import pandas as pd

# 读取电站数据函数

def read\_station\_data(station\_id, generation\_file, environment\_file, weather\_file, capacity):

# 读取数据

generation\_data = pd.read\_excel(generation\_file)

environment\_data = pd.read\_excel(environment\_file)

weather\_data = pd.read\_excel(weather\_file)

# 统一时间格式，并按小时聚合

generation\_data['时间'] = pd.to\_datetime(generation\_data['时间']).dt.round('H')

environment\_data['时间'] = pd.to\_datetime(environment\_data['时间']).dt.round('H')

weather\_data['时间'] = pd.to\_datetime(weather\_data['时间']).dt.round('H')

# 处理负值，将其视为缺失值（NaN）

generation\_data['当日累计发电量kwh'] = generation\_data['当日累计发电量kwh'].apply(lambda x: x if x >= 0 else None)

environment\_data['辐照强度w/m2'] = environment\_data['辐照强度w/m2'].apply(lambda x: x if x >= 0 else None)

weather\_data[['当前温度', '最高温度', '最低温度', '风速', '湿度']] = weather\_data[['当前温度', '最高温度', '最低温度', '风速', '湿度']].applymap(lambda x: x if x >= 0 else None)

# 填充缺失值，使用线性插值法

generation\_data['当日累计发电量kwh'] = generation\_data['当日累计发电量kwh'].interpolate()

environment\_data['辐照强度w/m2'] = environment\_data['辐照强度w/m2'].interpolate()

weather\_data[['当前温度', '最高温度', '最低温度', '风速', '湿度']] = weather\_data[['当前温度', '最高温度', '最低温度', '风速', '湿度']].interpolate()

# 按小时聚合数据

generation\_data = generation\_data.groupby('时间').agg({'当日累计发电量kwh': 'sum'}).reset\_index()

environment\_data = environment\_data.groupby('时间').agg({'辐照强度w/m2': 'mean'}).reset\_index()

weather\_data = weather\_data.groupby('时间').agg({'当前温度': 'mean', '最高温度': 'mean', '最低温度': 'mean', '风速': 'mean', '湿度': 'mean'}).reset\_index()

# 合并数据

merged\_data = pd.merge(generation\_data, environment\_data, on='时间', how='inner')

merged\_data = pd.merge(merged\_data, weather\_data, on='时间', how='inner')

# 添加电站编号和装机容量

merged\_data['电站编号'] = station\_id

merged\_data['装机容量'] = capacity

return merged\_data

# 电站1的装机容量与数据文件路径

station\_1\_capacity = 4998.30

station\_1\_generation\_file = '/mnt/data/电站1发电数据.xlsx'

station\_1\_environment\_file = '/mnt/data/电站1环境检测仪数据.xlsx'

station\_1\_weather\_file = '/mnt/data/电站1天气数据.xlsx'

# 电站2的装机容量与数据文件路径

station\_2\_capacity = 5581.00

station\_2\_generation\_file = '/mnt/data/电站2发电数据.xlsx'

station\_2\_environment\_file = '/mnt/data/电站2环境检测仪数据.xlsx'

station\_2\_weather\_file = '/mnt/data/电站2天气数据.xlsx'

# 电站3的装机容量与数据文件路径

station\_3\_capacity = 4456.00

station\_3\_generation\_file = '/mnt/data/电站3发电数据.xlsx'

station\_3\_environment\_file = '/mnt/data/电站3环境检测仪数据.xlsx'

station\_3\_weather\_file = '/mnt/data/电站3天气数据.xlsx'

# 电站4的装机容量与数据文件路径

station\_4\_capacity = 1794.61

station\_4\_generation\_file = '/mnt/data/电站4发电数据.xlsx'

station\_4\_environment\_file = '/mnt/data/电站4环境监测仪数据.xlsx'

station\_4\_weather\_file = '/mnt/data/电站4天气数据.xlsx'

# 处理每个电站的数据

station\_1\_data = read\_station\_data(1, station\_1\_generation\_file, station\_1\_environment\_file, station\_1\_weather\_file, station\_1\_capacity)

station\_2\_data = read\_station\_data(2, station\_2\_generation\_file, station\_2\_environment\_file, station\_2\_weather\_file, station\_2\_capacity)

station\_3\_data = read\_station\_data(3, station\_3\_generation\_file, station\_3\_environment\_file, station\_3\_weather\_file, station\_3\_capacity)

station\_4\_data = read\_station\_data(4, station\_4\_generation\_file, station\_4\_environment\_file, station\_4\_weather\_file, station\_4\_capacity)

# 合并四个电站的数据

final\_merged\_data = pd.concat([station\_1\_data, station\_2\_data, station\_3\_data, station\_4\_data], ignore\_index=True)

# 查看处理后的数据

final\_merged\_data.head()

# 保存清洗后的数据为新的 Excel 文件

final\_merged\_data.to\_excel('/mnt/data/四电站\_清洗后数据.xlsx', index=False)