链家武汉二手房数据分析

常世俊

2023-10-17

目录

1 你的主要发现

发现 1. 从建筑类型和房屋单价这两个变量进行分析，发现了” 板塔结合” 这类建筑类型的房屋整体单价偏高，“平房” 这类建筑类型的房屋整体单价偏低的结论。

发现 2. 从附近有无地铁和房屋单价这两个变量进行分析，发现了附近有地铁的房子整体单价都比附近没有地铁的单价高的结论。

发现 3. 从房屋总价和房屋面积这两个变量进行分析，发现了房屋总价大部分集中在 0 至 500 万元中间，房屋面积大部分集中在 0 至 200 平方米中间。在此区间范围内，随着房屋总价的提升，房屋面积也在不断变大。

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。

```
## [1] "directions2"     "property_height" "near_subway"     "if_2y"
## [5] "has_key"         "vr"
```

通过数据清洗，得到 2515 行链家二手房源信息，并发现”directions2”、“property_height”、“near_subway” 、“if_2y”、“has_key” 和”vr” 这 6 个变量里有空缺值。# 数据概览

数据表（lj1) 共包括 18 个变量, 去除重复项后共 2515 行。表的前 10 行示例如下：

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", "~
```

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

```
## $ 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> "1", "0", "1", "1", "0", "0", "1", "1", "1", "1", "1~
## $ if_2y             <chr> "0", "1", "0", "1", "1", "1", "0", "1", "0", "1", "0~
## $ has_key           <chr> "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1~
## $ vr                <chr> "0", "1", "0", "0", "1", "0", "1", "0", "0", "0", "0~
```

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：武汉二手房屋总价最贵的 1380 万元，最便宜的 10.6 万元；武汉二手房屋单价最贵的每平方米 44656 元，最便宜的 1771 元；武汉二手房屋房间数最多的 7 间，最少的 1 间；武汉二手房屋客厅数最多的 4 间，最少的 0 间；武汉二手房屋建筑面积最大的 588.66 平方米，最小的 22.77 平方米；武汉二手房楼层总层数最高 62 层，最矮的 2 层；武汉二手房最热门的房子有 262 个关注者。
- 直观结论 2：武汉二手房屋总价均价为 136 万元；武汉二手房屋单价均价为每平方米 14309 元；武汉二手房屋房间数平均有 3 间；武汉二手房屋客厅数平均有 2 间；武汉二手房屋建筑面积平均 95.46 平方米；武汉二手房楼层总层数平均 24 层。

3 探索性分析

3.1 变量 1 的数值描述与图形：房屋单价价格特点

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

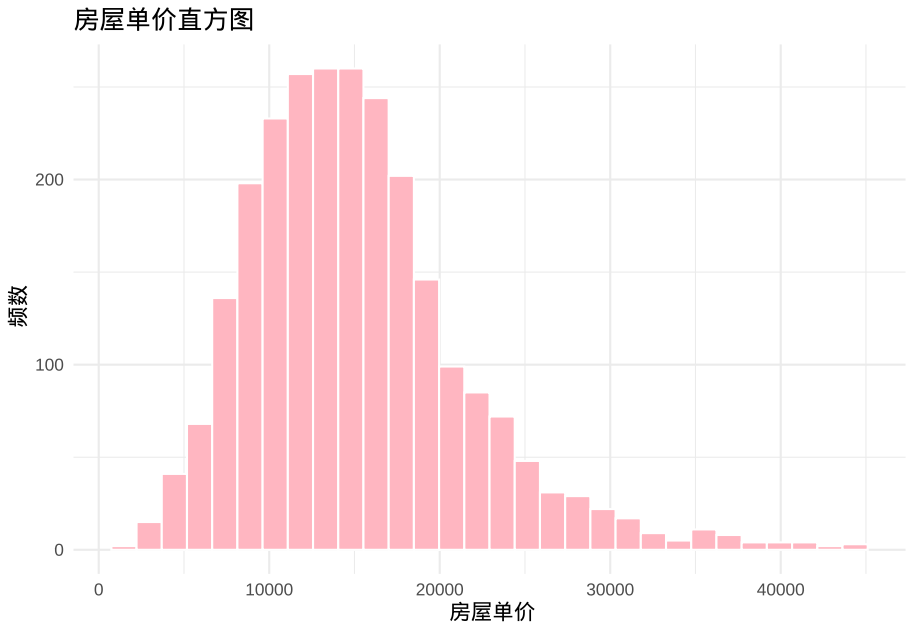
```
##  75%
## TRUE
```

```
## [1] 7447.5
```

```
##     75%
## 18212.5
```

```
##   25%
## 10765
```

```
## [1] 6347.032
```

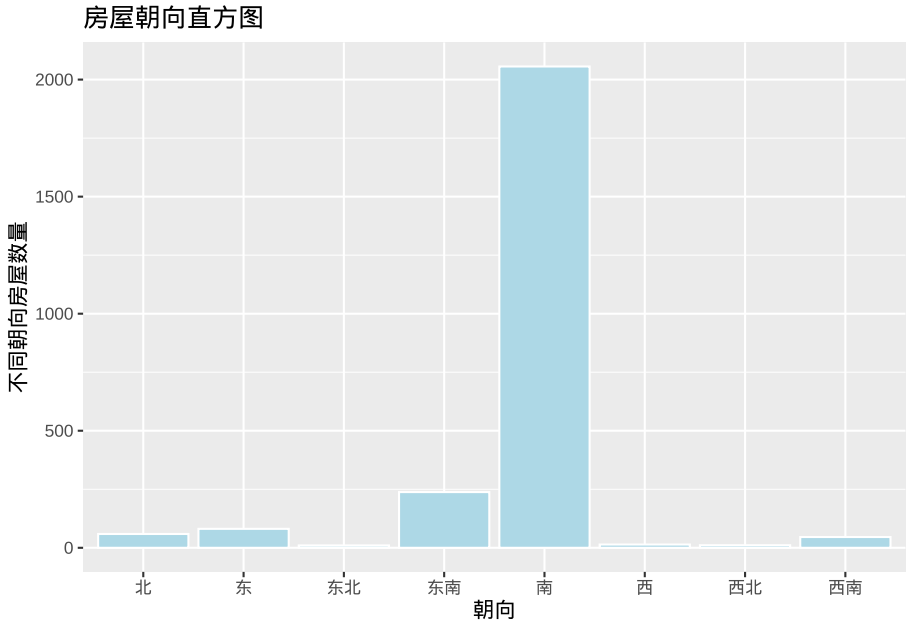


发现：

- 发现 1：房屋单价的数值分析：从集中趋势上，房屋单价众数为 10196；平均值为 15110.42；中位数为 14309。从离散趋势上，房屋单价四分位数间距为 7447.5, 上分位数为 18212.5，下分位数为 10765。二手房单价的标准差为 6347，表示房屋单价的观测值较分散，集中趋势较差。
- 发现 2 房屋单价直方图可以看出来房屋单价的累计频数分布是偏度为 1.06 的右偏，并且价格范围在 10765 到 18212.5 的房子数量最多。

3.2 变量 2 的数值描述与图形：房屋主要朝向

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

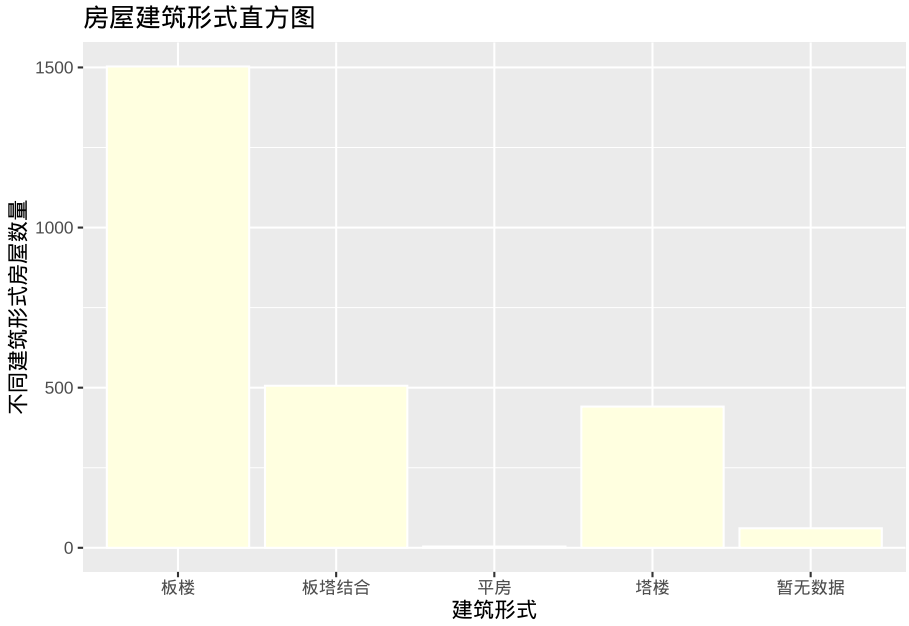


发现：

武汉的链家二手房的朝向分为” 北”,“东”,“东北”,“东南”,“南”,“西”,“西北”,“西南” 八种。其中” 南” 朝向的房子数量最多。“东北” 和” 西北” 朝向的房子数量最小。

3.3 变量 3 的数值描述与图形：房屋的建筑形式

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



发现：

武汉的链家二手房的建筑形式分为” 板楼”,“板塔结合”,“平房”,“塔楼” 四种。其中” 板楼” 形式的房子数量最多。“平房” 形式的房子数量最小。

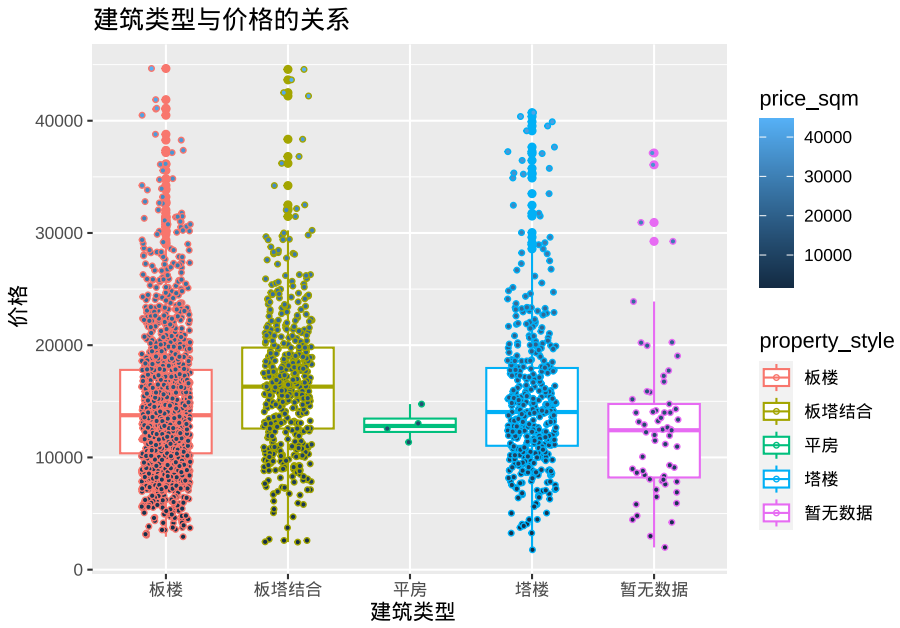
3.4 探索问题 1：房屋单价和建筑类型的关系

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2934   10368   13761   14650   17801   44656

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2456   12574   16308   16684   19784   44574

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11361   12273   12804   12930   13462   14752

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1771   11038   14045   15189   17970   40721
```



发现：

首先把” 板楼”,“板塔结合”,“平房”,“塔楼” 四种不同建筑类型的房子划分成四个表格，然后分别获取描述性统计量得到表一。从表一可以看出不同建筑类型的房屋单价均值和中位数大小顺序是一样的，顺序如下：板塔结合 > 塔楼 > 板楼 > 平房。可以得到” 板塔结合” 这类建筑类型的房子整体价格最高。“平房” 这类建筑类型的房子不仅数量较少且整体价格偏低。

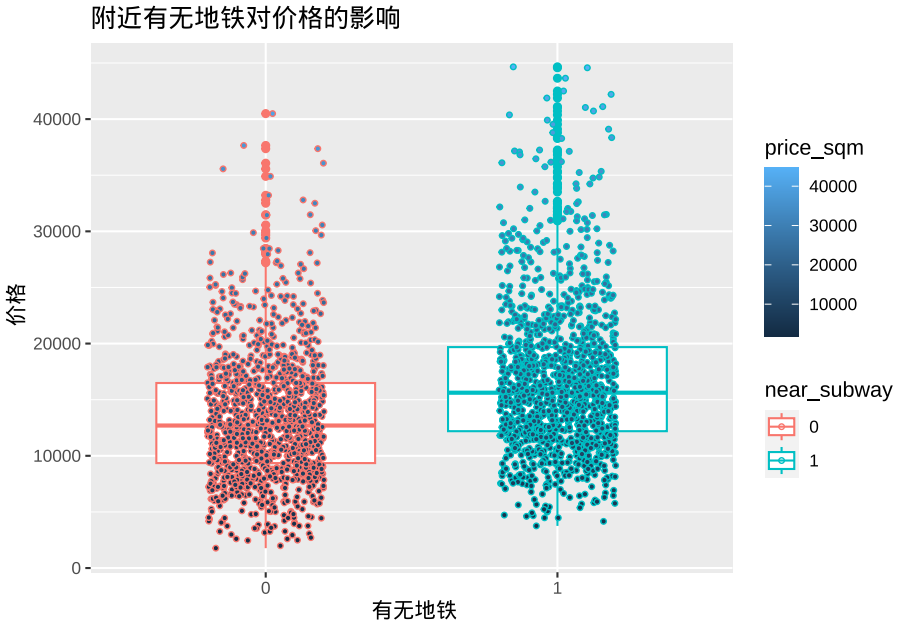
表二按照建筑类型和房屋单价绘制箱型图和散点图。可以看出四种类型的房屋数量排列如下：板楼 > 板塔结合 > 塔楼 > 平房。通过箱型图矩形框的长度，可以看出来” 板楼”,“板塔结合” 和 “塔楼” 的长度差不多，代表三个建筑类型的离散程度接近。

3.5 探索问题 2：附近有无地铁对价格的影响

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1771    9351   12698   13411   16486   40492

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    3755   12194   15623   16680   19685   44656
```

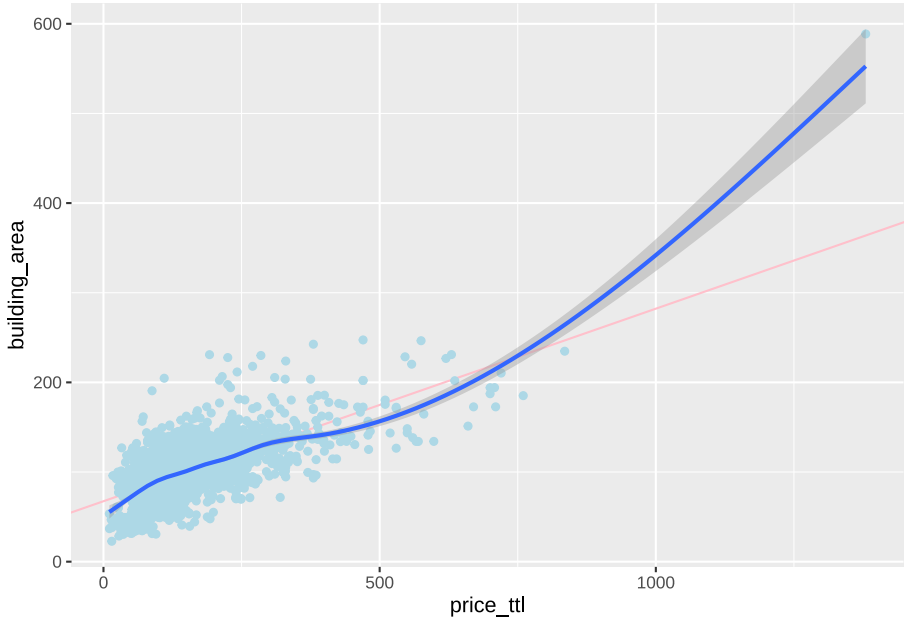


发现：

首先把数据库中除了 0 和 1 之外的无关数据去除，得到了关于有无地铁和价格的箱型图。得到了附近有地铁的房子整体单价都比附近没有地铁的单价高的结论。

3.6 探索问题 3：房屋总价和房屋面积的关系

```
## [1] 67.4238502  0.2147854
```



发现：

- 发现 1：通过牛顿-拉夫逊搜索优化得到房屋总价和房屋面积的一元二次方程式为 y=67.4238502+0.2147854x。
- 发现 2：房屋总价大部分集中在 0-500 万元中间，房屋面积大部分集中在 0-200 平方米中间。在此区间范围内，一元线性回归模型和平滑曲线很接近, 随着房屋总价的提升，房屋面积也在不断变大。

---

4 发现总结

通过数据的清洗，得到 2515 行武汉链家二手房源数据，首先对 18 个变量进行描述性统计量分析；其次选取单价、朝向和建筑形式进行单变量分析；最后探索房屋单价和建筑类型的关系、附近有无地铁对价格的影响和房屋总价与房屋面积的关系。得到了以下发现和总结：

- 1、房屋单价的众数为 10196 元，平均值为 15110.42 元，中位数为 14309

元，且且价格范围在 10765 到 18212.5 的房子数量最多。

- 2、武汉的链家二手房的朝向分为” 北”,“东”,“东北”,“东南”,“南”,“西”,“西北”,“西南” 八种。其中” 南” 朝向的房子数量最多。“东北” 和” 西北” 朝向的房子数量最小。
- 3、武汉的链家二手房的建筑形式分为” 板楼”,“板塔结合”,“平房”,“塔楼” 四种。其中” 板楼” 形式的房子数量最多。“平房” 形式的房子数量最小。
- 4、板塔结合这类建筑类型的房子整体单价偏高。平房这类建筑类型的房整体单价偏低。
- 5、附近有地铁的房子整体单价都比附近没有地铁的单价高。
- 6、房屋总价大部分集中在 0-500 万元中间，房屋面积大部分集中在 0-200 平方米中间。在此区间范围内，随着房屋总价的提升，房屋面积也在不断变大。