关于二手房的数据分析研究

任一轩

2023-10-20

目录

1 主要发现

1. 房屋所在地段是影响房屋单价的主要因素
2. 四居室、五居室房屋单价较高

3. 邻近地铁的房屋单价会略高于不邻近地铁的房屋

2 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。

```
## # A tibble: 3,000 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 2,990 more rows
## # 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>
```

3 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
```

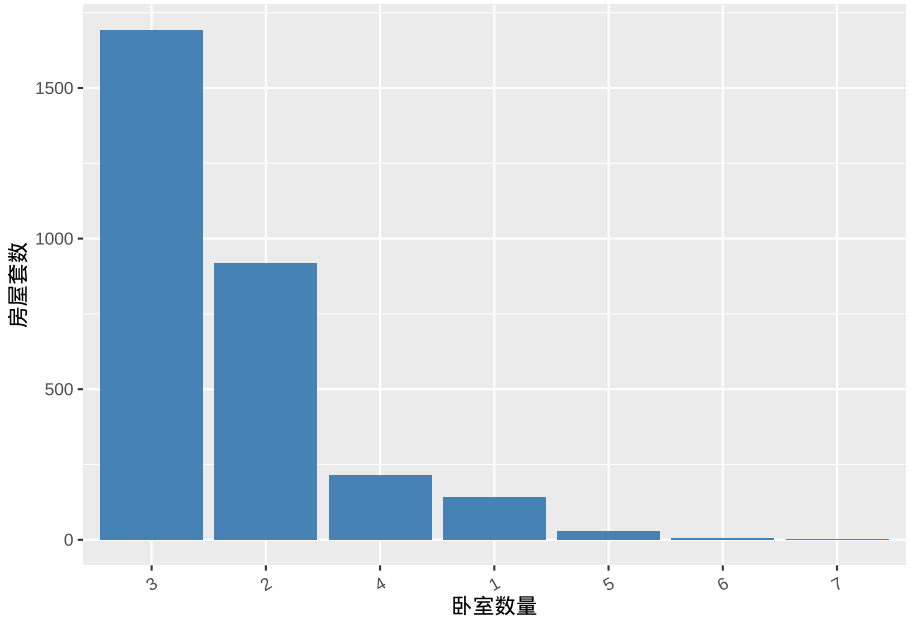
```
##                                      Mean   :24.22
##                                      3rd Qu.:33.00
##                                      Max.   :62.00
##  property_style       followers       near_subway           if_2y
##  Length:3000        Min.   :  0.000   Length:3000        Length:3000
##  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：

- 表中房屋售价总价的最大值为 1380 万元，而最小值仅为 10.6 万元，相差极大，最大值为最小值的 100 倍以上。
- 表中房屋单价最大值为 44656 元/㎡，最小值为 1771 元/㎡，最大值为最小值的 25 倍左右。
- 表中房屋建筑面积最大值为 588 ㎡，最小值为 23 ㎡，中位数和平均数接近，约 100 ㎡。

4 探索性分析

4.1 查看数据集户型和建筑面积的分布



发现：

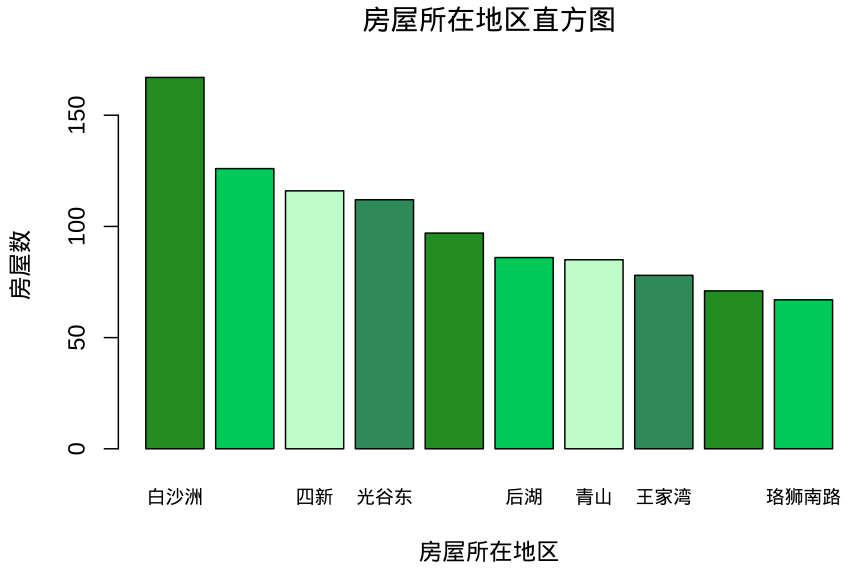
1. 三居室和二居室是本数据集中最常见的两种户型。

2. 样本房屋面积集中于 100 ㎡左右。

4.2 统计二手房所在地区的分布情况

```
##
##  CBD西北湖  VR看装修    白沙洲    百步亭  宝丰崇仁  蔡甸城区  蔡甸其它
##          35         1       167        57        22        37         3
##      藏龙岛  常青花园    常青路  楚河汉街    大智路      堤角  东湖东亭
##          35        36        10        15         7        37        38
##  东西湖其它       沌口       二七       古田  关山大道  关西长职    光谷东
```

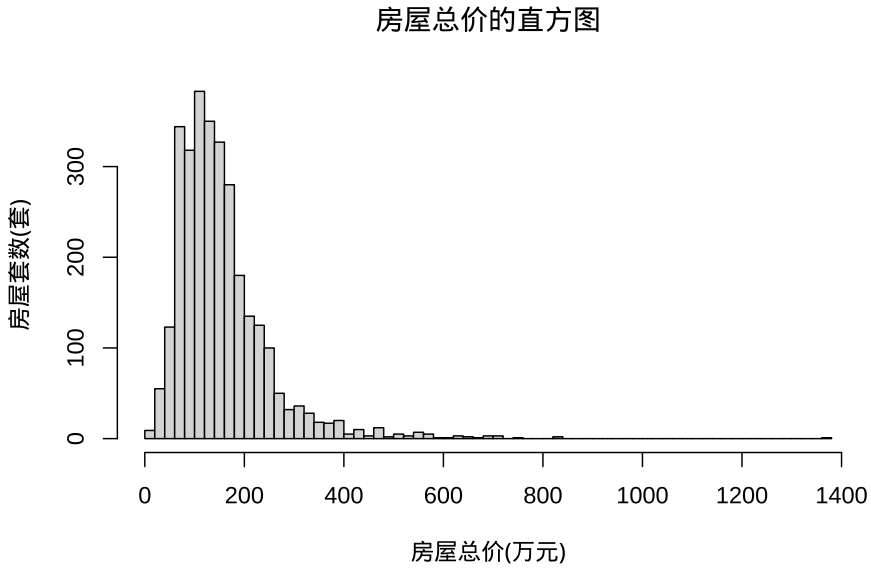
```
##       7      54      43      41      25      23     112
##  光谷广场  光谷南  国际百纳  汉口北  汉南其它  汉正街  洪山其它
##      25      27       1       1      16      10       6
##  后官湖  后湖  虎泉杨家湾  华科大  黄陂其它  黄家湖  黄埔永清
##       3      86      21      26       6      10      23
##  积玉桥  集贤  江夏其它  将军路  街道可  街道口  金融港
##      63      10       6      19       2      13      52
##  金银湖  近地铁  老南湖  珞狮南路  庙山  民族大道  南湖沃尔玛
##      97       3      54      67      32      61      33
##  盘龙城  七里庙  前进江汉  青山  三环南  三阳路  沙湖
##     126      35      19      85      18      16      32
##  首义  水果湖  四新  随时看  随时看房  塔子湖  台北香港路
##      24       9     116       2       3      71      24
##  唐家墩  团结大道  王家湾  文化大道  吴家山  武昌火车站  武广万松园
##      18      65      78      66      21      21       9
##  武湖  新华路万达  新南湖  徐东  阳逻  杨汉湖  杨园
##      15      16      29      47      66      33      22
##  育才花桥  长丰常码头  长港路  纸坊  中北路  中法生态城  中南丁字桥
##      18      35      42      31      18      12      51
##  钟家村  卓刀泉  宗关
##      65      30      34
```

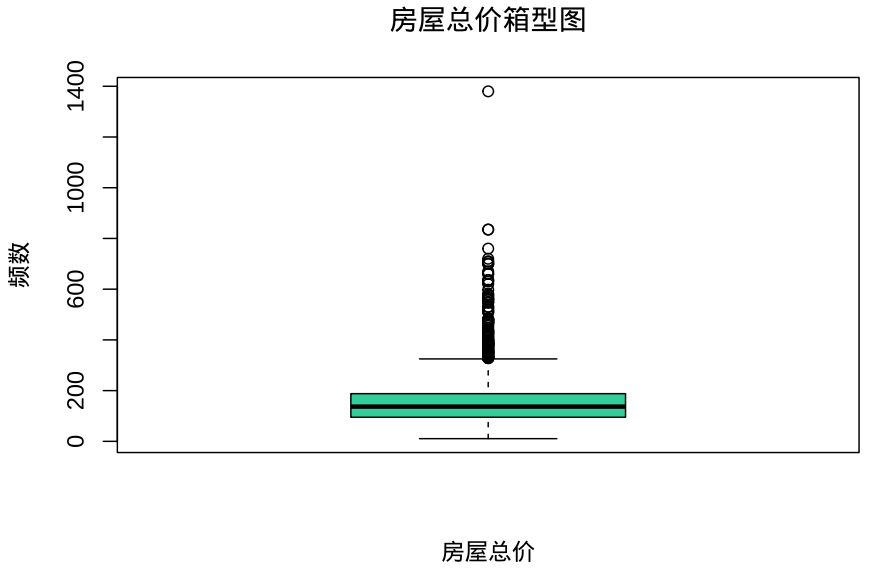


发现：样本中的房屋位置大多在三环外，远离市中心的位置。

4.3 房屋总价的数值描述与图形

```
## # A tibble: 3,000 x 1
##    `lj$price_ttl`
##             <dbl>
##  1            237
##  2            127
##  3             75
##  4            188
##  5            182
##  6            122
##  7             99
##  8            194.
##  9            325
## 10            192
## # i 2,990 more rows
```





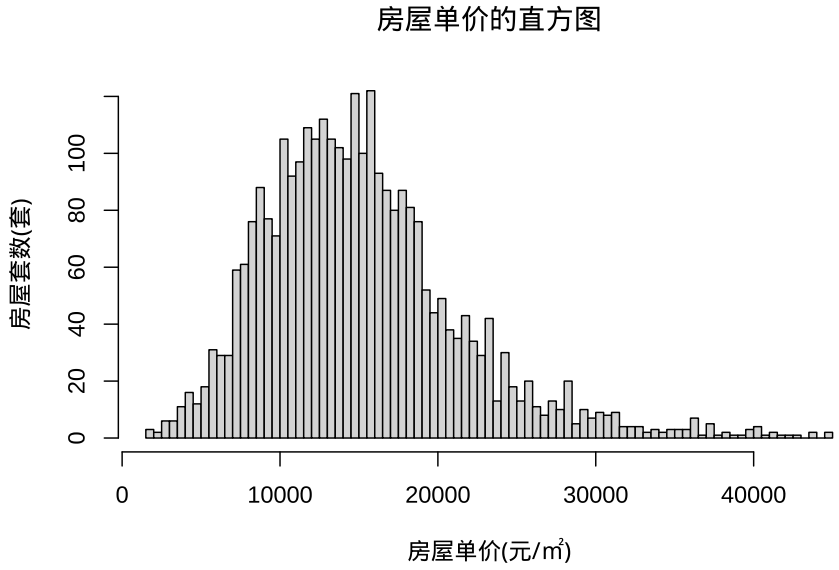
发现：

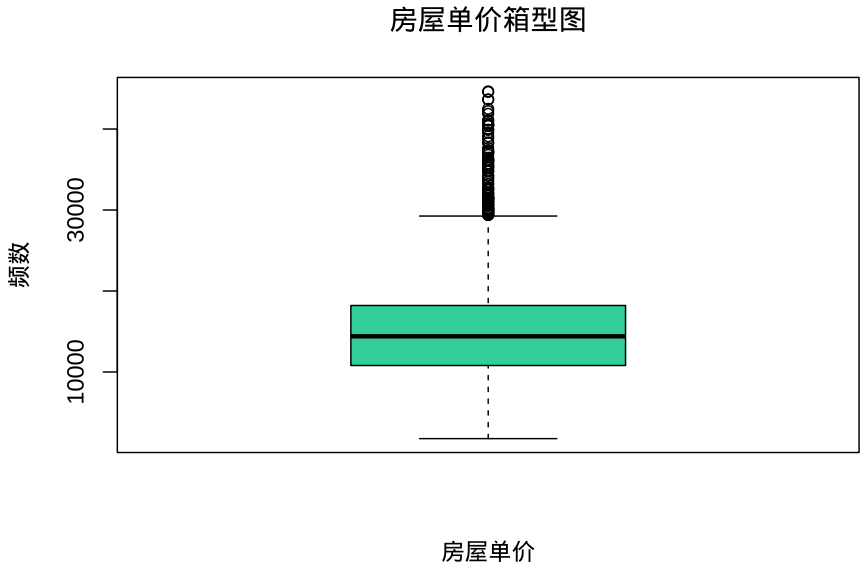
1. 样本集内房屋总价集中于 150 万-200 万的区间

2. 样本所在地和户型及面积的情况与总价的分布情况是一致的。

4.4 房屋单价的数值描述与图形

```
## # A tibble: 3,000 x 1
##    `lj$price_sqm`
##             <dbl>
##  1          18709
##  2          14613
##  3          15968
##  4          15702
##  5          17509
##  6          10376
##  7          12346
##  8          16336
##  9          32631
## 10          17403
## # i 2,990 more rows
```





发现：样本集内房屋单价集中于 1 万/㎡-2 万/㎡的区间。

4.5 影响房屋单价的主要因素是什么？

```
## # A tibble: 10 x 18
##    property_name      property_region price_ttl price_sqm bedrooms livingrooms
##    <chr>              <chr>                <dbl>     <dbl>    <dbl>       <dbl>
##  1 中商宿舍           中南丁字桥             320     44656        2           1
##  2 复地东湖国际五六期 中北路               598     44574        3           2
##  3 复地东湖国际一期   中北路               660     43643        4           2
##  4 复地东湖国际一期   中北路               660     43643        4           2
##  5 复地东湖国际五六期 中北路               570     42503        3           2
##  6 复地东湖国际五六期 中北路               566     42205        3           2
##  7 华发外滩首府       黄埔永清             530     41878        3           2
##  8 华发外滩首府       黄埔永清             710     41110        4           2
##  9 华发中城荟         CBD西北湖            760     41037        4           2
## 10 复地东湖国际二期   中北路               380     40721        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>
```

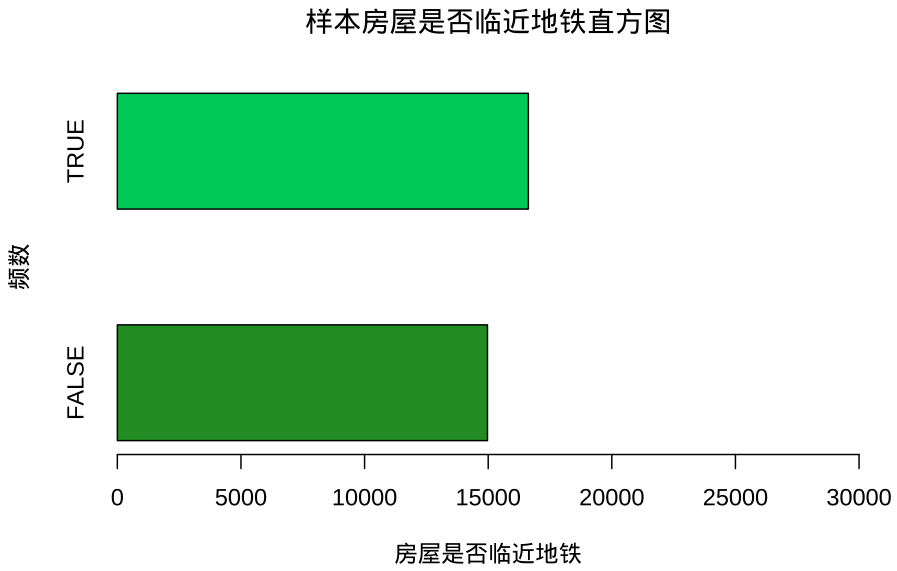
```
## # A tibble: 10 x 18
##    property_name property_region price_ttl price_sqm bedrooms livingrooms
##    <chr>         <chr>                <dbl>     <dbl>    <dbl>       <dbl>
##  1 天益华庭      阳逻                  17        1771        2           2
##  2 天益华庭      阳逻                  17        1771        2           2
##  3 农垦街        汉南其它              10.6      1984        1           1
##  4 常乐新城      阳逻                  23        2456        2           2
##  5 锦绣家园      阳逻                  23        2484        2           2
##  6 常乐新城      阳逻                  33        2599        3           2
##  7 精致阳光嘉园  阳逻                  26        2601        2           2
##  8 精致阳光嘉园  阳逻                  26        2601        2           2
##  9 桃园居小区    阳逻                  25        2710        3           2
## 10 武汉SOHO      阳逻                  10.8      2934        1           1
## # 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>
```

通过对比，可以发现：

- 1. 单价高的房屋地段集中于市中心，多为三室以上的大户型、精装修。
- 2. 反之，单价低的房屋地段位于偏远地段，面积较小。

4.6 邻近地铁这个因素是否对房价有影响？

```
##    FALSE     TRUE
## 14970.60 16623.93
```

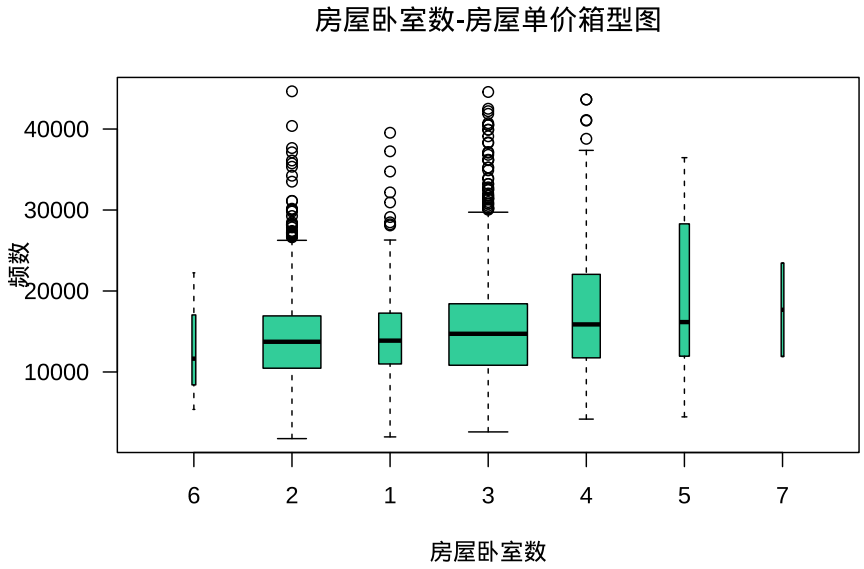


发现：邻近地铁的因素在本样本集中对房屋单价有一定影响，但影响程度大小有待进一步量化分析。

4.7 卧室数量这个因素是否对房价有影响？

```
##
##    1    2    3    4    5    6    7
##  142  919 1692  214   27    4    2
```

```
##        1        2        3        4        5        6        7
## 14924.05 14412.80 15208.88 17392.69 19969.41 12726.75 17679.50
```



发现：四居室、五居室房屋明显单价较二居室、三居室房屋高。因前者多为大面积、高质量房屋，此结论也符合常识。

---

5 发现总结

经由分析发现，房屋所在地段是影响房屋单价的主要因素，邻近地铁的房屋单价会略高于不邻近地铁的房屋，四居室、五居室房屋单价较高。

总体而论，地段好、出行方便、房型好、面积大的房屋更宜居，单价也更昂贵。