大作业

贺建坤

2020211277

大数据专业

实验时间2023,12,30

目录

1数据获取

2数据预处理

3数据分析--统计性分析和可视化分析

4.数据分析--分析影响房价的主要因素有哪些

5数据分析--分析二手房价格的地理分布特征

6 选做题：

1数据获取

爬取方法

1.网页分析

对链家北京二手房信息进行爬取

[https://bj.lianjia.com/ershoufang/](https://bj.lianjia.com/ershoufang/dongcheng/)

点进去看具体信息发现



<https://bj.lianjia.com/ershoufang/101121960406.html>

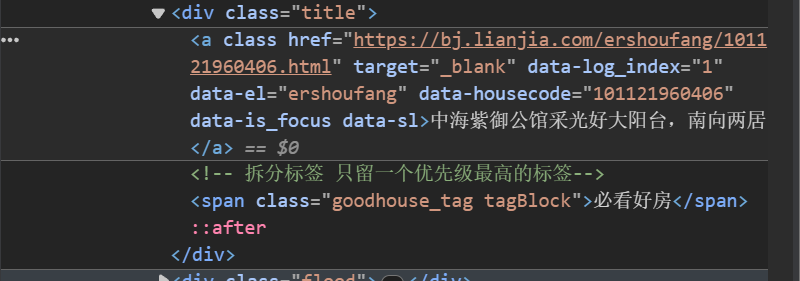
<https://bj.lianjia.com/ershoufang/101122021423.html>

<https://bj.lianjia.com/ershoufang/101122021175.html>

具体信息的网址都是类似这样的链接,

<https://bj.lianjia.com/ershoufang/> +<地区名> +<数字代号>.html

回到刚刚的界面使用f12打开调试窗口点击进入详情页的界面可以看到这样的信息



通过对上图进行分析，可以通过正则表达式来获取到每个二手房详情页的url

re\_f = '<a class="" href="(.\*?)" target="\_blank"'

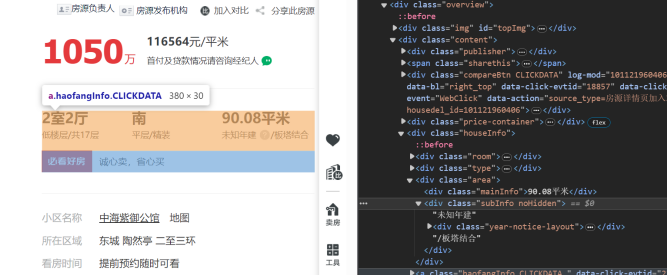
    url\_list = re.findall(re\_f, res\_text)

    return url\_list



输出了详情页的url

进入详情页



使用同样的方法获得其他属性

 # 获取房屋的标题

    title = res\_text.xpath("//div[@class='sellDetailHeader']//h1/@title")

    # 获取房屋的总价

    total\_price = res\_text.xpath("//div[@class='overview']//div/span/text()")[2]

    # 获取房屋的单价

    price = res\_text.xpath("//div[@class='overview']//div/span/text()")[3]

    # 获取房屋的地段

    place = res\_text.xpath("//div[@class='overview']//div/span/a/text()")

    ## 房屋基本信息获取

    # 获取房屋基本信息的标题

    lab = res\_text.xpath("//div[@class='base']//span/text()")

    # 获取房屋基本信息的内容

    val = res\_text.xpath("//div[@class='base']//li/text()")

    ## 获取房源交易信息

    # 获取房源交易标题

    key1 = res\_text.xpath("//div[@class='transaction']//span[1]//text()")

    # 获取房源交易信息内容

    trans = res\_text.xpath("//div[@class='transaction']//span[2]//text()")

    ## 获取房源特色信息

    # 获取房源特色标题

    key = res\_text.xpath("//div[@class='baseattribute clear']/div[@class='name']/text()")

    # 获取房源特色内容

    val1 = res\_text.xpath("//div[@class='baseattribute clear']/div[@class='content']/text()")



现在我们已经获得了同一页所有二手房的信息,接下来我们获得不同页不同区的信息

通过测试翻页和选择地区我们发现

https://bj.lianjia.com/ershoufang/dongcheng/

<https://bj.lianjia.com/ershoufang/dongcheng/pg2/>https://bj.lianjia.com/ershoufang/dongcheng/pg3/

链家网址是

A = <https://bj.lianjia.com/ershoufang/>

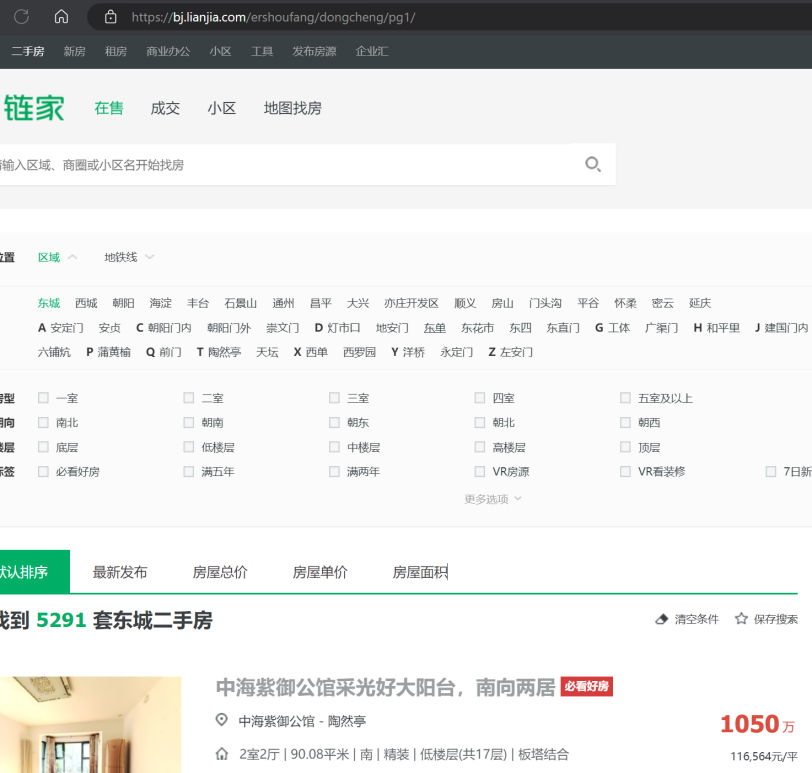
B = 地区拼音

C = 数字页码(从1开始)

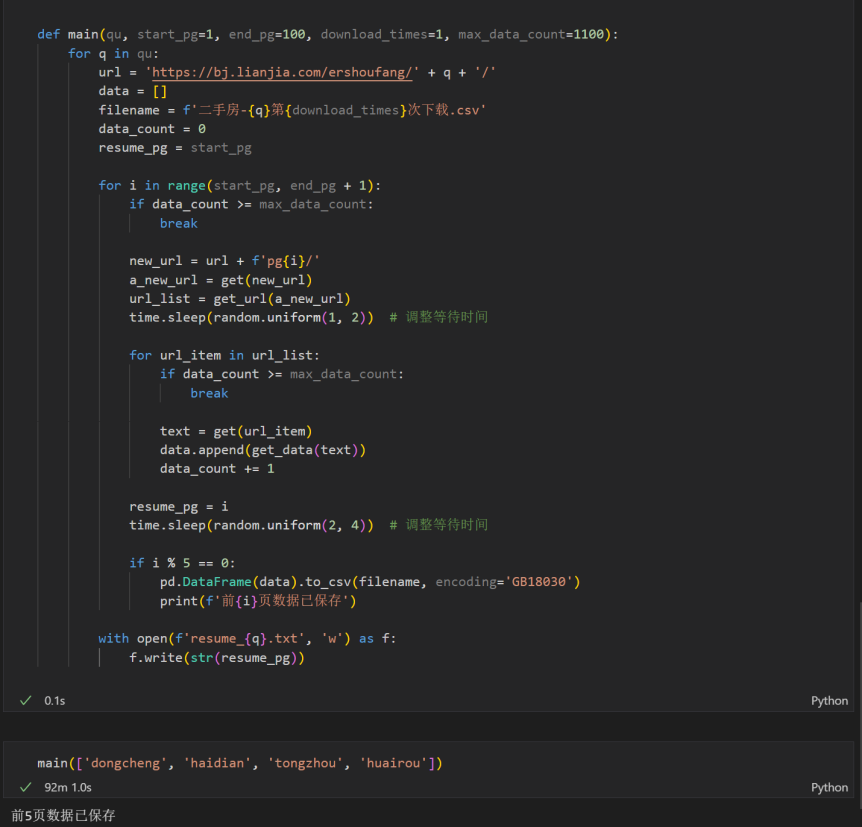
形如

A+B+/pg+C

同时虽然第一页看起来不符合这个范式,但是我们可以通过https://bj.lianjia.com/ershoufang/dongcheng/pg1访问第一页的



那么就可以爬取了



在长时间的爬取之后终于获得了数据

每个城区1000条且属性约有30条的数据了

'标题', '总价格', '单价', '地段', '房屋户型', '所在楼层', '建筑面积', '户型结构',

'套内面积', '建筑类型', '房屋朝向', '建筑结构', '装修情况', '梯户比例', '供暖方式', '配备电梯', '挂牌时间',

'交易权属', '上次交易', '房屋用途', '房屋年限', '产权所属', '抵押信息', '房本备件', '核心卖点', '小区介绍',

'户型介绍', '装修描述', '税费解析', '周边配套', '售房详情', '交通出行', '适宜人群', '权属抵押', '用水类型',

'用电类型'



2数据预处理

数据预处理： 删除不重要的属性只保留需要的

# 保留指定列

selected\_columns = ['总价格', '单价', '地段', '房屋户型', '所在楼层', '建筑面积', '户型结构', '装修情况']

# 去除建筑面积列中的"㎡"字符并转换为float类型

# 将所在楼层()里面的内容删去

# 去除建筑面积列中的"㎡"字符并转换为float类型

# 将所在楼层()里面的内容删去

# 去除建筑面积列中的"㎡"字符并转换为float类型

dongcheng\_df['建筑面积'] = dongcheng\_df['建筑面积'].str.rstrip('㎡').astype(float)

haidian\_df['建筑面积'] = haidian\_df['建筑面积'].str.rstrip('㎡').astype(float)

huairou\_df['建筑面积'] = huairou\_df['建筑面积'].str.rstrip('㎡').astype(float)

tongzhou\_df['建筑面积'] = tongzhou\_df['建筑面积'].str.rstrip('㎡').astype(float)

# 使用正则表达式删除括号及其内部的内容

dongcheng\_df['所在楼层'] = dongcheng\_df['所在楼层'].str.replace(r'\(.\*\)', '', regex=True)

haidian\_df['所在楼层'] = haidian\_df['所在楼层'].str.replace(r'\(.\*\)', '', regex=True)

huairou\_df['所在楼层'] = huairou\_df['所在楼层'].str.replace(r'\(.\*\)', '', regex=True)

tongzhou\_df['所在楼层'] = tongzhou\_df['所在楼层'].str.replace(r'\(.\*\)', '', regex=True)

对采集到的数据集进行重复值处理:删除重复值

# 1. 处理重复值

# 删除重复行，要求所有列都相同才删除

dongcheng\_df = dongcheng\_df.drop\_duplicates(subset=dongcheng\_df.columns.tolist(), keep='first')

haidian\_df = haidian\_df.drop\_duplicates(subset=haidian\_df.columns.tolist(), keep='first')

huairou\_df = huairou\_df.drop\_duplicates(subset=huairou\_df.columns.tolist(), keep='first')

tongzhou\_df = tongzhou\_df.drop\_duplicates(subset=tongzhou\_df.columns.tolist(), keep='first')

、缺失值处理: 删除有缺失值的行

# 2. 处理缺失值

# 删除缺失值

dongcheng\_df = dongcheng\_df.dropna()

haidian\_df = haidian\_df.dropna()

huairou\_df = huairou\_df.dropna()

tongzhou\_df = tongzhou\_df.dropna()

3数据分析--统计性分析和可视化分析

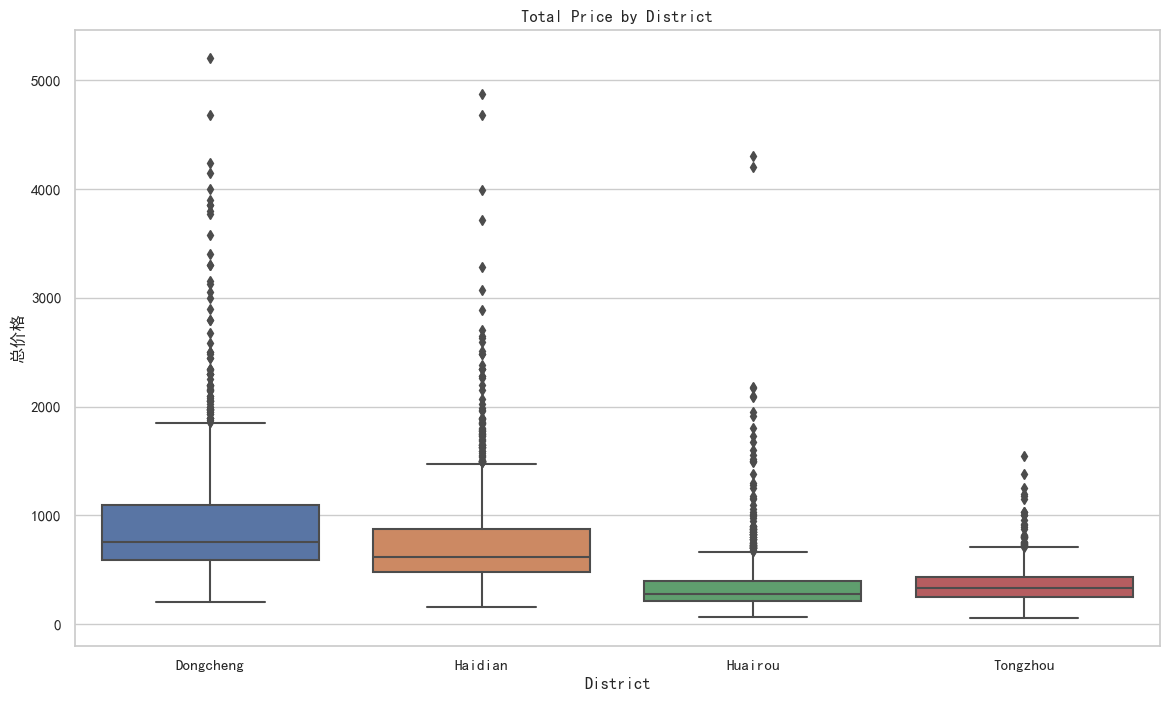
可视化思路:数值型变量采用箱线图

字符型采用柱状图或饼状图

这四个城区在售的二手房有哪些特点

总体来说,东城区房价最高

下面是总价,单价和面积



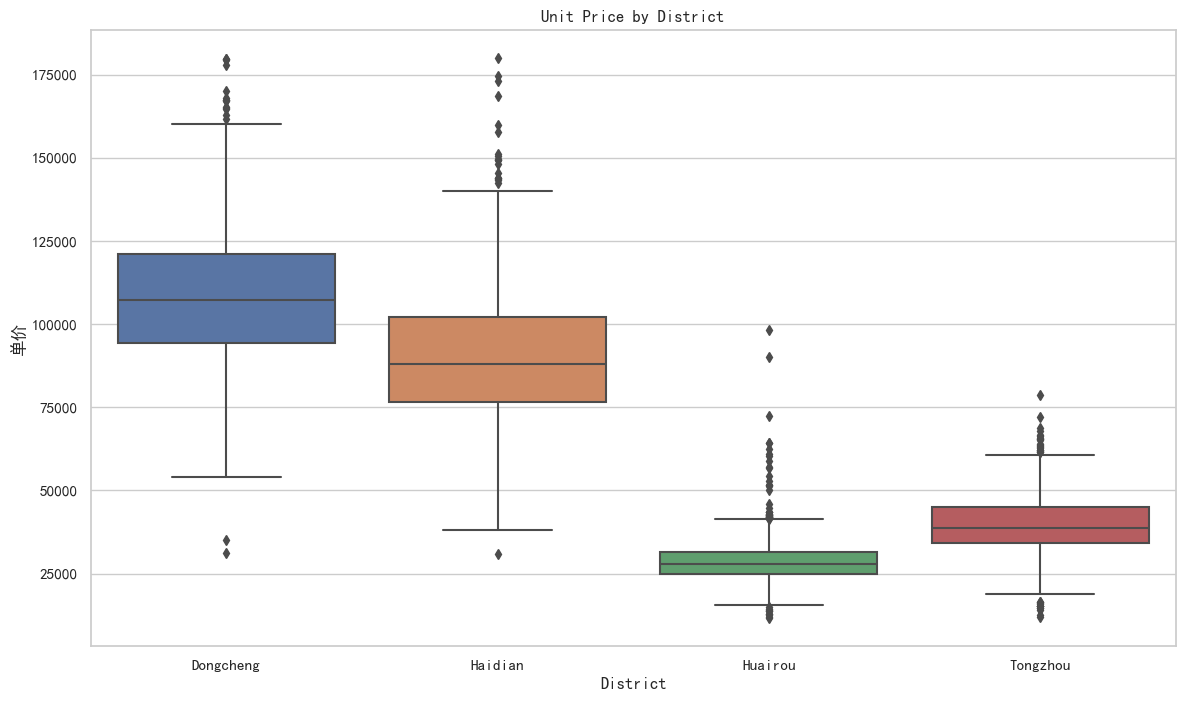
# 总价格 - 箱线图

plt.figure(figsize=(14, 8))

sns.boxplot(x='District', y='总价格', data=all\_data)

plt.title('Total Price by District')

plt.show()



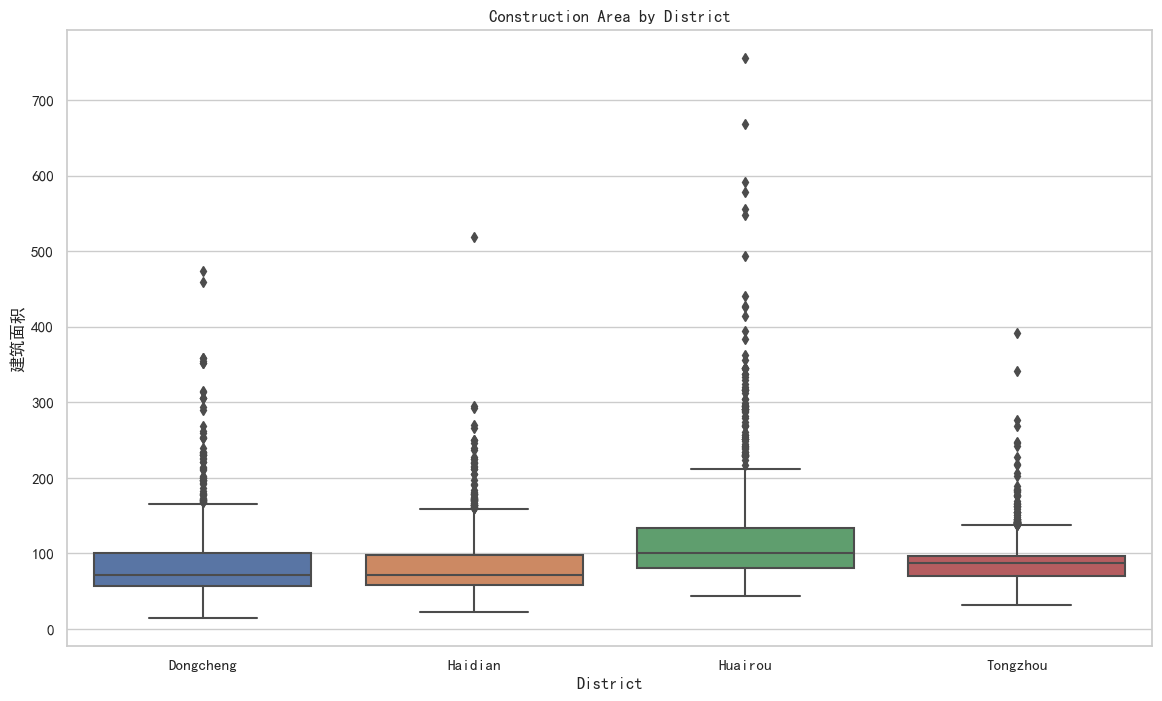
# 单价 - 箱线图

plt.figure(figsize=(14, 8))

sns.boxplot(x='District', y='单价', data=all\_data)

plt.title('Unit Price by District')

plt.show()



# 建筑面积 - 箱线图

plt.figure(figsize=(14, 8))

sns.boxplot(x='District', y='建筑面积', data=all\_data)

plt.title('Construction Area by District')

plt.show()

、户型类别

# 遍历每个区并绘制房屋户型分布饼图

districts = all\_data['District'].unique()

for district in districts:

    district\_data = all\_data[all\_data['District'] == district]

    house\_type\_counts = district\_data['房屋户型'].value\_counts()

    # 绘制饼图

    plt.figure(figsize=(10, 8))

    plt.pie(house\_type\_counts, labels=house\_type\_counts.index, autopct='%1.1f%%', startangle=140)

    plt.title(f'{district}房屋户型分布')

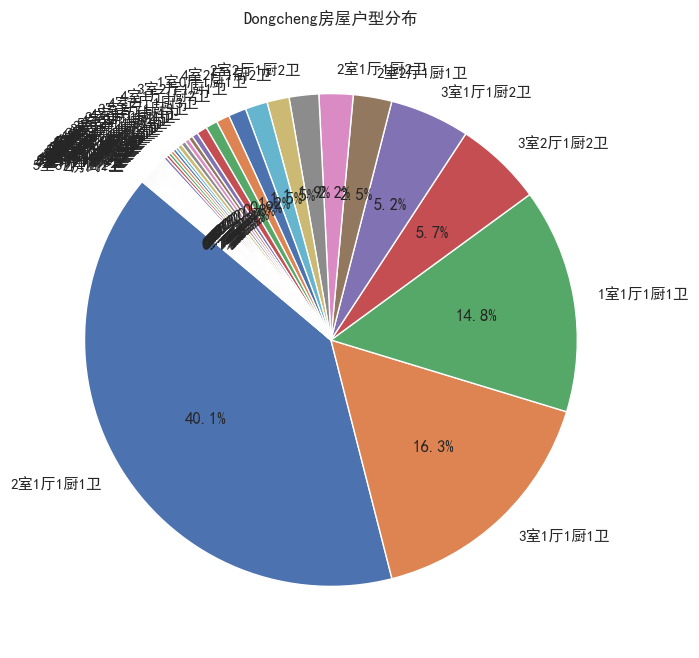
    plt.show()

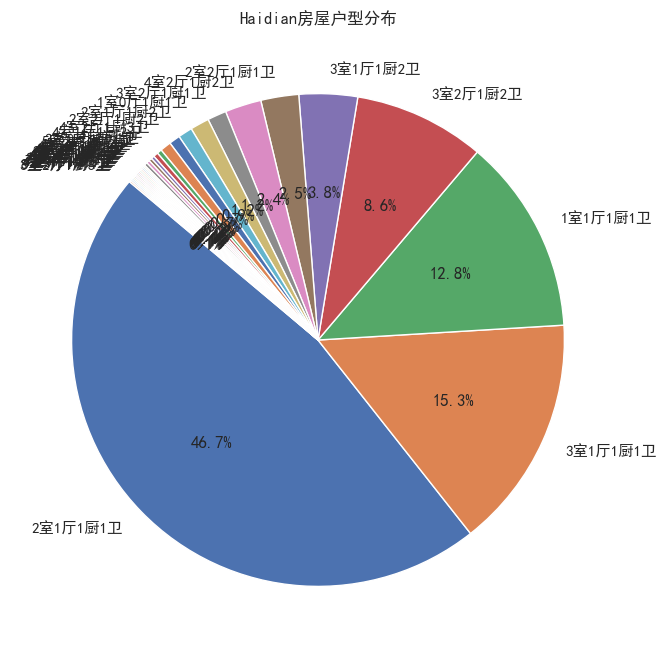
    # 输出分布情况

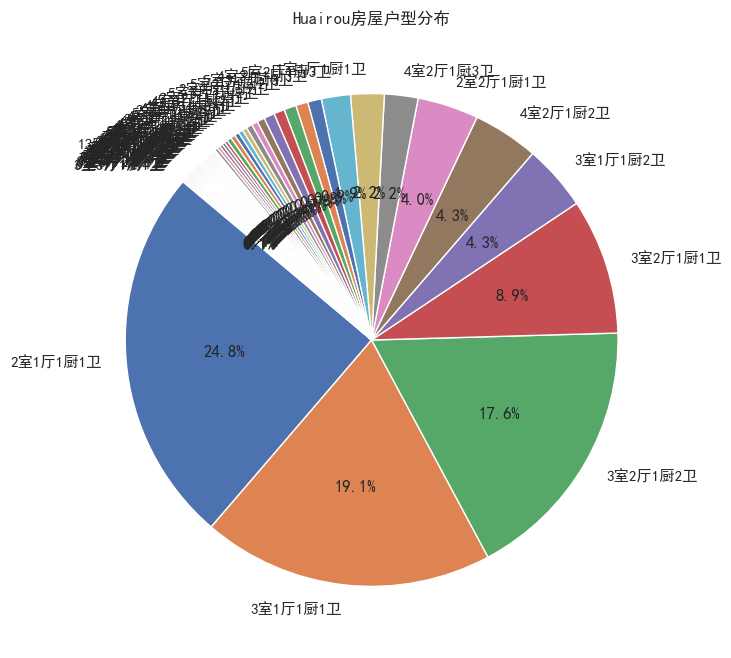
    print(f'{district}房屋户型分布情况:')

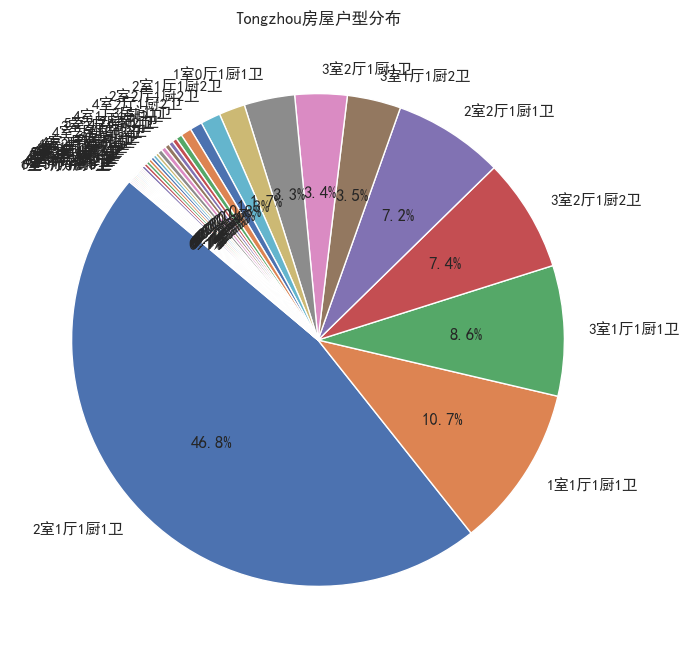
    print(house\_type\_counts)

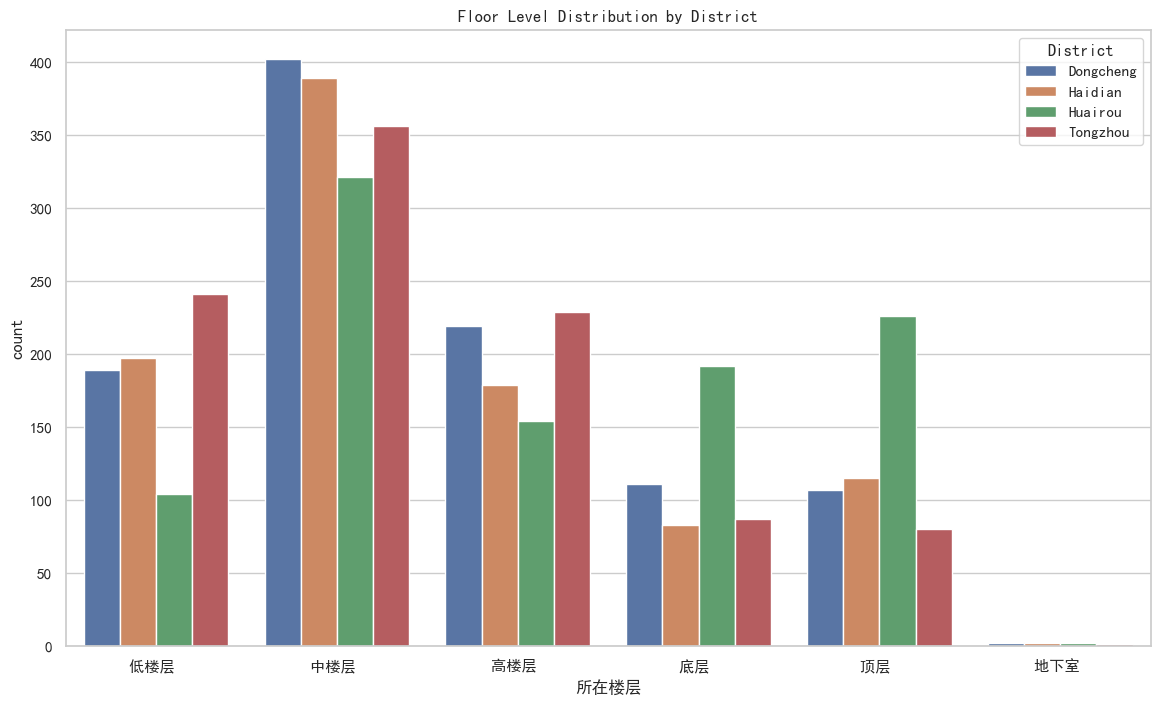
    print("\n")







、楼层



# 所在楼层 - 柱状图

plt.figure(figsize=(14, 8))

sns.countplot(x='所在楼层', hue='District', data=all\_data)

plt.title('Floor Level Distribution by District')

plt.show()

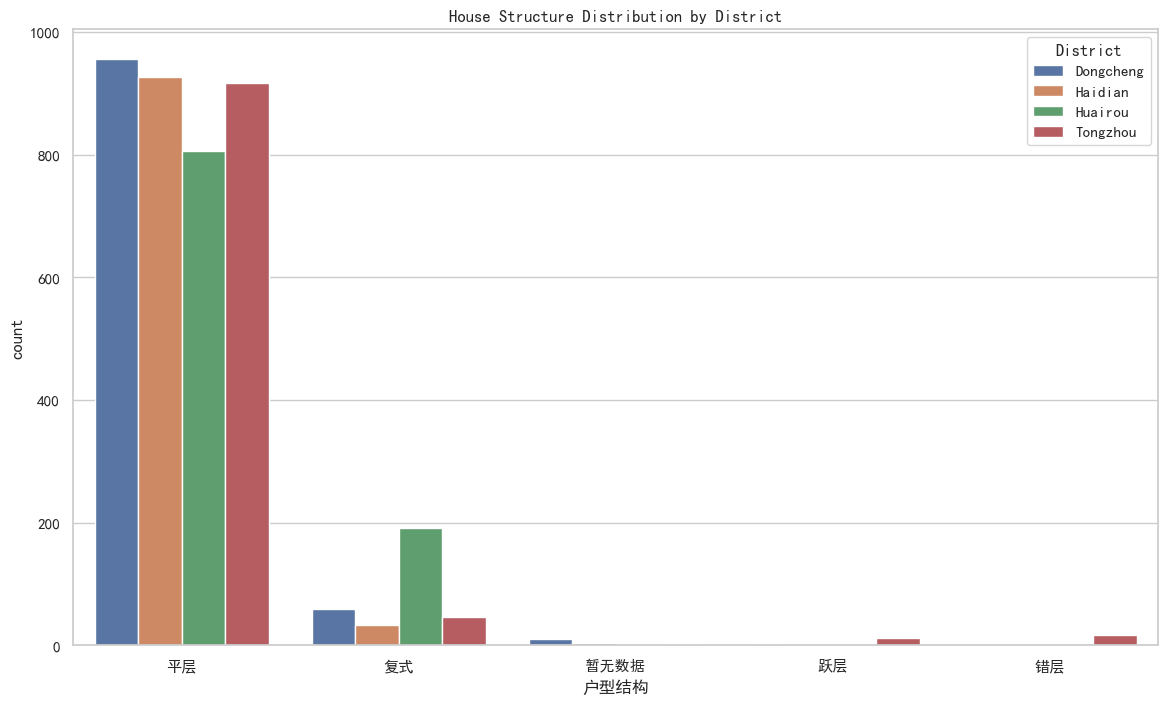
for district in districts:

    district\_data = all\_data[all\_data['District'] == district]

    # 输出分布情况

    print(f'{district}所在楼层分布情况:')

    print(district\_data['所在楼层'].value\_counts())

、户型结构  


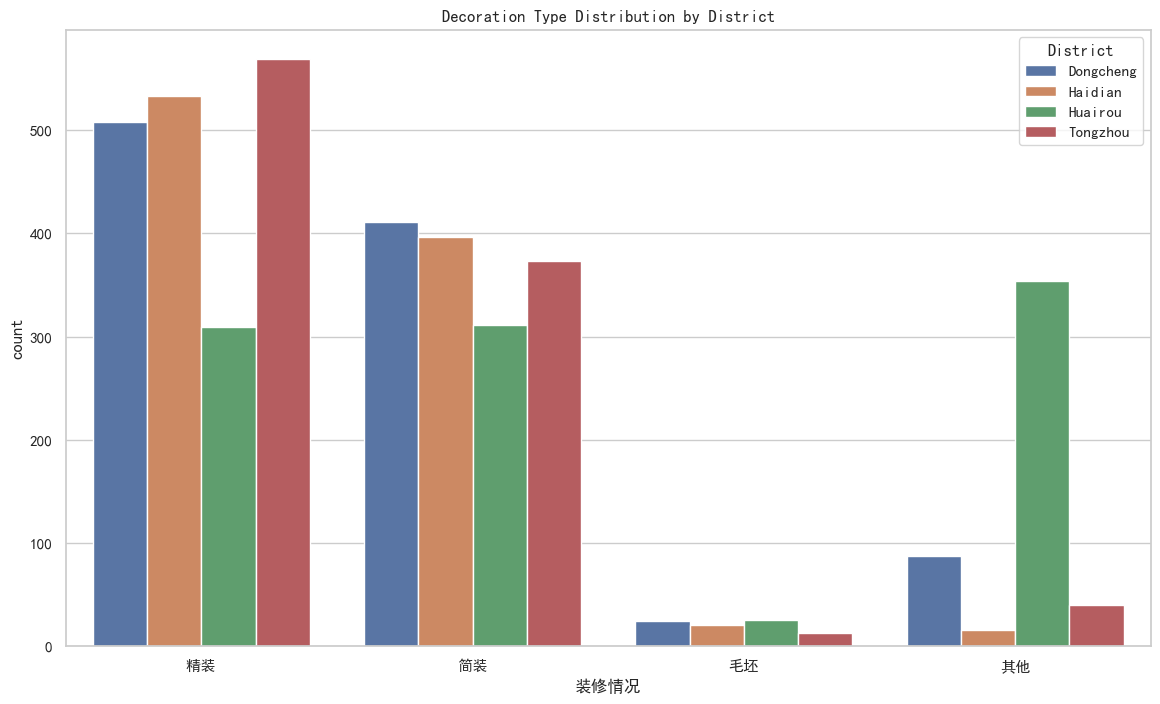
# 户型结构 - 柱状图

plt.figure(figsize=(14, 8))

sns.countplot(x='户型结构', hue='District', data=all\_data)

plt.title('House Structure Distribution by District')

plt.show()

、装修情况

# 装修情况 - 柱状图

plt.figure(figsize=(14, 8))

sns.countplot(x='装修情况', hue='District', data=all\_data)

plt.title('Decoration Type Distribution by District')

plt.show()

### 二手房特点分析

总价格、单价、建筑面积、所在楼层和房屋户型

#### 东城区 (Dongcheng)

价格与面积：东城区有着最高的总价格和单价均值，体现了其作为老城区的地段优势和房产价值。建筑面积虽然分布广泛，但多数集中在较小面积，符合市中心老房子的特征。

所在楼层分布：中楼层最多，其次是高楼层和低楼层，较少的底层和顶层可能是由于老城区楼房的建筑特点决定的。

房屋户型：2室1厅1厨1卫最为常见，符合核心城区家庭居住的需求，同时也有较多的3室和1室户型，多样的户型反映了多元化的居住需求。建筑面积: 平均值87.86平方米，最大值接近474平方米。这说明东城区房产类型多样，从小户型到大户型都有涵盖。

## 海淀区 (Haidian)

价格与面积：海淀区的总价格和单价略低于东城区，但依然位于较高水平。建筑面积平均值略小，反映了海淀区作为学区和科技中心，中小户型需求旺盛。

所在楼层分布：中楼层和低楼层较多，符合海淀区较多的中低层住宅和教育区的建筑特点。

房屋户型：2室1厅1厨1卫同样是最常见的户型，适合中小家庭和在校学生群体。

#### 怀柔区 (Huairou)

价格与面积：怀柔区总价格和单价明显低于城区，但平均建筑面积最大，反映出怀柔区的宽敞空间和较低的人口密度，适合度假和居住。

所在楼层分布：中楼层最多，其次是顶层和底层，可能由于郊区的多层建筑和别墅较多。

房屋户型：3室和2室户型较多，显示出怀柔区居民偏好更大的生活空间，适合家庭居住。

#### 通州区 (Tongzhou)

价格与面积：通州区的总价格和单价都处于较低水平，但随着城市扩展和交通改善，这里的房产正在逐渐升值。建筑面积分布较均匀，多数为中小户型。

所在楼层分布：中楼层最多，其次是低楼层和高楼层，显示出通州区高层住宅的普遍性。

房屋户型：2室1厅1厨1卫户型最为普遍，反映了中等收入群体的居住需求

东城区和海淀区因其优越的地理位置和丰富的教育资源，有较高的房价和单价，户型多样，主要集中在小到中等面积的住宅。

怀柔区由于其远郊的位置和自然环境，房价相对较低，但提供了更宽敞的居住环境，适合追求安静和舒适生活空间的居民。

通州区作为北京的副中心，正在逐步发展，房价和单价适中，主要提供中小户型的住宅，吸引了大量通勤和中等收入人群。

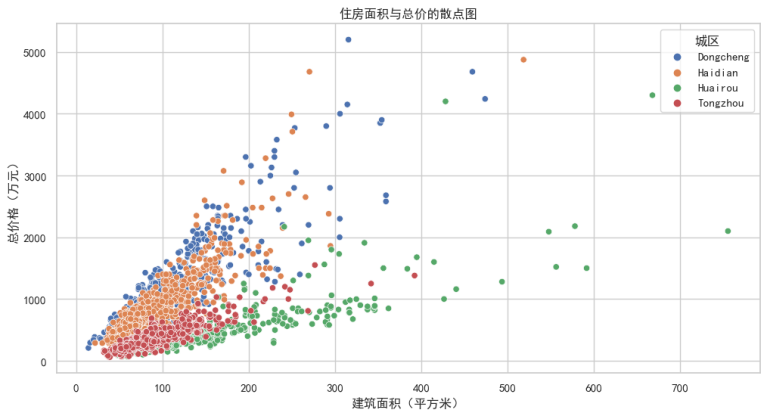
4.数据分析--分析影响房价的主要因素有哪些

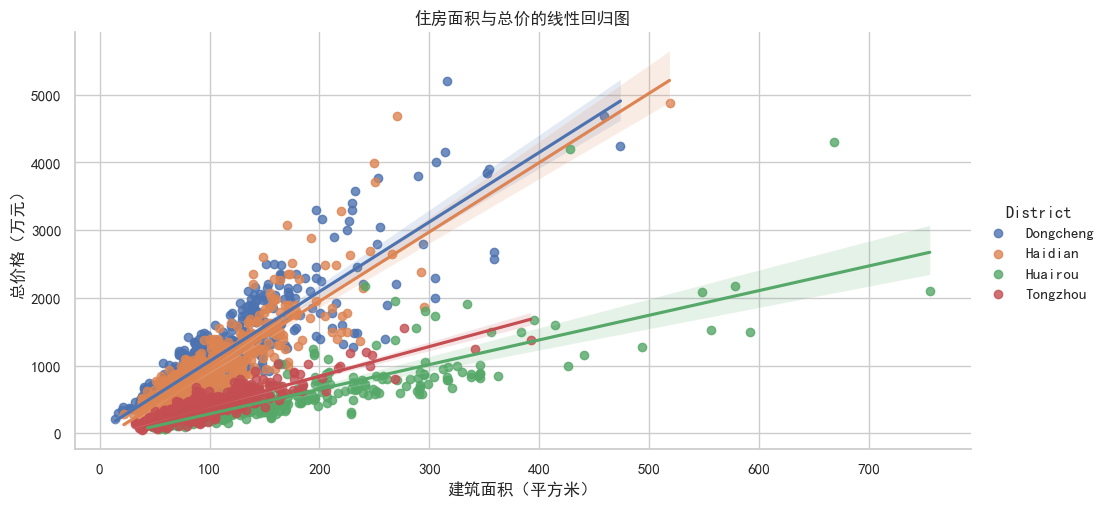
分析影响房价的主要因素有哪些，即分析房价与其他变量之间的关系。

房价与面积:

可以看到房价与面积大体上是正相关的,同时正相关的程度有所不同,东城区在同等面积下大概率是4个区中最贵的,同时怀柔是最便宜的.

这与怀柔地处郊区关系较深





图形绘制代码

# 散点图 - 展示住房面积与总价的关系

plt.figure(figsize=(12, 6))

sns.scatterplot(x='建筑面积', y='总价格', hue='District', data=all\_data)

plt.title('住房面积与总价的散点图')

plt.xlabel('建筑面积（平方米）')

plt.ylabel('总价格（万元）')

plt.legend(title='城区')

plt.show()

# 线性回归图 - 展示住房面积与总价的关系及趋势线

plt.figure(figsize=(12, 6))

sns.lmplot(x='建筑面积', y='总价格', hue='District', data=all\_data, aspect=2)

plt.title('住房面积与总价的线性回归图')

plt.xlabel('建筑面积（平方米）')

plt.ylabel('总价格（万元）')

plt.show()

房价与装修情况:

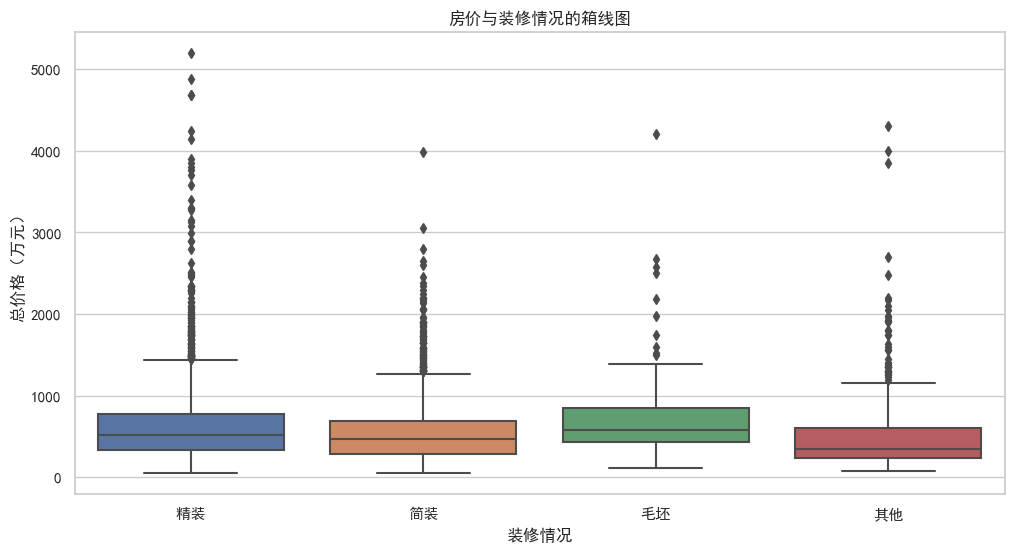
预计的结果,装修情况越好房价越高

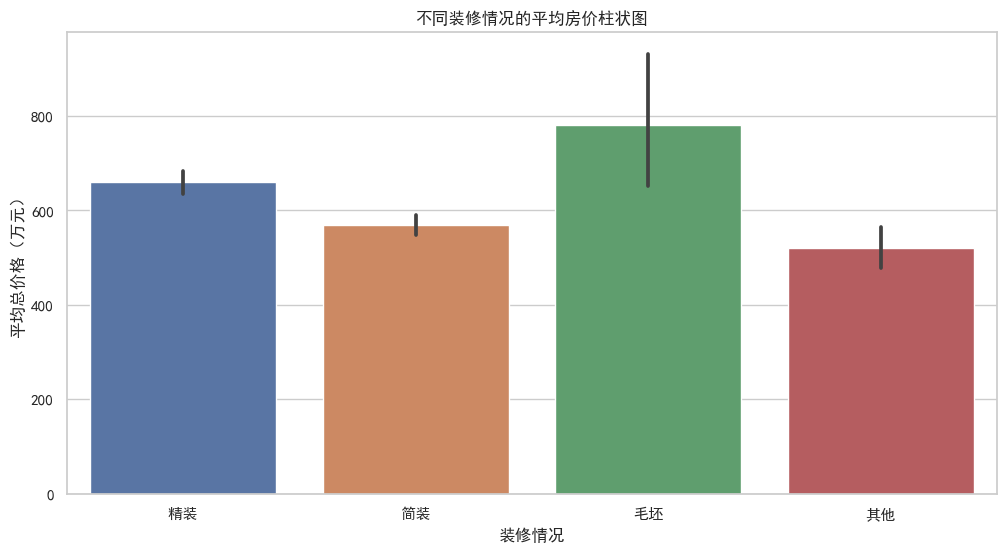
实际上相关性不是一个正相关或者负相关

这里我们不考虑其他

我们可以看到毛坯甚至大体上比精装修还贵一些

精装修又比简装修贵一些

这里我们猜想北京的房价具备一定的金融产品属性,人们买来很大一部分不是买来居住,而是作为商品低买高卖导致的结果  




图形绘制代码

# 箱线图 - 展示总价与装修情况的关系

plt.figure(figsize=(12, 6))

sns.boxplot(x='装修情况', y='总价格', data=all\_data)

plt.title('房价与装修情况的箱线图')

plt.xlabel('装修情况')

plt.ylabel('总价格（万元）')

plt.show()

# 柱状图 - 展示不同装修情况的平均房价

plt.figure(figsize=(12, 6))

sns.barplot(x='装修情况', y='总价格', data=all\_data, estimator=mean)

plt.title('不同装修情况的平均房价柱状图')

plt.xlabel('装修情况')

plt.ylabel('平均总价格（万元）')

plt.show()

5数据分析--分析二手房价格的地理分布特征

6 选做题：