1st_assignment_report

武汉二手房链家数据分析

2023281051013_ 周萌

主要发现

1. 二手房单价受多种因素影响，且影响关系复杂。同时总计影响因子较为单纯，随面积呈正相关。
2. 二手房市场上在售房源的单价分布和受关注程度与区域特性、便利程度（是否近地铁、楼层高度、是否装修）有很大关系。
3. 二手房市场的供需双方选择上具有一致性。

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：

```
## # A tibble: 10 x 18
##    property_name    property_region price_ttl price_sqm bedrooms livingrooms
##    <chr>            <chr>                <dbl>     <dbl>    <dbl>       <dbl>
##  1 南湖名都A区      南湖沃尔玛             237     18709        3           1
##  2 万科紫悦湾       光谷东                 127     14613        3           2
```

```
##  3 东立国际          二七                 75      15968        1           1
##  4 新都汇            光谷广场            188      15702        3           2
##  5 保利城一期        团结大道            182      17509        3           2
##  6 加州橘郡          庙山                122      10376        3           2
##  7 省建筑五公司西区  光谷广场             99      12346        2           1
##  8 保利上城东区      白沙洲              194.     16336        3           2
##  9 石化大院          中南丁字桥          325      32631        4           1
## 10 阳光花园          杨汊湖              192      17403        3           2
## # i 12 more variables: building_area <dbl>, directions1 <chr>,
## #   directions2 <chr>, decoration <chr>, property_t_height <dbl>,
## #   property_height <chr>, property_style <chr>, followers <dbl>,
## #   near_subway <chr>, if_2y <chr>, has_key <chr>, vr <chr>
```

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：

- 我们所获取到的 18 种数据类型按类别可简单划分为” 区域数据”、“价格数据”、“房屋数据”、“看房数据”。
- 3000 套房源房屋总价大多集中在 95-188 万，最高价 1380 万，总价均值右偏（155.9 万），可见房源总体价格均衡，但高价房源总价偏高明显。
- 房源单价集中在 1-1.8 万/平，均值 1.5 万/平，相对来说较为均衡，但也不乏一些特例。
- 二手房在售户型主要是三室两厅，建筑面积大多在 84-117 平，楼栋高度集中在 11-33 楼，整体右偏。可见偏好中等大小标准户型，高层住房。

探索性分析

1. 房屋单价、总价、面积分布

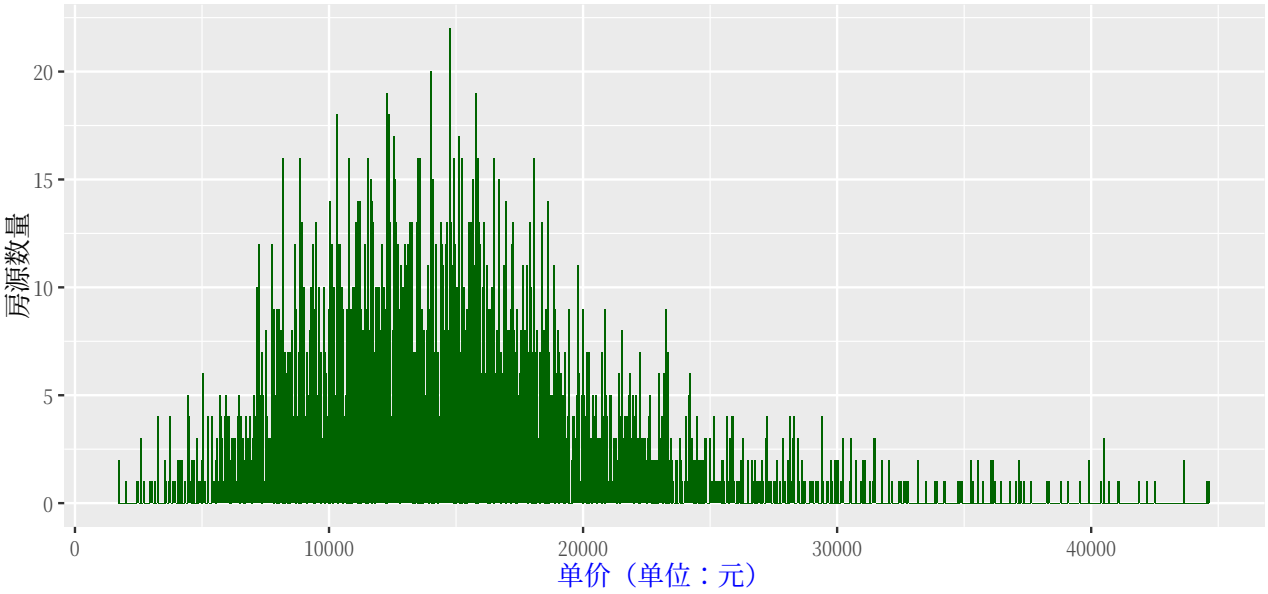
直接观察可见：

1. 二手房单价与总价左偏明显。
2. 单价集中在 1.5 万/平左右，大多单价低于 2 万/平，但高单价房源存在不少。
3. 总价分布较为单纯，在 500 万内，集中分布在 95-188 万。

- **单价数据简要如下：**

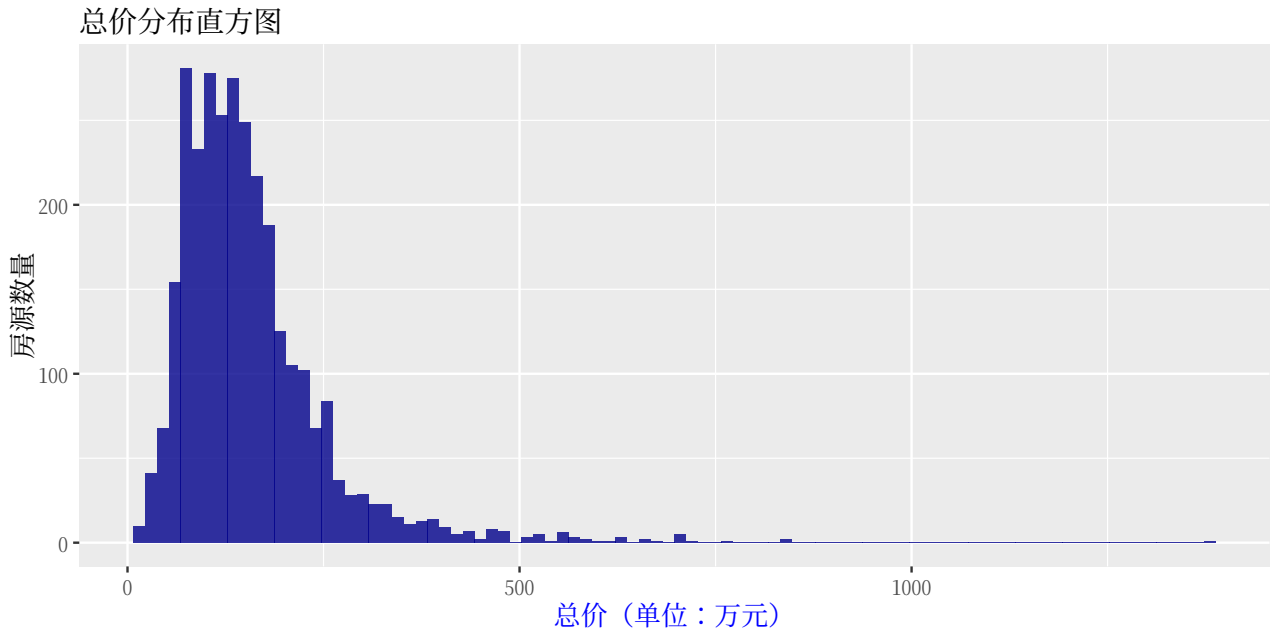
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1771   10799   14404   15148   18211   44656
```

单价分布直方图



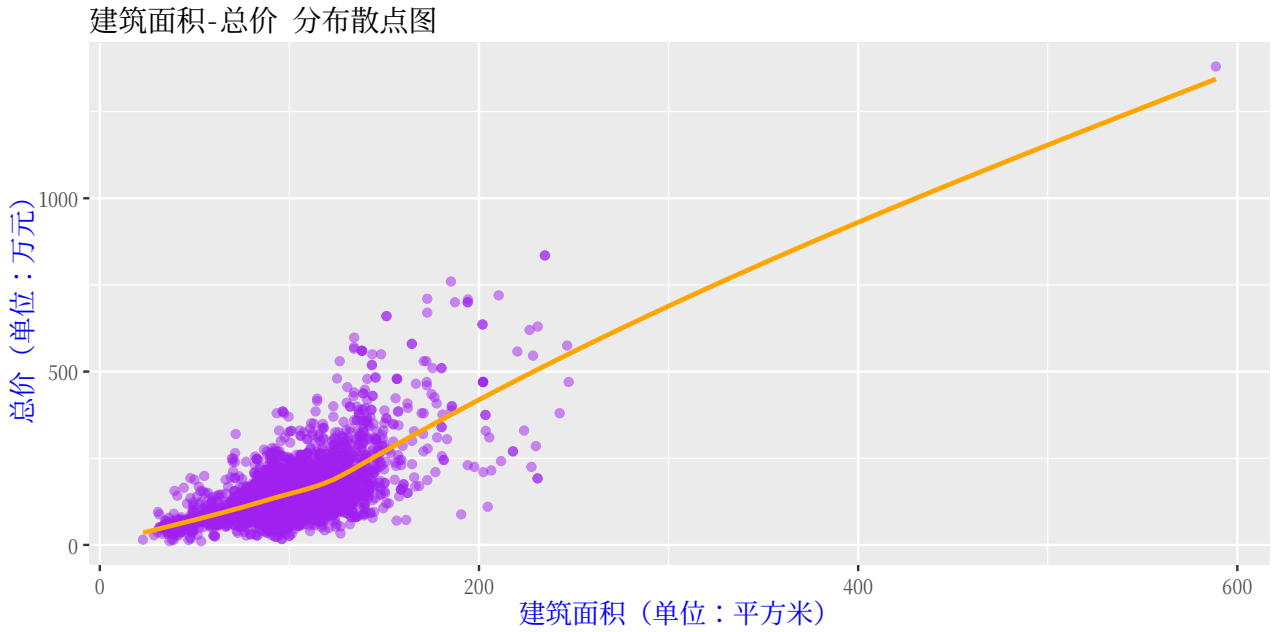
- **总价数据简要如下：**

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.6    95.0   137.0   155.9   188.0  1380.0
```



探索发现：

1. 由图可看出总价和房屋面积间正相关，建筑面积越大，总价越高。建筑面积主要集中在 200m² 内，总价集中在 500 万内，由**总价 = 单价 × 面积**可推测单价集中在 2.5 万/平内，且左偏，符合最初统计的单价数据。
2. 由曲线转点可见，建筑面积小于 100m² 时总价与面积的相关系数比在建筑面积大于 100m² 时小。



2. 装修状况分布

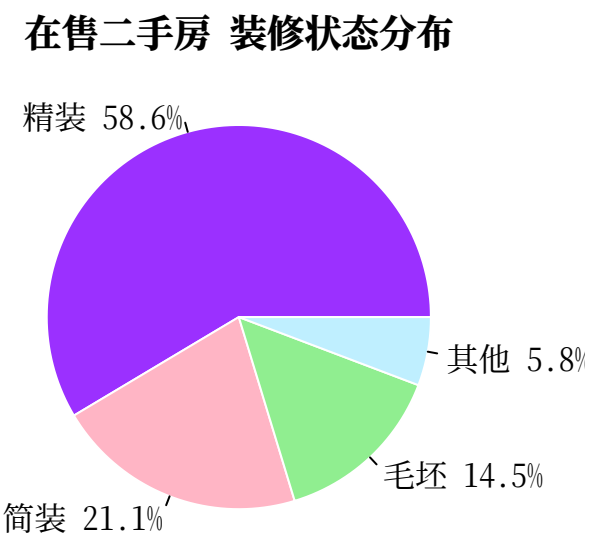
发现：

1. 在售二手房装修状况分为 4 种，分别为：**“精装”，“简装”，“毛坯”，“其他”**。大多在售二手房” 有装修”，其中” 精装” 占比最多（58.6%），同时关注度也是最高的。可见，大家在买房需求上和投资偏好达到共性，更倾向于” 有装修” 的房源。

2. 装修状况对房价有着显著影响。精装**单价均值**明显比简装或毛坯的房屋单价高。
3. 有装修的所有房源中”精装”和”简装”房源单价偏离程度相近，但”毛坯”房单价偏离较大，说明其他因素对”毛坯”房价影响高于”有装修”房价。

- **各装修状况下的单价均值、偏离程度概览如下：**

```
## # A tibble: 4 x 5
##   decoration     n   mean    sd follower_sum
##   <fct>      <int>  <dbl> <dbl>        <dbl>
## 1 精装        1757 16077. 6161.        11061
## 2 简装         634 13993. 5704.         4410
## 3 毛坯         436 13819. 7205.         3862
## 4 其他         173 13304. 6085.          508
```



3. 户型分布

数据处理：合并在售二手房样本中房间数与客厅数，生成新变量“**户型（roomtype）**”

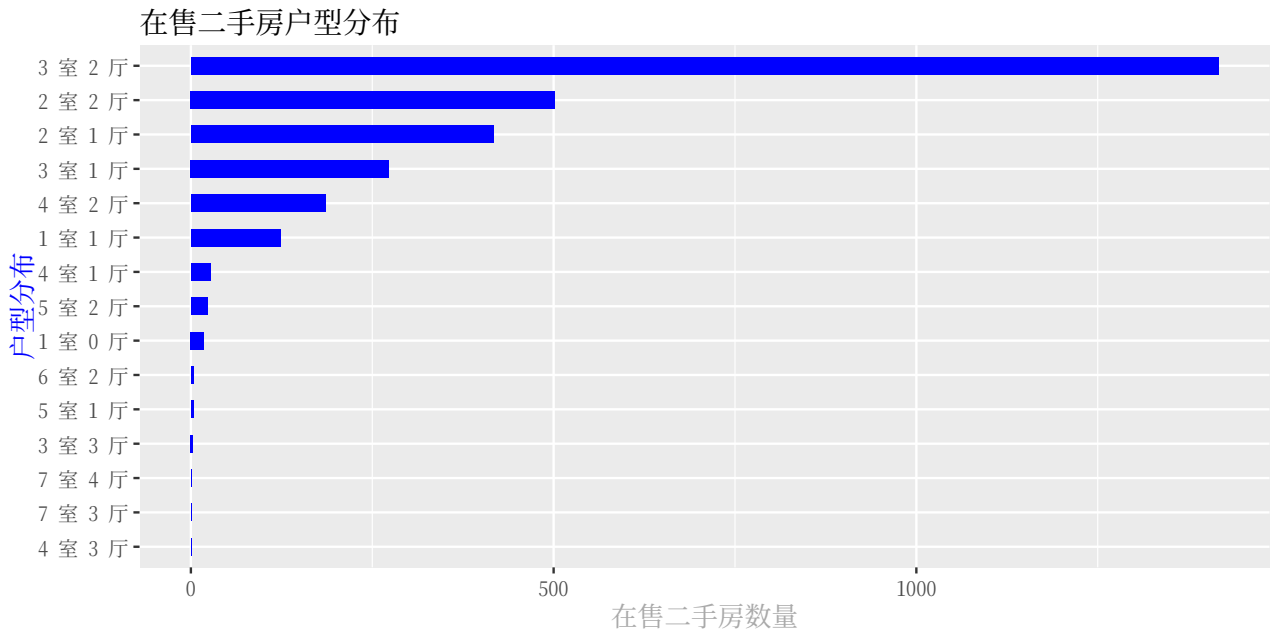
发现：

1. 在售二手房中占比最多房型为 3 室 2 厅，并且关注人数最多，说明大家在房型选择上更偏向这种标准结构。该房型的平均建筑面积 110m²，说明在售此房型多为大户型。各类房型均价大多在 1.5 万/平左右，且最低和最高房价差值较大，房屋单价可能受其他非房型因素影响更大。
2. 在售二手房中占比 top 房型为 3 室 2 厅，2 室 2 厅，2 室 1 厅，3 室 1 厅。只有少数几种户型数量比较多，其余的都非常少，明显属于长尾分布类型（严重偏态）。

- **各类户型的在售房源数、单价均值、最小值、最大值，平均建筑面积、关注人数概览如下：**

```
## # A tibble: 15 x 7
```

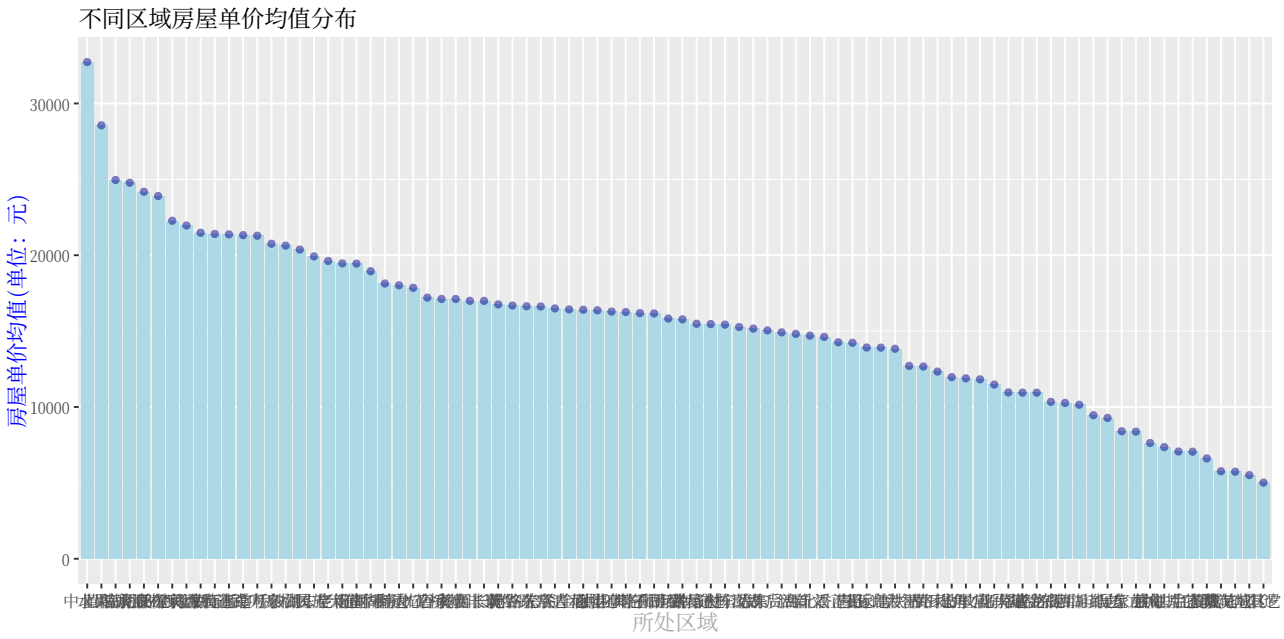
```
##    roomtype      n   mean   min   max building_area_aver follower_sum
##    <fct>     <int>  <dbl> <dbl> <dbl>              <dbl>        <dbl>
##  1 3 室 2 厅  1416 15347.  2599 44574              111.          9475
##  2 2 室 2 厅   502 14527.  1771 37126               85.2         3660
##  3 2 室 1 厅   417 14275.  3068 44656               74.8         2127
##  4 3 室 1 厅   273 14518.  3536 39909               99.1         1734
##  5 4 室 2 厅   186 17557.  4660 43643              150.          1533
##  6 1 室 1 厅   124 15601.  1984 39534               49.4          779
##  7 4 室 1 厅    27 16736.  4179 32631              132.            99
##  8 5 室 2 厅    23 18945.  4456 36463              185.           149
##  9 1 室 0 厅    18 10257.  4074 14921               40.9           94
## 10 5 室 1 厅     4 25861  11749 30565              145.            54
## 11 6 室 2 厅     4 12727.  5376 22241              195.            36
## 12 3 室 3 厅     3 12684. 11187 13433              124.            40
## 13 4 室 3 厅     1  4617   4617  4617              191.            10
## 14 7 室 3 厅     1 23444  23444 23444              589.            32
## 15 7 室 4 厅     1 11915  11915 11915               96.5           19
```



4. 各区域房源关注度、在售情况、高价/低价区域分布

- 按区域划分房源单价均值概览:

```
##       mean
##  Min.   : 5016
##  1st Qu.:11869
##  Median :15801
##  Mean   :15482
##  3rd Qu.:18337
##  Max.   :32728
```



发现：

1. 单价 top2 区域与其他区域断崖式价格差别，均价 1-2 万/平的区域较为集中。
2. 各区域单价均值集中在 1.2-1.8 万/平，中位数与均值相差不大。但最低价区域和最高价区域相差明显。

- **均价 top10 区域数据：**

```
## # A tibble: 10 x 5
##    property_region     n   mean     sd followers_sum
##    <fct>           <int>  <dbl>  <dbl>         <dbl>
##  1 中北路             18 32728. 10615.           203
##  2 水果湖              9 28562.  4997.            36
##  3 黄埔永清           23 24957. 11602.           307
##  4 三阳路             16 24777.  7935.            73
##  5 南湖沃尔玛         33 24181.  7704.           223
##  6 虎泉杨家湾         21 23902.  5184.           179
##  7 CBD西北湖          35 22272.  9058.           475
##  8 楚河汉街           15 21958.  4703.            92
##  9 关山大道           25 21480.  5191.           268
## 10 积玉桥             63 21403.  5955.           893
```

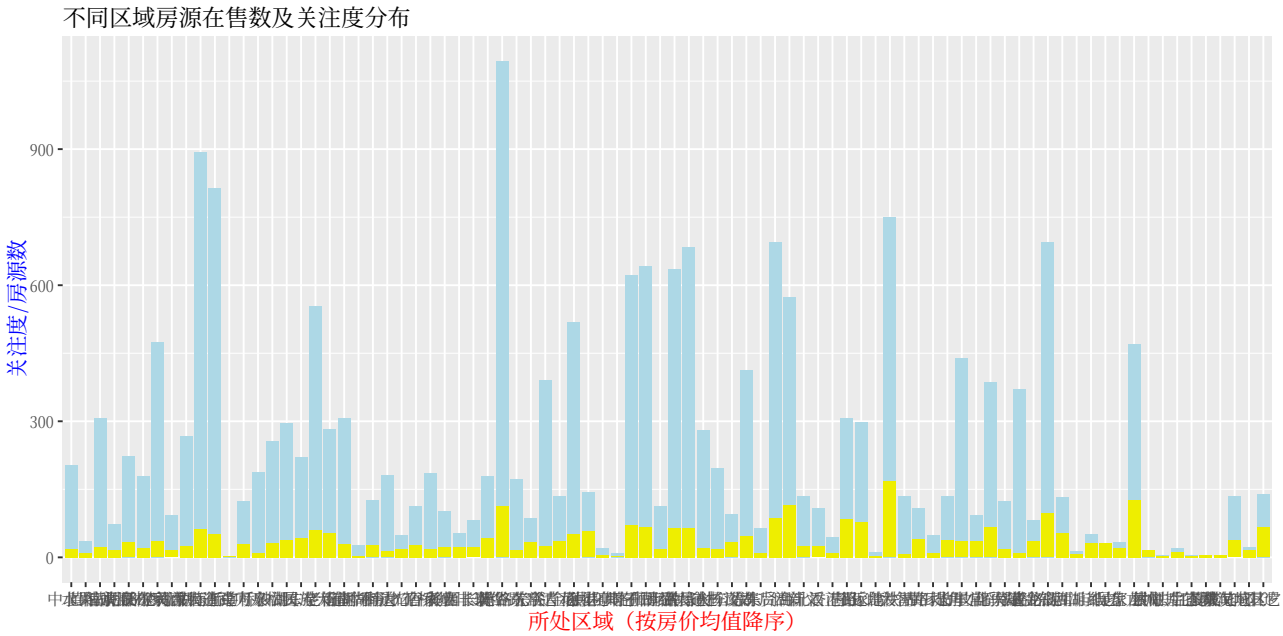
- **均价最低 top5 区域数据：**

```
## # A tibble: 5 x 5
##   property_region     n  mean    sd followers_sum
##   <fct>           <int> <dbl> <dbl>         <dbl>
## 1 阳逻               66 5016. 1487.           139
## 2 汉南其它           16 5512  2021.            23
## 3 蔡甸城区           37 5736. 1633.           134
## 4 黄陂其它            6 5764. 1974.             3
## 5 江夏其它            6 6612.  770.             2
```

发现：

1. 单价 top 区域是中北路，均价 3.2 万/平，中心城区，顶级商圈，地域优势明显，但很明显这个区域房源单价均值偏离程度大，说明其他因素会导致该区域房价差异大。top2 高价区是水果湖，学区房圈加成，但该区域房价单价均值偏离程度小，房源性质相差不大。
2. 低价区域是阳逻、汉南其他、蔡甸城区、黄陂其他、江夏其他。可以看出远城区房价和中心城区差异明显，而且低价区域各房源均价偏离程度小，看出其他因素对低价区域房价影响远没有地理位置这一点对房价影响大。

- **不同区域关注度与在售情况：**



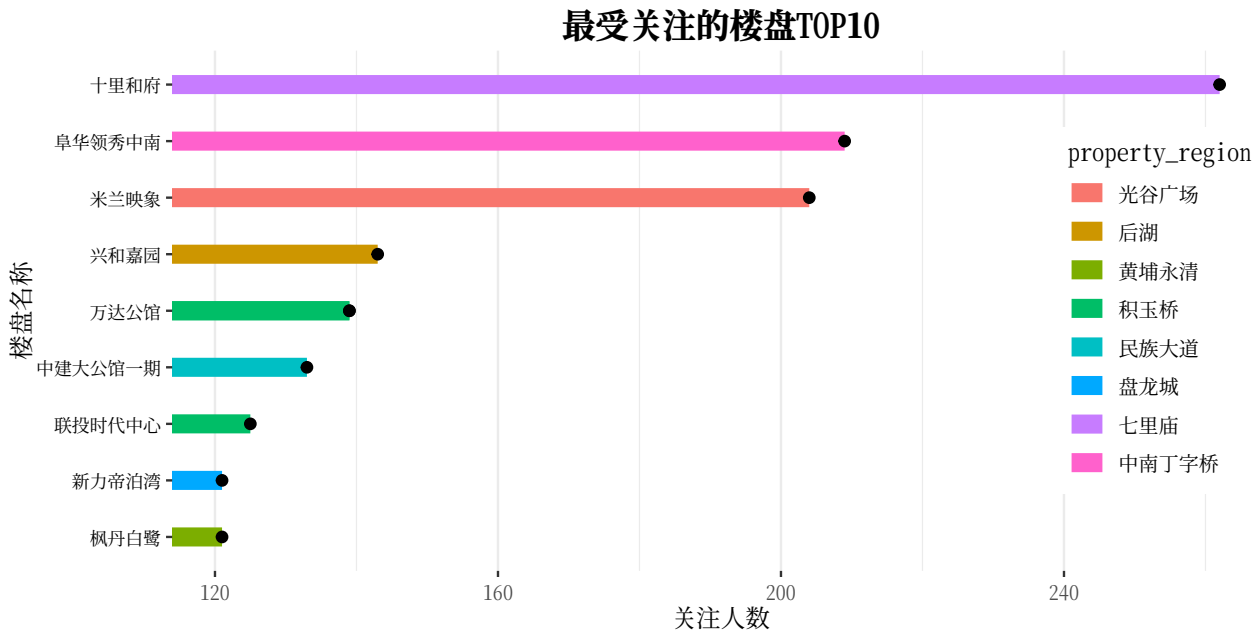
发现：

1. 由图可见区域关注度和在售房源分布有一定拟合，关注度高的区域，在售房源数也不低。但关注度和在售数明显与区域价格没有太强关联。
2. 关注度 top 的楼盘看区域主要在 “七里庙”、“中南丁字桥”、“光谷广场”，商圈地域特色明显。

- **top3 关注度楼盘信息概览：**

```
##   property_name property_region price_ttl price_sqm bedrooms livingrooms
## 1      十里和府           七里庙        90     10765        2           2
## 2  阜华领秀中南       中南丁字桥        68     11711        2           1
## 3      米兰映象         光谷广场       175     16948        3           2
##   building_area directions1 directions2 decoration property_t_height
## 1         83.61         东南        <NA>       精装                23
## 2         58.07           北        <NA>       简装                30
## 3        103.26           南        <NA>       简装                33
##   property_height property_style followers near_subway      if_2y  has_key
## 1               高           塔楼       262      近地铁 房本满两年 随时看房
## 2               中           塔楼       209        <NA>       <NA> 随时看房
## 3               高       板塔结合       204        <NA>       <NA> 随时看房
##         vr  roomtype
```

```
## 1 VR看装修 2 室 2 厅
## 2 VR看装修 2 室 1 厅
## 3 VR看装修 3 室 2 厅
```



5. 近地铁情况对二手房单价的影响

- “近地铁”二手房单价数据概览：

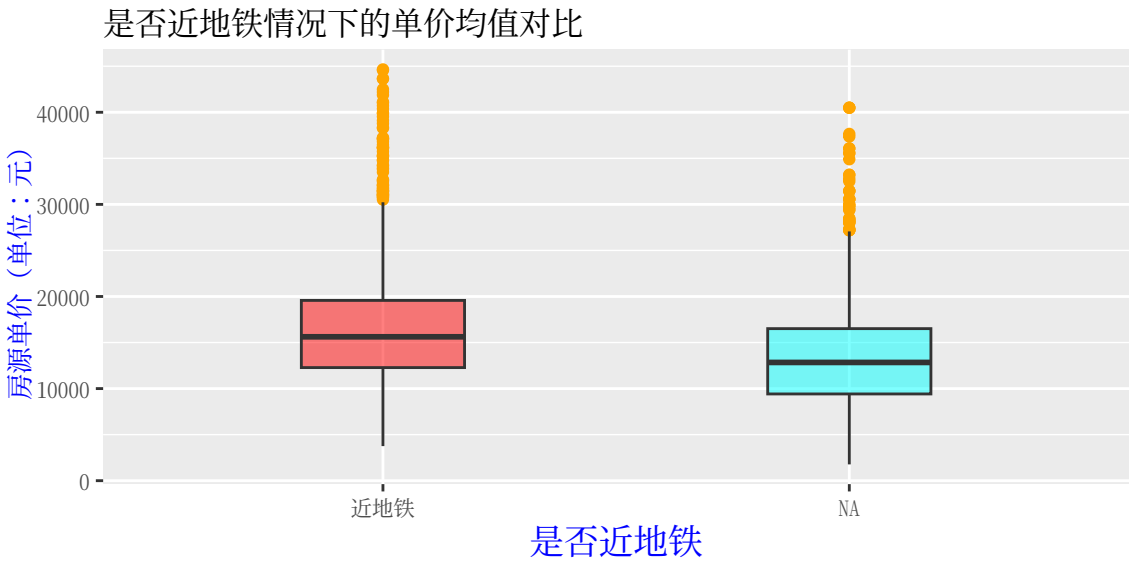
```
##    price_sqm
##  Min.   : 3755
##  1st Qu.:12280
##  Median :15622
##  Mean   :16624
##  3rd Qu.:19573
##  Max.   :44656
```

- “非近地铁”二手房单价数据概览：

```
##    price_sqm
##  Min.   : 1771
##  1st Qu.: 9418
##  Median :12840
##  Mean   :13558
##  3rd Qu.:16509
##  Max.   :40492
```

发现：

1. 近地铁的二手房单价区间高于非近地铁情况，同时近地铁二手房均价明显高于非近地铁情况。
2. 游离数据中仍有许多非近地铁房源处于高价区，说明这些房源其他优势更甚，高于是否近地铁的影响程度。



6. 不同高度楼栋下房源所处位置对单价影响

发现：

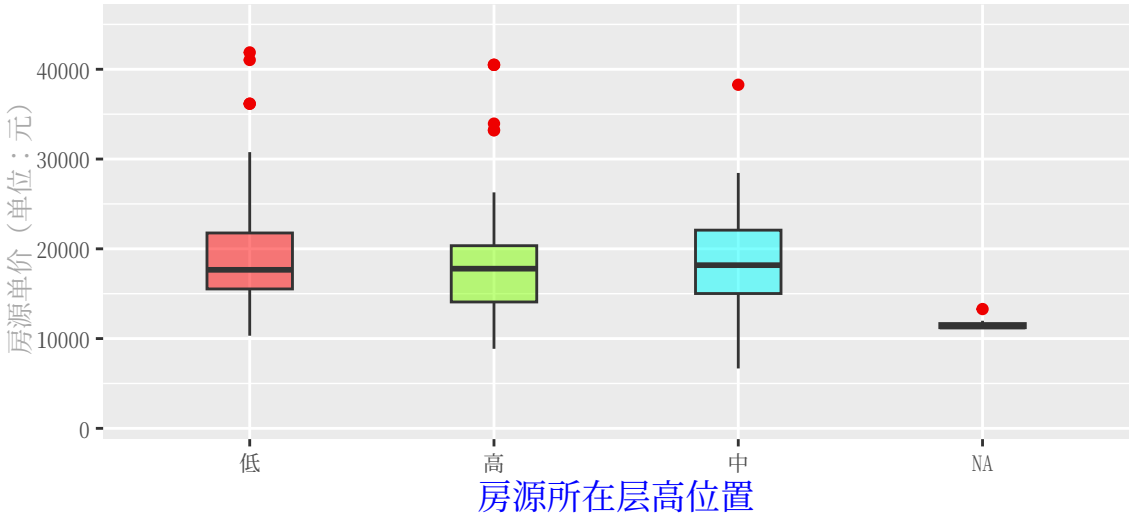
1. 横向对比不同高度规格的楼栋，发现楼栋高度对二手房单价有一定影响，楼栋总高度越高，单价区间越高。对于超高层（楼栋层高 >35 层）单价区间与其他相差明显，超高层单价集中在 1.4-2.1 万/平区间。其他楼栋高度下单价区间偏差不大，高层集中在 1-1.8 万/平；小高层集中在 0.9-1.6 万/平；低层集中在 0.8-1.6 万/平。
2. 对比在同一高度规格楼栋下，房源所处楼栋不同位置单价存在一定差异。其中不同房源位置在“超高层”、“低层”楼栋差异较大。“超高层”楼栋中间位置、“低层”楼栋低位单价区间较高。可能与超高层房屋结构下电梯、水压，或“低层”楼栋可能无电梯需爬楼，带庭院等因素影响有关。

不同高度的楼栋-房屋所在楼层位置与单价分布概览：

- **超高层 super_high** *(property_t_height >=35)*

```
##    price_sqm
##  Min.   : 6674
##  1st Qu.:14812
##  Median :17789
##  Mean   :18644
##  3rd Qu.:21721
##  Max.   :41878
```

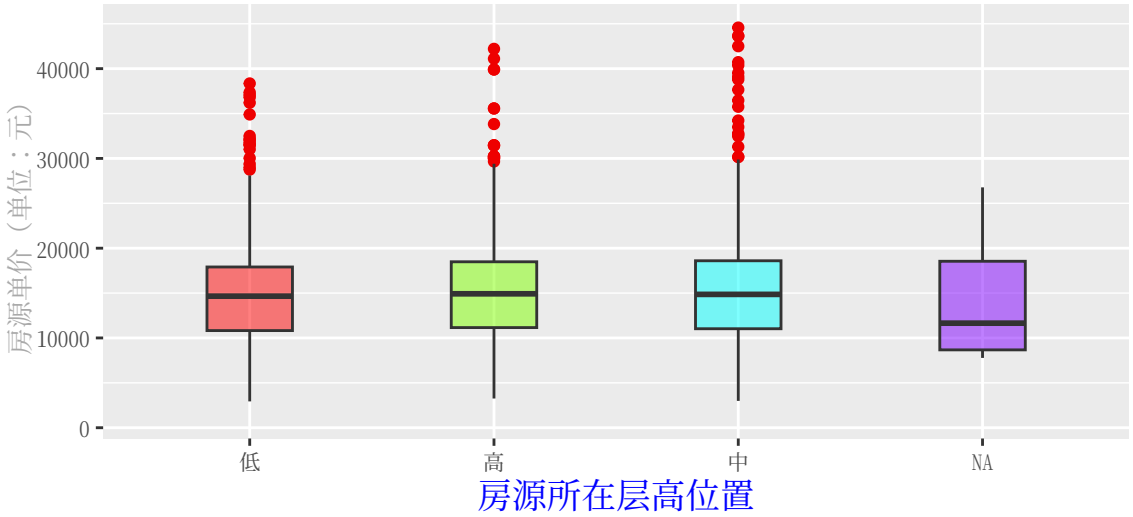
“超高层”楼栋-房屋所在楼层位置与单价分布



- 高层 **high** *(16<= property_t_height <35)*

```
##    price_sqm
##  Min.   : 2934
##  1st Qu.:10980
##  Median :14804
##  Mean   :15296
##  3rd Qu.:18294
##  Max.   :44574
```

“高层”楼栋-房屋所在楼层位置与单价分布

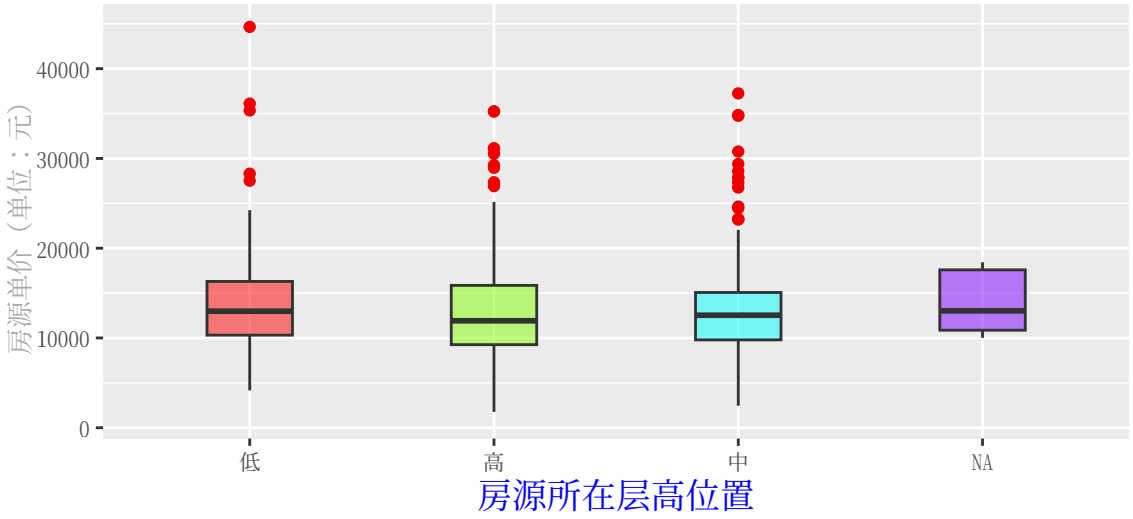


- 小高层 **little_high** *(7<= property_t_height <16)*

```
##    price_sqm
##  Min.   : 1771
##  1st Qu.: 9704
##  Median :12378
##  Mean   :13358
##  3rd Qu.:15805
```

```
##  Max.    :44656
```

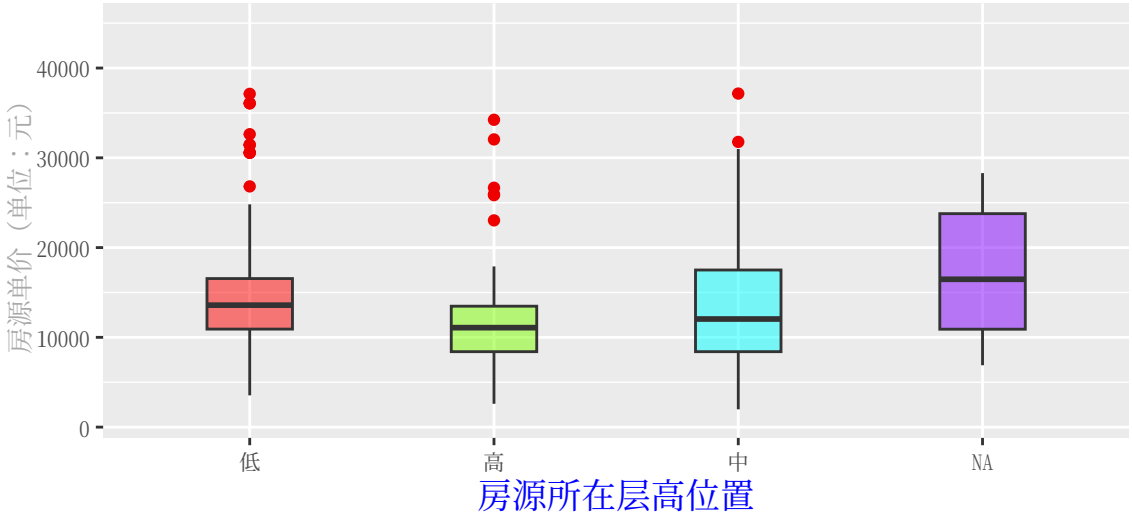
“小高层”楼栋-房屋所在楼层位置与单价分布



- 低层 **lower** *(property_t_height <7)*

```
##    price_sqm
##  Min.    : 1984
##  1st Qu.: 8959
##  Median :12422
##  Mean    :14092
##  3rd Qu.:16389
##  Max.    :37160
```

“低层”楼栋房屋所在层高与单价分布



发现总结

从总体数据来看，15000 元/平的房源是武汉二手市场交易的主流，在这个价格区间内，房源单价差异不大。可以综合户型，是否临近地铁，装修状况等因素来选择，但如果对楼栋高度、楼层位置有要求，可选范围大大缩小。从区域特征来看，武汉二手房主城区、远城区，商圈、学区房这些地域特征对房源单价影响较大。可见，这种附加属性带来的价值高于房屋本身价值。从关注和在售程度来看，单价或总价往往不是第一选择要素，比如精装往往关注和在售程度高于毛坯，近地铁高于非近地铁。在二手房选择上考虑的是多方面结合体，而在售和关注度往往相连紧密，说明二手房供需方的选择具有同向性。