第一次作业：基于武汉链家的数据分析报告

吴淇

2023-10-19

目录

1 主要发现

发现 1：

在装修的不同等级中，精装修对房屋单价的影响最为显著。而与简装的平均单价相比，精装的平均单价增加了 2084.4 元/平米，毛坯、简装和其他情况之间房屋平均单价差异不大。

发现 2：

房屋建筑面积和房屋单价存在正相关关系。根据线性回归模型，每增加一个单位的建筑面积，房屋单价预期会上涨 33.272 元。

发现 3：

房屋的楼栋高度与房屋单价存在正相关关系。并且在预测房屋单价方面，可能是一个比建筑面积更显著的因素。

发现 4：

装修情况同样是影响房屋关注度的最显著因素之一，装修等级越高，关注人数也越多。

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。

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
##
##
##
```

可以得出直观结论：

（一）房屋总价 (price_ttl)

1. 最低价为 10.6 万元，最高价为 1380 万元，中位数为 137 万元。

2. 平均价为 155.9 万元。

（二）房屋单价 (price_sqm)

1. 最低单价为 1771 元/平方米，最高单价为 44656 元/平方米。

2. 大部分房子的单价集中在 10799 元/平方米到 18211 元/平方米之间，其中中位数为 14404 元/平方米。

3. 平均单价为 15148 元/平方米。

（三）房间数量 (bedrooms)

1. 大部分房子有 2 至 3 个卧室。

2. 中位数为 3，意味着大多数房屋都是三室的。

（四）客厅数量 (livingrooms)

1. 大部分房子有 1 至 2 个客厅。

2. 中位数为 2，表示大多数房屋都是二厅的。

（五）建筑面积 (building_area)

1. 房子的建筑面积范围从 22.77 平方米到 588.66 平方米。

2. 中位数面积为 95.55 平方米，表示大部分房子的面积接近这个数值。

（六）楼栋总层数 (property_t_height)

1. 最小的楼栋只有 2 层，而最高的有 62 层。

2. 大部分楼栋的楼层数集中在 11 层到 33 层之间。

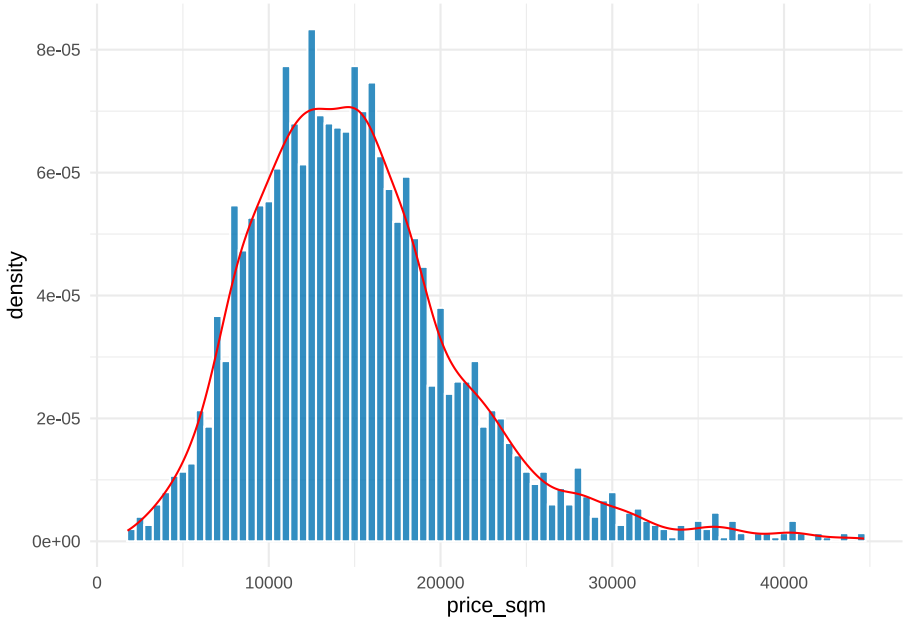
4 探索性分析

4.1 房屋单价的数值描述与图形

房屋单价的数值描述：

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

房屋单价的分布情况：

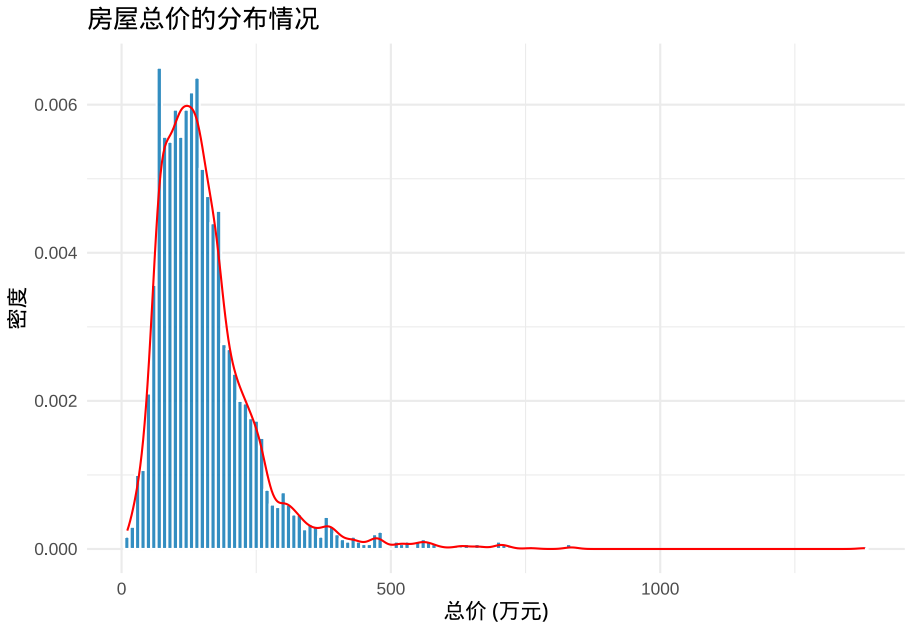


发现：房屋单价呈现为右偏正态分布的情形, 异常值似乎对均值影响不大。

4.2 房屋总价的数值描述与图形

房屋总价的描述性分析：

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.77   84.92   95.55  100.87  117.68  588.66
```



房屋总价的分布情况

- 发现：

  1. 房价的分布是右偏的，且较房屋单价更为集中。有一些高价的房子拉高了平均价。

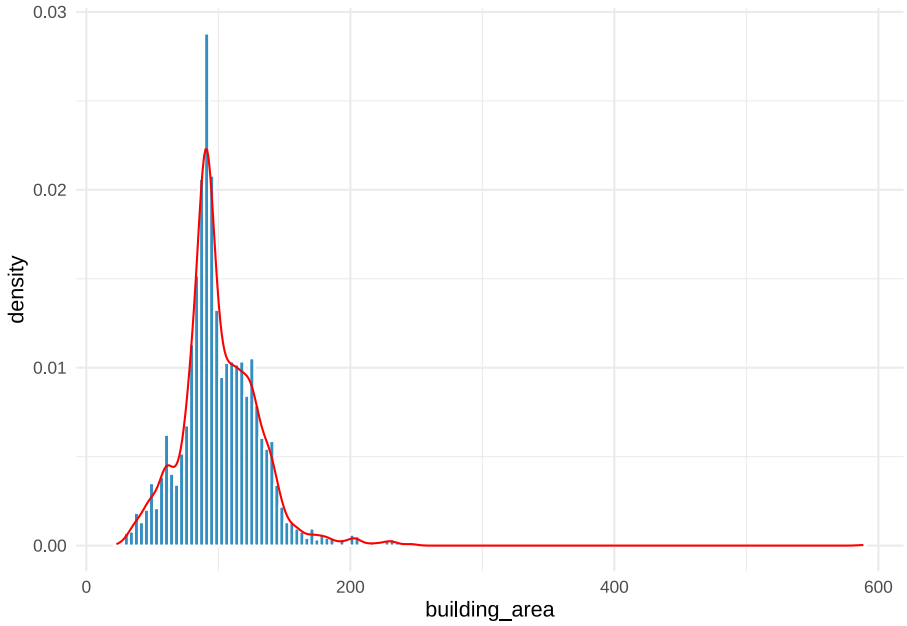
  2. 大部分房子的价格集中在在 95 万元到 188 万元之间。

4.3 房屋建筑面积的数值描述与图形

房屋建筑面积的描述性统计

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.77   84.92   95.55  100.87  117.68  588.66
```

房屋建筑面积的分布情况：



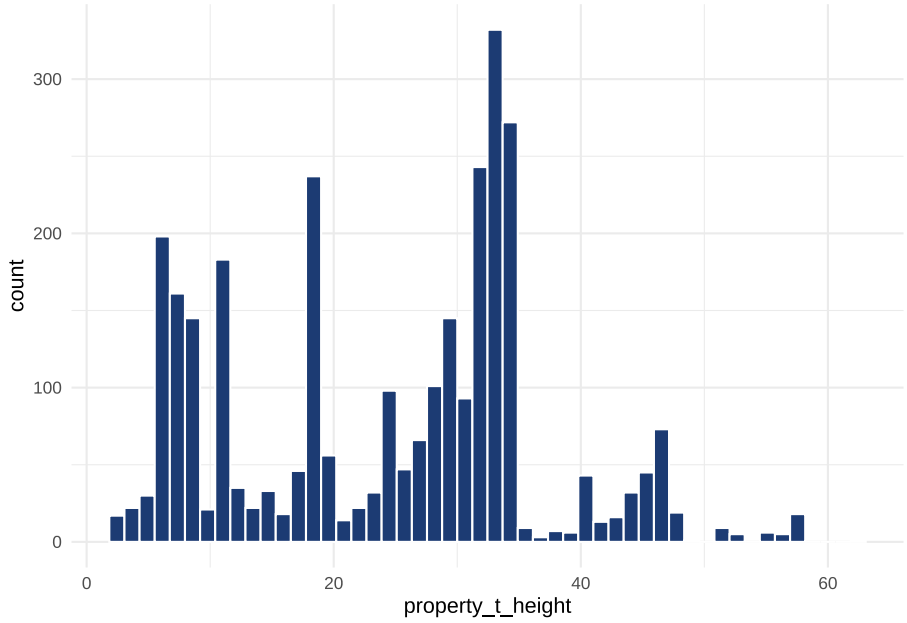
发现：

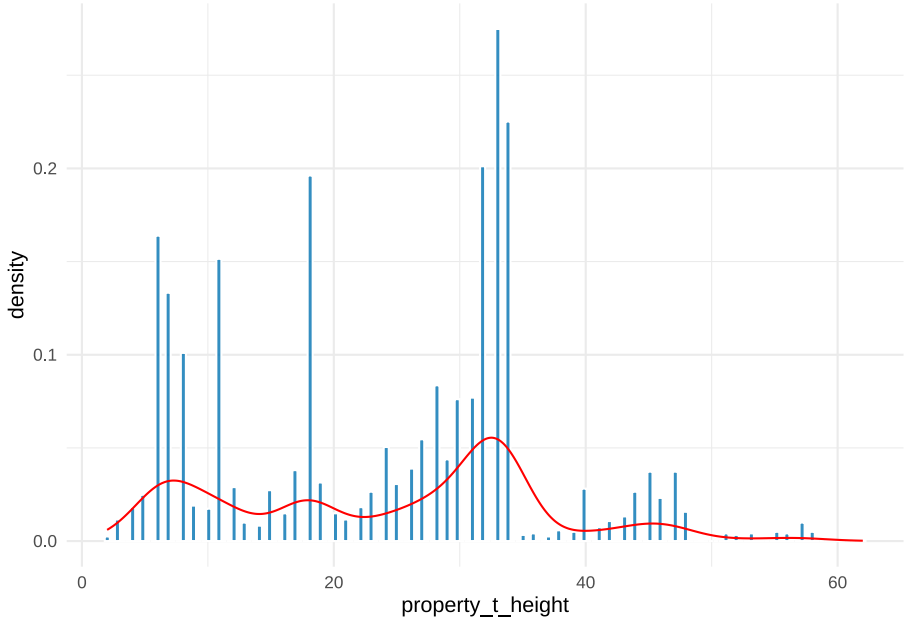
1. 房屋建筑面积大部分集中在 84.92 平方米到 117.68 平方米之间。

2. 相较于房屋单价和总价，房屋建筑面积右偏更显著。

4.4 楼栋总层数的数值描述与图形

楼栋总层数的计数和分布：





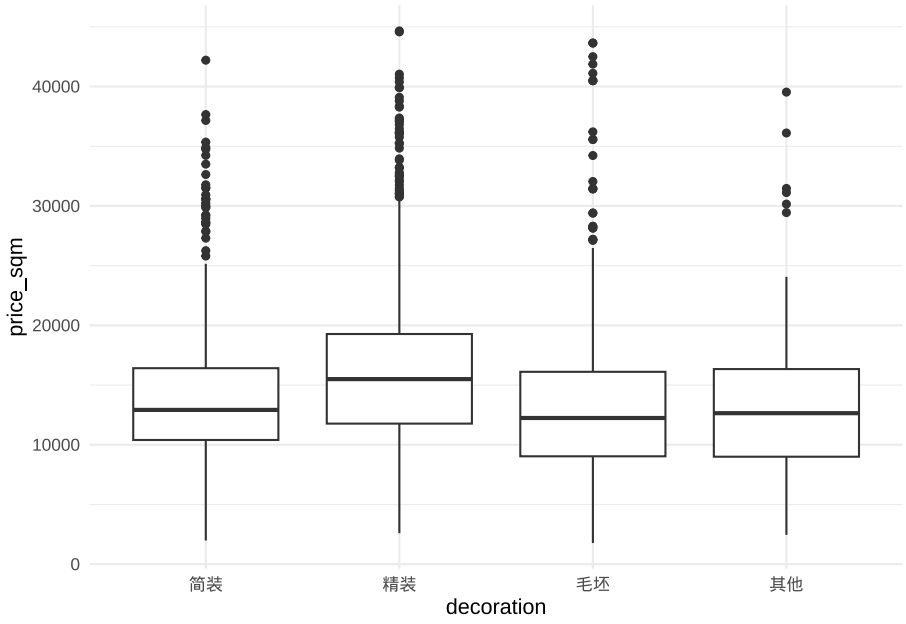
发现：

1. 大部分楼层集中在 11-33 层之间，并且在几个特殊楼层数时概率密度显著增高。

2. 楼层数呈现出左偏的情形，这和前几个右偏的数据存在明显区别。

3. 究其原因或许是因为建筑水平和政策方面的考量对高于某一特定层数的建筑有不同的限制和要求。

4.5 探索问题 1：装修对房屋单价有何影响？



通过 lm() 函数创建线性模型，查看装修 decoration 是如何影响单价 price_sqm 的：

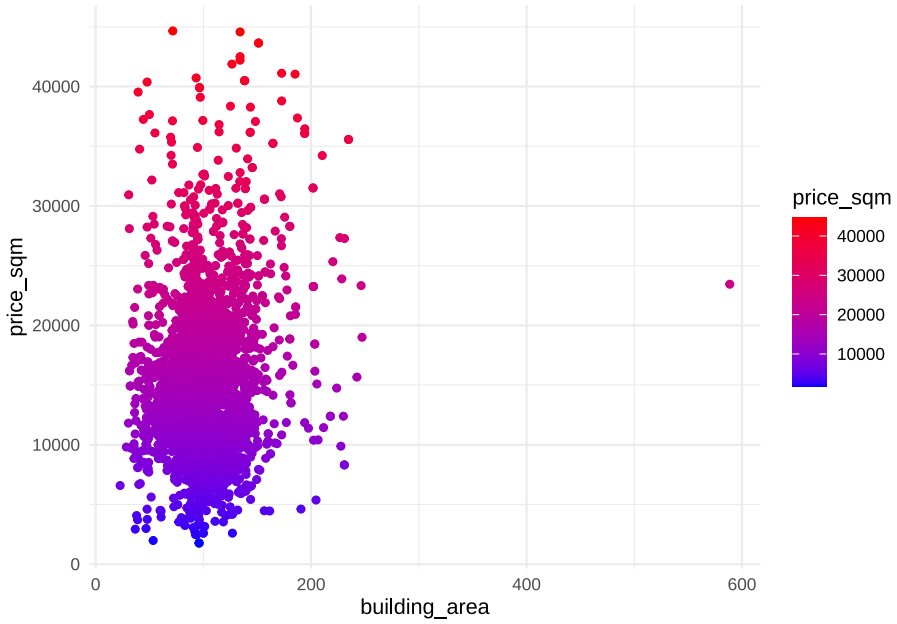
```
##
## Call:
## lm(formula = price_sqm ~ decoration, data = lj)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13478.1  -4152.3   -818.7   2919.3  29824.1
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     13992.7      247.3  56.577  < 2e-16 ***
## decoration精装   2084.4      288.5   7.224 6.35e-13 ***
## decoration毛坯   -173.8      387.4  -0.448    0.654
## decoration其他   -688.6      534.2  -1.289    0.197
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6227 on 2996 degrees of freedom
## Multiple R-squared:  0.03103,    Adjusted R-squared:  0.03006
## F-statistic: 31.98 on 3 and 2996 DF,  p-value: < 2.2e-16
```

- 发现：

与简装的平均单价 (Intercept) 相比，精装的平均单价增加了 2084.4 元/平米。这一结果是显著的（Pr(>|t|) 列，其值远小于 0.05）。而其他装修类型与简装相比在价格上的差异不显著。但需要注意模型只解释了房屋单价 price_sqm 的 3.103% 的变异，这意味着还有其他很多因素影响 price_sqm。

4.6 探索问题 2：房屋建筑面积和单价之间存在怎样的关系？



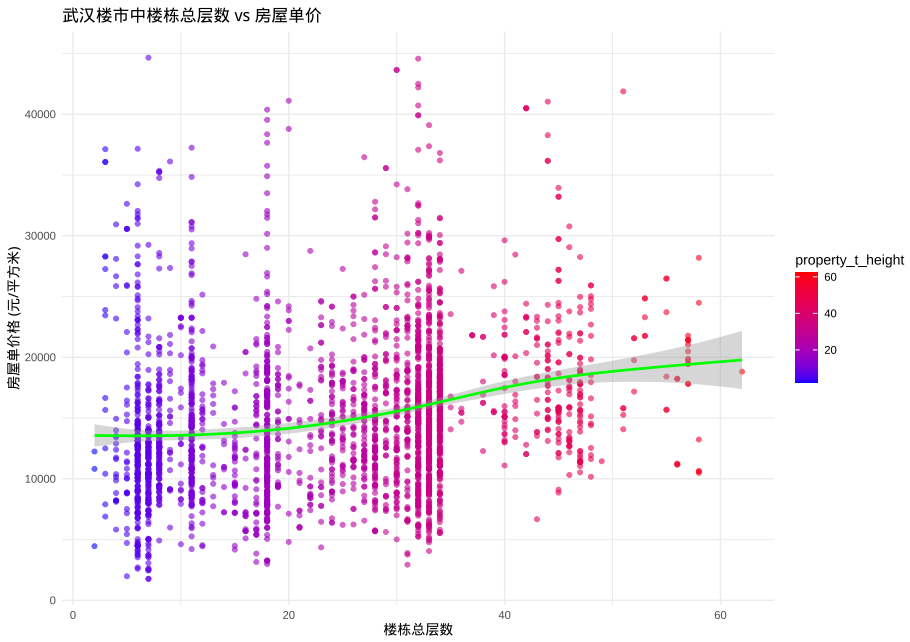

```
##
## Call:
## lm(formula = building_area ~ price_sqm, data = lj)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.28 -16.53  -4.19  17.61 481.42
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.923e+01  1.422e+00  62.758   <2e-16 ***
## price_sqm   7.680e-04  8.662e-05   8.867   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.99 on 2998 degrees of freedom
```

```
## Multiple R-squared:  0.02555,    Adjusted R-squared:  0.02523
## F-statistic: 78.62 on 1 and 2998 DF,  p-value: < 2.2e-16
```

-发现：

模型结果显示，price_sqm 与 building_area 有正向关系，并且这种关系是统计显著的。但 Multiple R-squared 值为 0.02555，这意味着线性模型仅解释了 2.555% 的 building_area 的变化。这意味着还有其他未考虑的因素也可能影响 building_area。

4.7 探索问题 3：楼栋总层数和房屋单价之间是否存在关联？



通过一个简单的线性回归模型，基于 property_t_height（楼栋总层数）来预测 price_sqm（每平方米的房屋单价）。

```
##
## Call:
## lm(formula = price_sqm ~ property_t_height, data = lj)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13004.4  -4114.3   -809.8   2811.7  31512.8
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       12327.92     245.88   50.14   <2e-16 ***
## property_t_height   116.47       9.03   12.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 6156 on 2998 degrees of freedom
## Multiple R-squared:  0.05257,    Adjusted R-squared:  0.05225
## F-statistic: 166.4 on 1 and 2998 DF,  p-value: < 2.2e-16
```

- 发现：

从这个模型可以看出，房屋的楼栋高度与每平方米的价格有正相关关系。模型解释了数据中的 5.257% 变异，比之前的模型（基于 building_area）解释的变异要高。这意味着物业的高度可能是一个比建筑面积更重要的因素来预测每平方米的价格。但 5.257% 仍然只是一个很小的比例，还存在着其它诸多影响因素。

4.8 探索问题 4：影响房屋受欢迎程度最重要的因素是什么？

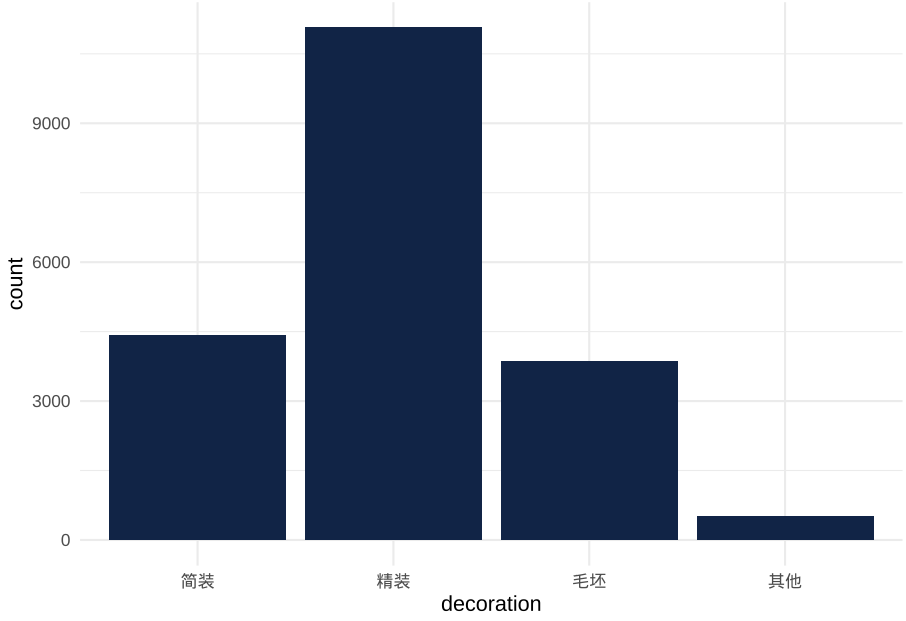
对于二手房而言，房型的受欢迎程度将在很大程度上影响房屋的保值率和流通性，因此具有一定研究意义和选购时的参考价值。所以我们以关注度为受欢迎程度的衡量标准，基于链家的 3000 条数据，分析其相关性，并筛选出其中相关度最靠前的几个指标。

```
##
## Call:
## lm(formula = followers ~ ., data = lj_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.702  -5.992  -3.493   0.522 252.510
##
## Coefficients:
##                           Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)              2.106e+01           NA  1.155e+01   1.823  0.06846 .
## price_sqm               -4.485e-04   -1.815e-01  1.944e-04  -2.307  0.02116 *
## price_ttl                5.361e-02    3.286e-01  1.728e-02   3.103  0.00195 **
## building_area           -7.040e-02   -1.330e-01  3.574e-02  -1.970  0.04902 *
## property_t_height       -3.637e-03   -2.895e-03  3.462e-02  -0.105  0.91635
## livingrooms              1.397e+00    4.237e-02  9.708e-01   1.439  0.15037
## decoration精装           3.534e-01    1.079e-02  1.025e+00   0.345  0.73023
## decoration毛坯           5.196e+00    9.719e-02  1.586e+00   3.276  0.00107 **
## decoration其他          -3.125e+00   -4.399e-02  1.894e+00  -1.650  0.09911 .
## property_style板塔结合 -9.446e-01   -2.317e-02  1.086e+00  -0.870  0.38445
## property_style平房       5.926e+01    1.628e-01  9.022e+00   6.569 6.93e-11 ***
## property_style塔楼       2.742e+00    7.165e-02  1.003e+00   2.733  0.00635 **
## property_style暂无数据 -3.961e+00   -4.202e-02  2.369e+00  -1.672  0.09472 .
## near_subway近地看       -1.071e+01   -1.700e-02  1.911e+01  -0.560  0.57527
## near_subway近地铁       -1.149e+01   -4.071e-02  1.100e+01  -1.044  0.29682
## near_subway珞狮南       -2.322e-01   -3.686e-04  1.903e+01  -0.012  0.99027
## near_subway太子湖1号    -1.630e+01   -2.588e-02  1.904e+01  -0.856  0.39186
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.53 on 1542 degrees of freedom
##   (因为不存在，1441个观察量被删除了)
## Multiple R-squared:  0.0624, Adjusted R-squared:  0.05267
## F-statistic: 6.414 on 16 and 1542 DF,  p-value: 3.777e-14
```

从表中我们可以发现:影响关注人数的最显著的三个变量依次为:decoration、price_ttl 房屋总价、建筑形式 property_style。

接下来对装修和房屋关注人数之间的关系做进一步分析：



发现：装修等级越高，受关注程度也越高，其中精装修的房屋关注人数远高于毛坯或简装。

5 发现总结

1. 在装修的不同等级中，精装修对房屋单价的影响最为显著。而与简装的平均单价相比，精装的平均单价增加了 2084.4 元/平米，毛坯、简装和其他情况之间房屋平均单价差异不大。

2. 房屋建筑面积和房屋单价存在正相关关系。根据线性回归模型，每增加一个单位的建筑面积，房屋单价预期会上涨 33.272 元。

3. 房屋的楼栋高度与房屋单价存在正相关关系。并且在预测房屋单价方面，可能是一个比建筑面积更显著的因素。

4. 装修情况同样是影响房屋关注度的最显著因素之一，装修等级越高，关注人数也越多。