第一次作业你的报告题目

夏章印

2024-10-30

你的主要发现

1. 发现 1
2. 发现 2
3. 发现 3

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。

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：

property_name	property_region	price_ttl	price_sqm	bedrooms	livingrooms	building_area	dire
南湖名都 A 区｜南湖沃尔	｜237.	｜1870	｜3		1｜1	6.68｜南	｜北
万科紫悦湾｜光谷东	｜12	.0｜14	13｜	3｜	2｜	86.91｜南	|NA
东立国际｜二七		75.0｜	5968｜	1｜	1｜	46.97｜南	|NA
新都汇｜光谷	场｜1	8.0｜1	702｜	3｜	2｜	119.73｜北	｜东
保利城一期｜团结大道	｜182	0｜175	9｜	｜	2｜	03.95｜东南	|NA
加州橘郡｜庙山		22.0｜	0376｜	3｜	2｜	117.59｜南	｜北
省建筑五公司西区｜光谷广场	｜99.0｜	12346｜	2｜		｜80	19｜南	NA

保利上城东区	白沙洲	193	8	163	6		2	18.64	南	北
石化大院	中南丁	桥	325	0	326	1		1	99.60	南	北
阳光花园	杨汉湖	1	2.0	1	403	3	2	110.33	南	北

各变量的简短信息:

```
## Rows: 3,000
## Columns: 18
## $ property_name     <fct> 南湖名都A区, 万科紫悦湾, 东立国际, 新都汇, 保利城一~
## $ property_region   <fct> 南湖沃尔玛, 光谷东, 二七, 光谷广场, 团结大道, 庙山, ~
## $ 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       <fct> 南, 南, 南, 北, 东南, 南, 南, 南, 南, 南, 南, 南, 东~
## $ directions2       <fct> 北, NA, NA, 东, NA, 北, NA, 北, 北, 北, 北, NA, 西南~
## $ decoration        <fct> 精装, 精装, 简装, 精装, 简装, 精装, 简装, 其他, 简装~
## $ property_t_height <dbl> 17, 28, 18, 32, 34, 34, 7, 34, 5, 7, 25, 32, 8, 31, ~
## $ property_height   <fct> 中, 中, 低, 高, 中, 低, 低, 中, 低, 低, 高, 高, 中, ~
## $ property_style    <fct> 塔楼, 板楼, 塔楼, 塔楼, 板塔结合, 板楼, 板楼, 板塔结~
## $ followers         <dbl> 3, 1, 3, 2, 3, 1, 0, 0, 2, 0, 0, 0, 10, 0, 0, 1, 0, ~
## $ near_subway       <fct> 近地铁, NA, 近地铁, 近地铁, NA, NA, 近地铁, 近地铁, ~
## $ if_2y             <fct> NA, 房本满两年, NA, 房本满两年, 房本满两年, 房本满两~
## $ has_key           <fct> 随时看房, 随时看房, 随时看房, 随时看房, 随时看房, 随~
## $ vr                <fct> NA, VR看装修, NA, NA, VR看装修, NA, VR看装修, NA, NA~
```

各变量的简短统计:

```
##        property_name  property_region   price_ttl        price_sqm
##  东立国际      :  22   白沙洲 : 167    Min.   :  10.6   Min.   : 1771
##  保利中央公馆 :  16   盘龙城 : 126    1st Qu.:  95.0   1st Qu.:10799
##  朗诗里程      :  16   四新   : 116    Median : 137.0   Median :14404
##  恒大名都      :  15   光谷东 : 112    Mean   : 155.9   Mean   :15148
##  阳光100大湖第:  15   金银湖 :  97    3rd Qu.: 188.0   3rd Qu.:18211
##  保利城一期    :  13   后湖   :  86    Max.   :1380.0   Max.   :44656
##  (Other)       :2903   (Other):2296
##     bedrooms      livingrooms    building_area     directions1    directions2
##  Min.   :1.000   Min.   :0.000   Min.   : 22.77   南     :2454   北     :1189
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.: 84.92   东南   : 281   南     :  66
##  Median :3.000   Median :2.000   Median : 95.55   东     :  98   西     :  25
##  Mean   :2.695   Mean   :1.709   Mean   :100.87   北     :  68   东南   :  15
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:117.68   西南   :  57   西南   :  12
##  Max.   :7.000   Max.   :4.000   Max.   :588.66   西     :  19   (Other):  21
##                                                   (Other):  23   NA's   :1672
##  decoration  property_t_height property_height  property_style
##  其他: 173   Min.   : 2.00     中  :1218       塔楼    : 527
##  毛坯: 436   1st Qu.:11.00     低  : 816       平房    :   5
##  简装: 634   Median :27.00     高  : 906       暂无数据:  72
##  精装:1757   Mean   :24.22     NA's:  60       板塔结合: 615
##              3rd Qu.:33.00                     板楼    :1781
##              Max.   :62.00
##
```

```
##    followers          near_subway          if_2y          has_key
##  Min.   :  0.000   VR看装修 :   2   房本满两年:1264   随时看房:2525
##  1st Qu.:  1.000   太子湖1号:   1   NA's      :1736   近地铁  :   7
##  Median :  3.000   珞狮南   :   1                    VR看装修:   4
##  Mean   :  6.614   近地看   :   1                    世纪花园:   1
##  3rd Qu.:  6.000   近地铁   :1554                    仁厚社区:   1
##  Max.   :262.000   NA's     :1441                    (Other) :   4
##                                                      NA's    : 458
##                    vr
##  VR看装修           :2084
##  VR\xe7甸城\xe5\x8c:   1
##  保利拉\xe8\x8f    :   1
##  塔子湖             :   1
##  江景湾             :   1
##  (Other)            :   6
##  NA's               : 906
```

- 直观结论 1

  从 glimpse(lj) 的结果可以看出：

  1. 整个数据集一共有 3000 行，18 列
  2. 其中数值类型数据 7 列，字符类型数据有 9 列
  3.direction2/property_height/near_subway/if_2y/vr 数据中存在空值 (NA)

- 直观结论 2

  从 summary(lj) 的结果可以看出：

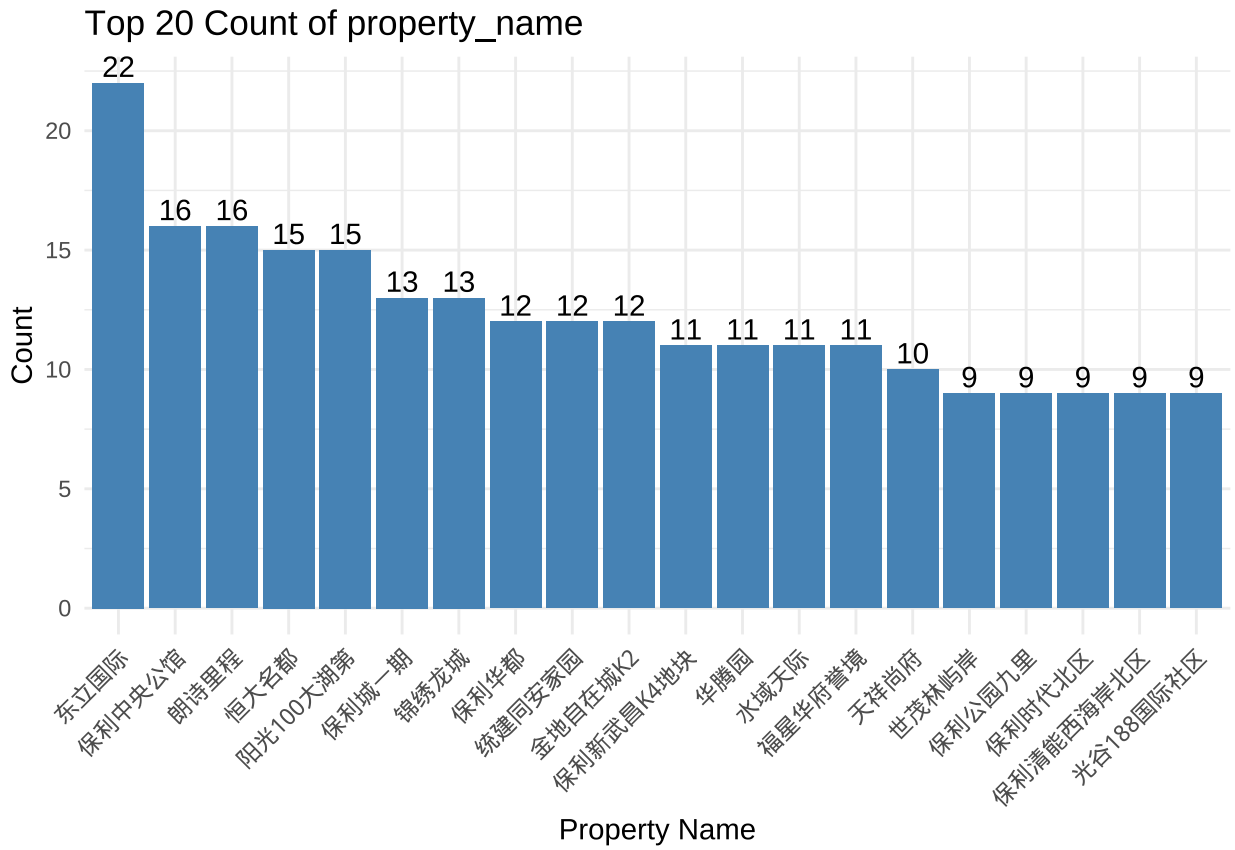
  1. 数据集中 chr 字符类型转换为 factor 因子型后的变量，显示了部分变量类型的频数
     1.1. 按小区名称汇总：东立国际有 22 套、保利中央公馆有 16 套、朗诗里程 16 套等等
     1.2. 按小区区域汇总：白沙洲有 167 套、盘龙城 126 套、四新 116 套等等
     1.3. 按小区主要和次要朝向，统计，绝大部分房子是坐北朝南的户型
     1.4. 其中简装 634 套、精装 1757 套，毛坯 436 套
     1.5. 板楼居多，有 1781 套
     1.6. 所挂的房源一半都离较近，说明该市的地铁较发达，或者房源标注近地铁有利于房源销售
     1.7. 所挂的房源绝大部分都有钥匙，随时可以看房
     1.8. 所挂的房源绝大部分支持 VR 看房
     1.9. 满二的房源接近一半
  2. 数据集中每个 num 数值型变量都做了基本的统计，包括：最小值、第一个四分位、中位数、均值、第三个四分位、最大值
     2.1. 房屋的总价：在 10.6w-1380w 之间，中位数和均值相近；第一个四分位和第三个四分位与最小值和最大值相差很大，并且有异常值
     2.2. 房屋的均价：50% 访问的均价在 10799-18211 之间，与均价 15148 比较吻合，最高的有 44656 的单价
     2.3. 房屋的房间数为 1、2、3，最大有 7 个房间
     2.4. 层高，最小 2 层，最高 62 层
     2.5. 房源的关注数，数据存在右偏，最小均值和中位数，有一些变异数据

探索性分析

变量 1(property_name) 的数值描述与图形

```
## # A tibble: 1,345 x 2
## # Groups:   property_name [1,345]
```

```
##    property_name     n
##    <fct>         <int>
##  1 东立国际         22
##  2 保利中央公馆     16
##  3 朗诗里程         16
##  4 恒大名都         15
##  5 阳光100大湖第    15
##  6 保利城一期       13
##  7 锦绣龙城         13
##  8 保利华都         12
##  9 统建同安家园     12
## 10 金地自在城K2     12
## # i 1,335 more rows
```



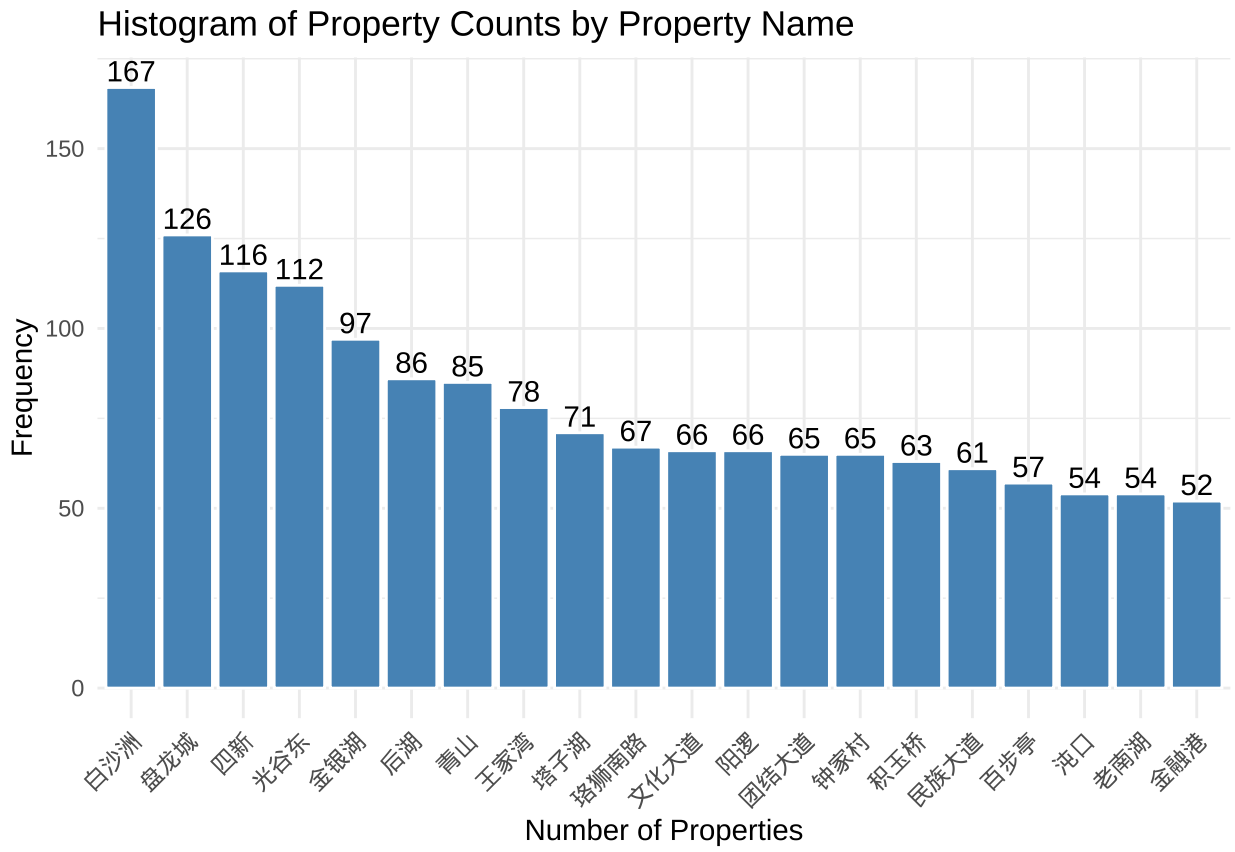
发现：

- 东立国际挂的房源最多有 22 套，其余依次是暴力中央公馆 16 套、朗诗里程 16 套、恒大名都 15 套……
- 根据词云显示，小区的名称中喜欢用国际、保利、一期、小区、万科、金地等词

变量 2(property_region) 的数值描述与图形

```
## # A tibble: 87 x 2
## # Groups:   property_region [87]
##    property_region     n
```

```
##    <fct>           <int>
##  1 白沙洲            167
##  2 盘龙城            126
##  3 四新              116
##  4 光谷东            112
##  5 金银湖             97
##  6 后湖               86
##  7 青山               85
##  8 王家湾             78
##  9 塔子湖             71
## 10 珞狮南路           67
## # i 77 more rows
```



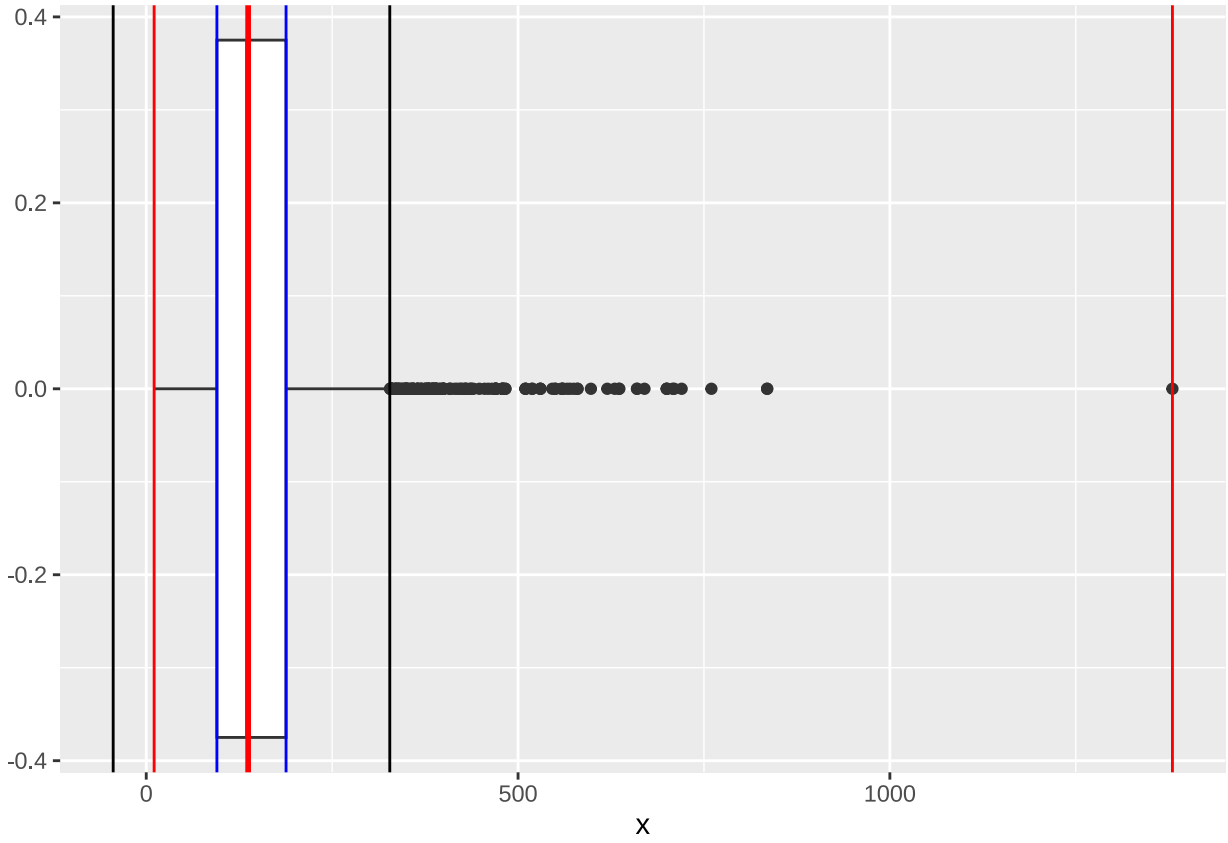
发现:

- 热门的房源区域依次为白沙洲（167 套）、盘龙城（126 套）、四新（116 套）、光谷东（112 套）、金银湖（97 套）……

变量 (price_ttl) 的数值描述与图形

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.6    95.0   137.0   155.9   188.0  1380.0
## range: 10.6 1380NULL
## IQR: 93NULL
```

```
## var: 9129.445NULL
## sd: 95.54813NULL
## skewness: 2.753223NULL
## kurtosis: 16.11672NULL
```

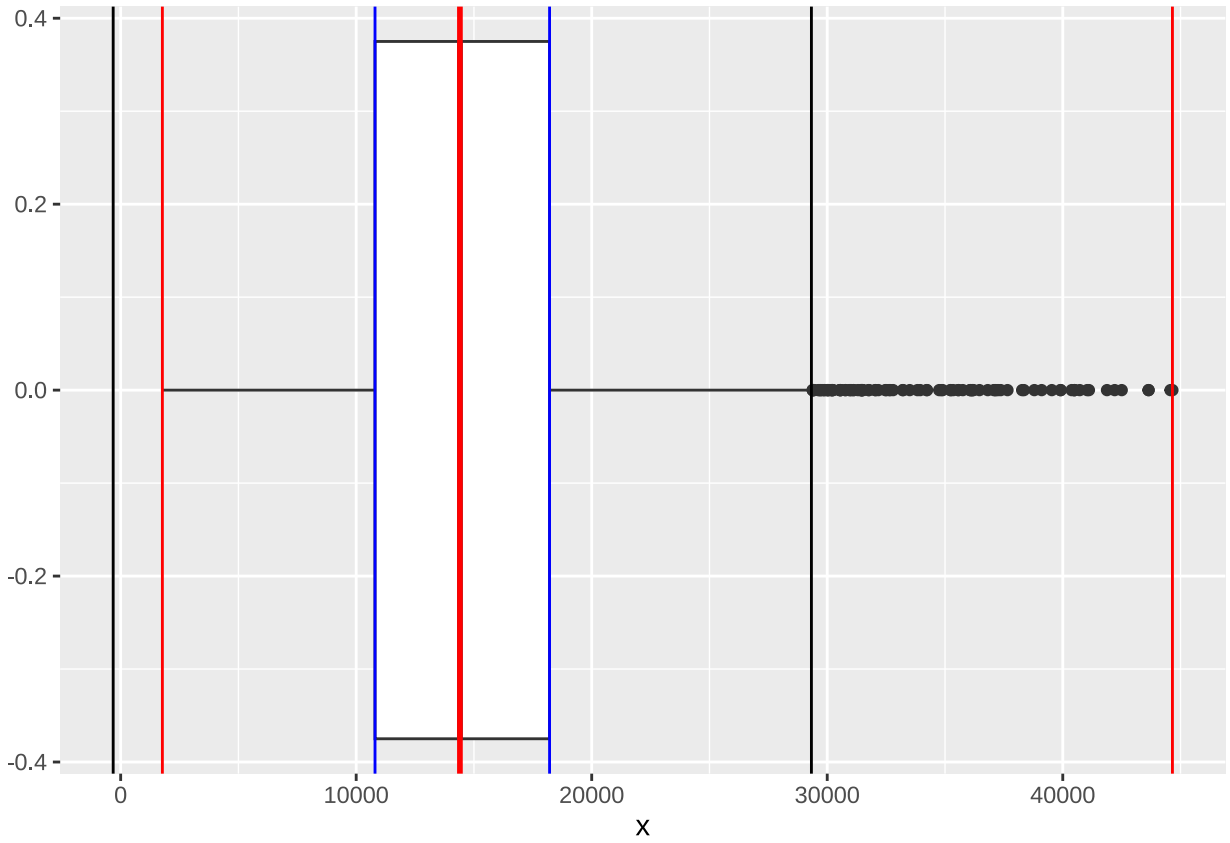


发现：

- 房屋总价存在多个右偏的异常高值（离群点）
- 中位数（137.0）和均值（155.9）相差不大，数据分布相对对称给，略有右偏
- 最大值（1380.0）远大于上限值，为极端值

变量 (price_sqm) 的数值描述与图形

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1771   10799   14404   15148   18211   44656
## range: 1771 44656NULL
## IQR: 7411.75NULL
## var: 39982547NULL
## sd: 6323.175NULL
## skewness: 1.079464NULL
## kurtosis: 2.025625NULL
```

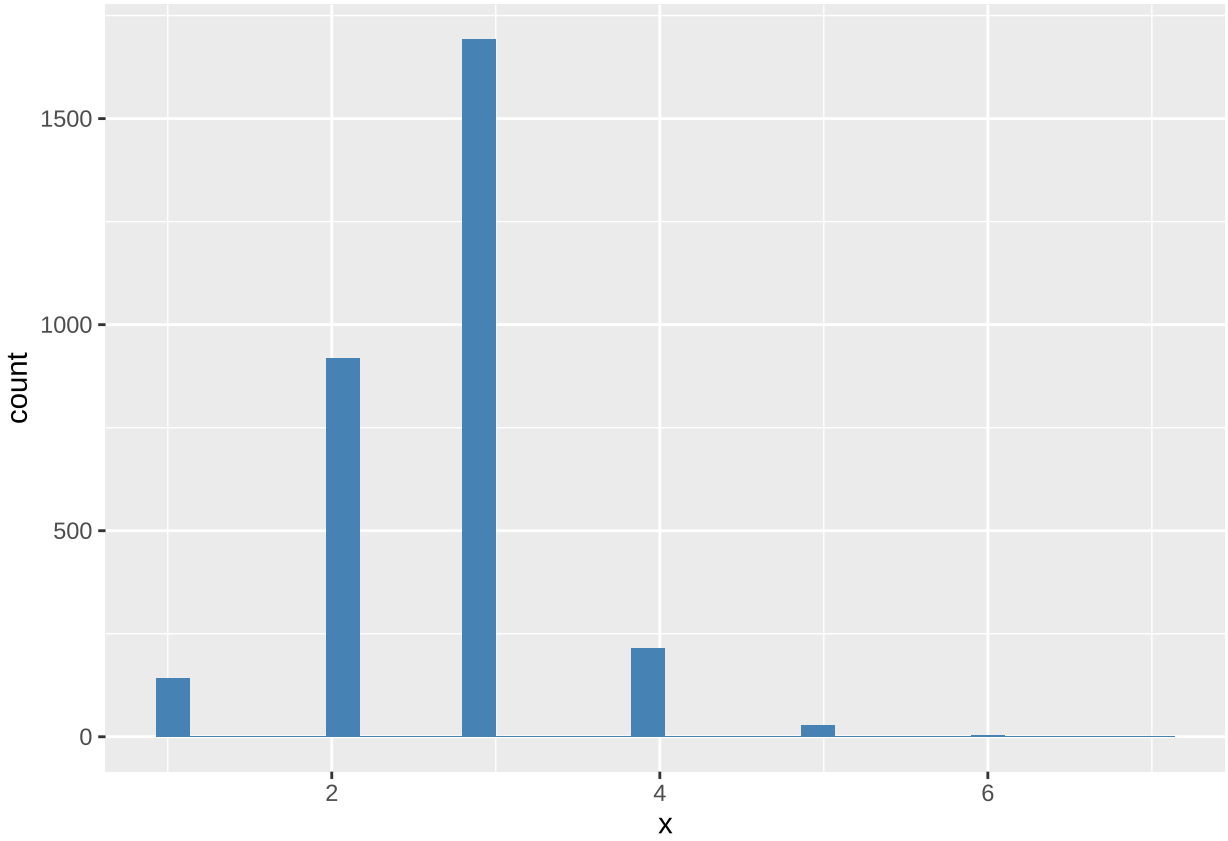


发现:

- 房屋单价跟房屋总价的特征基本一致

变量 (bedrooms) 的数值描述与图形

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.695   3.000   7.000
## range: 1 7NULL
## IQR: 1NULL
## var: 0.5328193NULL
## sd: 0.7299447NULL
## skewness: 0.1356027NULL
## kurtosis: 1.635711NULL
```

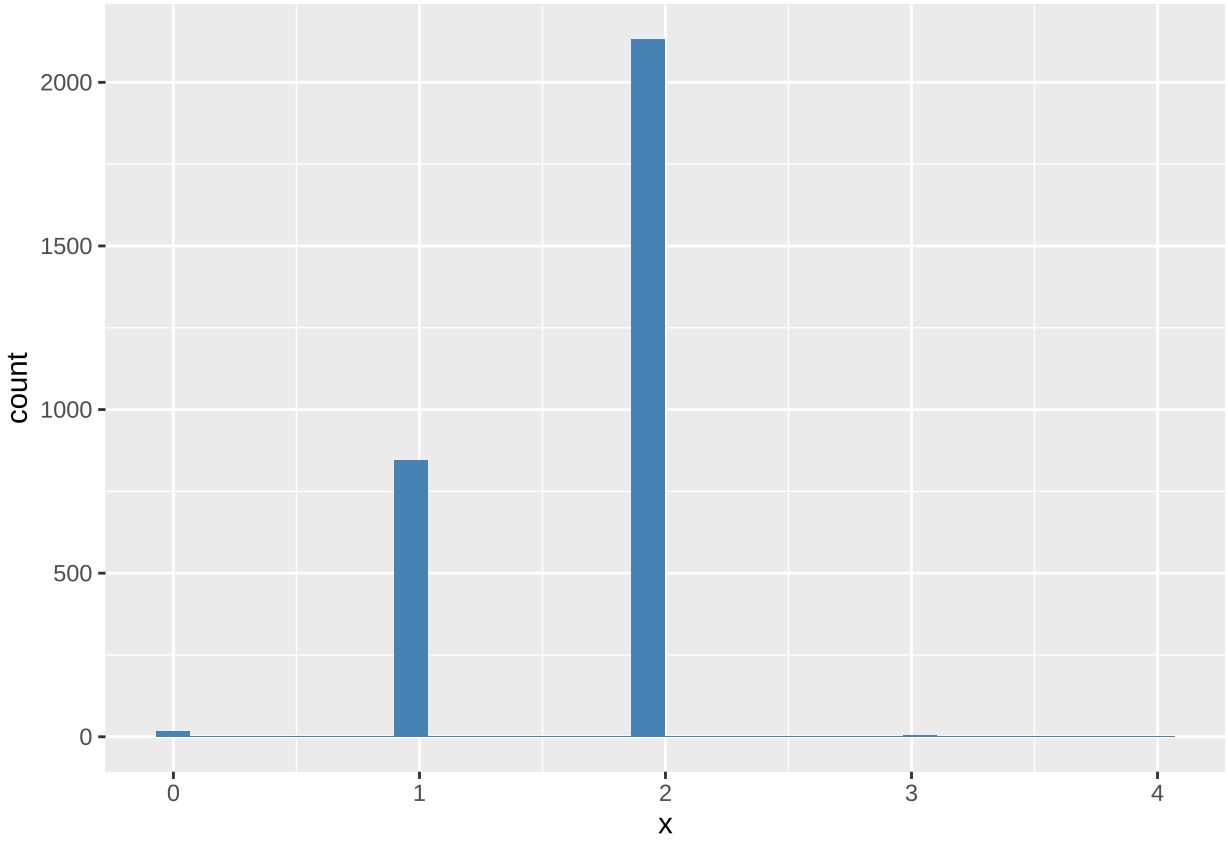


发现:

- 房源 2 房和 3 房居多

变量 (livingrooms) 的数值描述与图形

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   1.709   2.000   4.000
## range: 0 4NULL
## IQR: 1NULL
## var: 0.2238662NULL
## sd: 0.473145NULL
## skewness: -0.991422NULL
## kurtosis: -0.1840918NULL
```

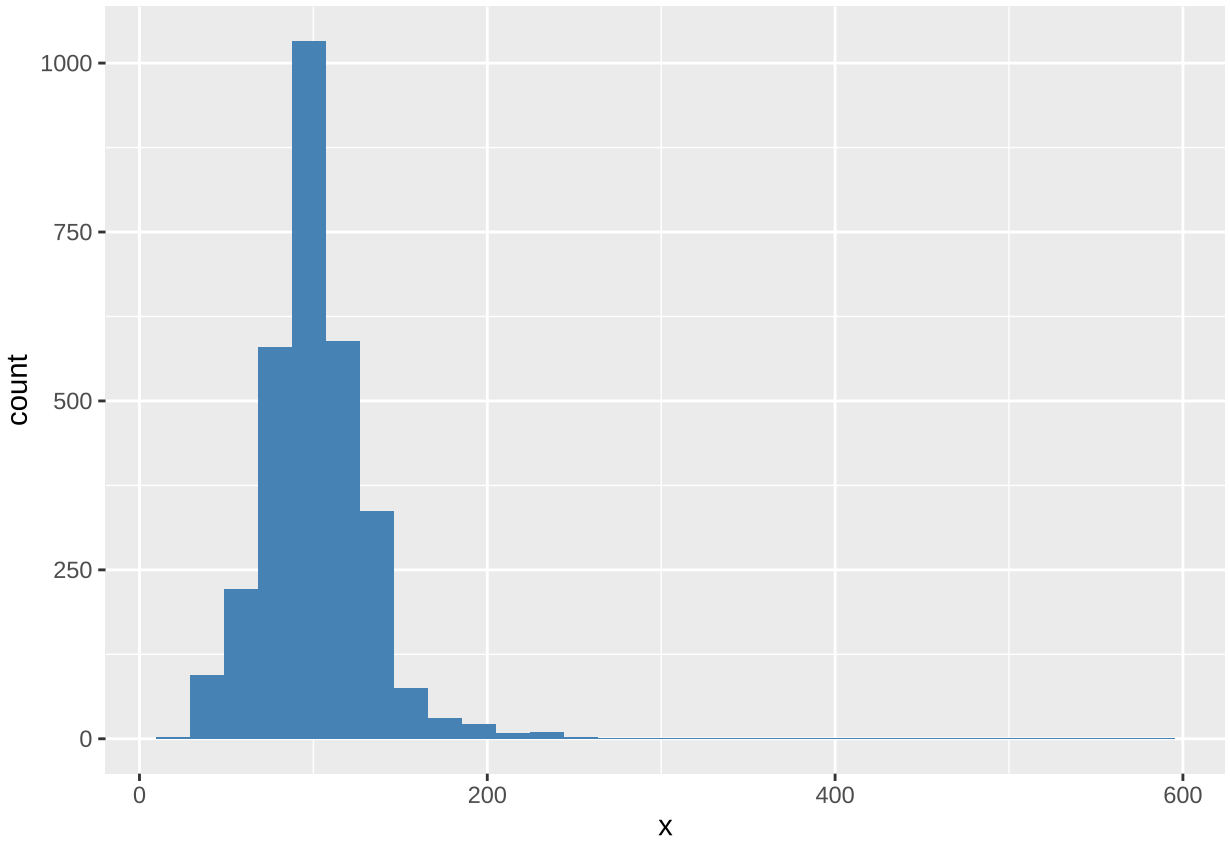


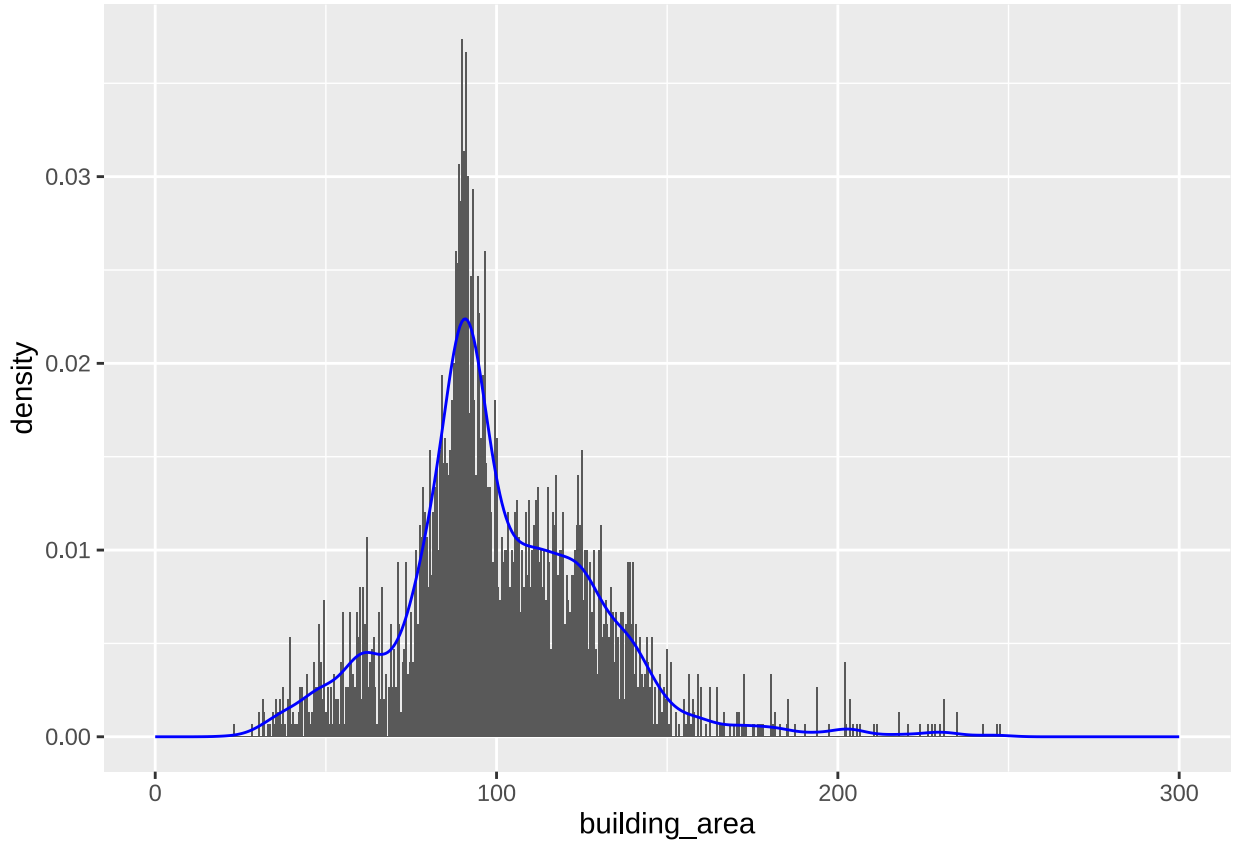
发现：

- 大部分是 2 厅或者 1 厅的房源，2 厅较多，其次是 1 厅

变量 (building_area) 的数值描述与图形

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.77   84.92   95.55  100.87  117.68  588.66
## range: 22.77 588.66NULL
## IQR: 32.7625NULL
## var: 922.9442NULL
## sd: 30.38NULL
## skewness: 2.079785NULL
## kurtosis: 23.63585NULL
```

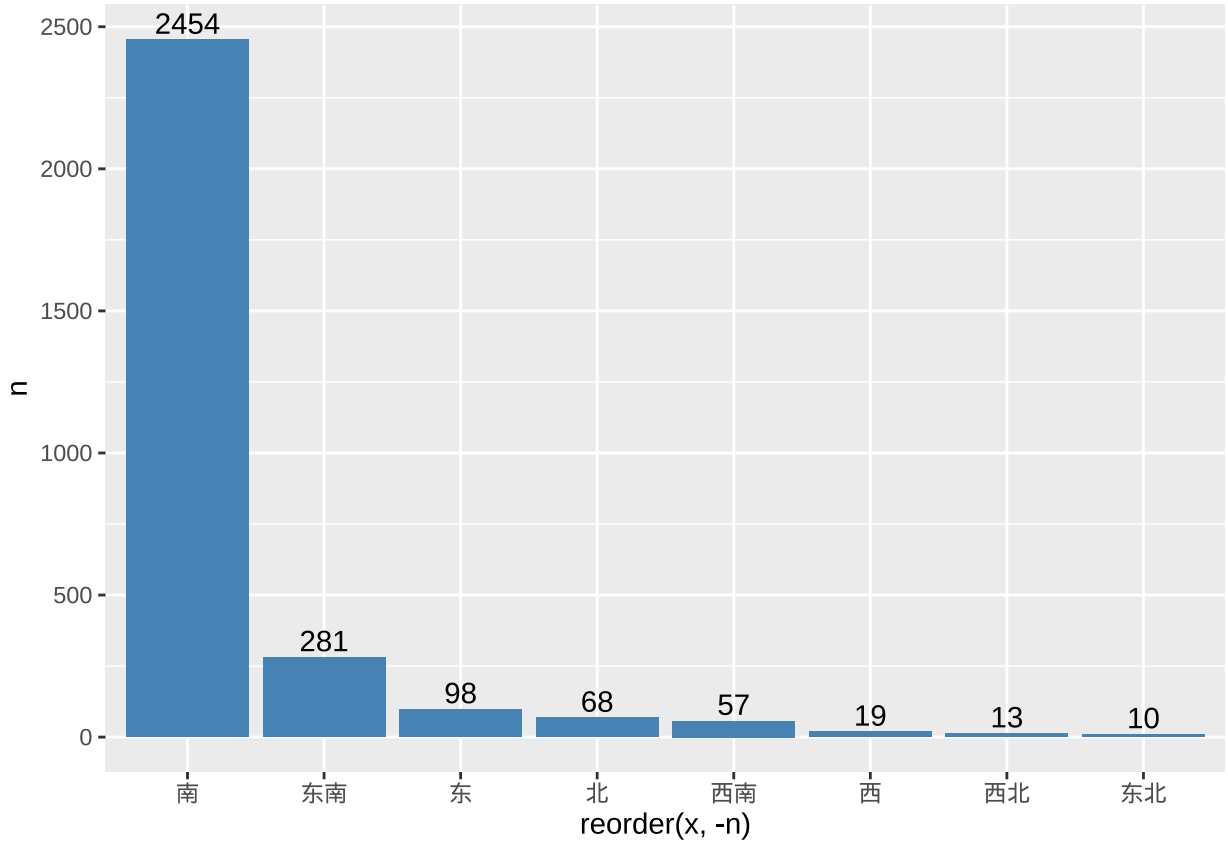

1000
750
500
250
0
count
0
200
400
600
x



发现:

- 房屋面积分布呈正态分布

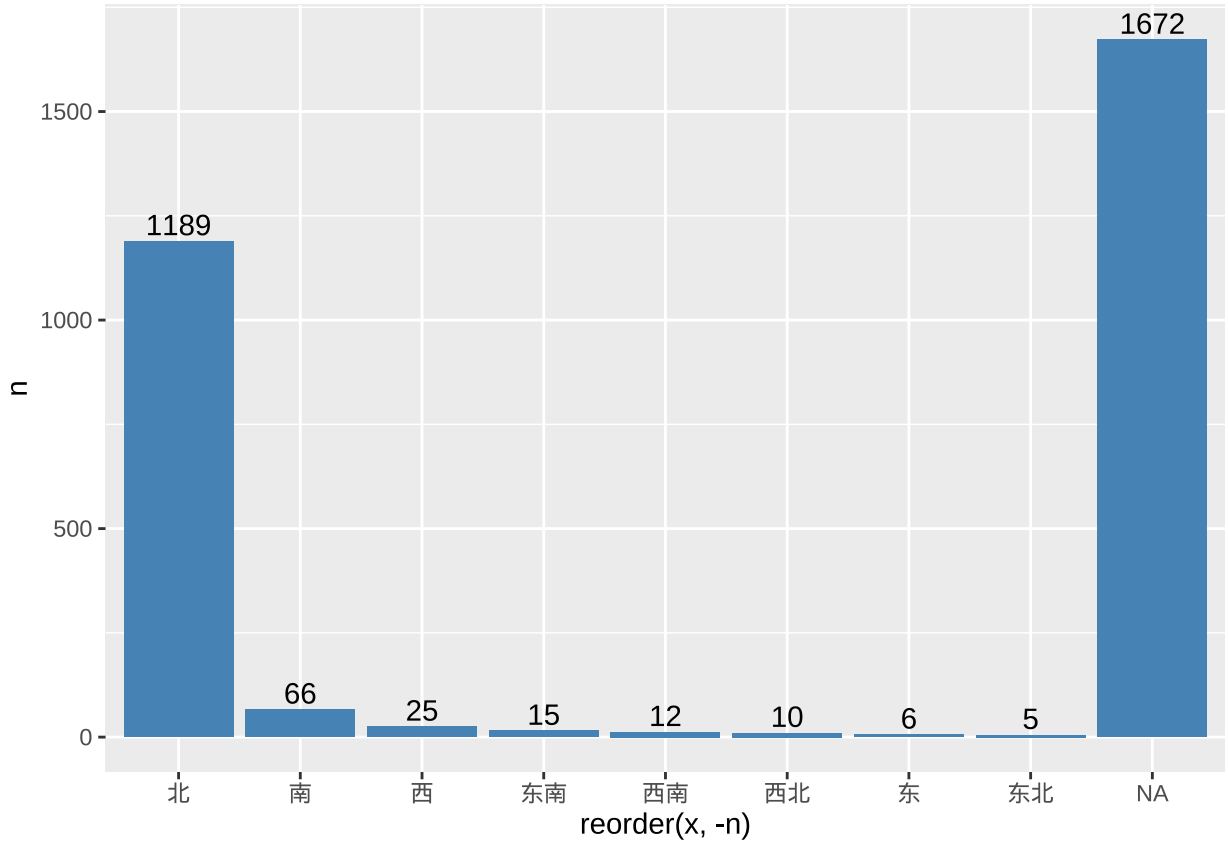
变量 (directions1) 的数值描述与图形



发现:

- 主要朝向 80% 朝南

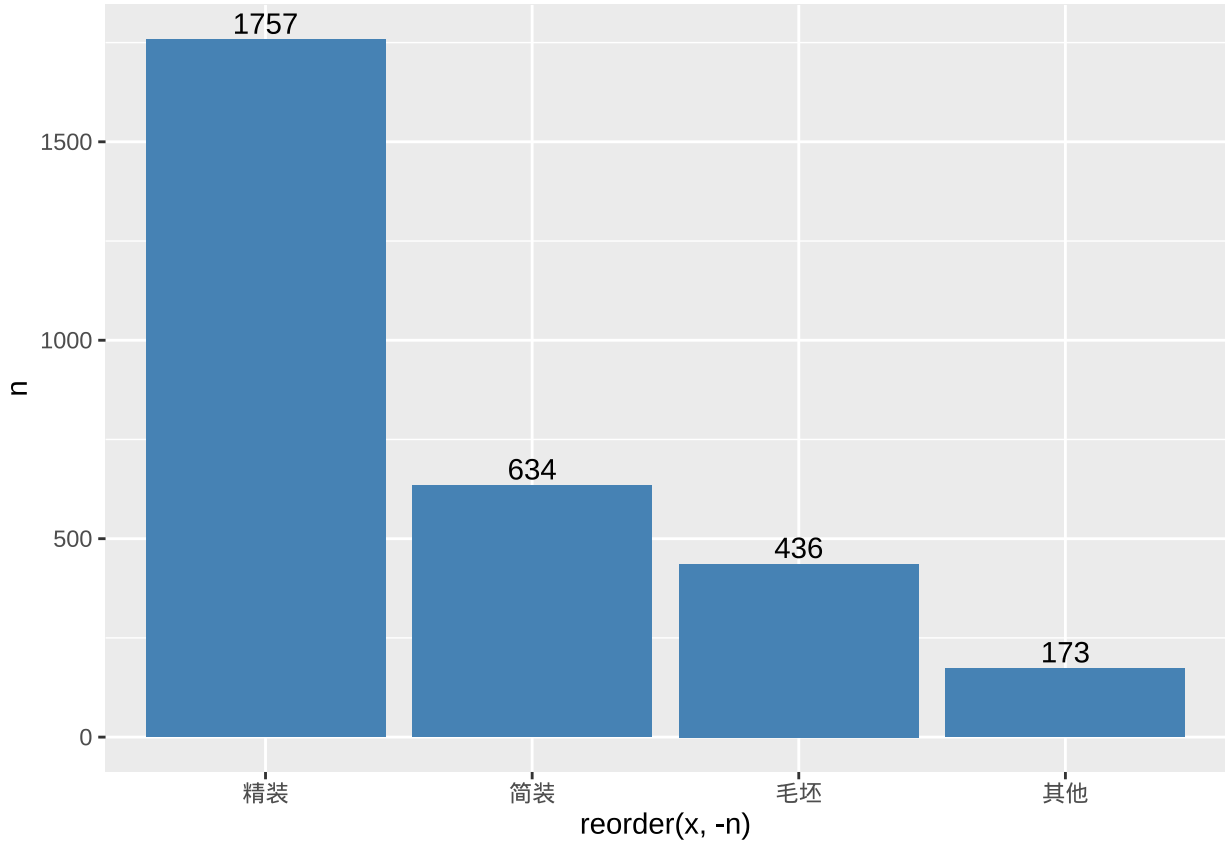
变量 (directions2) 的数值描述与图形



发现:

- 次要朝向 55% 的数据缺失，剩余中有 90% 朝北

变量 (decoration) 的数值描述与图形

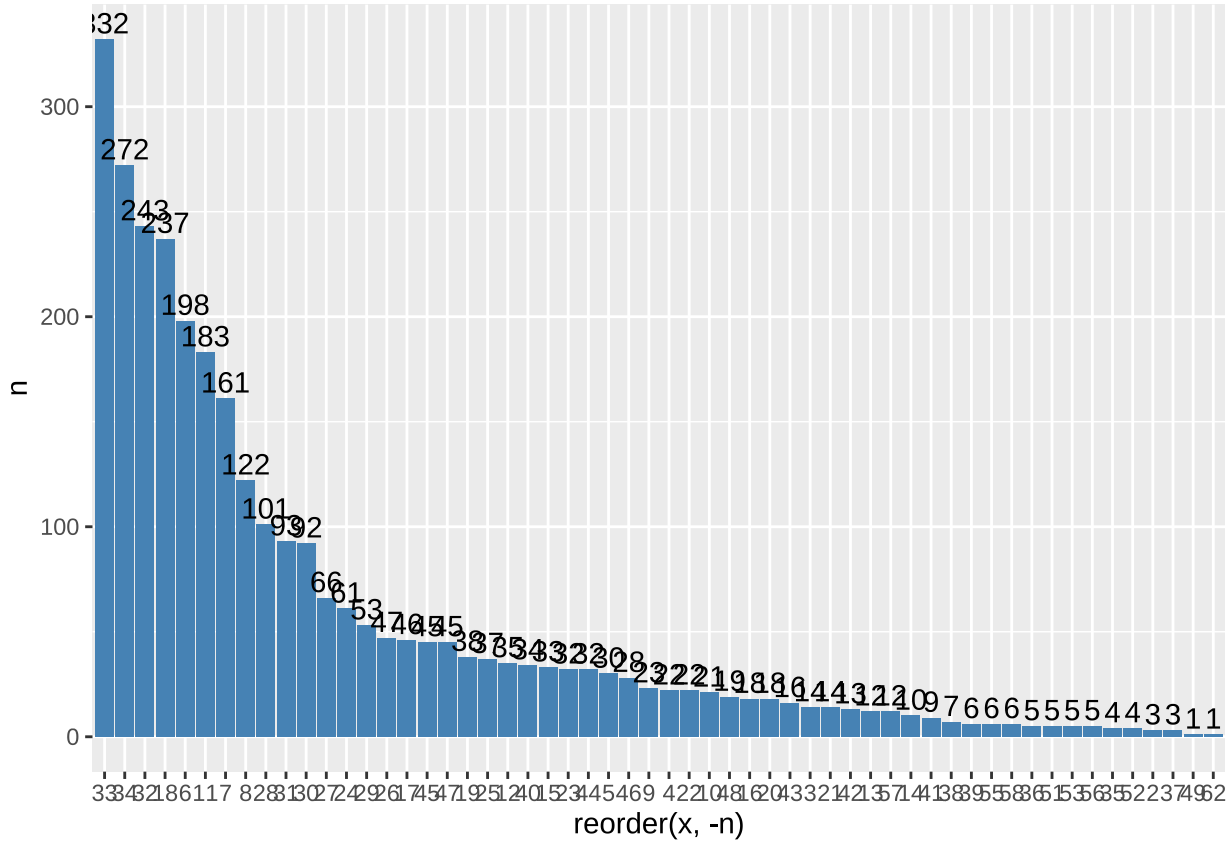


发现:

- 接近 60% 的房子为精装，20% 简装，14% 毛坯

变量 (property__t__height) 的数值描述与图形

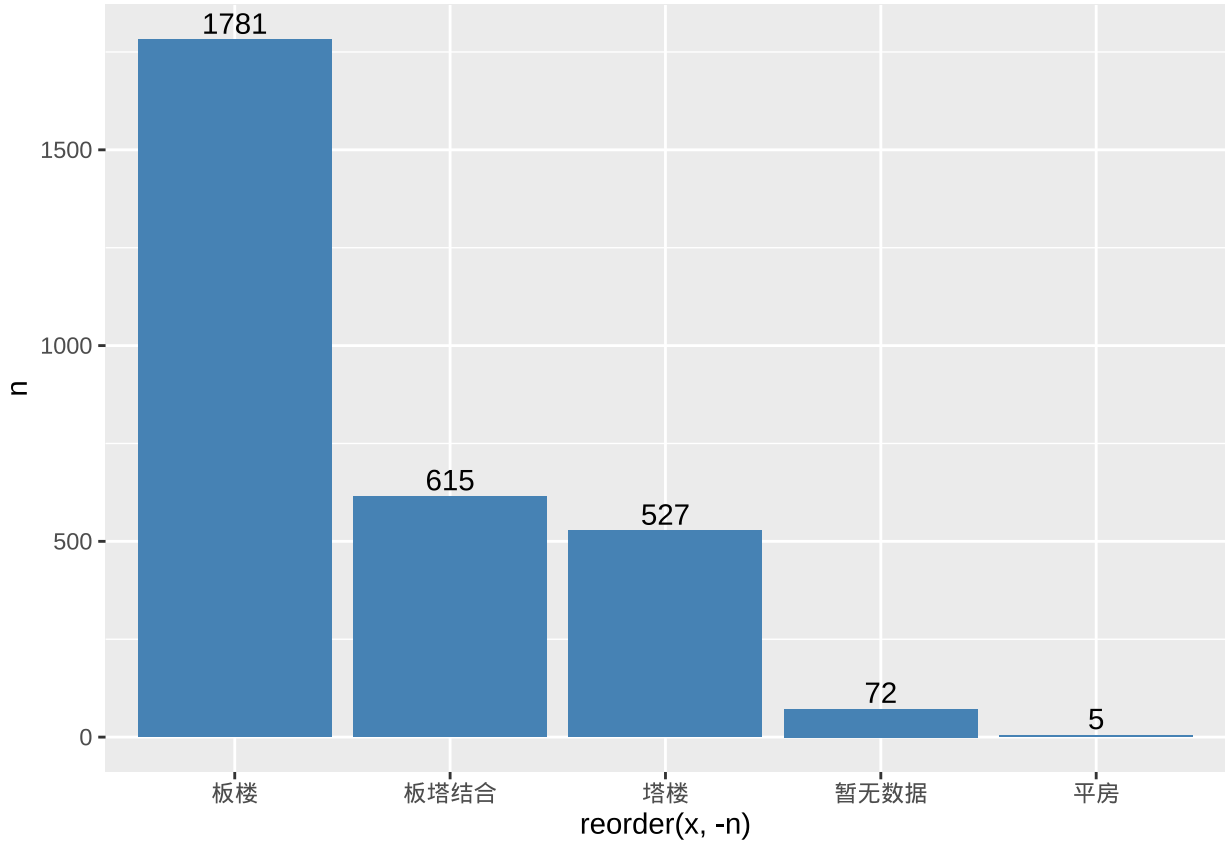
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   11.00   27.00   24.22   33.00   62.00
## range: 2 62NULL
## IQR: 22NULL
## var: 154.9588NULL
## sd: 12.44824NULL
## skewness: 0.04850289NULL
## kurtosis: -0.797201NULL
```



发现：

- 高层 32/33/34 层的楼房居多，其次是 18/6/11/楼的居多

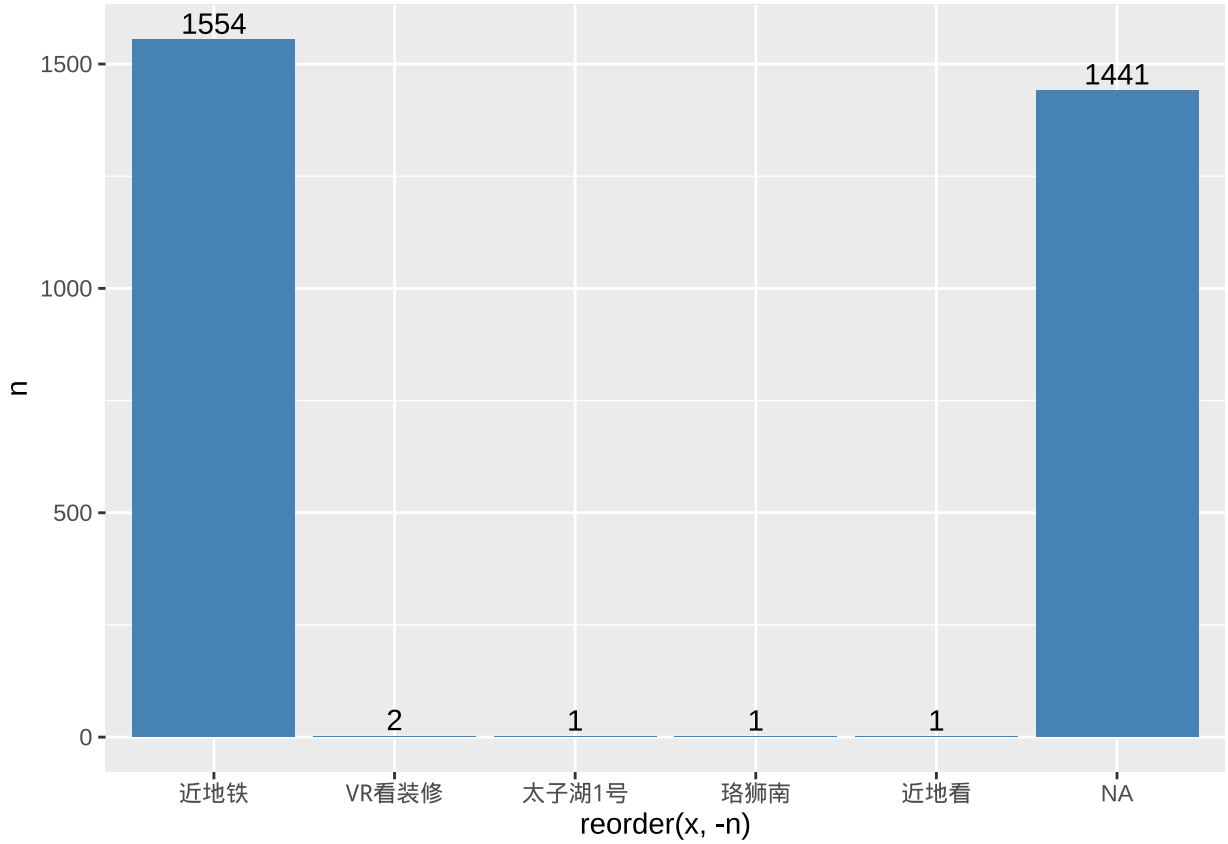
变量 (property_style) 的数值描述与图形



发现：

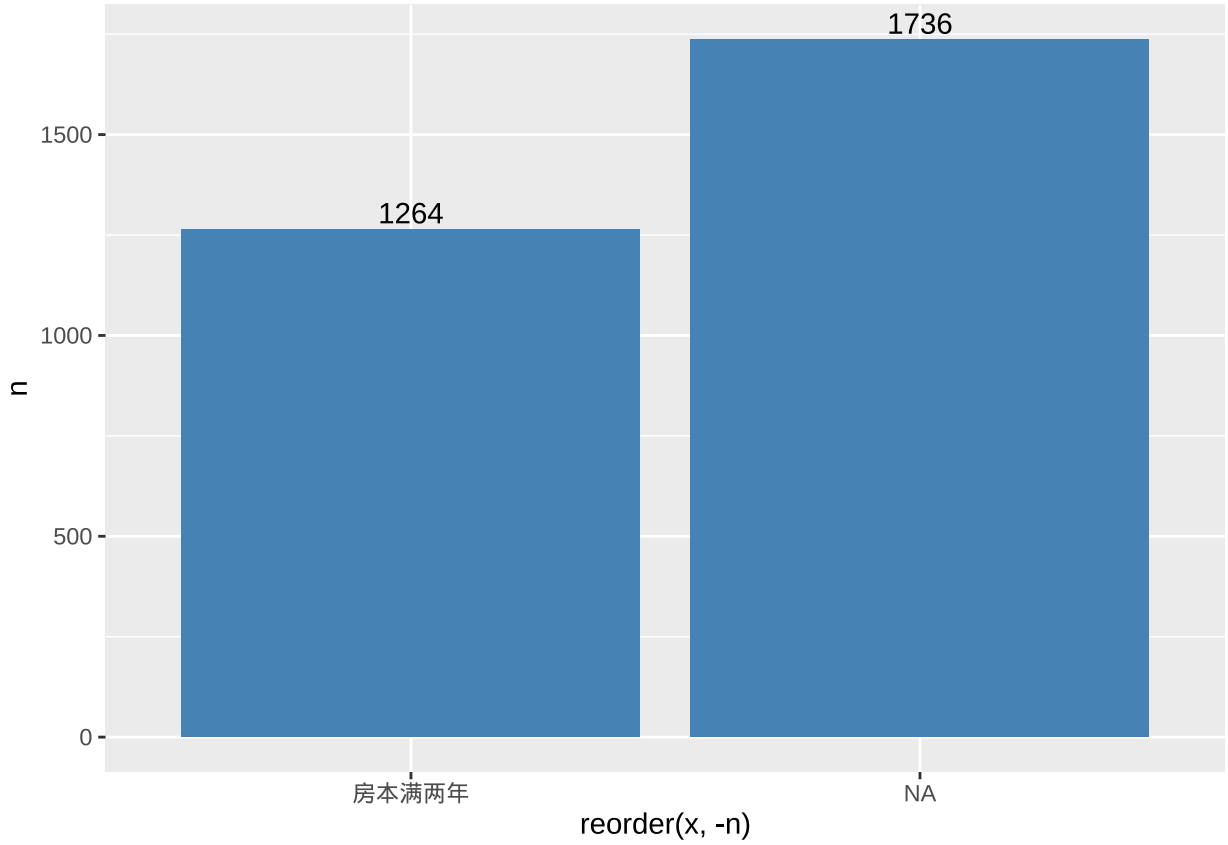
- 房源中建筑形式为板楼居多，共 1781 套，约 60%

变量 (near_subway) 的数值描述与图形



发现:

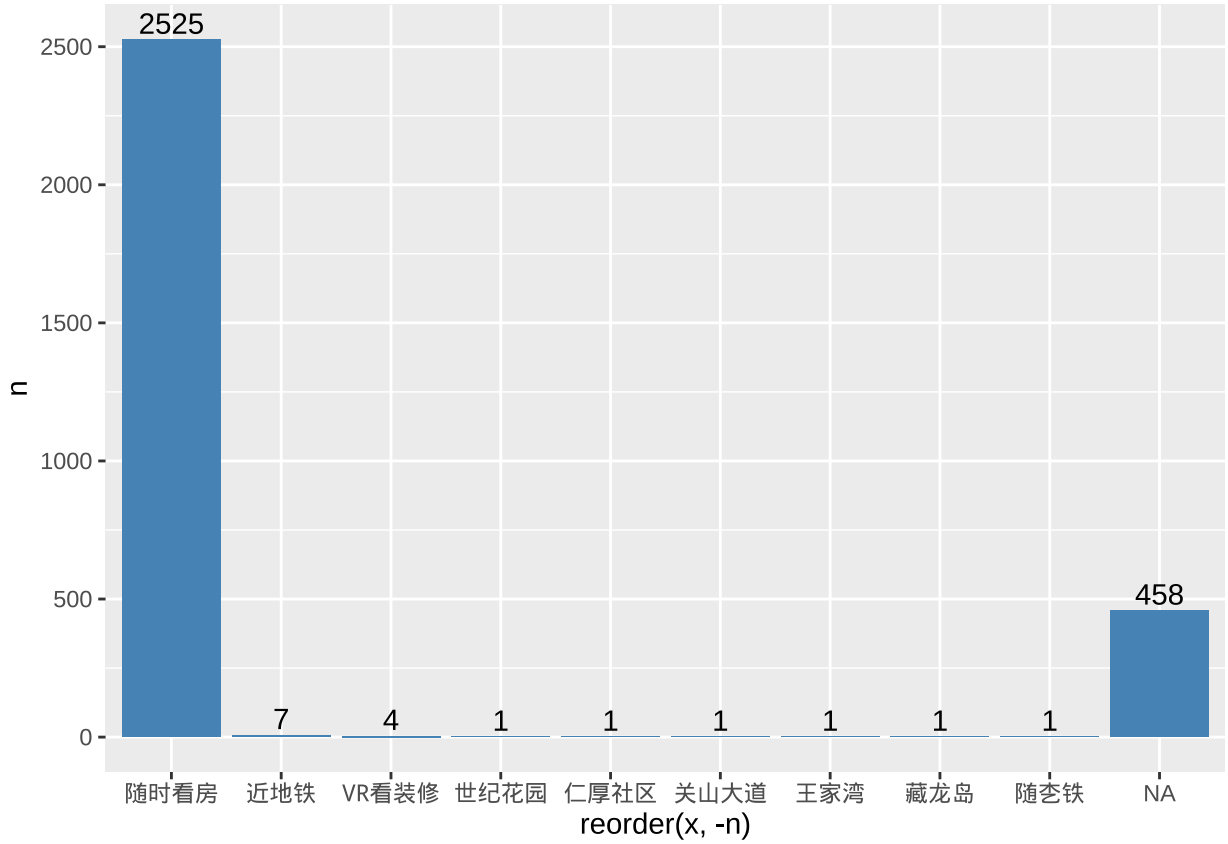
变量 (if_2y) 的数值描述与图形



发现:

- 满两年的房源共 1264 套，占比 42%

变量 (has_key) 的数值描述与图形



发现:

- 有钥匙可随时看房的有 2525 套，占比 84%

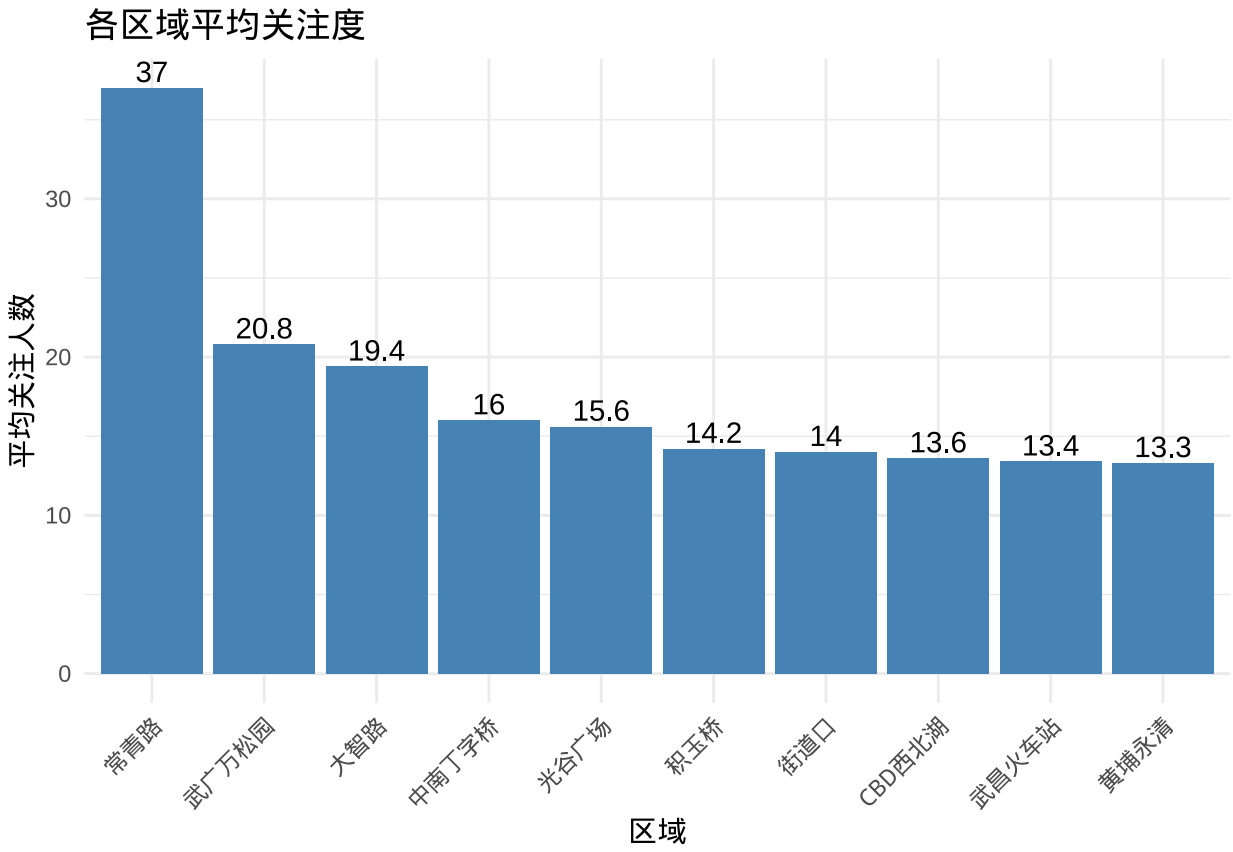
变量 (vr) 的数值描述与图形

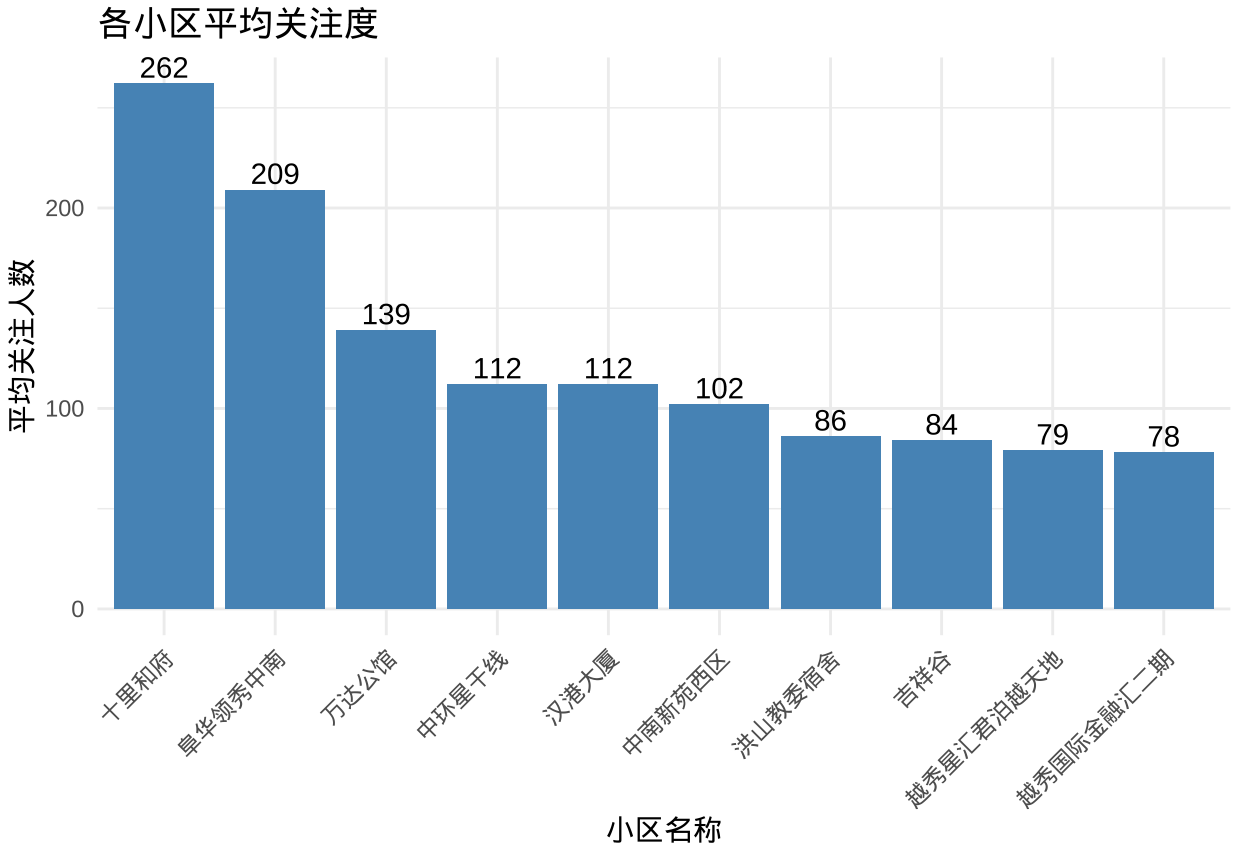
```
##          VR看装修 VR\xe7甸城\xe5\x8c      保利拉\xe8\x8f          塔子湖
##              2084                 1                 1                 1
##           江景湾            泓悦府          珞狮南路 育才花\xe6\xa1
##                1                 1                 1                 1
##           近地铁          随时看房 随时看\xe6\x88          NA's
##                1                 1                 1               906
```

发现:

- 可 VR 看装修的有 2084 套，占比 70%
- 数据中存在一些异常数据

探索问题 1：哪些区域和哪些小区最受欢迎

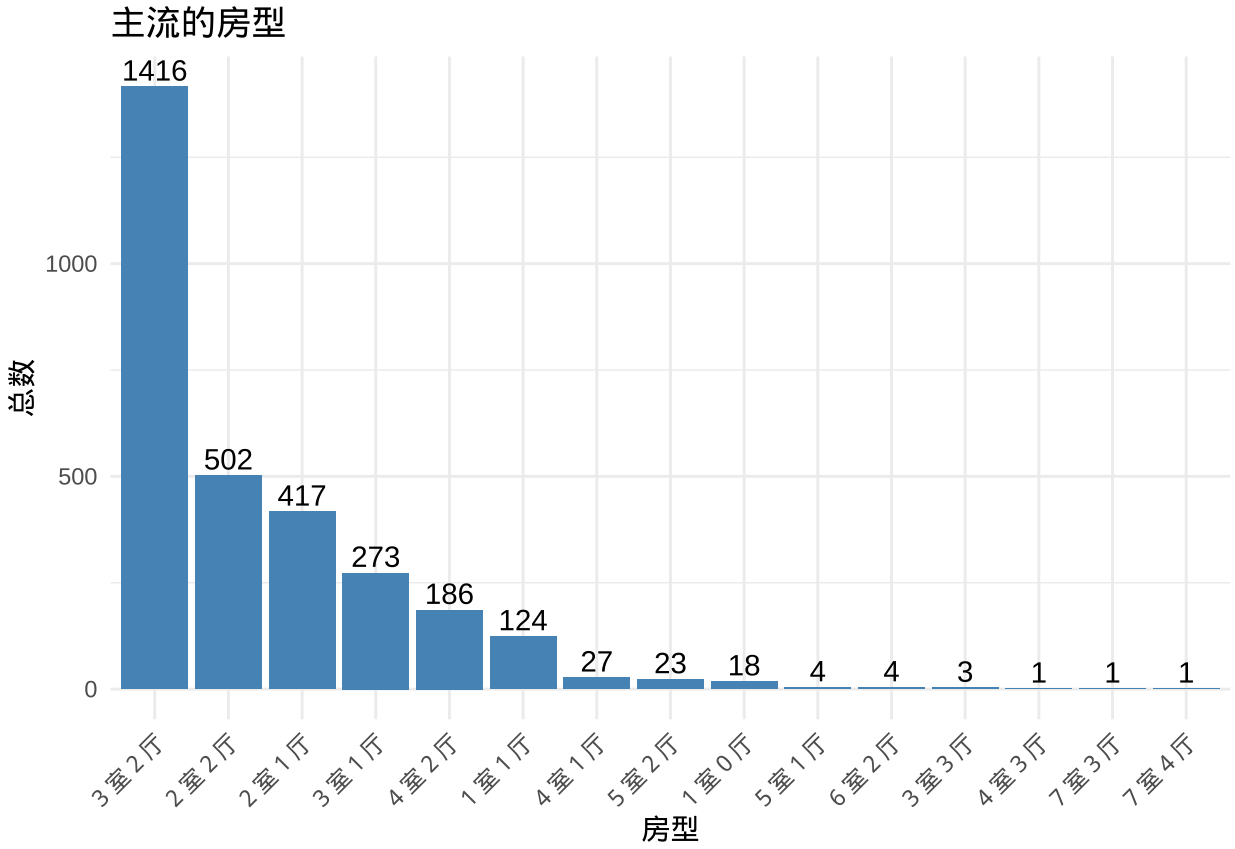




发现：

- 最受欢迎的区域位置依次为：常青路、武广万松园、大智路、中南丁字桥…
- 最后欢迎的小区依次为：十里合府、阜华领秀中南、万达公馆、中环星干线…

探索问题 2：市场上主流的房型是几室几厅



发现：

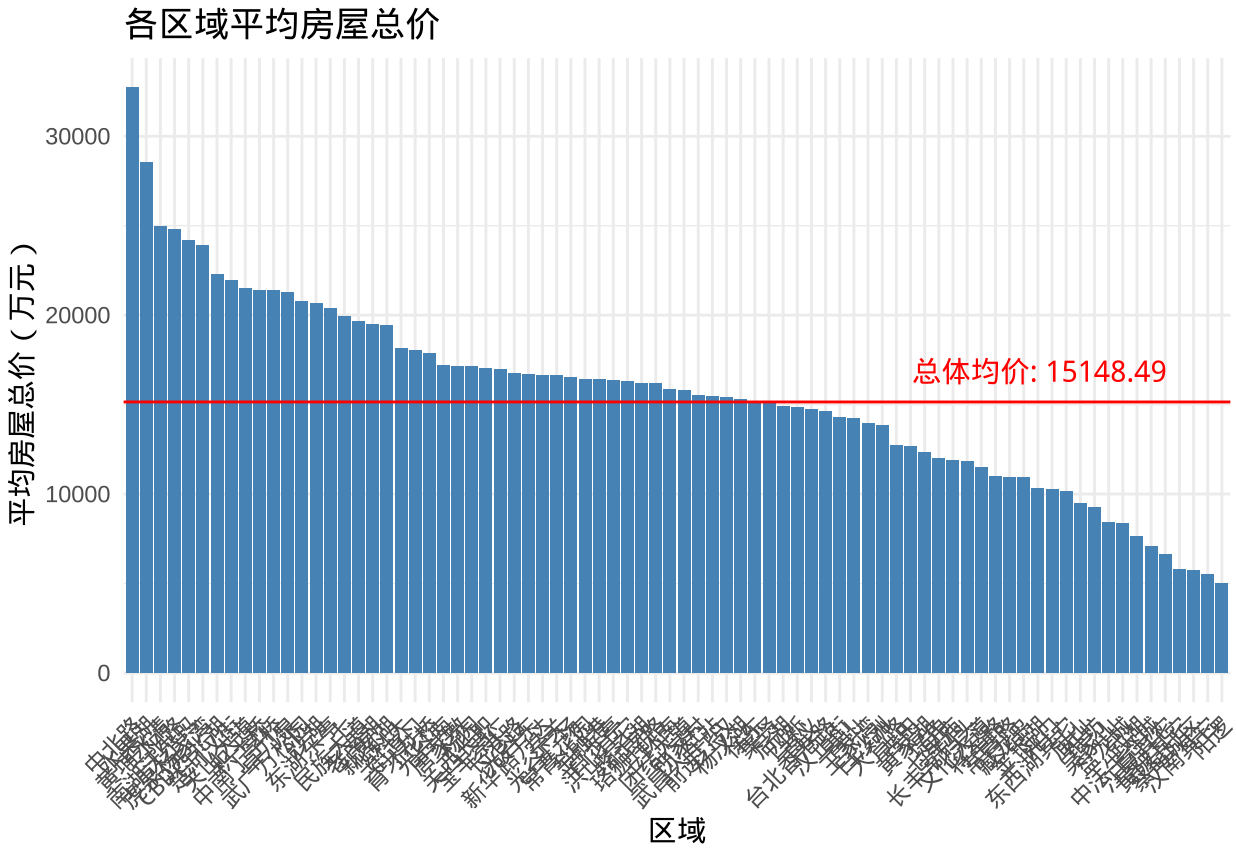
- 市面上主流的房型排名第一的是三室两厅，剩下的依次是 2 室 2 厅、2 个 1 厅，3 室 1 厅…

探索问题 3：区域与房价的关系

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

```
## # A tibble: 78 x 3
##    property_region avg_price_sqm   cnt
##    <fct>                   <dbl> <int>
##  1 中北路                 32728.    18
##  2 水果湖                 28562.     9
##  3 黄埔永清               24957.    23
##  4 三阳路                 24777.    16
##  5 南湖沃尔玛             24181.    33
```

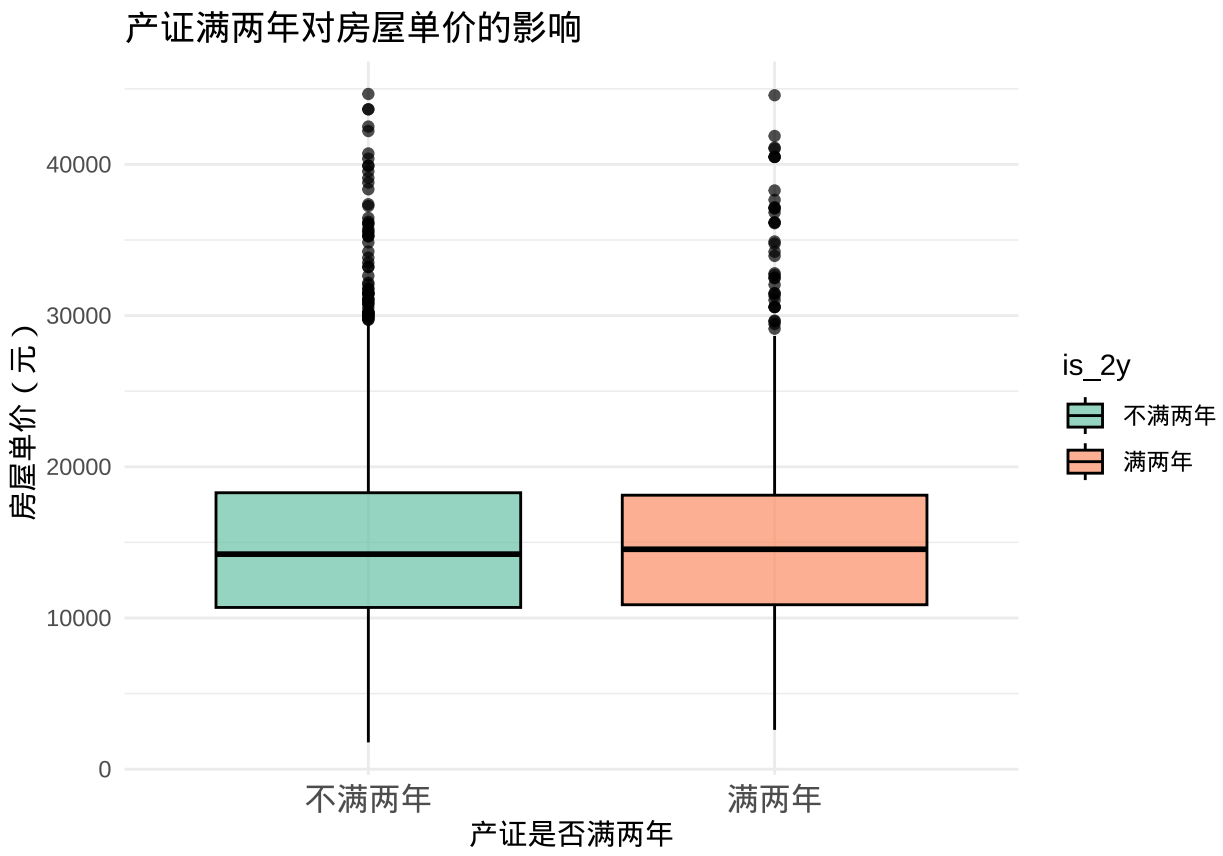
```
##  6 虎泉杨家湾                23902.    21
##  7 CBD西北湖                 22272.    35
##  8 楚河汉街                  21958.    15
##  9 关山大道                  21480.    25
## 10 积玉桥                    21403.    63
## # i 68 more rows
```



发现：

- 各区域的核心位置均价较高，依次是中北路（32727）、水果湖（28561）、黄埔永清（24956）、三阳路（24777）…
- 有一半区域的均价在总体均价上

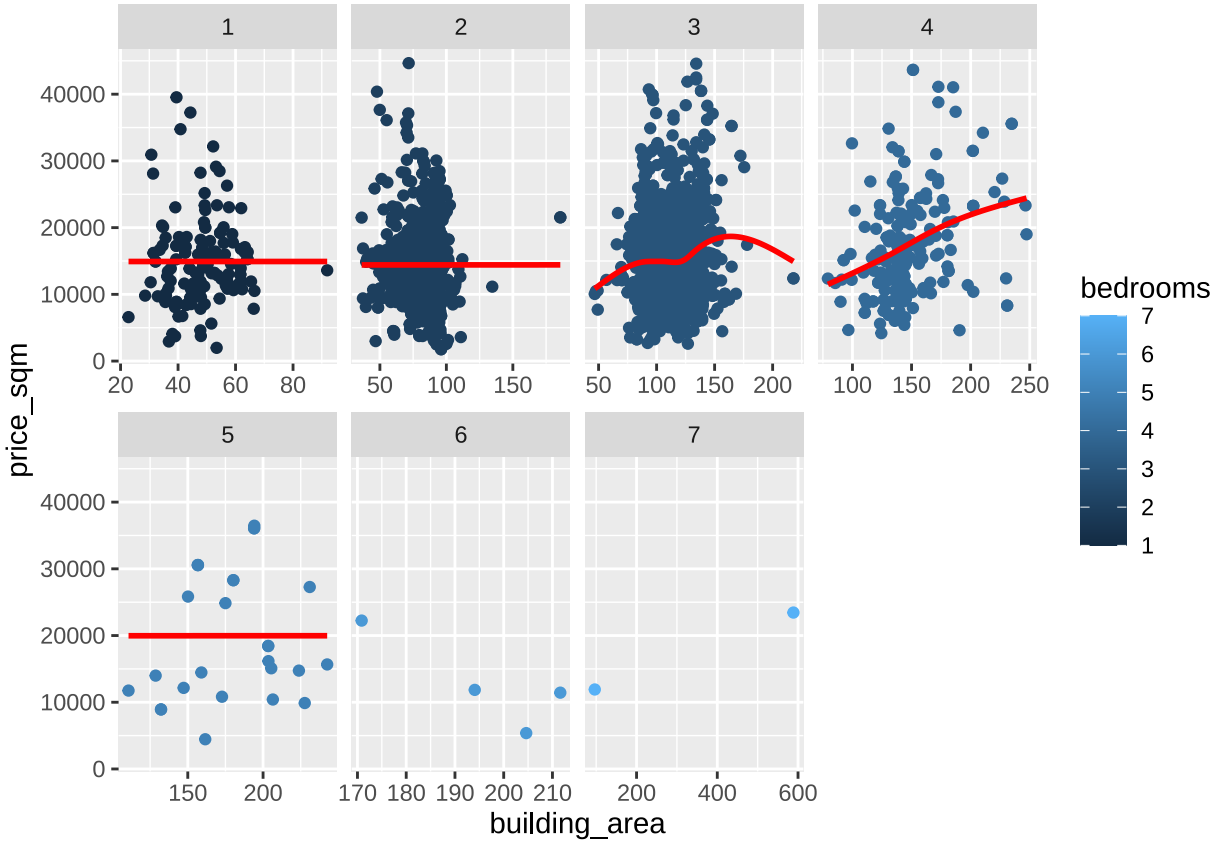
探索问题 4：产证满两年对房屋单价的影响



发现：

- 房屋的产证是否满两年对房屋的价格基本没有大的影响

探索问题 5：房子面积和房间数，对房屋单价的影响



发现：

- 三房和四房的面积与均价成正比，面积越大，房子的单价越高

发现总结

1. 购房者对位置比较看重，核心区域的房子比较受欢迎，关注量多，均价也高；
2. 三室两厅是市面上主流的户型，比较受购房者的青睐；
3. 产证是否满二对房屋单价的影响不大；
4. 小房子的均价差异不大，大房子主要是改善，房子的面积越大，单价越高。