上海二手房屋（公寓）属性分析

高三2班 张宇昂

目录

1. 前言
2. 数据采集、预处理

数据采集

预处理

1. 数据分析

准备工作

单位统一

图表形式简介

图表

索引

表1~33

1. 总结
2. 前言

房屋价格的影响因素非常多，影响因素之间的关系错综复杂。本文选择如下因素进行分析：厅室数量（几室几厅）、建筑面积、装修（精装、简装、毛坯）、建筑类型（板楼、塔楼）、朝向、所属区、地段（内环内、内环至中环、中环至外环、外环外）、建造年份。分析 每个影响因素对房屋单价和总价的影响。别墅样本数量不足、较为特殊，故不在本文分析范围之内。

1. 数据采集、预处理

房源信息采集自链家二手房（https://sh.lianjia.com/ershoufang/），共找到 58894 套上海二手房。本次数据采集于2018年2月11日星期日，采集共53007条在售房源信息，5887条信息未能采集，原因不明。

数据采集

链家二手房页面如下所示，共18个区域



每个区域有若干子区域



每个子区域有若干页



每页有若干条房源信息，每条信息对应一个在售房屋



链接内包含所需数据，本文仅收集便于量化处理的数据



以上各部均可自动化实现，大致步骤如下：

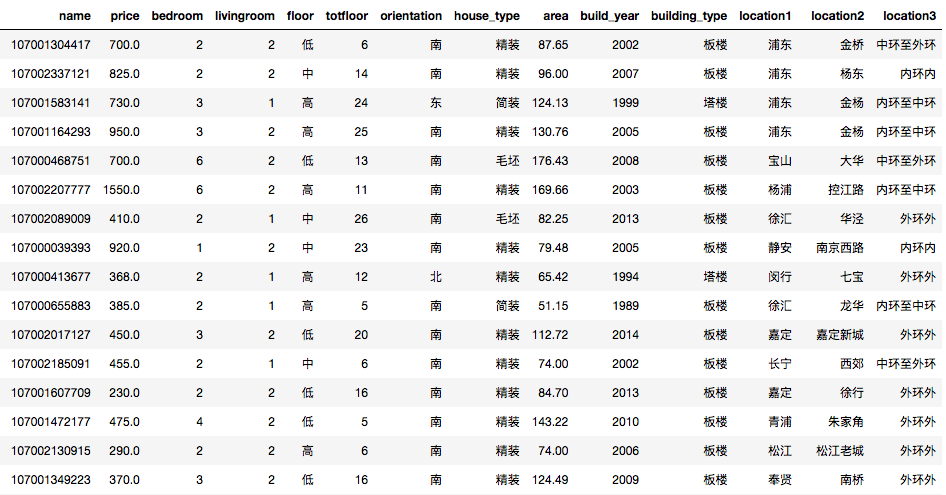
1. 收集每个子区域、每页、每条信息的链接（共53007条，约10分钟）
2. 下载每个链接对应的网页（共7.3GB，耗时一小时）

数据预处理

1. 分析html结构，提取总价、房厅数量、面积、装修（精装、简装、毛坯）、建筑类型（板楼、塔楼）、朝向、所属区、地段（内环内、内环至中环、中环至外环、外环外）、建造年份、总楼层、所在楼层（低、中、高）、子区域。其中楼层和子区域信息过于复杂，本文不做分析。
2. 对所需信息进行过滤，排除异常数据：
   1. 总楼层小于四层，或总楼层数量不明的
   2. 建筑类型不是板楼或塔楼的
   3. 装修类型不是精装、简装、毛坯之一的
   4. 朝向不明、建造年份不明的
   5. 价格高于6000万的
3. 根据总价和面积计算每平米价格

进行上述处理后，仍有42609样本。保存为csv表格文件，大小为3.7MB。

下图为处理后的样本样例，程序见附录：



1. 数据分析

准备工作

单位统一

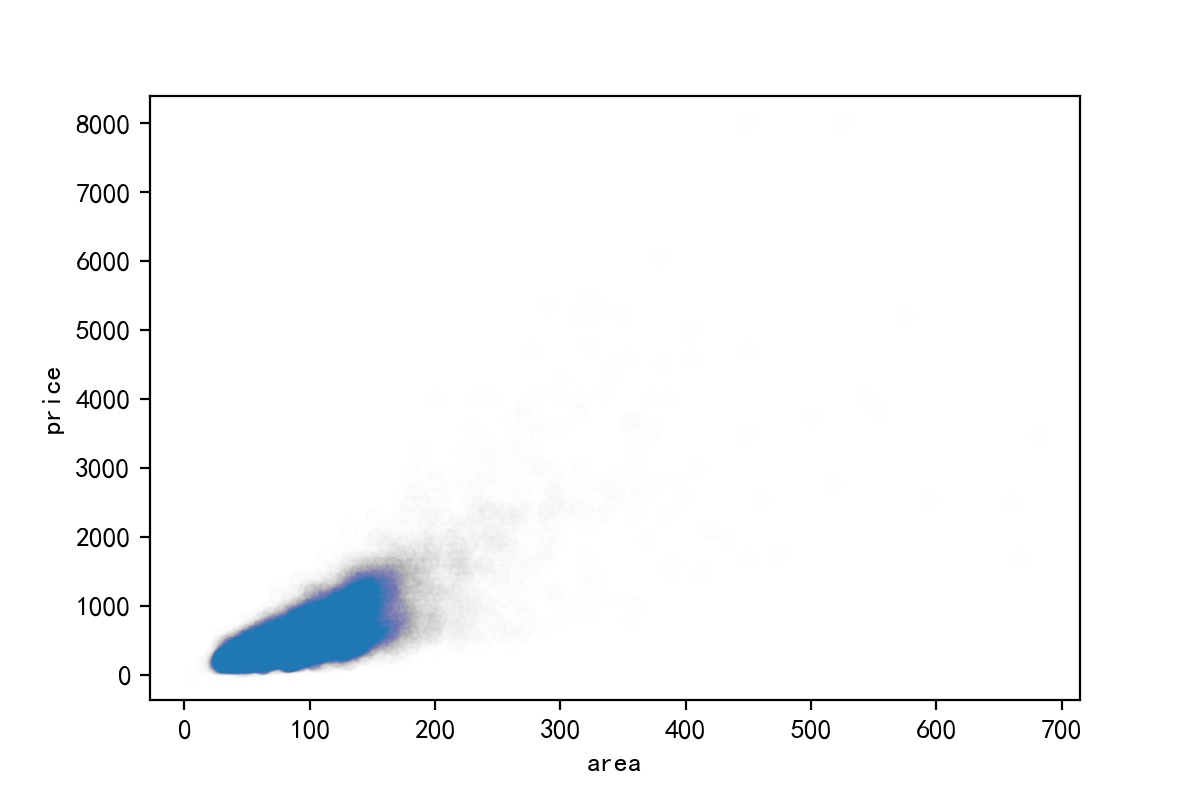
|  |  |  |
| --- | --- | --- |
| 名称 | 单位 | 图标、程序内名称 |
| 房屋总价 | 万元 | price |
| 卧室数量 | 间 | bedroom |
| 餐厅、客厅数量 | 间 | livingroom |
| 总楼层 | 层 | totfloor (total floor) |
| 总面积 | 平方米 | area |
| 一平米价格 | 万元/平方米 | unit\_price |

图表形式简介

1. 散点图（每一个房屋的一对数据都对应平面上的一个点）

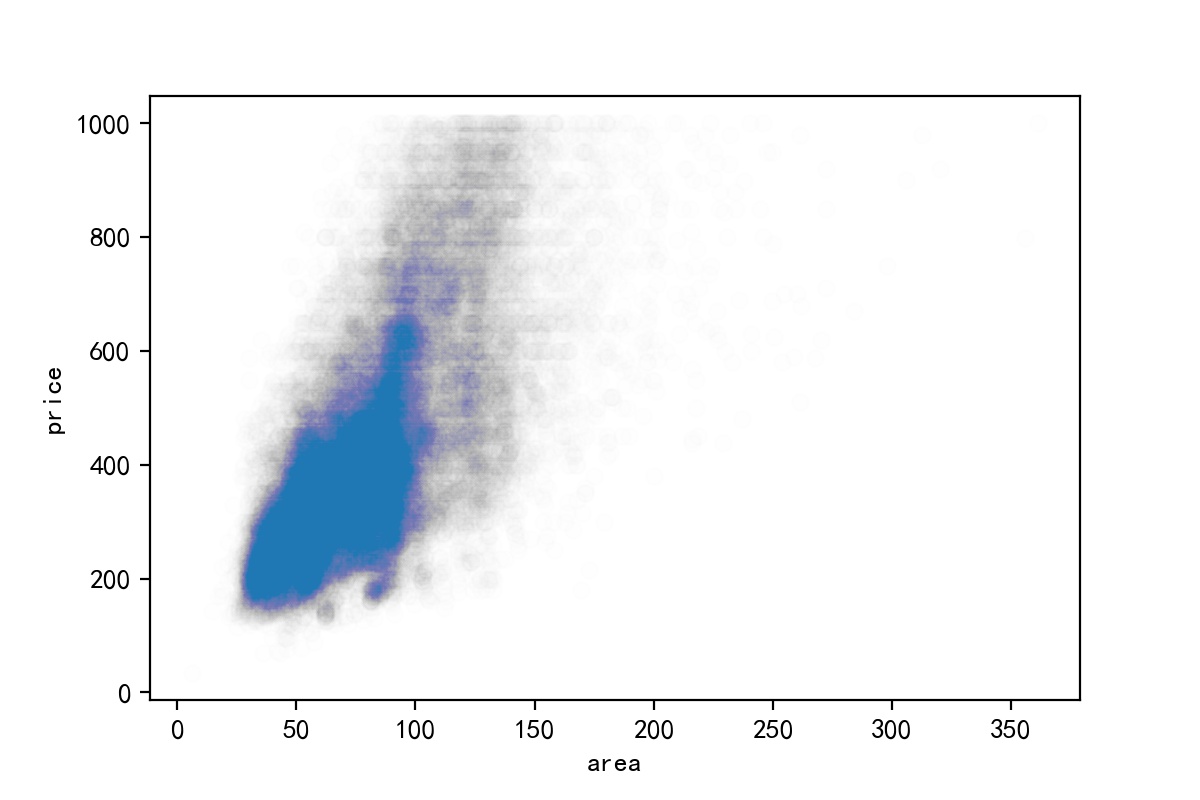
表现连续自变量对连续因变量的影响

举例：房屋总价-面积图

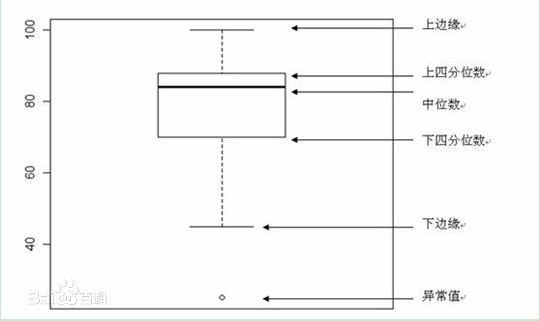


由于样本数量太大，所以采用了低透明度的设置，颜色越深表示该处房屋越多。上图表现了房屋总价和面积的关系，可以看出房屋总价与房屋面积正相关，绝大部分房屋价格小于2000万元，约94%房屋小于1000万元。后文部分图表为便于观察，仅采用小于1000万元的房屋数据。

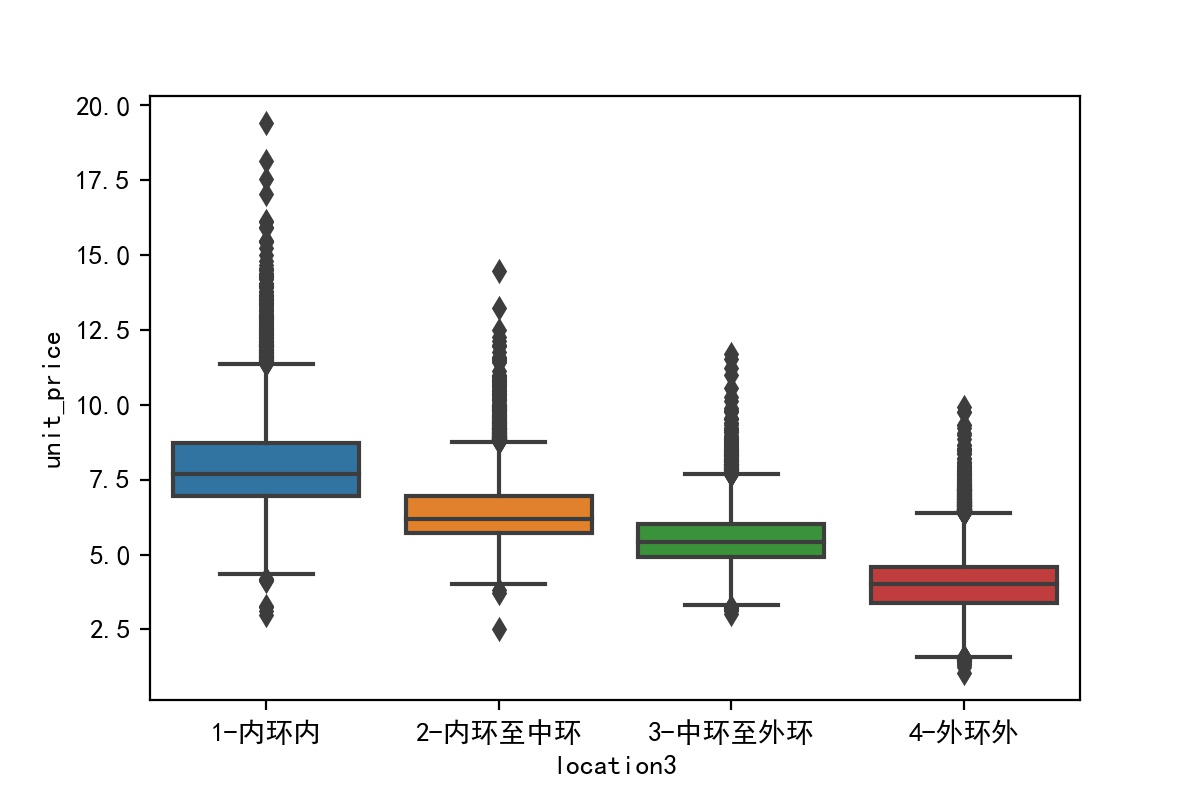
举例：房屋总价-面积图（总价小于1000万元）



1. 箱线图（表现离散自变量对连续因变量的影响）



举例：单价-地段图



由上图可见，在各个四分位上，房屋单价随距市中心距离递减。图表

索引

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 变量 | 室 | 厅 | 地段 | 区划 | 建筑类型 | 装修类型 |
| 总价 | 1 | 2 | 3 | 4 | 5 | 6 |
| 面积 | 9 | 10 | 11 | 12 | 13 | 无关 |
| 单价 | 17 | 18 | 19 | 20 | 21 | 22 |
| 建造年份 | 24 | 25 | 26 | 27 | 28 | 29 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 变量 | 总价 | 面积 | 单价 | 建造年份 |
| 总价 | \ | 7 | 无关 | 8 |
| 面积 | 7 | \ | 15 | 16 |
| 单价 | 无关 | 15 | \ | 23 |
| 建造年份 | 8 | 16 | 23 | \ |

附：

图30 楼高-建造年份

图31 楼高-建筑类型

图32 楼高-地段

图33 楼高-区划

图34 室-厅

图1 总价-卧室数量

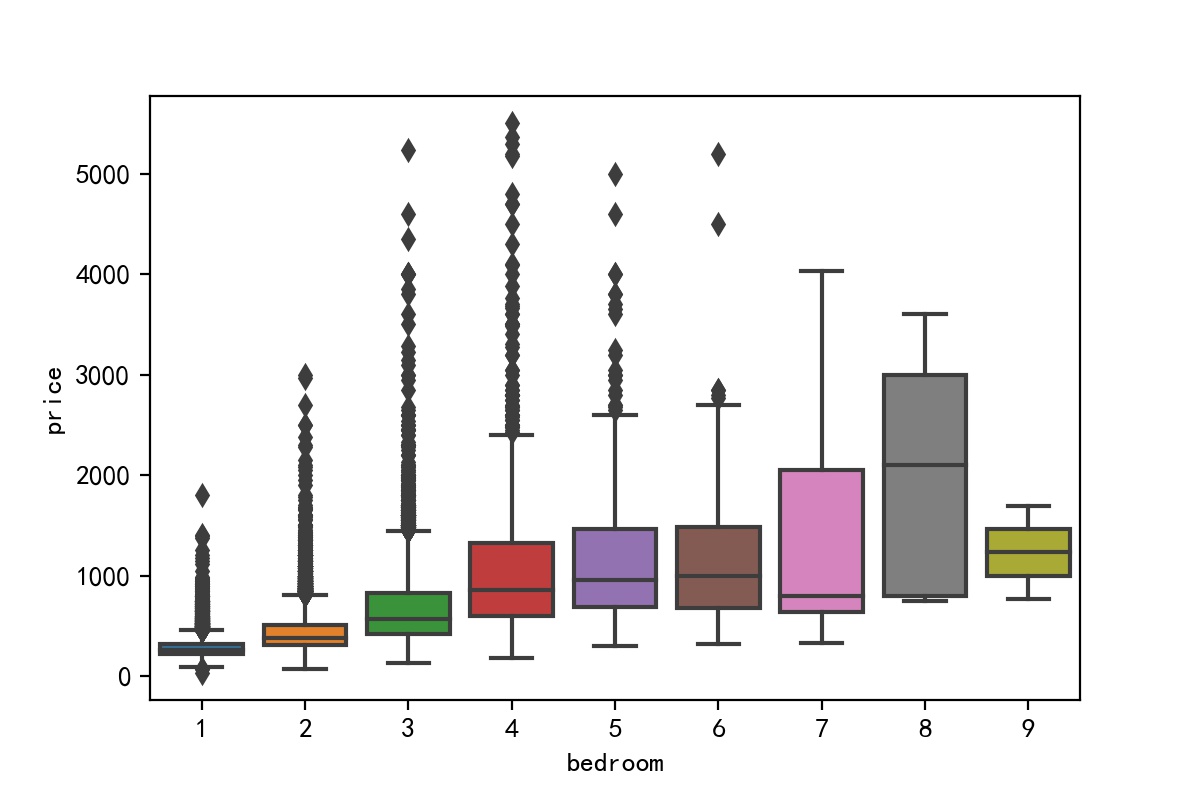
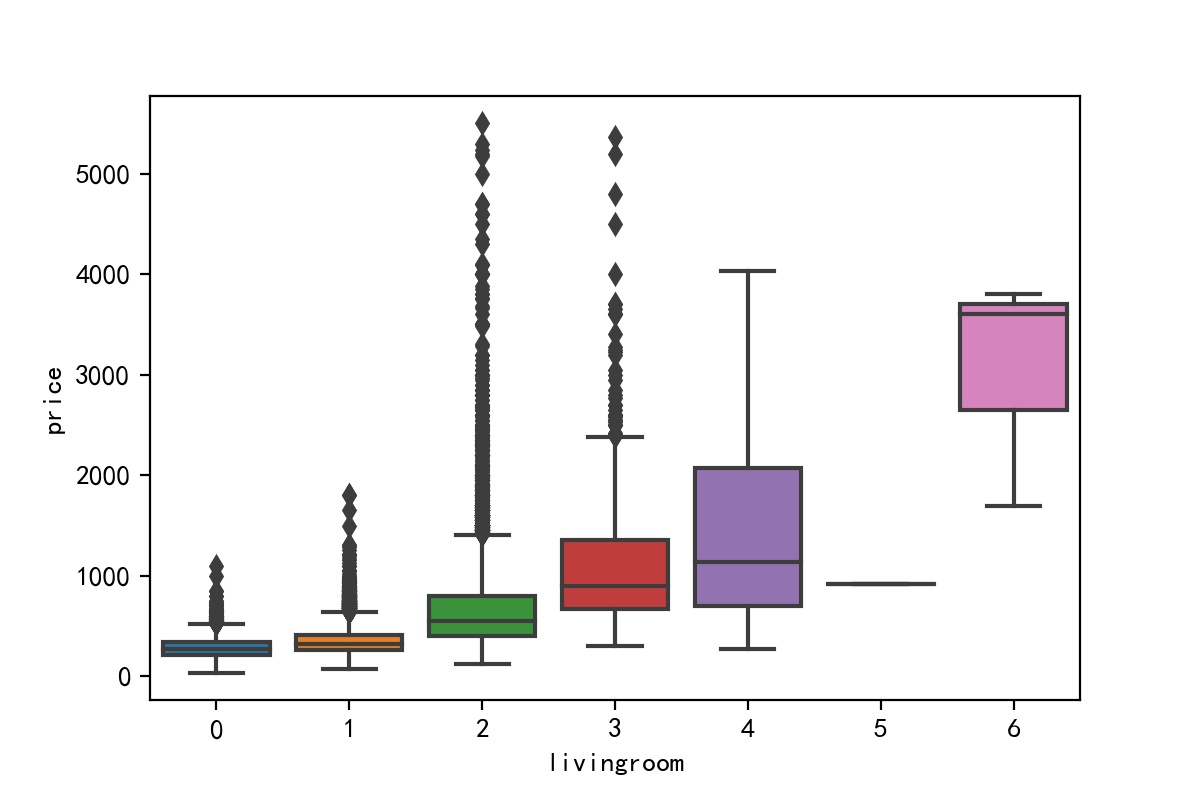


图2 总价-厅数量



总价与厅事数量呈正相关，零厅和一厅价格明显低于两厅。

图3 总价-地段

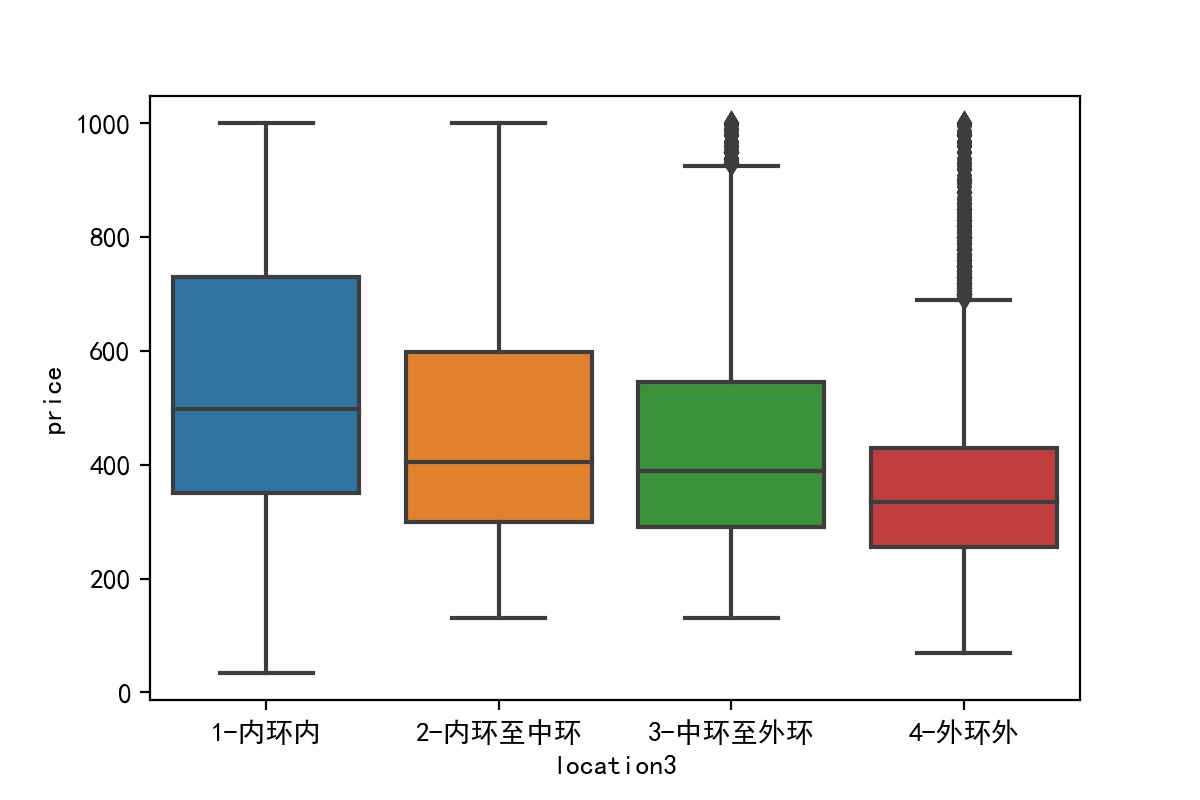
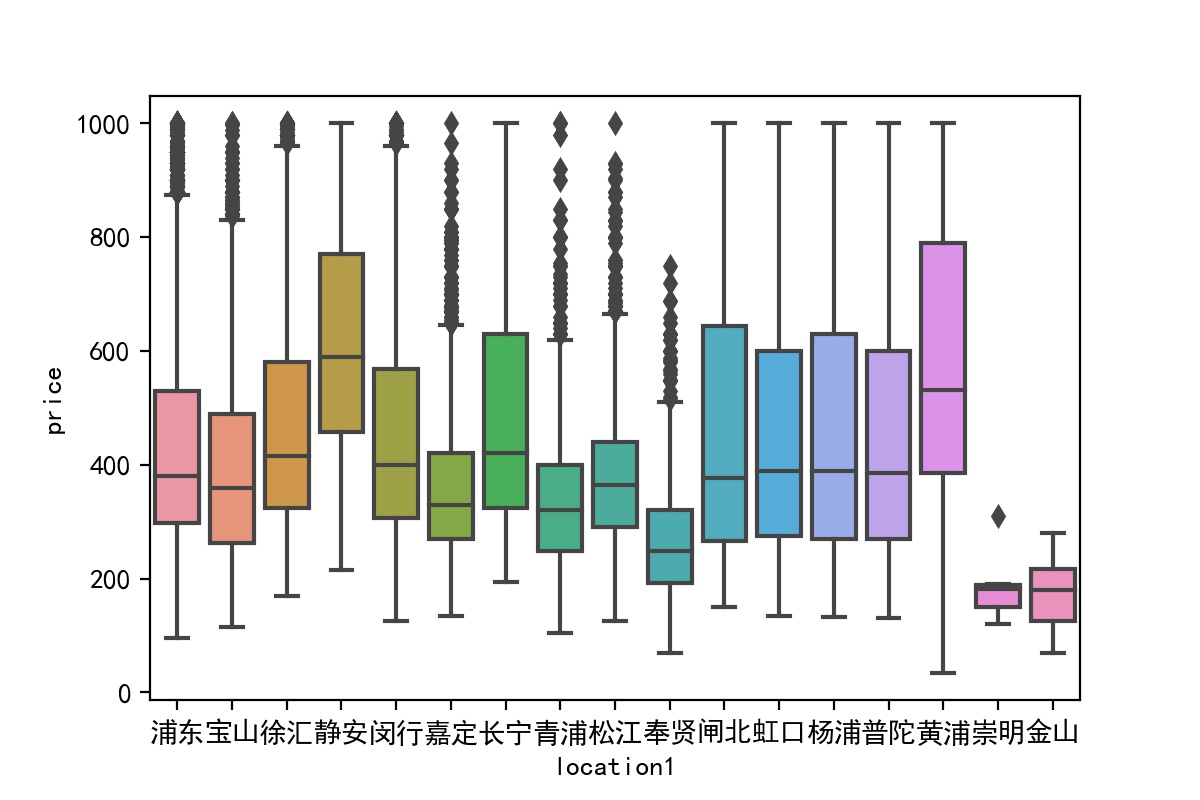


图4 总价-区划



距市中心越近，房价越高。崇明金山房价远低于其他区。

图5 总价-建筑类型

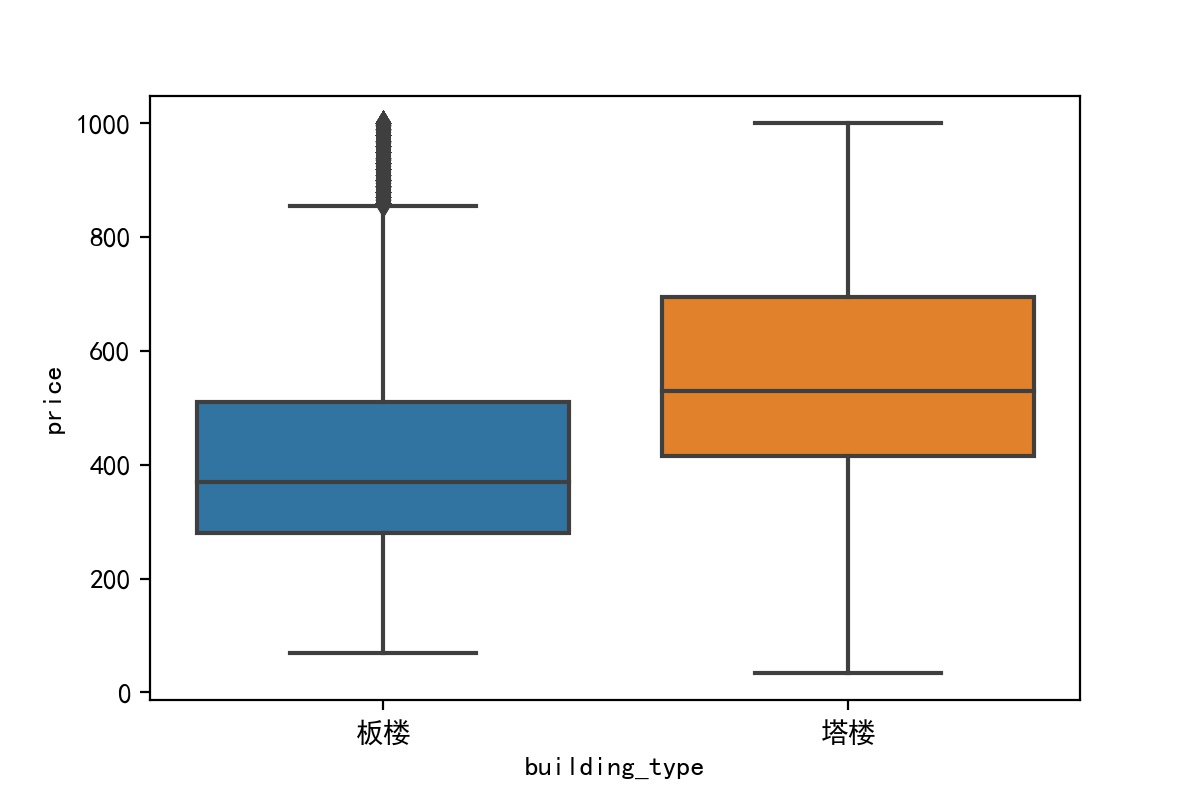
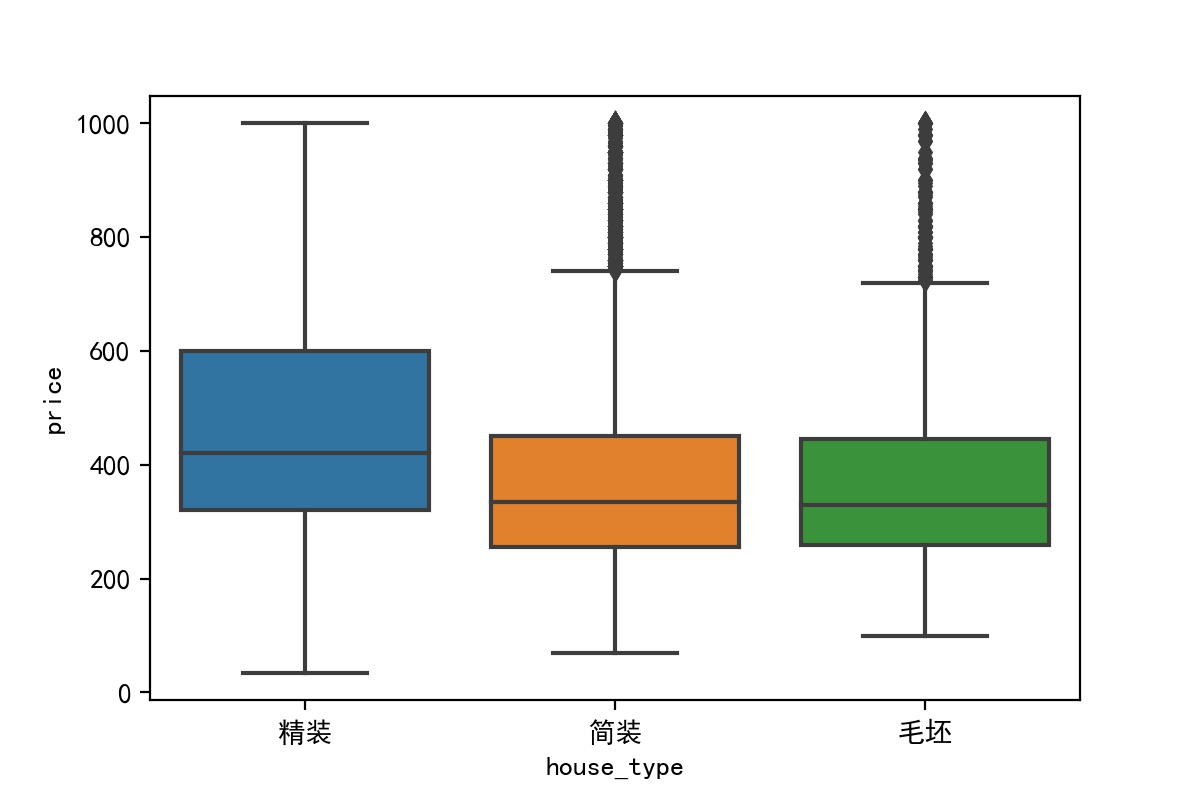
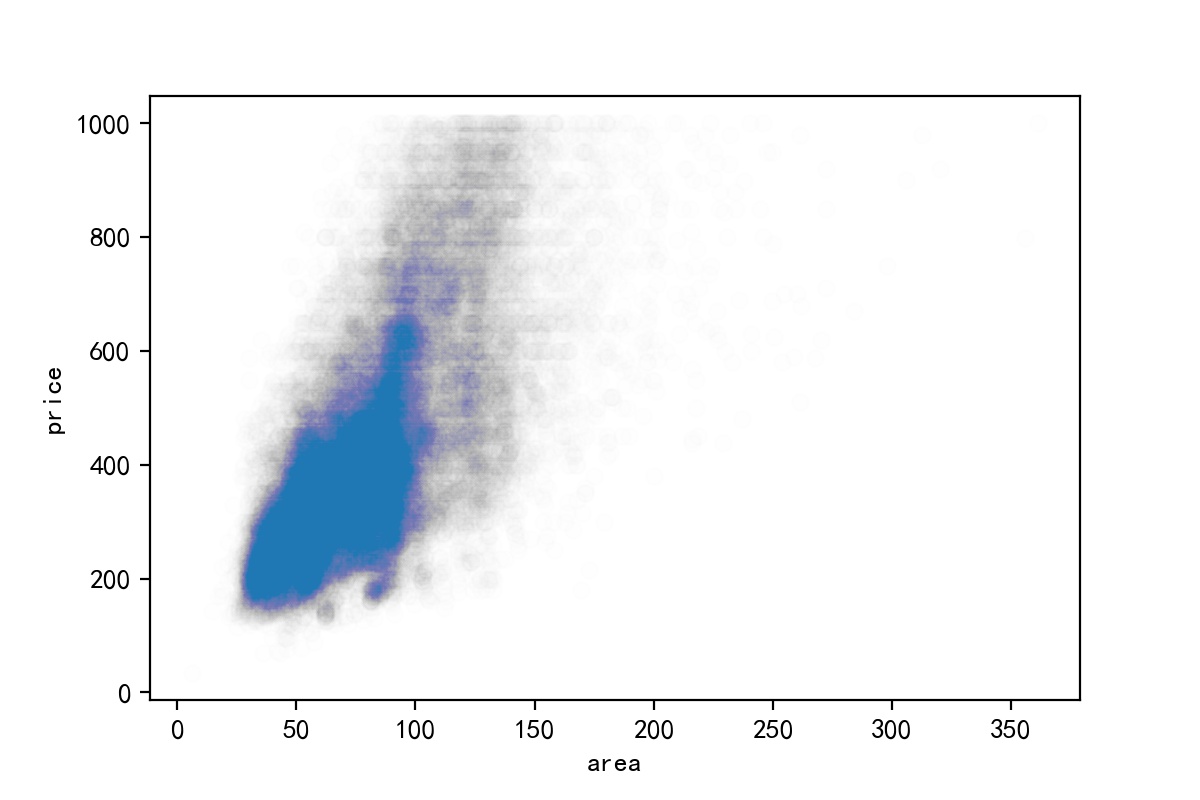


图6 总价-装修类型



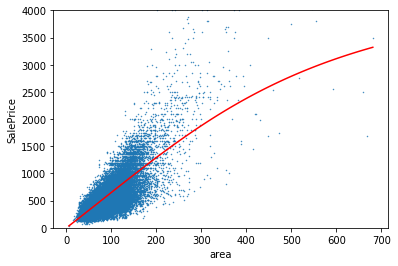
塔楼、精装房总价更高。

图7 总价-面积



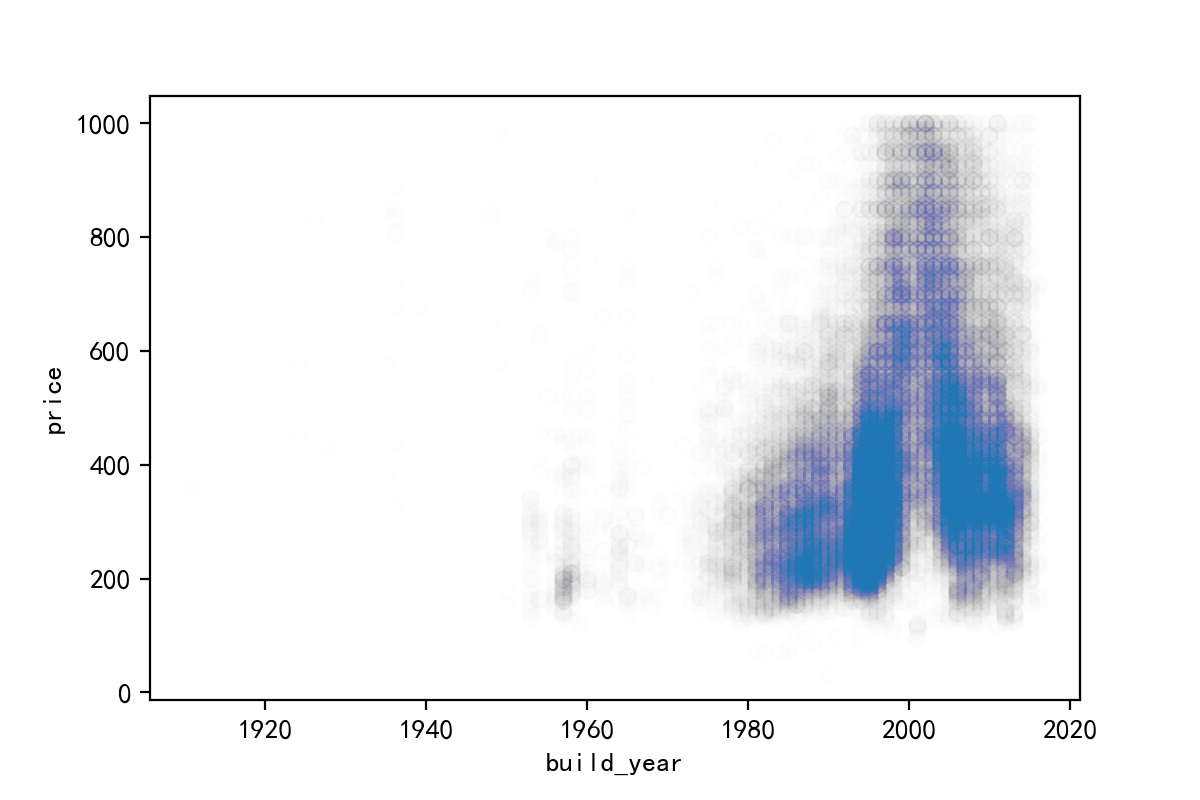
比值(单价)在一定区间内，大部分房屋在50-150平方米。

一般认为面积是房价在主要影响因素之一，这里使用深度学习对总价-面积关系进行拟合，得到结果如下(注意两张图的坐标范围不同)：



可见总价与面积基本成正比，而面积大于400平方米后，总价增幅放缓。

图8 总价-建造年份



建于1990年后的房价显著升高。建造后25年内，房屋价格基本不受时间影响。

图9 面积-室

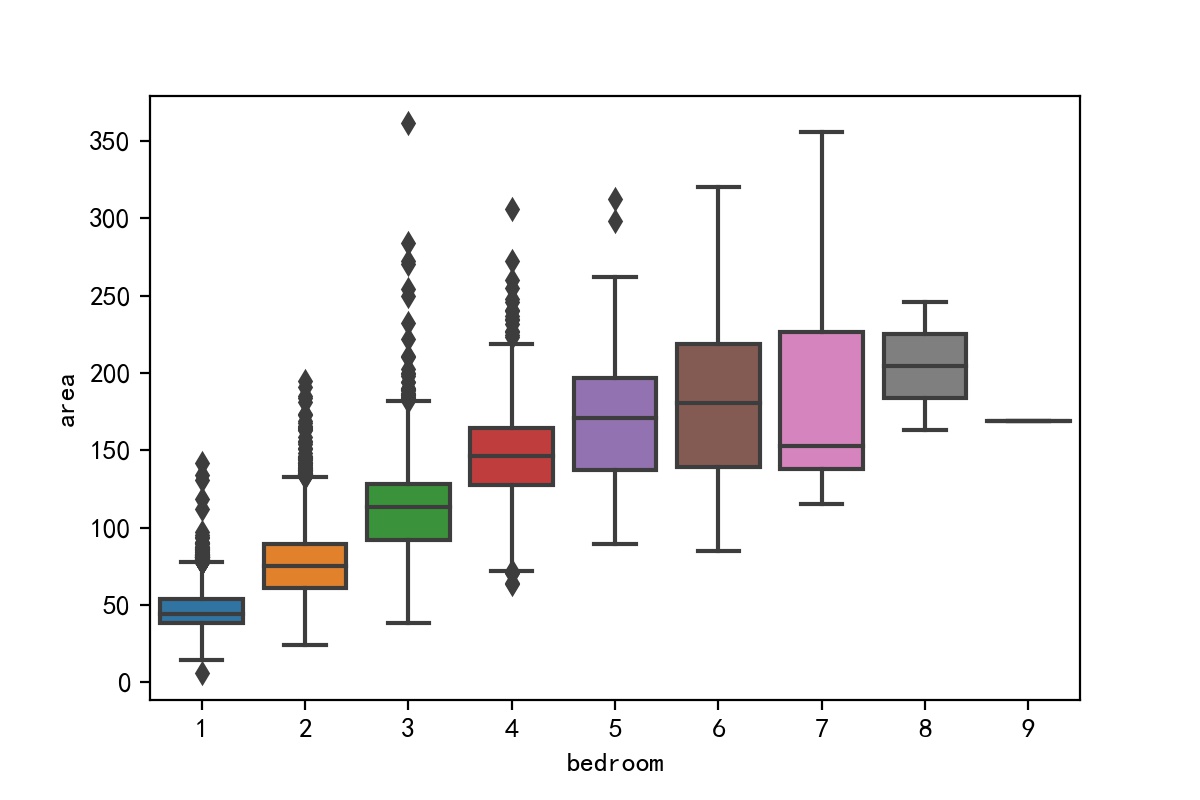


图10 面积-厅

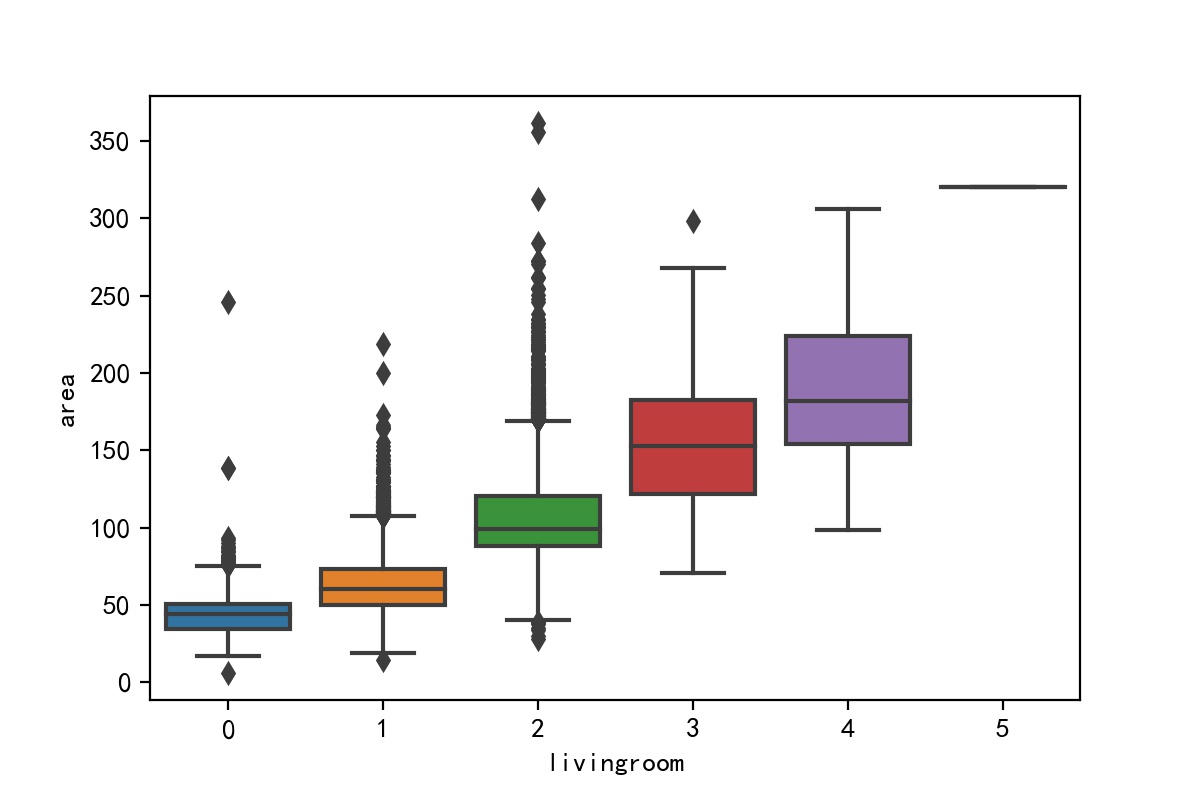
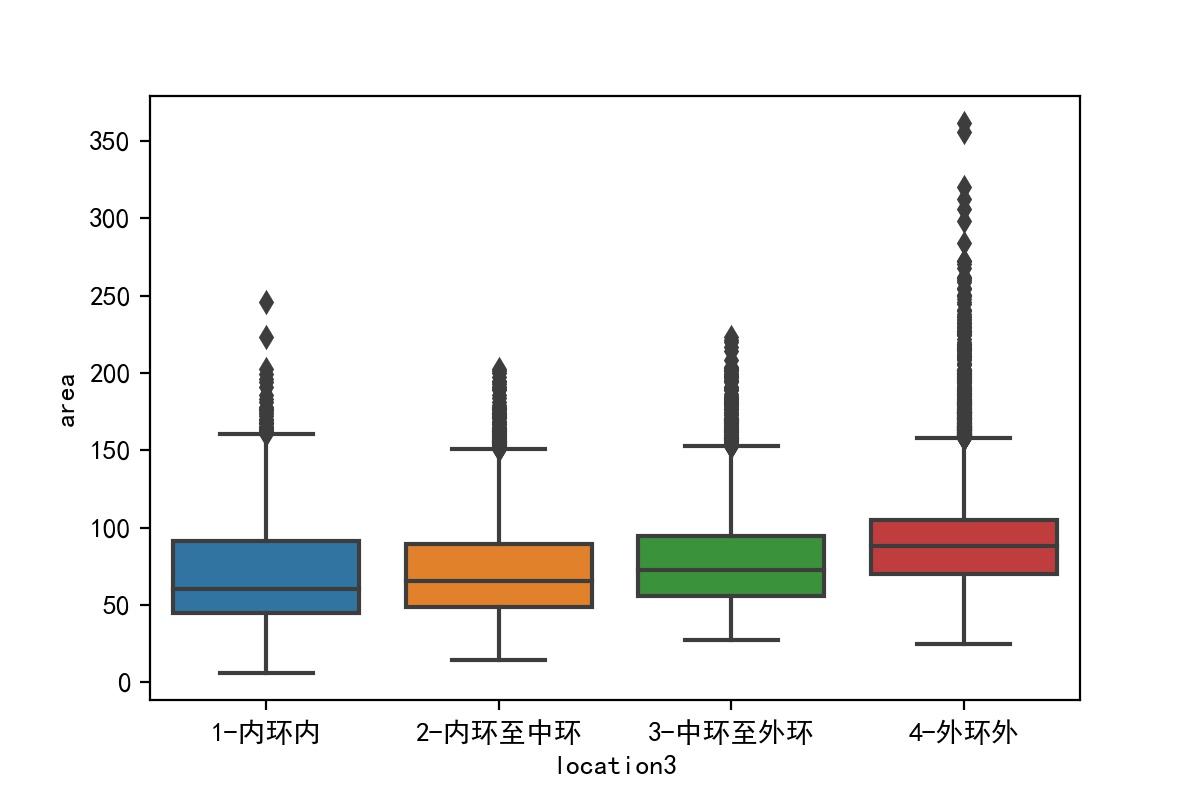
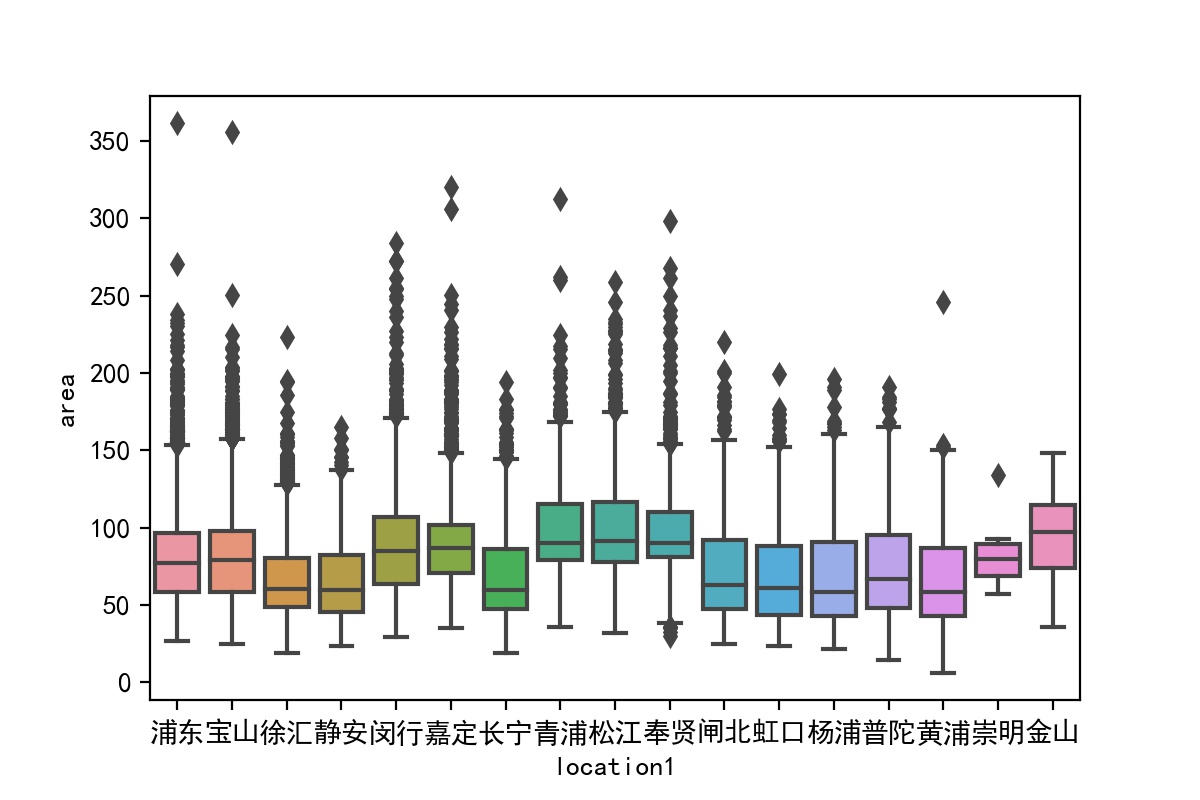


图11 面积-地段



外环外房屋面积较大。

图12 面积-区划



郊区房屋面积较大。

图13 面积-建筑类型

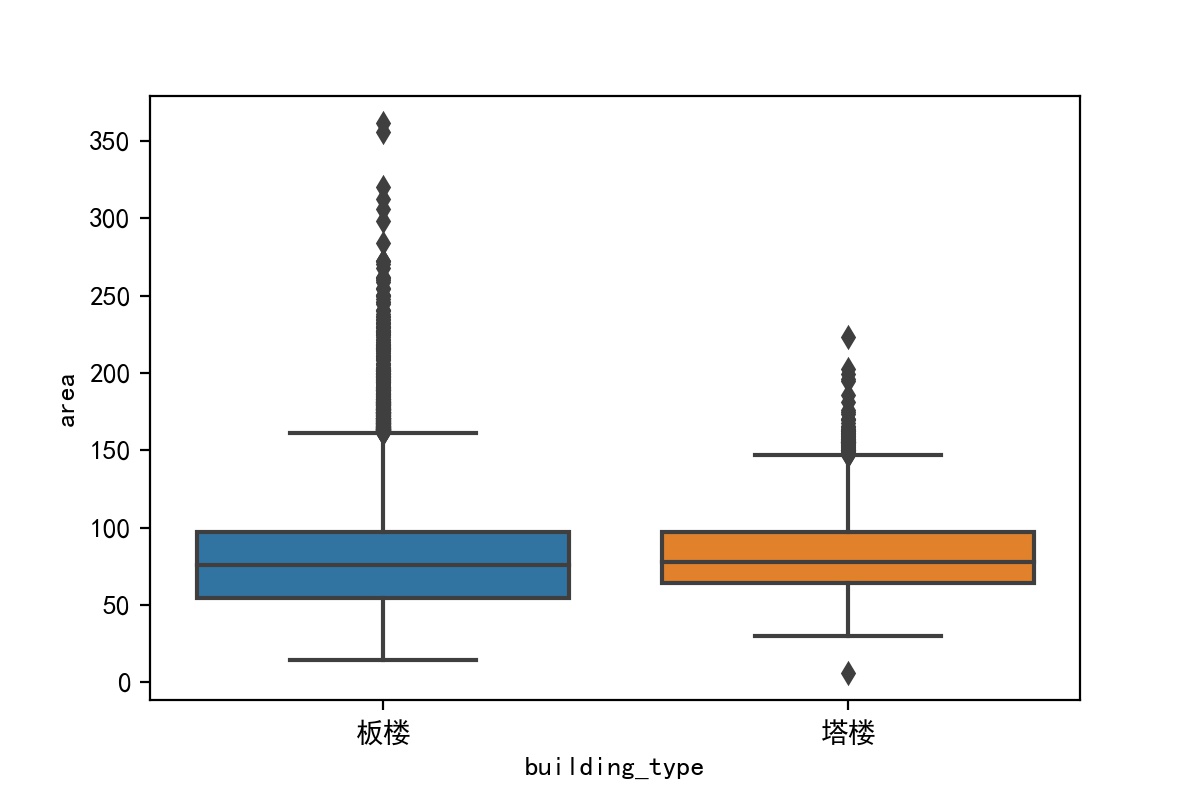
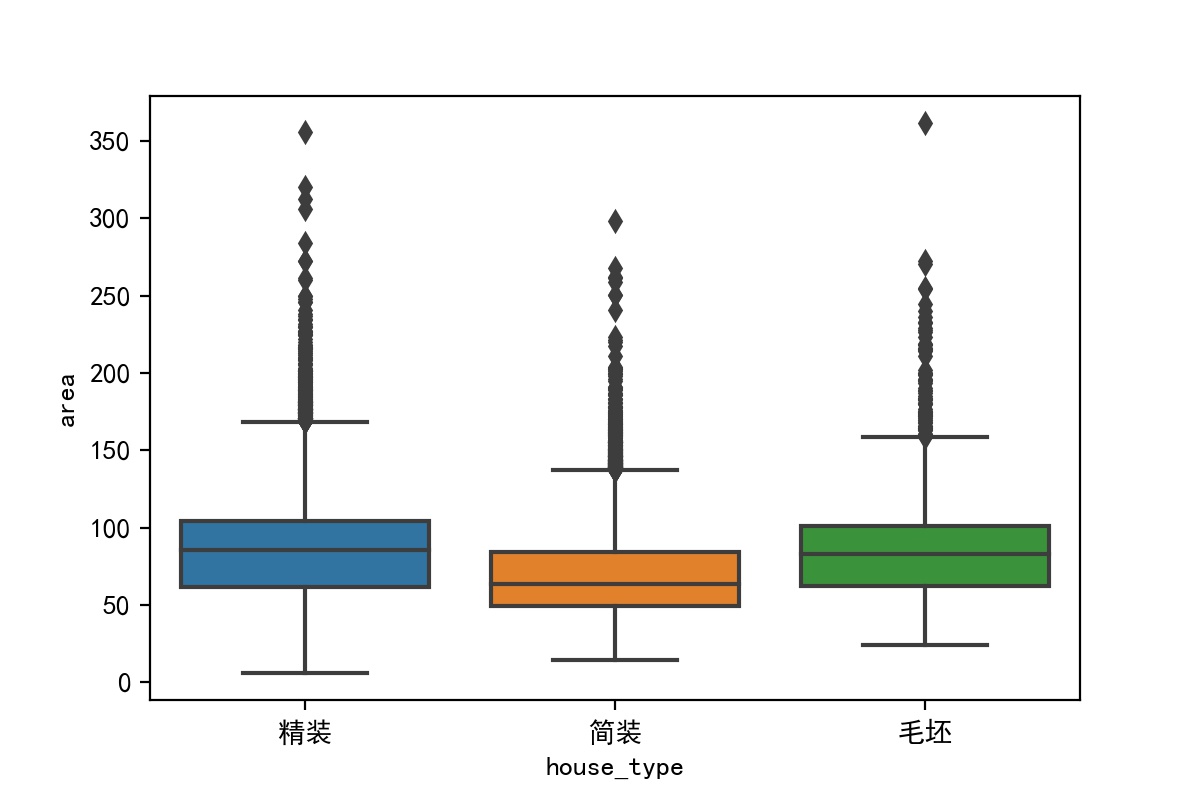
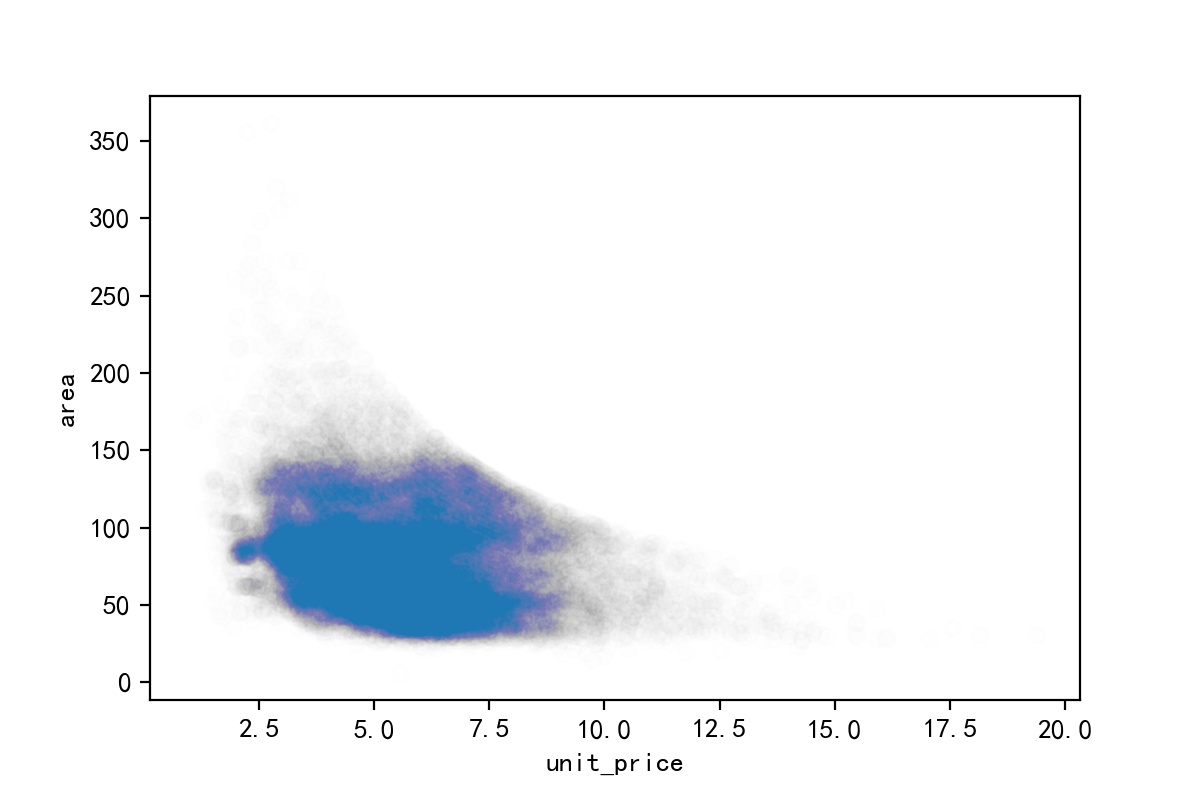


图14 面积-装修类型



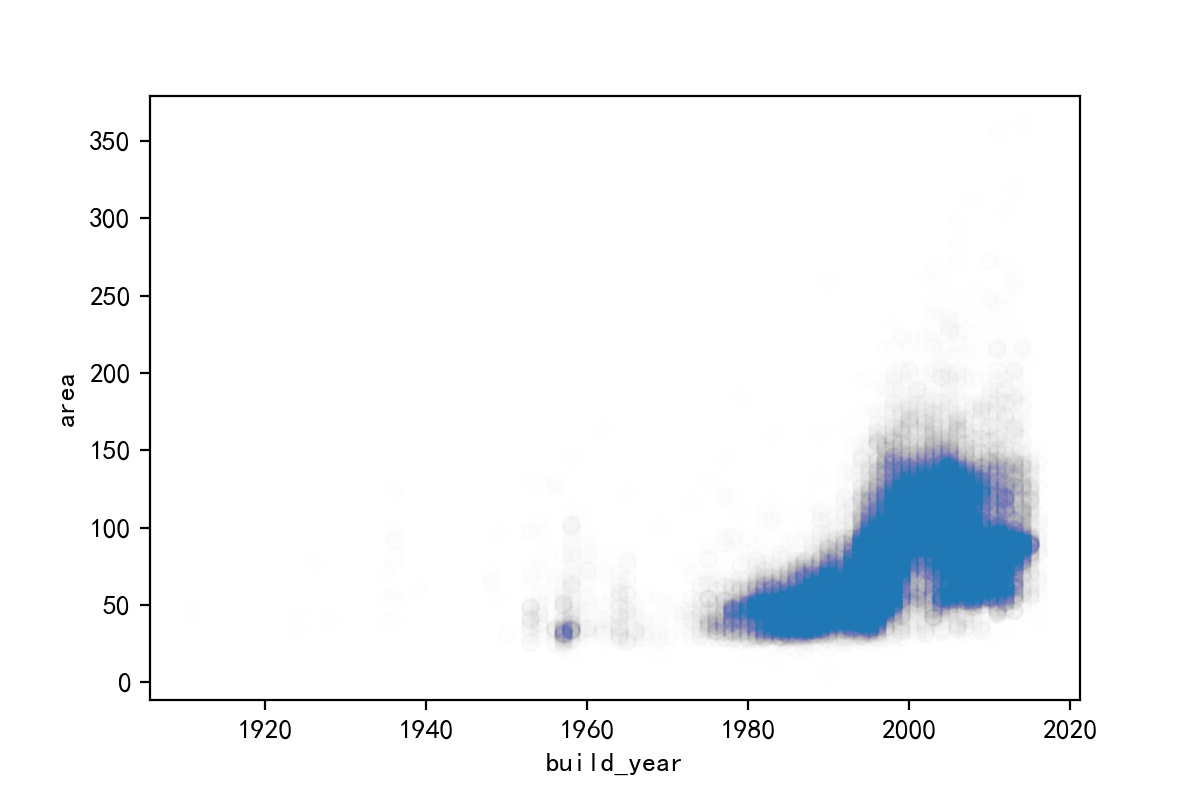
房屋面积与建筑类型、装修类型无关。面积小的房屋更倾向简装。

图15 面积-单价



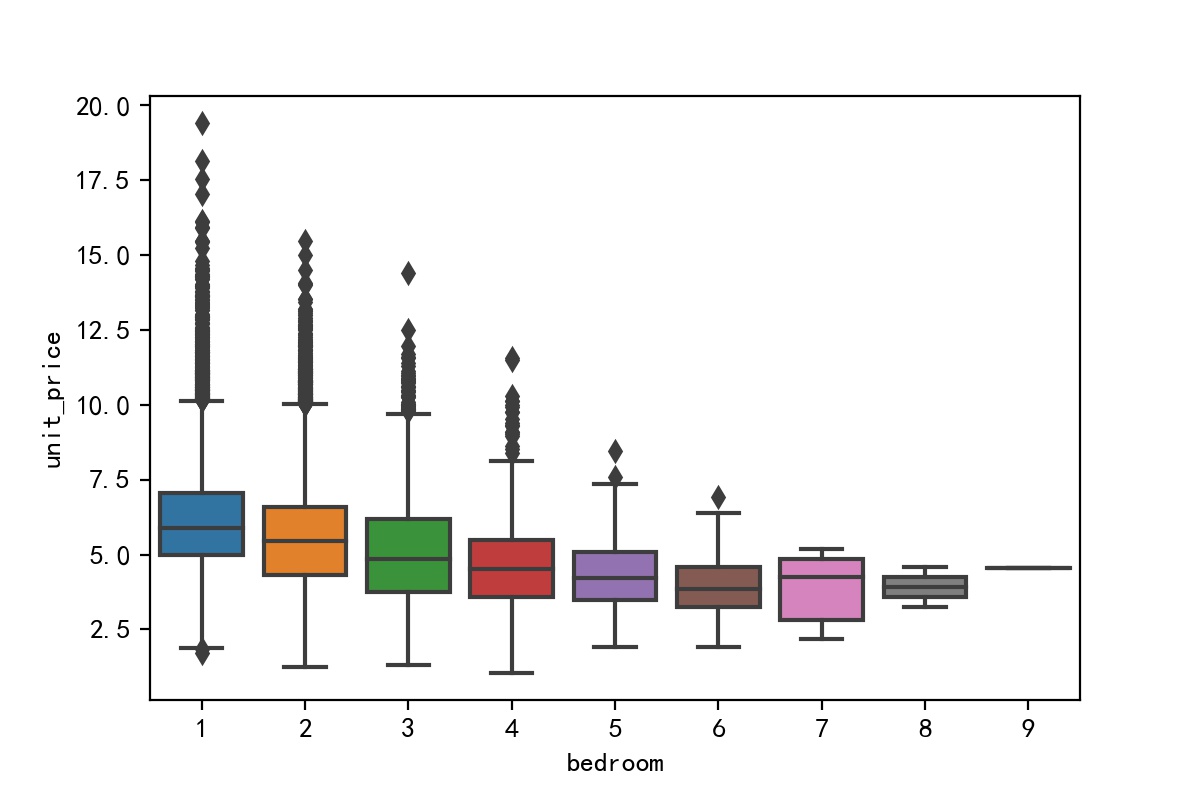
部分呈现负相关，面积大的房屋单价总体上低，反之亦然。

图16 面积-建造年份



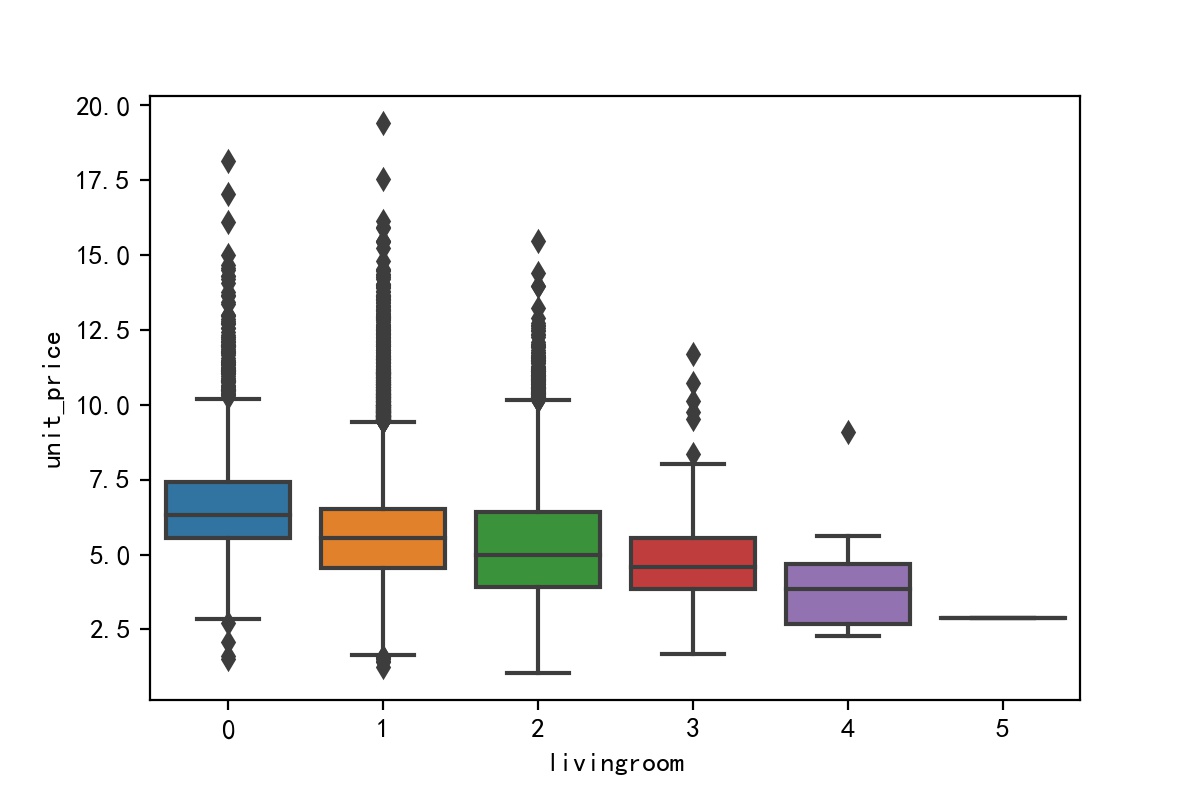
近40年房屋面积提升明显，生活质量改善。

图17 单价-室



卧室数量越多，单价越低。

图18 单价-厅



厅数量越多，单价越低。

图19 单价-地段

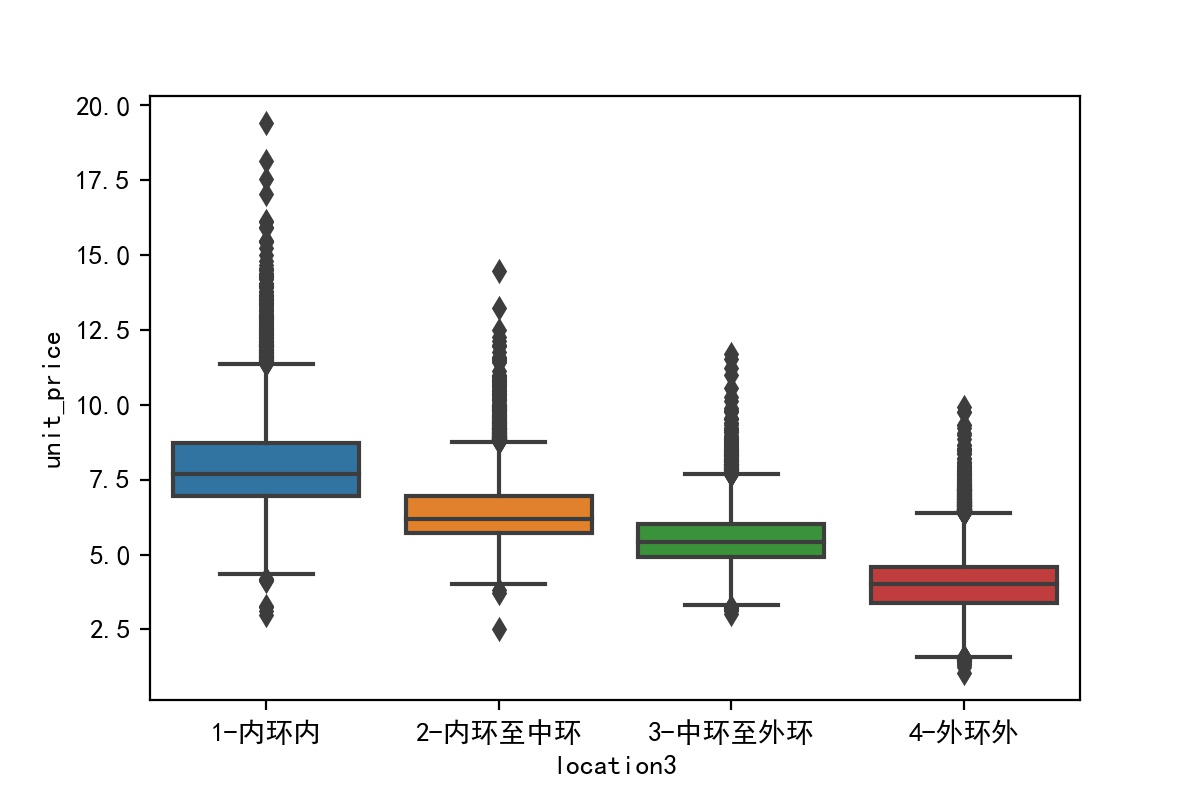
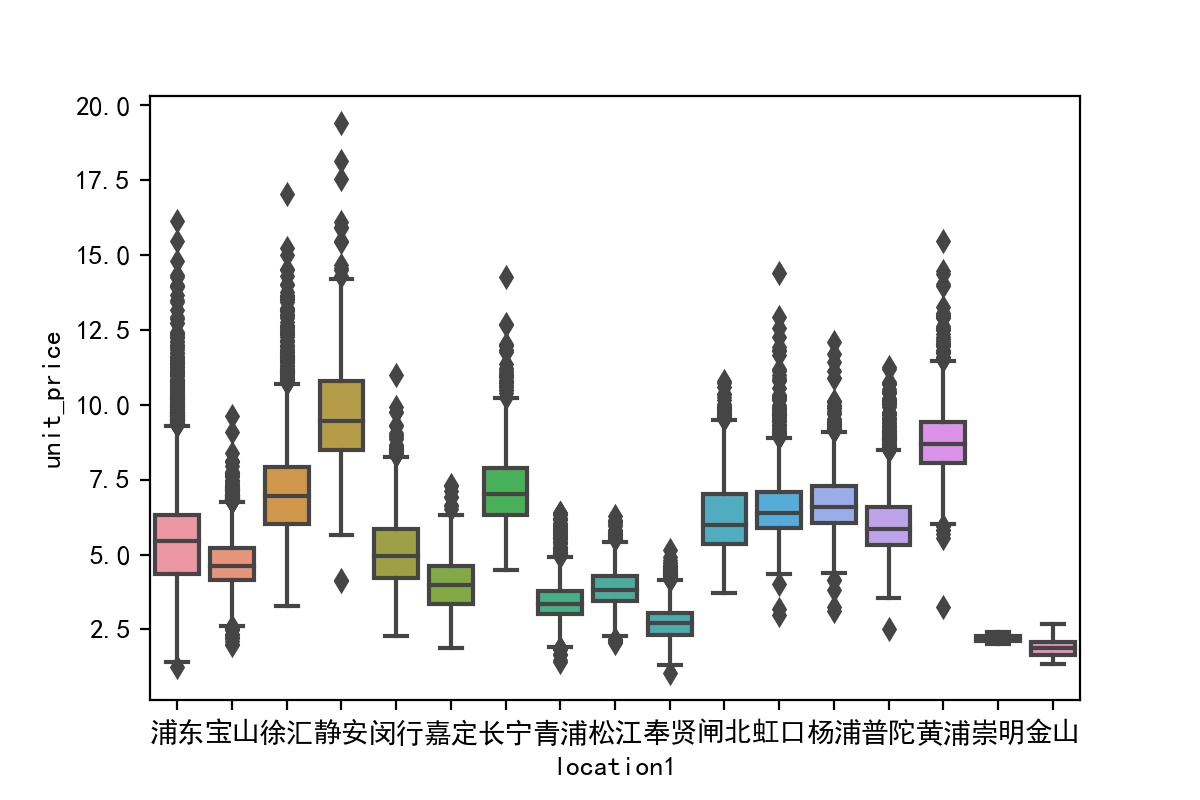
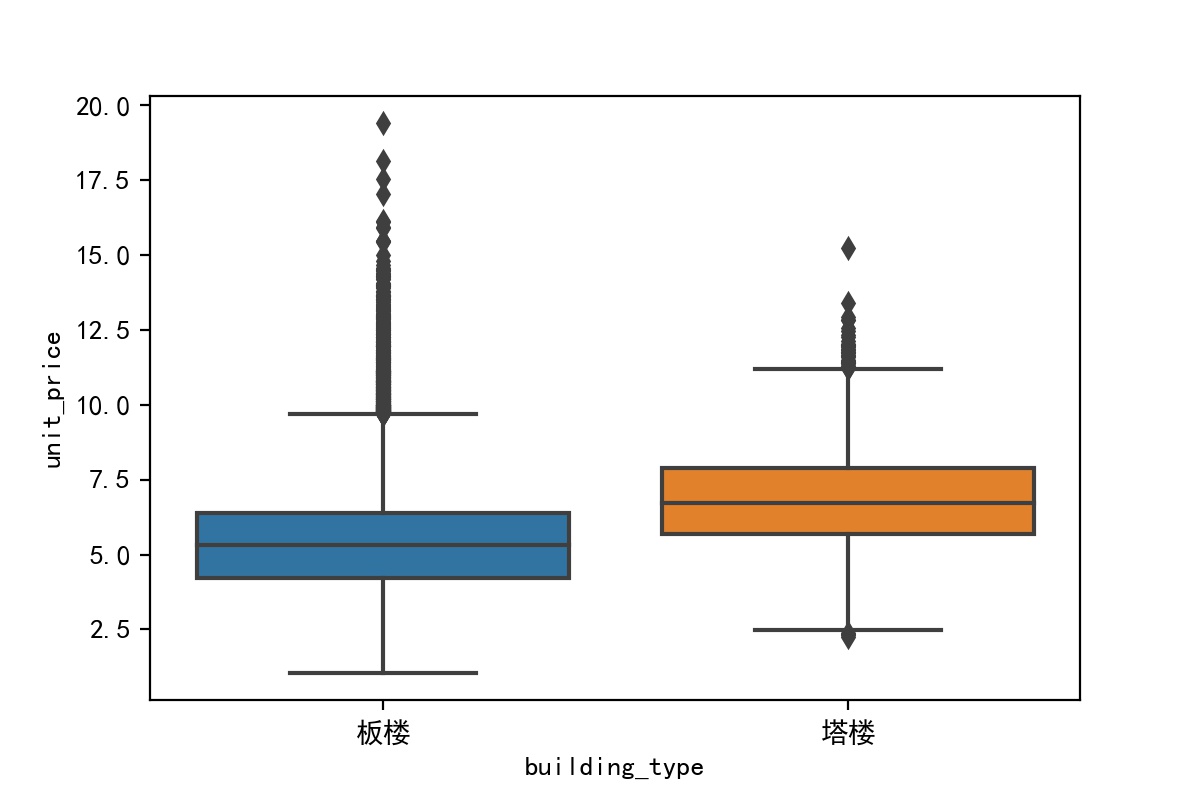


图20 单价-区划



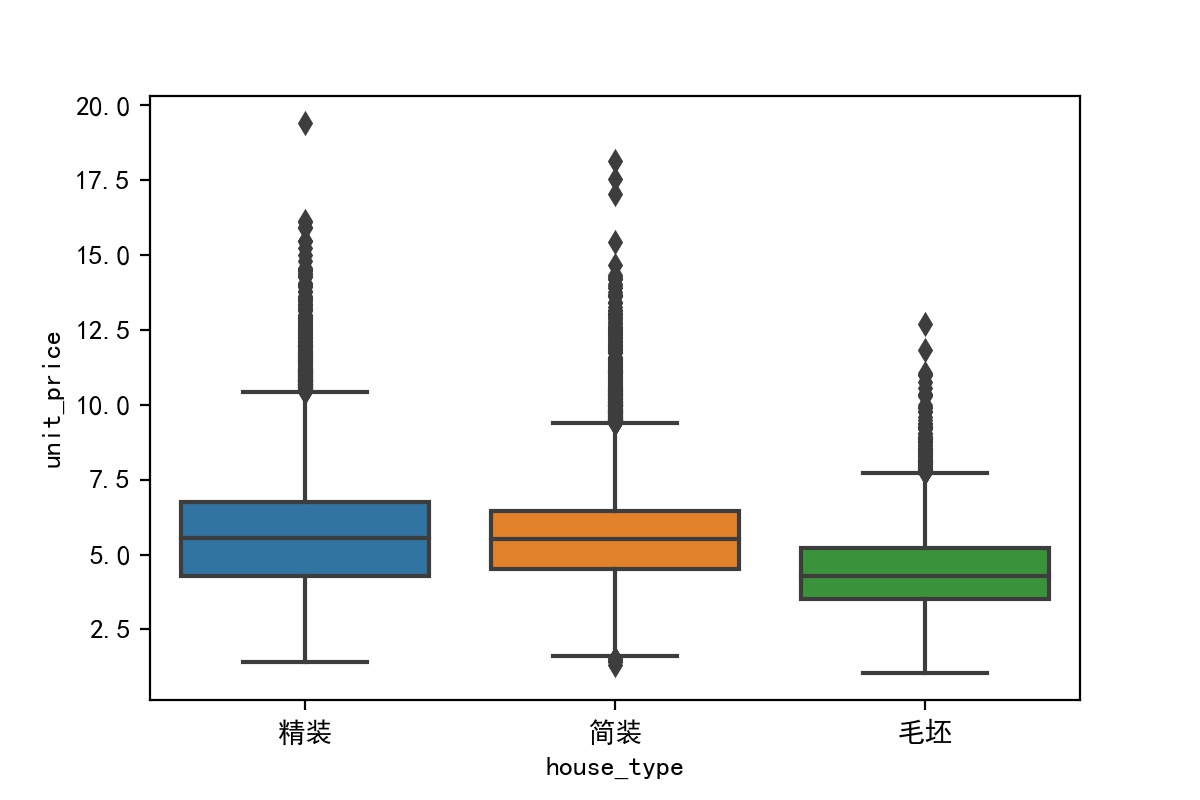
距市中心越近，单价越高。崇明金山房价远低于其他区。差距大于总价。

图21 单价-建筑类型



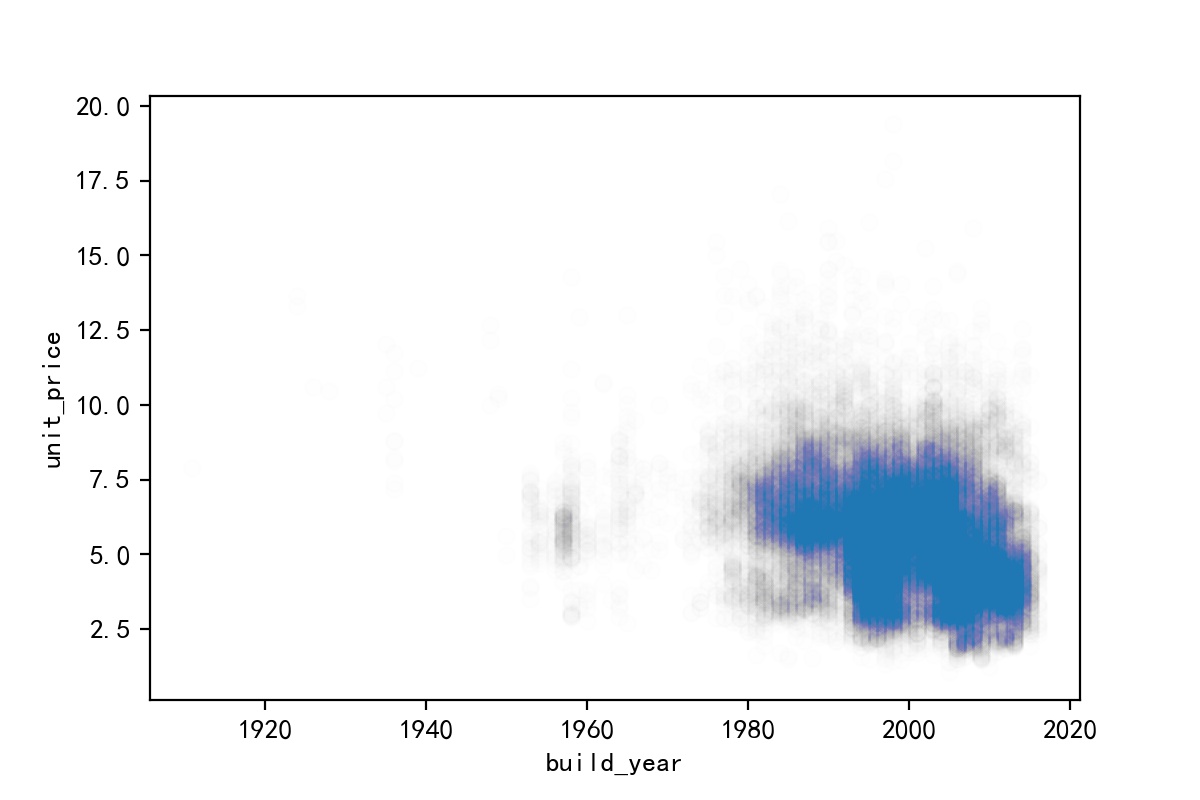
塔楼单价明显高于板楼。

图22 单价-装修类型



精装简装单价差距不大，毛培单价较低。

图23 单价-建造年份



20年前建造的房屋单价较高。可能是因为面积较小而地段比较好。

图24 建造年份-室

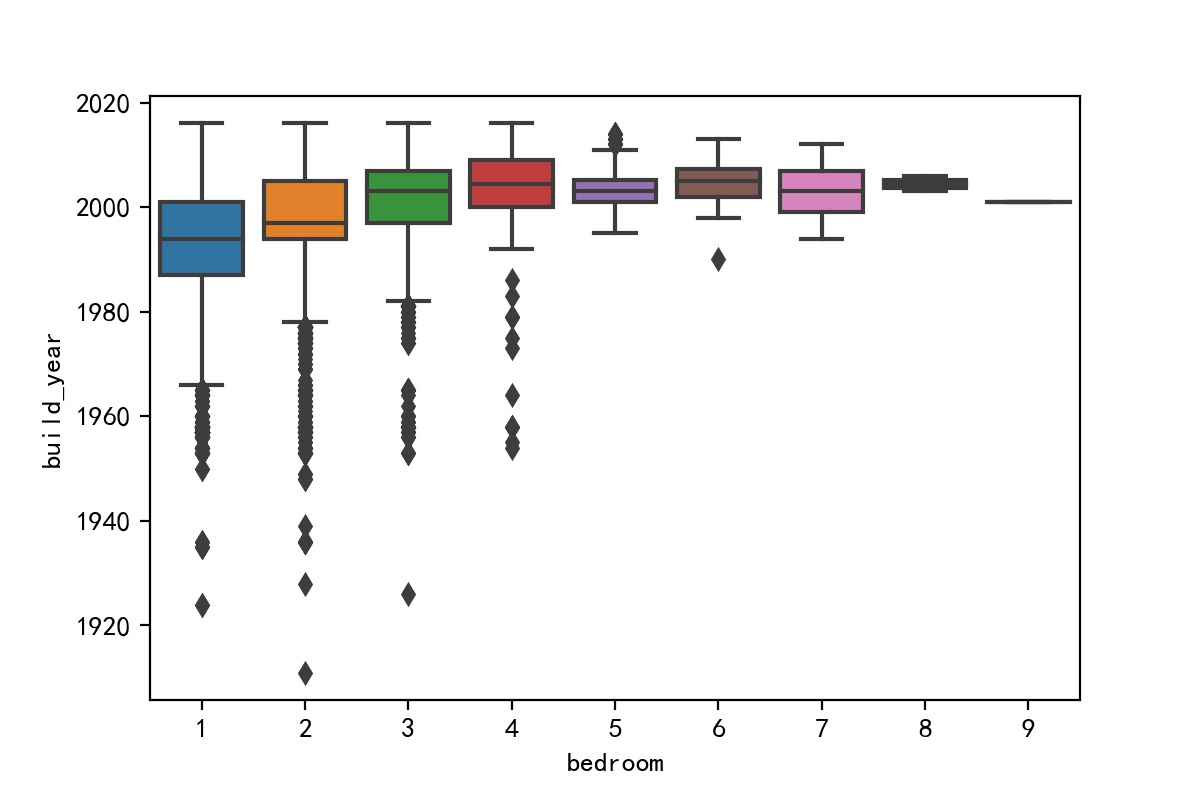


图25 建造年份-厅

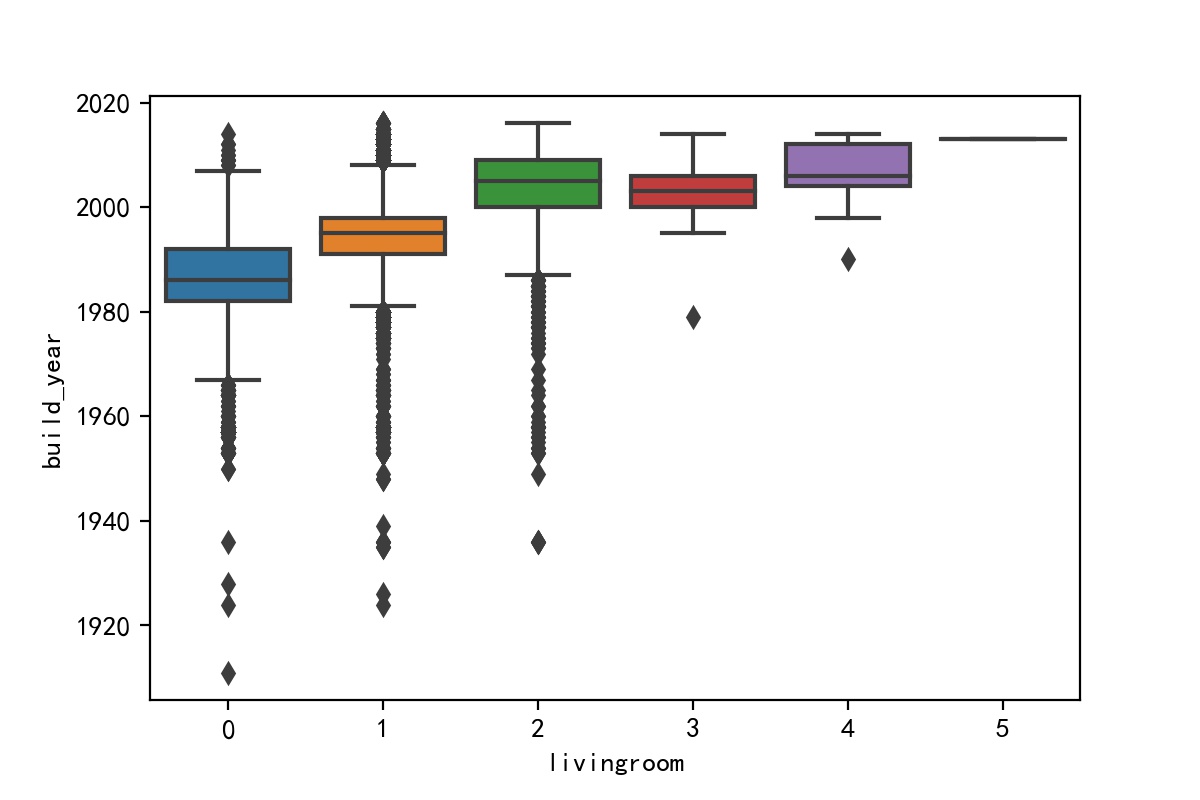


图26建造年份-地段

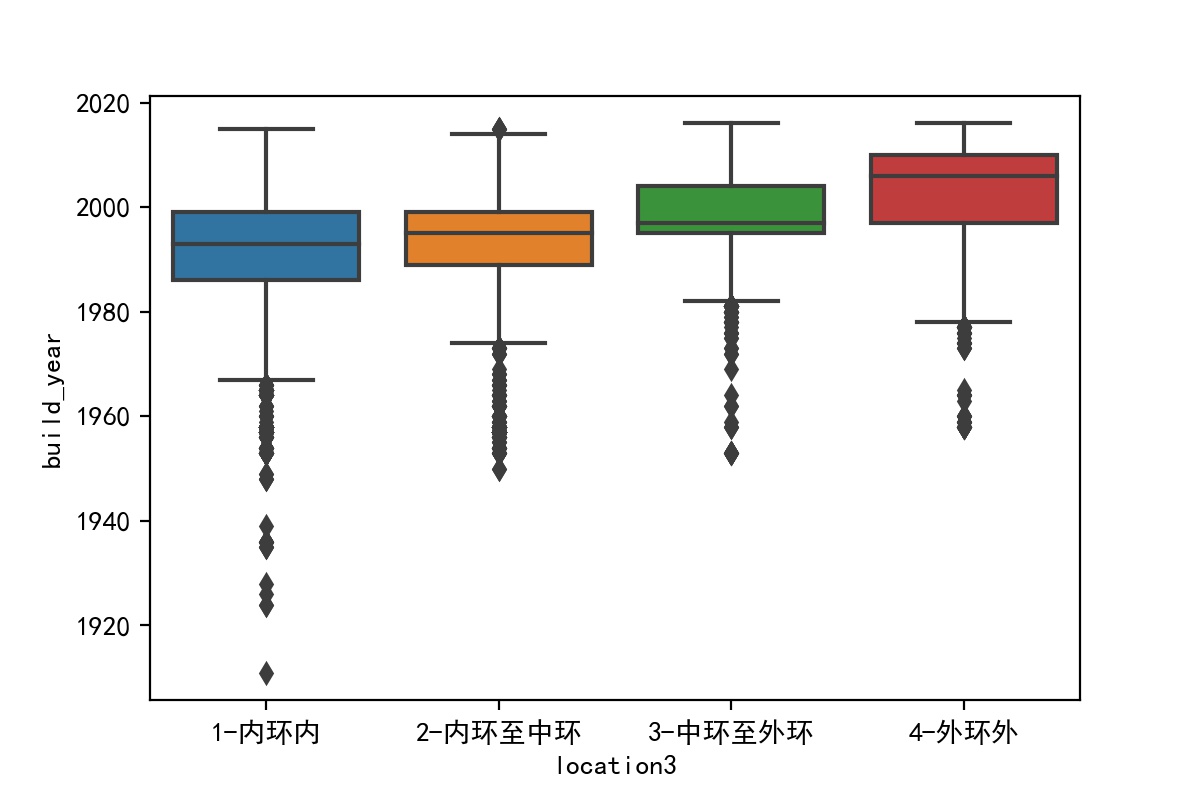
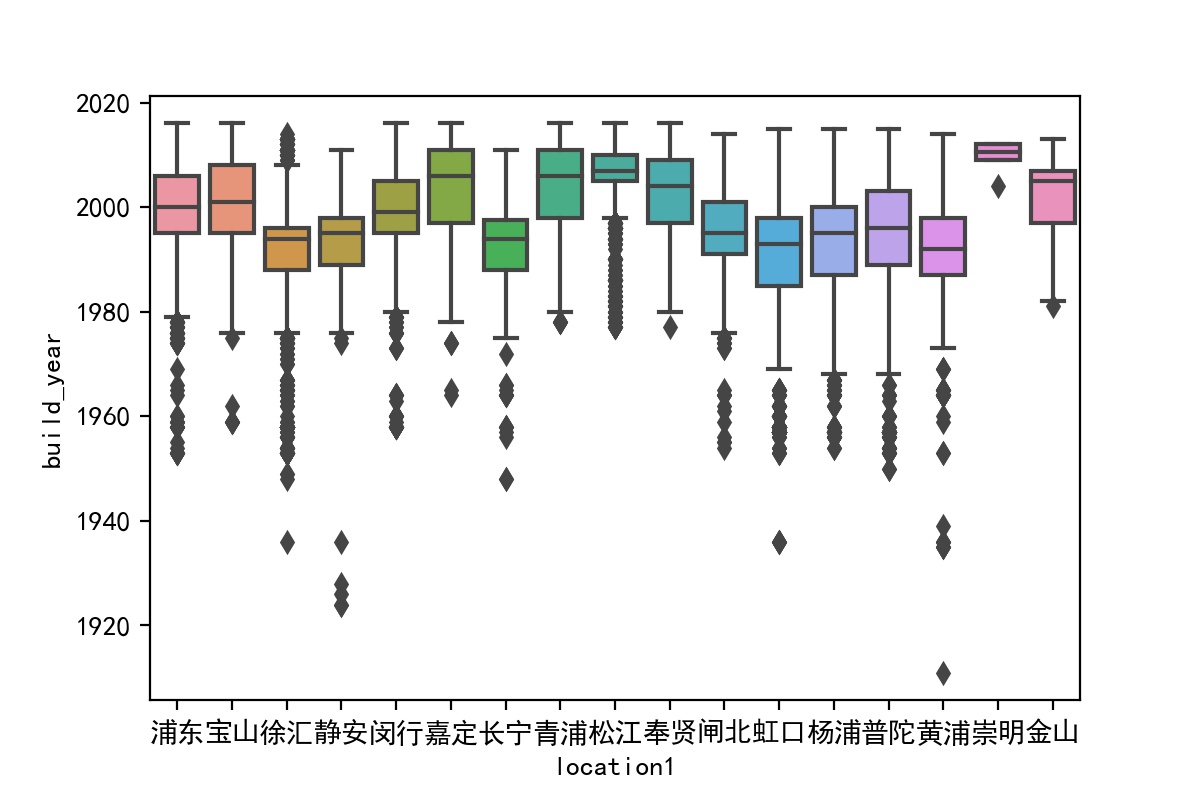


图27建造年份- 区划



市区房屋建造早于郊区。

图28建造年份-建筑类型

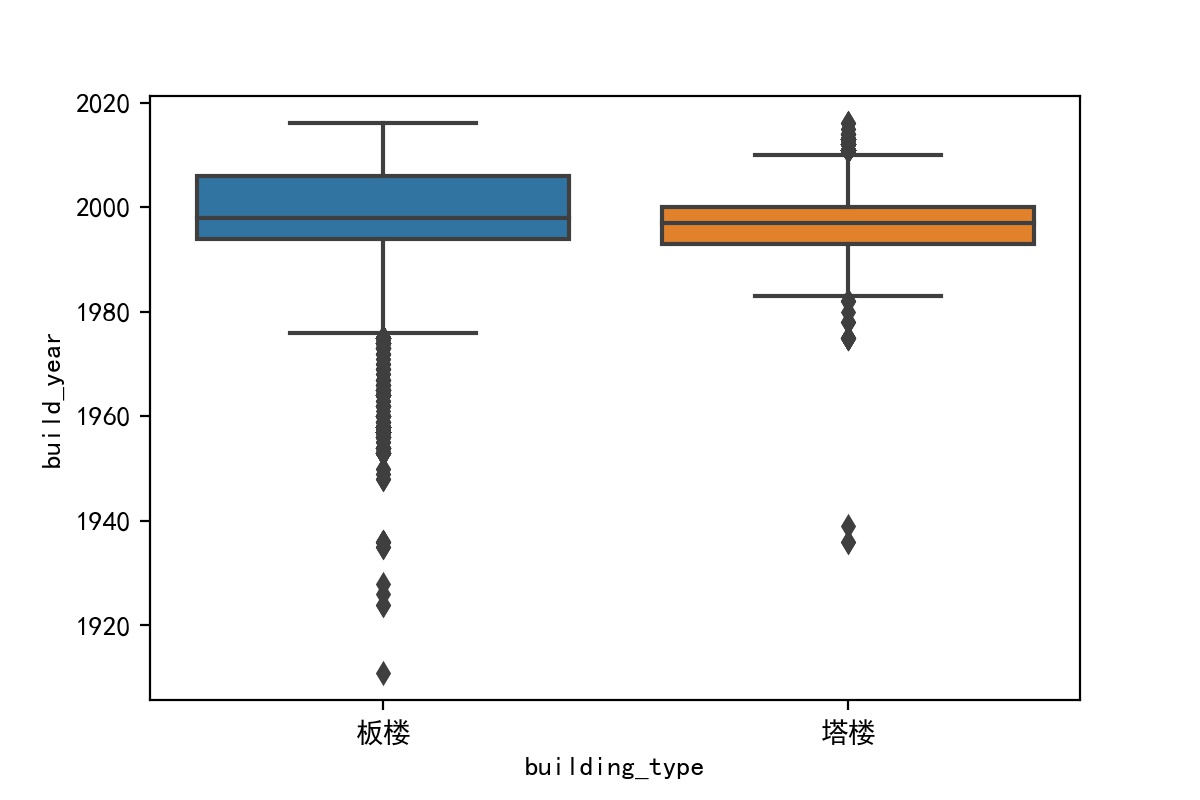
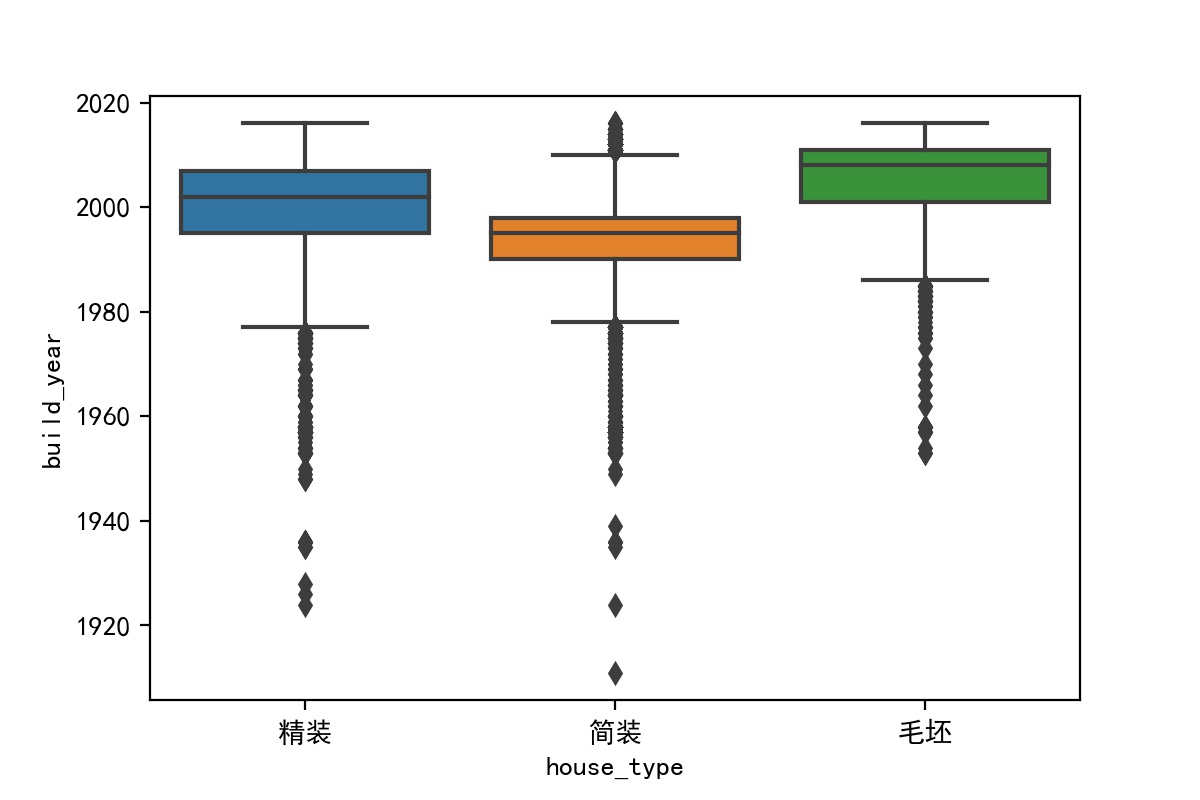
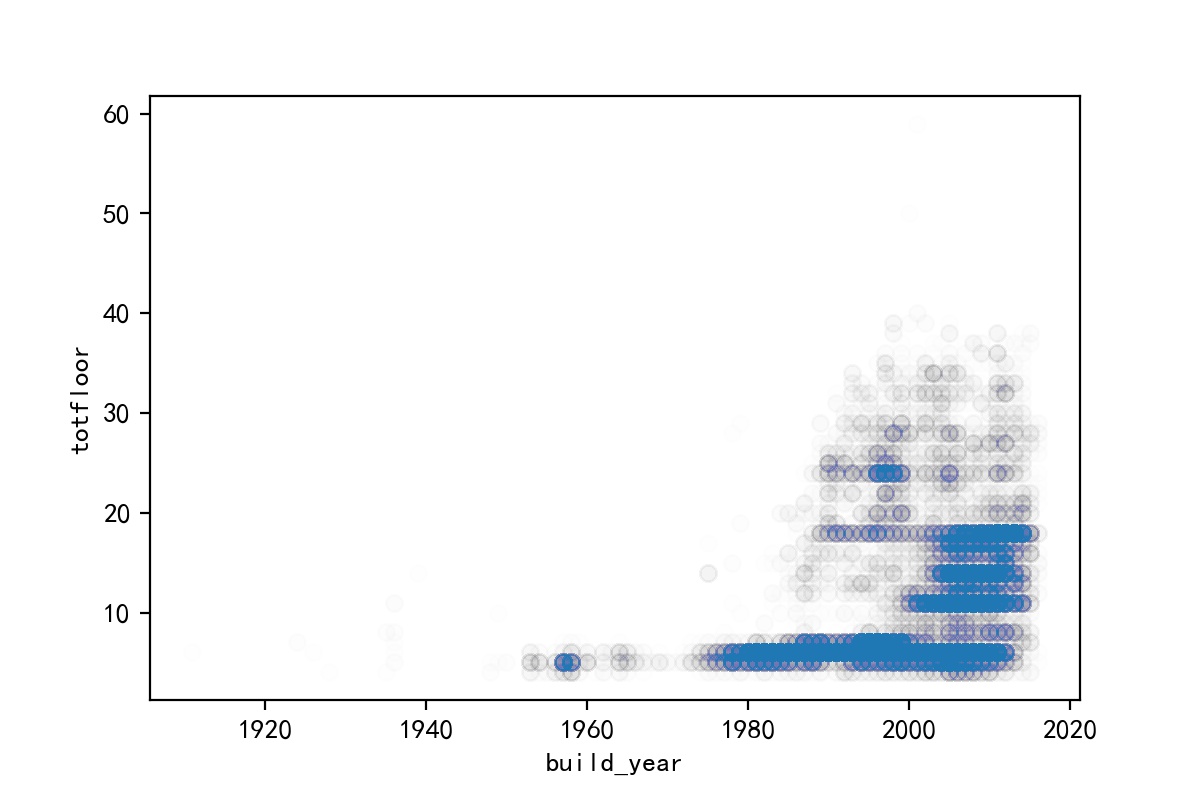


图29建造年份-装修类型



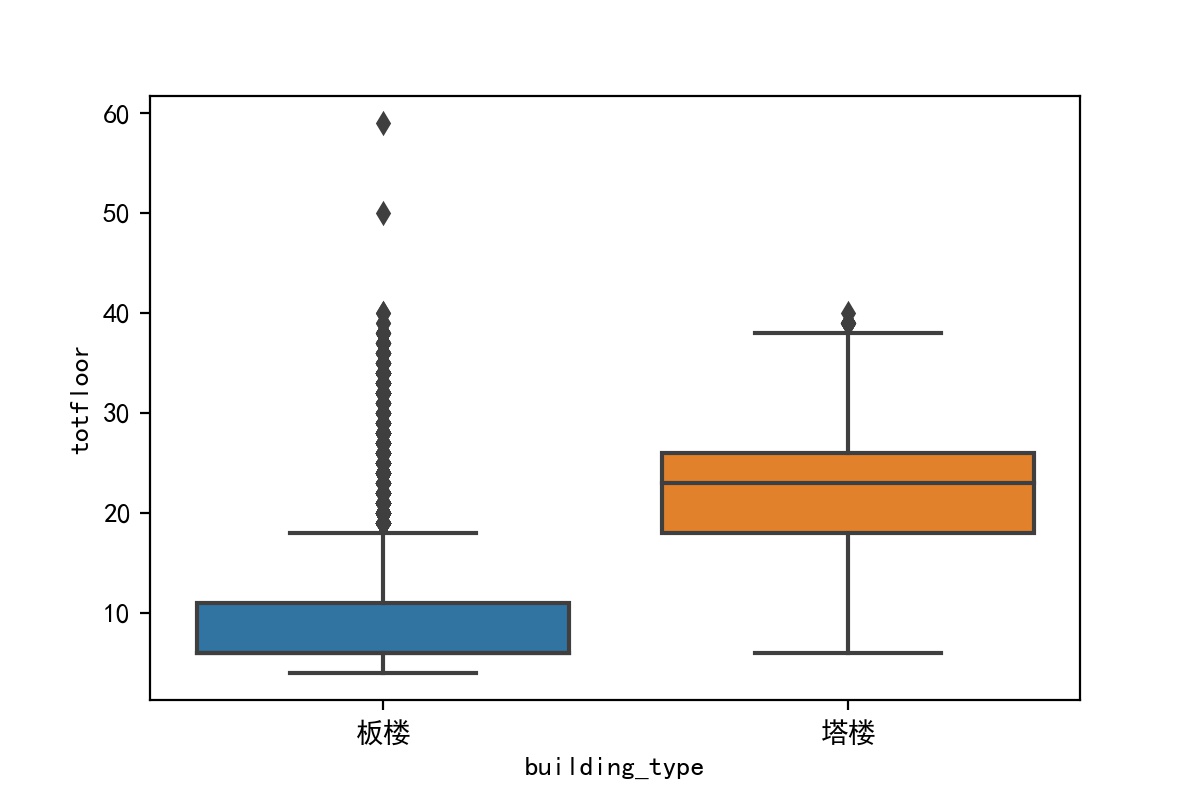
建造较早的多为简装，建造较晚的多为毛坯。

图30 楼高-建造年份



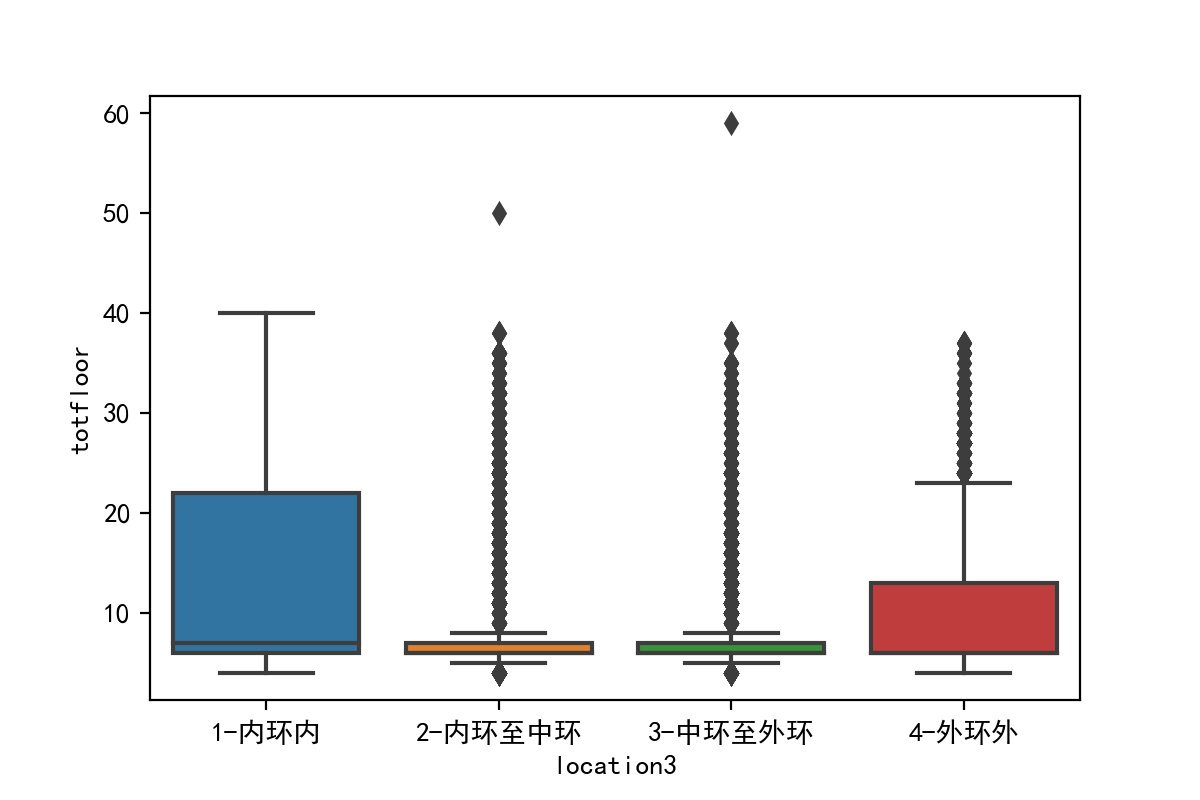
2000年后开始出现大量高住宅楼。

图31 楼高-建筑类型



塔楼普遍高于板楼。

图32 楼高-地段



绝大多数为6层，市中心和郊区高楼较多。

图33 楼高-区划

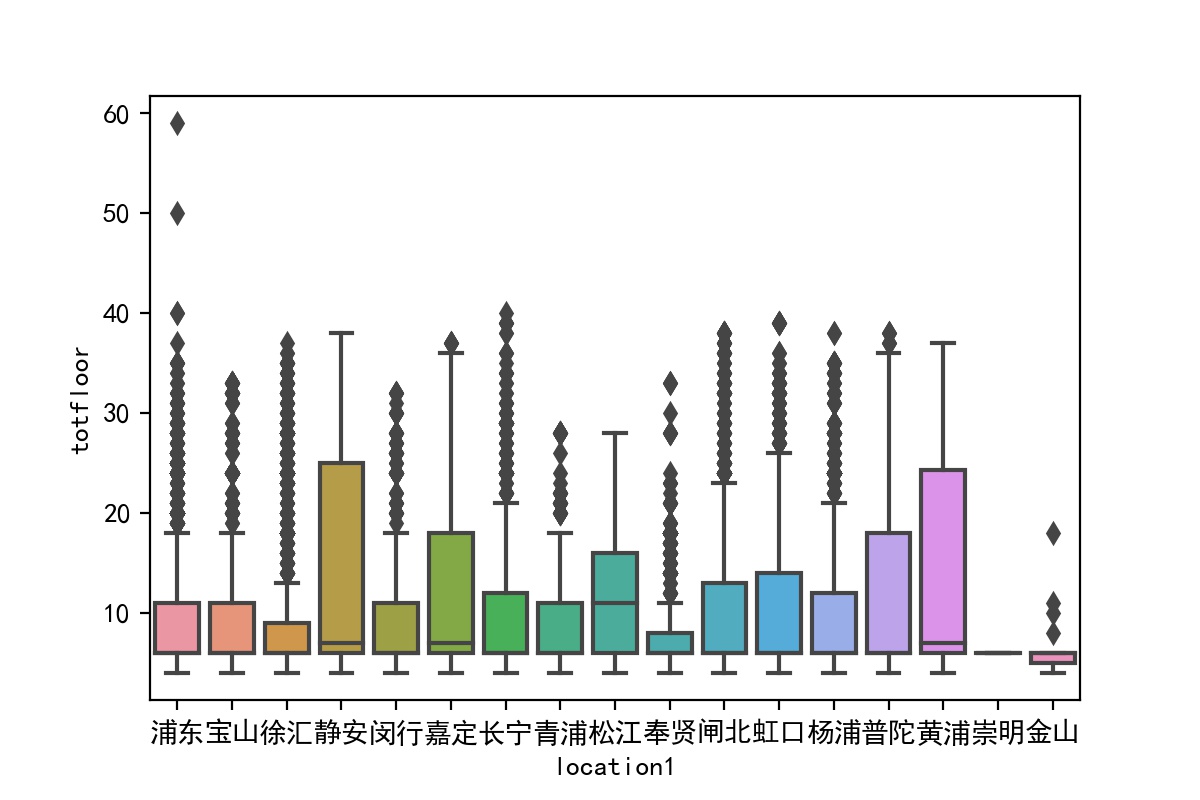
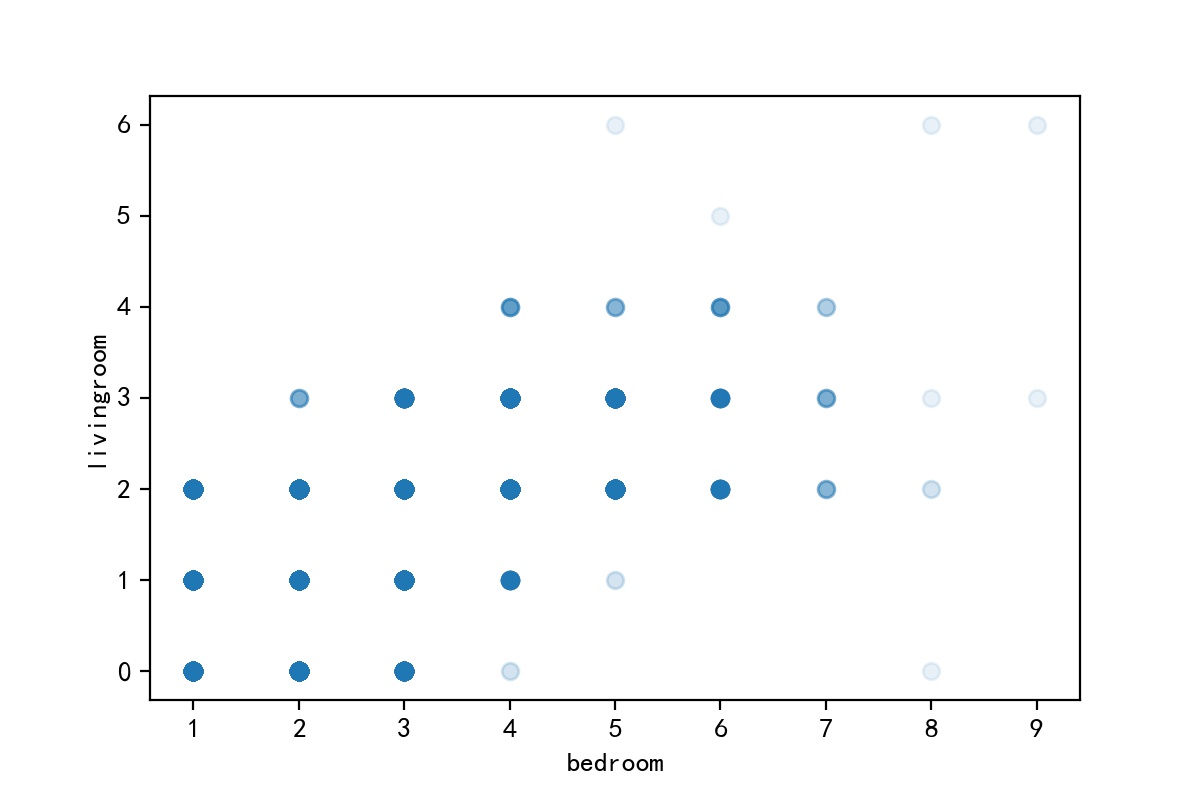


图34 室-厅



存在八室零厅的失衡情况： sh.lianjia.com/ershoufang/107002272897.html



1. 总结

本文以图表的形式展现了住房总价、厅室数量、面积、装修、建筑类型、所属区划、地段、建造年份之间的关系。大部分图表内容与一般认识相符，本文用更加直观的方式展现了各个因素之间的相互关系。

人民住房面积不断扩大、住房支出的提升体现了生产发展和社会进步。

从图表内容看，面积和地段是房屋总价的主要因素，装修、建筑类型、建造年份等是次要因素。

现代信息技术使得每个人都可以获得大量房产信息，在大量的房产信息中匹配最符合自己需求的资源，从而使得住房资源得到更加充分的利用。

附：

1. 采集及解释绘图程序（带少量注释，国外网站打开可能较慢）：

<https://github.com/zyayoung/lianjia/blob/master/phrase.ipynb>

1. 数据表（csv格式，逗号分隔，下载后excel可能可以打开）：

<https://github.com/zyayoung/lianjia/blob/master/data.csv>

1. 100张自动生成的原始统计图：

<https://github.com/zyayoung/lianjia/tree/master/plots>

1. 102张自动生成的原始统计图（房价小于一千万）：

<https://github.com/zyayoung/lianjia/tree/master/plots_small>

1. 链家robots.txt爬虫协议

<https://www.lianjia.com/robots.txt>

本文爬取地址sh.lianjia.com/\*，未违反爬虫协议。