Python数据预处理

Scrapy链家新房数据爬取预处理

目的

- 通过爬虫爬取链家的新房数据，并进行预处理。
- 输出格式：csv文件，包含：
    - 名称
    - 地理位置（行政区、街道、地址3个字段分别存储）
    - 房型（只保留最小房型）
    - 面积（按照最小值，整数）
    - 均价（元，整数）
    - 总价（万元，保留小数点后4位）

环境

- Windows10 Pro 2004，macOS 10.15。
- PyCharm，Python3.7，Scrapy 2.4.1

分析

- 在`items.py`中，定义需要爬取的每一项房源的属性。
- 在`spider.py`中，根据网页源码结构，定义爬虫的 URL 以及对爬取数据的解析和清洗操作。
- 在`pipelines.py`中，通过`process_item()`将清洗完毕的数据按照格式写到磁盘，并且把需要记录值以供运算的数据放在全局变量中。在爬取结束调用`close_spider()`时，对全局变量记录的房价信息进行排序整理输出。
- 在`settings.py`中，定义爬虫工程的名称、UA 伪装、机器人协议遵守、管道初始化信息。
- 在`begin.py`中，调用执行整个爬虫程序。

项目结构

- PreProcessLianJia
    - PreProcLianJia
        - spiders
            - __init__.py
            - spider.py
        - items.py
        - middlewares.py
        - pipelines.py
        - settings.py

- begin.py
- LJ.csv
- Price.txt
- Report.md
- scrapy.cfg

源码

begin.py

```
from scrapy import cmdline

cmdline.execute("scrapy crawl PPLianJia".split())
```

items.py

```
import scrapy


class PreProcessLianJiaItem(scrapy.Item):
    Name = scrapy.Field()
    LocationDistrict = scrapy.Field()
    LocationBlock = scrapy.Field()
    LocationAddr = scrapy.Field()
    LDK = scrapy.Field()
    AreaSize = scrapy.Field()
    PricePerSqM = scrapy.Field()
    PricePerSuite = scrapy.Field()
    pass
```

spider.py

```
import scrapy
from PreProcessLianJia.PreProcLianJia.items import PreProcessLianJiaItem


page = 19

class PreProcessLianJiaSpider(scrapy.Spider):
    name = "PPLianJia"
    allowed_domains = ["bj.fang.lianjia.com"]
    start_urls = []
    if page >= 1:  # 根据待爬取页面数量，添加起始链接
        for i in range(1, page + 1):
            start_urls.append(f"https://bj.fang.lianjia.com/loupan/pg{i}")
```

```
    def parse(self, response, **kwargs):
        item = PreProcessLianJiaItem()
        for each in response.xpath("/html/body/div[4]/ul[2]/*"):  # 将每一条房源
数据进行整理
            item['Name'] = each.xpath("div/div[1]/a/text()").extract()[0]
            item['LocationDistrict'] =
each.xpath('div/div[2]/span[1]/text()').extract()[0]
            item['LocationBlock'] =
each.xpath('div/div[2]/span[2]/text()').extract()[0]
            item['LocationAddr'] = each.xpath('div/div[2]/a/text()').extract()
[0]
            try:
                item['LDK'] = each.xpath('div/a/span[1]/text()').extract()[0]
            except Exception as e:
                item['LDK'] = ''
            item['AreaSize'] = each.xpath('div/div[3]/span/text()').extract()
[0]
            item['AreaSize'] = int(item['AreaSize'][3:-1].split('-')[0])
            if '均价' in
each.xpath('div/div[6]/div[1]/span[2]/text()').extract()[0]:
                item['PricePerSqM'] =
int(each.xpath('div/div[6]/div[1]/span[1]/text()').extract()[0])
                item['PricePerSuite'] = item['AreaSize'] * item['PricePerSqM']
/ 10000
            else:
                item['PricePerSuite'] =
float(each.xpath('div/div[6]/div[1]/span[1]/text()').extract()[0])
                item['PricePerSqM'] = int(item['PricePerSuite'] * 10000 /
item['AreaSize'])
            yield item  # yield传递数据给下一个
```

pipelines.py

```
import csv

totalPrice = []
unitPrice = []


def get_num(x):
    return int(''.join(ele for ele in x if ele.isdigit()))


class PreProcessLianJiaPipeline:
    def process_item(self, item, spider):
        d_item = dict(item)
        totalPrice.append(d_item['PricePerSuite'])
        unitPrice.append(d_item['PricePerSqM'])
```

```
        d_item['PricePerSuite'] = format(d_item['PricePerSuite'], '.4f')

        with open("LJ.csv", 'a+') as f:
            csv_write = csv.writer(f)
            data_value = [d_item['Name'], d_item['LocationDistrict'],
d_item['LocationBlock'], d_item['LocationAddr'], d_item['LDK'],
d_item['AreaSize'], d_item['PricePerSqM'], d_item['PricePerSuite']]
            csv_write.writerow(data_value)

        return item

    def open_spider(self, spider):
        with open("LJ.csv", 'a+') as f:
            csv_write = csv.writer(f)

    def close_spider(self, spider):
        totalPrice.sort()
        unitPrice.sort()
        totalmin = totalPrice[0]
        totalmax = totalPrice[-1]
        totalmid = totalPrice[int((int((len(totalPrice)) / 2) +
int((len(totalPrice) + 1) / 2)) / 2)]
        unitmin = unitPrice[0]
        unitmax = unitPrice[-1]
        unitmid = unitPrice[int((int((len(unitPrice)) / 2) +
int((len(unitPrice) + 1) / 2)) / 2)]
        with open("Price.txt", 'a+') as ptxt:
            outstr = "{:.4f}".format(totalmin) + '\t' + "
{:.4f}".format(totalmid) + '\t' + "{:.4f}".format(totalmax) + '\t' +
 str(unitmin) + '\t' + str(unitmid) + '\t' + str(unitmax) + '\n'
            ptxt.write(outstr)
```

settings.py

```
BOT_NAME = 'PreProcLianJia'

SPIDER_MODULES = ['PreProcLianJia.spiders']
NEWSPIDER_MODULE = 'PreProcLianJia.spiders'

USER_AGENT = 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/86.0.4240.198 Safari/537.36'

ROBOTSTXT_OBEY = False

ITEM_PIPELINES = {
    'PreProcLianJia.pipelines.PreProcessLianJiaPipeline': 300,
}
```

运行结果

- 因为链家网站上很多房源信息不完善，程序在输出到 csv 文件时做了记录，但在计算房屋单价、总价时忽略了信息不完整的数据。程序输出的信息中删除了不该出现的空格，但是对于地址中带有中文标点符号的信息，进行原样保存。

LJ.csv

```
远洋新天地,门头沟,门头沟其它,长安街西延线与滨河路南延交汇处（东南侧）,1
室,1118,33000,3689.4000
北京书院,朝阳,惠新西街,北三环以北，惠新东街与北土城东路交汇处西行200米路北,1
室,67,112000,750.4000
金隅上城郡,昌平,北七家,北亚花园东路50米,3室,212,45000,954.0000
御河壹号庄园,顺义,顺义城,顺福路2号,6室,519,37000,1920.3000
旭辉城,房山,房山其它,西南六环良乡出口往南约2公里,2室,75,32000,240.0000
首创远洋禧瑞天著,亦庄开发区,通州其它,科创十一街与经海九路交汇处西南角（亦庄线次渠南站200
米）,2室,89,52695,468.9855
首开保利熙悦林语,大兴,瀛海,旧宫镇五环路与德贤路交叉口东150米,3室,220,52315,1150.9300
檀香府,门头沟,门头沟其它,京潭大街与潭柘十街交叉口,3室,124,43000,533.2000
金融街金悦府,大兴,黄村火车站,4号线义和庄地铁站往北400米,2室,83,50000,415.0000
中海丽春湖墅·合院,昌平,沙河,地铁昌平线沙河地铁站南600米,4室,304,54553,1658.4112
祥云赋,顺义,后沙峪,京承高速6出口（中央别墅区）罗马环岛北800米,3室,90,55583,500.2470
中海金樾和著,房山,房山其它,良常路官道路口西800米,3室,89,33000,293.7000
北京经开汀塘,亦庄开发区,通州其它,科创十一街与经海九路交汇处（地铁亦庄线次渠南站400米）,2
室,82,53980,442.6360
燕西华府,丰台,丰台其它,王佐镇青龙湖公园东1500米， 泉湖西路1号院（七区）， 泉湖西路1号院（六
区）,4室,60,49800,298.8000
国锐金岚,亦庄开发区,亦庄,荣华南路1号院,4室,285,78000,2223.0000
首开香溪郡,通州,通州其它,宋庄镇荷香街2号院,3室,90,38000,342.0000
天润福熙大道,朝阳,北苑,清河营东路1号院， 清河营东路3号院,3室,180,90000,1620.0000
通州万国城MOMA,通州,乔庄,果园环岛以东，运河西大街与玉桥西路交汇处,3室,112,71000,795.2000
领秀翡翠墅,丰台,丰台其它,王佐镇长青路南侧， 长青路88号院,4室,270,65000,1755.0000
奥园北京源墅,密云,密云其它,密关路与313省道交汇处,3室,110,21000,231.0000
北京城建北京合院,顺义,顺义其它,燕京街与通顺路交汇口东800米(仁和公园南),3
室,95,47000,446.5000
中骏西山天璟,门头沟,城子,永定楼北300米,4室,140,63000,882.0000
新光大中心,通州,北关,滨惠北一街1号院2号楼,1室,97,49800,483.0600
首开缇香郡,通州,通州其它,京津高速与张采路交叉口，京津高速于家务出口即到,3
室,89,29000,258.1000
珠光御景西园,丰台,丰台其它,长辛店长云路2号,3室,117,38600,451.6200
和光尘樾,朝阳,东坝,东五环七棵树桥出口向东1.5千米,4室,290,79800,2314.2000
北京城建北京合院,顺义,顺义其它,燕京街与通顺路交汇口东800米(仁和公园南),4
室,210,45000,945.0000
天恒水岸壹号,房山,良乡,良乡大学城西站地铁南侧400米，刺猬河旁,3室,185,58000,1073.0000
檀香府,门头沟,门头沟其它,京潭大街与潭柘十街交叉口,3室,208,43000,894.4000
观唐云鼎,密云,密云其它,溪翁庄镇密溪路39号院（云佛山度假村对面）,3室,171,30000,513.0000
运河铭著,通州,北关,商通大道与榆东一街交叉口，温榆河森林公园东500米,2室,100,46000,460.0000
万年广阳郡九号,房山,长阳,长阳清苑南街与汇商东路交汇处西北角,3室,139,48500,674.1500
华远裘马四季,门头沟,大峪,增产路16号院,3室,150,60000,900.0000
```

合景映月台,海淀,清河,安宁庄西路与小营西路交汇处西南侧，安宁庄西路与小营西路36号院项目,4室,182,120879,2200.0000
V7九间堂,通州,潞苑,通燕高速耿庄桥北出口中化石油对面,4室,220,68000,1496.0000
御汤山熙园,昌平,昌平其它,紧邻安泗路，距离北六环61号出口约2000米,4室,300,40000,1200.0000
华远和墅,大兴,南中轴机场商务区,南六环磁各庄桥沿南中轴向南2公里,5室,295,54000,1593.0000
天资华府,房山,长阳,房山区CSD政务大厅5号门,2室,94,45000,423.0000
观山源墅,房山,良乡,阳光北大街与多宝路交汇处西南（理工大学北校区西侧）,3室,290,47500,1377.5000
西山甲一号,丰台,丰台其它,长辛店生态城园博园南路路北500米,3室,118,57000,672.6000
电建金地华宸,门头沟,门头沟其它,长安街西延线南侧约500米,3室,180,64000,1152.0000
首创天阅西山,海淀,海淀北部新区,永丰路与北清路交汇处东北角，中关村壹号北,3室,175,80000,1400.0000
西山燕庐,门头沟,门头沟其它,长安街延线南约500米,3室,137,65000,890.5000
北科建泰禾丽春湖院子,昌平,沙河,中关村北延新核心，沙河水库边（地铁昌平线沙河站向南800米）,4室,379,50000,1895.0000
首开国风尚樾,朝阳,望京,望京南湖南路三帆中学对面,3室,146,120000,1752.0000
绿地海珀云翡,大兴,大兴其它,兴亦路京开高速东侧（黄村镇第一中心小学对面）,2室,102,70000,714.0000
中粮京西祥云,房山,长阳,地铁稻田站北800米，西邻京深路,4室,115,58000,667.0000
中冶德贤公馆,大兴,旧宫,德贤东路6号院（南四环榴乡桥东南角800米）,3室,186,77000,1432.2000
燕西华府,丰台,丰台其它,王佐镇青龙湖公园东1500米，泉湖西路1号院（七区），泉湖西路1号院（六区）,3室,195,52000,1014.0000
天恒京西悦府,房山,阎村,燕房线阎村地铁站东南角约189米,3室,119,38000,452.2000
尚峯壹號,顺义,顺义其它,中央别墅北区京承高速11号出口，天承环路8号院,0室,107,36000,385.2000
合景寰汇公馆,通州,武夷花园,滨河中路和新华大街交汇处西南角，往南200米,2室,77,65000,500.5000
K2十里春风,通州,通州其它,永乐店镇漷小路百菜玛工业园对面,2室,74,24500,181.3000
K2十里春风,通州,通州其它,永乐店镇漷小路百菜玛工业园对面,3室,155,28000,434.0000
奥园北京源墅,密云,密云其它,溪翁庄镇,3室,120,24000,288.0000
金隅·金麟府,大兴,大兴其它,景园北街贵派大厦首层（荣昌东街地铁站对面）,3室,89,52695,468.9855
住总正华·时代广场,大兴,天宫院,地铁4号线生物医药基地站南200米,0室,61,55000,335.5000
天恒京西悦府,房山,阎村,燕房线阎村地铁站东南角约189米,3室,175,50000,875.0000
东方太阳城,顺义,顺义其它,潮白河畔（顺义新城滨河森林公园西300米）,5室,411,54000,2219.4000
和悦华锦,大兴,大兴其它,博兴八路与兴海一街交叉口西南侧100米,3室,89,52695,468.9855
尊悦光华,朝阳,CBD,光华东里甲1号院3号楼,3室,133,130000,1729.0000
新潮嘉园二期,通州,潞苑,潞苑南大街185号,1室,65,58000,377.0000
中海云筑,大兴,大兴新机场,京开高速庞各庄桥西1500米（庞各庄镇宏轩饺子馆儿对面）,3室,89,37000,329.3000
中海云筑,大兴,大兴新机场,京开高速庞各庄桥西1500米，团结路北（庞各庄镇宏轩饺子馆儿对面）,4室,266,38000,1010.8000
远洋五里春秋,石景山,石景山其它,五里坨黑石头路前行500m,3室,290,52024,1508.6960
绿城西府海棠,石景山,石景山其它,隆恩寺路北人大附石景山分校西侧,,90,52024,468.2160
和锦薇棠,朝阳,常营,黄渠东路与朝阳北路辅路交汇处,3室,290,68921,1998.7090
首创·河著,顺义,顺义其它,京承高速11出口（昌金路）向东900 米路北,4室,248,53000,1314.4000
华萃西山,门头沟,门头沟其它,永定镇地铁S1号线石厂西南700米,3室,83,47000,390.1000
山屿西山著,海淀,海淀北部新区,海淀区温泉镇太舟坞,3室,89,53000,471.7000
林肯时代,亦庄开发区,亦庄,荣华路与荣京西街交叉口西侧300米（荣京东街站西侧600米）,2室,103,42000,432.6000
信达·国子郡,昌平,沙河,沙河镇北沙河中路与高教园附近,3室,79,48810,385.5990

华业玫瑰东筑,通州,临河里,梨园南街与临河里路交汇处，通州地铁土桥站南行800米,2室,97,50000,485.0000
北京东湾,通州,万达,通惠北路98号,1室,58,68500,397.3000
天恒世界集,大兴,高米店,盛坊路1号院,0室,60,55000,330.0000
中铁诺德阅墅,顺义,中央别墅区,中央别墅区优山美地D区西北侧,4室,220,62000,1364.0000
懋源·璟岳,丰台,玉泉营,南三环西路99号院,2室,132,113000,1491.6000
懋源·璟岳,丰台,玉泉营,南三环西路99号院,4室,465,113000,5254.5000
山屿西山著,海淀,海淀北部新区,北清路山屿西山著,4室,225,53000,1192.5000
御世佳府,通州,潞苑,朝阳北路延长线北侧,4室,222,70000,1554.0000
公园十七区,顺义,后沙峪,中央别墅区火沙路和裕庆路交汇口北500米,3室,90,55711,501.3990
首开熙悦观湖,房山,房山其它,房山区青龙湖镇中心区,西至魏各庄路,北至青龙湖管理处,3室,89,32500,289.2500
国誉府,顺义,马坡,聚源西路7号内205号,2室,80,43467,347.7360
金辰府,昌平,北七家,北五环定泗路北七家镇政府正南,3室,89,53000,471.7000
建邦·顺颐府,顺义,后沙峪,空港B区裕民大街30号千里马国际一层建邦·顺颐府售楼中心,3室,89,55583,494.6887
中铁·诺德春风和院,丰台,花乡,丰台区白盆窑六圈路和樊羊路交汇处东南,3室,89,67702,602.5478
萬橡悦府,昌平,回龙观,京藏高速与北清路交汇口东2公里,2室,73,54197,395.6381
中铁诺德·逸府,丰台,科技园区,樊羊路与六圈南路交汇处往南500米,2室,69,67702,467.1438
华萃西山,门头沟,门头沟其它,门头沟永定镇地铁S1号线石厂站西南700米,4室,135,52000,702.0000
碧桂园·京源著,延庆,延庆其它,百泉街与延康路交叉口北200米,3室,89,28000,249.2000
城市之光·东望,通州,通州其它,东五环化工桥向东京津高速第一出口通马路向南约500米,2室,70,51585,361.0950
合景泰富天汇,顺义,马坡,昌金路与通顺路交汇处,3室,89,38000,338.2000
金地旭辉·江山风华,大兴,黄村中,地铁4号线清源路站西侧800米,3室,89,55800,496.6200
水岸雁栖,怀柔,怀柔,雁栖镇京加路雁栖桥西北500米处,0室,56,34000,190.4000
懋源·璟玺,朝阳,中央别墅区,孙河京密路与京平辅路交叉口西行1000米,4室,262,86000,2253.2000
金隅学府,大兴,大兴其它,景园北街2号51-1号贵派大厦一楼,2室,82,52695,432.0990
台湖金茂悦,通州,通州其它,台湖镇亦庄新城惠民路与生态公园路交汇处,2室,80,50000,400.0000
中国铁建国际公馆,大兴,大兴其它,博兴十路中国铁建国际公馆,2室,80,52695,421.5600
迈宇平墅,顺义,顺义城,顺平路迈宇平墅,4室,217,38000,824.6000
中建国望府,丰台,丰台其它,射击场路中建国望府,5室,314,63630,1998.0000
阳光上东,朝阳,酒仙桥,东四环北路6号,3室,189,83000,1568.7000
北科建·翡翠华府,怀柔,怀柔,中高路与乐园大街交汇处（雁栖河生态廊道旁）,3室,95,40567,385.3865
保利绿城·和锦诚园,大兴,瀛海,8号线南段地铁站东1.6公里,3室,111,62000,688.2000
金悦郡,通州,通州其它,亦庄新城环景西一路与景盛南二街交叉口珠江逸景家园南区西侧,2室,72,36000,259.2000
北京恒大上河院,密云,密云其它,科技路东侧,2室,79,25500,201.4500
路劲·御合院,大兴,大兴新机场洋房别墅区,采育镇彩凤路与育进街交叉口,3室,89,29000,258.1000
熙红印,大兴,西红门,宏福路3号,2室,65,64400,418.6000
长安和玺,石景山,古城,古城南里长安和玺,4室,125,74000,925.0000
观承·望溪,顺义,后沙峪,京承高速8号出口东800米,4室,257,55800,1434.0600
富力首开·金禧璞瑆,顺义,顺义其它,高泗路四村段23号,5室,360,36000,1296.0000
熙悦天寰,丰台,丰台其它,鑫博西路与大灰厂东路交口西行20米,3室,89,58000,516.2000
金融街武夷·融御,通州,武夷花园,通胡大街与东六环西侧路交汇处东南角,2室,50,62000,310.0000
中骏云景台,房山,房山其它,官道北21号,2室,75,28000,210.0000
北京金茂府二期,丰台,宋家庄,宋家庄地铁站F口旁,2室,75,103000,772.5000
禧瑞金海,平谷,平谷其它,平蓟路与环镇东路交汇,2室,87,22000,191.4000
禧瑞金海,平谷,平谷其它,平蓟路与环镇东路交汇,2室,125,22000,275.0000

```
中国铁建·山语澜廷,房山,房山其它,阎吕路中国铁建·山语澜廷,4室,223,30000,669.0000
京澜誉府,通州,万达,玉带河大街乙72号,3室,99,60000,594.0000
电建·洺悦湾,大兴,旧宫,旧宫北路旧宫地铁站东侧300米,2室,75,54715,410.3625
观承·望溪,顺义,后沙峪,京承高速8号出口东800米,4室,130,55800,725.4000
西山上品湾MOMA,昌平,昌平其它,阳坊镇温南路与阳八路交汇处,3室,130,37500,487.5000
中骏天峰,门头沟,滨河西区,石龙东路中骏天峰,3室,89,53000,471.7000
中海寰宇时代,大兴,瀛海,8号线地铁瀛海站南200米,2室,49,55128,270.1272
合景·领汇长安,门头沟,滨河西区,北京市门头沟区S1号线桥户营站西300米,3室,78,50000,390.0000
大苑·海淀府,海淀,田村,西四环外2000米，田村路北大苑·海淀府,3室,170,110000,1870.0000
禹洲朗廷湾,通州,通州其它,朝阳北路和通顺路交汇处,3室,80,52508,420.0640
金茂北京国际社区,顺义,顺义其它,水色西路西侧,1室,50,36903,184.5150
住总如院,大兴,大兴新机场洋房别墅区,采育镇采华路南端,2室,98,31136,305.1328
北科建翡翠华庭,怀柔,怀柔,京密路辅路与怀杨路交叉口东侧100米,2室,88,29438,259.0544
中海甲叁號院,丰台,玉泉营,丰台恒丰路,3室,145,117500,1703.7500
熙悦宸著,大兴,西红门,福欣路与京开路交叉口东南侧,3室,81,64400,521.6400
亦庄橡树湾,亦庄开发区,马驹桥,马驹桥镇亦庄新城漷马路与兴华南街交汇处东300米,2
室,78,37000,288.6000
海淀幸福里,海淀,海淀北部新区,北清路海淀幸福里,3室,98,80000,784.0000
京投发展公园悦府,昌平,回龙观,回南北路与科星西路交汇路口处（8号线平西府站向东200米）,3
室,86,58500,503.1000
兴创屹墅,大兴,高米店,双高路兴创屹墅小区， 高米店南里,5室,464,70000,3248.0000
华贸铂金墅,朝阳,北苑,清苑路华贸天地北侧,3室,326,69000,2249.4000
天恒乐墅,房山,阎村,京港澳高速阎村出口西行1500米,4室,184,43370,798.0080
丽景长安二期,门头沟,冯村,永定镇石龙西路与三石路交叉口向南200米西侧,4室,155,52000,806.0000
观承别墅·大家,顺义,后沙峪,高丽营镇京承高速8号出口路东,4室,300,55000,1650.0000
卓越万科翡翠山晓,石景山,石景山其它,石景山石门路东侧500米,3室,90,49999,449.9910
```

Price.txt

- 从左到右依次为：最低总价、中位总价、最高总价、最低均价、中位均价、最高均价。

```
181.3000  513.0000  5254.5000 21000 52695 130000
```

Pandas北京空气质量数据处理

目的

- 计算北京空气质量数据：
    - 汇总计算 PM 指数年平均值的变化情况
    - 汇总计算 10-15 年 PM 指数和温度月平均数据的变化情况

环境

- Windows10 Pro 2004，macOS 10.15。
- PyCharm，Python3.7，Pandas 1.1.4

分析

- 使用 pandas 库：
- 利用`read_csv()`读取文件，用`groupby()`将数据分组整合，用`mean()`方法计算平均值，用`to_csv()`方法将得到的结果输出到文件。pandas 的整体使用思路与 SQL 数据库类似。

项目结构

- PreProcessPM25
  - pm25-data-for-five-chinese-cities
    - BeijingPM20100101_20151231.csv
    - ChengduPM20100101_20151231.csv
    - GuangzhouPM20100101_20151231.csv
    - ShanghaiPM20100101_20151231.csv
    - ShenyangPM20100101_20151231.csv
  - PM25_month.py
  - PM25_year.py
  - month_avg.csv
  - year_avg.csv

源码

PM25_year.py

```
import pandas as pd

# open file
FileNameStr = './pm25-data-for-five-chinese-
cities/BeijingPM20100101_20151231.csv'
df = pd.read_csv(FileNameStr, encoding='utf-8', usecols=[1, 6, 7, 8, 9])

# create avg row
# mean(axis=1) to get avg row
df['PM_avg'] = df.iloc[:, 1:5].mean(axis=1)
# group by year, calculate average of PM. output to file
df.groupby('year')['PM_avg'].mean().to_csv("year_avg.csv")
# output to console
print(df.groupby('year')['PM_avg'].mean())
```

PM25_month.py

```
import pandas as pd

# open file
FileNameStr = './pm25-data-for-five-chinese-
cities/BeijingPM20100101_20151231.csv'
df = pd.read_csv(FileNameStr, encoding='utf-8', usecols=[1, 2, 6, 7, 8, 9])

# create avg row
# mean(axis=1) to get avg row
df['PM_avg'] = df.iloc[:, 2:6].mean(axis=1)
# group by year, calculate average of PM. output to file
df.groupby(['year', 'month'])['PM_avg'].mean().to_csv("month_avg.csv")
# output to console
print(df.groupby(['year', 'month'])['PM_avg'].mean())
```

运行结果

year_avg.csv 年平均 PM 指数变化

```
year,PM_avg
2010,104.04572982326042
2011,99.0932403834184
2012,90.53876763535511
2013,98.40266354428444
2014,93.91770369524673
2015,85.85894216133943
```

month_avg.csv 月平均 PM 指数变化。

```
year,month,PM_avg
2010,1,90.40366972477064
2010,2,97.23994038748137
2010,3,94.04654442877292
2010,4,80.0724233983287
2010,5,87.0719131614654
2010,6,109.03893805309734
2010,7,123.4260752688172
2010,8,97.68343195266272
2010,9,122.79273504273505
2010,10,118.78436657681941
2010,11,138.38403614457832
2010,12,97.1157469717362
2011,1,44.87369985141159
2011,2,150.29017857142858
2011,3,57.99198717948718
2011,4,91.72067039106145
2011,5,65.10814606741573
```

```
2011,6,108.79465541490858
2011,7,107.38648648648649
2011,8,103.7338003502627
2011,9,94.96940194714882
2011,10,145.5568181818182
2011,11,109.43496503496503
2011,12,108.72139973082099
2012,1,118.92238805970149
2012,2,84.44202898550725
2012,3,96.47432432432433
2012,4,87.83588317107093
2012,5,90.96671490593343
2012,6,96.63418079096046
2012,7,80.64970930232558
2012,8,81.1653290529695
2012,9,59.95224719101124
2012,10,94.95135135135135
2012,11,87.43696275071633
2012,12,109.18729641693811
2013,1,183.19527027027036
2013,2,113.56646825396813
2013,3,114.57269265232975
2013,4,63.04780092592594
2013,5,89.14852150537635
2013,6,111.3548611111111
2013,7,74.93283917340521
2013,8,67.92361111111111
2013,9,85.7178240740741
2013,10,102.20878136200719
2013,11,85.1462962962963
2013,12,90.31776433691755
2014,1,107.9117383512545
2014,2,160.5138888888889
2014,3,103.18324372759857
2014,4,92.16064814814811
2014,5,64.95855734767025
2014,6,59.15462962962963
2014,7,91.7999551971326
2014,8,65.66823687752354
2014,9,68.23263888888889
2014,10,135.26971326164866
2014,11,106.3375
2014,12,76.62253584229393
2015,1,110.02273745519707
2015,2,103.44556051587303
2015,3,94.4834229390681
2015,4,79.39699074074069
2015,5,61.16756272401436
2015,6,60.33240740740742
```

```
2015,7,60.22950268817203
2015,8,45.89605734767028
2015,9,50.92476851851854
2015,10,77.25784050179212
2015,11,125.80312500000011
2015,12,162.17898745519724
```