陈向向 +2023281051020+ 第一次作业

陈向向

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。

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。

```
## spc_tbl_ [3,000 x 18] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ property_name    : Factor w/ 1345 levels "027社区","08经典",..: 770 1027 220 1106
##  $ property_region  : Factor w/ 87 levels "CBD西北湖","VR看装修",..: 49 21 17 22 65
##  $ price_ttl        : num [1:3000] 237 127 75 188 182 ...
##  $ price_sqm        : num [1:3000] 18709 14613 15968 15702 17509 ...
##  $ bedrooms         : num [1:3000] 3 3 1 3 3 3 2 3 4 3 ...
##  $ livingrooms      : num [1:3000] 1 2 1 2 2 2 1 2 1 2 ...
##  $ building_area    : num [1:3000] 126.7 86.9 47 119.7 104 ...
##  $ directions1      : Factor w/ 8 levels "北","东","东北",..: 5 5 5 1 4 5 5 5 5 5 ..
##  $ directions2      : Factor w/ 8 levels "北","东","东北",..: 1 NA NA 2 NA 1 NA 1 1
##  $ decoration       : Factor w/ 4 levels "简装","精装",..: 2 2 1 2 1 2 1 4 1 4 ...
##  $ property_t_height: num [1:3000] 17 28 18 32 34 34 7 34 5 7 ...
##  $ property_height  : Factor w/ 3 levels "低","高","中": 3 3 1 2 3 1 1 3 1 1 ...
##  $ property_style   : Factor w/ 5 levels "板楼","板塔结合",..: 4 1 4 4 2 1 1 2 1 1 .
##  $ followers        : num [1:3000] 3 1 3 2 3 1 0 0 2 0 ...
##  $ near_subway      : Factor w/ 5 levels "VR看装修","近地看",..: 3 NA 3 3 NA NA 3 3
##  $ if_2y            : chr [1:3000] NA "房本满两年" NA "房本满两年" ...
##  $ has_key          : chr [1:3000] "随时看房" "随时看房" "随时看房" "随时看房" ...
##  $ vr               : chr [1:3000] NA "VR看装修" NA NA ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   property_name = col_character(),
##   ..   property_region = col_character(),
##   ..   price_ttl = col_double(),
##   ..   price_sqm = col_double(),
##   ..   bedrooms = col_double(),
##   ..   livingrooms = col_double(),
##   ..   building_area = col_double(),
##   ..   directions1 = col_character(),
##   ..   directions2 = col_character(),
```

```
##   ..   decoration = col_character(),
##   ..   property_t_height = col_double(),
##   ..   property_height = col_character(),
##   ..   property_style = col_character(),
##   ..   followers = col_double(),
##   ..   near_subway = col_character(),
##   ..   if_2y = col_character(),
##   ..   has_key = col_character(),
##   ..   vr = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

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     <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, 东~
## $ directions2       <fct> 北, NA, NA, 东, NA, 北, NA, 北, 北, 北, 北, NA, 西南~
```

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	杨汊湖	192.0	17403	3	2	110

```
## $ 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             <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
##  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
##  简装: 634   Min.   : 2.00     低  : 816       板楼    :1781
##  精装:1757   1st Qu.:11.00     高  : 906       板塔结合: 615
##  毛坯: 436   Median :27.00     中  :1218       平房    :   5
##  其他: 173   Mean   :24.22     NA's:  60       塔楼    : 527
##              3rd Qu.:33.00                     暂无数据:  72
##              Max.   :62.00
##
##    followers          near_subway      if_2y              has_key
##  Min.   :  0.000   VR看装修 :   2   Length:3000        Length:3000
##  1st Qu.:  1.000   近地看   :   1   Class :character   Class :character
##  Median :  3.000   近地铁   :1554   Mode  :character   Mode  :character
##  Mean   :  6.614   珞狮南   :   1
##  3rd Qu.:  6.000   太子湖1号:   1
##  Max.   :262.000   NA's     :1441
##
##       vr
##  Length:3000
##  Class :character
##  Mode  :character
##
##
##
##
```

可以看到：

表 2: Data summary

Name	lj
Number of rows	3000
Number of columns	18
Column type frequency:	
numeric	1
Group variables	None

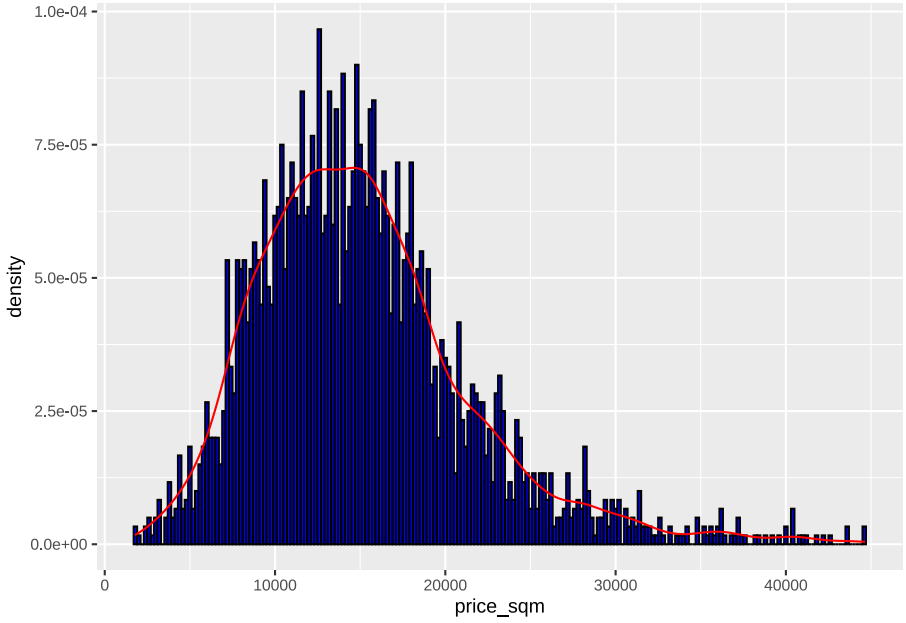
- 直观结论 1: 此数据中的房屋总价主要集中在 95 万-188 万之间，单价主要集中在 10799 元/平米到 18211 元/平米之间，面积主要集中在 85 平米到 117 平米之间，房屋户型主要以三室或两室，两厅或者一厅为主，挂牌房产以总层数为 11 到 33 层的电梯房为主。
- 直观结论 2：均价的中位数比均值低，说明相对于市场均价，高于均价的挂牌房屋数量比低于均价挂牌房屋数量要多，即房屋单价是右偏的；同理，总价也是如此。
- 直观结论 3：房间数为三室的最多、客厅数为 2 厅的最多、朝向为南北数量最多、装修为精装的最多、房产为中楼层的最多、房屋类型为板楼的最多
- ...

3 探索性分析

3.1 变量房屋单价（price_sqm）的数值描述与图形

Variable type: numeric

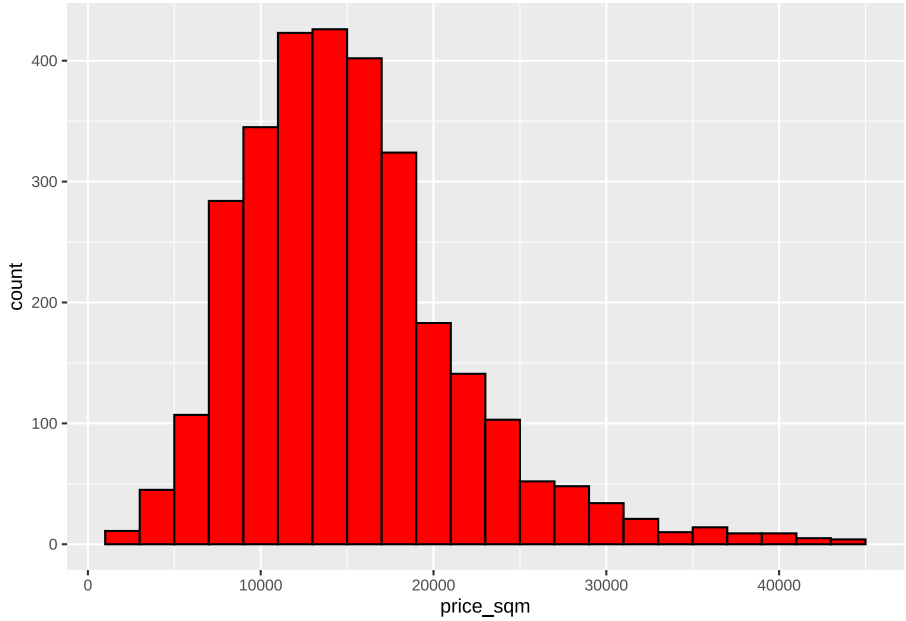
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
price_sqm	0	1	15148.49	6323.18	1771	10799.25	14404	18211




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

```
## [1] 1.080004
```

```
## [1] 5.028977
```




```
## # A tibble: 22 x 2
##    `cut_width(lj$price_sqm, 2000)`     n
##    <fct>                           <int>
##  1 [1e+03,3e+03]                      11
##  2 (3e+03,5e+03]                      45
##  3 (5e+03,7e+03]                     107
##  4 (7e+03,9e+03]                     284
##  5 (9e+03,1.1e+04]                   345
##  6 (1.1e+04,1.3e+04]                 423
##  7 (1.3e+04,1.5e+04]                 426
##  8 (1.5e+04,1.7e+04]                 402
##  9 (1.7e+04,1.9e+04]                 324
## 10 (1.9e+04,2.1e+04]                 183
## # i 12 more rows
```

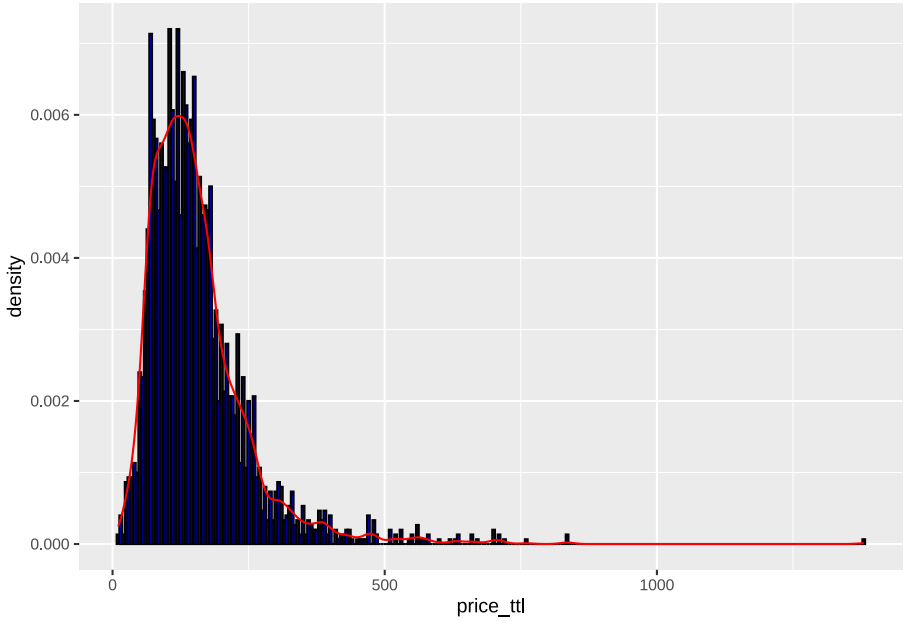
发现：

- 1. 总体均价在 15148 元/平方米，挂牌单价中位数为 14404 元，挂牌房屋的单价主要集中在 10799 元/平方米-18211 元/平方米之间，最高

价与最低价之间的差值大，说明挂牌房屋的单价较为分散

- 2. 房屋单价与其出现的频次呈现正态分布的关系，房屋单价的偏度约为 1.08，峰度约为 5.03，其分布为整体右偏且较为陡峭。
- 3. 样本中房屋挂牌单价分别在 [11000,13000]、[13000,15000]、[15000,17000] 元/平方米区间段的频次最高

3.2 变量房屋总价（price_ttl）的数值描述与图形



Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
price_ttl	0	1	155.86	95.55	10.6	95	137	188	1380	

```
## [1] 2.7546
```

```
## [1] 19.12947
```

```
## # A tibble: 28 x 2
##    `cut_width(lj$price_ttl, 30)`     n
```

表 3: Data summary

Name	lj
Number of rows	3000
Number of columns	18
Column type frequency:	
numeric	1
Group variables	None

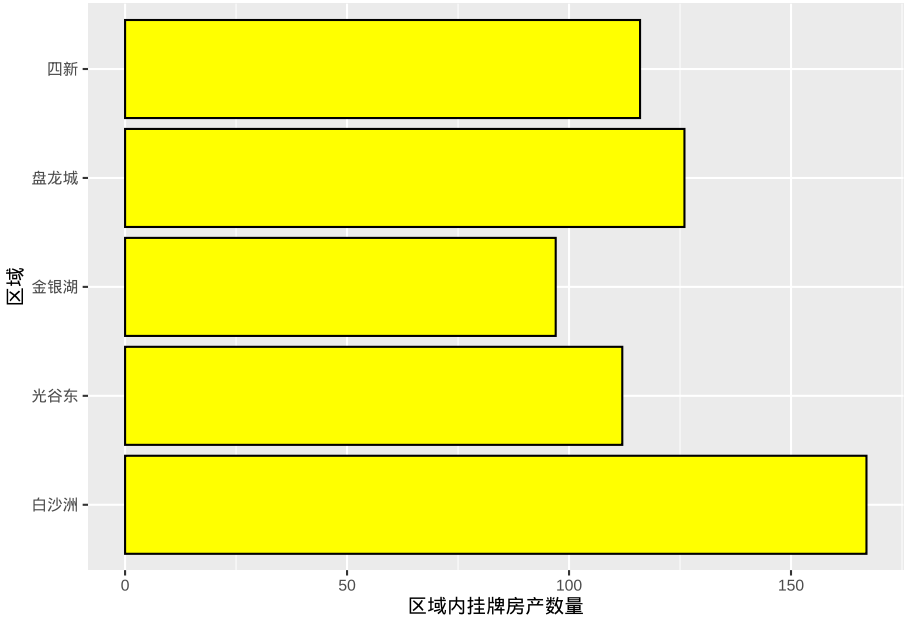
```
##    <fct>                          <int>
##  1 [-15,15]                           5
##  2 (15,45]                           71
##  3 (45,75]                          365
##  4 (75,105]                         510
##  5 (105,135]                        530
##  6 (135,165]                        488
##  7 (165,195]                        352
##  8 (195,225]                        212
##  9 (225,255]                        164
## 10 (255,285]                         87
## # i 18 more rows
```

发现：

- 1. 挂牌房屋的总价平均值为 155 万，中位数为 137 万，挂牌房屋总价的密集区间主要在 75 万到 165 万之间
- 2. 房屋总价与其出现的频次呈现正态分布的关系，房屋总价的偏度约为 2.75，峰度约为 19.13，其分布为整体右偏且非常陡峭，即挂牌的房屋总价在中位数左边集中，且中位数附近出现的频率更多，更为集中。

3.3 变量房产区域（property_region）的数值描述与图形

```
## # A tibble: 5 x 2
## # Groups:   property_region [5]
##   property_region     n
##   <fct>           <int>
## 1 白沙洲            167
## 2 光谷东            112
## 3 金银湖             97
## 4 盘龙城            126
## 5 四新              116
```

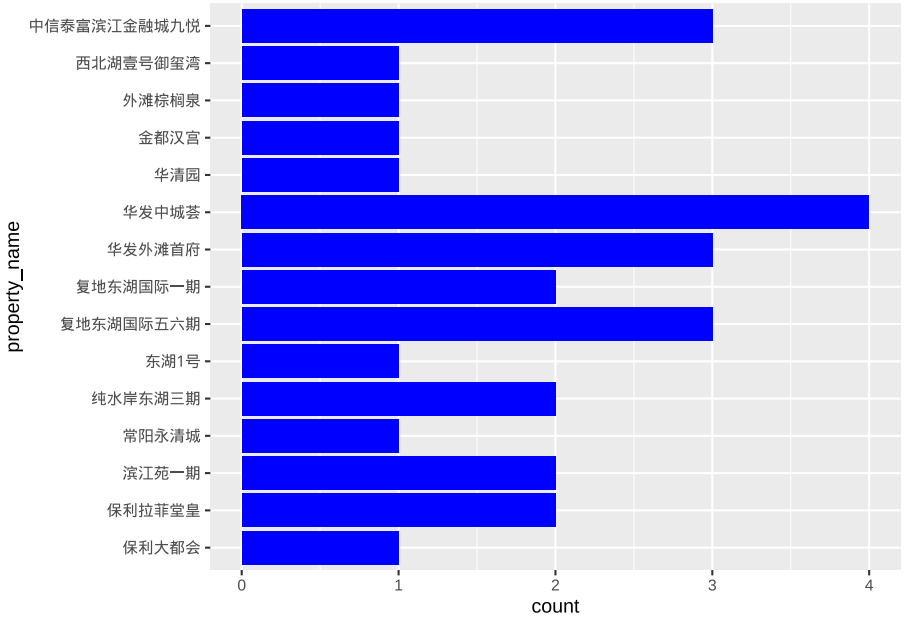


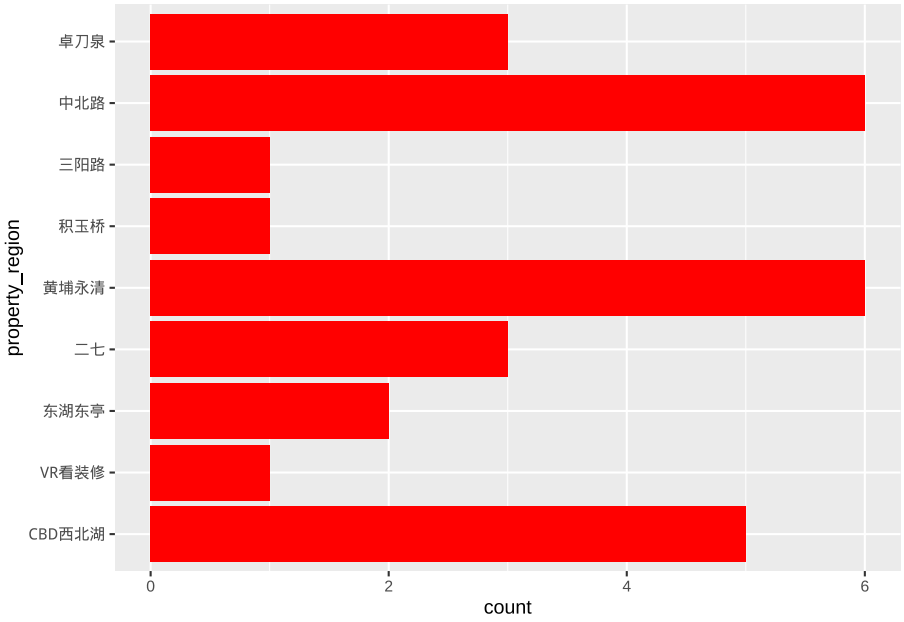
发现：

- 挂牌数量最多的几个区域依次是：白沙洲、盘龙城、四新、光谷东、金银潭

3.4 武汉地区的一些高房价区域和小区在哪里？

```
## # A tibble: 28 x 18
##    property_name       property_region price_ttl price_sqm bedrooms livingrooms
##    <fct>               <fct>                <dbl>     <dbl>    <dbl>       <dbl>
##  1 华发中城荟          CBD西北湖              519     36163        3           2
##  2 华发中城荟          CBD西北湖              519     36163        3           2
##  3 华发中城荟          CBD西北湖              550     38275        3           2
##  4 华发中城荟          CBD西北湖              760     41037        4           2
##  5 西北湖壹号御玺湾    CBD西北湖              720     34223        4           2
##  6 华清园              VR看装修               480     38351        3           2
##  7 纯水岸东湖三期      东湖东亭               835     35567        4           2
##  8 纯水岸东湖三期      东湖东亭               835     35567        4           2
##  9 中信泰富滨江金融城~ 二七                   560     40492        3           2
## 10 中信泰富滨江金融城~ 二七                   560     40492        3           2
## # i 18 more rows
## # i 12 more variables: building_area <dbl>, directions1 <fct>,
## #   directions2 <fct>, decoration <fct>, property_t_height <dbl>,
## #   property_height <fct>, property_style <fct>, followers <dbl>,
## #   near_subway <fct>, if_2y <chr>, has_key <chr>, vr <chr>
```





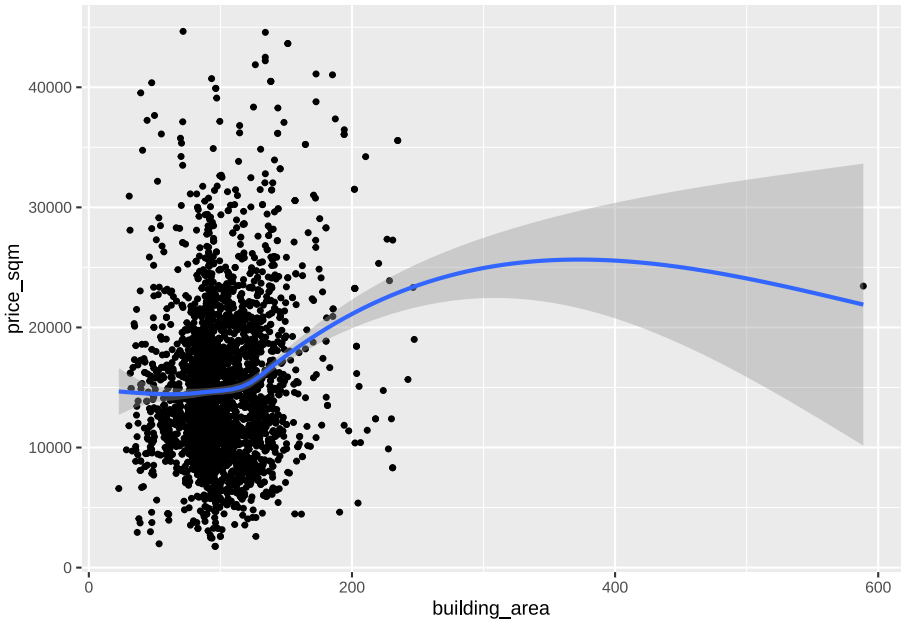
发现：

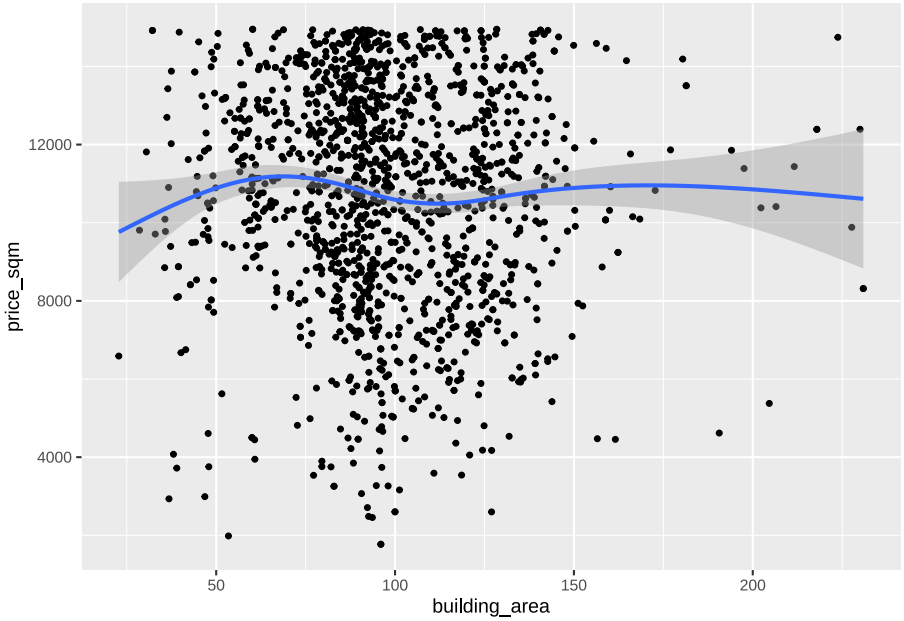
- 1. 挂牌房产中的一些总价大于 422 万且均价大于 3.3 万/平米的高档

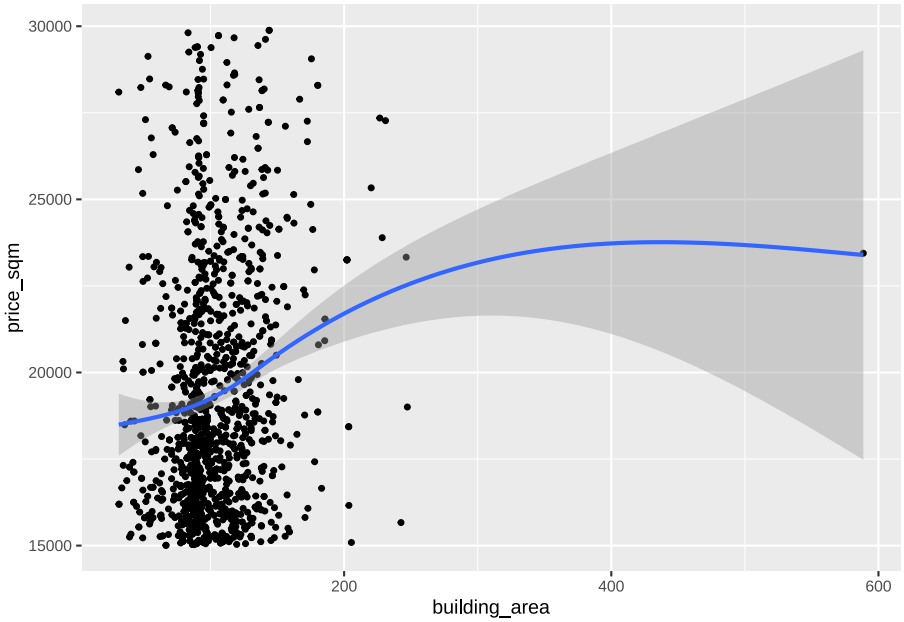
小区有：中信泰富滨江金融城九悦、西北湖壹号、外滩棕榈泉、金都汉宫、华清园、华发中城荟、华发外滩首府、复地东湖国际、东湖 1 号、纯水岸东湖、常阳永清城、滨江苑、保利拉菲堂皇、保利大都会。

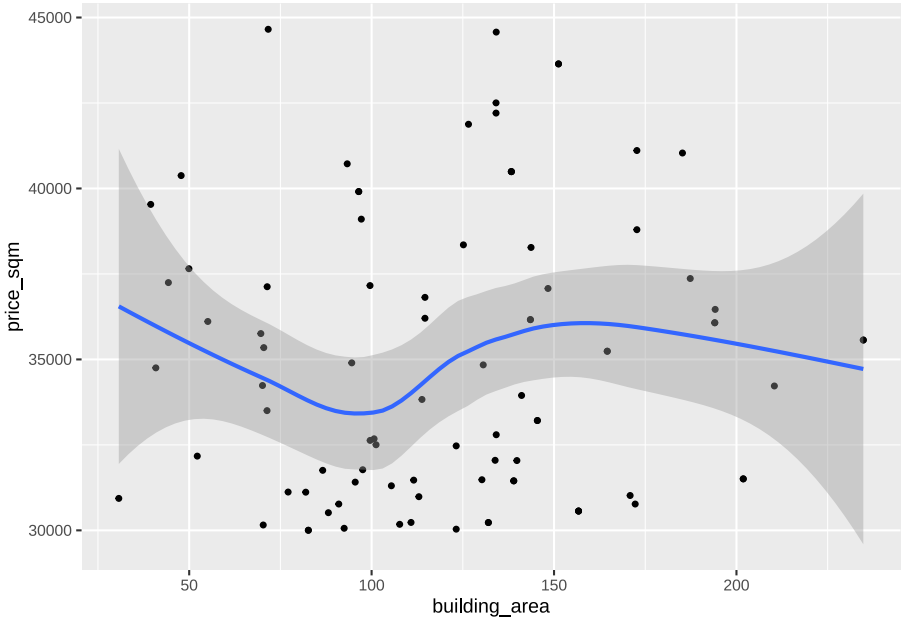
- 2. 其主要分布在以下区域：卓刀泉、中北路、三阳路、积玉桥、黄埔永清、二七、东湖东亭、CBD 西北湖；

3.5 均价会随着面积的增加而减少吗




price_sqm
12000
8000
4000
50
100
150
200
building_area


price_sqm
30000
25000
20000
15000
200
400
600
building_area



发现：

- 均价并不一定会随着面积的增加而减少, 但是平米数超过一定数值，均价会随着面积的增加而减少
- 在一定面积段内或者在价格段内，均价与面积会有一定的相关线性关系

---

4 发现总结

1. 总体均价在 15148 元/平方米，挂牌单价中位数为 14404 元，挂牌房屋的单价主要集中在 10799 元/平方米-18211 元/平方米居多；挂牌总价平均值为 155 万，中位数为 137 万，挂牌房屋总价的密集区间主要在 75 万到 165 万之间。

2. 挂牌数量最多的几个区域依次是：白沙洲、盘龙城、四新、光谷东、金银潭；高档小区主要分布在以下区域：卓刀泉、中北路、三阳路、积玉桥、黄

埔永清、二七、东湖东亭、CBD 西北湖。

3. 均价并不一定会随着面积的增加而减少, 但是平米数超过一定数值，均价会随着面积的增加而减少；在一定面积段内或者在价格段内，均价与面积会有一定的相关线性关系。