

```

#1.1',
'# 读取TSV格式的地震数据文件',
'# 1.1 计算每个国家的总死亡人数，并输出死亡人数最多的前20个国家',
'# 忽略无效数据',
'# 提取location_name中："前的国家名',
'# 按国家分组，计算每个国家的总死亡人数',
'# 按死亡人数降序排列，并选择前20个国家',
'# 输出前20个国家的结果',
'# 计算所有国家的总死亡人数',
'# 输出全球因地震导致的总死亡人数',

'# 1.2 计算全球每年震级大于3.0的地震总数，并绘制时间序列图',
'# 将 'Mag' 列转换为数字类型，并忽略无效数据',
'# 选出震级大于3.0的地震',
'# 按年份分组，并计算每年发生的地震次数',
'# 绘制时间序列图',

'#Q: Do you observe any trend? Explain why or why not?',
'#A: As time going on , the number of eqs with mag > 3.0 each year becomes more and more. In my
views, there are mainly caused by 3 reasons.',
'#1. Historical Record Limitations: Earthquake records from the ancient past, especially before
the Common Era, are limited and often incomplete.',
'#2. Advancements in Earthquake Monitoring Technology.',
'#3. Data Collection and Integration: Modern datasets tend to include more earthquake events,
especially smaller magnitude ones. ',

'# 1.3',
'# 定义函数 CountEq_LargestEq，计算地震总数',
'# 筛选出该国家从公元前2150年至今的数据，并排除震级为空的记录',
'# 计算地震总数',
'# 找到最大震级及其对应的日期和位置',
'# 获取最大震级的整行数据',
'# 如果没有符合条件的数据',
'# 返回地震总数、最大震级、最大震级地震的日期和位置',
'# 获取所有唯一国家',
'# 存储每个国家的结果',
'# 遍历每个国家并调用 CountEq_LargestEq 函数计算地震统计信息',
'# 转换结果为 DataFrame 并按最大震级降序排序',

'#2',
'# 读取 CSV 文件',
'# 拆分 TMP 列，将温度和质量控制代码分别存储在新列中',
'# 将温度值 (TMP_Value) 转换为数值，并处理缩放因子',
'# 将质量控制代码 (TMP_Quality) 转换为整数',
'# 过滤掉无效温度值（例如 NaN 值）和质量控制不合格的数据',
'# 将时间数据转换为日期时间格式',
'# 只保留有效日期的数据',
'# 提取年月并计算每个月的平均温度',
'# 绘制月平均气温随时间的变化图'

```

```
#3.1',
'# 加载数据并指定要解析的列',
'# 将 'WMO_WIND' 列转换为数值类型, 并处理不能转换的数据为 NaN'',
'# 过滤掉 'WMO_WIND' 列中包含 NaN 的行'',
'# 3.1 根据风速(WMO_WIND)按风暴(SID)分组, 并报告前10个风速最大的飓风的名称',
'# 按照风暴标识符 (SID) 分组',
'# 按照风速列(WMO_WIND)降序排序, 并选择风速最大的10个飓风',
'# 打印风速最大的10个飓风的名称和对应的风速',
'#3.2',
'# 加载数据并指定要解析的列',
'# 将 'WMO_WIND' 列转换为数值类型, 并处理不能转换的数据为 NaN'',
'# 过滤掉 'WMO_WIND' 列中包含 NaN 的行'',
'# 将 'NAME' 列中的 NaN 值替换为 'Unnamed' ',
'# 按照风速降序排序, 选择风速最大的20个飓风, 同时去掉重复风速数据',
'# 确保 'NAME' 列为字符串类型'',
'# 使用 SID 绘制柱状图',
'# 设置 x 轴的刻度标签为对应的 NAME',
'# 图表标题和标签',
'#3.3',
'# 加载数据并指定要解析的列',
'# 检查 'BASIN' 列是否有缺失值, 如果有则处理'',
'# 统计每个流域的数据点数量',
'# 绘制柱状图',
'#3.4',
'# 加载数据并指定要解析的列',
'# 处理 LAT 和 LON 中的缺失值',
'# 绘制六边形图, 显示纬度和经度的分布情况',
'#3.5',
'# 加载数据并指定要解析的列',
'# 处理 LAT, LON 和 NAME 中的缺失值',
'# 筛选出台风山竹 (MANGKHUT) 在 2018 年的数据',
'# 绘制台风山竹的路径图 (经纬度散点图)',
'#3.6',
'# 加载数据并指定要解析的列',
'# 筛选1970年之后的数据, 并且流域为 WP 或 EP',
'# 查看过滤后的数据框',
'#3.7',
'# 加载数据并指定要解析的列',
'# 筛选1970年之后的数据, 并且流域为 WP 或 EP',
'# 计算每一天的数据点数量',
'# 绘制时间序列图',
'#3.8',
'# 加载数据并指定要解析的列',
'# 筛选1970年之后的数据, 并且流域为 WP 或 EP',
'# 添加一年中的天数 (day of year)',
'# 按每年中的天数分组, 计算每一天的平均数据点数量 (气候学)',
'# 绘制气候学图表',
'#3.9',
'# 加载数据并指定要解析的列',
'# 筛选1970年之后的数据, 并且流域为 WP 或 EP',
'# 添加一年中的天数 (day of year)',
'# 计算每一天的气候学 (按 dayofyear)',
'# 计算每一天的实际数据点数量',
'# 映射每天的 dayofyear 到 climatology 中, 确保长度匹配',
'# 计算异常值: 每天的数据点数量减去气候学值',
'# 绘制异常值图表',
'#3.10',
'# 加载数据并指定要解析的列',
'# 筛选1970年之后的数据, 并且流域为 WP 或 EP',
'# 添加一年中的天数 (day of year)',
'# 计算每一天的气候学 (按 dayofyear)',
'# 计算每一天的实际数据点数量',
'# 映射每天的 dayofyear 到 climatology 中, 确保长度匹配',
'# 计算异常值: 每天的数据点数量减去气候学值',
'# 按年重新采样异常值',
'# 绘制年度异常值时间序列',
'# 打印出异常值最大的前几个年份',
```

```
'#4.1',
'# 加载CSV文件 # 使用正确的文件路径',
"# 将 'DATE' 列转换为日期格式，确保时间数据有效",
"# 删除在关键列 'DATE', 'TAVG', 'TMAX', 'TMIN' 中存在缺失值的数据行",
'# 显示清理后的数据概况',
'# 显示清理后的前几行数据',
'
```

```
#4.2',
'# 绘制平均气温（TAVG）的时间序列',
'# 显示图表',
```

```
'# 4.3 对 TAVG 变量进行至少5项简单统计检查，并报告结果',
'# 1. 计算 TAVG 的均值',
'# 2. 计算 TAVG 的中位数',
'# 3. 计算 TAVG 的标准差',
'# 4. 计算 TAVG 的最小值',
'# 5. 计算 TAVG 的最大值',
'# 打印统计结果']
```

1、均值 (Mean): 7.620333333333333

均值为7.62可能表示在观测期间，该大气参数（如温度或污染物浓度）的平均水平。这个值可以用于评估环境质量标准或与历史数据进行比较。

2、中位数 (Median): 7.58

中位数略低于均值，这可能表明数据集中有一些较高的值，这些值拉高了均值。在环境数据中，这可能意味着某些时段的污染物浓度或温度较高，但大多数时间的浓度或温度接近或低于中位数。

3、标准差 (Standard Deviation): 1.0976816349704557

相对较大的标准差表明大气参数的值在平均值周围有较大的波动。这可能反映了气象条件的变化、污染源的不稳定性或测量误差。

4、最小值 (Minimum): 5.3

最小值为5.3可能表示在观测期间记录到的最低环境参数值。这可能是由于气象条件的变化、污染源的减少或控制措施的效果。

5、最大值 (Maximum): 10.34

最大值为10.34可能表示在观测期间记录到的最高环境参数值。这可能是由于污染事件、气象条件的极端变化或特定污染源的高强度排放。

综合这些统计数据，我们可以得出以下结论：

环境波动性：较大的标准差表明大气参数有显著的日变化或季节变化，这可能与气象条件（如风速、温度、湿度）的变化有关。

污染控制效果：如果最小值和中位数相对较低，可能表明在某些时段内污染控制措施是有效的。

潜在的高风险时段：较高的最大值可能指示存在特定的高风险时段，需要进一步调查和监测以确定污染源并采取相应的控制措施。