关于某二手房网站有关武汉的 3000 套二手房的数据分析

吴梦洁

目录

主要发现

1. 发现 1：通过对武汉二手房链家数据的数据概览，可以得出：房屋总价的均价为 154.8 万元/套；房屋单价均价为 15110 元/m^2；平均建筑面积为 100.67m^2，且多为朝南方向、中层、板楼和近地铁的二手房。

2. 发现 2：通过对房屋单价、房屋总价和建筑面积的单变量分析，可以得出：数据分布均向右偏斜，但房屋单价的右偏程度较房屋总价的右偏程度小，而房屋面积是房屋总价和房屋单价右偏相差较大的原因。

3. 发现 3：通过多变量的相关性分析，可以得出：低层位置对房屋单价的影响度相对更为明显，但在整体上较为接近；板塔结合的房屋单价高是板塔结合的房屋总价高的主要原因；在所有房屋所处区域中，金银湖房屋总价最高，而其建筑面积大是其主要原因。

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 2515 行。表的前 10 行示例如下：

表 1: 武汉链家二手房

property_name	property_region	price_ttl	price_sqm	bedrooms	livingrooms	building_area	directions1
南湖名都 A 区	南湖沃尔玛	237.0	18709	3	1	126.68	南
万科紫悦湾	光谷东	127.0	14613	3	2	86.91	南
东立国际	二七	75.0	15968	1	1	46.97	南
新都汇	光谷广场	188.0	15702	3	2	119.73	北
保利城一期	团结大道	182.0	17509	3	2	103.95	东南
加州橘郡	庙山	122.0	10376	3	2	117.59	南
省建筑五公司西区	光谷广场	99.0	12346	2	1	80.19	南
保利上城东区	白沙洲	193.8	16336	3	2	118.64	南
石化大院	中南丁字桥	325.0	32631	4	1	99.60	南
阳光花园	杨汉湖	192.0	17403	3	2	110.33	南

各变量的简短信息：

```
## Rows: 2,515
## 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", "0", "近地铁", "近地铁", "0", "0", "近地铁~
## $ if_2y             <chr> "0", "房本满两年", "0", "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> "0", "VR看装修", "0", "0", "VR看装修", "0", "VR看装~
```

```
##
##   北   东 东北 东南   南   西 西北 西南
##   59   81   10  238 2056   14   11   46
```

```
##
##   北   东 东北 东南   南   西 西北 西南
## 1001    5    5   12   54   21    9   11
```

```
##
## 简装 精装 毛坯 其他
##  536 1476  356  147
```

```
##
##   低   高   中
##  662  750 1050
```

```
##
##     板楼 板塔结合     平房     塔楼 暂无数据
##     1503      506        4      441       61
```

```
##
##         0  VR看装修    近地看    近地铁    珞狮南 太子湖1号
##      1205         2         1      1305         1         1
```

各变量的简短统计：

```
##  property_name      property_region      price_ttl        price_sqm
##  Length:2515        Length:2515        Min.   :  10.6   Min.   : 1771
##  Class :character   Class :character   1st Qu.:  95.0   1st Qu.:10765
##  Mode  :character   Mode  :character   Median : 136.0   Median :14309
##                                        Mean   : 154.8   Mean   :15110
##                                        3rd Qu.: 188.0   3rd Qu.:18213
```

```
##                                       Max.   :1380.0   Max.   :44656
##     bedrooms      livingrooms    building_area    directions1
##  Min.   :1.000   Min.   :0.000   Min.   : 22.77   Length:2515
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.: 84.45   Class :character
##  Median :3.000   Median :2.000   Median : 95.46   Mode  :character
##  Mean   :2.689   Mean   :1.706   Mean   :100.67
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:118.03
##  Max.   :7.000   Max.   :4.000   Max.   :588.66
##  directions2         decoration        property_t_height property_height
##  Length:2515        Length:2515        Min.   : 2.00     Length:2515
##  Class :character   Class :character   1st Qu.:11.00     Class :character
##  Mode  :character   Mode  :character   Median :27.00     Mode  :character
##                                        Mean   :24.05
##                                        3rd Qu.:33.00
##                                        Max.   :62.00
##  property_style       followers       near_subway           if_2y
##  Length:2515        Min.   :  0.000   Length:2515        Length:2515
##  Class :character   1st Qu.:  1.000   Class :character   Class :character
##  Mode  :character   Median :  2.000   Mode  :character   Mode  :character
##                     Mean   :  6.326
##                     3rd Qu.:  6.000
##                     Max.   :262.000
##    has_key              vr
##  Length:2515        Length:2515
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

可以看到：

- 直观结论 1
  去重后，数据表（lj) 共包括 property_name, property_region, price_ttl, price_sqm, bedrooms, livingrooms, build-ing_area, directions1, directions2, decoration, property_t_height, property_height, property_style, follow-ers, near_subway, if_2y, has_key, vr 等 18 个变量, 共 2515 行。

- 直观结论 2
  武汉二手房链家数据结果显示：房屋主要朝向包含北、东、东北、东南、南、西、西北、西南等 8 个类别，其中朝南方向的二手房数量最多；装修状况包括简装、精装、毛坯、其他等 4 个类别，其中精

装二手房数量最多；房屋在所在楼栋所处位置包含高中低 3 个类别，其中中层二手房数量最多；除此之外，建筑形式为板楼和近地铁的二手房数量最多。

- 直观结论 3
  武汉二手房房屋总价的均价为 154.8 万元/套，房屋总价最大值为 1380 万元/套；武汉二手房房屋单价均价为 15110 元/m²，房屋单价最大值为 44656 元/m²；武汉二手房平均建筑面积为 100.67m²，建筑面积最大值为 588.66m²。

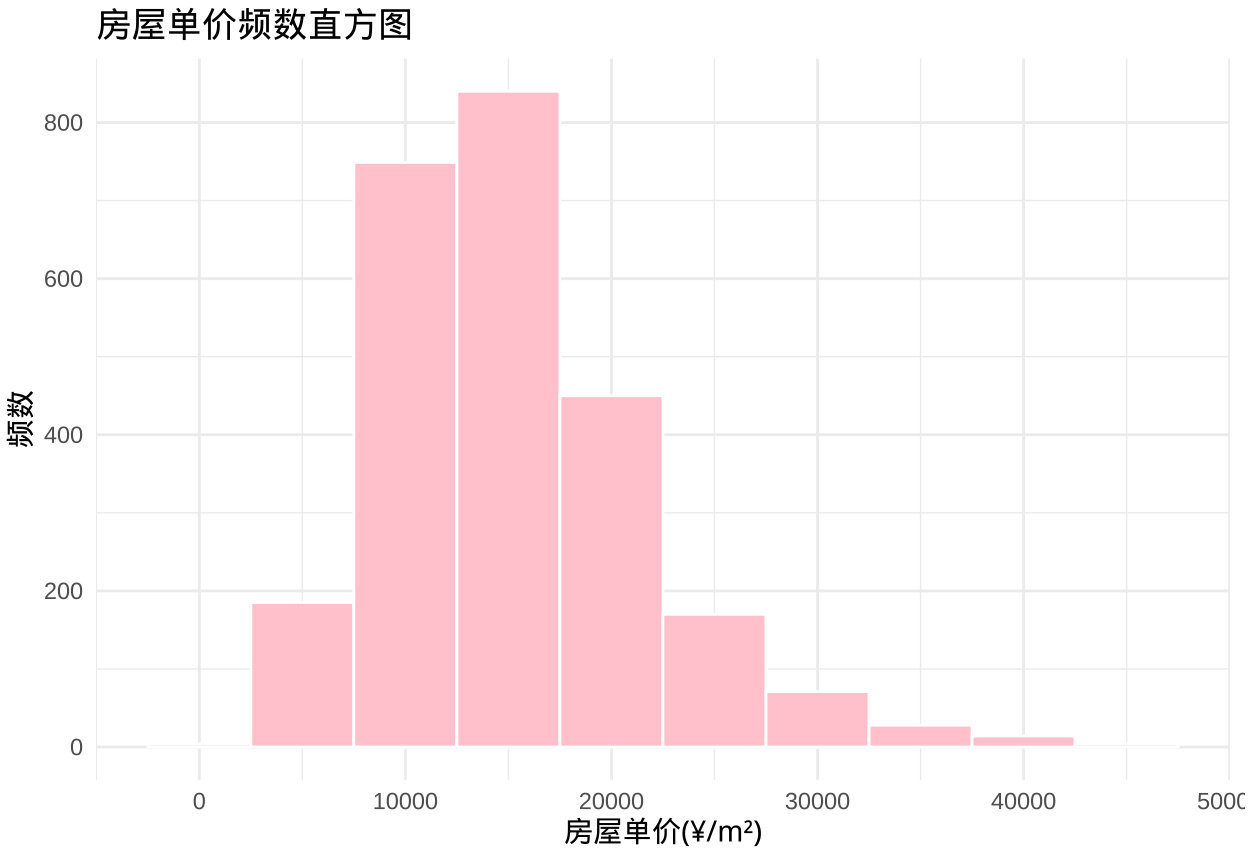
探索性分析

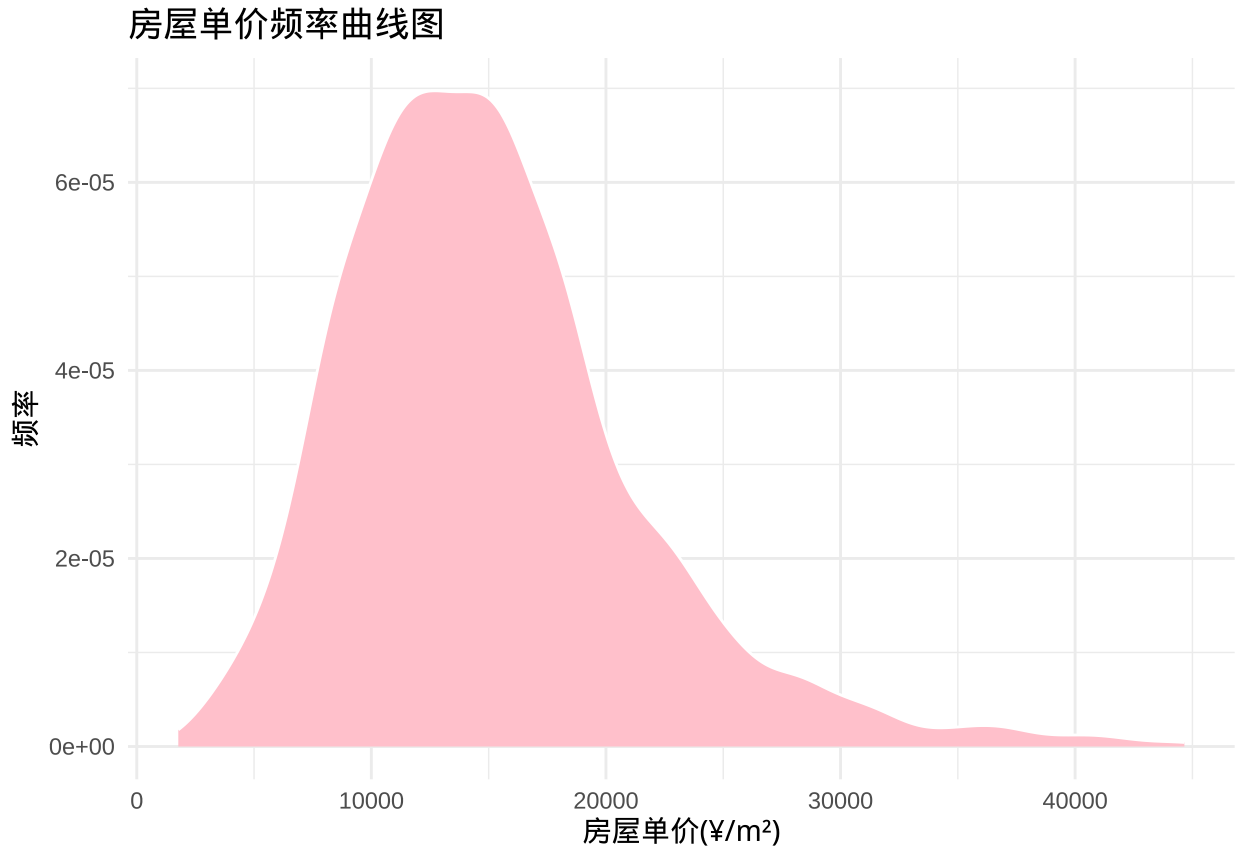
变量 1 的数值描述与图形：房屋单价

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1771   10765   14309   15110   18213   44656
```

```
## 10196
##   510
```

```
## [1] "10196"
```






```
## [1] 1.067103
```

发现：

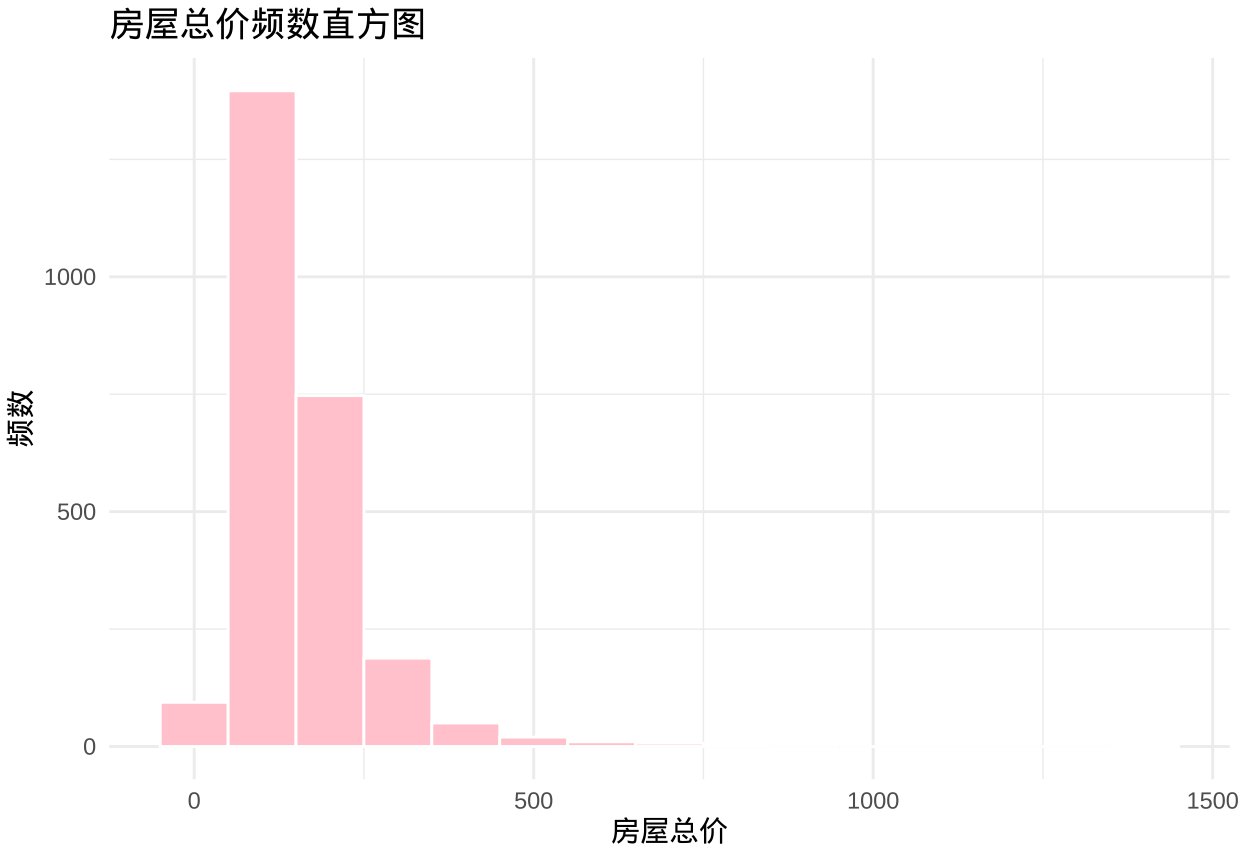
- 发现 1：数据存在偏斜。集中趋势的数值指标显示房屋单价数据的最小值为 1771 元/m²，第一四分位数为 10765 元/元/m²，中位数为 14309 元/元/m²，平均值为 15110 元/m²，第三四分位数为 18213 元/m²，最大值为 44656 元/m²，表明数据存在一定的偏斜；频数直方图显示房屋单价主要集中在 10000-20000 元/m² 之间，频率曲线图进一步展示了这一分布趋势，这些图表都表明数据存在一定的偏斜。
- 发现 2：数据分布向右偏斜，即存在正偏分布。通过偏度计算公式表明偏度值为 1.06，说明数据分布向右偏斜，即存在正偏分布。

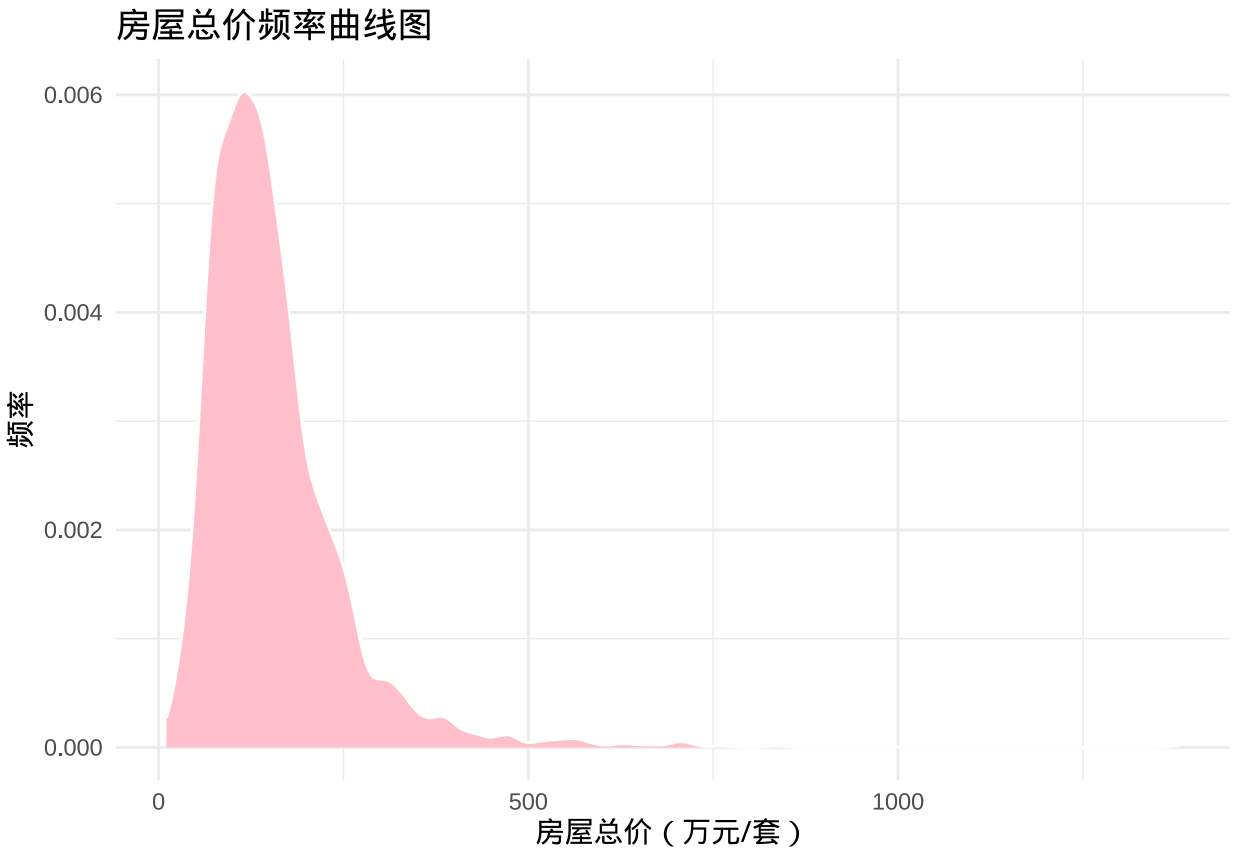
变量 2 的数值描述与图形: 房屋总价

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.6    95.0   136.0   154.8   188.0  1380.0
```

```
## 105
## 140
```

```
## [1] "105"
```


房屋总价频数直方图
1000
500
0
频数
0
500
1000
1500
房屋总价


房屋总价频率曲线图
0.006
0.004
0.002
0.000
频率
0
500
1000
房屋总价（万元/套）

```
## [1] 2.806202
```

发现：

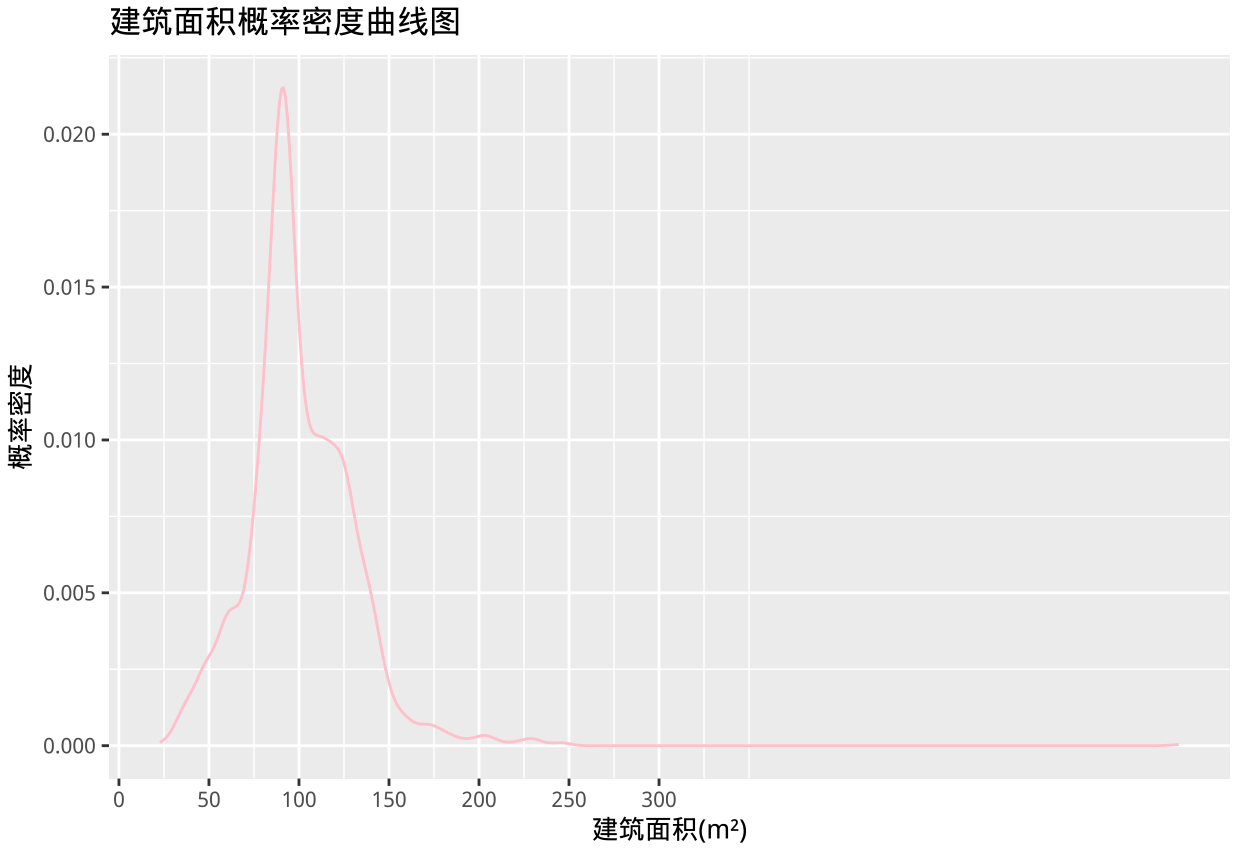
- 发现 1：数据存在偏斜。集中趋势的数值指标显示房屋总价数据的最小值为 10.6 万元/m²，第一四分位数为 95 万元/m²，中位数为 136 万元/m²，平均值为 154.8 万元/m²，第三四分位数为 188 万元/m²，最大值为 1380 万元/m²，表明数据存在一定的偏斜；频数直方图显示房屋总价主要集中在 50-150 万元/m² 之间，频率曲线图进一步展示了这一分布趋势，这些图表都表明数据存在一定的偏斜。
- 发现 2：数据分布向右偏斜，即存在正偏分布。通过偏度计算公式表明偏度值为 2.8，说明数据分布向右偏斜，即存在正偏分布。
- 发现 3：初步猜测房屋面积是房屋总价和房屋单价右偏程度相差较大的原因。房屋单价与房屋总价均呈现右偏分布，但房屋单价的右偏程度较房屋总价的右偏程度小，初步猜测房屋面积是房屋总价和房屋单价右偏相差较大的原因，下一步单因素变量分析可选择建筑面积这一变量的差异程度来进一步证实此猜想。

变量 3 的数值描述与图形：建筑面积

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.77   84.45   95.46  100.67  118.03  588.66
```

```
## 88.58
##   701
```

```
## [1] "88.58"
```

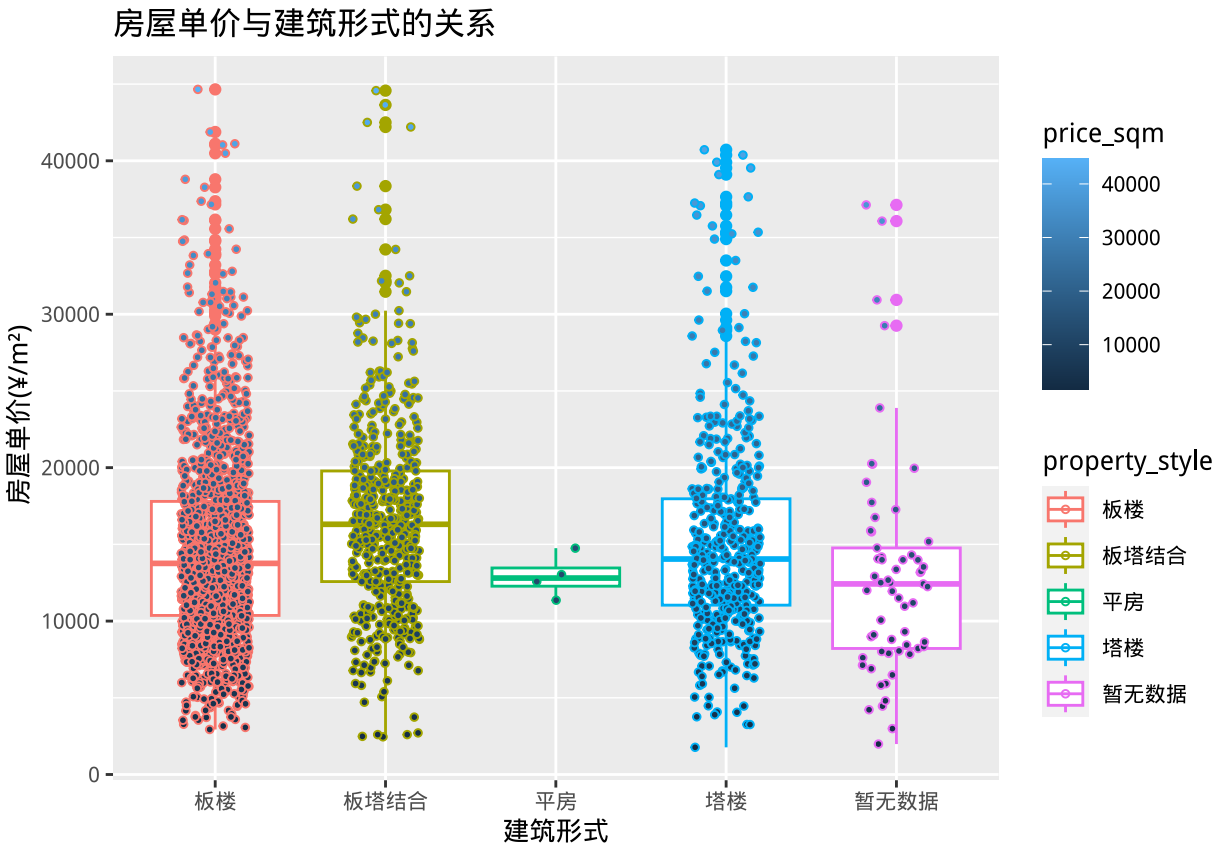


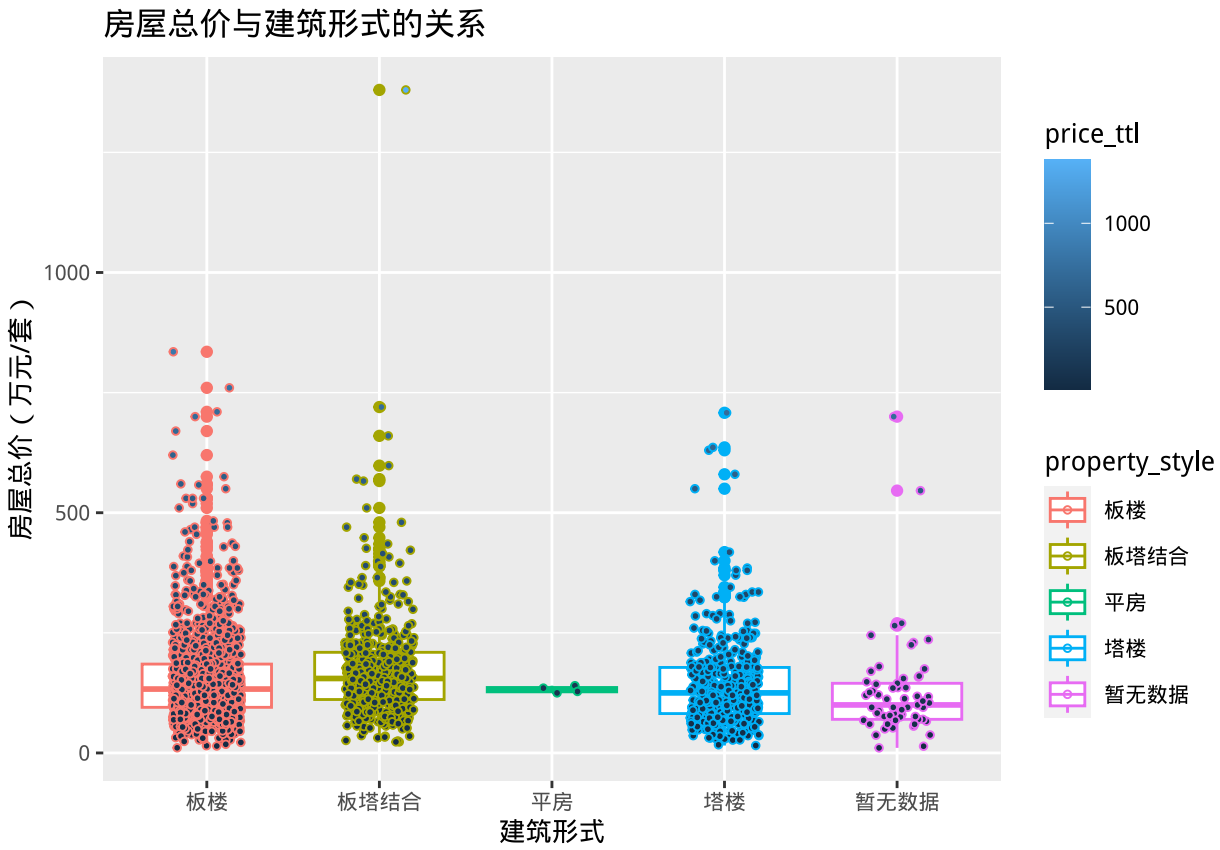
```
## [1] 2.238489
```

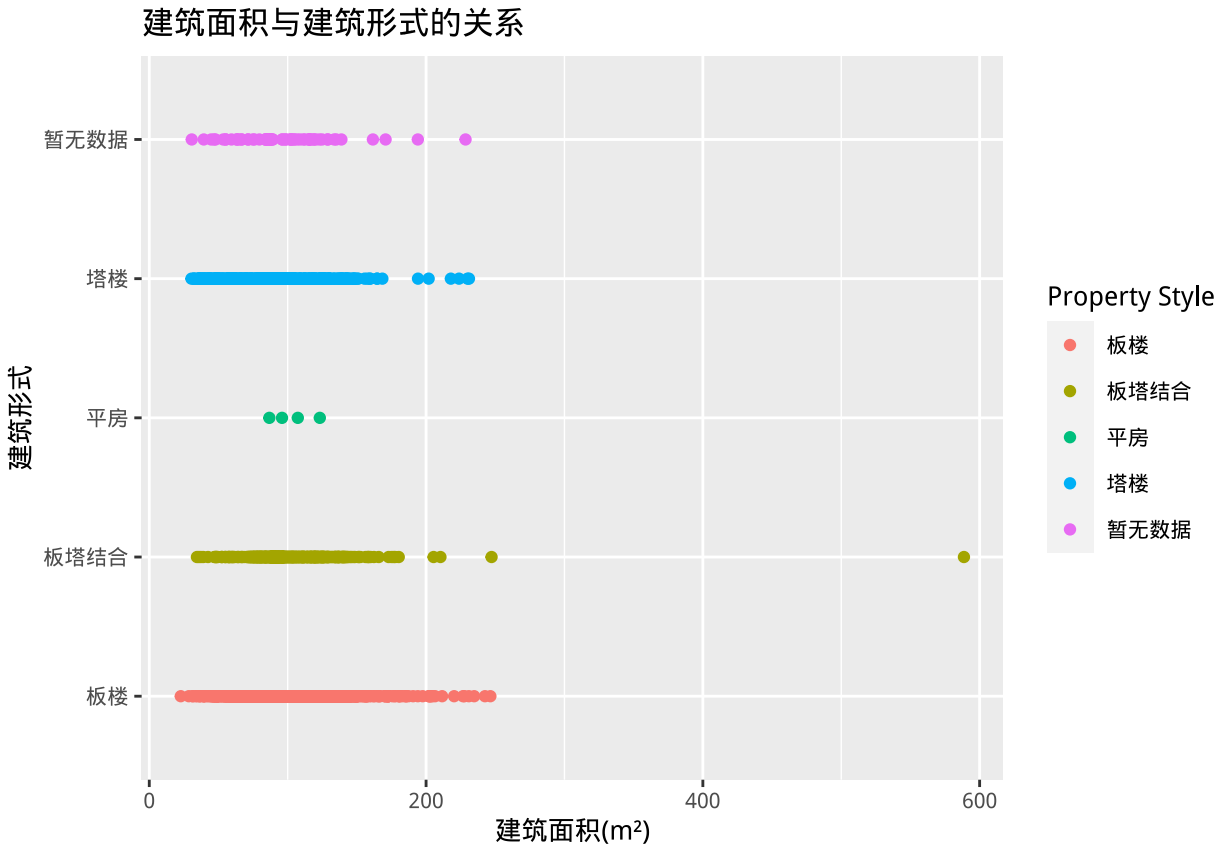
发现：

- 发现 1 数据存在偏斜。集中趋势的数值指标显示建筑面积数据的最小值为 22.77m²，第一四分位数为 84.45m²，中位数为 95.46m²，平均值为 100.67m²，第三四分位数为 118.03m²，最大值为 588.66m²，最小值和最大值之间的差距非常大，表明数据存在一定的偏斜；建筑面积概率密度曲线图显示建筑面积主要集中在 75-125m² 之间，进一步表明数据存在一定的偏斜。
- 发现 2 数据分布向右偏斜，差异程度较大，造成房屋总价和房屋单价右偏相差较大。通过偏度计算公式表明偏度值为 2.24，说明数据分布向右偏斜，即存在正偏分布。符合上述房屋面积对房屋总价的影响的初步猜测。

探索问题 1：板塔结合房屋单价是影响板塔结合房屋总价的主要原因吗？







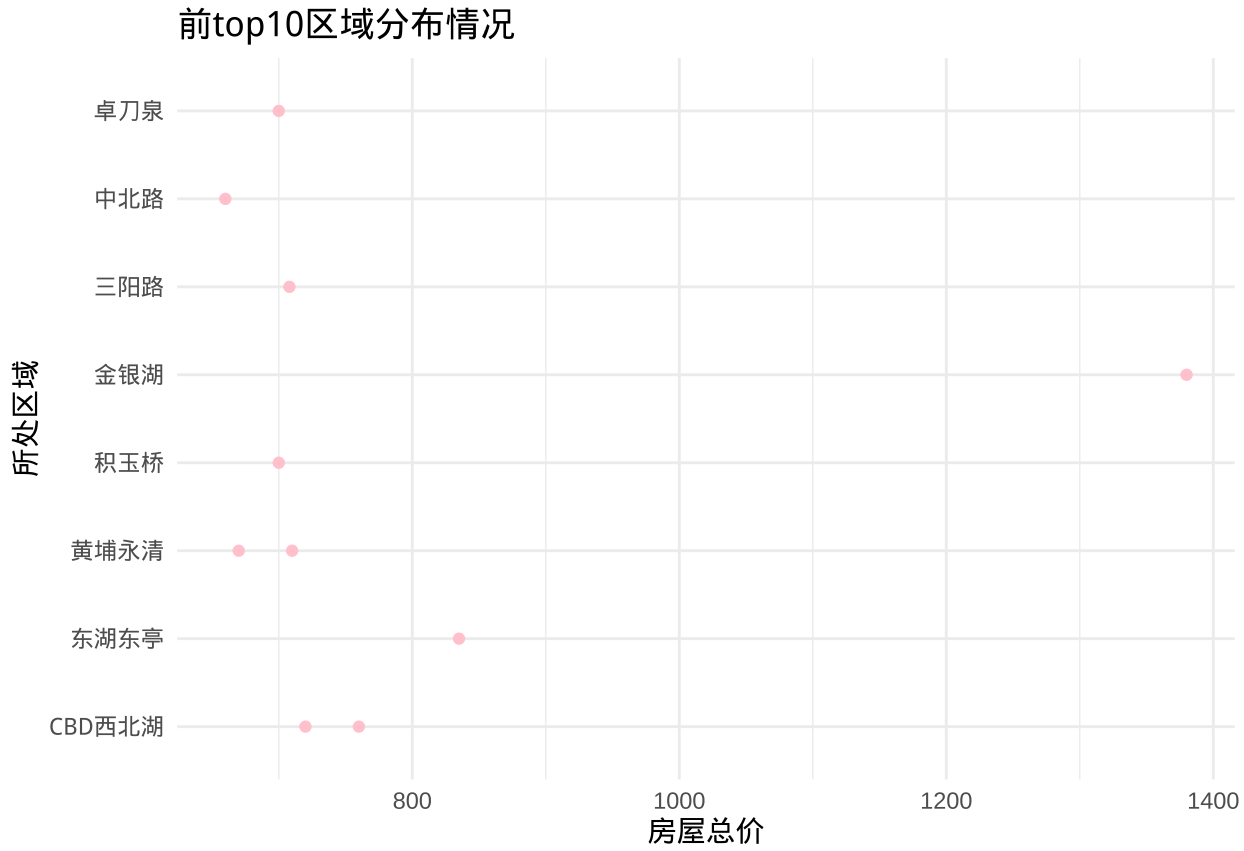
发现：

- 发现 1 板塔结合的房屋总价和房屋单价均是最高的。通过房屋单价与建筑形式的关系图，可以直观

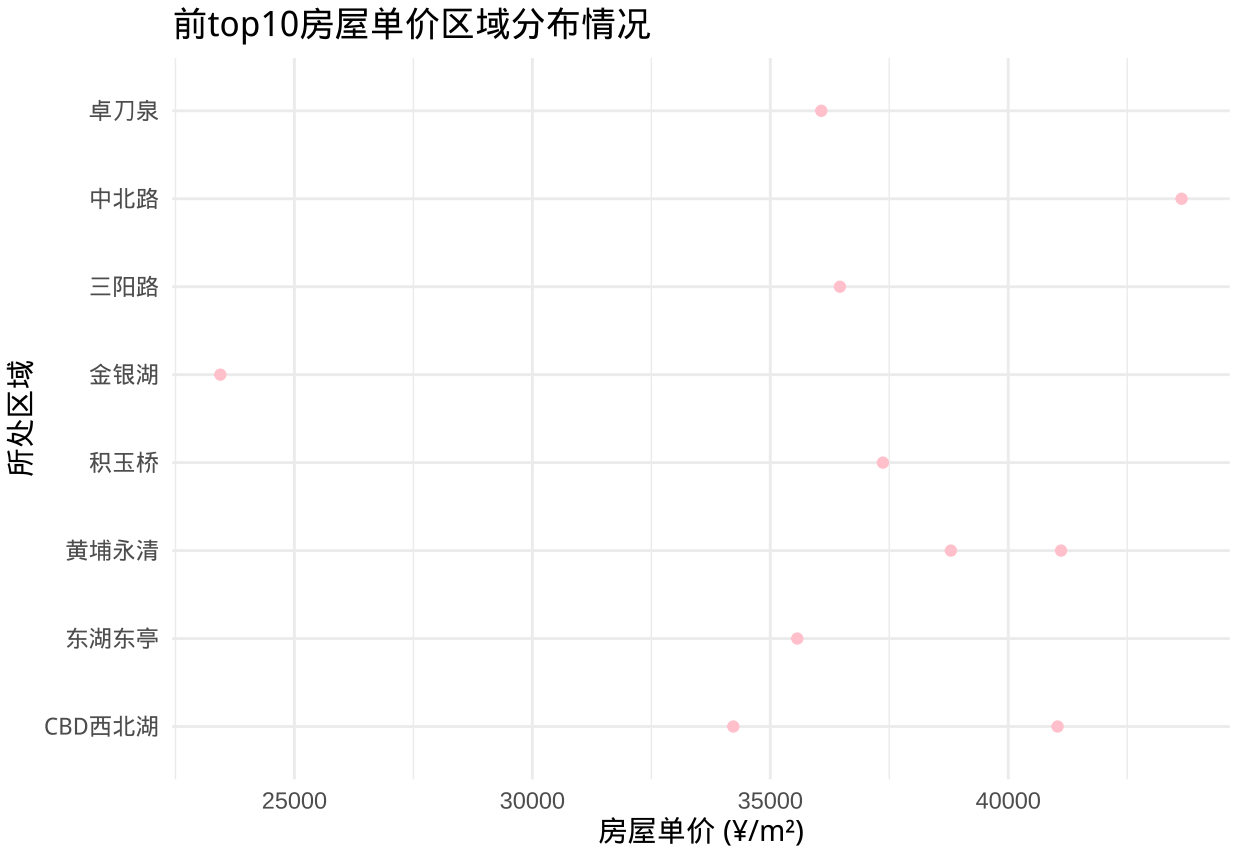
地观察到：板塔结合的整体房屋单价是最高的。通过房屋总价与建筑形式的关系图，可以直观地观察到：板塔结合的整体房屋总价是最高的。初步猜想，板塔结合的房屋单价高导致了板塔结合的整体房屋总价高。

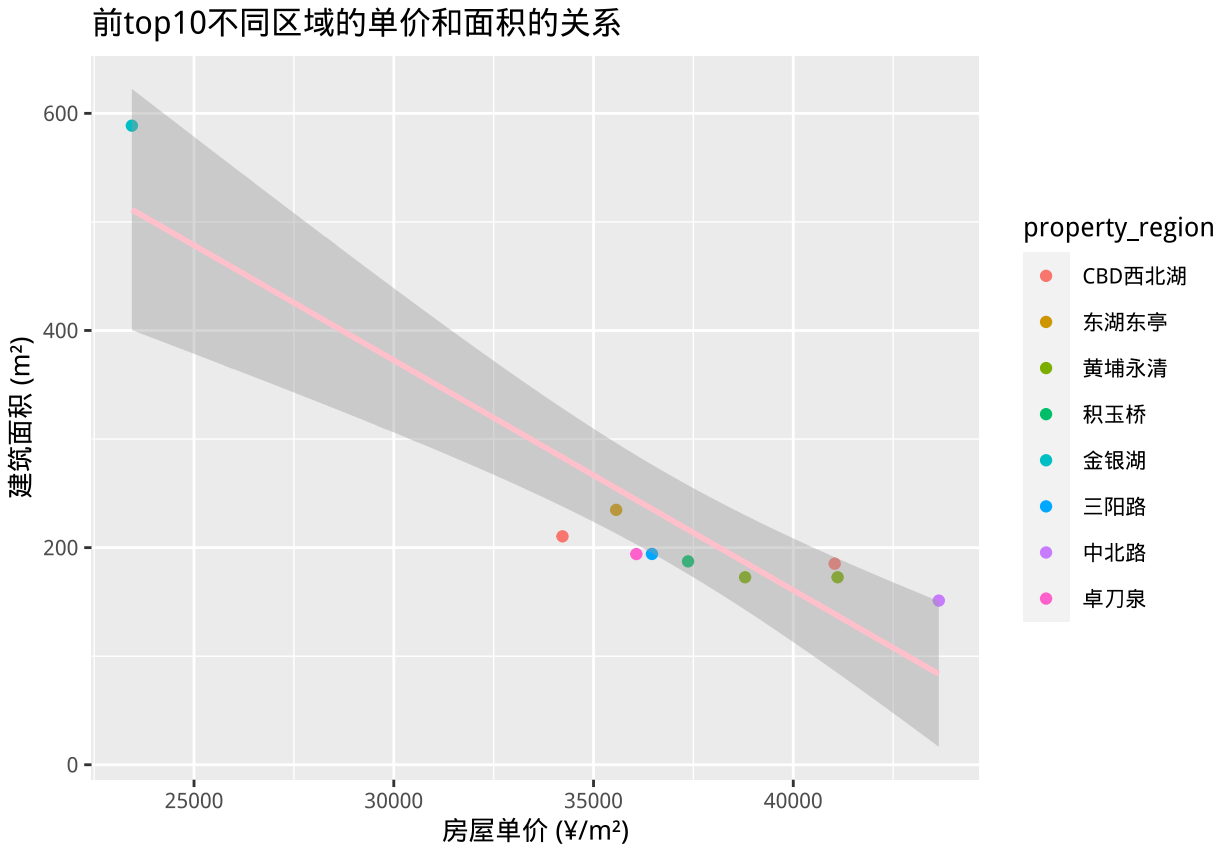
- 发现 2 板塔结合的房屋单价高确实导致了板塔结合的整体房屋总价高。通过建筑面积与建筑形式的关系图，可以大致得到板塔结合与其他各种建筑形式的建筑面积的范围相差不大，因此，可以验证板塔结合的房屋单价高确实导致了板塔结合的整体房屋总价高。

探索问题 2：造成金银湖房屋总价最高，甚至超过西北湖、黄浦路及东湖等区域的原因是什么？




```
## # A tibble: 1 x 18
##   property_name      property_region price_ttl price_sqm bedrooms livingrooms
##   <chr>              <chr>                <dbl>     <dbl>    <dbl>       <dbl>
## 1 万科高尔夫城市花园 金银湖                1380     23444        7           3
## # 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>
```





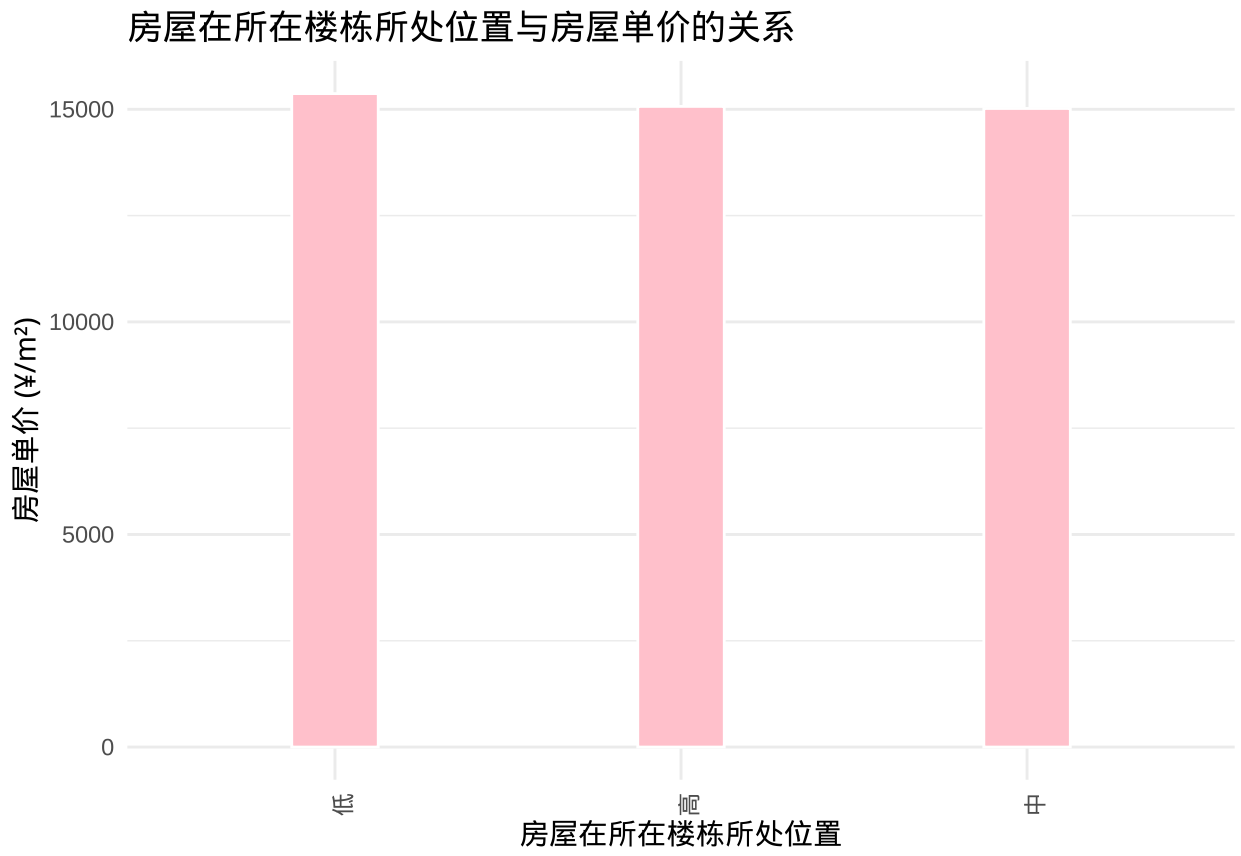
发现：

- 发现 1 金银湖所在区域的房屋总价最高。通过散点图-“前 top10 区域分布情况” 可以直观看出：房屋

总价 top-10 的区域在 CBD 西北湖、东湖东亭、黄埔永清、积玉桥、金银湖、三阳路、中北路及卓刀泉，其中金银湖所在区域的房屋总价最高。除了金银湖，基本满足消费者对武汉房价的基本认知。

- 发现 2 金银湖建筑面积大是影响其房屋总价高的主要原因。通过散点图-“前 top10 房屋单价区域分布情况” 可以直观看出：金银湖的房屋单价在 top-10 所处的 8 个区域中属于最小值，远远低于其他区域，因此房屋单价不是它总价最高的原因；通过进一步比较不同区域的单价和面积的关系可以看出：与房屋单价的区域分布情况相反，金银湖的建筑面积远远高于其他区域。因此，金银湖建筑面积大是影响其房屋总价高的主要原因。

探索问题 3：房屋在所在楼栋所处位置对房屋单价的影响程度大吗？哪一个位置相对较高？




```
## [1] 15110.42
```

```
## [1] 15369.01
```

```
## [1] 15019.14
```

```
## [1] 15067.72
```

```
## [1] 258.59
```

```
## [1] -91.28
```

```
## [1] -42.7
```

发现：

- 发现 1 房屋低、中、高的位置因素并不是房屋单价的主导因素。低、中、高三类房屋的平均房屋单价存在差距，但差距不大。这可能表明这三类房屋的市场价值在不同程度上有所波动，但整体上较为接近，那么可能说明房屋的位置因素并不是房屋单价的主导因素，影响更大的可能是其他因素，例如所处区域、建筑面积、装修等。
- 发现 2 低层位置对房屋单价的影响度相对更为明显。通过计算每一类房屋平均房屋单价与整体平均房屋单价的偏离程度或差异程度，可以进一步理解房屋在所在楼栋所处位置与房屋单价之间的关系，结果显示：低层房屋的平均价格大大高于整体平均价格，而中、高层房屋的平均价格低于整体平均价格，说明低层位置对房屋单价的影响度相对更为明显。

发现总结

总结 1：
武汉二手房链家的房屋总价平均价格为 154.8 万元/套，房屋单价平均价格为 15110 元/m^2。这些数据提供了对武汉地区二手房市场的基本认识，并为潜在买家或研究者提供了参考；
总结 2：
数据分析显示，房屋单价和房屋总价的数据分布向右偏斜，这意味着大部分房屋的价格在平均价格以下，存在一定的价格下偏分布。然而，房屋单价的右偏程度较房屋总价的小，这可能是因为房屋单价的变化幅度较小，或者低价房屋在总样本中占据的比例较小。房屋建筑面积的数据分布也向右偏斜，但与房屋总价和房屋单价的分布有所不同。这可能是因为在相同的价格水平上，不同的房屋类型或地理位置等因素会影响到建筑面积；
总结 3：
通过多变量的相关性分析，发现低层位置对房屋单价的影响度相对更为明显。这可能是因为低层房屋通常具有更好的便利性和视野，从而在价格上有所体现。另外，板塔结合的房屋单价高是板塔结合的房屋总价高的主要原因，这也可能是因为板塔结合的房屋通常具有更好的建筑质量和居住体验。在所有房屋所处区域中，金银湖区域的房屋总价最高，而其建筑面积大是其主要原因。这可能是因为金银湖区域有着较好的环境和生活设施，从而吸引了较高的房价。所以房屋单价的影响因素是多样且交叉的，每个影响因素对于潜在买家、卖家和政策制定者都具有参考价值。