链家二手房的数据探索

张矜 2023281051017

目录

1 你的主要发现

1. 随房屋面积的上升，房屋总价明显上升。且房屋建筑面积越大时，房屋总价的置信区间明显变宽。其中相同房屋面积的情况下，房本满 2 年的房屋总价，在建筑面积 135 以上的区间，整体趋势高于房本未满 2 年的房型。

2. 在本次数据范围的 8 类不同房屋主要朝向下，房价与房屋面积关系拟合曲线整体呈现上升趋势，除了朝向东和朝向西的拟合曲线出现了一定程度的下降区间。

3. 探索各房屋建筑形式中装修状况的比例，主流趋势为：装修状况为 “精装” 的房屋所占比例最大，占比约为 50%，装修状况为 “简装” 的房屋所占比例次之。

2 数据介绍

本报告**链家**数据获取方式如下：

报告人在 2023 年 9 月 12 日获取了链家武汉二手房网站数据。

- 链家二手房网站默认显示 100 页，每页 30 套房产，因此本数据包括 3000 套房产信息；

- 数据包括了页面可见部分的文本信息，具体字段及说明见作业说明。

说明：数据仅用于教学；由于不清楚链家数据的展示规则，因此数据可能并不是武汉二手房市场的随机抽样，结论很可能有很大的偏差，甚至可能是错误的。

3 数据概览

数据表（lj) 共包括 property_name, property_region, price_ttl, price_sqm, bedrooms, livingrooms, building_area, directions1, directions2, deco-

表 1: 武汉链家二手房

property_name	property_region	price_ttl	price_sqm	bedrooms	livingrooms	building_a
南湖名都 A 区	南湖沃尔玛	237.0	18709	3	1	126
万科紫悦湾	光谷东	127.0	14613	3	2	86
东立国际	二七	75.0	15968	1	1	46
新都汇	光谷广场	188.0	15702	3	2	119
保利城一期	团结大道	182.0	17509	3	2	103
加州橘郡	庙山	122.0	10376	3	2	117
省建筑五公司西区	光谷广场	99.0	12346	2	1	80
保利上城东区	白沙洲	193.8	16336	3	2	118
石化大院	中南丁字桥	325.0	32631	4	1	99
阳光花园	杨汉湖	192.0	17403	3	2	110

ration, property_t_height, property_height, property_style, followers, near_subway, if_2y, has_key, vr 等 18 个变量, 共 3000 行。表的前 10 行示例如下：

各变量的简短信息：

```
## Rows: 3,000
## Columns: 18
## $ property_name     <chr> "南湖名都A区", "万科紫悦湾", "东立国际", "新都汇", "~
## $ property_region   <chr> "南湖沃尔玛", "光谷东", "二七", "光谷广场", "团结大~
## $ price_ttl         <dbl> 237.0, 127.0, 75.0, 188.0, 182.0, 122.0, 99.0, 193.8~
## $ price_sqm         <dbl> 18709, 14613, 15968, 15702, 17509, 10376, 12346, 163~
## $ bedrooms          <dbl> 3, 3, 1, 3, 3, 3, 2, 3, 4, 3, 5, 3, 4, 3, 3, 2, 3, 4~
## $ livingrooms       <dbl> 1, 2, 1, 2, 2, 2, 1, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2~
## $ building_area     <dbl> 126.68, 86.91, 46.97, 119.73, 103.95, 117.59, 80.19,~
## $ directions1       <chr> "南", "南", "南", "北", "东南", "南", "南", "南", "~
## $ directions2       <chr> "北", NA, NA, "东", NA, "北", NA, "北", "北", "北", ~
## $ decoration        <chr> "精装", "精装", "简装", "精装", "简装", "精装", "简~
## $ property_t_height <dbl> 17, 28, 18, 32, 34, 34, 7, 34, 5, 7, 25, 32, 8, 31, ~
## $ property_height   <chr> "中", "中", "低", "高", "中", "低", "低", "中", "低"~
```

```
## $ property_style    <chr> "塔楼", "板楼", "塔楼", "塔楼", "板塔结合", "板楼", ~
## $ followers         <dbl> 3, 1, 3, 2, 3, 1, 0, 0, 2, 0, 0, 0, 10, 0, 0, 1, 0, ~
## $ near_subway       <chr> "近地铁", NA, "近地铁", "近地铁", NA, NA, "近地铁", ~
## $ if_2y             <chr> NA, "房本满两年", NA, "房本满两年", "房本满两年", "~
## $ has_key           <chr> "随时看房", "随时看房", "随时看房", "随时看房", "随~
## $ vr                <chr> NA, "VR看装修", NA, NA, "VR看装修", NA, "VR看装修", ~
```

各变量的简短统计：

```
##  property_name      property_region      price_ttl        price_sqm
##  Length:3000        Length:3000        Min.   :  10.6   Min.   : 1771
##  Class :character   Class :character   1st Qu.:  95.0   1st Qu.:10799
##  Mode  :character   Mode  :character   Median : 137.0   Median :14404
##                                        Mean   : 155.9   Mean   :15148
##                                        3rd Qu.: 188.0   3rd Qu.:18211
##                                        Max.   :1380.0   Max.   :44656
##     bedrooms      livingrooms    building_area    directions1
##  Min.   :1.000   Min.   :0.000   Min.   : 22.77   Length:3000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.: 84.92   Class :character
##  Median :3.000   Median :2.000   Median : 95.55   Mode  :character
##  Mean   :2.695   Mean   :1.709   Mean   :100.87
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:117.68
##  Max.   :7.000   Max.   :4.000   Max.   :588.66
##  directions2        decoration        property_t_height property_height
##  Length:3000        Length:3000        Min.   : 2.00     Length:3000
##  Class :character   Class :character   1st Qu.:11.00     Class :character
##  Mode  :character   Mode  :character   Median :27.00     Mode  :character
##                                        Mean   :24.22
##                                        3rd Qu.:33.00
##                                        Max.   :62.00
##  property_style       followers       near_subway          if_2y
##  Length:3000        Min.   :  0.000   Length:3000        Length:3000
##  Class :character   1st Qu.:  1.000   Class :character   Class :character
##  Mode  :character   Median :  3.000   Mode  :character   Mode  :character
```

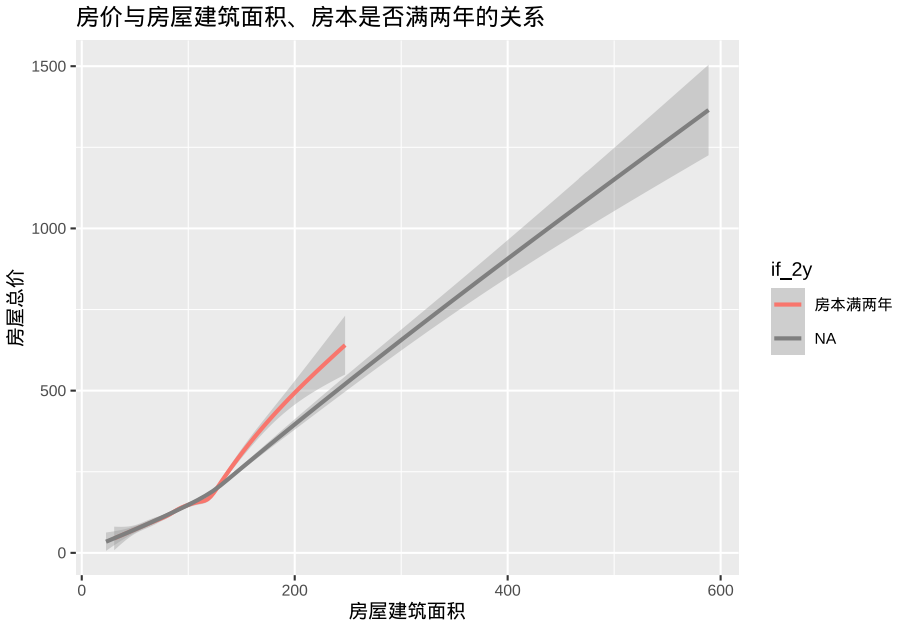
```
##                     Mean   :  6.614
##                     3rd Qu.:  6.000
##                     Max.   :262.000
##    has_key              vr
##  Length:3000        Length:3000
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

可以看到：

- 在 3000 套二手房数据中，房屋的总价分布在 10.6（万）与 1380.0（万）之间，平均房价为 155.9（万）；
- 在 3000 套二手房数据中，房屋的建筑面积分布在 22.77（平方米）与 588.66（平方米）之间，平均建筑面积为 100.87（平方米）；
- 在 3000 套二手房数据中，卧室数量的中位数为三居室，卧室数量的平均值为 2.695；起居室数量的中位数为两厅，起居室数量的平均值为 1.709

4 探索性分析

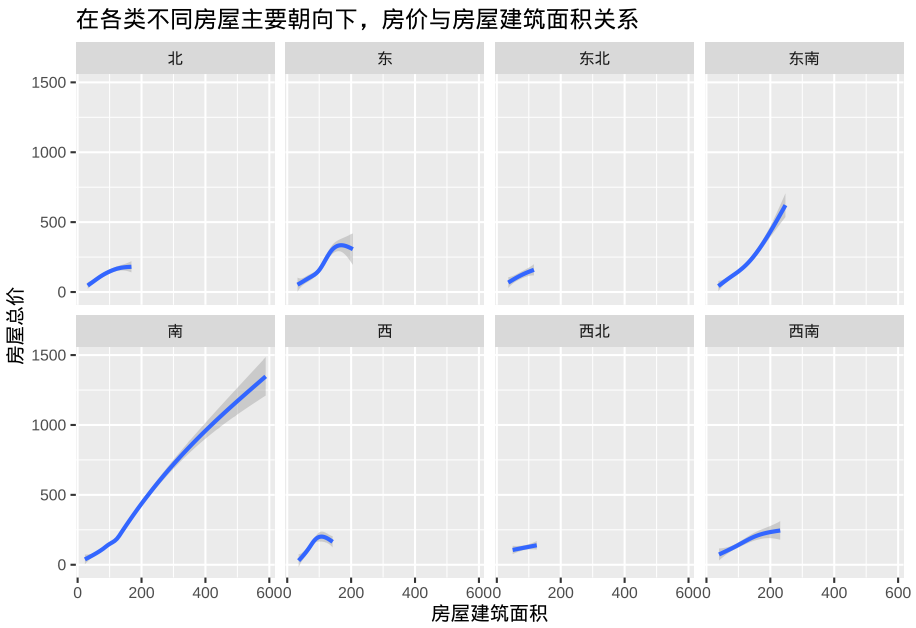
4.1 变量 1 的数值描述与图形



发现：

- 发现 1：随房屋面积的上升，房屋总价明显上升，且房屋建筑面积越大时，房屋总价的置信区间明显变宽。
- 发现 2：其中相同房屋面积的情况下，房本满 2 年的房屋总价，明显高于房本未满 2 年的房型。

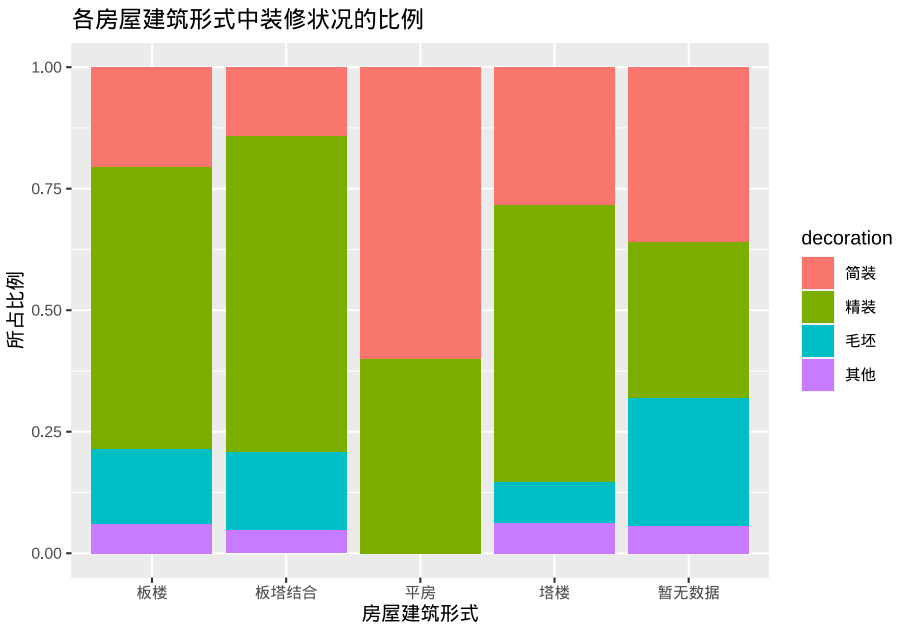
4.2 变量 2 的数值描述与图形



发现：

- 发现 1：房屋主要朝向为南的房子最多，且建筑面积 200 以上的房屋，主要朝向只有南、西南或东南. 建筑面积 300 以上的房屋，主要朝向只有南。
- 发现 2：相同房屋建筑面积下，所有方向中，房屋主要朝向的为南房屋总价最高。

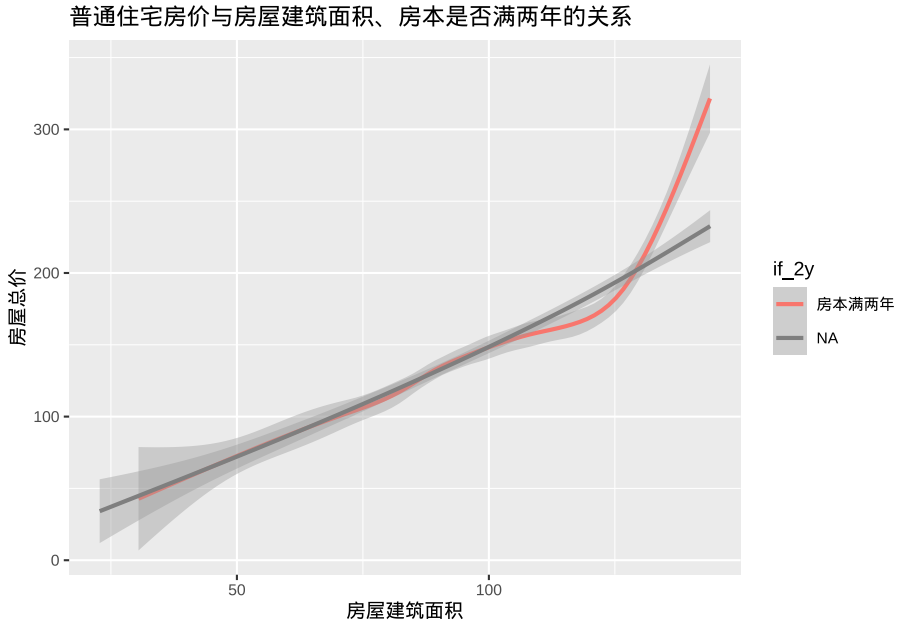
4.3 变量 3 的数值描述与图形



发现：

- 发现 1：在建筑形式为板楼、板塔结合、塔楼的房屋中，装修状况为“精装”的房屋所占比例最大，装修状况为“简装”的房屋所占比例次之；在建筑形式为平房还有缺失建筑形式数据的房屋中，装修状况为“简装”的房屋所占比例最大，装修状况为“精装”的房屋所占比例次之。
- 发现 2：在建筑形式为平房的房屋中，装修状况不存在毛坯或其他的情况。

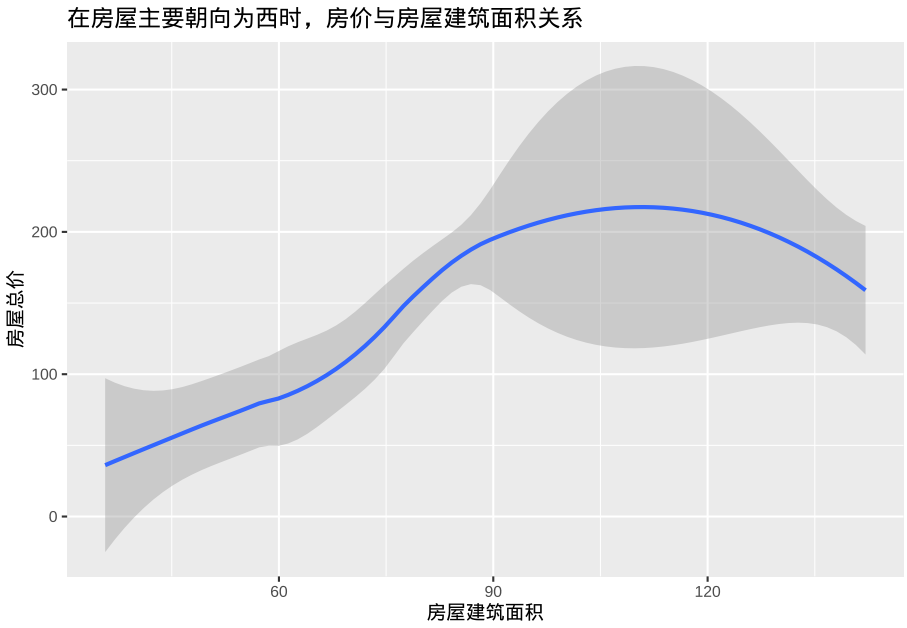
4.4 探索问题 1：在房屋建筑面积 144 以内（普通住宅），房本满 2 年的房屋总价和房本未满 2 年的价格差异分析



发现：

- 发现 1：在房屋建筑面积 130 以内，房本满两年和房本未满两年的房屋总价几乎没有差异，甚至房屋建筑面积 100~130 之间，房本满两年的房屋总价出现高于房本未满两年的房屋总价的异常现象，可能由于数据量不足导致。
- 发现 2：在房屋建筑面积 130 以上，在相同的房屋建筑面积下，房本满两年的房屋总价明显高于未满两年的，且趋势稳定。

4.5 探索问题 2：在各类不同房屋主要朝向下，房屋总价与房屋面积关系拟合曲线出现了一定程度的下降区间，以朝西为例进行分析



发现：

- 发现 1: 筛选在房屋主要朝向为西时，房价与房屋建筑面积关系出现先上升后下降的趋势。查看房屋主要朝向为西的数据发现，3000 套二手房中主要朝向为西的房屋仅有 19 套，在共计八个朝向中，该数据严重不足，房屋总价与房屋面积关系出现了一定程度的下降区间，可能是由于数据量不足导致拟合不够精确。
- 发现 2：建筑面积在 90~135 的区间内，房屋总价的置信区间异常宽大，从一定程度印证了数据量不足的上述发现。

4.6 探索问题 3：在 3000 套二手房房屋中，平房的装修状况与其他建筑形式的房屋明显有差异，仅存在简装和精装两种情况，且其比例也和其他建筑形式的房屋有差异。

```
##  property_name      property_region      price_ttl       price_sqm
##  Length:5           Length:5           Min.   :125.0   Min.   :11361
##  Class :character   Class :character   1st Qu.:128.0   1st Qu.:12577
##  Mode  :character   Mode  :character   Median :135.0   Median :12577
##                                        Mean   :132.6   Mean   :12860
##                                        3rd Qu.:135.0   3rd Qu.:13032
##                                        Max.   :140.0   Max.   :14752
##     bedrooms    livingrooms building_area    directions1
##  Min.   :2.0   Min.   :2    Min.   : 86.77   Length:5
##  1st Qu.:3.0   1st Qu.:2    1st Qu.: 95.92   Class :character
##  Median :3.0   Median :2    Median :107.34   Mode  :character
##  Mean   :2.8   Mean   :2    Mean   :104.12
##  3rd Qu.:3.0   3rd Qu.:2    3rd Qu.:107.34
##  Max.   :3.0   Max.   :2    Max.   :123.23
##  directions2         decoration        property_t_height property_height
##  Length:5           Length:5           Min.   : 6.0      Length:5
##  Class :character   Class :character   1st Qu.: 6.0      Class :character
##  Mode  :character   Mode  :character   Median :18.0      Mode  :character
##                                        Mean   :19.4
##                                        3rd Qu.:33.0
##                                        Max.   :34.0
##  property_style       followers    near_subway           if_2y
##  Length:5           Min.   : 0.0   Length:5           Length:5
##  Class :character   1st Qu.: 6.0   Class :character   Class :character
##  Mode  :character   Median :25.0   Mode  :character   Mode  :character
##                     Mean   :40.6
##                     3rd Qu.:86.0
##                     Max.   :86.0
##    has_key              vr
```

```
##  Length:5           Length:5
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

发现：

- 发现 1：在 3000 套二手房房屋中，平房仅有 5 套，导致平房的装修状况与其他建筑形式的房屋明显有差异。
- 发现 2：从平房的数据量来看，可能是数据本身有偏，收集不全面；另一方面，由于该数据集缺失房屋年代数据，从建筑年代上推测，2000 年后的房屋建筑形式为平房的数量较少，且平房自住的比例较高，所以在二手房交易市场的平房数据结论不具有普遍性。

---

5 发现总结

用 1-3 段话总结你的发现。

- 随房屋面积的上升，房屋总价呈现明显上升趋势；且房屋建筑面积越大时，房屋总价的置信区间明显变宽，即波动变大。在房屋建筑面积 130 以上，在相同的房屋建筑面积下，房本满两年的房屋总价明显高于未满两年的，且趋势稳定。
- 在房屋主要朝向为南的房子最多，且建筑面积 200 以上的豪宅（注：面积大于 144 平方米为非普通住宅，又名豪宅），主要朝向只有南、西南或东南. 建筑面积 300 以上的房屋，主要朝向只有南。筛选在房屋主要朝向为西时，房价与房屋建筑面积关系出现先上升后下降的趋势。查看房屋主要朝向为西的数据发现，3000 套二手房中主要朝向为西的房屋仅有 19 套，在共计八个朝向中，该数据严重不足，房屋总价与房屋面积关系出现了一定程度的下降区间，可能是由于数据量不足导致拟

合不够精确。

- 在建筑形式为板楼、板塔结合、塔楼的房屋中，装修状况为“精装”的房屋所占比例最大，装修状况为“简装”的房屋所占比例次之；在建筑形式为平房还有缺失建筑形式数据的房屋中，装修状况为“简装”的房屋所占比例最大，装修状况为“精装”的房屋所占比例次之。在 3000 套二手房房屋中，平房仅有 5 套，从平房的数据量来看，可能是数据本身有偏，收集不全面；另一方面，由于该数据集缺失房屋年代数据，从建筑年代上推测，2000 年后的房屋建筑形式为平房的数量较少，且平房自住的比例较高，所以在二手房交易市场的平房数据不具有普遍性。