

# 2023 年链家二手房数据分析报告

胡渝

2024-10-31

## 你的主要发现

- 1. 价格偏高的二手房所在区域与挂牌数量多的区域没有重合
- 2. 普遍认为的房屋朝向对房屋单价的影响，在样本数据中表现的不明显
- 3. 黄埔永清、CBD 西北路、中北路，这三个属于比较热门的片区，单价高而且关注人数不低

## 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：

Table 1: 武汉链家二手房

| 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    |
|---------------|-----------------|-----------|-----------|----------|-------------|---------------|-------------|-------------|------------|-------------------|-----------------|----------------|-----------|-------------|-------|---------|-------|
| 南湖名都A区        | 南湖沃尔玛           | 237.01    | 8709      | 3        | 1           | 126.68        | 南           | 北           | 精装         | 17                | 中               | 塔楼             | 3         | 近地铁         | NA    | 随时看房    | NA    |
| 万科紫悦湾         | 光谷东             | 127.01    | 4613      | 3        | 2           | 86.91         | 南           | NA          | 精装         | 28                | 中               | 板楼             | 1         | NA          | 房本满两年 | 随时看房    | VR看装修 |

| property       | property | price | reg   | file | bed | div    | sqm | building | direction | distance | ratio | property | property | type | height | full | style | year | if | days | key |
|----------------|----------|-------|-------|------|-----|--------|-----|----------|-----------|----------|-------|----------|----------|------|--------|------|-------|------|----|------|-----|
| 东立国际           | 二七       | 75.0  | 15968 | 1    | 1   | 46.97  | 南   | NA       | 简装        | 18       | 低     | 塔楼       | 3        | 近地铁  | NA     | 随时看房 | NA    |      |    |      |     |
| 新都汇            | 光谷广场     | 188.0 | 15702 | 3    | 2   | 119.73 | 北   | 东        | 精装        | 32       | 高     | 塔楼       | 2        | 近地铁  | 房本满两年  | 随时看房 | NA    |      |    |      |     |
| 保利城一期          | 团结大道     | 182.0 | 17509 | 3    | 2   | 103.95 | 东南  | NA       | 简装        | 34       | 中     | 板塔结合     | 3        | NA   | 房本满两年  | 随时看房 | VR看装修 |      |    |      |     |
| 加州橘郡           | 庙山       | 122.0 | 10376 | 3    | 2   | 117.59 | 南   | 北        | 精装        | 34       | 低     | 板楼       | 1        | NA   | 房本满两年  | 随时看房 | NA    |      |    |      |     |
| 省建筑五公司西区保利上城东区 | 光谷广场白沙洲  | 99.0  | 12346 | 2    | 1   | 80.19  | 南   | NA       | 简装        | 7        | 低     | 板楼       | 0        | 近地铁  | NA     | 随时看房 | VR看装修 |      |    |      |     |
|                |          | 193.8 | 16336 | 3    | 2   | 118.64 | 南   | 北        | 其他        | 34       | 中     | 板塔结合     | 0        | 近地铁  | 房本满两年  | 随时看房 | NA    |      |    |      |     |
| 石化大院           | 中南丁字桥    | 325.0 | 32631 | 4    | 1   | 99.60  | 南   | 北        | 简装        | 5        | 低     | 板楼       | 2        | 近地铁  | NA     | 随时看房 | NA    |      |    |      |     |
| 阳光花园           | 杨汊湖      | 192.0 | 17403 | 3    | 2   | 110.33 | 南   | 北        | 其他        | 7        | 低     | 板楼       | 0        | 近地铁  | 房本满两年  | 随时看房 | NA    |      |    |      |     |

各变量的简短信息：

各变量的简短统计：

可以看到：

- 房屋总价跨度从 10.6w 到 1380w, 整体跨度区间极大。第一分位数 (95w) 和第三分位数 (188w), 估计大部分房屋总价集中在 100-200w 之间, 房屋单价分布在 10000-20000 居多。
- 从房屋挂牌数上看, 市场上比较多的房屋类型是: 精装、南北朝向、中楼层、板楼
- 从房屋房间数看, 大部分是 2 室或 3 室

## 探索性分析

### price\_ttl 的数值描述与图形

Table 2: price\_ttl 数值分析

| Statistic          | Value            |
|--------------------|------------------|
| Mean               | 155.86278        |
| Median             | 137              |
| Mode               | 105              |
| Variance           | 9129.44477979821 |
| Standard Deviation | 95.5481280810786 |
| Range              | 10.6 to 1380     |
| IQR                | 93               |
| Skewness           | 2.75322271453055 |
| Kurtosis           | 16.11671561457   |

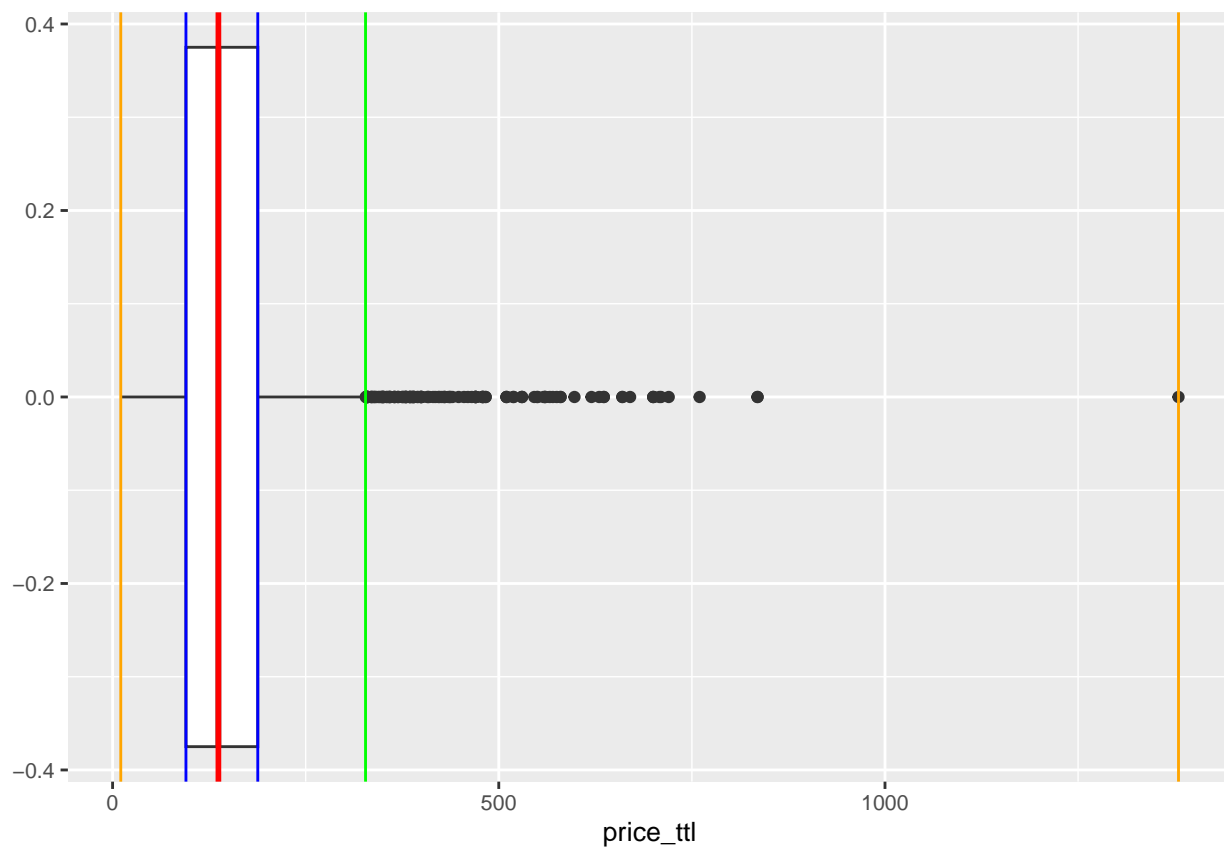


Table 3: 挂牌数前 5 名的区域

| property_region | count |
|-----------------|-------|
| 白沙洲             | 167   |
| 盘龙城             | 126   |
| 四新              | 116   |
| 光谷东             | 112   |

| property_region | count |
|-----------------|-------|
| 金银湖             | 97    |

Table 4: 高离散值前 5 名的区域

| property_region | count |
|-----------------|-------|
| 积玉桥             | 21    |
| 中北路             | 11    |
| CBD 西北湖         | 9     |
| 中南丁字桥           | 9     |
| 黄埔永清            | 8     |

发现:

- 二手房均价为 155.9w, IQR 为 93w, 数据相对还是比较集中, 50% 的数据在 95w-188w 之间。
- 偏度 >0, 长尾在右侧, 较多总价集中在左侧, 这与线框图的图形特征相符。从箱线图看, 右须较长, 存在较多的离散值
- 挂牌数量前 5 位的区域是白沙洲、盘龙城、四新、光谷东、金银湖; 而价格偏高的房屋所在区域, 排在前五的是: 积玉桥、中北路、CBD 西北湖、中南丁字桥、黄埔永清。价格高的房屋都不在挂牌数量多的区域。

## property\_name 的数值描述与图形

Table 5: 小区挂牌数 top10

| property_name | count |
|---------------|-------|
| 东立国际          | 22    |
| 保利中央公馆        | 16    |
| 朗诗里程          | 16    |
| 恒大名都          | 15    |
| 阳光 100 大湖第    | 15    |
| 保利城一期         | 13    |
| 锦绣龙城          | 13    |
| 保利华都          | 12    |
| 统建同安家园        | 12    |
| 金地自在城 K2      | 12    |

Table 6: 小区关注度 top10

| property_name | count |
|---------------|-------|
| 中环星干线         | 336   |
| 万达公馆          | 278   |
| 十里和府          | 262   |
| 米兰映象          | 241   |
| 联投花山郡一期 (香颂)  | 236   |
| 江汉人家          | 215   |
| 保利心语六期        | 212   |
| 阜华领秀中南        | 209   |

| property__name | count |
|----------------|-------|
| 中建东湖锦城         | 208   |
| 保利中央公馆         | 197   |

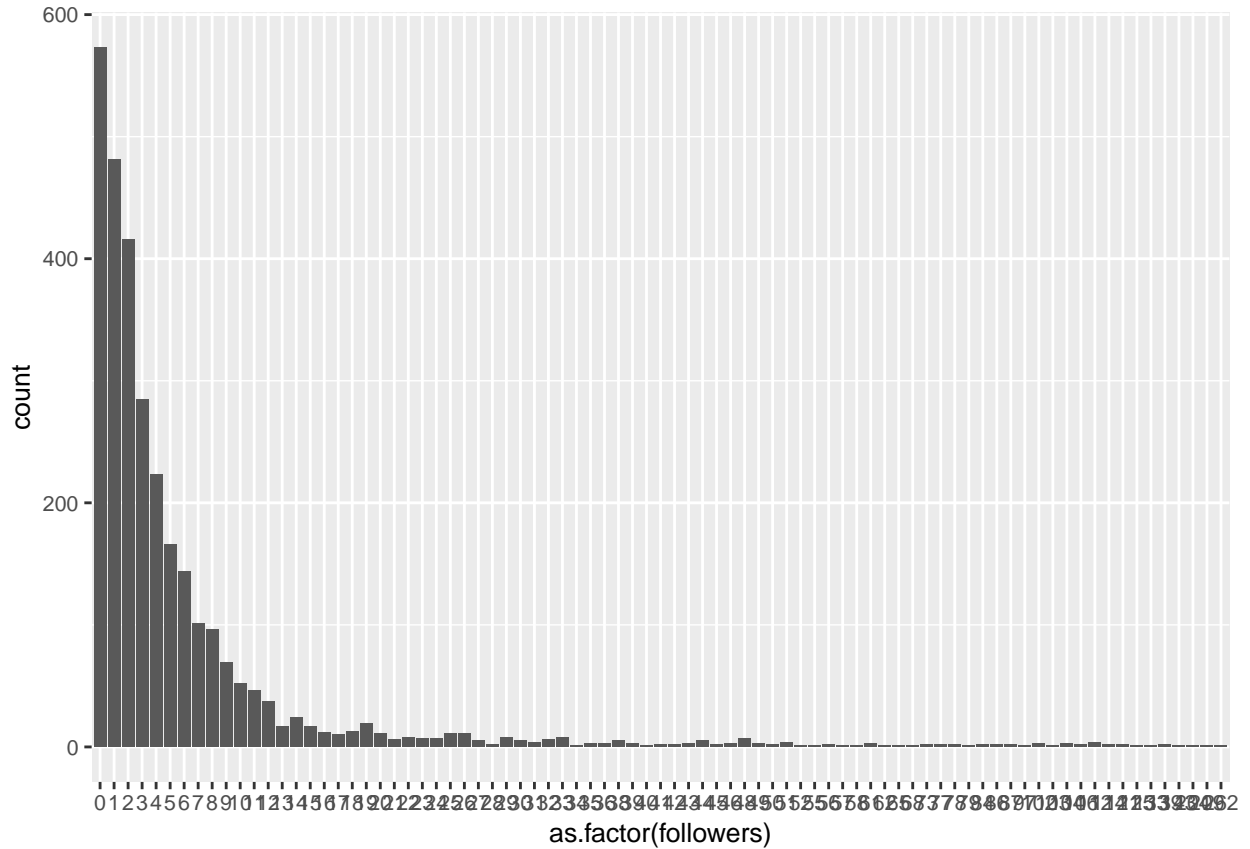
发现:

- 挂牌小区词条出现最多的是” 国际”
- 小区挂牌数前 3 位是: 东立国际、保利中央公馆、朗诗里程
- 小区关注度最高的前 3 位是: 中环星干线、万达公馆、十里河府

## 变量 followers 的数值描述与图形

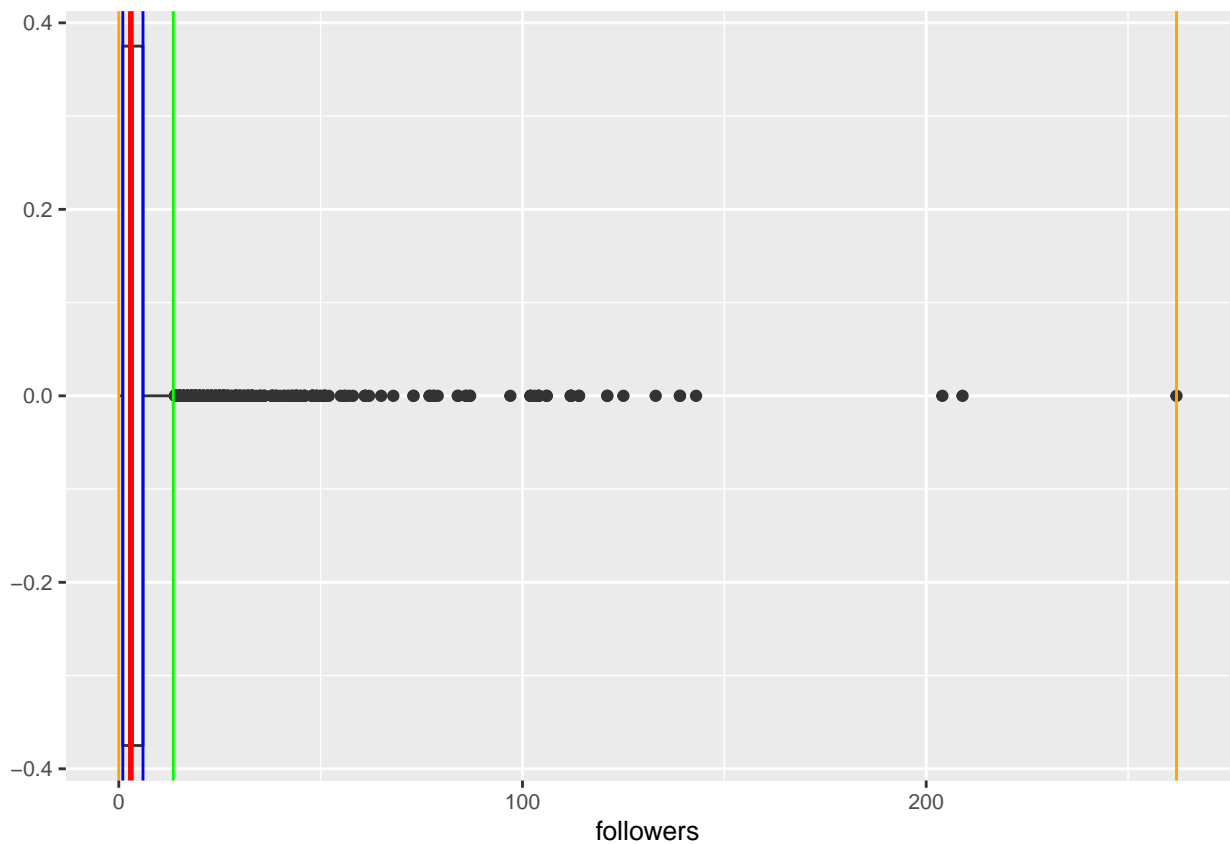
Table 7: followers 数值分析

| Statistic          | Value            |
|--------------------|------------------|
| Mean               | 6.6136666666667  |
| Median             | 3                |
| Mode               | 0                |
| Variance           | 231.557599088586 |
| Standard Deviation | 15.2170167604753 |
| Range              | 0 to 262         |
| IQR                | 5                |
| Skewness           | 6.90076729401956 |
| Kurtosis           | 68.1705707707924 |



```
## <ggproto object: Class ScaleDiscretePosition, ScaleDiscrete, Scale, gg>
##   aesthetics: x xmin xmax xend
##   axis_order: function
##   break_info: function
##   break_positions: function
##   breaks: waiver
##   call: call
##   clone: function
##   dimension: function
##   drop: TRUE
##   expand: waiver
##   get_breaks: function
##   get_breaks_minor: function
##   get_labels: function
##   get_limits: function
##   get_transformation: function
##   guide: waiver
##   is_discrete: function
##   is_empty: function
##   labels: waiver
##   limits: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 ...
##   make_sec_title: function
##   make_title: function
##   map: function
##   map_df: function
##   n.breaks.cache: NULL
```

```
##      na.translate: TRUE
##      na.value: NA
##      name: waiver
##      palette: function
##      palette.cache: NULL
##      position: bottom
##      range: environment
##      range_c: environment
##      rescale: function
##      reset: function
##      train: function
##      train_df: function
##      transform: function
##      transform_df: function
##      super: <ggproto object: Class ScaleDiscretePosition, ScaleDiscrete, Scale, gg>
```

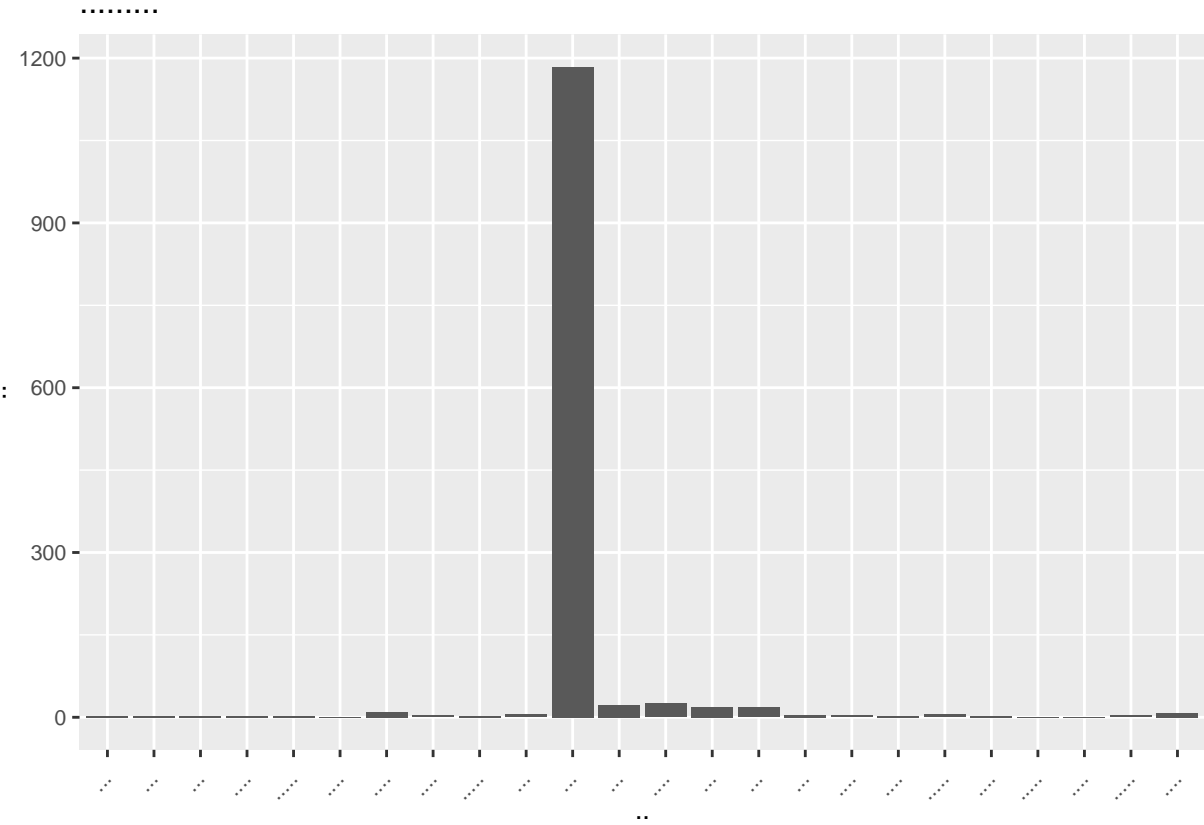


```
## [1] 4.004219
```

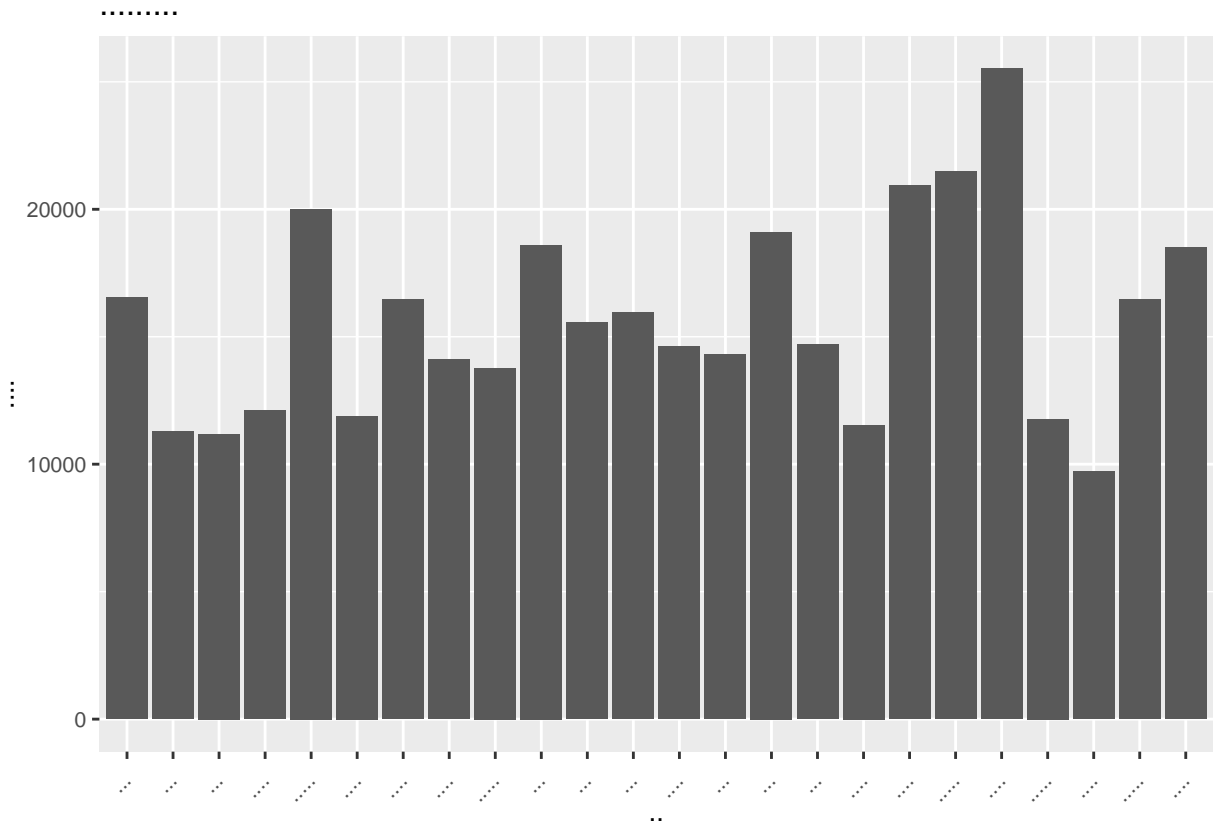
发现:

- 房屋关注人数范围从 0-262, 差别极大
- 偏度 >0, 长尾在右侧, 较多总价集中在左侧, 这与线框图的图形特征相符。从箱线图看, 右须较长, 存在较多的离散值
- 整体的关注人数平均值为 6.6, 除去离散点后的平均值为 4, 两者相差较大, 说明在数据分析过程中不仅要看整体, 也需要分析是否存在个别离散数据对整体造成较大影响

探索问题 1: 房屋朝向对单价是否存在影响





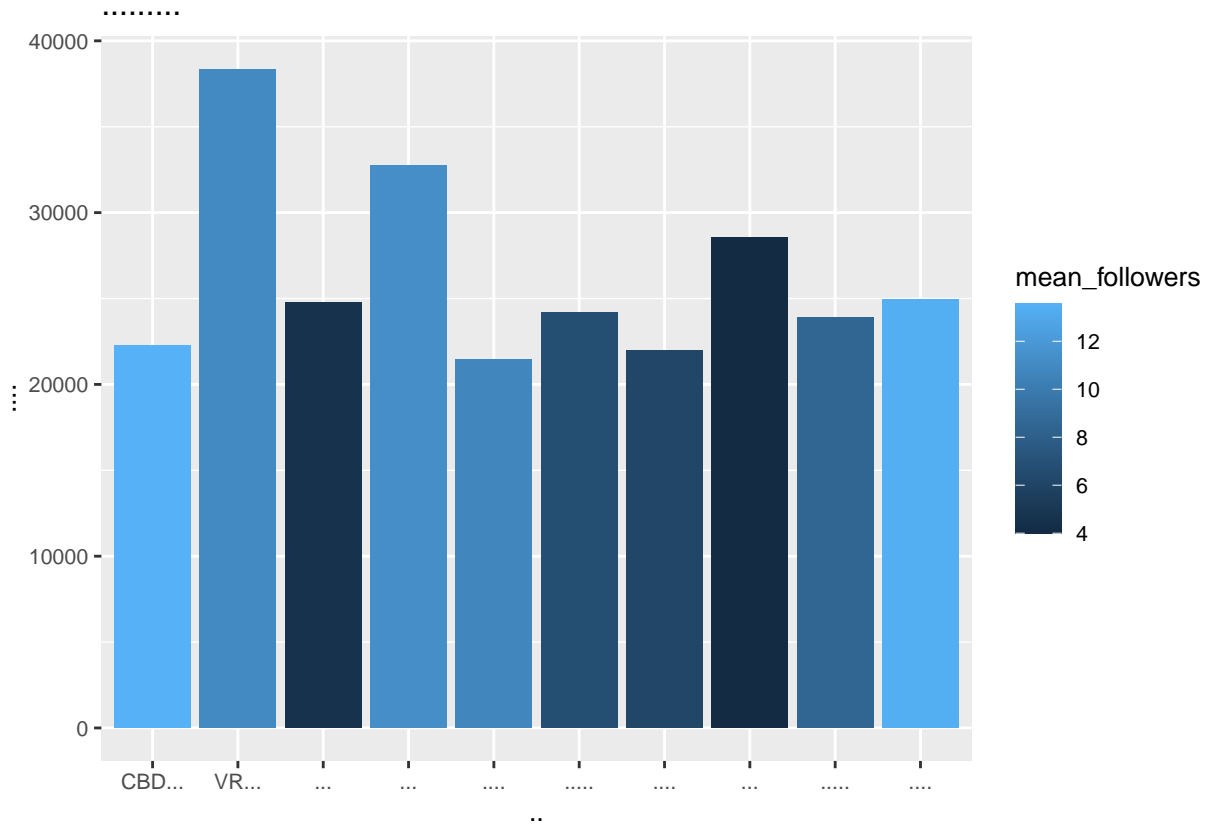


发现:

- 南北朝向的二手房挂牌数最多
- 朝向对房屋的单价影响不大

## 探索问题 2: 单价高和关注高的主要在哪些片区

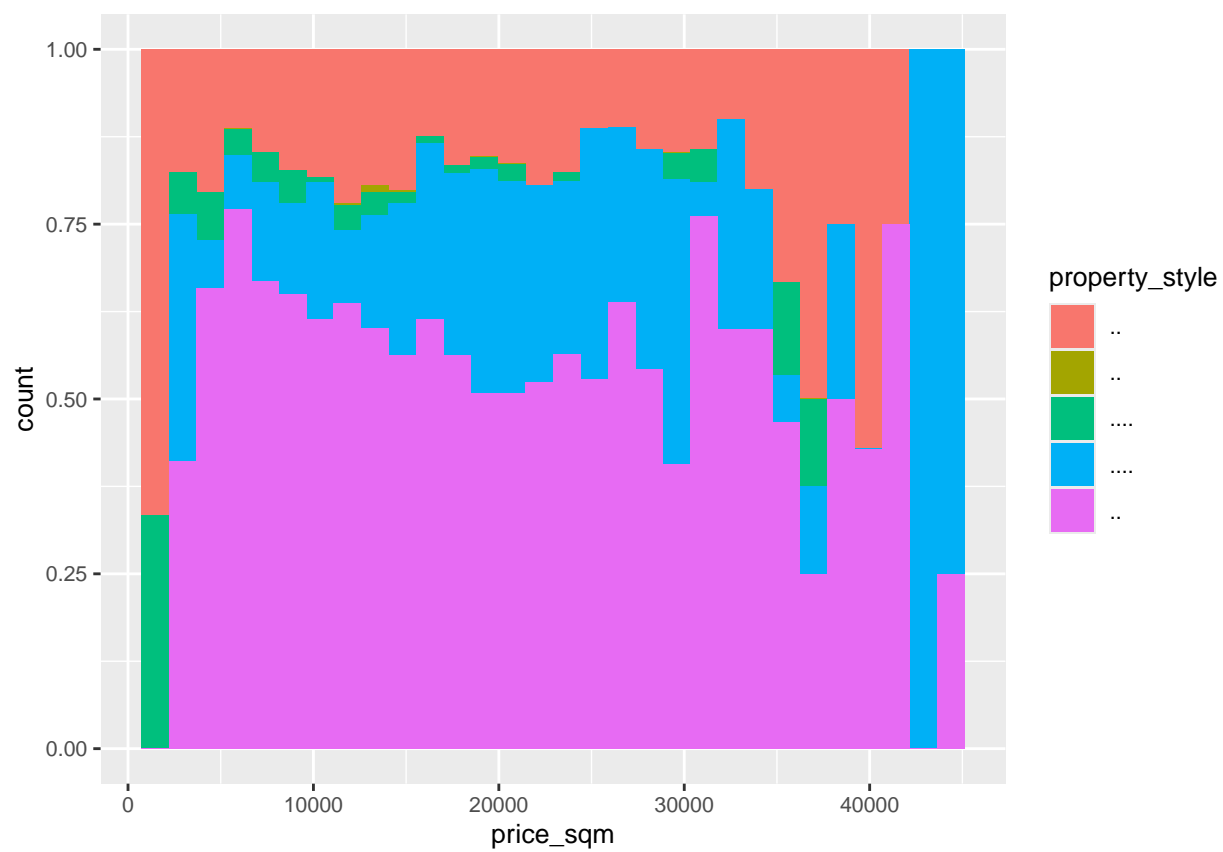
```
## # A tibble: 87 x 4
##   property_region mean_price mean_followers count
##   <chr>          <dbl>         <dbl> <int>
## 1 VR看装修      38351           11         1
## 2 中北路        32728.          11.3        18
## 3 水果湖        28562.           4         9
## 4 黄埔永清      24957.          13.3        23
## 5 三阳路        24777.           4.56        16
## 6 南湖沃尔玛    24181.           6.76        33
## 7 虎泉杨家湾    23902.           8.52        21
## 8 CBD西北湖     22272.          13.6        35
## 9 楚河汉街      21958.           6.13        15
## 10 关山大道      21480.          10.7        25
## # i 77 more rows
```

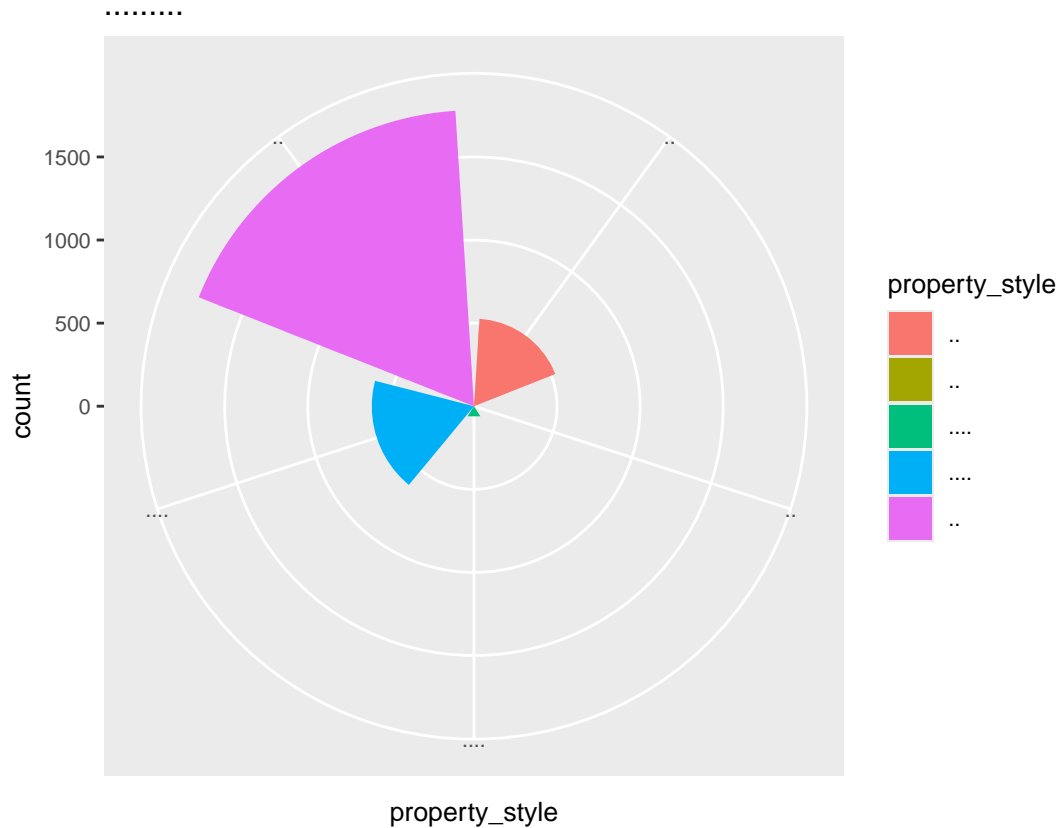


发现：

- 单价高而且关注高的片区是：黄埔永清、CBD 西北路、VR 看装修 (可能是脏数据)、中北路

### 探索问题 3: 建筑类型对房价的影响





发现：

- 建筑形式最多的是板楼，其次是板塔结合，再次是塔楼。
- 房屋单价上看不出与建筑类型有比较强的关联关系

## 发现总结

从上面的数据分析中，数据分析是一个比较抽象的过程，特别是在探索性发现的过程中，你常识中预测的时间结果可能与最终分析的结果不一致。