PS2

第一题

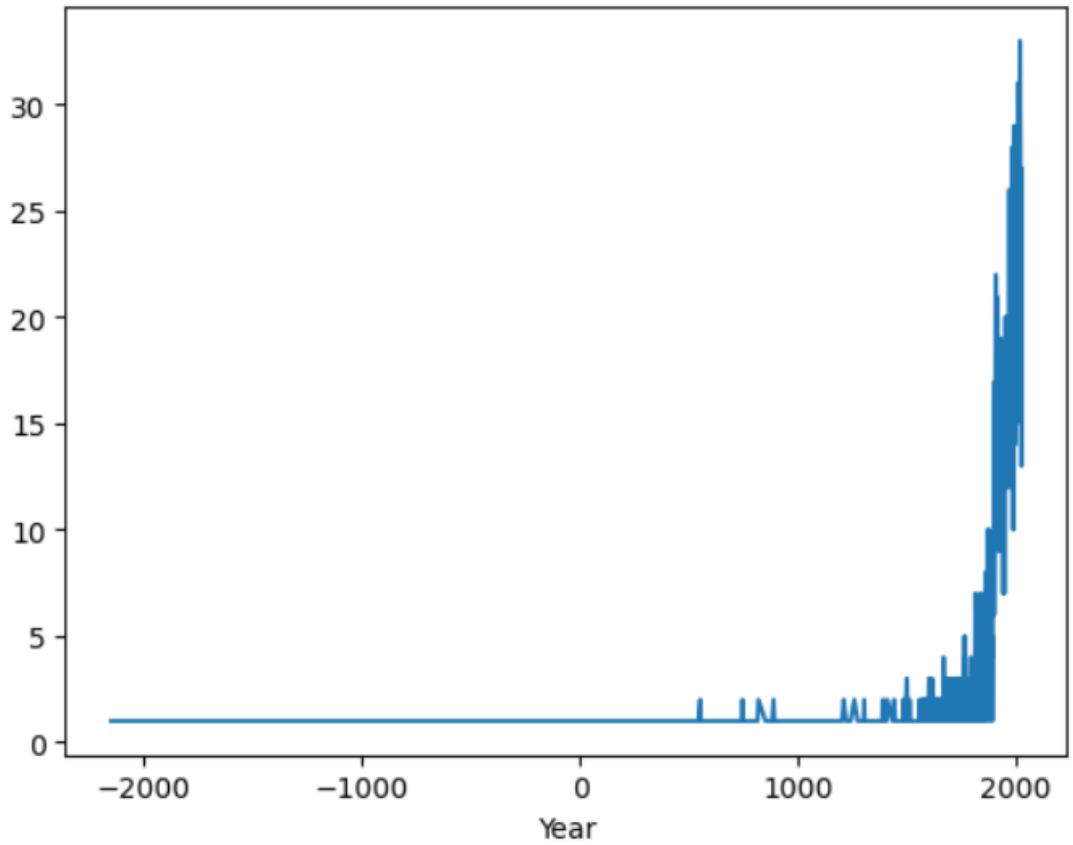
1.1

```
#1
#1.1
import pandas as pd
#读取数据
Sig_Eqs = pd.read_csv("earthquakes-2025-10-29_21-11-27_+0800.tsv",sep ='\t')
#查看前10行
Sig_Eqs.head(10)
#查看列名
Sig_Eqs.columns
#以国家进行分组，求和，再根据Total Deaths列降序排列，取前10行
top_10 = Sig_Eqs.groupby(['Country']).sum().sort_values('Total Deaths',ascending=False).head(10)
#列出前10的国家名和Total Deaths
top_10['Total Deaths']
```

```
Country
CHINA         2106524.0
TURKEY        1008863.0
IRAN           761654.0
SYRIA          487726.0
ITALY          423280.0
HAITI          323782.0
JAPAN          319443.0
AZERBAIJAN     319251.0
INDONESIA      282838.0
ARMENIA        189000.0
Name: Total Deaths, dtype: float64
```

1.2

```
#1.2
import pandas as pd
from matplotlib import pyplot as plt
#增加一列数值为1，便于求和计算地震总数
Sig_Eqs['Count'] = (1)
#筛选出大于六级的地震，按年求和,画出Count列数值变化
Sig_Eqs.loc[(Sig_Eqs['Mag'] > 6.0)].groupby(['Year']).sum()['Count'].plot()
plt.show()
```



观察到上升的趋势，原因：科技进步后，越来越多的地震被检测到。

1.3

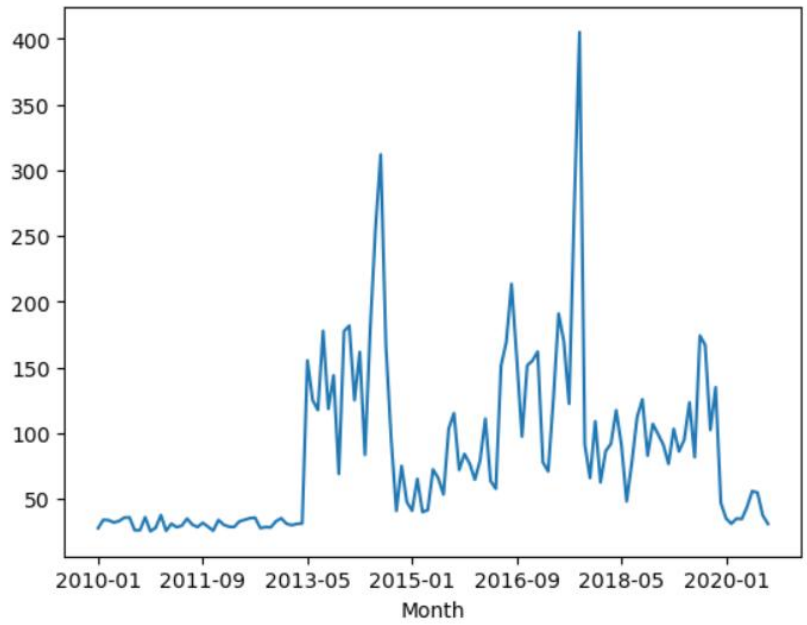
```
#1.3
def CountEq_LargestEq(x):
    #增加日期列
    Sig_Eqs['Date'] = Sig_Eqs['Year'].astype(str) + Sig_Eqs['Mo'].astype(str) + Sig_Eqs['Dy'].astype(str)
    #返回最大震级的索引
    max_mag_idx = Sig_Eqs.groupby(['Country'])['Mag'].idxmax()
    #输出地震次数和最大震级的日期
    return Sig_Eqs.groupby(['Country']).sum().loc[x]['Count'], Sig_Eqs.loc[max_mag_idx.loc[x]]['Date']
CountEq_LargestEq('CHINA')
```

```
(np.int64(623), '1668.07.025.0')
```

第二题

```
#2
#读取文件
speed = pd.read_csv("2281305.csv",sep =',')
#查看列名
speed.columns
#筛选出需要用到到的两列数据
data = speed[['DATE','WND']]
#需要进一步筛选，将WND列按逗号分割成多列，该函数使用参考deepseek
new_columns = data['WND'].str.split(',', expand=True)
#重命名新列
new_columns.columns = ['Direction_angle', 'Direction_quality', 'Type_code','Speed','Speed_quality']
#将新列合并回原表格
data = pd.concat([data, new_columns], axis=1)
#同样，对DATE列分隔，得到Month列
month = data['DATE'].str.rsplit('-', n=1, expand=True)
month.columns = ['Month', 'Time']
data = pd.concat([data, month], axis=1)
data
#进行筛选，根据 user guide，剔除风速缺失数据、错误数据
fin_data = data.loc[(data['Speed'] != 9999) & (data['Speed_quality'] != 3) & (data['Speed_quality'] != 7)]
fin_data

from matplotlib import pyplot as plt
#将Speed列转换成数值，否则画图时会报错，这一步来自deepseek的建议
fin_data['Speed'] = pd.to_numeric(fin_data['Speed'], errors='coerce')
#按Month分组，进行画图
fin_data[['Month','Speed']].groupby(['Month']).mean()['Speed'].plot()
plt.show()
```



2010 到 2020 年，月平均风速有上升趋势和周期循环规律

第三题

3.1

```
#3
#3.1
#安装xlrd库，否则文件读取不成功，这一步求助了deepseek
#pip install xlrd
#在网页下载中国大陆面2011-2020年碳排放量
data = pd.read_excel("carbon.xls")
#整体查看数据情况，没看到缺失值和类似999的数据
data
#检查是否有NaN/None的数据，来自deepseek
print(data.isnull().any().any())
#检查最大值，查看是否有999、9999
data.max()
#检查最小值，查看是否有-999、-9999
data.min()
#经过以上检查，发现没有缺失值等需要清洗的,但列名需要修改一下
data.columns= ['province', 'city', '2011','2012','2013', '2014','2015','2016', '2017','2018','2019', '2020']
data
```

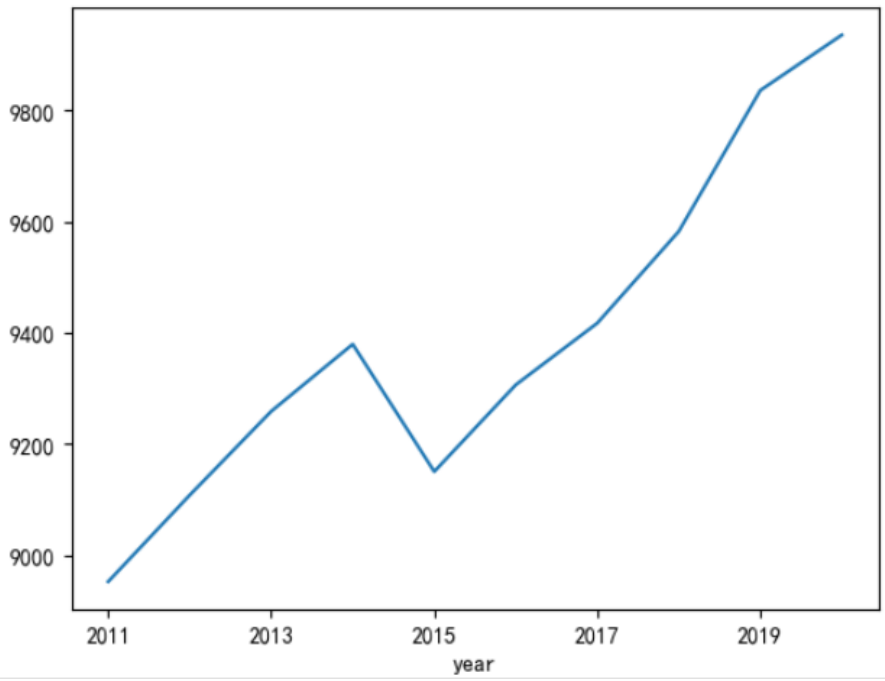
```
False
```

	province	city	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
0	四川省	阿坝藏族羌族自治州	6.682069	7.008357	7.019117	7.014884	6.395414	6.433615	6.511219	6.588822	6.666425	6.744028
1	新疆维吾尔自治区	阿克苏地区	29.252485	29.477133	31.567110	32.847546	26.105312	25.158390	44.393181	26.239975	57.610630	62.720737
2	新疆维吾尔自治区	阿拉尔市	1.638740	1.606597	1.872859	1.926499	1.790960	1.838021	2.187019	2.536018	2.885016	3.234015
3	内蒙古自治区	阿拉善盟	15.410883	15.917135	16.226202	16.186244	15.427383	15.457511	15.379101	15.300692	15.222282	15.143872
4	新疆维吾尔自治区	阿勒泰地区	8.849788	8.891016	8.818304	9.160530	10.836302	10.937434	9.036594	11.489783	14.813093	16.119106

3.2

```
#3.2
#数据的时间是宽格式，画图前先转换成长格式,melt函数用法求助deepseek
id_vars = ['province', 'city']  # 标识符列
value_vars = ['2011', '2012', '2013', '2014', '2015', '2016', '2017', '2018', '2019', '2020']
# 使用melt函数转换
data_long = data.melt(id_vars=id_vars,
                      value_vars=value_vars,
                      var_name='year',
                      value_name='carbon_emission')
data_long
#绘制全国碳排放总量随时间的变化
data_long.groupby(['year']).sum()['carbon_emission'].plot()
```

```
<Axes: xlabel='year'>
```



3.3

```
#3.3
#2011-2020年中全国最低的排放量
data_long.groupby(['year']).sum()['carbon_emission'].min()
#2011-2020年中全国最高的排放量
data_long.groupby(['year']).sum()['carbon_emission'].max()
#设置支持中文的字体，来自deepseek
from matplotlib import font_manager
plt.rcParams['font.sans-serif'] = ['SimHei', 'Microsoft YaHei', 'DejaVu Sans']
#2011-2020年每个省的排放总量
data_long.groupby(['province']).sum()['carbon_emission'].plot(
    kind='barh', figsize=(5, 7))
plt.show()
#省排放总量最高的
data_long.groupby(['province']).sum()['carbon_emission'].max()
#省排放总量最低的
data_long.groupby(['province']).sum()['carbon_emission'].min()
```

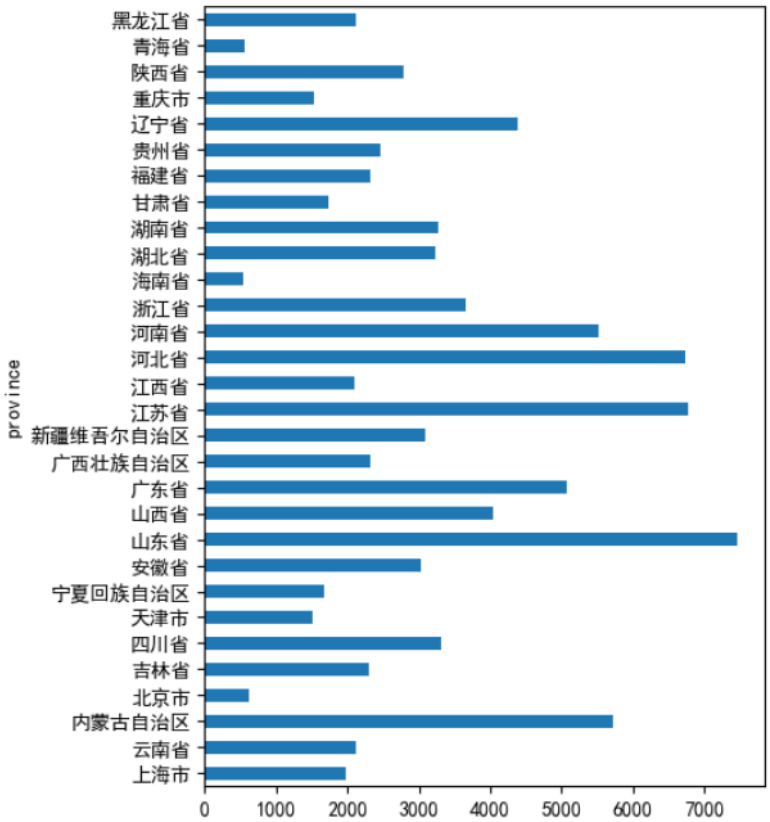
```
data_long.groupby(['year']).sum()['carbon_emission'].min()
```

```
8952.4913362073
```

```
data_long.groupby(['year']).sum()['carbon_emission'].max()
```

```
9935.3517857237
```

```
data_long.groupby(['province']).sum()['carbon_emission'].plot(
    kind='barh', figsize=(5, 7))
plt.show()
```



```
data_long.groupby(['province']).sum()['carbon_emission'].max()
```

```
6930.639837249901
```

```
data_long.groupby(['province']).sum()['carbon_emission'].min()
```

```
535.3457303922
```

在 2011-2020 年的全国碳排放总量中，2011 年碳排放量最低，是 8952 百万吨，2020 年碳排放量最高，9935 百万吨。在 2011-2020 年总排放量中，山东省贡献最多，6931 百万吨，海南省最少，535 百万吨。