第一次作业你的报告题目

Code

蔡亚东

2024-10-29

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 (https://wh.lianjia.com/ershoufang/)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。

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：

武汉链家二手房

property_name	property_region	price_ttl	price_sqm	bedrooms	livingrooms	building_area	directions1	directions2	decoration	property_t_heig
南湖名都A区 | 南湖沃尔	237.									

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 lj这个数据表里，有3000条数据，18个变量。包含小区名称，小区位置，房屋总价，房屋每平方的价格，几个卧室，几个厅，建筑面积，主要朝向，次要朝向， 装修状况，楼栋总层数，房屋楼层，房屋风格，关注人数，是否靠近地铁，是否满2年，是否有钥匙，是否有VR
- 直观结论2 字符型数据：小区名称，小区位置，主要朝向，次要朝向，装修状况，房屋楼层，房屋风格，长度都为3000 逻辑型数据：是否靠近地铁，是否满2年，是否有钥匙，是否有VR 数值类数据：房屋总价，房屋每平方的价格，几个卧室，几个厅，建筑面积，楼栋总层数，关注人数 数值类数据分别展示了最小值，1/4分位数，中位数，平均值，3/4分位数，最大值
- …

3 探索性分析

3.1 变量1的数值描述与图形

4 变量1：房屋每平米的价格 的数值描述

Hide

```
        #最小值和最大值
        range(lj$price_sqm)
```

```
## [1]  1771 44656
```

Hide

```
        #平均数
        mean(lj$price_sqm)
```

```
## [1] 15148.49
```

Hide

```
        #中位数
        median(lj$price_sqm)
```

```
## [1] 14404
```

Hide

```
#四分位距
IQR(lj$price_sqm) #quantile(lj$price_sqm,0.75)-quantile(lj$price_sqm,0.25)
```

```
## [1] 7411.75
```

Hide

```
#众数
mad(lj$price_sqm)
```

```
## [1] 5465.605
```

Hide

```
#方差
var(lj$price_sqm)
```

```
## [1] 39982547
```

Hide

```
#标准差
sd(lj$price_sqm)
```

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

Hide

```
#峰度和偏度
e1071::skewness(lj$price_sqm)
```

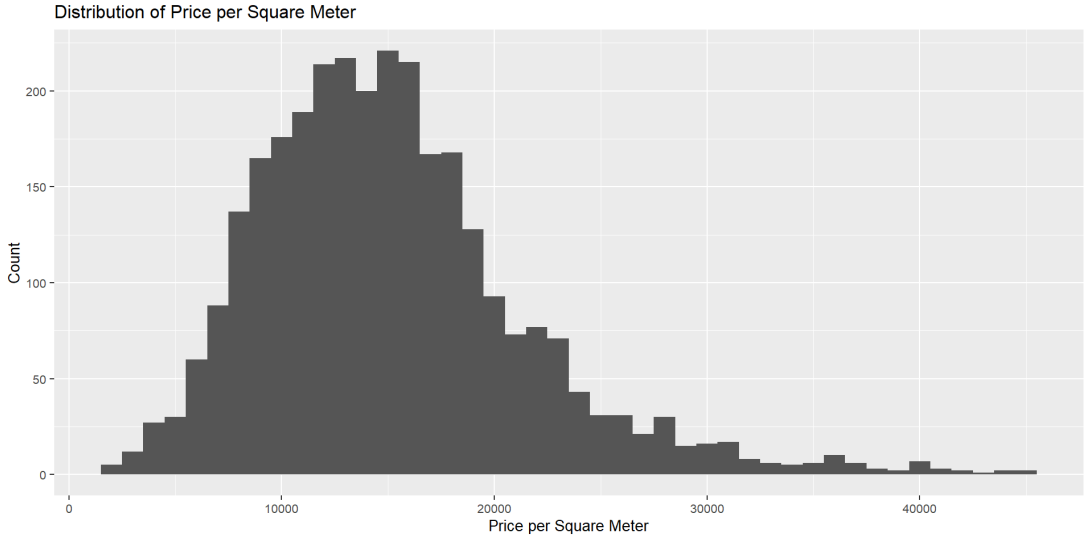
```
## [1] 1.079464
```

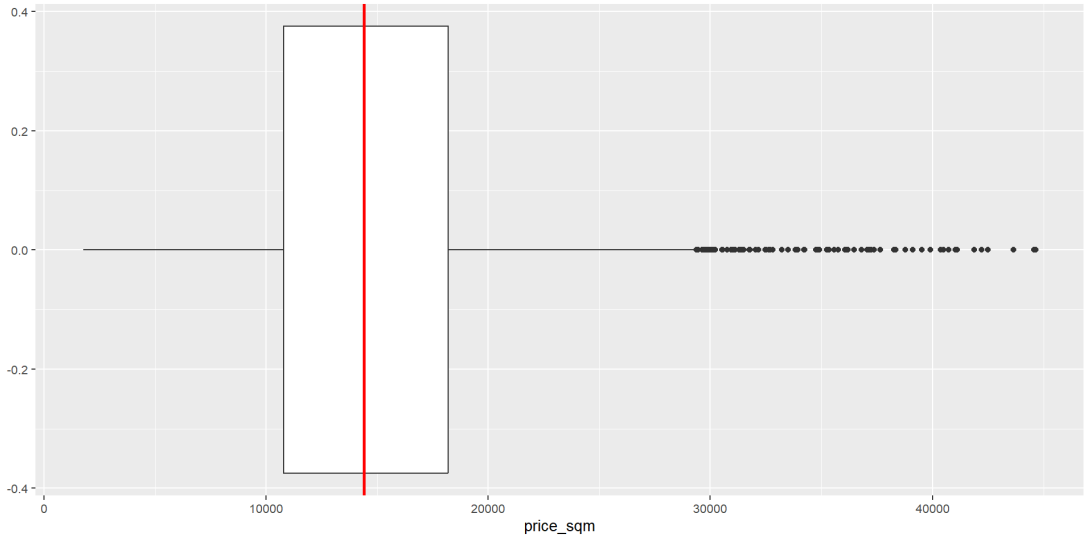
Hide

```
e1071::kurtosis(lj$price_sqm)
```

```
## [1] 2.025625
```

5 变量1：房屋每平米的价格 的图形





发现:

- 发现1 lj数据中“每平米价格”，服从于 近似的正太分布（μ = 15148.49，sd = 6323.175），右偏，偏度为1.079464 中位数为14404，小于平均数，意味着存在着部分高价格，拉高了整体平均水平
- 发现2 箱线图右端存在部分异常值，已经超出 Q3 + 1.5 * IQR

5.1 变量2的数值描述与图形

6 变量2： bedrooms的数值描述与图形

Hide

```
#最小值和最大值
range(lj$bedrooms)
```

```
## [1] 1 7
```

Hide

```
#平均数
mean(lj$bedrooms)
```

```
## [1] 2.695
```

Hide

```
#中位数
median(lj$bedrooms)
```

```
## [1] 3
```

Hide

```
#四分位距
IQR(lj$bedrooms) #quantile(lj$price_sqm,0.75)-quantile(lj$price_sqm,0.25)
```

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

Hide

```
#众数
mad(lj$bedrooms)
```

```
## [1] 0
```

Hide

```
#方差
var(lj$bedrooms)
```

```
## [1] 0.5328193
```

Hide

```
        #标准差
        sd(lj$bedrooms)
```

```
## [1] 0.7299447
```

Hide

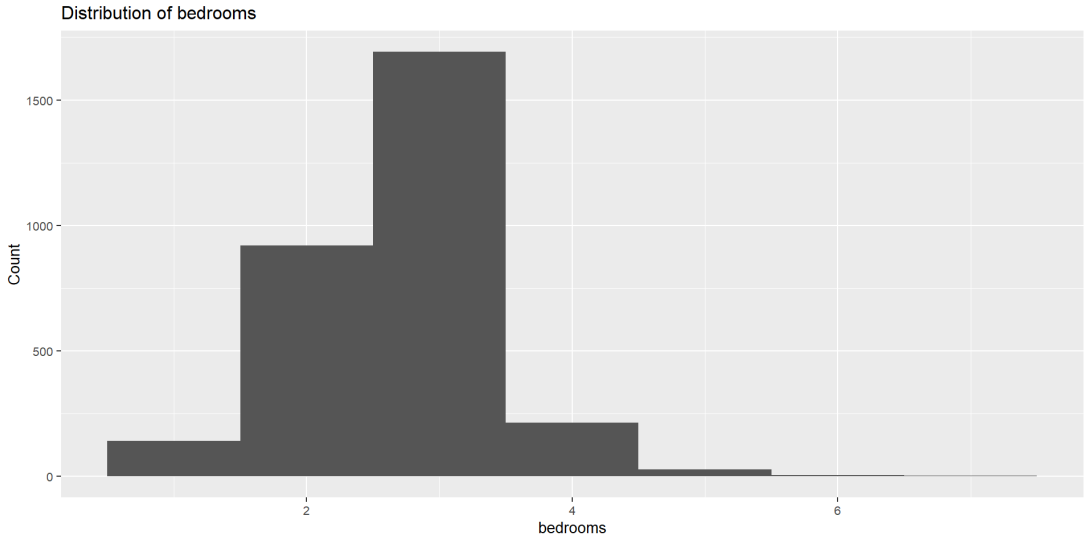
```
        #峰度和偏度
        e1071::skewness(lj$bedrooms)
```

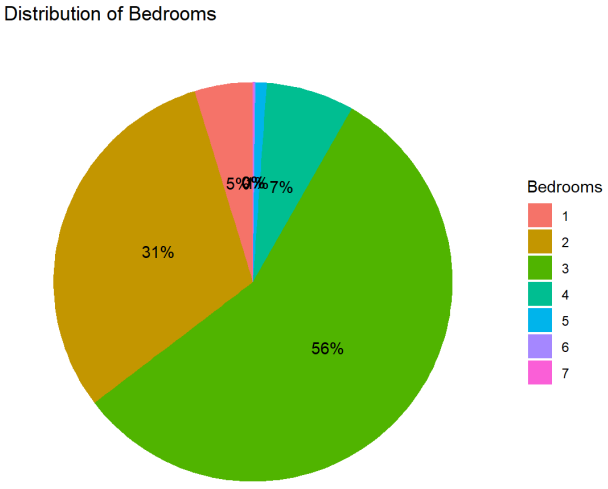
```
## [1] 0.1356027
```

Hide

```
        e1071::kurtosis(lj$bedrooms)
```

```
## [1] 1.635711
```





发现:

- 发现1 人们更钟情于购买3个卧室的房子，占比为56%。其次是2个卧室的房子，占比31%
- 发现2 同时也有少部分房屋存在多间卧室，最多的是7间

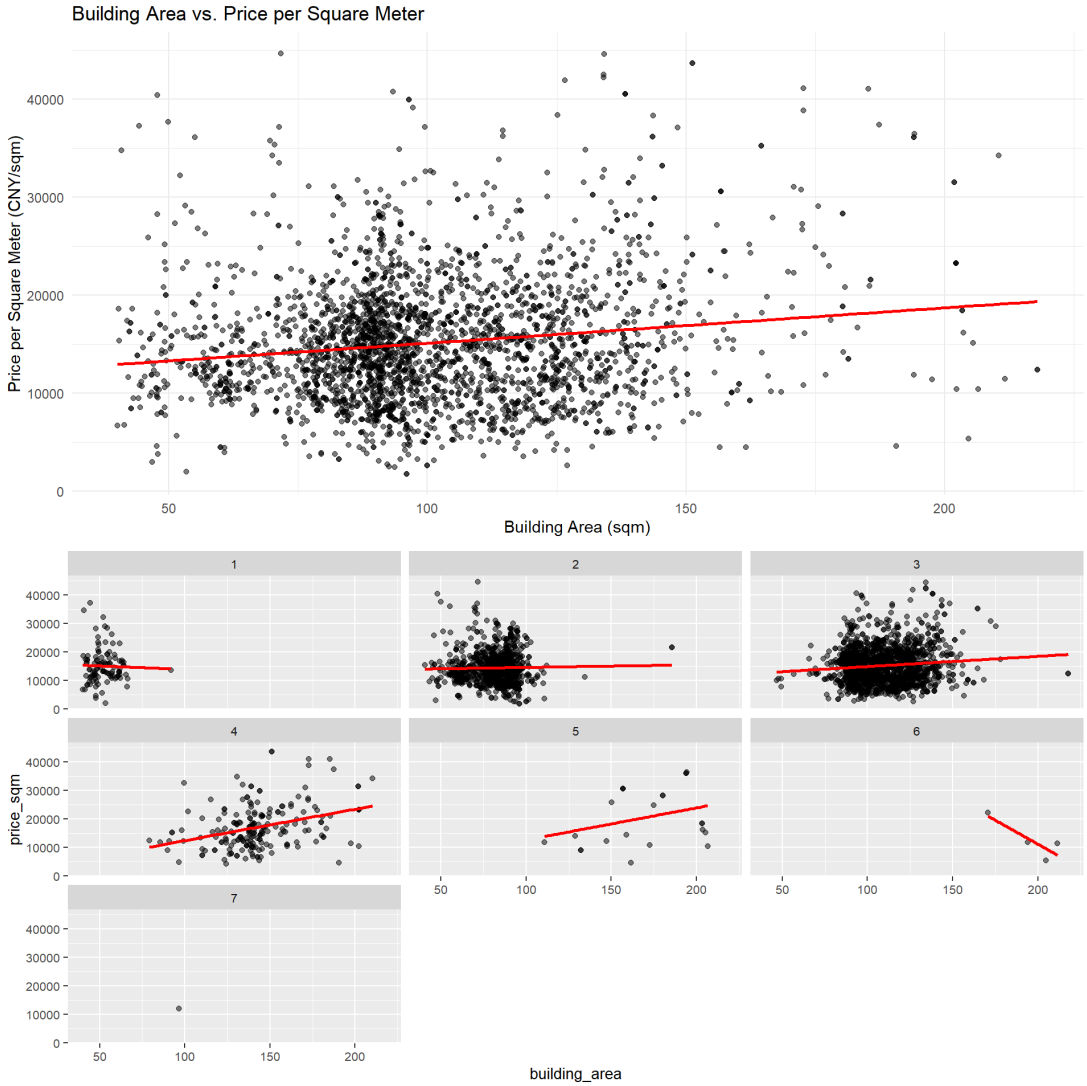
6.1 变量...的数值描述与图形

7 property_region 的词条频次图

发现：

- 发现1 白沙洲的房子供应最多，其次是盘龙城，四新，光谷东等。
- 发现2 供应房子多的地方大多都位于郊区，意味着市区土地有限，人口朝郊区外扩

7.1 探索问题1：面积大的房子，每平米的价格更低还是更高？

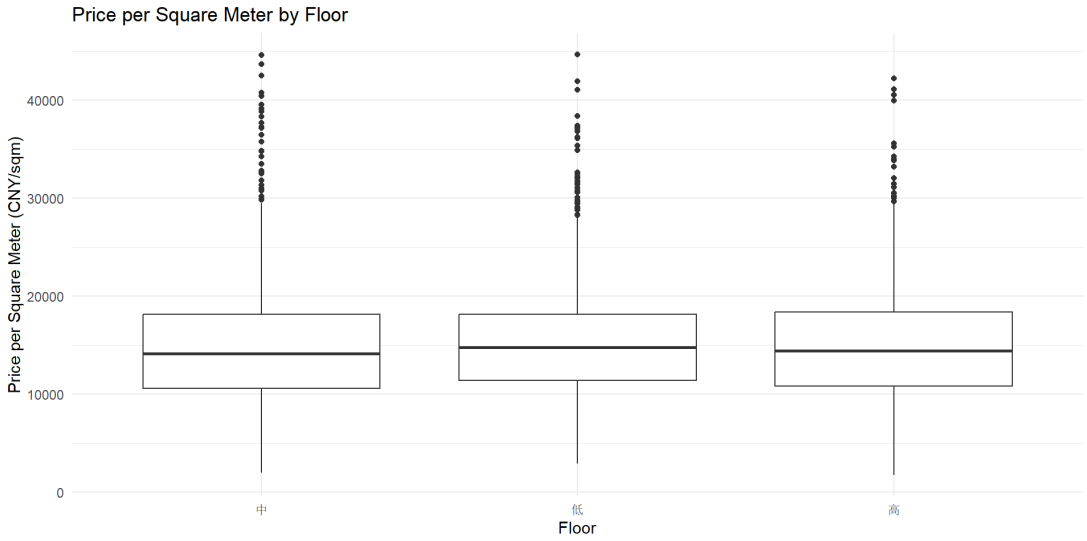


```
## NULL
```

发现：总体面积大的房子，每平米价格更高。房间越多，越明显。

- 发现1 总体上，面积大的房子，每平米价格更高
- 发现2 1，2，3室的房子，面积对每平米价格的影响不大。 4，5室的房子随着面积越大，每平米价格更高。6，7室的样本少于30，无法得出结论

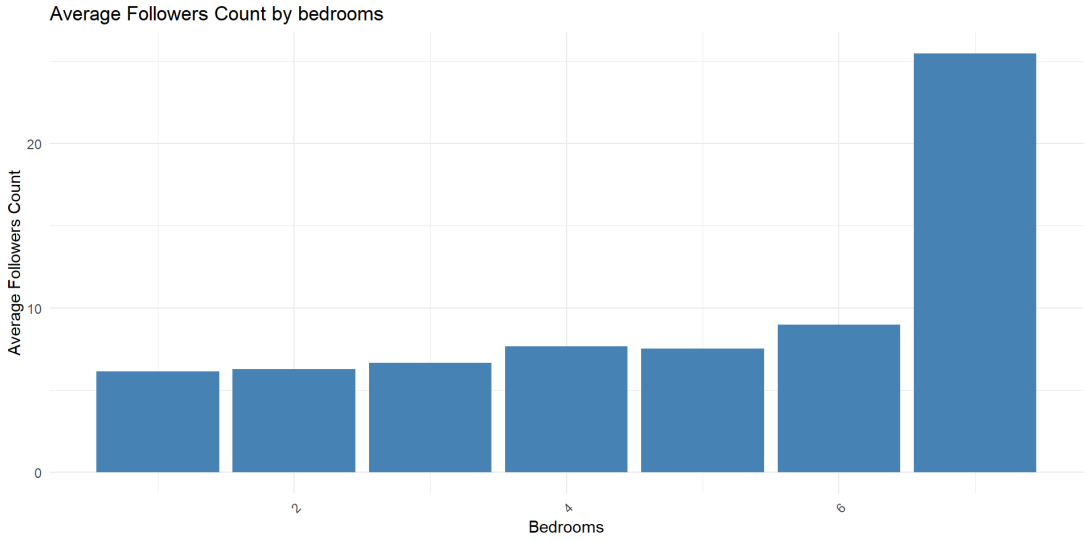
7.2 探索问题2：楼层高低对每平米价格的影响?

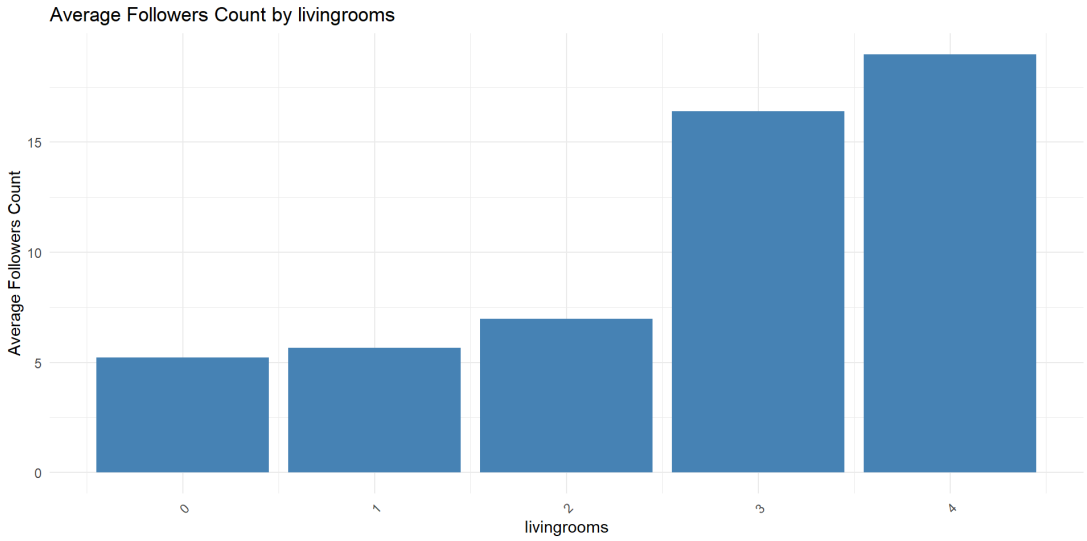


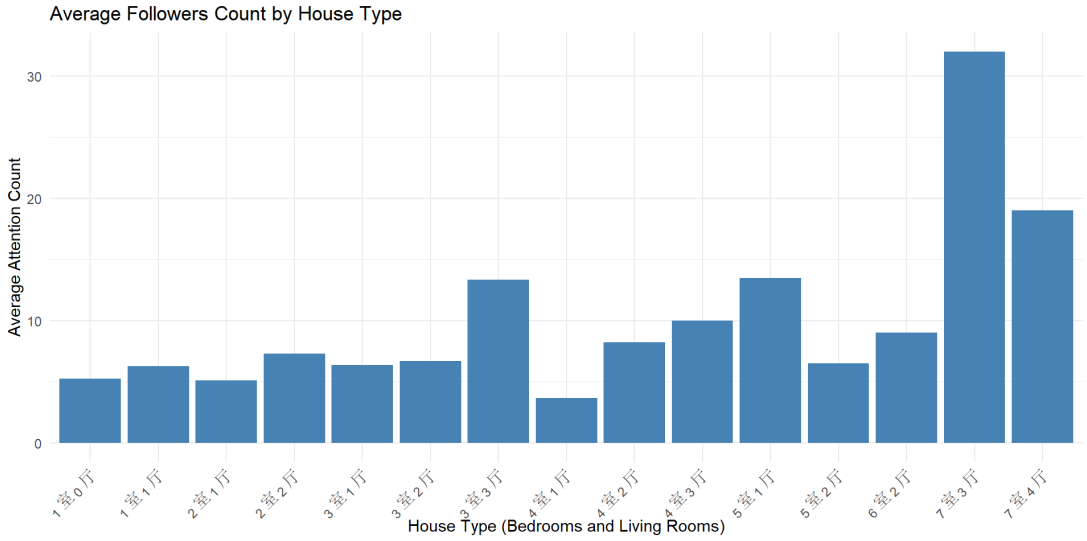
发现：目前3000个样本得出来的初步结论：楼层越低，房价越高。然而缺失"房屋是否带有电梯"这个因素，所以暂时对结论存疑。

- 发现1 低层的房子每平米价格更高
- 发现2 中高成的房子每平米价格标准差更大，

7.3 探索问题3：什么类型的房子（几室几厅）更受人关注?







发现：人们更想要住越大的房子，有更多的卧室和厅。人人都向往富翁的生活

- 发现1 多室多厅的房子更受人关注
- 发现2 7个卧室的房子，关注度相比1-6个卧室 陡然提示。可能存在其他因素影响

---

8 发现总结

用1-3段话总结你的发现。

总体面积大的房子，每平米价格更高。房间越多，越明显。倾向于买房间越多的房子的人，家底厚实，越不在乎价格。

楼层越低，房价越高。有可能会因为是否带有电梯这个因素影响这个结论。

人们更想要住越大的房子，有更多的卧室和厅。人人都向往富翁的生活，但现实是2-3室的房源更多，占到80%左右。