Documentation for Air Quality Index Analysis

毛思文 吴逸涵 | 学号：2201211276 2201211243

1/2/2023

一、 项目简介

该项目目的为 Lecture5 和 Lecture6 中爬虫学习的深化：利用爬取的历史天气数据分析雾霾天气近五年的发展趋势及与风力、风向的关系。

为达该目的，本项目爬取了“https://www.aqistudy.cn/historydata”网站上北京市 2018.01.01 至 2022.12.31 的雾霾情况数据和天气数据。后续，对爬取得到的相关数据进行统计分析，发现了近五年来北京空气质量逐渐好转的趋势，同时分析得到空气质量与风力、风向等天气因素之间的关系，并通过可视化呈现分析结果。

二、 流程描述

4.1 获取 AQI 数据

主要函数功能如下：

函数名	功能
get_renewed_ctx()	Python 程序执行 JavaScript 的库，运行 JS 代码
get_historyapi()	获取北京 AQI 数据
save_text2df	将文本数据保存为 csv 格式，便于分析管理

后续添加 headers，设置 Users-Agent 等参数，scrapy 应对反爬机制。

```
if headers is None:
    headers = {
        'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8',
        'Accept-Encoding': 'gzip, deflate',
        'Accept-Language': 'zh-CN,zh;q=0.8,zh-TW;q=0.7,zh-HK;q=0.5,en-US;q=0.3,en;q=0.2',
        'Content-Type': 'application/x-www-form-urlencoded',
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/71.0.3578.80 Safari/537.
    }
```

Request.post()方法将 POST 请求发送到指定的 URL，得到 html_info

```
    html_info = requests.post(url,
                              data={ ajax_data_key: ctx.eval(js) },
                              headers=headers)
```

根据时间获取 AQI 数据，getParameter 为 js 文件中自定义函数

```
for date_time in date_times:
    js = f'getParameter("{city}", "{date_time}")'

    html_info = requests.post(url,
                              data={ ajax_data_key: ctx.eval(js) },
                              headers=headers)

    if (len(html_info.text) < 500):
        print('Error:', len(html_info.text), '\n', html_info.text)

        return temp_text

    else:
        kw = f'{city}_{date_time}'
        temp_text[kw] = html_info.text
        i += 1; progressbar(i, Ni, msg=kw)
return temp_text
```

最后使用 save_text2df 函数将文本数据保存为 csv 格式。

4.2 获取天气数据

运用类似的方法获取天气数据，并将数据同样保存为 csv 格式。

4.3 数据合并

将 AQI 数据与 weather 数据进行合并处理，最终结果如下图所展示：

date	aqi	aqi2345	pm2_5	pm10	so2	no2	co	o3	rank	quality	aqiInfo	aqiLevel	fengxiang	fengli	bWendu	yWendu	tianqi
2018-01-01	57	59	34	63	9	44	1.0	38	84.0	良	良	2	东北风	1-2级	3℃	-6℃	晴~多云
2018-01-02	50	49	28	50	7	33	0.8	46	57.0	优	优	1	东北风	1-2级	2℃	-5℃	阴~多云

4.4 数据分析

对爬取到的 2018-2022 年北京市雾霾数据和天气数据进行统计分析，得到 AQI 的变化趋势和与风力、风向等天气因素的关系，并对分析结果进行可视化处理。

三、 数据分析结果

4.1 雾霾近五年变化趋势

我们以月为分析粒度，研究月度下北京空气质量指数（AQI）的最大值、最小值和均值在近五年的变化趋势。

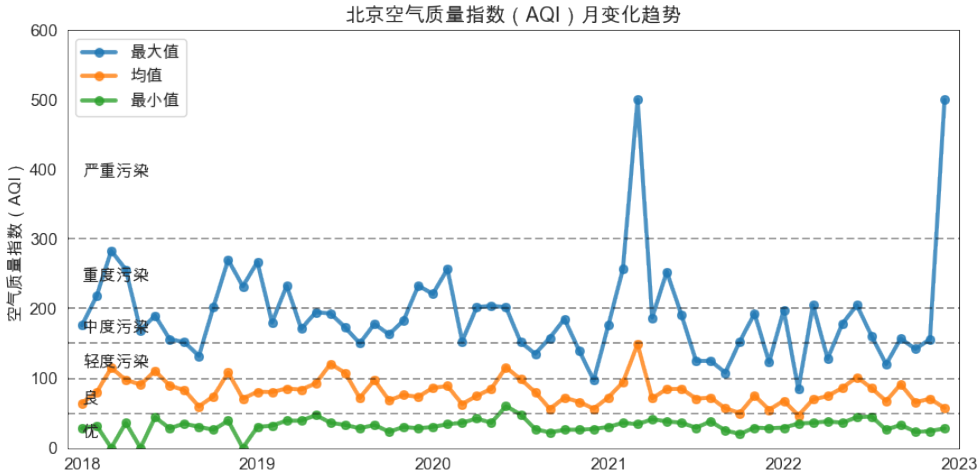

图 1 北京空气质量指数（AQI）月变化趋势

从图 1 中可以得到看出，总体上，北京月度空气质量指数的均值和最小值在近五年内维持稳定，而最大值呈下降趋势，但在 2021 年初和 2022 年末出现了极端值，空气质量达到了严重污染的程度。

进一步分析各空气质量等级占比的变化趋势。

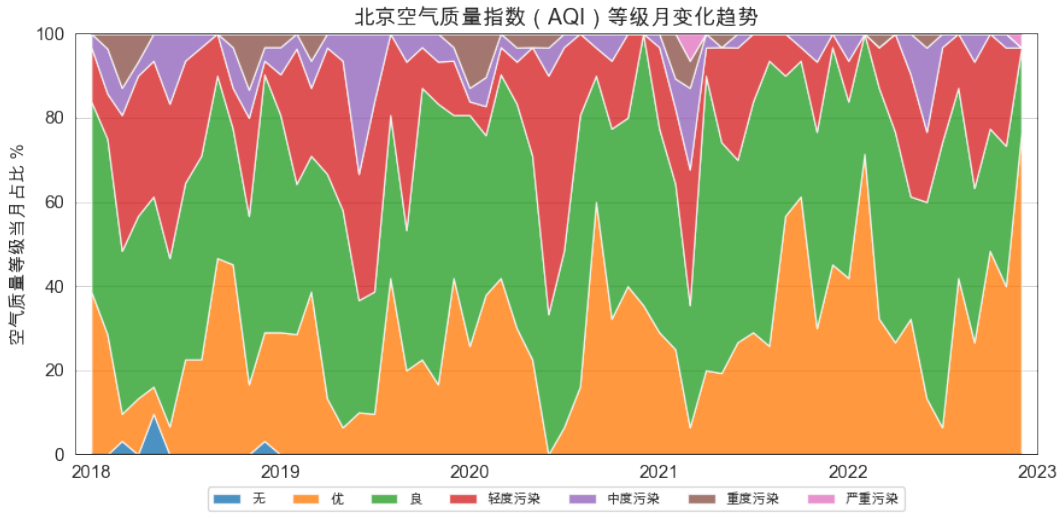

图 2 北京空气质量指数（AQI）等级月变化趋势

从图 2 中可以看出，北京空气质量等级呈现明显的季节性，冬春时节的空气质量显著差于夏秋时节，各程度污染占比显著高于夏秋时节对应占比。同季节下对比，总体上看，轻度污染、中度污染、重度污染的天数占比呈现下降趋势，2022 年下降尤为明显，而空气质量优良的天数占比相应呈现上升趋势，北京的空气质量得到改善。

4.2 雾霾与风向/风力的关系

首先分析空气质量指数（AQI）与风向的关系，我们绘制了北京空气质量指数（AQI）与八个不同风向的小提琴图。

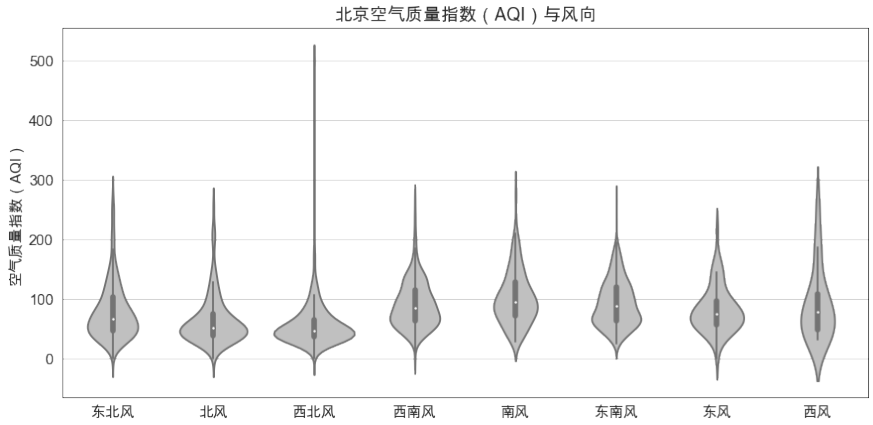

图 3 北京空气质量指数（AQI）与风向关系

从图 3 可以看出，总体上，北风或偏北风天气下的空气质量指数显著低于其他风向，空气质量更好，而南风或偏南风天气下的空气质量指数相对更高，空气质量更差。接着我们将风向和风力结合，共同分析其对北京空气质量的影响。

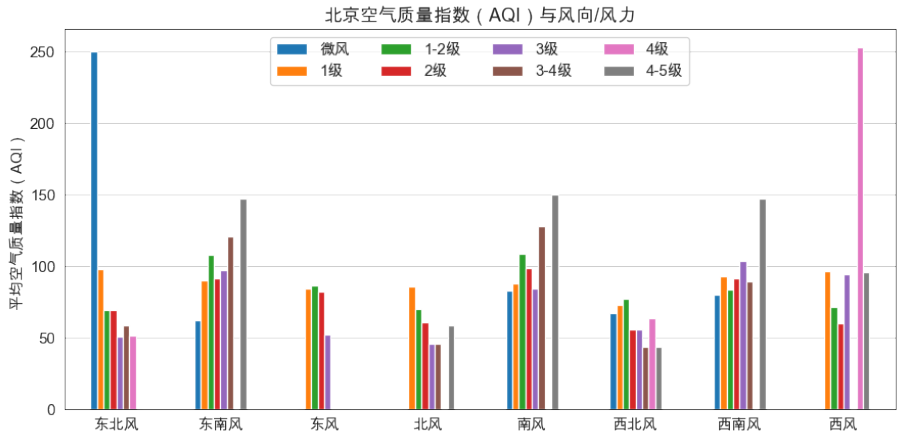

图 4 北京空气质量指数（AQI）与风向/风力柱形图

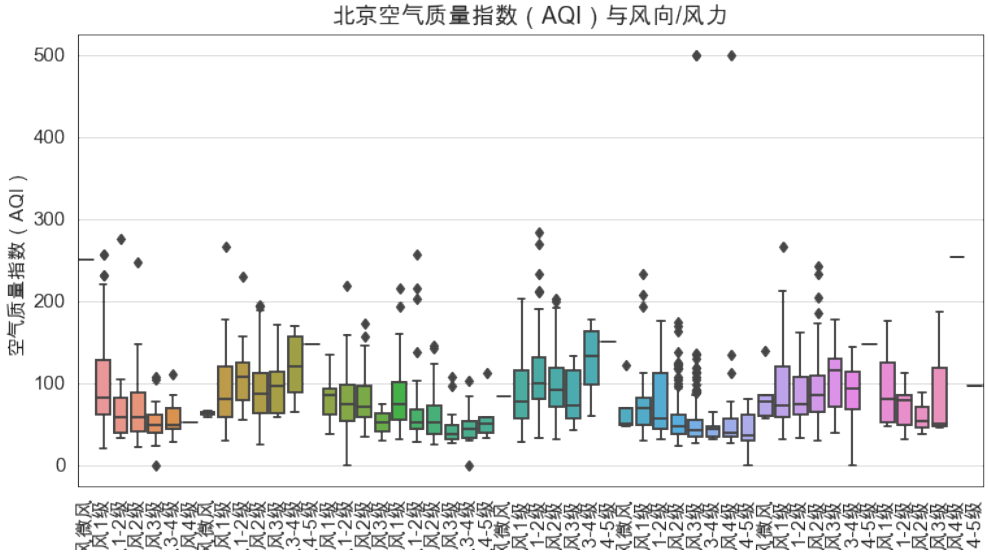

图 5 北京空气质量指数（AQI）与风向/风力箱线图

从图 4 和图 5 可以看出，与图 3 分析类似，北风和偏北风空气质量指数较低，南风和偏南风空气质量指数较高。在相同风向下对风力的影响进行进一步分析，北风或偏北风天气下，风力等级越高，污染指数越低，总体空气质量指数在 50 上下波动；而在南风或偏南风天气下，风力等级越高，污染指数也越高，总体空气质量指数在 100 上下波动；在东风和西风天气下，风力等级对污染指数的影响不显著。

四、 总结与展望

本项目基于课程的爬取天气数据项目，进一步爬取近五年北京市的空气质量数据和天气数据，并对得到的数据进行分析处理和可视化。通过数据分析发现，近五年来北京市的空气质量指数呈下降趋势，空气质量逐渐好转。同时，分析还发现了空气质量指数与风力和风向之间的关系：北风或偏北风天气下，空气质量指数较低，空气较好，而南风或偏南风天气下，空气质量指数较高，空气较差。更进一步，在北风或偏北风天气下，风力等级越高，空气质量指数越低，空气越好；而南风或偏南风天气下，风力等级越高，空气质量指数越高，空气越差。

本研究还存在一些不足和需要后续改进的地方：首先，我们可以尝试定量分析和更严谨的统计学回归方法给出风力、风向对于空气质量的定量影响。其次，

本研究的数据仅聚焦于北京一个城市，后续可以尝试采用并发爬虫爬取多个城市的数据，让研究样本更加丰富。