**在data/Exam目录下有数据文件bj\_aq\_17\_18.csv数据，其中包含以下字：id,station\_id,time,PM25\_Concentration,PM10\_Concentration,NO2\_Concentration,CO\_Concentration,O3\_Concentration,SO2\_Concentration，请按照要求完成以下数据处理工作：**

1. 读取数据文件，删除id字段，将其它字段的名称转化为：stationId,utc\_time,PM2.5,PM10,NO2,CO,O3,SO2 （10分）
2. 选择任意一个stationId,利用pandas-profiling查看数据报告，并对数据结果进行描述分析。（10分）
3. 将第1步处理后的数据转华为以时间为索引，stationId+原列名为列名的形式，比如：'gucheng\_aq\_PM2.5', 'gucheng\_aq\_PM10', 'gucheng\_aq\_NO2', 'gucheng\_aq\_CO', 'gucheng\_aq\_O3', 'gucheng\_aq\_SO2'等，共计210列。（20分）



1. 查看并输出最早日期和最晚日期，统计重复日期的行，并将重复日期的行删除，输出删除前和删除后的总记录数。（10分）
2. 对缺失值进行分析，包括以下内容：
   1. a) 在给定的时间段内，分别统计应该有多少个小时节点？，实际有多少个时间节点？，缺失的时间节点有多少？（10分）
   2. b) 查看哪些时间点上所有站点的所有特征都缺失了，要求输出缺失的时间点以及整小时缺失数量。（10分）
   3. c) 判断缺失时间点是否在长度等于5小时范围内，并输出缺失时间点在长度等于5小时范围内的时间点和大于5小时范围外的时间点。例如：假设2022年10月20日10点是缺失的，如果7、8、9、11、12、13点都是缺失的，则该时间点为超过5个小时，如果7、8、9三个时间段是缺失的，则在5小时范围内。（10分）
   4. d) 对于缺失时间点在5小时范围内的时间点，利用前后最近的未缺失时间点的数据平均值填充；对于缺失时间点大于5小时范围内的时间点，利用空值np.nan填充。（20分）

**要求：**

1. 每个同学独立完成，上交源代码+详细的报告word版，输出内容截图放在报告中，并进行说明。Word和源代码均以“学号\_姓名”格式命名。
2. 完成时间：2周内，11月7号24点之前，由学习委员收齐，统一发送到邮箱：[441005514@qq.com](mailto:441005514@qq.com)
3. 学习委员将自己班的word文档和源代码分别保存。