1st_assignment_hujiaming

hujiaming

目录

1 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。

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

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。

2 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, decoration, 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：

- 直观结论 1 price_ttl 房屋总价最大值 1380 万元，最小值 10.6 万元，中位数值 137 万元，均值 155.9 万元。
- 直观结论 2 price_sqm 每平方米均价最大值 44656 元，最小值 1771 元，中位数值 14404 元，均值 15148 元。
- 直观结论 3 -部分数据存在异常值需要清洗，如房屋朝向（direction2）未填写正确值，部分数据填充为 NA。在整个表格中，数值类型数据 7 列，字符类型数据 11 列，字符类型数据需要进一步处理分析。# 探索性分析

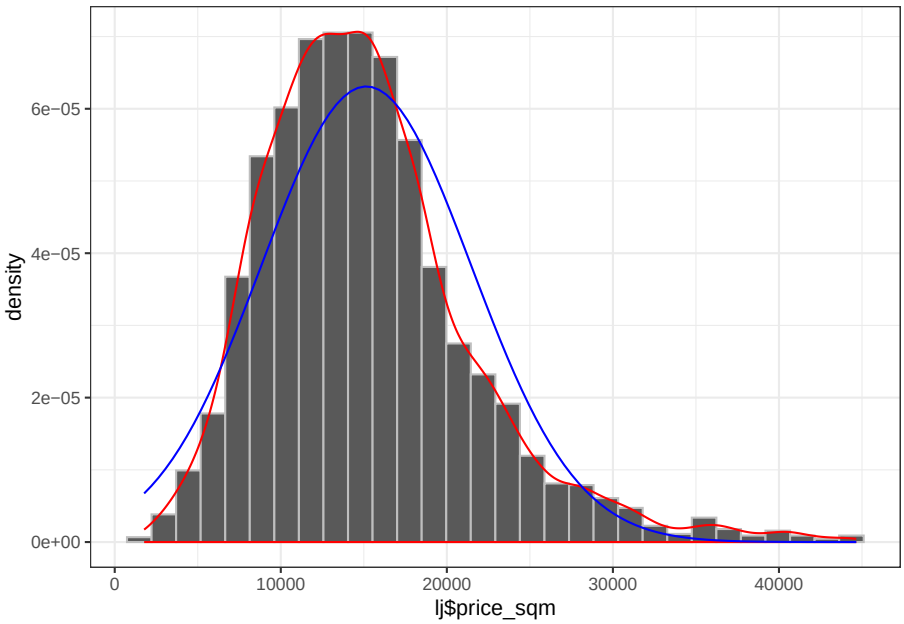
2.1 变量 price_sqm 的数值描述与图形

- 发现 1
- price_sqm 变量数值描述最大值：44656 最小值：1771 中位数值：14404 均值：15148 标准差：6323.175 极差：1369.4

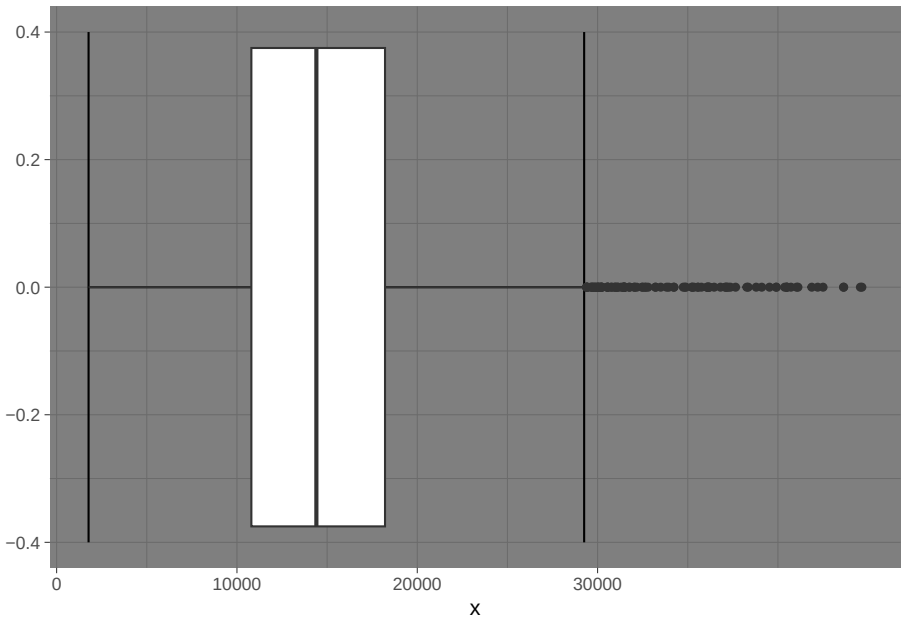
```
## [1] 6323.175
```

```
## [1] 1369.4
```

- price_sqm 变量图形描述 price_sqm 变量图形描述：直方图描述与概率密度曲线将 price_sqm 的数据用直方图展示结果类似卡方分布，红色线条为该数据的概率密度曲线，蓝色线条为该数据在正态分布下的概率密度曲线，可以看出房屋的每平方米单价趋向于正态分布

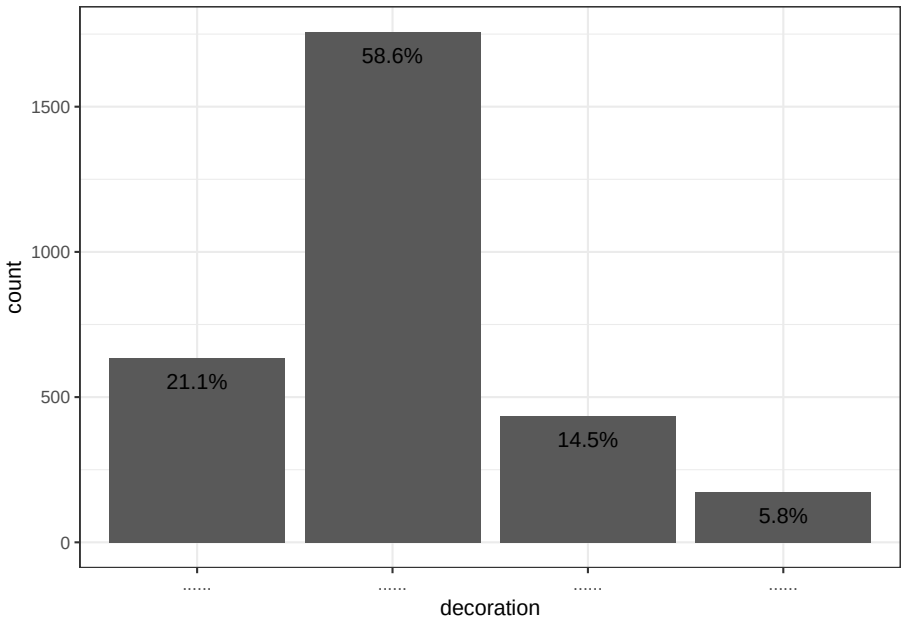


-price_ttl 变量图形描述：箱线图箱线图展示了变量的四分位、上下分界等数据特征，后续将把离群点数据作为异常值处理。



2.2 变量 decoration 的数值描述与图形

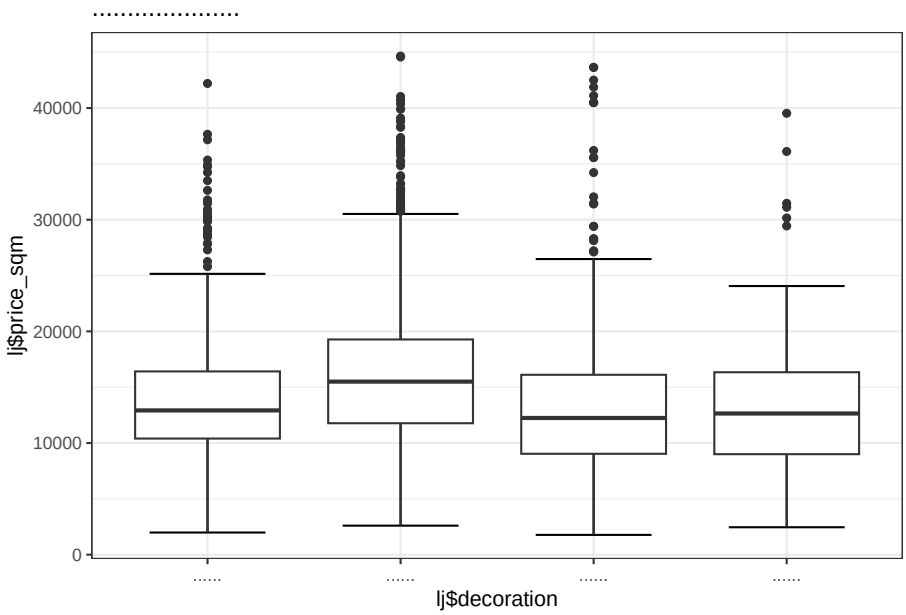
- decoration 变量数值描述数据类型：character 是否有空值：无数据内容：：“精装” “简装” “其他” “毛坯”
- decoration 变量图形描述：直方图通过条形图可以看到精装是占比最高的 58.6%，其次是简装 21.1%，最后是毛坯 14.5%, 其他占比 5.8%

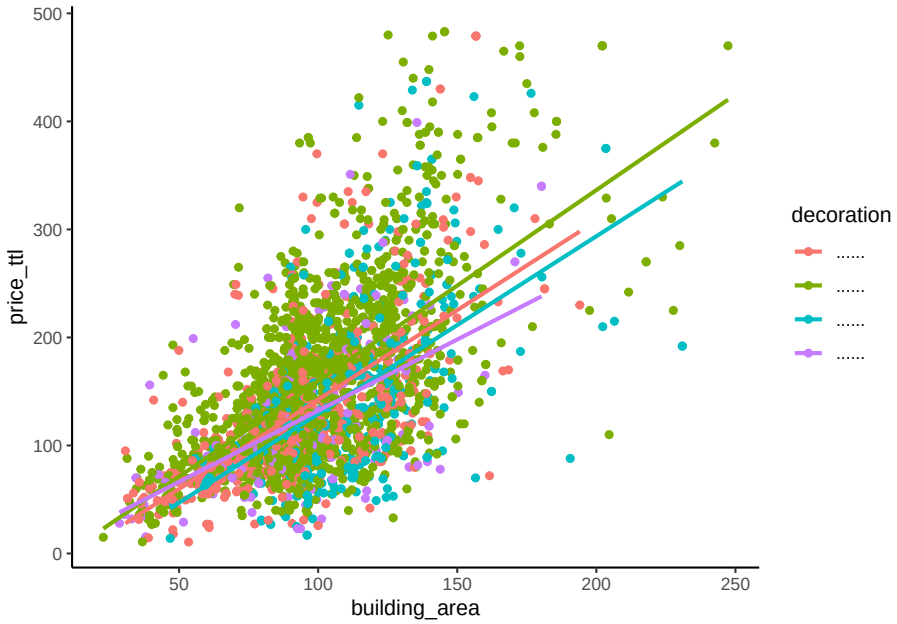


2.3 探索问题 1

-装修与房屋价格之间是否存在某种关系

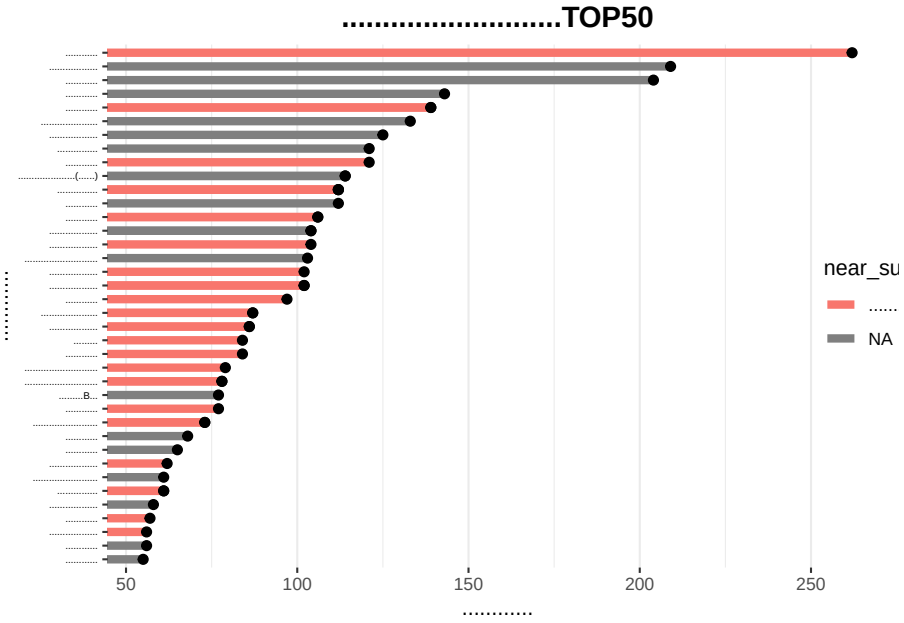
- 发现 1 以下是 4 种装修类别的房屋的单价展示出的箱线图，以及面积/房屋总价散点图，从箱线图中的四分位、散点图拟合直线的斜率中可以看出：精装修确实会让房屋的每平方米单价增高，但简装对房屋单价的影响并不大。


.....
40000
30000
20000
10000
0
lj$price_sqm
.....
.....
.....
.....
lj$decoration


500
400
300
200
100
0
price_ttl
50
100
150
200
250
building_area
decoration
.....
.....
.....
.....

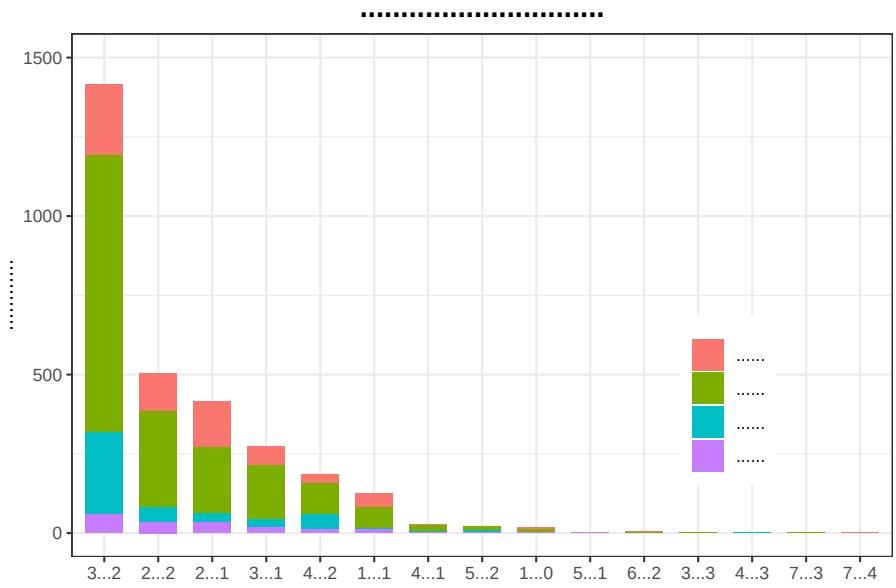
2.4 探索问题 2

- 在 3000 个样本中，热度最高的楼盘是哪些？它们的热度是否与地铁相关？
- 以下是武汉最受关注的楼盘 TOP50，其中最受欢迎的楼盘为十里和府，但它们的热度与是否靠近地铁的相关性不高。



2.5 探索问题 3

- 二手房户型分布与装修情况与样本数量的关系，3 室两厅且为精装是样本数量最大的类型



3 发现总结

对链家武汉二手房网站的 3000 套房产信息分析后可以得出以下结论：1. 样本数据中的二手房平均每平方米单价波动幅度较大，但整体近似服从平均值为 15148，标准差为 6323.175 的正态分布 2. 装修情况以精装修居多，精装修的样本占到了样本总量的 58.6% 3. 装修程度与房屋单价呈现正相关（装修程度越好，房屋总价/面积的拟合直线斜率越大）4. 从房屋结构的角度上分析，三室两厅是是样本数量最大的类型