Python 程序设计: 数据预处理作业
作业报告

吴镇均
2020211448
北京邮电大学　计算机科学与技术

2022 年 12 月 13 日

1　作业 1

1.1　题目要求

爬取并存储链家的新房数据，并进行预处理。

- 爬取起始网页：https://bj.fang.lianjia.com/loupan/
- 爬取信息的提取及存储要求
  - 信息以 csv 文件存储，应包括以下字段：名称，地理位置（3 个字段分别存储），房型（只保留最小房型），面积（按照最小值），均价（元，整数），总价（万元，保留小数点后 4 位）。有均价者按均价计算总价；无均价者按总价计算均价。
  - 对于所有字符串字段，要求去掉所有的前后空格
  - 删除面积缺失的房屋数据
- 数据统计
  - 找出总价最贵和最便宜的房子，以及总价的中位数。
  - 找出均价最贵和最便宜的房子，以及均价的中位数。
- 异常值处理
  - 列出总价在均值三倍标准差以外的房屋，展示其基本信息（如果太多可以只展示一部分），并分析其原因（找 4 条数据即可）
  - 通过箱型图原则判断并列出均价为异常值的房屋，展示其基本信息（如果太多可以只展示一部分），并分析其原因（找 4 条数据即可）
- 离散化处理
  - 对房屋的均价进行离散化处理，自行设定每个区间的长度并给出设置的理由，给出每个区间的房屋数量和所占比例

1.2 信息爬取

1.2.1 核心代码

这里只爬取前十八页，是由于本代码爬取时，只有前十八页有有效楼盘信息。

```
import requests
import csv
import parsel
import time

f1 = open('loupan.csv', mode='x', encoding='utf-8', newline='')
csv_writer1 = csv.DictWriter(f1, fieldnames=['名称', '地理位置（字段1）',
    '地理位置（字段2）', '地理位置（字段3）', '房型', '面积（㎡）', '均价（元/平米）',
    '总价（万元/套）'])
csv_writer1.writeheader()

for page in range(1, 19):
    time.sleep(1)
    url = 'https://bj.fang.lianjia.com/loupan/' + 'pg' + str(page) + '/'
    headers = {
        'User-Agent':
        'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko)
            Chrome/81.0.4044.138 Safari/537.36'
    }
    response = requests.get(url=url, headers=headers)
    selector = parsel.Selector(response.text)
    lis = selector.css('.resblock-list-wrapper li')
    dit = {}
    for li in lis:
        title = li.css('.resblock-name a::text').get().strip()
        dit['名称'] = title
        location1 = li.css('.resblock-location span::text').get().strip()
        dit['地理位置（字段1）'] = location1
        location2 = li.css('.resblock-location span::text').getall()[1].strip()
        dit['地理位置（字段2）'] = location2
        location2 = li.css('.resblock-location a::text').get().strip()
        dit['地理位置（字段3）'] = location2
        room = li.css('.resblock-room span::text').get()
        if room is not None:
            room = room.strip()
        dit['房型'] = room
        area = li.css('.resblock-area span::text').get()
        '''如果面积为空，那么不写入该行'''
        if area is None:
            continue
        area = area.lstrip('建面 ').rstrip('㎡').split('-')[0]
        dit['面积（㎡）'] = area
        unitPrice = li.css('.main-price span::text').get()
        total = li.css('.second::text').get()
```

```
        '''如果total为空，说明没有均价'''
        if total is None:
            total = unitPrice.split('-')[0]
            unitPrice = str(round((eval(total) / eval(area)) * 10000))
            total = str(format(eval(total), '.4f'))
        else:
            '''如果total不为空，就用单价来算总价'''
            total = str(format((eval(unitPrice) * eval(area)) / 10000, '.4f'))
        dit['均价（元/平米）'] = unitPrice
        dit['总价（万元/套）'] = total
        csv_writer1.writerow(dit)
```

1.2.2 文件展示

保存得到的 csv 文件如下：

	A	B	C	D	E	F	G	H
1	名称	地理位置（字段1）	地理位置（字段2）	地理位置（字段3）	房型	面积（㎡）	均价（元/	总价（万元/套）
2	北京岭秀	平谷	平谷其它	环山路与主环路交叉口东南(南一路)	3室	220	16000	352.0000
3	水岸壹号	房山	良乡	良乡大学城西站地铁南侧800米，刺猬河旁	3室	185	58000	1073.0000
4	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	2室	107	27000	288.9000
5	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	49000	490.0000
6	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	166	50000	830.0000
7	天恒世界集	大兴	高米店	西红门镇广平大街与盛坊路交叉口	1室	45	27000	121.5000
8	御汤山熙园	昌平	昌平其它	北京市昌平区小汤山镇顺沙路99号院	4室	300	40000	1200.0000
9	天资华府	房山	长阳	房山区CSD政务大厅5号门	3室	115	38000	437.0000
10	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	45000	936.0000
11	韩建·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	40000	1160.0000
12	国门智慧城	顺义	后沙峪	后沙峪枯柳树环岛西100米国门一号西侧	2室	50	30000	150.0000
13	中国铁建花语金郡	大兴	瀛海	南海子公园西侧(南五环旧忠桥向南第二个红绿灯西300米)	3室	150	70000	1050.0000
14	北辰墅院1900	顺义	马坡	顺兴街11号院望尊园	4室	251	35000	878.5000
15	香江别墅	昌平	昌平其它	北京市昌平区马池口镇百葛路366号院	4室	210	39300	825.3000
16	都丽华府	平谷	平谷其它	新平南路与林荫南街交汇处向西100米	2室	86	29000	249.4000
17	燕西华府	丰台	丰台其它	王佐镇青龙湖公园东1500米，	2室	60	42000	252.0000
18	水岸壹号	房山	良乡	良乡大学城西站地铁南侧800米，刺猬河旁	3室	122	48000	585.6000
19	鲁能钓鱼台美高梅公馆	丰台	刘家窑	南苑乡石榴庄(地铁宋家庄站D出口西150米)	4室	332	163000	5411.6000
20	尊悦日坛	朝阳	朝阳门外	日坛北路19号	1室	45	63000	283.5000
21	天恒摩墅	房山	房山其它	周口店镇政府东200米	3室	140	23000	322.0000
22	鲁能·格拉斯小镇	通州	通州其它	北京市通州区宋庄镇格拉斯小镇营销中心	3室	246	62000	1525.2000
23	兴创荣墅	大兴	大兴新机场洋房别墅区	北京市大兴区育胜街	3室	188	25000	470.0000
24	丽都壹号	朝阳	酒仙桥	将台路与驼房营路交叉口向北150米，将府家园北里	1室	61	95000	579.5000
25	韩建·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）	4室	93	40000	372.0000
26	北辰墅院1900	顺义	马坡	顺兴街11号院望尊园	2室	83	36000	298.8000
27	润泽御府	朝阳	北苑	北京市朝阳区北五环顾家庄桥向北约2.6公里	4室	630	120000	7560.0000
28	中骏西山天璟	门头沟	城子	西山永定楼北300米	4室	117	65000	760.5000
29	兴创国际中心	大兴	西红门	西红门镇欣宁街与宏康路交叉口向西500米	0室	1450	38000	5510.0000
30	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）	0室	195	52000	1014.0000
31	京西悦府	房山	阎村	燕房线阎村地铁站东南角约189米	3室	120	33000	396.0000
32	合景寰汇公馆	通州	武夷花园	北京市通州区滨河中路西侧（合景寰汇公馆）	2室	77	42000	323.4000
33	K2十里春风	通州	通州其它	北京市通州区	2室	74	24500	181.3000
34	K2十里春风	通州	通州其它	北京市通州区	3室	155	28000	434.0000
35	玺萌壹號院	丰台	草桥	西南三环嘉园路与镇国寺北街交叉口	5室	320	90000	2880.0000
36	北京书院	朝阳	惠新西街	北京市朝阳区北土城东路辅路	1室	67	145000	971.5000
37	旭辉26号街区	顺义	顺义其它	临空经济核心区南法信地铁站南300米，信中北街16号院	1室	27	27000	72.9000
38	金隅上城郡	昌平	北七家	北亚花园东路50米	4室	212	45000	954.0000
39	中铁华侨城和园	大兴	瀛海	南五环南海子公园西侧约500米	3室	154	60000	924.0000
40	顺鑫颐和天璟	顺义	顺义其它	新城右堤路与昌金路交汇处向北200米	3室	110	33000	363.0000
41	顺鑫颐和天璟	顺义	顺义其它	新城右堤路与昌金路交汇处向北200米	4室	278	33000	917.4000
42	未来公元	昌平	北七家	京承高速北七家出口800米路南	3室	104	61183	636.3032
43	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
44	复地运河公馆	通州	武夷花园	通州运河核心区临滨河西路	2室	89	43000	382.7000
45	珠光御景西园	丰台	丰台其它	北京市丰台区长辛店长云路2号珠江御景营销中心	3室	117	40000	468.0000
46	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
47	月亮河七星公馆	通州	武夷花园	通燕高速耿庄桥出口南200米月亮河，河滨路1号	1室	55	68000	374.0000
48	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	1室	65	120000	780.0000
49	元熙华府	丰台	宋家庄	东南三环东铁营桥向南600米	3室	126	87000	1096.2000
50	京贸国际公馆	通州	九棵树(家乐福)	怡乐中路299号院（广渠快速路二期出口向南1000米）	1室	72	64000	460.8000
51	凯德麓语	昌平	昌平其它	兴寿镇京承高速G11出口向西怀昌路北侧	3室	280	35000	980.0000

图 1: 获取到的楼盘数据

1.3 数据预处理

1.3.1 数据统计

核心代码 数据统计的核心代码如下：

```
'''总价'''
```

```
print('\n【总价最贵】')
print(loupan.loc[loupan['总价（万元/套）'] == loupan['总价（万元/套）'].max()])
print('\n【总价最便宜】')
print(loupan.loc[loupan['总价（万元/套）'] == loupan['总价（万元/套）'].min()])
print('\n【总价中位数】')
print(loupan['总价（万元/套）'].median())

'''均价'''
print('\n【均价最贵】')
print(loupan.loc[loupan['均价（元/平米）'] == loupan['均价（元/平米）'].max()])
print('\n【均价最便宜】')
print(loupan.loc[loupan['均价（元/平米）'] == loupan['均价（元/平米）'].min()])
print('\n【均价中位数】')
print(loupan['均价（元/平米）'].median())
```

结果 统计结果如下：

```
【总价最贵】
      名称 地理位置（字段1） 地理位置（字段2）          地理位置（字段3）  房型  面积（㎡）  均价（元/平米）  总价（万元/套）
102  北京壹号总部        大兴       亦庄  台湖镇光机电一体化产业基地科创东二街5号  1室   3127    28000 8755.6000

【总价最便宜】
      名称 地理位置（字段1） 地理位置（字段2）                地理位置（字段3）  房型  面积（㎡）  均价（元/平米）  总价（万元/套）
35  旭辉26号街区        顺义      顺义其它  临空经济核心区南法信地铁站南300米，信中北街16号院  1室     27     27000   72.9000

【总价中位数】
534.0

【均价最贵】
       名称 地理位置（字段1） 地理位置（字段2）          地理位置（字段3）  房型  面积（㎡）  均价（元/平米）  总价（万元/套）
17  鲁能钓鱼台美高梅公馆        丰台       刘家窑  南苑乡石榴庄(地铁宋家庄站D出口西150米)  4室    332    163000 5411.6000

【均价最便宜】
     名称 地理位置（字段1） 地理位置（字段2）          地理位置（字段3）  房型  面积（㎡）  均价（元/平米）  总价（万元/套）
0  北京岭秀        平谷      平谷其它  环山路与主环路交叉口东南(南一路)  3室    220     16000  352.0000

【均价中位数】
47500.0
```

图 2: 统计结果

1.3.2 异常值处理

对总价在三倍标准差以外的房屋认为是异常数据，对均价在箱型图原则以外的房屋认为是异常数据。

核心代码 异常值处理的核心代码如下：

```
'''总价异常值'''
print("\n【总价异常值】")
min_mask = loupan['总价（万元/套）'] < (loupan['总价（万元/套）'].mean() - 3 *
    loupan['总价（万元/套）'].std())
max_mask = loupan['总价（万元/套）'] > (loupan['总价（万元/套）'].mean() + 3 *
    loupan['总价（万元/套）'].std())
mask = min_mask | max_mask
print(loupan.loc[mask])

'''均价异常值'''
```

```
print('\n【均价异常值】')
plt.boxplot(x=loupan['均价（元/平米）'])
plt.show()
q1 = loupan['均价（元/平米）'].quantile(q=0.25)
q3 = loupan['均价（元/平米）'].quantile(q=0.75)
low_mask = loupan['均价（元/平米）'] < q1 - 1.5*(q3-q1)
high_mask = loupan['均价（元/平米）'] > q3 + 1.5*(q3-q1)
mask = low_mask | high_mask
print(loupan.loc[mask])
```

结果 均价箱型图：

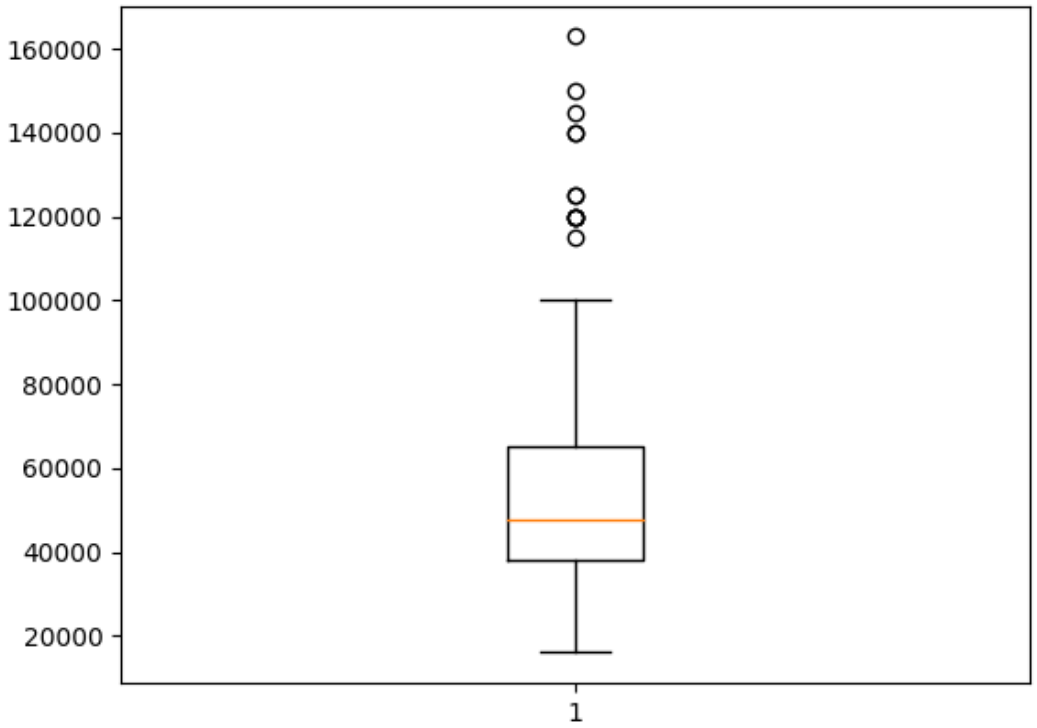
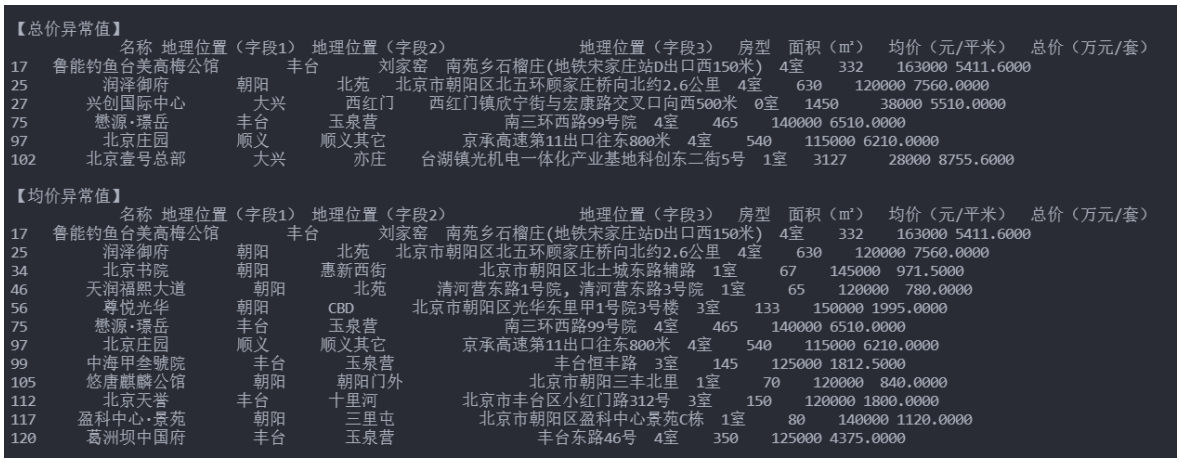


图 3: 均价箱型图

异常值处理结果如下：



【总价异常值】

	名称	地理位置（字段1）	地理位置（字段2）	地理位置（字段3）	房型	面积（㎡）	均价（元/平米）	总价（万元/套）
17	鲁能钓鱼台美高梅公馆	丰台	刘家窑	南苑乡石榴庄(地铁宋家庄站D出口西150米)	4室	332	163000	5411.6000
25	润泽御府	朝阳	北苑	北京市朝阳区北五环顾家庄桥向北约2.6公里	4室	630	120000	7560.0000
27	兴创国际中心	大兴	西红门	西红门镇欣宁街与宏康路交叉口向西500米	0室	1450	38000	5510.0000
75	懋源·璟岳	丰台	玉泉营	南三环西路99号院	4室	465	140000	6510.0000
97	北京庄园	顺义	顺义其它	京承高速第11出口往东800米	4室	540	115000	6210.0000
102	北京壹号总部	大兴	亦庄	台湖镇光机电一体化产业基地科创东二街5号	1室	3127	28000	8755.6000

【均价异常值】

	名称	地理位置（字段1）	地理位置（字段2）	地理位置（字段3）	房型	面积（㎡）	均价（元/平米）	总价（万元/套）
17	鲁能钓鱼台美高梅公馆	丰台	刘家窑	南苑乡石榴庄(地铁宋家庄站D出口西150米)	4室	332	163000	5411.6000
25	润泽御府	朝阳	北苑	北京市朝阳区北五环顾家庄桥向北约2.6公里	4室	630	120000	7560.0000
34	北京书院	朝阳	惠新西街	北京市朝阳区北土城东路辅路	1室	67	145000	971.5000
46	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	1室	65	120000	780.0000
56	尊悦光华	朝阳	CBD	北京市朝阳区光华东里甲1号院3号楼	3室	133	150000	1995.0000
75	懋源·璟岳	丰台	玉泉营	南三环西路99号院	4室	465	140000	6510.0000
97	北京庄园	顺义	顺义其它	京承高速第11出口往东800米	4室	540	115000	6210.0000
99	中海甲叁號院	丰台	玉泉营	丰台恒丰路	3室	145	125000	1812.5000
105	悠唐麒麟公馆	朝阳	朝阳门外	北京市朝阳三丰北里	1室	70	120000	840.0000
112	北京天誉	丰台	十里河	北京市丰台区小红门路312号	3室	150	120000	1800.0000
117	盈科中心·景苑	朝阳	三里屯	北京市朝阳区盈科中心景苑C栋	1室	80	140000	1120.0000
120	葛洲坝中国府	丰台	玉泉营	丰台东路46号	4室	350	125000	4375.0000

图 4: 异常值处理结果

总价异常值原因分析

- **鲁能钓鱼台美高梅公馆**：户型格局方面很不错，而且地理位置也好，出行便利，市区的别墅。同时占地面积不小，因此总价高。
- **润泽御府：**地理位置比较好，挨着北苑天通苑，算是很靠近市区的别墅。同时占地面积不小，因此总价高。
- **兴创国际中心：**周围地铁和商场等较多，同时占地面积大。
- **北京庄园：**精装满配，独栋现房，带花园，有泳池，有会所。是配套别墅，占地面积大，因此总价高。

均价异常值原因分析

- **北京书院：**位置好，五证齐全，在三环内，因此均价高。
- **天润福熙大道：**环境不错，周边还在发展，配套已经很全面，去望京很近，因此均价高。
- **尊悦光华：**位置好，在儿研所和清华附小附近，适合金融人士的需求，周边繁华，因此均价高。
- **盈科中心 · 景苑：**位于三里屯，位置好，配套齐全。因此均价高。

1.3.3 离散化处理

核心代码 离散化处理的核心代码如下：

```
print('\n【均价离散化】')
bins = [0, 30000, 40000, 50000, 60000, 80000, 200000]
cuts = pd.cut(loupan['均价（元/平米）'], bins)
print('离散房屋数量：')
print(pd.value_counts(cuts, sort=False))
pd.set_option('display.float_format', lambda x: format(x, '.2%'))
```

```
print('\n离散所占比例：')
print(pd.value_counts(cuts, normalize=True, sort=False))
```

结果 离散化处理结果如下：

```
【均价离散化】
离散房屋数量：
(0, 30000]         27
(30000, 40000]     39
(40000, 50000]     27
(50000, 60000]     21
(60000, 80000]     35
(80000, 200000]    25
Name: 均价（元/平米）, dtype: int64

离散所占比例：
(0, 30000]        15.52%
(30000, 40000]    22.41%
(40000, 50000]    15.52%
(50000, 60000]    12.07%
(60000, 80000]    20.11%
(80000, 200000]   14.37%
Name: 均价（元/平米）, dtype: float64
```

图 5: 离散化处理结果

区间设置理由 一开始设置区间为 30000 是由于 0～20000 之间仅有个位数房屋，所占比例低，没有参考价值。之后设置每次递增 10000 是由于均价在 30000～60000 的房屋较多，适合相对更细的切割。而 60000～80000 设置是由于这一段房屋数量较少，因此采用 20000 的区间设置。之后设置区间 80000～200000 是由于最高均价为 163000，这个区间可以囊括之后的所有房屋，同时所占比例没有极低，具有参考价值。

1.3.4 完整核心代码

整体核心代码如下：

```
import matplotlib.pyplot as plt
import pandas as pd

pd.set_option('display.max_columns', 500)
pd.set_option('display.width', 1000)
pd.set_option('display.float_format', lambda x: '%.4f' % x)
loupan = pd.read_csv('loupan.csv', sep=',', encoding="utf-8")

'''总价'''
print('\n【总价最贵】')
print(loupan.loc[loupan['总价（万元/套）'] == loupan['总价（万元/套）'].max()])
print('\n【总价最便宜】')
print(loupan.loc[loupan['总价（万元/套）'] == loupan['总价（万元/套）'].min()])
print('\n【总价中位数】')
print(loupan['总价（万元/套）'].median())

```

```
'''均价'''
print('\n【均价最贵】')
print(loupan.loc[loupan['均价（元/平米）'] == loupan['均价（元/平米）'].max()])
print('\n【均价最便宜】')
print(loupan.loc[loupan['均价（元/平米）'] == loupan['均价（元/平米）'].min()])
print('\n【均价中位数】')
print(loupan['均价（元/平米）'].median())

'''总价异常值'''
print("\n【总价异常值】")
min_mask = loupan['总价（万元/套）'] < (loupan['总价（万元/套）'].mean() - 3 *
    loupan['总价（万元/套）'].std())
max_mask = loupan['总价（万元/套）'] > (loupan['总价（万元/套）'].mean() + 3 *
    loupan['总价（万元/套）'].std())
mask = min_mask | max_mask
print(loupan.loc[mask])

'''均价异常值'''
print('\n【均价异常值】')
plt.boxplot(x=loupan['均价（元/平米）'])
plt.show()
q1 = loupan['均价（元/平米）'].quantile(q=0.25)
q3 = loupan['均价（元/平米）'].quantile(q=0.75)
low_mask = loupan['均价（元/平米）'] < q1 - 1.5*(q3-q1)
high_mask = loupan['均价（元/平米）'] > q3 + 1.5*(q3-q1)
mask = low_mask | high_mask
print(loupan.loc[mask])

print('\n【均价离散化】')
bins = [0, 30000, 40000, 50000, 60000, 80000, 200000]
cuts = pd.cut(loupan['均价（元/平米）'], bins)
print('离散房屋数量：')
print(pd.value_counts(cuts, sort=False))
pd.set_option('display.float_format', lambda x: format(x, '.2%'))
print('\n离散所占比例：')
print(pd.value_counts(cuts, normalize=True, sort=False))
```

2 作业 2

2.1 题目要求

分析处理 2015 年北京市 PM2.5 指数数据集空值。

- 原始数据集：BeijingPM20100101_20151231.csv
- 数据抽取及存储：从原始数据集中抽取 2015 年度数据，存储为新的 csv 文件
- 找出空值：对新的 csv 文件，找出存在的空值列及相应的空值数量

- 空值处理方法：对所有存在空值的列，给出空值的处理方法及理由，要求处理方法必须可在本数据集范围内执行
- 空值处理并存储：按照自己的处理方法，通过 pandas、numpy 或 python 方法对空值进行处理，完成后给出新的空值列信息，并将处理后的数据（不涉及空值的列应原样保留）存储为新的 csv 文件

2.2 核心代码

该代码包括以下三个部分：

- 抽取 2015 年的数据；
- 统计含有空值的列及对应的空值数量；
- 空值处理。

```
import pandas as pd

pd.set_option('display.max_columns', 500)
pd.set_option('display.width', 1000)

# 数据抽取，注意这里的读取会读取出 index 信息
raw = pd.read_csv('BeijingPM20100101_20151231.csv')
condition = raw['year'] == 2015
after = raw.loc[condition]
# 存储 2015 的 PM 信息。不存储 index 信息
after.to_csv('BeijingPM2015.csv', index=False)

new = pd.read_csv('BeijingPM2015.csv')

# 含空值列及对应空值数量统计输出
print('【缺失值统计】')
print(new.isnull().sum())

# 删除缺失率过高的列
new = new.drop('PM_Dongsihuan', axis=1)

# 前向填充
new['cbwd'] = new['cbwd'].fillna(method='ffill')
new['precipitation'] = new['precipitation'].fillna(method='ffill')
new['Iprec'] = new['Iprec'].fillna(method='ffill')
new['DEWP'] = new['DEWP'].fillna(method='ffill')
new['HUMI'] = new['HUMI'].fillna(method='ffill')
new['PRES'] = new['PRES'].fillna(method='ffill')
new['TEMP'] = new['TEMP'].fillna(method='ffill')

# 线性插值
new = new.interpolate(method='linear').round(2)
```

```
33
34 # 精度调整
35 new['PM_Dongsi'] = new['PM_Dongsi'].round()
36 new['PM_Nongzhanguan'] = new['PM_Nongzhanguan'].round()
37 new['PM_US Post'] = new['PM_US Post'].round()
38
39 new.to_csv('BeijingPM2015_Process.csv', index=False)
```

统计得到的缺失值信息　如图，其中每行包含列名和空值个数两个属性。

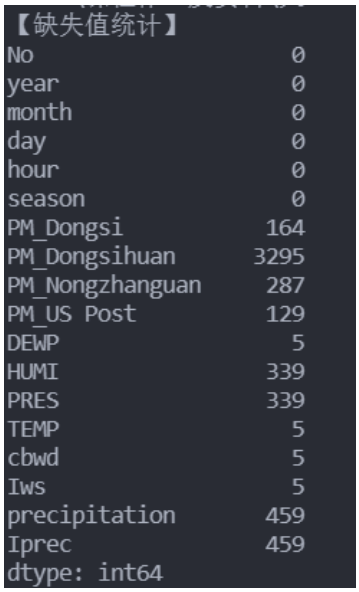


图 6: 缺失值信息

采用的空值处理方式　主要分为以下三种：

- **删除**。该空值处理方法只针对 PM_Dongsihuan 列，该列重要性高、缺失率高。由于该列缺失值接近 40% ，因此进行插值或填充都容易严重偏离实际，且该列与其他列没有非常紧密的联系，不适合通过其他列来进行估算，因此采取直接删除的方式。

- **前向填充**。对 cbwd 、precipitation 及 Iprec 列采用这种空值处理方式。

  对于 cbwd 而言，该列总体上重要性低、缺失率低。空值少且空值分布不连续，同时该列在同一段时间内大概率保持不变，因此适合前向填充。

  而对于 precipitation、Iprec，总体上重要性低、缺失率低。同时由于这两列大部分情况下为 0 ，而不为 0 时在同一段时间内浮动也较小。因此对这三列采用前向填充。

  对于 DEWP 、TEMP 、HUMI 、PRES，重要性低、缺失率低。同时相邻时间内波动小，因此采用前向填充。

- **线性插值**。对未删除的其他三个地点的 PM 值和 Iws 列采用这种空值处理方式。

其中，对于记录 PM 值的其余三个列而言，重要性高、缺失率低。通过分析可知其空值分布不集中，同时由于数据是每小时记录一次，同一段时间内的变化趋势一致，因此采用线性插值的空值处理方式。

而对于 Iws 列而言，重要性低、缺失率低。但由于精度较高，且时时发生变化且变化趋势相邻时间内一致，因此采用线性插值。

同时，为了确保和原数据保持每列的精度统一，会在线性插值后进行精度调整。

2.3 文件展示

2.3.1 提取到的 year 为 2015 的 PM 数据文件

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
7401	51224	2015	11	5	7	3	209		211	203	9	93	1027	10	cv	3.56	0	0
7402	51225	2015	11	5	8	3	184		194	175	9	87	1028	11	cv	4.45	0	0
7403	51226	2015	11	5	9	3	134		155	117	8	81	1029	11	NE	4.02	0	0
7404	51227	2015	11	5	10	3	108		104	90	8	87	1030	10	NE	8.04	0	0
7405	51228	2015	11	5	11	3	70		75	62	6	76	1030	10	NE	12.06	0	0
7406	51229	2015	11	5	12	3	55	58	60	49	5	75	1030	9	NE	16.08	0	0
7407	51230	2015	11	5	13	3	37	34	35	33	5	75	1030	9	NE	20.1	0	0
7408	51231	2015	11	5	14	3	26	24	27	28	4	75	1030	8	NE	24.12	0	0
7409	51232	2015	11	5	15	3	20	24	23	24	3	75	1031	7	NE	28.14	0	0
7410	51233	2015	11	5	16	3	21	24	23	26	2	75	1031	6	NE	33.06	0	0
7411	51234	2015	11	5	17	3	23	25	24	23	1	70	1032	6	NE	38.87	0	0
7412	51235	2015	11	5	18	3	24	23	25	20	1	75	1032	5	NE	44.68	0	0
7413	51236	2015	11	5	19	3	22	25	26	24	0	75	1032	4	NE	51.83	0	0
7414	51237	2015	11	5	20	3	24	20	18	24	0	75	1033	4	NE	57.64	0	0
7415	51238	2015	11	5	21	3	23	22	18	18	0	75	1033	4	NE	62.56	0	0
7416	51239	2015	11	5	22	3	19	19	18	16	0	75	1033	4	NE	66.58	0	0
7417	51240	2015	11	5	23	3	20	16	17	17	0	80	1032	3	NE	73.73		
7418	51241	2015	11	6	0	3	20	16	15	17	−1	69	1032	4	NE	78.65		
7419	51242	2015	11	6	1	3	15	17	14	20	−1	69	1031	4	NE	83.57		
7420	51243	2015	11	6	2	3	19	12	11	17	−2	64	1031	4	NE	88.49	0	0
7421	51244	2015	11	6	3	3	20	16	19	18	−2	64	1031	4	NE	93.41		
7422	51245	2015	11	6	4	3	20	17	16	21	−2	64	1030	4	NE	99.22		
7423	51246	2015	11	6	5	3	19	16	21	14	0	86	1031	2	NE	105.03		
7424	51247	2015	11	6	6	3	16	20	18	18	0	93	1032	1	NE	110.84	1.8	1.8
7425	51248	2015	11	6	7	3	12	17	16	13	0	93	1032	1	NE	116.65	1.8	3.6
7426	51249	2015	11	6	8	3	12	13	17	12	0	93	1032	1	NE	122.46	1.3	4.9
7427	51250	2015	11	6	9	3	15		16	16	0	93	1032	1	NE	127.38	1.9	6.8
7428	51251	2015	11	6	10	3	15		12	12	0	93	1032	1	NE	131.4	1	7.8
7429	51252	2015	11	6	11	3	12		18	17	1	100	1032	1	NW	4.92	0	0
7430	51253	2015	11	6	12	3	17		19	21	1	100	1031	1	NW	9.84	0	0
7431	51254	2015	11	6	13	3	21	16	19	26	1	100	1030	1	NW	14.76	0	0
7432	51255	2015	11	6	14	3	17		20	23	1	100	1030	1	NW	18.78	0	0
7433	51256	2015	11	6	15	3	18		23	25	1	100	1029	1	NE	4.02	0	0
7434	51257	2015	11	6	16	3	15		16	12	1	93	1029	2	NE	9.83	0	0
7435	51258	2015	11	6	17	3	16		15	12	0			2	NE	15.64	0	0
7436	51259	2015	11	6	18	3	13		12	15	0			2	NE	20.56	0	0
7437	51260	2015	11	6	19	3	17		16	19	1			2	NE	24.58	0	0
7438	51261	2015	11	6	20	3	18		19	19	0			2	NE	29.5	0	0
7439	51262	2015	11	6	21	3	13		10	17	0			2	NE	35.31	0	0
7440	51263	2015	11	6	22	3	9		12	16	0			2	NE	41.12	0	0
7441	51264	2015	11	6	23	3	10		7	14	0			2	NE	46.04	0	0
7442	51265	2015	11	7	0	3	11		12	8	−1			2	NE	51.85	0	0
7443	51266	2015	11	7	1	3	9		12	12	−1			2	NE	57.66	1	1
7444	51267	2015	11	7	2	3	11		17	14	0	93	1029	1	NW	4.92	0.6	1.6
7445	51268	2015	11	7	3	3	15		12	12	0	93	1029	1	NE	4.02	0.6	2.2
7446	51269	2015	11	7	4	3	14		11	8	0	93	1028	1	NE	9.83	0.1	2.3
7447	51270	2015	11	7	5	3	10		10	9	0	100	1028	0	NW	5.81	0	0
7448	51271	2015	11	7	6	3	5		8	7	0	100	1028	0	NE	5.81	0	0
7449	51272	2015	11	7	7	3	8		5	9	−1	86	1028	1	NE	11.62	0	0
7450	51273	2015	11	7	8	3	8		7	14	−1	86	1029	1	NW	4.02	0	0

图 7: 2015 的 PM 数据文件

2.3.2 经过空值处理后的 year 为 2015 的 PM 数据文件

注意：对重要性高、缺失率高的列 PM_Dongsihuan 进行了删除，详细可查看2.2。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
7401	51224	2015	11	5	7	3	209	211	203	9	93	1027	10	cv	3.56	0	0
7402	51225	2015	11	5	8	3	184	194	175	9	87	1028	11	cv	4.45	0	0
7403	51226	2015	11	5	9	3	134	155	117	8	81	1029	11	NE	4.02	0	0
7404	51227	2015	11	5	10	3	108	104	90	8	87	1030	10	NE	8.04	0	0
7405	51228	2015	11	5	11	3	70	75	62	6	76	1030	10	NE	12.06	0	0
7406	51229	2015	11	5	12	3	55	60	49	5	75	1030	9	NE	16.08	0	0
7407	51230	2015	11	5	13	3	37	35	33	5	75	1030	9	NE	20.1	0	0
7408	51231	2015	11	5	14	3	26	27	28	4	75	1030	8	NE	24.12	0	0
7409	51232	2015	11	5	15	3	20	23	24	3	75	1031	7	NE	28.14	0	0
7410	51233	2015	11	5	16	3	21	23	26	2	75	1031	6	NE	33.06	0	0
7411	51234	2015	11	5	17	3	23	24	23	1	70	1032	6	NE	38.87	0	0
7412	51235	2015	11	5	18	3	24	25	20	1	75	1032	5	NE	44.68	0	0
7413	51236	2015	11	5	19	3	22	26	24	0	75	1032	4	NE	51.83	0	0
7414	51237	2015	11	5	20	3	24	18	24	0	75	1033	4	NE	57.64	0	0
7415	51238	2015	11	5	21	3	23	18	18	0	75	1033	4	NE	62.56	0	0
7416	51239	2015	11	5	22	3	19	18	16	0	75	1033	4	NE	66.58	0	0
7417	51240	2015	11	5	23	3	20	17	17	0	80	1032	3	NE	73.73	0	0
7418	51241	2015	11	6	0	3	20	15	17	−1	69	1032	4	NE	78.65	0	0
7419	51242	2015	11	6	1	3	15	14	20	−1	69	1031	4	NE	83.57	0	0
7420	51243	2015	11	6	2	3	19	11	17	−2	64	1031	4	NE	88.49	0	0
7421	51244	2015	11	6	3	3	20	19	18	−2	64	1031	4	NE	93.41	0	0
7422	51245	2015	11	6	4	3	20	16	21	−2	64	1030	4	NE	99.22	0	0
7423	51246	2015	11	6	5	3	19	21	14	0	86	1031	2	NE	105.03	0	0
7424	51247	2015	11	6	6	3	16	18	18	0	93	1032	1	NE	110.84	1.8	1.8
7425	51248	2015	11	6	7	3	12	16	13	0	93	1032	1	NE	116.65	1.8	3.6
7426	51249	2015	11	6	8	3	12	17	12	0	93	1032	1	NE	122.46	1.3	4.9
7427	51250	2015	11	6	9	3	15	16	16	0	93	1032	1	NE	127.38	1.9	6.8
7428	51251	2015	11	6	10	3	15	12	12	0	93	1032	1	NE	131.4	1	7.8
7429	51252	2015	11	6	11	3	12	18	17	1	100	1032	1	NW	4.92	0	0
7430	51253	2015	11	6	12	3	17	19	21	1	100	1031	1	NW	9.84	0	0
7431	51254	2015	11	6	13	3	21	19	26	1	100	1030	1	NW	14.76	0	0
7432	51255	2015	11	6	14	3	17	20	23	1	100	1030	1	NW	18.78	0	0
7433	51256	2015	11	6	15	3	18	23	25	1	100	1029	1	NE	4.02	0	0
7434	51257	2015	11	6	16	3	15	16	12	1	93	1029	2	NE	9.83	0	0
7435	51258	2015	11	6	17	3	16	15	12	0	93	1029	2	NE	15.64	0	0
7436	51259	2015	11	6	18	3	13	12	15	0	93	1029	2	NE	20.56	0	0
7437	51260	2015	11	6	19	3	17	16	19	1	93	1029	2	NE	24.58	0	0
7438	51261	2015	11	6	20	3	18	19	19	0	93	1029	2	NE	29.5	0	0
7439	51262	2015	11	6	21	3	13	10	17	0	93	1029	2	NE	35.31	0	0
7440	51263	2015	11	6	22	3	9	12	16	0	93	1029	2	NE	41.12	0	0
7441	51264	2015	11	6	23	3	10	7	14	0	93	1029	2	NE	46.04	0	0
7442	51265	2015	11	7	0	3	11	12	8	−1	93	1029	2	NE	51.85	0	0
7443	51266	2015	11	7	1	3	9	12	12	−1	93	1029	2	NE	57.66	1	1
7444	51267	2015	11	7	2	3	11	17	14	0	93	1029	1	NW	4.92	0.6	1.6
7445	51268	2015	11	7	3	3	15	12	12	0	93	1029	1	NE	4.02	0.6	2.2
7446	51269	2015	11	7	4	3	14	11	8	0	93	1028	1	NE	9.83	0.1	2.3
7447	51270	2015	11	7	5	3	10	10	9	0	100	1028	0	NW	5.81	0	0
7448	51271	2015	11	7	6	3	5	8	7	0	100	1028	0	NE	5.81	0	0
7449	51272	2015	11	7	7	3	8	5	9	−1	86	1028	1	NE	11.62	0	0
7450	51273	2015	11	7	8	3	8	7	14	−1	86	1029	1	NW	4.02	0	0

图 8: 处理后的 2015 的 PM 数据文件