实战案例1-1：中国五大城市PM2.5数据分析 (1)

作者： Robin
日期： 2018/02
提问： 小象问答
数据集来源： kaggle
声明： 小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利

1. 案例描述

作为中国政治经济发展中非常重要的五个大型城市，北京、上海、广州、成都和沈阳分别位于京津冀、长三角、珠三角、川渝和辽宁经济区，这五大区域的经济总量占据全国总量的50%以上，能源消耗量巨大，也是大气污染和雾霾天气灾害较为严重的区域。雾霾的主要成分 PM2.5对人类健康有着极大的危害，也不同程度地影响着农业、生态、气候和居民的生活质量，因此研究中国这五个主要城市 PM2.5的污染状况及其影响因素，将为中国大气污染的预防和治理提供重要的实证依据。

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

- 五城市污染状态
- 五城市每个区空气质量的月度差异

4. 主要代码解释

- 代码结构

```
lect01_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
├── lect01_proj_readme.pdf  # 案例讲解文档
```

- **main.py**

字典的遍历

```
def main():
    ...
    # config.data_config_dict 为提前构造的字典数据
    # key为城市拼音
    # value为tuple，其中tuple中第一个元素为文件名，第二个元素为该城市对应的区名称列表
    # 如：北京
    # key为 'beijing'
    # value为 ('BeijingPM20100101_20151231.csv', ['Dongsi', 'Dongsihuan', 'Nongzhanguan'])

    # 遍历字典时，可以通过以下方法同时取出城市拼音、文件名及对应的区名称列表
    for city_name, (filename, cols) in config.data_config_dict.items():
        ...
    ...
```

- **main.py**

列表推导式的使用

```
def main():
    ...
    # ['PM_' + col for col in cols]是将字符串'PM_'和区的名称进行拼接，返回以'PM_'开头的字符串list
    # list相加，返回需要使用的列
    usecols = config.common_cols + ['PM_' + col for col in cols]
    ...
```

- **main.py**

条件表达式的使用

```
def load_data(data_file, usecols):
    ...
    for col in usecols:
        str_val = row[col]
        # 数据类型转换为float，如果是'NA'，则返回nan
        row_data.append(float(str_val) if str_val != 'NA' else np.nan)
    ...
```

- **main.py**

NumPy中条件索引

```
def get_avg_pm_per_month(data_arr):
    ...
    # 获取当前年份数据
    # data_arr[:, 0] == year 表示判断data_arr中的第0列中的数据是否等于year，如果是返回True，否则为False，即布尔值数组
    # 然后将上述结果作为mask作用于原始数组data_arr中，过滤出符合条件的数据
    year_data_arr = data_arr[data_arr[:, 0] == year]
    ...
    # 获取月份的所有数据
    month_data_arr = year_data_arr[year_data_arr[:, 1] == month]
    ...
```

- **main.py**

格式化字符串

```
def get_avg_pm_per_month(data_arr):
    ...
    # 格式化字符串
    # '{:.0f}-{:02.0f}'.format(year, month) 将年和月组合成字符串
    # 如 2013, 1 -> '2013-01'
    row_data = ['{:.0f}-{:02.0f}'.format(year, month)] + mean_vals
    ...
```

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Python的进阶技巧及科学计算库NumPy:
  - 字典的遍历
  - csv数据读写操作
  - NumPy的使用
  - 列表推导式的使用
  - 条件表达式的使用

6. 课后练习

- 五城市每个区空气质量的季度差异

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