Assignment 02

1. Significant earthquakes since 2150 B.C.

解题思路如下：

（1）按国家分组统计死亡人数并排序

按国家分组：使用“groupby”方法按“Country”列对数据进行分组。

计算每个国家的死亡人数总和：对每个分组的“Deaths”列求和，得到每个国家的死亡人数总和。

重置索引：使用“reset_index”方法将分组后的结果转换为一个新的DataFrame，方便后续操作。

排序并获取前 20 个国家：按照“Deaths”列的值降序排序，并使用“head(20)”方法获取死亡人数最多的前 20 个国家。

（2）绘制地震数量随年份变化的折线图

筛选地震数据：从原始数据中筛选出震级大于 3.0 的地震。

按年份分组统计地震数量：使用“groupby”方法按“Year”列对筛选后的数据进行分组，并使用“size”方法计算每个分组的大小，即每年地震的数量。

重置索引并命名新列：使用“reset_index”方法将分组后的结果转换为一个新的 DataFrame，并使用“name='Count'”为计数列命名。

（3）计算每个国家的地震数量和最大地震信息

定义函数：定义了一个名为“CountEq_LargestEq”的函数，该函数接收一个国家名称作为参数。

筛选特定国家的地震数据：使用条件筛选出特定国家的地震数据。

计算地震总数：使用“shape[0]”获取筛选后 DataFrame 的行数，即该国家的地震总数。

查找最大地震：如果该国家有地震数据，并且“MS”列中有有效值，使用“idxmax”方法找到震级最大的地震的索引，然后使用“loc”方法获取该地震的详细信息。

处理日期信息：确保“Mo”和“Dy”列的值为整数，并处理任何 NaN 值，将日期格式化为“YYYY-MM-DD”的形式。

返回结果：函数返回该国的地震总数、最大地震的日期和位置。

应用函数到所有国家：遍历数据集中所有唯一的国家名称，对每个国家应用“CountEq_LargestEq”函数，并将结果存储在“results”列表中。

创建结果 DataFrame：将“results”列表转换为一个新的 DataFrame，并指定列名。

排序结果：按照“Total Earthquakes”列的值降序排序结果。

2. Air temperature in Shenzhen during the past 25 years

解题思路如下：

提取温度和质量控制列：从原始数据中提取“DATE”（日期）、“TMP”（温度）和“QUALITY_CONTROL”（质量控制）三列，存储在“temperature_data”变量中。

处理缺失值：由于温度数据中的缺失值被标记为“+9999,9”，代码通过条件筛选去除这些值。

处理温度列：使用正则表达式“str.extract”提取“TMP”列中的数字部分，并将其转换为数值类型。“errors='coerce'”参数将无法转换的值设置为 NaN。

筛选 NaN 值：去除“TMP”列中的 NaN 值，确保后续计算的准确性。

温度值缩放：将温度值从原始单位转换为摄氏度，通过除以 10 实现。

日期转换：将“DATE”列转换为 datetime 格式，方便后续的时间序列分析。

设置索引：将“DATE”列设置为 DataFrame 的索引，这样可以更方便地进行时间序列分析。

按月计算平均温度：使用“resample('ME').mean()”方法按月（“ME”代表月末）计算平均温度。这里使用“ME”代替“M”是为了确保每个月的最后一天都被包括在内。

3. Global collection of hurricanes

解题思路如下：

（1）找出最强的 10 个飓风

转换“WMO_WIND”列为数值类型并去除缺失值：将“WMO_WIND”列转换为数值类型，无法转换的值将被设置为 NaN，并去除这些缺失值。

按飓风标识符（SID）和名称（NAME）分组，然后聚合得到最大风速和相关名称：使用“groupby”方法按“SID”和“NAME”列对数据进行分组，然

后计算每个飓风的最大风速。

按风速降序排序并获取前 10 个飓风：对结果按“WMO_WIND”列的值降序排序，并使用“head(10)”方法获取风速最大的前 10 个飓风。

（2）绘制风速最强的前 20 个飓风的条形图

获取风速最强的前 20 个飓风：使用“nlargest(20, 'WMO_WIND')”方法获取风速最强的前 20 个飓风。

（3）按流域统计数据点数量并绘制条形图

按流域统计数据点数量：使用“value_counts”方法统计每个流域的数据点数量。

（4）创建经纬度的六边形图

确保纬度和经度列是数值类型：将“LAT”和“LON”列转换为数值类型。

创建六边形图：使用“plt.hexbin”函数创建六边形图，显示数据点的分布密度。

（5）绘制 2018 年台风山竹的路径

筛选 2018 年的台风山竹数据：筛选出名称为“MANGKHUT”且季节为 2018 的数据。

去除缺失经纬度值的行：去除“LAT”或“LON”列中有缺失值的行。

（6）筛选 1970 年以后在 WP 或 EP 流域的记录

筛选记录：筛选出 1970 年以后在 WP 或 EP 流域的记录。

（7）绘制每天的数据点数量

转换“ISO_TIME”列为日期：去除时间部分，只保留日期。

按日期统计记录数量：使用“groupby”方法按日期分组，并计算每天的记录数量。

（8）绘制每年数据点数量的气候学

确保“ISO_TIME”列为日期时间格式：将“ISO_TIME”列转换为日期时间格式。

提取每年中的第几天：使用“dt.dayofyear”提取每年中的第几天。

按每年中的第几天分组并计算平均计数：使用“groupby”方法按每年中的第几天分组，并计算每天的平均数据点数量。

（9）绘制每日计数与气候学的差异

转换“DATE”列为日期时间格式：确保“DATE”列为日期时间格式。

提取每年中的第几天：使用“dt.dayofyear”提取每年中的第几天。

合并气候学数据：将每日计数与气候学数据合并。

计算差异：计算每日计数与气候学的平均计数之间的差异。

（10）绘制年度差异

重采样年度差异时间序列：使用“resample”方法按年度重采样差异时间序列，并计算每年的平均差异。

识别异常年份：定义一个阈值（例如，绝对值大于 5）来识别异常年份。通过“annual_anomalies[abs(annual_anomalies) > threshold]”筛选出超过阈值的年份。

显示异常年份及其对应的异常值：使用“print”函数输出异常年份和它们对应的异常值。

通过图表识别异常年份：通过观察图表，识别出高于零线的峰值表示高于平均飓风活动的年份，低于零线的显著下降表示低于平均活动的年份。从图表中，可以识别出早期 1980 年代、晚期 1990 年代、早期 2010 年代的高异常年份，以及早期 2000 年代、大约 2020 年的低异常年份。

4. Explore a data set

解题思路如下：

（1）数据清除

删除缺失值：使用“dropna”方法删除 DataFrame 中任何包含缺失值的行，并将清除后的数据存储在“df_cleaned”中。

保存清除后的数据集：使用“to_csv”方法将清除后的数据保存为“cleaned_data.csv”文件，“index=False”参数表示不保存行索引。

（2）绘制温度时间序列图

转换“DATE”列为日期时间格式：使用“pd.to_datetime”函数将“DATE”列转换为日期时间格式，“errors='coerce'”参数将无法转换的值设置为 NaT。

将温度从华氏度转换为摄氏度：使用公式“(°F - 32) * 5/9”将“TEMP”列中的华氏温度转换为摄氏度，并存储在新的“Temperature (C)”列中。

（3）计算温度的基本统计数据

计算温度的描述性统计量：计算“Temperature (C)”列的平均值、中位数、标准差、最小值、最大值和四分位数。

报告发现：根据统计数据，对景德镇的温度数据进行分析和解释，包括平均温度、中位数、标准差、最低和最高温度以及四分位数，从而得出关于季节变化和温度分布的结论。

5. Acknowledgments

Ouyang wenjian explained to me what is asked in problem set 2,3,4.

I am very grateful to the student for his careful answer.