《Python 程序设计》

小作业 2：链家、PM2.5

学院：计算机学院（国家示范性软件学院）

班级：2021211313

姓名：吕子健

学号：2023523012

一、 实验目的

作业 1：爬取并存储链家的新房数据，并进行预处理 （本题 7 分）

（1）爬取起始网页：https://bj.fang.lianjia.com/loupan/

（2）爬取信息的提取及存储要求（单条数据示例在第 4 页） • 信息以 csv 文件存储，应包括以下字段：名称，类别，地理位置（3 个字段分别存储），房 型（只保留最小房型），面积（区间取中值并取整），均价（元，整数），总价（万元， 区间取中值并取整） • 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

（3）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

（4）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）

（5）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

作业 2：分析处理 2015 年北京市 PM2.5 指数数据集空值（本题 3 分）

（1）原始数据集：BeijingPM20100101_20151231.csv（列信息见第 6 页说明）

（2）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

（3）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

（4）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

（5）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

二、 实验内容（代码）

作业 1

```
# 导入所需的库
import requests
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
import time
import matplotlib.pyplot as plt
import os

os.chdir(os.path.dirname(__file__))

plt.rcParams['font.sans-serif']=['SimHei']
plt.rcParams['axes.unicode_minus'] = False
```

```python
# 定义一个函数，用于从字符串中提取数字
def extract_number(s):
    # 如果字符串为空，返回 None
    if not s:
        return None
    # 否则，去掉所有的非数字字符，包括小数点和逗号
    s = s.replace('.', '').replace(',', '')
    # 尝试将字符串转换为整数，如果失败，返回 None
    try:
        return int(s)
    except ValueError:
        return None

# 定义一个函数，用于从面积区间中取中值并取整
def get_area_midpoint(s):
    # 如果字符串为空，返回 None
    if not s:
        return None
    # 否则，去掉所有的非数字字符，包括平方米和逗号
    s = s.replace('平方米', '').replace(',', '').replace('㎡',
'').replace('建面', '')
    # 尝试将字符串分割为两个数字，如果失败，返回 None
    try:
        a, b = map(int, s.split('-'))
    except ValueError:
        return None
    # 返回两个数字的平均值并取整
    return round((a + b) / 2)

# 定义一个函数，用于从总价区间中取中值并取整
def get_price_midpoint(s):
    # 如果字符串为空，返回 None
    if not s:
        return None
    # 否则，去掉所有的非数字字符，包括万元和逗号
    s = s.replace('(万/套)', '').replace('总价', '')
    # 尝试将字符串分割为两个数字，如果失败，返回 None
    if s.count('-') != 0:
        try:
            a, b = map(int, s.split('-'))
        except ValueError:
            return None
        # 返回两个数字的平均值并取整
```

```
        return round((a + b) / 2)
    else:
        return int(s)

# 定义一个空列表，用于存储房屋数据
data = []

# 定义起始网页的 url
url = 'https://bj.fang.lianjia.com/loupan/'
pg=1

# 定义一个循环，用于遍历所有的网页
while url:
    pg += 1
    # 发送请求，获取网页内容
    response = requests.get(url)
    # 解析网页内容，使用 beautiful soup
    soup = BeautifulSoup(response.text, 'html.parser')
    # 找到所有的房屋信息的 div 标签
    divs = soup.find_all('div', class_='resblock-desc-wrapper')
    # 找到当前页面有几个房屋信息
    house_value = soup.find('span', class_='value')
    house_value_num = int(house_value.contents[0])
    if house_value_num == 0:
        break
    # 遍历每个 div 标签，提取房屋信息
    for div in divs:
        # 提取名称，去掉前后空格
        name = div.find('a', class_='name').text.strip()
        # 提取类别，去掉前后空格
        category = div.find('span', class_='resblock-type').text.strip()
        # 提取地理位置，分别存储区域，板块和地址，去掉前后空格
        location = div.find('div', class_='resblock-
location').text.strip()
        location = location.replace('\n/\n', ' ').replace(', ', ',')
        region, block, address = location.split(" ", 2)
        # 提取房型，只保留最小房型，去掉前后空格
        room_type = div.find('a', class_='resblock-room').text.strip()
        room_type = room_type.split('/')[0]
        room_type = room_type.replace('\n', '')
        # 提取面积，区间取中值并取整
        area = div.find('div', class_='resblock-area').text.strip()
        area = get_area_midpoint(area)
        # 提取均价，去掉前后空格
```

```
        unit_price = div.find('span', class_='number').text.strip()
        # 提取总价，区间取中值并取整
        total_price = div.find('div', class_='second')
        if total_price:
            total_price = total_price.text.strip()
        else:
            break
        total_price = get_price_midpoint(total_price)
        # 将房屋信息以字典的形式添加到列表中
        data.append({
            '名称': name,
            '类别': category,
            '区域': region,
            '板块': block,
            '地址': address,
            '房型': room_type,
            '面积': area,
            '均价': unit_price,
            '总价': total_price
        })
    time.sleep(2)
    url = 'https://bj.fang.lianjia.com/loupan/pg' + str(pg) + '/'
    #url = None


# 将列表转换为数据框
df = pd.DataFrame(data)

# 删除面积缺失的房屋数据
df = df.dropna(subset=['面积'])

# 将均价和总价的字符串转换为数字
df['均价'] = df['均价'].apply(extract_number)

# 将数据框保存为 csv 文件
df.to_csv('lianjia.csv', index=False)

# 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
max_total_price = df['总价'].max()
min_total_price = df['总价'].min()
median_total_price = df['总价'].median()
print(f'总价最贵的房子是：{df[df["总价"] == max_total_price]["名称
"].values[0]}，总价为：{max_total_price}万元')
```

```
print(f'总价最便宜的房子是：{df[df["总价"] == min_total_price]["名称
"].values[0]}，总价为：{min_total_price}万元')
print(f'总价的中位数是：{median_total_price}万元')

# 找出均价最贵和最便宜的房子，以及均价的中位数
max_unit_price = df['均价'].max()
min_unit_price = df['均价'].min()
median_unit_price = df['均价'].median()
print(f'均价最贵的房子是：{df[df["均价"] == max_unit_price]["名称
"].values[0]}，均价为：{max_unit_price}元')
print(f'均价最便宜的房子是：{df[df["均价"] == min_unit_price]["名称
"].values[0]}，均价为：{min_unit_price}元')
print(f'均价的中位数是：{median_unit_price}元')

# 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
mean_total_price = df['总价'].mean()
std_total_price = df['总价'].std()
outlier_total_price = df[(df['总价'] > mean_total_price + 3 *
std_total_price) | (df['总价'] < mean_total_price - 3 *
std_total_price)]
print(f'总价在均值三倍标准差以外的房屋有{len(outlier_total_price)}个，它们
的基本信息如下：')
print(outlier_total_price)

# 通过箱型图原则判断并列出均价为异常值的房屋
q1_unit_price = df['均价'].quantile(0.25)
q3_unit_price = df['均价'].quantile(0.75)
iqr_unit_price = q3_unit_price - q1_unit_price
outlier_unit_price = df[(df['均价'] > q3_unit_price + 1.5 *
iqr_unit_price) | (df['均价'] < q1_unit_price - 1.5 * iqr_unit_price)]
print(f'均价为异常值的房屋有{len(outlier_unit_price)}个，它们的基本信息如
下：')
print(outlier_unit_price)

# 离散化处理

bins = np.arange(int(df['均价'].min()/5000)*5000, df['均价'].max() +
10000, 5000)
labels = [f'{a}-{b}' for a, b in zip(bins[:-1], bins[1:])]
print(f'区间为：{labels}')
df['均价区间'] = pd.cut(df['均价'], bins=bins, labels=labels,
right=False)
counts = df['均价区间'].value_counts(sort=False)
```

```
percentages = df['均价区间'].value_counts(normalize=True) * 100

# 将结果保存为 csv 文件
result = pd.DataFrame({'均价区间': labels, '房屋数量': counts, '所占比例':
percentages})
result.to_csv('lianjia_result.csv', index=False)  # 不保存索引列

# 绘制直方图
plt.bar(result['均价区间'], result['房屋数量'])
plt.xticks(rotation=90)  # 旋转 x 轴刻度标签，以避免重叠
plt.xlabel('均价区间')
plt.ylabel('房屋数量')
plt.title('房屋均价离散图')
plt.show()

# 打印结果
print(result.to_string(index=False))
```

作业 2

```
import pandas as pd
import os

os.chdir(os.path.dirname(__file__))

df = pd.read_csv("BeijingPM20100101_20151231.csv")
# 筛选出 2015 年的数据
df_2015 = df[df["year"] == 2015]
# 保存为新的 csv 文件
df_2015.to_csv("BeijingPM2015.csv", index=False)
print(df_2015.isnull().sum())

# 读取新的 csv 文件
df_2015 = pd.read_csv("BeijingPM2015.csv")
# 对 PM_xx 列使用线性插值法填充缺失值
df_2015["PM_Dongsi"] =
df_2015["PM_Dongsi"].interpolate(method="linear")
df_2015["PM_Dongsihuan"] =
df_2015["PM_Dongsihuan"].interpolate(method="linear")
df_2015["PM_Nongzhanguan"] =
df_2015["PM_Nongzhanguan"].interpolate(method="linear")
df_2015["PM_US Post"] = df_2015["PM_US
Post"].interpolate(method="linear")
# 对 DEWP 和 TEMP 列使用均值填充缺失值
df_2015["DEWP"] = df_2015["DEWP"].fillna(df_2015["DEWP"].mean())
```

```
df_2015["TEMP"] = df_2015["TEMP"].fillna(df_2015["TEMP"].mean())
# 对 HUMI、PRES、cbwd 和 Iws 列使用相邻位置的数据填充缺失值
df_2015["HUMI"] = df_2015["HUMI"].fillna(df_2015["PM_Nongzhanguan"])
df_2015["PRES"] = df_2015["PRES"].fillna(df_2015["PM_Nongzhanguan"])
df_2015["cbwd"] = df_2015["cbwd"].fillna(df_2015["PM_Nongzhanguan"])
df_2015["Iws"] = df_2015["Iws"].fillna(df_2015["PM_Nongzhanguan"])
# 对 precipitation 和 Iprec 列使用 0 填充缺失值
df_2015["precipitation"] = df_2015["precipitation"].fillna(0)
df_2015["Iprec"] = df_2015["Iprec"].fillna(0)
# 保存为新的 csv 文件
df_2015.to_csv("BeijingPM2015_clean.csv", index=False)
print(df_2015.isnull().sum())
```

三、 实验结果及其分析

作业 1

爬出数据的 csv 文件:

linehome_chuli.py | Preview 'lianjia_result.csv' | Preview 'lianjia.csv' | test.py | get_linehome.py | Preview 'lianjia_result.csv'

名称	类别	区域	板块	地址	房型	面积	均价	总价
水岸壹号	别墅	房山	良乡	良乡大学城西站地铁南侧800米，邻濯河旁	3室	192	50000	1200
尚峯壹號	商业类	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	234	27000	1050
运河铭著	住宅	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	3室	153	49000	778
万年广阳那九号	住宅	房山	长阳	长阳清苑南街与汇商东路交汇处西北角	4室	197	50000	1000
首开璞堤隐园	住宅	丰台	方庄	紫芳西五区	4室	203	110000	2175
御汤山熙园	别墅	昌平	昌平其它	北京市昌平区小汤山镇顺沙路99号院	4室	418	40000	5402
天资华府	住宅	房山	长阳	房山区CSD政务大厅5号门	3室	155	38000	630
檀香府	别墅	门头沟	门头沟其它	京澄大街与澄拓十街交叉口	3室	264	42000	1250
韩建·观山源墅	别墅	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	310	40000	1250
北辰墅院1900	别墅	顺义	马坡	顺兴街11号院望尊园	4室	266	35000	1203
泰禾西府大院	住宅	丰台	丽泽	西三环丽泽桥西北角，玉璞公园东侧	5室	216	150000	3800
绿地海珀云翡	住宅	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	143	65000	900
燕西华府	住宅	丰台	丰台其它	王佐镇青龙湖公园东1500米	3室	174	50000	520
水岸壹号	住宅	房山	良乡	良乡大学城西站地铁南侧800米，邻濯河旁	3室	138	48000	606
天恒摩墅	别墅	房山	房山其它	周口店镇政府东200米	3室	150	23000	390
鲁能·格拉斯小镇	别墅	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	349	62000	2032
兴创荣墅	别墅	大兴	大兴新机场洋房	北京市大兴区育胜街	3室	380	25000	1115
温哥华森林	别墅	昌平	北七家	北五环外紧邻立汤路，北七家建材城向北第一个路口200米路东,枫树家园6区,枫树家园五区	5室	464	46000	2731
丽都壹号	住宅	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	4室	124	90000	1674
北辰墅院1900	住宅	顺义	马坡	顺兴街11号院望尊园	3室	109	36000	430
燕西华府	别墅	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	600	47000	2450
京西悦府	住宅	房山	阎村	燕房线阎村地铁站东南角约189米		128	33000	440
福景苑	住宅	朝阳	燕莎	亮马桥路46号	1室	206	83000	1775
合景寰汇公馆	住宅	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	97	35000	385
K2十里春风	住宅	通州	通州其它	北京市通州区	2室	82	23500	200
K2十里春风	别墅	通州	通州其它	北京市通州区	3室	156	28000	450
玺萌壹號院	别墅	丰台	草桥	西南三环嘉园路与镇国寺北街交叉口	5室	392	90000	3795
北京书院	住宅	朝阳	惠新西街	北京市朝阳区北土城东路辅路	1室	109	155000	1066
中铁华侨城和园	住宅	大兴	瀛海	南五环南海子公园西侧约500米	3室	169	60000	955
顺鑫颐和天璟	住宅	顺义	顺义其它	北京市顺义区牛栏山镇牛富路顺鑫颐和天璟临润售楼中心	4室	165	28000	410
顺鑫颐和天璟	别墅	顺义	顺义其它	新城右堤路与昌金路交汇处向北200米	4室	382	28000	1075
北京城建北京合	住宅	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	112	46000	561
复地运河公馆	住宅	通州	武夷花园	通州运河核心区临滨河西路	2室	117	43000	550
北京城建北京合	别墅	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	270	39000	1150

linehome_chuli.py | Preview 'lianjia_result.csv' | Preview 'lianjia.csv' | test.py | get_linehome.py | Preview 'lianjia_result.csv'

名称	类别	区域	板块	地址	房型	面积	均价	总价
中铁诺德阅墅	别墅	顺义	中央别墅区	顺义区后沙峪镇裕园路762乡龙湖滟澜山对面	4室	278	50000	1425
中铁华侨城和园	别墅	大兴	瀛海	南五环南海子公园西侧约500米	4室	329	50000	1870
懋源·璟岳	别墅	丰台	玉泉营	南三环西路99号院	4室	528	140000	7750
合景泰富天汇	住宅	顺义	马坡	顺义区昌金路与通顺路交汇处	2室	94	33000	310
懋源·璟玺	别墅	朝阳	中央别墅区	孙河京密路与京平辅路交叉口西行1000米	5室	608	100000	5579
万科雲庐	住宅	丰台	丰台其它	魏各庄路万科雲庐。售楼处位置在山湖路与泉湖西湖交叉位置	4室	202	39000	738
万科雲庐	别墅	丰台	丰台其它	魏各庄路万科雲庐	4室	265	30000	901
金茂北京国际社	住宅	顺义	顺义其它	顺义新城北小营昌金路水色时光路西	1室	84	30000	260
住总如院	住宅	大兴	大兴新机场洋房	北京市大兴区采华路(波尔多小镇南区西南侧约250米)	2室	166	31136	378
郎府书苑	住宅	通州	通州其它	西集镇京哈高速郎府出口南侧300米	3室	102	25800	286
和悦春风	住宅	大兴	大兴新机场	北京市大兴区隆延街与甜园路交叉路口往西北约150米	2室	92	38000	360
阳光城溪山悦	别墅	密云	密云其它	密溪路33号	2室	214	20000	410
金禧璞瑅	住宅	顺义	顺义其它	高酒路四村段23号	4室	150	39000	612
金禧璞瑅	别墅	顺义	顺义其它	高酒路四村段23号	5室	375	25000	1200
中骏云景台	住宅	房山	房山其它	官道西大街与良常路交叉口	2室	95	28000	235
禧瑞金海	住宅	平谷	平谷其它	平蓟路与环镇东路交汇	2室	98	22000	200
禧瑞金海	别墅	平谷	平谷其它	平蓟路与环镇东路交汇	2室	162	22000	342
电建·洺悦湾	住宅	大兴	旧宫	旧宫北路旧宫地铁站东侧300米	3室	104	60186	700
北科建·翡翠华府	住宅	怀柔	怀柔	中高路与乐园大街交叉口北150米	1室	90	36800	412
保利绿城·和璟诚	住宅	大兴	瀛海	8号线南段地铁站东1.6公里	4室	141	65000	1010
北京庄园	别墅	顺义	顺义其它	京承高速第11出口往东800米	4室	1195	115000	12950
中海甲叁號院	住宅	丰台	玉泉营	丰台恒丰路	4室	172	125000	2500
复地通运府	酒店式公寓	通州	万达	北京市通州区北大街148号	2室	114	35000	435
住总兴创如遇	住宅	大兴	大兴新机场洋房	采华路与育镇街交叉路口往西北约100米(采育文化广场西南侧约100米)	1室	84	28000	202
悠唐麒麟公馆	住宅	朝阳	朝阳门外	北京市朝阳三丰北里	1室	96	120000	1075
中海寰宇视界	住宅	房山	长阳	北京市房山区轨道交通房山线稻田地铁站南侧200米处	1室	166	55000	600
国祥府	住宅	密云	鼓楼街道	檀西路与新北路交汇处向北约400米	2室	101	28000	292
北京天誉	住宅	丰台	十里河	北京市丰台区小红门路312号	4室	185	120000	2575
航城壹号	住宅	大兴	大兴新机场	大广高速瓜乡桥西北300米航城壹号	2室	138	32000	435
中建·京西印玥	住宅	房山	良乡	阳光北大街中建·京西印玥	3室	90	43000	400
长城璞院	别墅	怀柔	怀柔其它	距离怀柔区黄花城水长城东侧1.5公里	5室	215	20000	574
玖瀛府	别墅	昌平	南部	南金路7号院	3室	257	48000	1425
国誉万和城	住宅	丰台	南苑	西四环14号线大瓦窑站北300m	4室	160	82000	1215

（部分）

数据统计：

```
PS G:\作业\BUPT-python> python -u "g:\作业\BUPT-python\smallwork\linehome_chuli.py"
总价最贵的房子是：北京庄园，总价为：12950万元
总价最便宜的房子是：汇豪公园里，总价为：180万元
总价的中位数是：755.0万元
均价最贵的房子是：北京书院，均价为：155000元
均价最便宜的房子是：和棠瑞著，均价为：16000元
均价的中位数是：50000.0元
```

异常值处理：

```
总价在均值三倍标准差以外的房屋有4个，它们的基本信息如下：
      名称  类别  区域    板块                地址  房型     面积     均价     总价
5   御汤山熙园  别墅  昌平   昌平其它     北京市昌平区小汤山镇顺沙路99号院  4室   418.0   40000   5402
61  懋源·璟岳  别墅  丰台   玉泉营          南三环西路99号院  4室   528.0  140000   7750
63  懋源·璟玺  别墅  朝阳  中央别墅区  孙河京密路与京平辅路交叉口西行1000米  5室   608.0  100000   5579
80   北京庄园  别墅  顺义   顺义其它      京承高速第11出口往东800米  4室  1195.0  115000  12950
均价为异常值的房屋有11个，它们的基本信息如下：
      名称  类别  区域    板块                地址  房型     面积     均价    总价
10  泰禾西府大院  住宅  丰台    丽泽          西三环丽泽桥西北角，玉璞公园东侧  5室  216.0  150000  3800
28    北京书院  住宅  朝阳   惠新西街          北京市朝阳区北土城东路辅路  1室  109.0  155000  1066
46    尊悦光华  住宅  朝阳    CBD      北京市朝阳区光华东里甲1号院3号楼  3室  152.0  150000  2500
56  葛洲坝中国府  住宅  丰台   玉泉营          北京市丰台东路46号  3室  204.0  125000  2600
61   懋源·璟岳  别墅  丰台   玉泉营          南三环西路99号院  4室  528.0  140000  7750
81  中海甲叁號院  住宅  丰台   玉泉营             丰台恒丰路  4室  172.0  125000  2500
84  悠唐麒麟公馆  住宅  朝阳  朝阳门外          北京市朝阳三丰北里  1室   96.0  120000  1075
87    北京天誉  住宅  丰台   十里河       北京市丰台区小红门路312号  4室  185.0  120000  2575
94  葛洲坝中国府  别墅  丰台   玉泉营             丰台东路46号  4室  470.0  125000  3750
155   懋源璟廷  住宅  丰台   七里庄   北京市丰台区望园东里西北（望园北路北）  3室  214.0  130000  2439
156  懋源煜泽台  住宅  丰台   七里庄  北京市丰台区七里庄中国农业银行（北京丰台支行）南  3室  164.0  120000  2050
```

总价异常共性分析：

1. 总面积大于平均，属于大面积住宅。
2. 都是别墅，房型比较豪华，整体价值高。
3. 异常值只有高于均值没有低于均值，分析为新房中小面积房子较少。

均价异常共性分析：

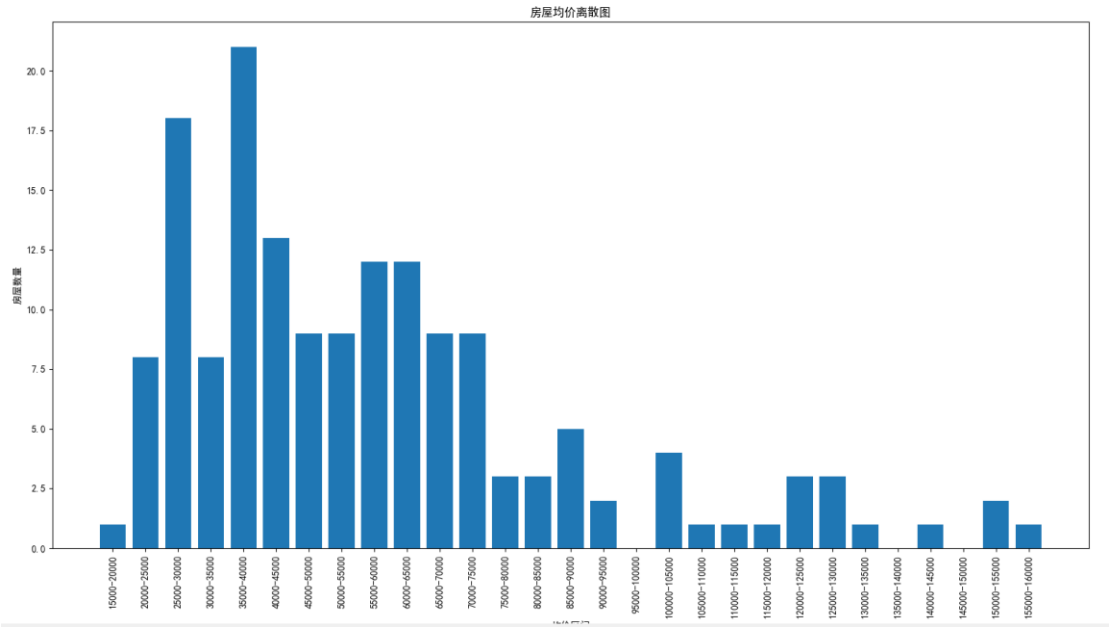
1. 异常值只有高于均值没有低于均值。

2. 仅位于丰台和朝阳两地，地理位置优越，配套资源齐全，且作为后发展辖区相较于类似老辖区有更多空地开发新房。

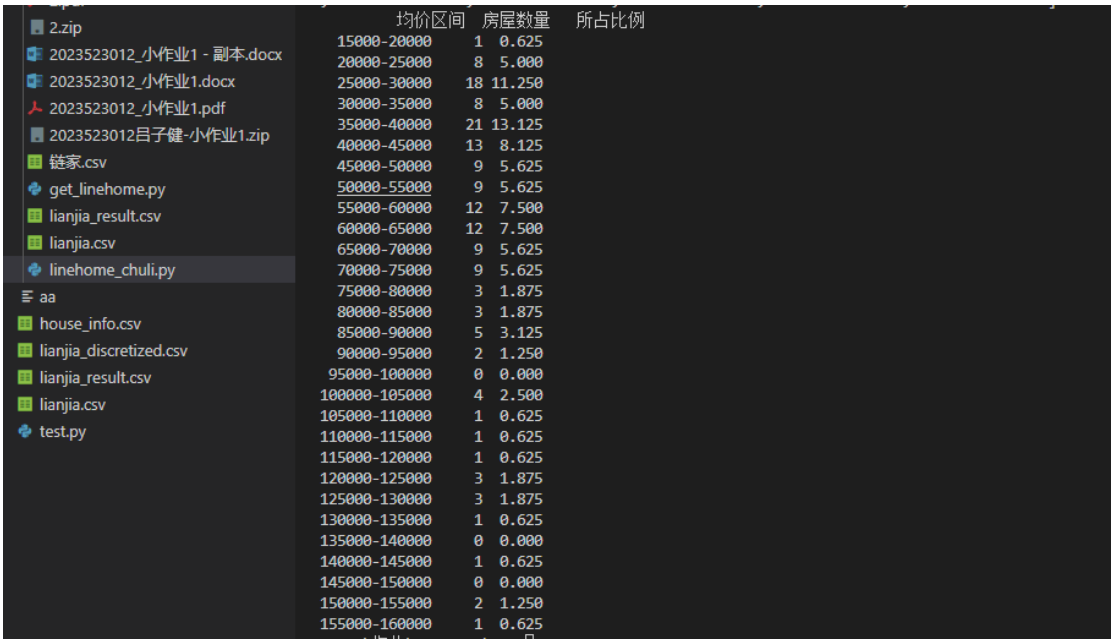
离散化处理:

我选择每个区间的长度为 5000 元, 因为这样可以保证区间的数量不会太多或太少，也可以反映均价的分布情况。

matplotlib.pyplot 输出柱状图:

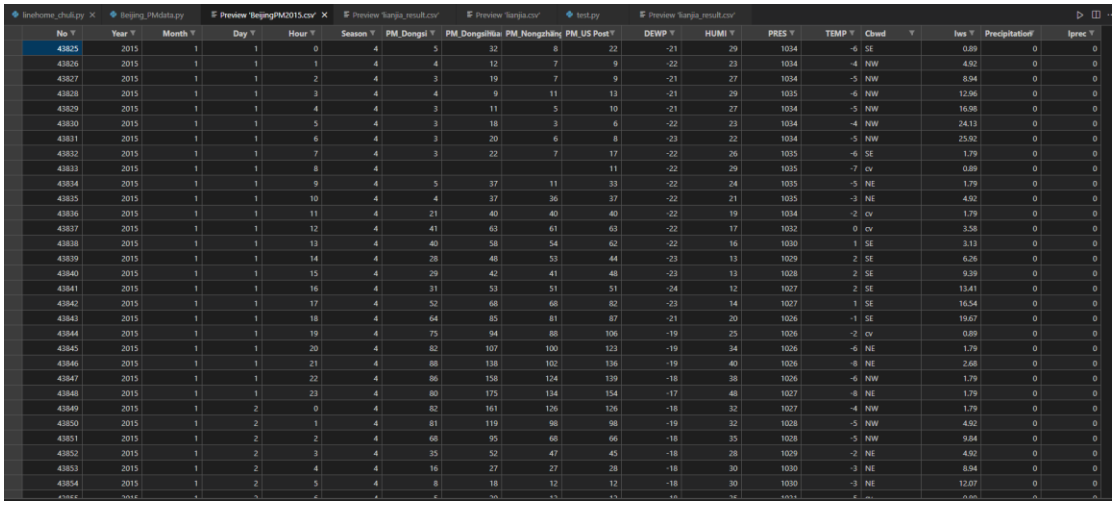


终端中显示比例:

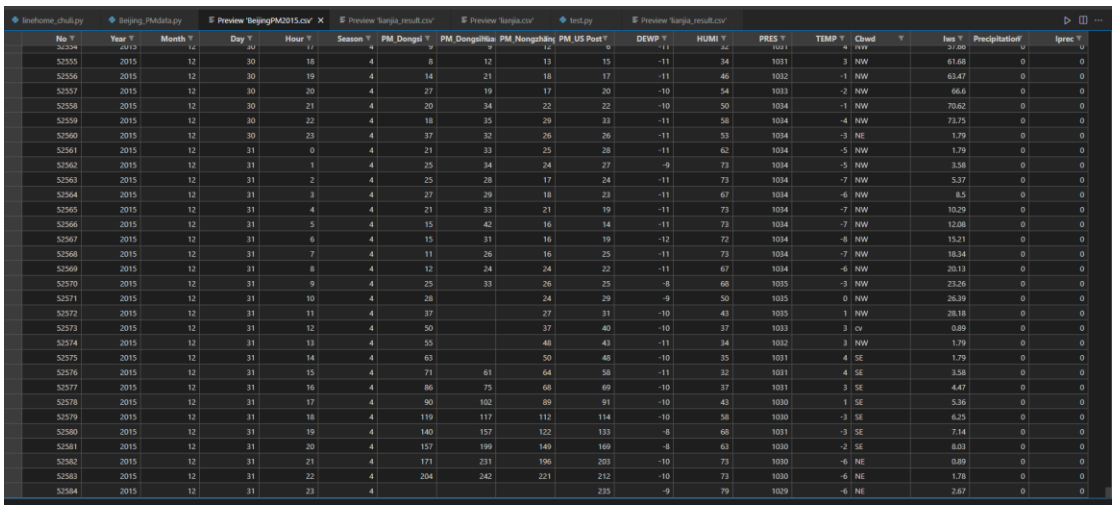


作业 2

1. 从原始数据集中抽取 2015 年度数据，存储为新的 csv 文件：

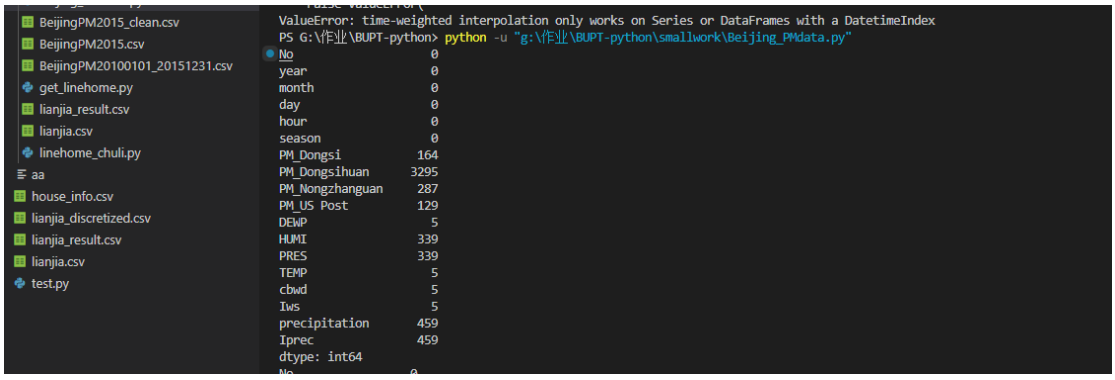


No	Year	Month	Day	Hour	Season	PM_Dongsi	PM_Dongsihuan	PM_Nongzhanguan	PM_US Post	DEWP	HUMI	PRES	TEMP	Cbwd	Iws	Precipitation	Iprec
43825	2015	1	1	0	4	5	32	8	22	-21	29	1034	-6	SE	0.89	0	0
43826	2015	1	1	1	4	4	12	7	9	-22	23	1034	-4	NW	4.92	0	0
43827	2015	1	1	2	4	3	19	7	9	-21	27	1034	-5	NW	8.94	0	0
43828	2015	1	1	3	4	4	9	11	13	-21	29	1035	-6	NW	12.96	0	0
43829	2015	1	1	4	4	3	11	5	10	-21	27	1034	-5	NW	16.98	0	0
43830	2015	1	1	5	4	3	18	3	6	-22	23	1034	-4	NW	24.13	0	0
43831	2015	1	1	6	4	3	20	6	8	-23	22	1034	-5	NW	25.92	0	0
43832	2015	1	1	7	4	3	22	7	17	-22	26	1035	-6	SE	1.79	0	0
43833	2015	1	1	8	4				11	-22	29	1035	-7	cv	0.89	0	0
43834	2015	1	1	9	4	5	37	11	33	-22	24	1035	-5	NE	1.79	0	0
43835	2015	1	1	10	4	4	37	36	37	-22	21	1035	-3	NE	4.92	0	0
43836	2015	1	1	11	4	21	40	40	40	-22	19	1034	-2	cv	1.79	0	0
43837	2015	1	1	12	4	41	63	61	63	-22	17	1032	0	cv	3.58	0	0
43838	2015	1	1	13	4	40	58	54	62	-22	16	1030	1	SE	3.13	0	0
43839	2015	1	1	14	4	28	48	53	44	-23	13	1029	2	SE	6.26	0	0
43840	2015	1	1	15	4	29	42	41	48	-23	13	1028	2	SE	9.39	0	0
43841	2015	1	1	16	4	31	53	51	51	-24	12	1027	2	SE	13.41	0	0
43842	2015	1	1	17	4	52	68	68	82	-23	14	1027	1	SE	16.54	0	0
43843	2015	1	1	18	4	64	85	81	87	-21	20	1026	-1	SE	19.67	0	0
43844	2015	1	1	19	4	75	94	88	106	-19	25	1026	-2	cv	0.89	0	0
43845	2015	1	1	20	4	82	107	100	123	-19	34	1026	-6	NE	1.79	0	0
43846	2015	1	1	21	4	88	138	102	136	-19	40	1026	-8	NE	2.68	0	0
43847	2015	1	1	22	4	86	158	124	139	-18	38	1026	-6	NW	1.79	0	0
43848	2015	1	1	23	4	80	175	134	154	-17	48	1027	-8	NE	1.79	0	0
43849	2015	1	2	0	4	82	161	126	126	-18	32	1027	-4	NW	1.79	0	0
43850	2015	1	2	1	4	81	119	98	98	-19	32	1028	-5	NW	4.92	0	0
43851	2015	1	2	2	4	68	95	68	66	-18	35	1028	-5	NW	9.84	0	0
43852	2015	1	2	3	4	35	52	47	45	-18	28	1029	-2	NE	4.92	0	0
43853	2015	1	2	4	4	16	27	27	28	-18	30	1030	-3	NE	8.94	0	0
43854	2015	1	2	5	4	8	18	12	12	-18	30	1030	-3	NE	12.07	0	0



No	Year	Month	Day	Hour	Season	PM_Dongsi	PM_Dongsihuan	PM_Nongzhanguan	PM_US Post	DEWP	HUMI	PRES	TEMP	Cbwd	Iws	Precipitation	Iprec
52555	2015	12	30	18	4	8	12	13	15	-11	34	1031	3	NW	61.68	0	0
52556	2015	12	30	19	4	14	21	18	17	-11	46	1032	-1	NW	63.47	0	0
52557	2015	12	30	20	4	27	19	17	20	-10	54	1033	-2	NW	66.6	0	0
52558	2015	12	30	21	4	20	34	22	22	-10	50	1034	-1	NW	70.62	0	0
52559	2015	12	30	22	4	18	35	29	33	-11	58	1034	-4	NW	73.75	0	0
52560	2015	12	30	23	4	37	32	26	26	-11	53	1034	-3	NE	1.79	0	0
52561	2015	12	31	0	4	21	33	25	28	-11	62	1034	-5	NW	1.79	0	0
52562	2015	12	31	1	4	25	34	24	27	-9	73	1034	-5	NW	3.58	0	0
52563	2015	12	31	2	4	25	28	17	24	-11	73	1034	-7	NW	5.37	0	0
52564	2015	12	31	3	4	27	29	18	23	-11	67	1034	-6	NW	8.5	0	0
52565	2015	12	31	4	4	21	33	21	19	-11	73	1034	-7	NW	10.29	0	0
52566	2015	12	31	5	4	15	42	16	14	-11	73	1034	-7	NW	12.08	0	0
52567	2015	12	31	6	4	15	31	16	19	-12	72	1034	-8	NW	15.21	0	0
52568	2015	12	31	7	4	11	26	16	25	-11	73	1034	-7	NW	18.34	0	0
52569	2015	12	31	8	4	12	24	24	22	-11	67	1034	-6	NW	20.13	0	0
52570	2015	12	31	9	4	25	33	26	25	-8	68	1035	-3	NW	23.26	0	0
52571	2015	12	31	10	4	28		24	29	-9	50	1035	0	NW	26.39	0	0
52572	2015	12	31	11	4	37		27	31	-10	43	1035	1	NW	28.18	0	0
52573	2015	12	31	12	4	50		37	40	-10	37	1033	3	cv	0.89	0	0
52574	2015	12	31	13	4	55		48	43	-11	34	1032	3	NW	1.79	0	0
52575	2015	12	31	14	4	63		50	48	-10	35	1031	4	SE	1.79	0	0
52576	2015	12	31	15	4	71	61	64	58	-11	32	1031	4	SE	3.58	0	0
52577	2015	12	31	16	4	86	75	68	69	-10	37	1031	3	SE	4.47	0	0
52578	2015	12	31	17	4	90	102	89	91	-10	43	1030	1	SE	5.36	0	0
52579	2015	12	31	18	4	119	117	112	114	-10	58	1030	-3	SE	6.25	0	0
52580	2015	12	31	19	4	140	157	122	133	-8	68	1031	-3	SE	7.14	0	0
52581	2015	12	31	20	4	157	199	149	169	-8	63	1030	-2	SE	8.03	0	0
52582	2015	12	31	21	4	171	231	196	203	-10	73	1030	-6	NE	0.89	0	0
52583	2015	12	31	22	4	204	242	221	212	-10	73	1030	-6	NE	1.78	0	0
52584	2015	12	31	23	4				235	-9	79	1029	-6	NE	2.67	0	0

2. 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：



3. 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：

对于 PM2.5 浓度的四个列（PM_Dongsi, PM_Dongsihuan, PM_Nongzhanguan, PM_US Post）它们是目标变量，不能随意填充或删除，所以我使用插值法来估计缺失值，例如使用线性插值或三次样条插值。这样可以保持数据的连续性和平滑性，避免引入额外的误差。

对于 DEWP 和 TEMP 列，由于它们只有很少的空值，使用均值或中位数来

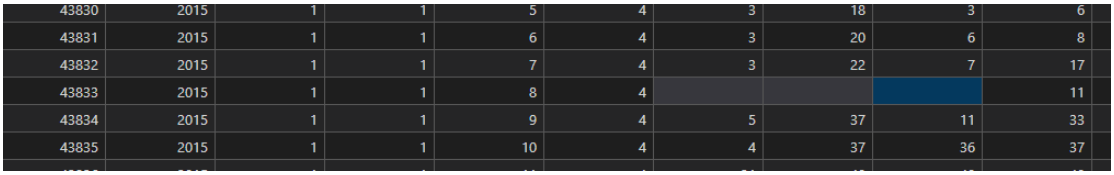
填充缺失值，这样可以保持数据的一致性和稳定性，避免影响数据的分布和统计特征。

对于 HUMI、PRES、cbwd 和 Iws 列，由于它们的空值数量相同，且与气象站的位置有关，使用相邻位置的数据来填充缺失值，例如使用 PM_Nongzhanguan 的数据来填充 PM_Dongsihuan 的数据。这样可以利用数据的空间相关性，避免引入不合理的数据。

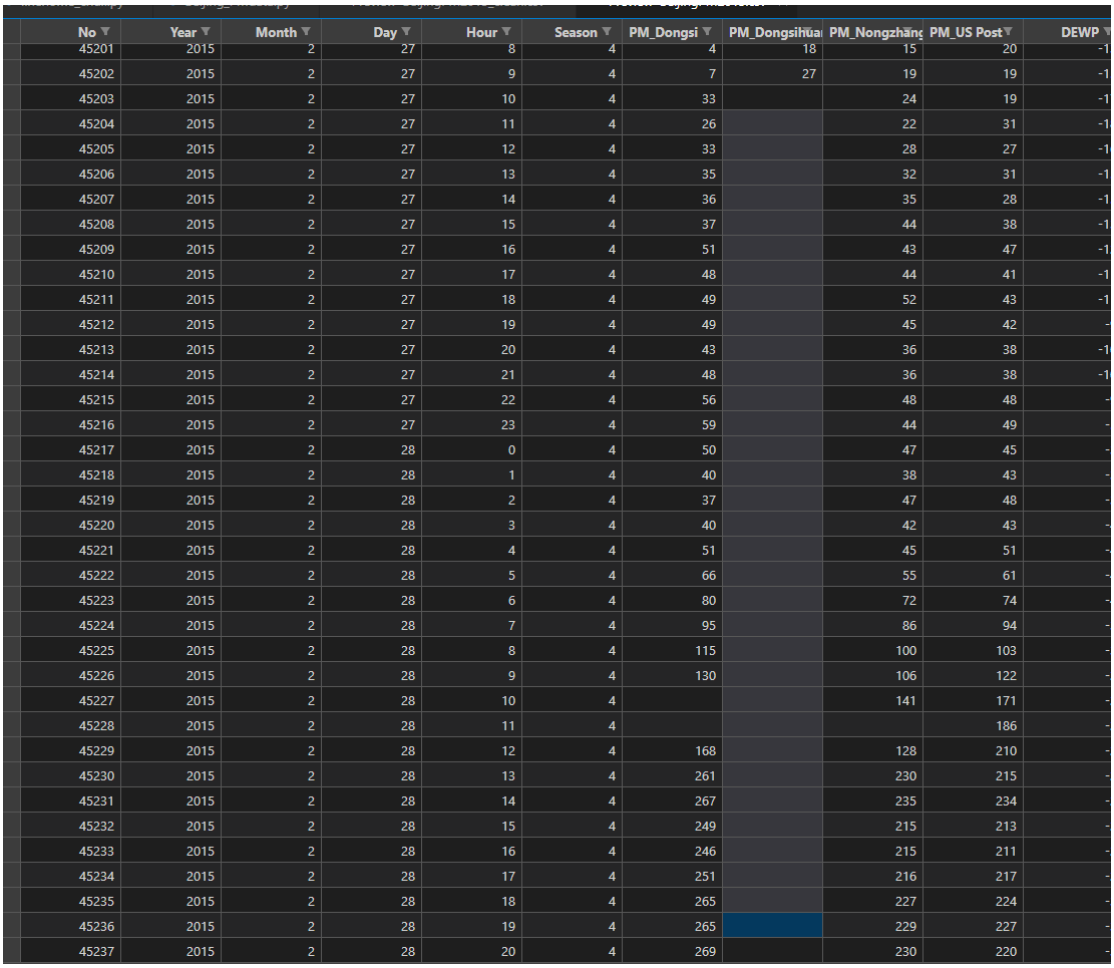
对于 precipitation 和 Iprec 列，由于它们的空值数量相同，且与降水量有关，我建议使用 0 来填充缺失值，这样可以假设缺失值代表没有降水，避免影响数据的总和和比例。

4. 处理结果:

原:



43830	2015	1	1	5	4	3	18	3	6
43831	2015	1	1	6	4	3	20	6	8
43832	2015	1	1	7	4	3	22	7	17
43833	2015	1	1	8	4				11
43834	2015	1	1	9	4	5	37	11	33
43835	2015	1	1	10	4	4	37	36	37



No	Year	Month	Day	Hour	Season	PM_Dongsi	PM_Dongsihua	PM_Nongzhang	PM_US Post	DEWP
45201	2015	2	27	8	4	4	18	15	20	-1
45202	2015	2	27	9	4	7	27	19	19	-1
45203	2015	2	27	10	4	33		24	19	-1
45204	2015	2	27	11	4	26		22	31	-1
45205	2015	2	27	12	4	33		28	27	-1
45206	2015	2	27	13	4	35		32	31	-1
45207	2015	2	27	14	4	36		35	28	-1
45208	2015	2	27	15	4	37		44	38	-1
45209	2015	2	27	16	4	51		43	47	-1
45210	2015	2	27	17	4	48		44	41	-1
45211	2015	2	27	18	4	49		52	43	-1
45212	2015	2	27	19	4	49		45	42	-
45213	2015	2	27	20	4	43		36	38	-1
45214	2015	2	27	21	4	48		36	38	-1
45215	2015	2	27	22	4	56		48	48	-
45216	2015	2	27	23	4	59		44	49	-
45217	2015	2	28	0	4	50		47	45	-
45218	2015	2	28	1	4	40		38	43	-
45219	2015	2	28	2	4	37		47	48	-
45220	2015	2	28	3	4	40		42	43	-
45221	2015	2	28	4	4	51		45	51	-
45222	2015	2	28	5	4	66		55	61	-
45223	2015	2	28	6	4	80		72	74	-
45224	2015	2	28	7	4	95		86	94	-
45225	2015	2	28	8	4	115		100	103	-
45226	2015	2	28	9	4	130		106	122	-
45227	2015	2	28	10	4			141	171	-
45228	2015	2	28	11	4				186	-
45229	2015	2	28	12	4	168		128	210	-
45230	2015	2	28	13	4	261		230	215	-
45231	2015	2	28	14	4	267		235	234	-
45232	2015	2	28	15	4	249		215	213	-
45233	2015	2	28	16	4	246		215	211	-
45234	2015	2	28	17	4	251		216	217	-
45235	2015	2	28	18	4	265		227	224	-
45236	2015	2	28	19	4	265		229	227	-
45237	2015	2	28	20	4	269		230	220	-

Month	Day	Hour	Season	PM_Dongsi	PM_Dongsihuan	PM_Nongzhang	PM_US Post	DEWP	HUMI	PRES	TEMP	Cbwd	Iws
3	27	8	1	98		218	97	0	49	1024	10	SE	11.63
3	27	9	1	100	112		95	0	40	1024	13	SE	13.42
3	27	10	1	113	122	276	106	1	41	1023	14	cv	0.89
3	27	11	1	117	134		118	0	35	1023	15	cv	1.78
3	27	12	1	118	132		115	0	31	1021	17	SE	4.02
3	27	13	1	116	123		109	1	31	1020	18	SE	7.15
3	27	14	1	112	117			1	28	1018	20	SE	11.17
3	27	15	1	118	123			2	30	1017	20	SE	15.19
3	27	16	1	131	133			2	30	1016	20	SE	20.11
3	27	17	1	142	150			2	30	1015	20	SE	24.13
3	27	18	1	152	155			3	34	1015	19	SE	27.26
3	27	19	1	155	163			2	38	1015	16	SE	30.39
3	27	20	1	137	142	152		3	39	1015	17	SE	32.18
3	27	21	1	110	114	120		3	41	1015	16	SE	36.2
3	27	22	1	99	105	102		3	39	1014	17	SE	42.01
3	27	23	1	88	87	89		2	36	1014	17	SE	47.82
3	28	0	1	89	97	98		2	41	1013	15	SE	51.84
3	28	1	1	95	99	101		3	47	1013	14	SE	53.63
3	28	2	1	98	106	96		4	50	1013	14	SE	54.52
3	28	3	1	112	110	113		4	54	1012	13	SE	56.31
3	28	4	1	113	115	115		4	57	1012	12	cv	0.89
3	28	5	1	114	113	122		3	61	1012	10	NW	0.89
3	28	6	1	80	93	95		3	66	1012	9	NW	2.68
3	28	7	1	27	43	34		0	35	1012	15	NW	9.83
3	28	8	1	22	37	28		-2	27	1013	17	NW	15.64
3	28	9	1	24	41	34		-2	25	1013	18	NW	23.69
3	28	10	1	51	86	61		-2	24	1013	19	NW	31.74
3	28	11	1	83	142	116		-1	24	1013	20	NW	40.68
3	28	12	1	84	156	120		0	24	1012	21	NW	50.51
3	28	13	1	77	132	99		-1	20	1012	23	NW	60.34
3	28	14	1	51	87	66		0	20	1011	24	NW	69.28
3	28	15	1	22	40	27		0	19	1011	25	NW	76.43
3	28	16	1	25	27	23		-2	18	1011	23	NW	80.45
3	28	17	1	23	29	25		-1	19	1012	24	NW	88.5

后:

43831	2015	1	1	6	4	3	20	6	8
43832	2015	1	1	7	4	3	22	7	17
43833	2015	1	1	8	4	4	29.5	9	11
43834	2015	1	1	9	4	5	37	11	33
43835	2015	1	1	10	4	4	37	36	37
43836	2015	1	1	11	4	21	40	40	40

No	Year	Month	Day	Hour	Season	PM_Dongsi	PM_Dongsihuan	PM_Nongzhang	PM_US Post
45201	2015	2	27	8	4	4	18	15	20
45202	2015	2	27	9	4	7	27	19	19
45203	2015	2	27	10	4	33	26.87	24	19
45204	2015	2	27	11	4	26	26.74	22	31
45205	2015	2	27	12	4	33	26.61	28	27
45206	2015	2	27	13	4	35	26.48	32	31
45207	2015	2	27	14	4	36	26.35	35	28
45208	2015	2	27	15	4	37	26.22	44	38
45209	2015	2	27	16	4	51	26.09	43	47
45210	2015	2	27	17	4	48	25.96	44	41
45211	2015	2	27	18	4	49	25.83	52	43
45212	2015	2	27	19	4	49	25.7	45	42
45213	2015	2	27	20	4	43	25.57	36	38
45214	2015	2	27	21	4	48	25.44	36	38
45215	2015	2	27	22	4	56	25.31	48	48
45216	2015	2	27	23	4	59	25.19	44	49
45217	2015	2	28	0	4	50	25.06	47	45
45218	2015	2	28	1	4	40	24.93	38	43
45219	2015	2	28	2	4	37	24.8	47	48
45220	2015	2	28	3	4	40	24.67	42	43
45221	2015	2	28	4	4	51	24.54	45	51
45222	2015	2	28	5	4	66	24.41	55	61
45223	2015	2	28	6	4	80	24.28	72	74
45224	2015	2	28	7	4	95	24.15	86	94
45225	2015	2	28	8	4	115	24.02	100	103
45226	2015	2	28	9	4	130	23.89	106	122
45227	2015	2	28	10	4	142.67	23.76	141	171
45228	2015	2	28	11	4	155.33	23.63	134.5	186
45229	2015	2	28	12	4	168	23.5	128	210
45230	2015	2	28	13	4	261	23.37	230	215
45231	2015	2	28	14	4	267	23.24	235	234
45232	2015	2	28	15	4	249	23.11	215	213
45233	2015	2	28	16	4	246	22.98	215	211
45234	2015	2	28	17	4	251	22.85	216	217
45235	2015	2	28	18	4	265	22.72	227	224

Month	Day	Hour	Season	PM_Dongsi	PM_Dongsihuai	PM_Nongzhang	PM_US Post	DEWP	HUMI	PRES
3	27	8	1	98	106	218	97	0	49	1024
3	27	9	1	100	112	247	95	0	40	1024
3	27	10	1	113	122	276	106	1	41	1023
3	27	11	1	117	134	263.6	118	0	35	1023
3	27	12	1	118	132	251.2	115	0	31	1021
3	27	13	1	116	123	238.8	109	1	31	1020
3	27	14	1	112	117	226.4	106.07	1	28	1018
3	27	15	1	118	123	214	103.14	2	30	1017
3	27	16	1	131	133	201.6	100.21	2	30	1016
3	27	17	1	142	150	189.2	97.28	2	30	1015
3	27	18	1	152	155	176.8	94.34	3	34	1015
3	27	19	1	155	163	164.4	91.41	2	38	1015
3	27	20	1	137	142	152	88.48	3	39	1015
3	27	21	1	110	114	120	85.55	3	41	1015
3	27	22	1	99	105	102	82.62	3	39	1014
3	27	23	1	88	87	89	79.69	2	36	1014
3	28	0	1	89	97	98	76.76	2	41	1013
3	28	1	1	95	99	101	73.83	3	47	1013
3	28	2	1	98	106	96	70.9	4	50	1013
3	28	3	1	112	110	113	67.97	4	54	1012
3	28	4	1	113	115	115	65.03	4	57	1012
3	28	5	1	114	113	122	62.1	3	61	1012
3	28	6	1	80	93	95	59.17	3	66	1012
3	28	7	1	27	43	34	56.24	0	35	1012
3	28	8	1	22	37	28	53.31	-2	27	1013
3	28	9	1	24	41	34	50.38	-2	25	1013
3	28	10	1	51	86	61	47.45	-2	24	1013
3	28	11	1	83	142	116	44.52	-1	24	1013
3	28	12	1	84	156	120	41.59	0	24	1012
3	28	13	1	77	132	99	38.66	-1	20	1012
3	28	14	1	51	87	66	35.72	0	20	1011
3	28	15	1	22	40	27	32.79	0	19	1011
3	28	16	1	25	27	23	29.86	-2	18	1011
3	28	17	1	23	29	25	26.93	-1	19	1012
3	28	18	1	27	30	30	24	-1	20	1012

查询结果：