实战案例1-2：中国五大城市PM2.5数据分析 (2)

作者： Robin
日期： 2018/02
提问： 小象问答

1. 案例描述

随着PM2.5污染的严重性被越来越多地认识，PM2.5数据的质量也成为人们关心的话题。目前，公众判断所在城市PM2.5污染程度最常用的两大数据源，一是美国驻华大使馆（或领事馆）所发布的数据，二是中国环保部的实时播报。然而，中国环保部所发布的数据真实性却不时遭到质疑，例如《华尔街日报》就曾在2012年的一篇报道中称：北京官方的PM2.5数据与美国大使馆的数据不一致！也有不少学者著文，研究探讨中国空气污染数据的人为干扰。一些公众也持怀疑态度，认为环保部门“美化”数据的讨论不绝于耳。

数据可靠性是研究的基石，如果没有高质量的数据真实反映一个城市大气的污染程度，大气污染防治就无从谈起。该案例选取北京、上海、广州、成都、沈阳五个城市美国使/领馆及其邻近的环保部站点在2013-2015三年间的PM2.5数据，运用统计学方法验证美国使/领馆和邻近的环保部站点数据的可靠性。

2. 数据集描述

- Kaggle提供的数据集包括北京、上海、广州、成都和沈阳的2010-2015的空气质量数据。每个城市的数据文件为CSV文件：
  - BeijingPM20100101_20151231.csv
  - ShanghaiPM20100101_20151231.csv
  - GuangzhouPM20100101_20151231.csv
  - ChengduPM20100101_20151231.csv
  - ShenyangPM20100101_20151231.csv
- 数据字典
  - **No**: 记录编号，整型
  - **year**: 年份，整型
  - **month**: 月份，整型
  - **day**: 日期，整型
  - **hour**: 小时，整型
  - **season**: 季度，整型
  - **PM_?**: 中国环保部发布的?区的PM2.5指数 (ug/m^3)，浮点型
  - **PM_US Post**: 美国驻华大使馆发布的PM2.5指数 (ug/m^3)，浮点型
  - **DEWP**: 露点温度 (摄氏度)
  - **TEMP**: 温度 (摄氏度)
  - **HUMI**: 湿度 (%)
  - **PRES**: 气压 (hPa)
  - **cbwd**: 合成风向
  - **Iws**: 合成风速 (m/s)

- **precipitation**: 每小时降水量 (mm)
- **Iprec**: 累积降水量 (mm)

3. 任务描述

- 统计每个城市每天的平均PM2.5的数值
- 基于天数对比中国环保部和美国驻华大使馆统计的污染状态

4. 主要代码解释

- 代码结构

```
lect02_proj
├── data
    ├── *.csv   # 数据文件
├── output
    ├── *.csv   # 分析结果保存
├── config.y    # 配置文件
├── main.py     # 主程序
├── lect02_proj_readme.pdf  # 案例讲解文档
```

- **main.py**

轴方向的使用

```
def get_china_us_pm_df(data_df, suburb_cols):
    ...
    # axis=1，表示横向计算
    data_df['PM_China'] = data_df[pm_suburb_cols].mean(axis=1)
    ...

def compare_state_by_day(day_stats):
    ...
    # 横向组合DataFrame
    comparison_result = pd.concat(city_comparison_list, axis=1)
    ...
```

- **main.py**

层级分组操作

```
def main():
    ...
    # 先按照city，然后再按照date做分组
    # 注意city和date的顺序
    day_stats = all_data_df.groupby(['city', 'date'])[['PM_China', 'PM_US Post']].mean()
    ...
```

- **main.py**

copy()的使用，以下两个函数都是用到了copy()操作，因为DataFrame作为参数传到函数中，对DataFrame的操作有时会影响到外界的DataFrame，所以使用copy()避免了误操作的危险。

```
def add_date_col_to_df(data_df):
    ...
    proc_data_df = data_df.copy()
    ...

def add_polluted_state_col_to_df(day_stats):
    ...
    proc_day_stats = day_stats.copy()
    ...
```

- **main.py**

没有边界的分箱操作

```
def add_polluted_state_col_to_df(day_stats):
    ...
    # 如果需要在分箱操作指定<=或>操作时，及没有边界，可以用numpy中的inf(无穷)表示
    # -np.inf -> 负无穷；np.inf -> 正无穷
    bins = [-np.inf, 35, 75, 150, np.inf]
    ...
```

5. 案例总结

- 该项目通过分析中国五大城市PM2.5的数据实巩固了Pandas的常用数据处理及分析技巧:
    - 数据清洗
    - 向量化字符串操作
    - 分组与聚合操作
    - 层级索引的使用
    - 离散化和分箱操作

6. 课后练习

- 按小时统计每个城市的PM2.5指数，并添加相应的污染状态

参考资料

1. 最新出炉！北京、上海、广州、成都、沈阳五城市PM2.5污染状况对比及分析
2. 北大研究团队：供暖为北方的冬季增加多少PM2.5？（以京、沈为例）
3. 五大城市PM2.5官方数据可靠性深度验证
4. PM 2.5 data reliability, consistency, and air quality assessment in five Chinese cities