商务统计分析 1st_assignment_shurui

shurui

2023-10-17

```
knitr::opts_chunk$set(echo = FALSE,error = FALSE, warning = FALSE, message = FALSE,
                      out.width = "100%", split = FALSE, fig.align = "center")
pdf.options(family="GB1")
#load library
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
library(lubridate)
library(scales)
```

```
##
## 载入程辑包: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```r
library(plotly)
```

```
##
## 载入程辑包: 'plotly'
##
## The following object is masked from 'package:ggplot2':
##
##     last_plot
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following object is masked from 'package:graphics':
##
##     layout
```

```r
library(patchwork)
library(ggrepel)
library(wordcloud2)
```

数据介绍

- 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 3000 行。

共有 18 个变量解释如下：| 变量 | 解释 | |:–|:–| |property__name| 小区名字 | |property__region| 所处区域 | |price__ttl| 房屋总价，单位万元 | |price__sqm| 房屋单价，单位元 | |bedrooms| 房间数 | |livingrooms| 客厅数 | |building__area| 建筑面积 | |directions1| 房屋主要朝向 | |directions2| 房屋次要朝向 | |decoration| 装修状况 | |property__t__height| 楼栋总层数 | |property__height| 房屋在所在楼栋所处位置，取值为高中低 | |property__style| 建筑形式，如板楼、塔楼等 | |followers| 在该二手房网站的关注人数 | |near__subway| 是

否靠近地铁 | |if_2y| 产证是否满 2 年 | |has_key| 中介是否有钥匙，标注 “随时看房” 表示有钥匙 | |vr| 是否支持 VR 看房 |

该表共有 3000 行数据，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
##
##
##
```

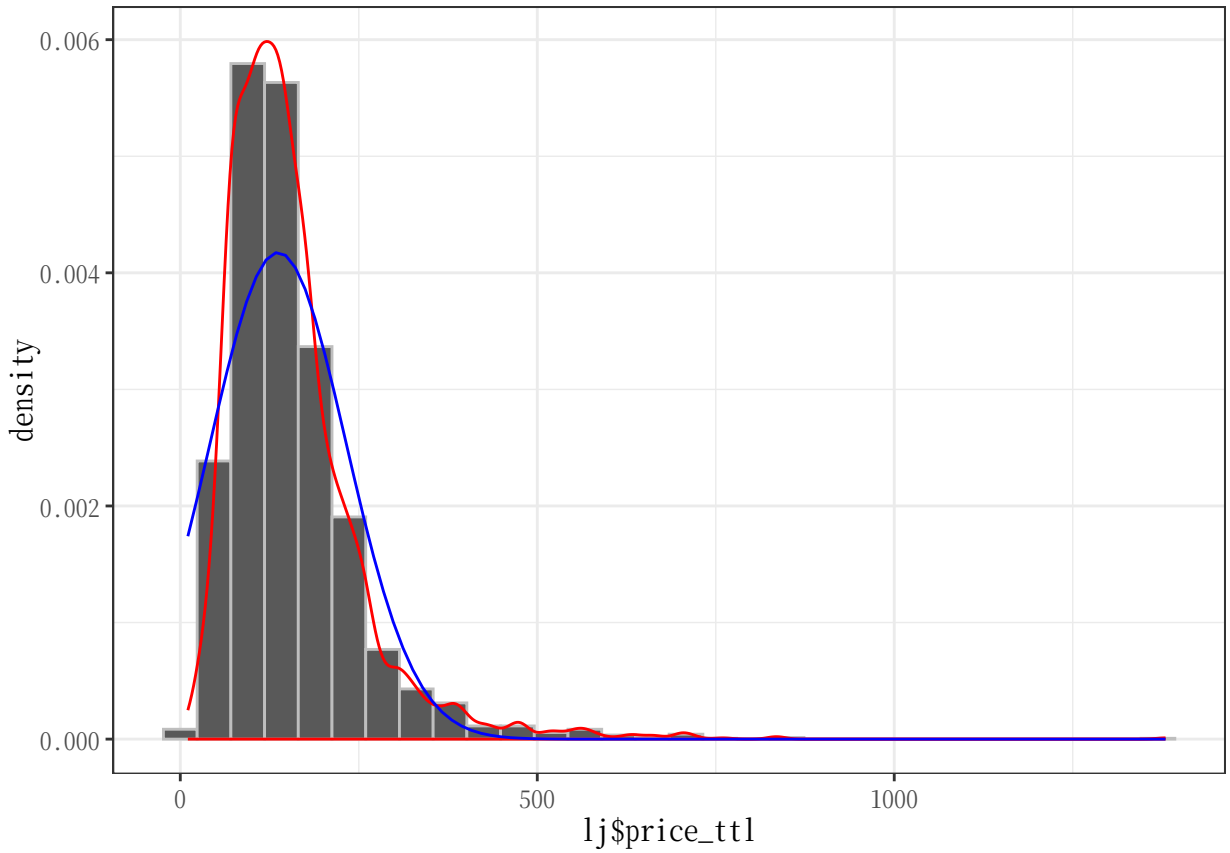
可以看到：

- 直观结论 1
- 价格特点：price_ttl 房屋总价最大值 1380 万元，最小值 10.6 万元，中位数值 137 万元，均值 155.9 万元。均值与最大值最小值有一定差距，数据分布可能比较分散，数据集中程度不高。也证明了各个区域下房价的价值分布有较大的差异性。
- 直观结论 2
- 部分数据存在异常值需要清洗，如 property_region 未填写正确数值，部分数据填充值为 NA，并且该 NA 具有业务属性，即非 NA 则为统一值，可进行转换而不做清除。
- 直观结论 3
- 数值类型数据 7 个，字符类型数据 11 个，字符类型数据需要进一步处理分析。

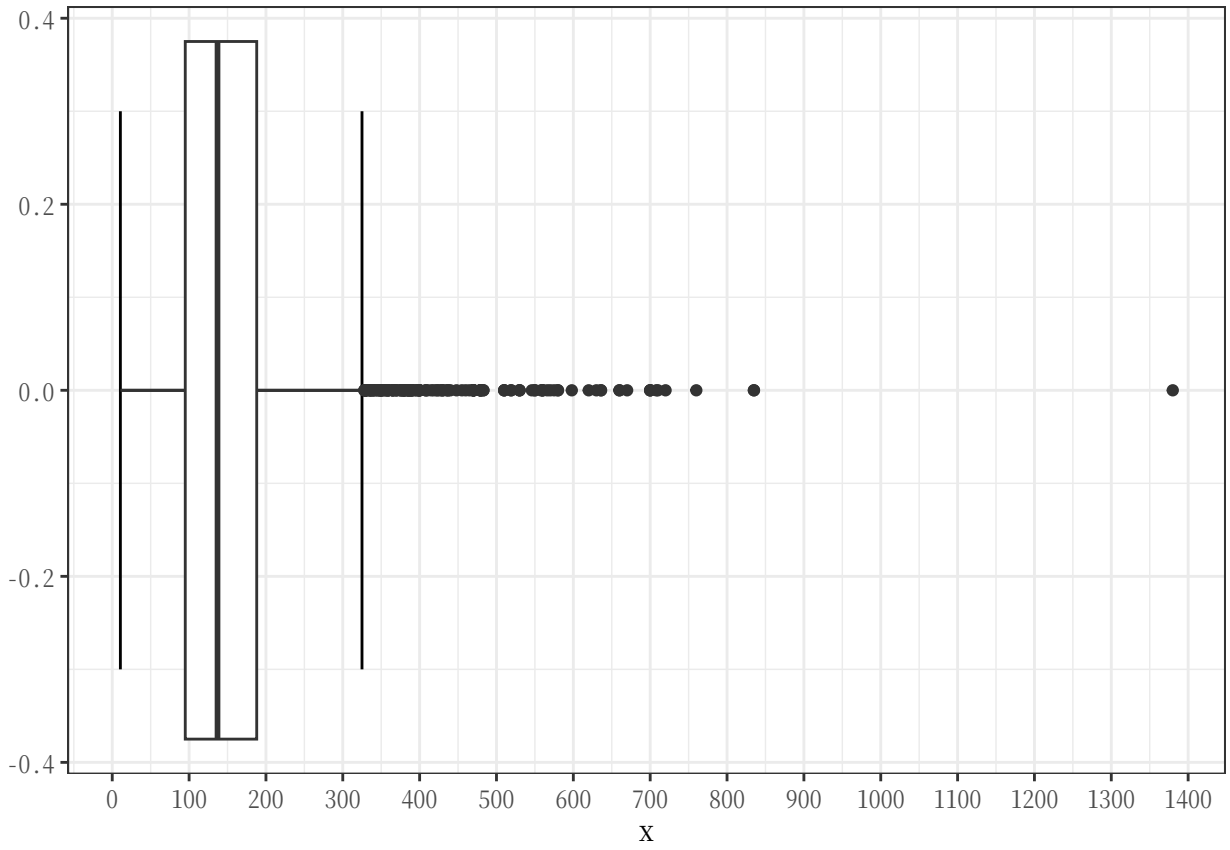
探索性分析

变量 price_ttl 的数值描述与图形

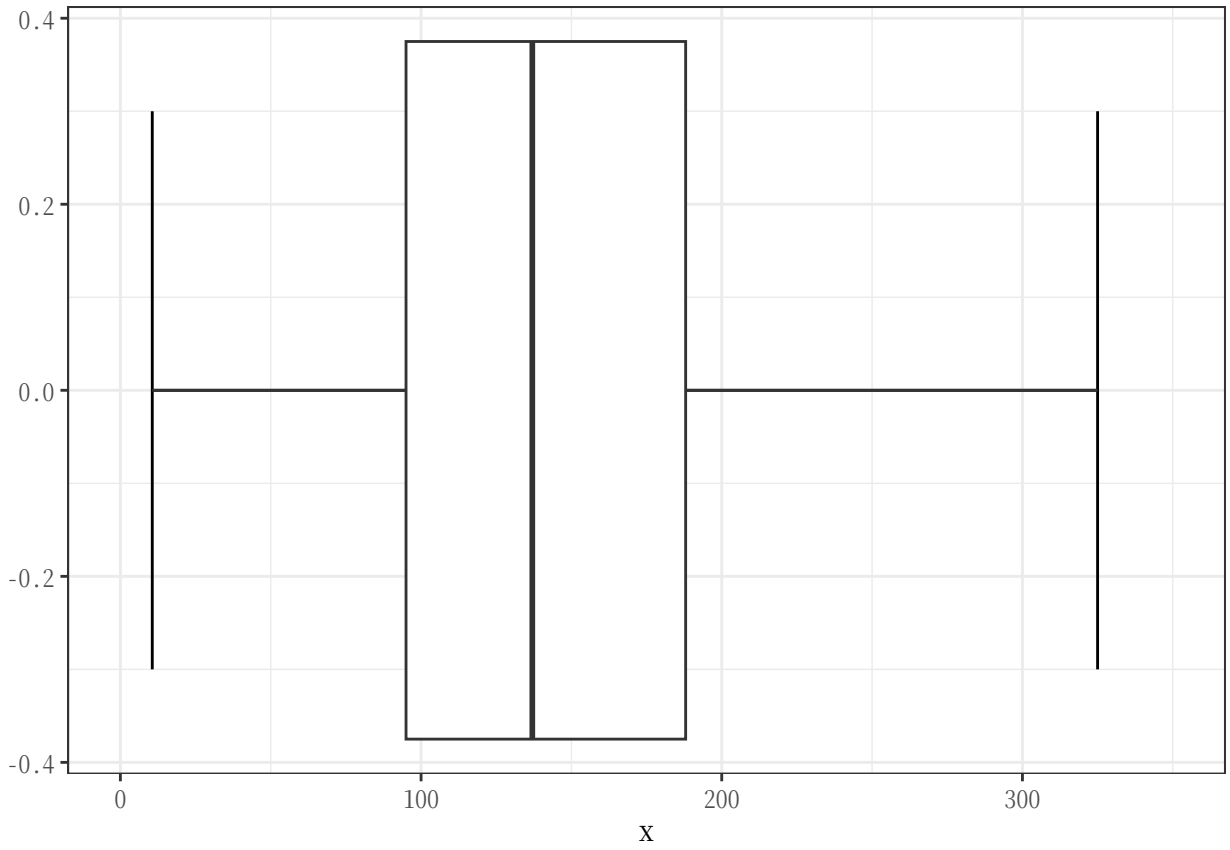
- 发现：
- 发现 1
- price_ttl 变量数值描述类型 numeric：numeric max：1380 min：10.6 mean：137 avg：155.9 四分位距：93 方差 95.5481281:95.54813 极差 1369.4:：1369.4 是否有空值 0：无
- 发现 2
- price_ttl 变量图形描述：直方图描述与概率密度曲线将 price_ttl 的数据用直方图展示结果类似卡方分布，红色线条为该数据的概率密度曲线，蓝色线条为该数据在正态分布下的概率密度曲线，能看出房屋的总价有点趋向于正态分布



price_ttl 变量图形描述：箱线图箱线图展示了变量的 4 分位以及上下分界，价格区间从十几万到 350 万间有 99.7% 的数据，有不少的离群点数据在分析时可作为异常值处理。



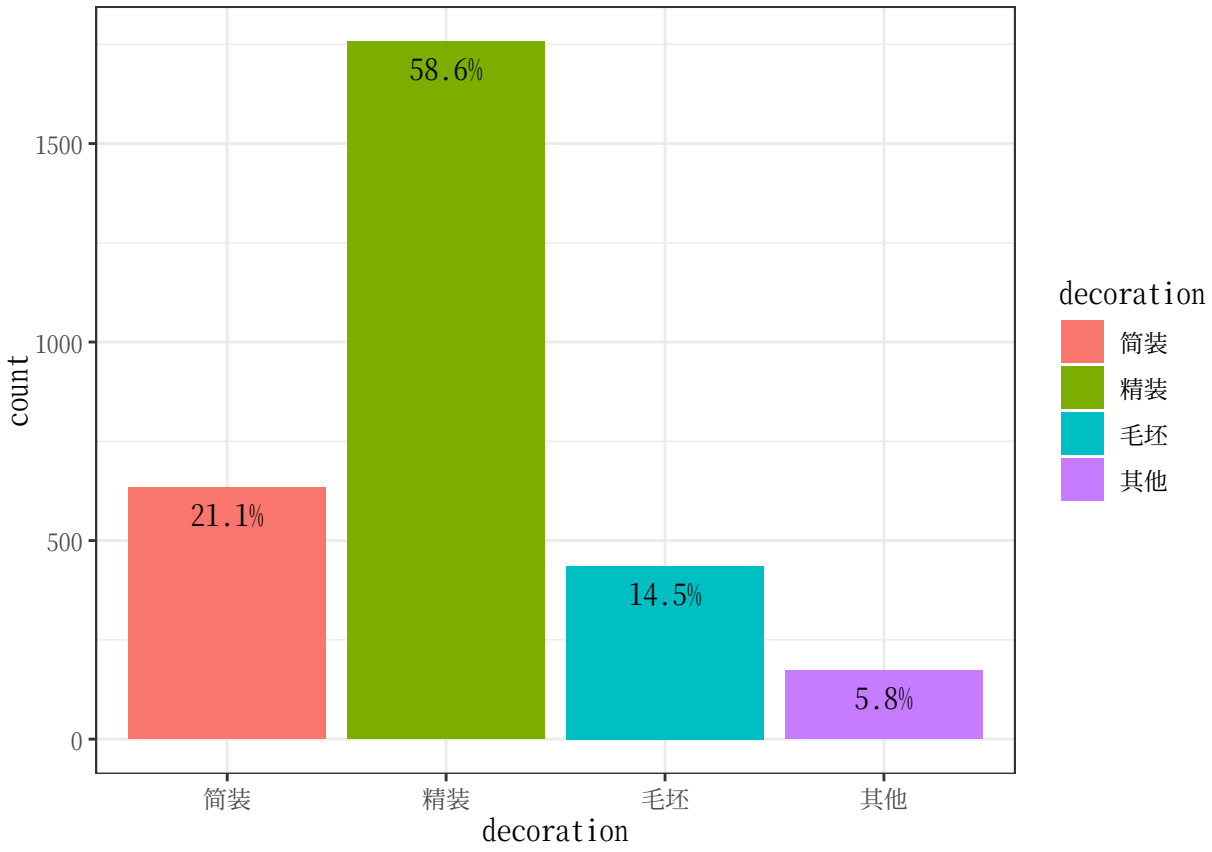
该箱线图剔除了异常值并缩小了 x 轴范围，比较直观的看到数据中整体房价分布区间在 90 多万至 190 万左右，包含了 50% 的数据，箱体比较扁，该部分数据较为集中，证明该部分数据展示的价格区间非常集中。



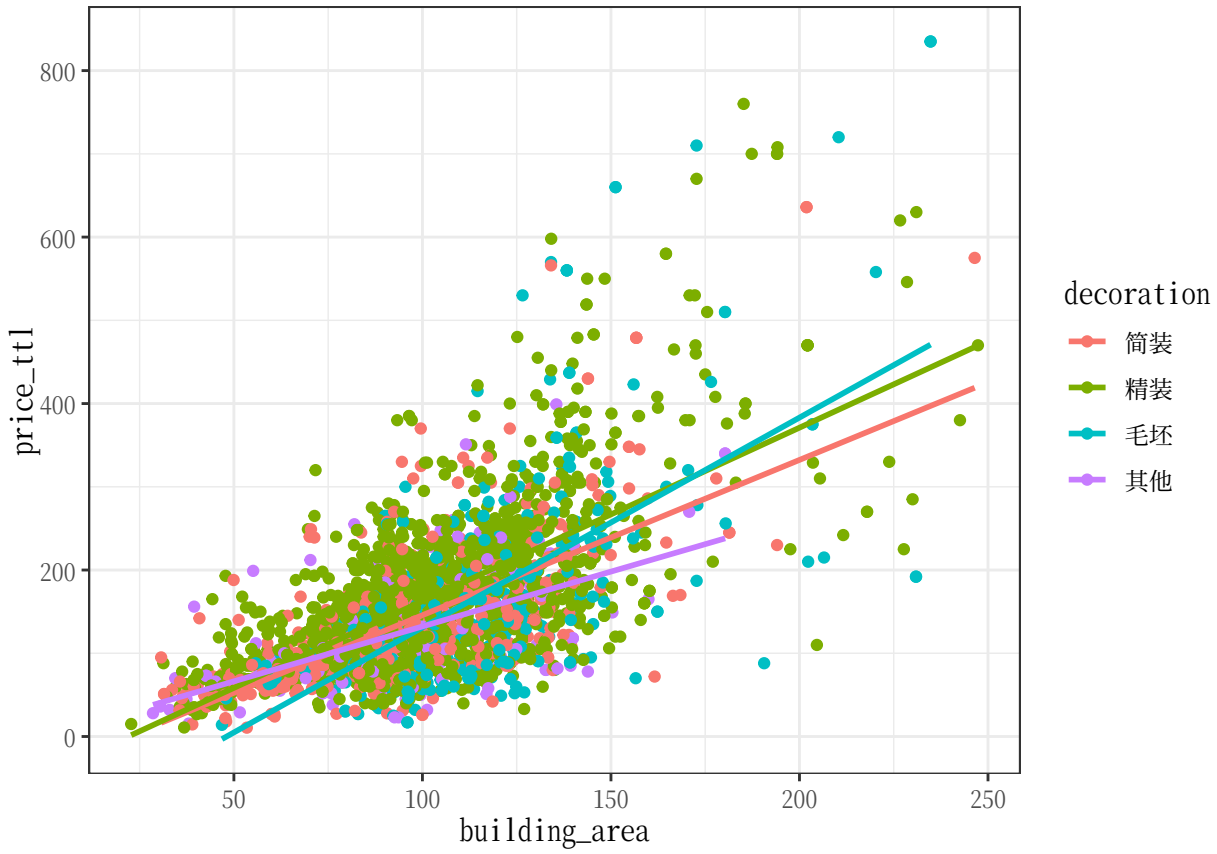
变量 decoration 的数值描述与图形

发现：- 发现 1

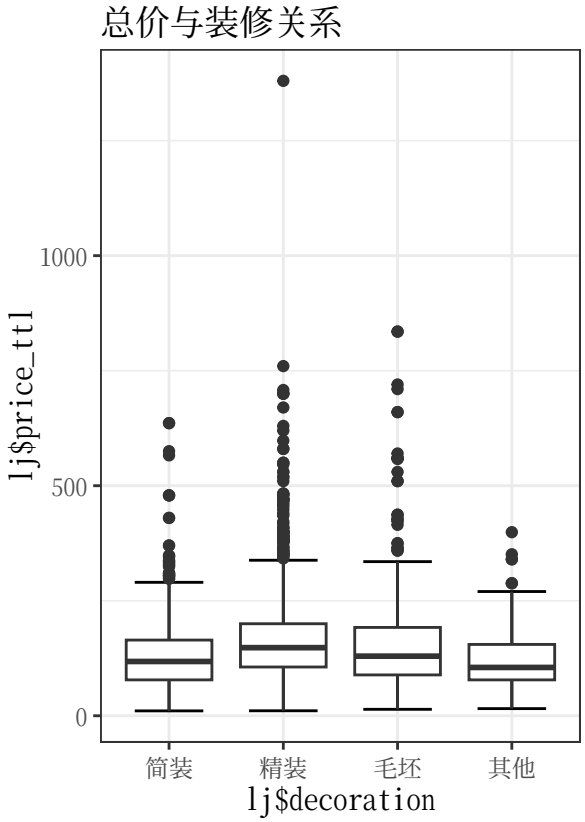
- decoration 变量数值描述类型 character：character 是否有空值 0：无数据类型一共有精装, 简装, 其他, 毛坯：“精装” “简装” “其他” “毛坯” 这 4 种
- 发现 2
- decoration 变量图形描述：直方图通过条形图可以看到精装是占比最高的 58.6%，其次是简装 21.1%，最后是毛坯 14.5%, 其他占比 5.8%

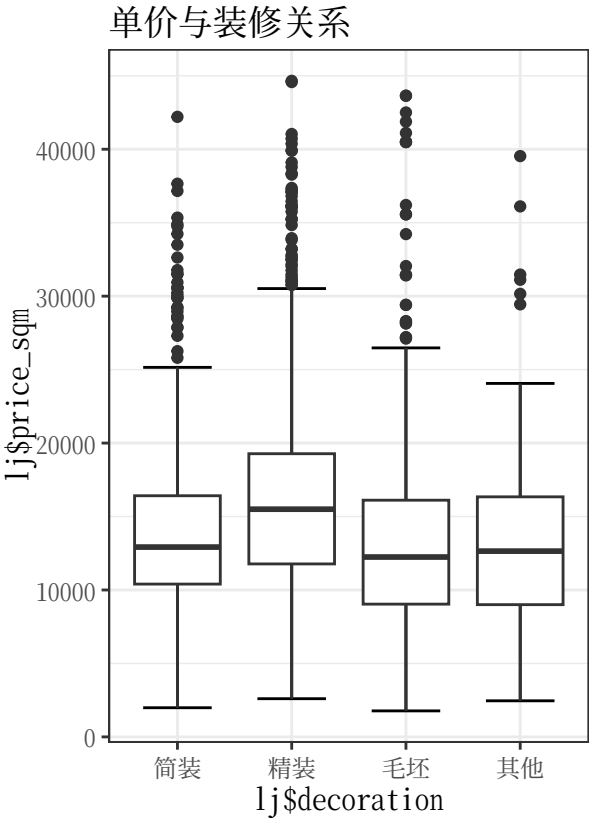


通过单价与面积的散点图，我们能看到单价与面积成线性增长关系，即面积越大单价越高，并且精装修的房屋线性关系是三类房屋中最平稳的



我们可以看到 4 种装修类别的房屋的总价和单价展示出的箱线图，可以看出精装修确实会让房屋的出售价格高一些，符合社会规律。

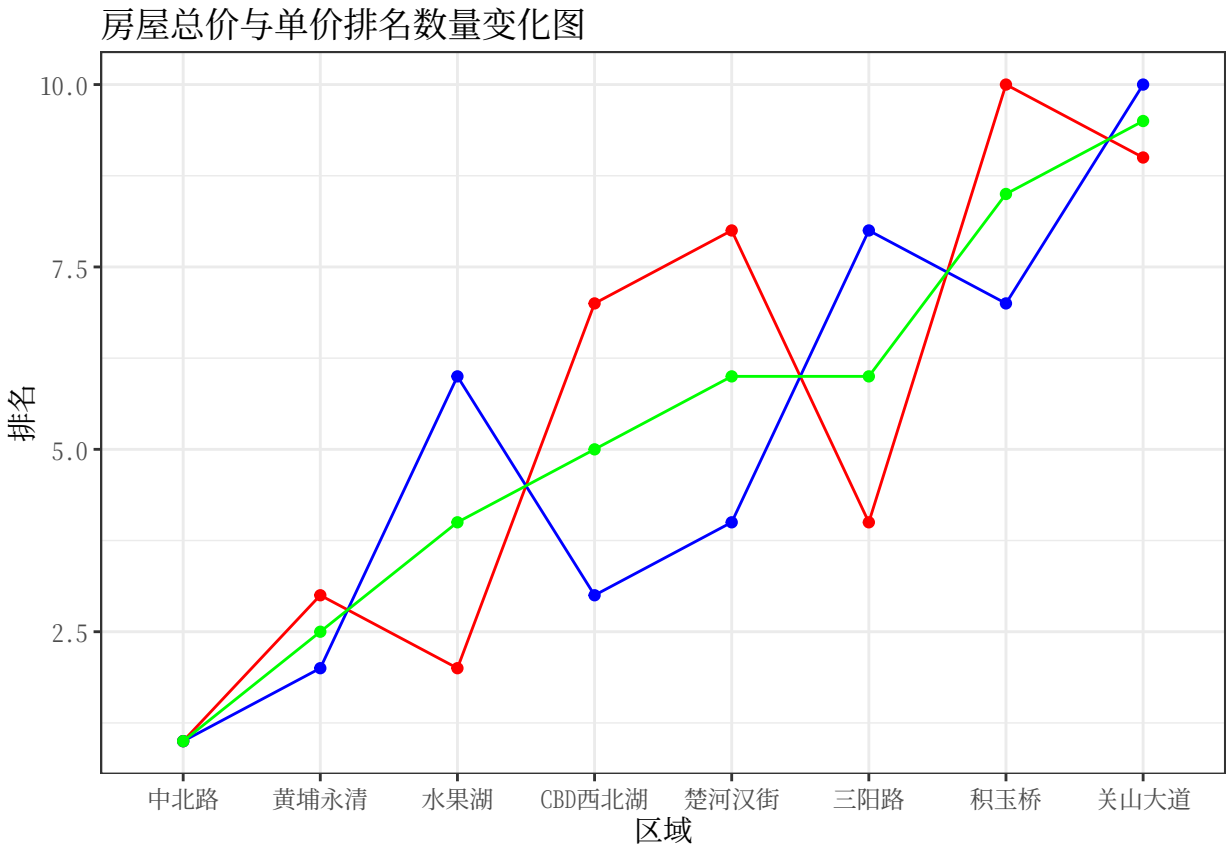




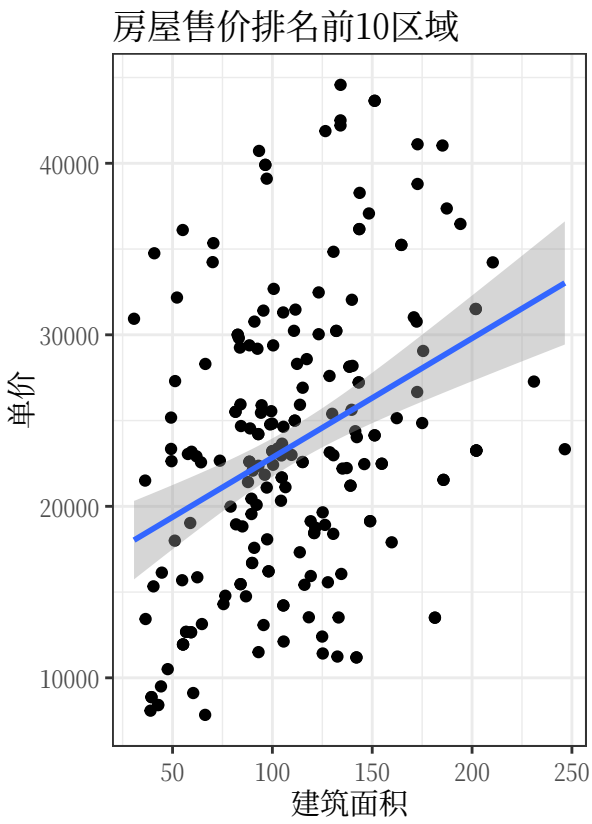
探索问题 1

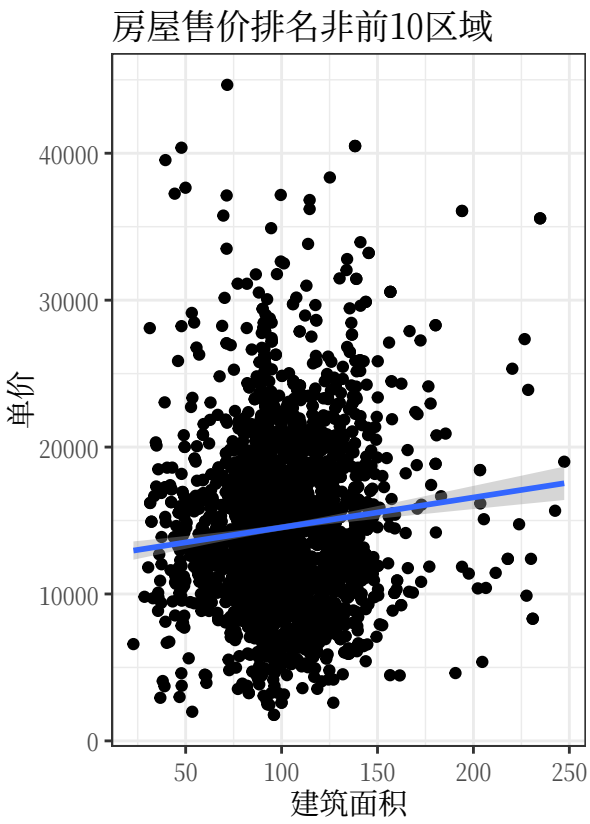
- 发现：探索房屋价格与所属区域的关系
- 发现 1

我们能找到房屋单价、总价排名靠前的房屋区域，我们将排名前 10 的区域找到，并将单价前 10 的数据与总价前 10 的数据做交集，得到有 8 个区域是既是单价排名靠前同时总价排名也是靠前，发现数据中单价与总价都会同时影响一个区域的表现关系，虽然排名有一定的变化（红色和蓝色线条），但是如绿色线条展示他们有一定的线性关系。



- 发现 2
- 对比前 10 区域的房屋与非前 10 区域房屋在单价与面积的关系，可以看出排名前 10 区域的房屋单价与面积的斜率更大, 而非前 10 区域房屋单价与面积关系线性关系比较平滑，说明前 10 区域内的房屋单价/面积的比值更大。

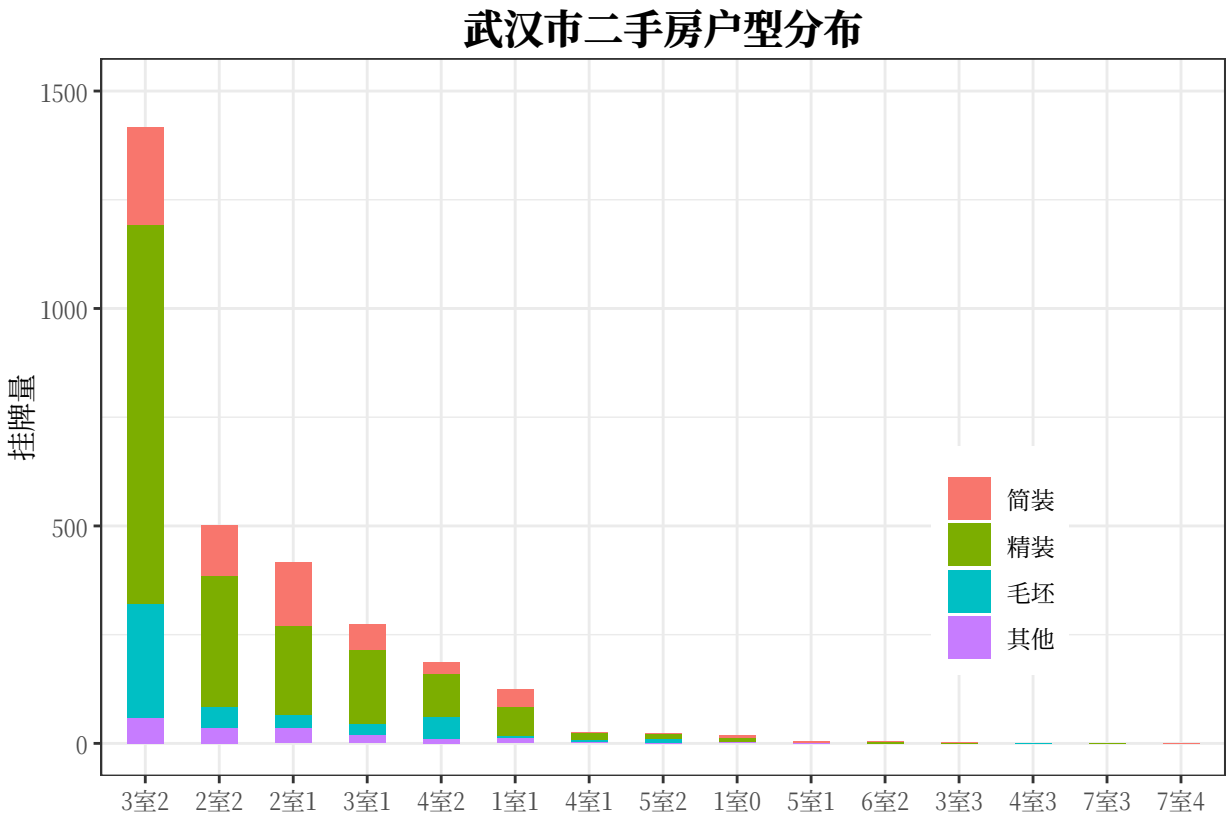




探索问题 2

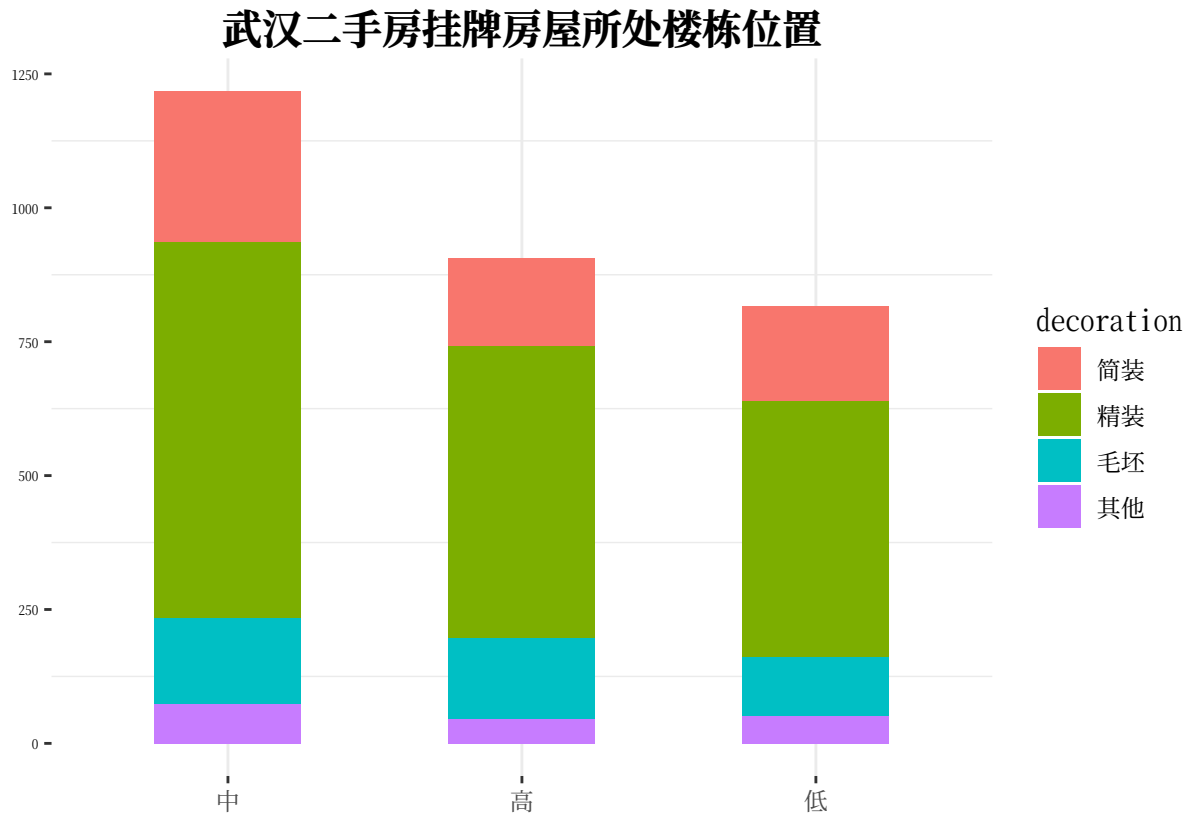
探索房屋的装修情况 decoration，房屋的户型 bedrooms、livingrooms 房屋在所在楼栋所处位置 property_height，等与挂牌数量的关系发现：- 发现 1

- 二手房户型分布与装修情况与挂牌数量的关系，3 室两厅且为精装是挂牌最多的类型



-发现 2

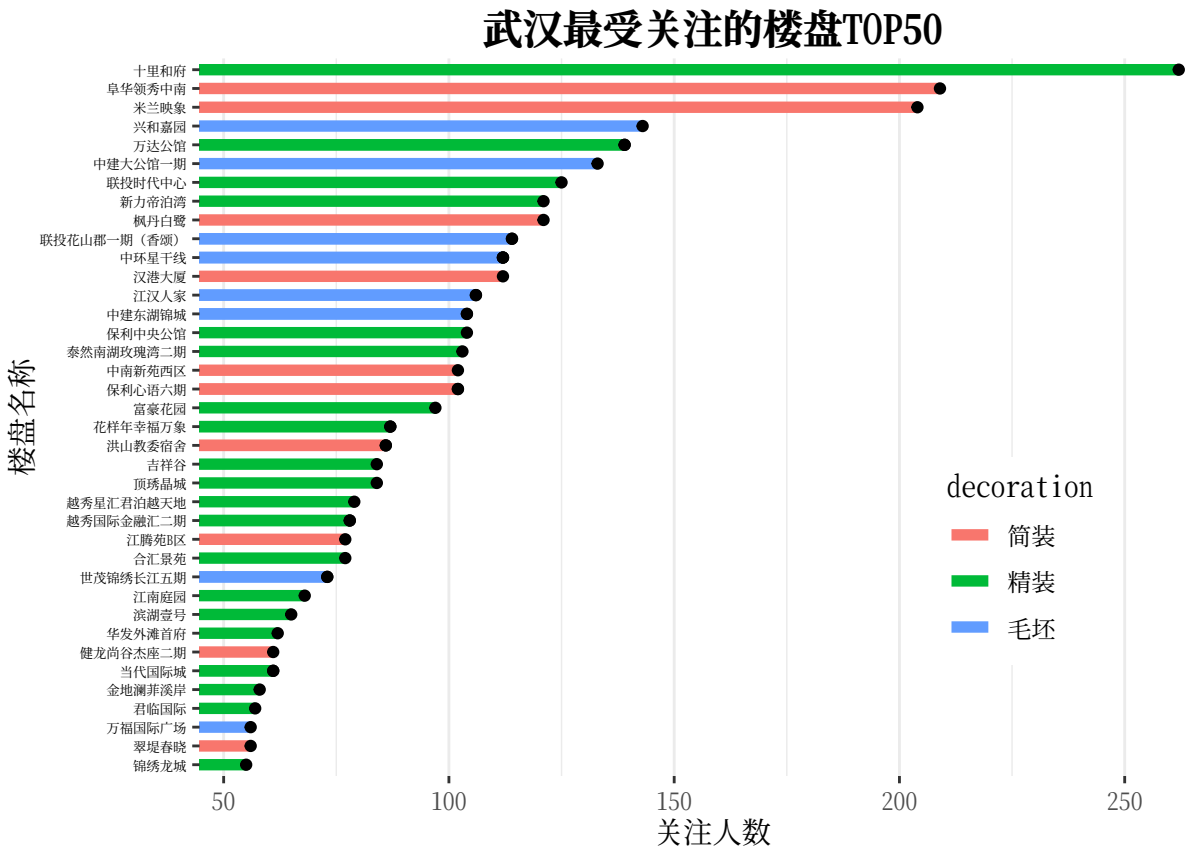
- 房屋在楼层中间位置挂牌数量最多，房屋所处楼层在较低或较高都不太影响房屋打算出售意向, 精装房不管在任何楼层都依然是二手市场的出售主力



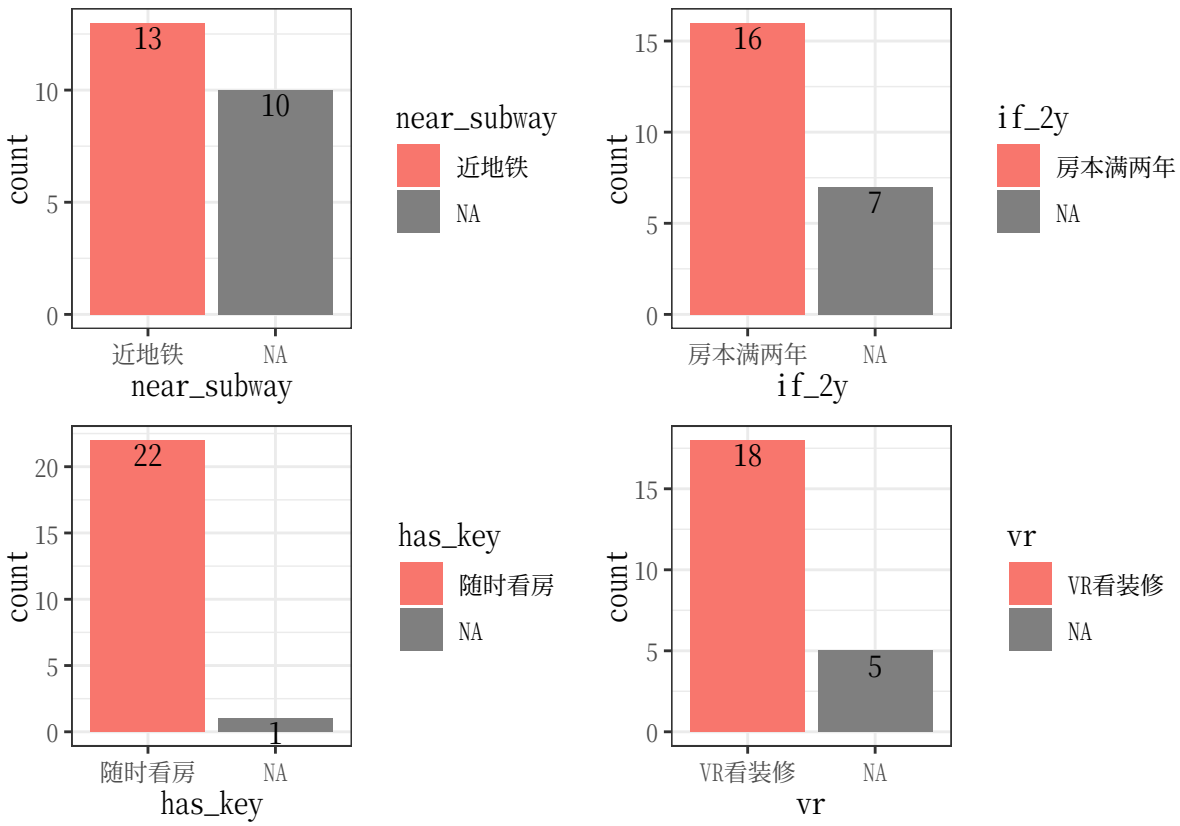
探索问题 3

探索最受关注楼盘是哪些

- 发现：
- 发现 1
- 最受欢迎楼盘为十里和府，毫无悬念的为精装房



- 发现 2
- 关注超过人数超过 100 的楼盘为什么受欢迎比如是否精装修，近地铁，房本满两年，随时看房，可 VR 看装修等。通过图能看出房本满两年、随时看房、可 VR 看装修等由房屋可控的因素对是否受欢迎影响非常大。而是否近地铁由于小区地理属性本身的限制，占比也超过 50%，也为正相关影响。



发现总结

通过对该数据包括的 3000 套房产信息进行数据分析，我发现：

- 1. 房屋价格在 90 万至 190 万之间数据比较集中，剔除离群点的数据，整体价格趋近正态分布，房屋单价与房屋面积成线性正相关关系。
- 2. 精装房是挂牌最多的房屋，精装房的房屋单价及总价的中值都比其他类型要高，符合市场规律。
- 3. 房屋总价排名靠前的区域与房屋单价排名靠前的区域有 80% 的重合性，关联分析后，单价与总价的合并排名成线性正相关关系，体现了单价高的区域房屋出售的面积也更大；房屋总价与单价排名靠前的区域单价/面积的比值更大，并且越靠前区域的比值越大，体现了高价值区域的楼面价值更大。
- 4. 三室两厅以及精装是挂牌最多的房屋，房屋所在楼层不影响挂牌量，可能由于数据中标注为房屋所在楼层而不是整体房屋楼层，但该组数据同样也表明了挂牌多为家庭改善型置换出售，即房屋较大，任何楼层出售数量都很平均，多为精装修。
- 5. 房本满两年、随时看房、可 VR 看装修等由房屋出售时的可控因素影响房屋挂牌的关注人数，因为这些属性可由售房者控制并且全对买方利好，数据的表现符合市场规律。