 10 countries with the most DAMS

```
COUNTRY
United States    1920
China             921
Japan             546
India             332
South Africa      269
Spain             262
Canada            234
Brazil            203
Australia         190
Turkey            142
Name: COUNTRY, dtype: int64
```

2. #The name and country of the longest dam

	COUNTRY	DAM_NAME	DAM_LEN_M
2554	Senegal	Diama	80000

3. #The name and country of the deepest dam

	COUNTRY	DAM_NAME	DEPTH_M
1958	United States	Structure 336	1000.0
4689	Tajikistan	Rogun	1000.0

4. #The main use of these dams

```
MAIN_USE
Irrigation          1896
Hydroelectricity    1822
Water supply         892
Flood control        577
Recreation           294
Other                208
Navigation            56
Fisheries             14
Name: MAIN_USE, dtype: int64
```

5. #The highest dam

	COUNTRY	DAM_NAME	ELEV_MASL
7096	Peru	Sibinacocha	4870

Reference:

1. Getting started — pandas 1.3.4 documentation (pydata.org)helped me solve many problems related to PANDAS. **In in problem set 1 2 and 3.**

2. 时间序列与日期用法 | Pandas 中文 (pypandas.cn) Help me understand how to use the datetime method in Pandas, **in problem set2**.