# 关于某家上武汉二手房的数据分析报告

2023281051040-王政-MEM

## 目录

**数据介绍** **2**

**一、配置环境** **3**

1、配置环境，导入数据： . . . . . 3

2、查看数据整体结构： . . . . . 3

**二、数据清洗** **5**

1、去重，并查看数据缺失情况 . . . . . 5

2、在对区域内容进行统计中发现歧义内容，进行修改 . . . . . 6

3、查看整体分布特征 . . . . . 6

**三、数据分析与可视化** **9**

1、关键字段数据分布情况 . . . . . 9

2、区域维度分析 . . . . . 13

3，主要区域的价格分布情况 . . . . . 15

4, 房屋单价与房屋总价模型 . . . . . 18

**四、总结** **20**

# 数据介绍

本报告链家数据获取方式如下：数据为 2023 年 9 月 12 日获取了链家武汉二手房网站中数据。

- 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**

| 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 |

# 一、配置环境

## 1、配置环境，导入数据：

```
knitr::opts_chunk$set(echo = TRUE)
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 1.9.2     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(pander)
library(modelr)
library(showtext)
```

```
## 载入需要的程辑包: sysfonts
## 载入需要的程辑包: showtextdb
```

```
showtext_auto(enable = TRUE)
lj <- read.csv("./data/2023-09-12_cleaned.csv")
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec = dec,
## : embedded nul(s) found in input
```

## 2、查看数据整体结构：

```
glimpse(lj)
```

```
## 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         <int> 18709, 14613, 15968, 15702, 17509, 10376, 12346, 163~
```

```
## $ bedrooms          <int> 3, 3, 1, 3, 3, 3, 2, 3, 4, 3, 5, 3, 4, 3, 3, 2, 3, 4~
## $ livingrooms       <int> 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", "", "", "东", "", "北", "", "北", "北", "北", ~
## $ decoration        <chr> "精装", "精装", "简装", "精装", "简装", "精装", "简~
## $ property_t_height <int> 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         <int> 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", ~
```

# 二、数据清洗

## 1、去重，并查看数据缺失情况

样本数据存在重复情况，由原来的 3000 个样本数，去重后得到实际可用样本数 2515 个。数值缺失度较低，对数值分析影响较小。

```
lj <- distinct(lj)
pander(summary(is.na(lj)))
```

表 2: Table continues below

| property_name | property_region | price_ttl | price_sqm |
|---|---|---|---|
| Mode :logical | Mode :logical | Mode :logical | Mode :logical |
| FALSE:2515 | FALSE:2515 | FALSE:2515 | FALSE:2515 |
| NA | NA | NA | NA |

表 3: Table continues below

| bedrooms | livingrooms | building_area | directions1 | directions2 |
|---|---|---|---|---|
| Mode :logical | Mode :logical | Mode :logical | Mode :logical | Mode :logical |
| FALSE:2515 | FALSE:2515 | FALSE:2515 | FALSE:2515 | FALSE:2515 |
| NA | NA | NA | NA | NA |

表 4: Table continues below

| decoration | property_t_height | property_height | property_style |
|---|---|---|---|
| Mode :logical | Mode :logical | Mode :logical | Mode :logical |
| FALSE:2515 | FALSE:2515 | FALSE:2462 | FALSE:2515 |
| NA | NA | TRUE :53 | NA |

| followers | near_subway | if_2y | has_key | vr |
|---|---|---|---|---|
| Mode :logical | Mode :logical | Mode :logical | Mode :logical | Mode :logical |
| FALSE:2515 | FALSE:1310 | FALSE:1050 | FALSE:2092 | FALSE:1754 |
| NA | TRUE :1205 | TRUE :1465 | TRUE :423 | TRUE :761 |

### 2、在对区域内容进行统计中发现歧义内容，进行修改

查看 property_region 字段内容：

```
print(filter(lj, property_region == "VR 看装修"))
```

```
##   property_name property_region price_ttl price_sqm bedrooms livingrooms
## 1         华清园         VR看装修       480     38351        3           2
##   building_area directions1 directions2 decoration property_t_height
## 1        125.16           南           北       精装                18
##   property_height property_style followers near_subway if_2y  has_key       vr
## 1              低        板塔结合        11       近地铁  <NA> 随时看房 VR看装修
```

对区域显示为"VR 看装修" 的字段内容进行修改，修改为" 三阳路"：

```
lj$property_region[lj$property_region=="VR 看装修"] <- " 三阳路"
print(filter(lj, property_region == "VR 看装修"))
```

```
##  [1] property_name     property_region   price_ttl         price_sqm
##  [5] bedrooms          livingrooms       building_area     directions1
##  [9] directions2       decoration        property_t_height property_height
## [13] property_style    followers         near_subway       if_2y
## [17] has_key           vr
## <0 行> (或0-长度的row.names)
```

### 3、查看整体分布特征

```
pander(summary(lj))
```

表 6: Table continues below

| 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 |
| NA | NA | Mean : 154.8 | Mean :15110 |
| NA | NA | 3rd Qu.: 188.0 | 3rd Qu.:18213 |
| NA | NA | Max. :1380.0 | Max. :44656 |

表 7: Table continues below

| 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 | NA |
| 3rd Qu.:3.000 | 3rd Qu.:2.000 | 3rd Qu.:118.03 | NA |
| Max. :7.000 | Max. :4.000 | Max. :588.66 | NA |

表 8: Table continues below

| 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 |
| NA | NA | Mean :24.05 | NA |
| NA | NA | 3rd Qu.:33.00 | NA |
| NA | NA | Max. :62.00 | NA |

表 9: Table continues below

| 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 |
| NA | Mean : 6.326 | NA | NA |
| NA | 3rd Qu.: 6.000 | NA | NA |
| NA | Max. :262.000 | NA | NA |

| has_key | vr |
|---|---|
| Length:2515 | Length:2515 |
| Class :character | Class :character |
| Mode :character | Mode :character |
| NA | NA |
| NA | NA |

| has_key | vr |
|---|---|
| NA | NA |

**可以直观看到**：2023 年 9 月 12 号链家武汉二手房数据呈现如下特征：

- 房屋总价：最低价 10.60 万元，最高价 1,380.00 万元，中位数 136.00 万元，平均数 154.80 万元，数据分布呈现右偏分布；
- 房屋单价：最低价 1,771.00 元/$m^2$，最高价 44,656 元/$m^2$，中位数 14,309 元/$m^2$，平均数元 15,110/$m^2$，数据分布呈现右偏分布，但右偏幅度较房屋总价较小，原因在于房屋面积不同影响在单价在总价上呈现的差异；
- 建筑面积：最小面积 22.77$m^2$，最大面积 588.66$m^2$，中位数 95.46$m^2$，平均数 100.67$m^2$，呈现右偏分布，符合上面房屋面积对房屋总价的初步猜测。
- 房间数、客厅数、楼栋整层数：观察数据平均数，武汉在售二手房多为二至三居室、一至二客厅、平均给楼层 24 层的中层建筑，较符合常规认知。

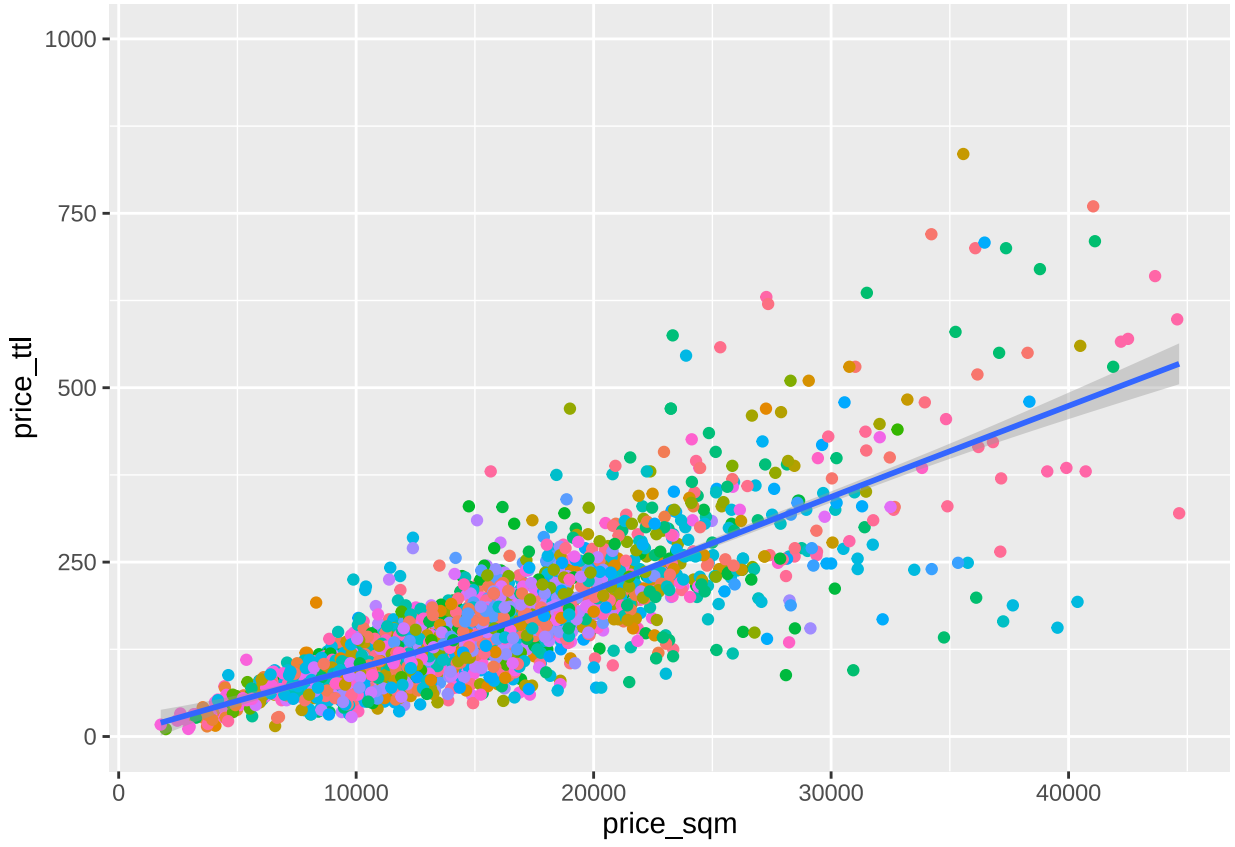
# 三、数据分析与可视化

## 1、关键字段数据分布情况

### （1）房屋单价和总价的关系

```
lj %>%
  ggplot() +
    geom_point(aes(x = price_sqm, y = price_ttl, color =  property_region,
                   group = property_region)) +
    theme(legend.position = "none") +
    geom_smooth(aes(x = price_sqm, y = price_ttl)) +
    coord_cartesian(ylim = c(0, 1000))
```

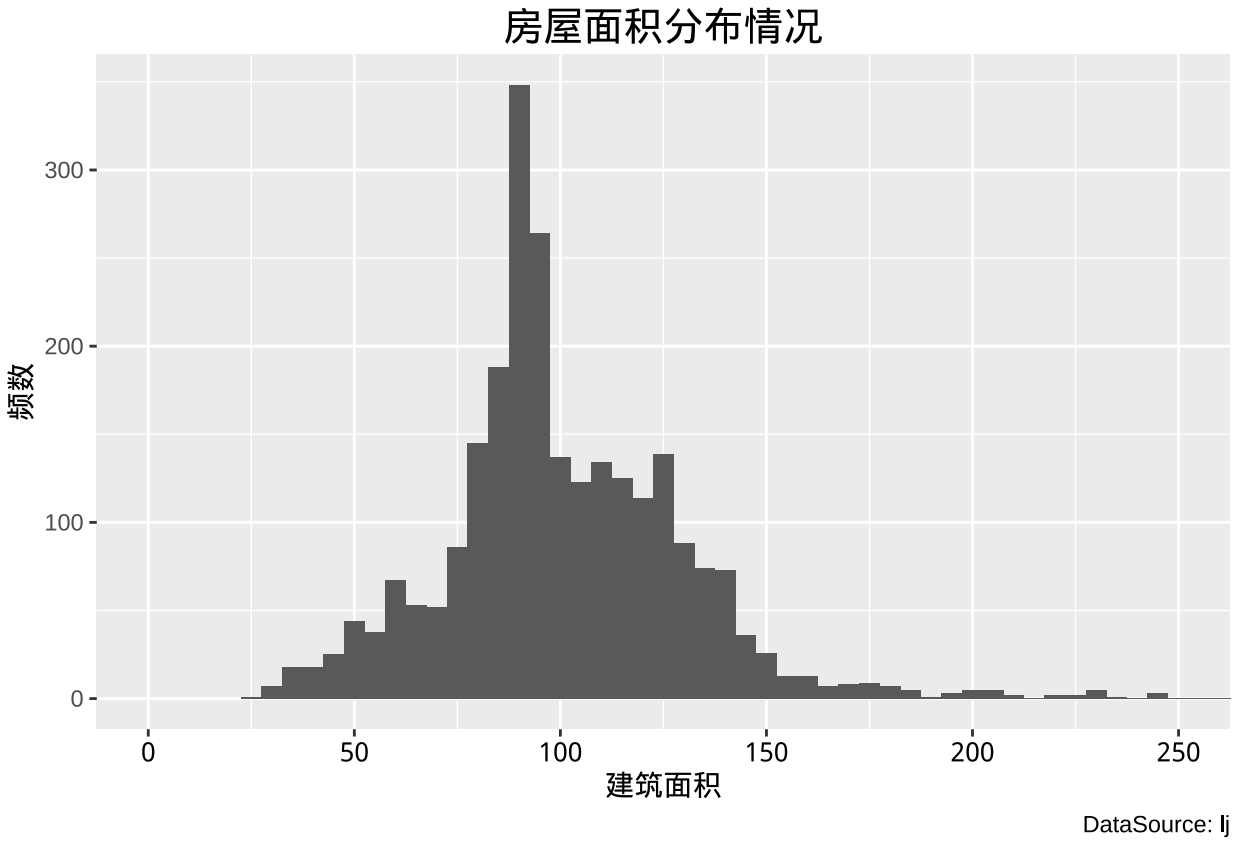
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




**结论：**房屋价格和房屋总价整体成正相关，存在极个别的异常数据。

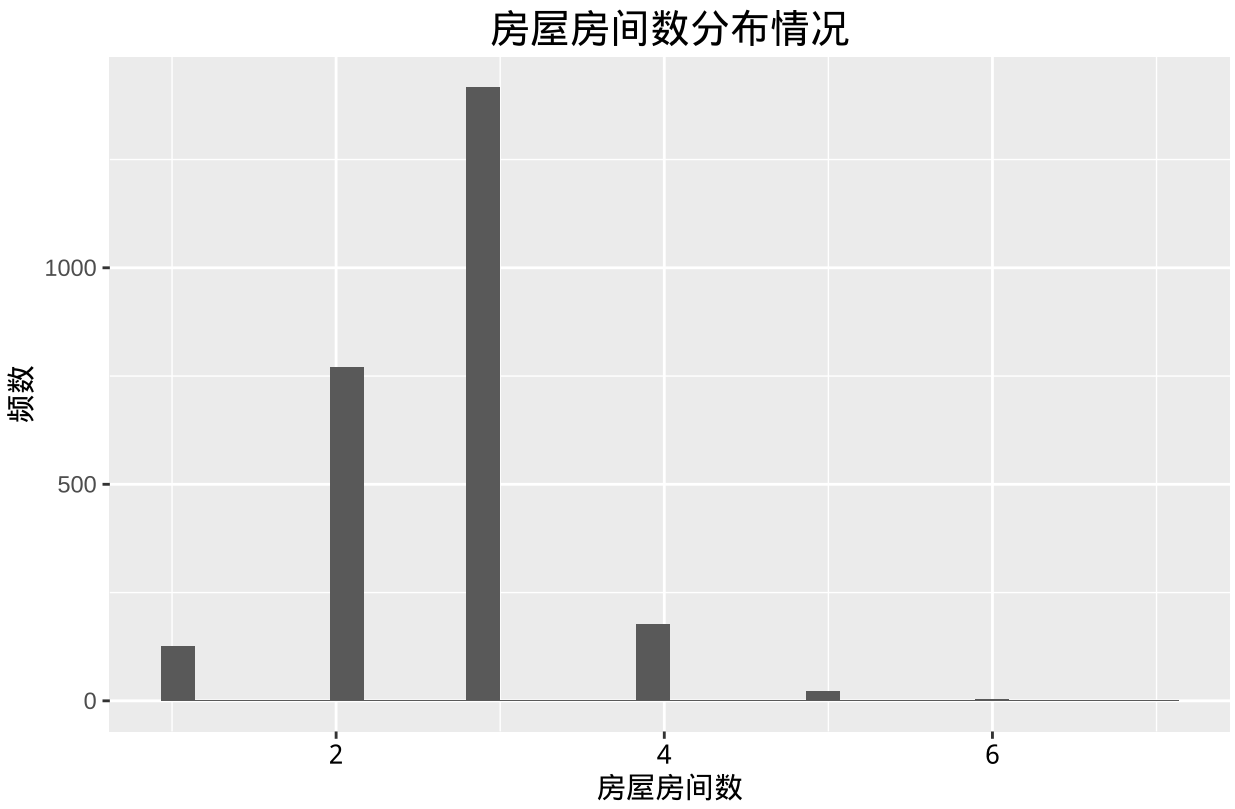
### （2）房屋建筑面积、房间数、客厅数分布情况

```
# 房屋建筑面积分布情况
ggplot(lj) +
  geom_histogram(aes(building_area), binwidth = 5) +
  coord_cartesian(xlim = c(0, 250)) +
  labs(title = " 房屋面积分布情况",  x = " 建筑面积",  y = " 频数",
       caption = "DataSource: lj") +
  theme(axis.text.x = element_text(family = "wqy-microhei", face = "bold",
                                   color = "black", size = 10),
        plot.title = element_text(family = "wqy-microhei", face = "bold",
                                  color = "black", size = 15, hjust = 0.5,
                                  vjust = 0,5))
```



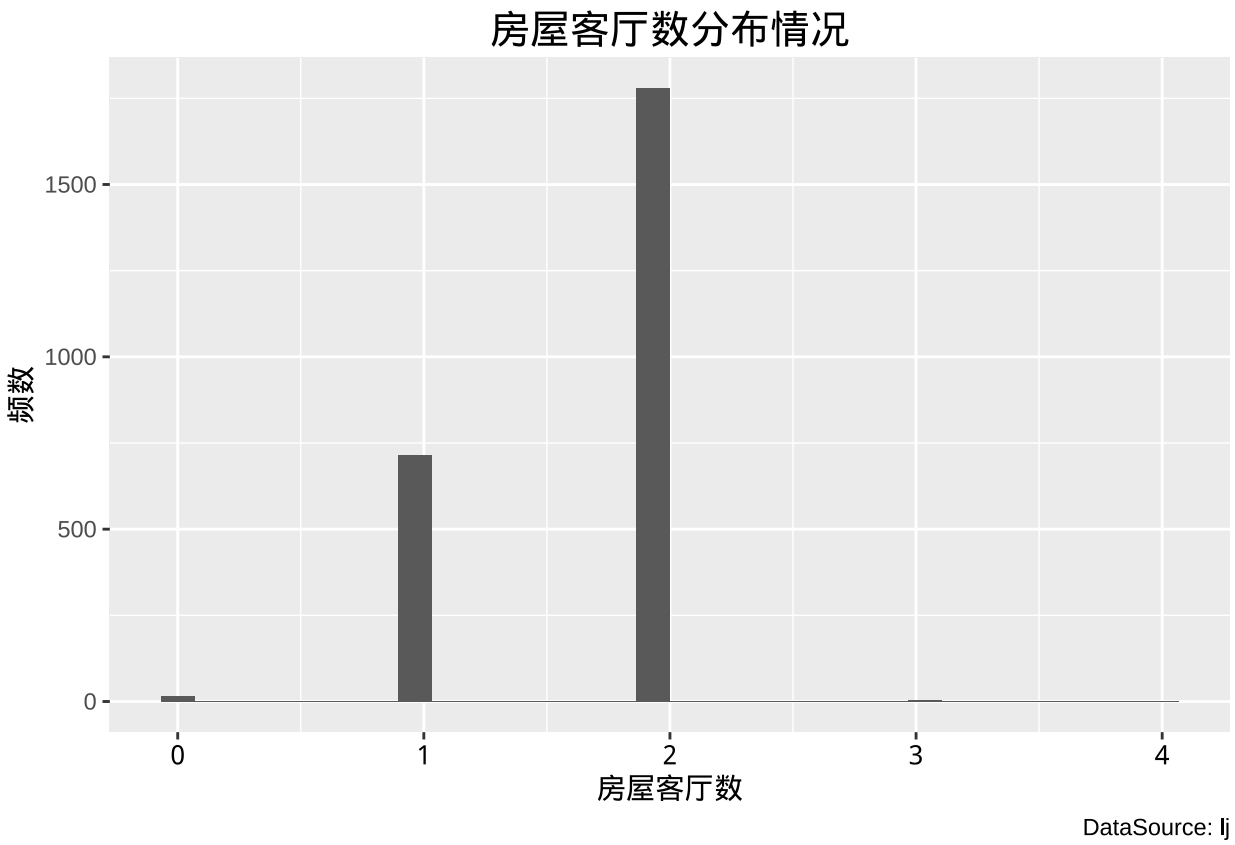

```
# 房屋房间数分布情况
ggplot(lj) +
  geom_histogram(aes(bedrooms), bins = 30) +
  labs(title = " 房屋房间数分布情况",  x = " 房屋房间数",  y = " 频数",
         caption = "DataSource: lj") +
  theme(axis.text.x = element_text(family = "wqy-microhei", face = "bold",
                                   color = "black", size = 10),
          plot.title = element_text(family = "wqy-microhei", face = "bold",
```

```
                                color = "black", size = 15, hjust = 0.5,
                                vjust = 0,5))
```




```
# 房屋客厅数分布情况
ggplot(lj) +
  geom_histogram(aes(livingrooms)) +
  labs(title = " 房屋客厅数分布情况",  x = " 房屋客厅数",  y = " 频数",
         caption = "DataSource: lj") +
  theme(axis.text.x = element_text(family = "wqy-microhei", face = "bold",
                                   color = "black", size = 10),
          plot.title = element_text(family = "wqy-microhei", face = "bold",
                                color = "black", size = 15, hjust = 0.5,
                                vjust = 0,5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

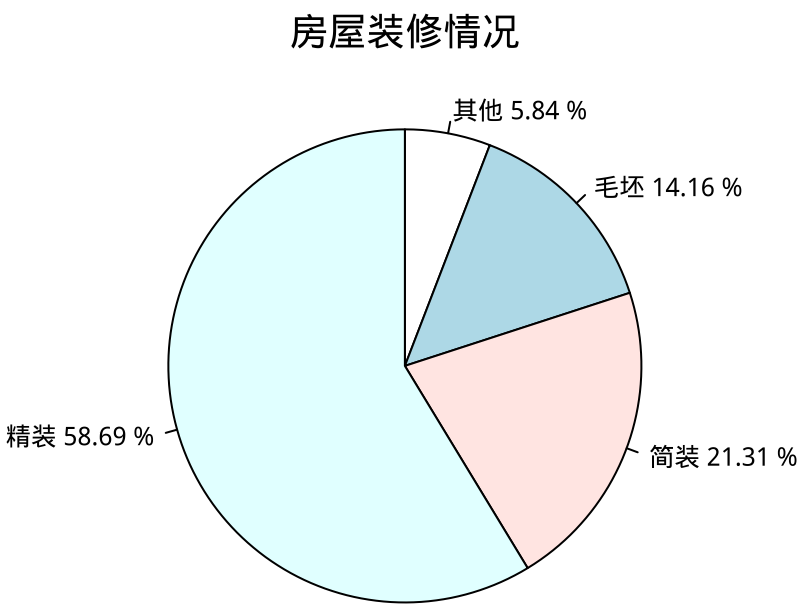



**结论**：在售二手房房屋面积多在 $100m^2$，存在超出 $400m^2$ 的异常数据；以三室两厅房型为主。

### （3）房屋装修情况

```
# 根据装修情况进行分组计数
decoration_count <- lj %>%
  group_by(decoration) %>%
  summarise(sum_decoration = n())

# 得到不同装修情况的百分比
get_rate <- function(x){
  j <- c(0)
  sum_count <- sum(x)
  for (i in 1:length(x)){
    j[i] <- round(x[i]/sum_count*100, 2)
  }
  return(j)
}

# 绘制饼状图
```

```
label_decoration <- get_rate(decoration_count$sum_decoration)
pie(decoration_count$sum_decoration, paste(decoration_count$decoration,
                                           label_decoration, "%"),
    radius = 1.0, clockwise=T, main = " 房屋装修情况", cex = 0.8)
```



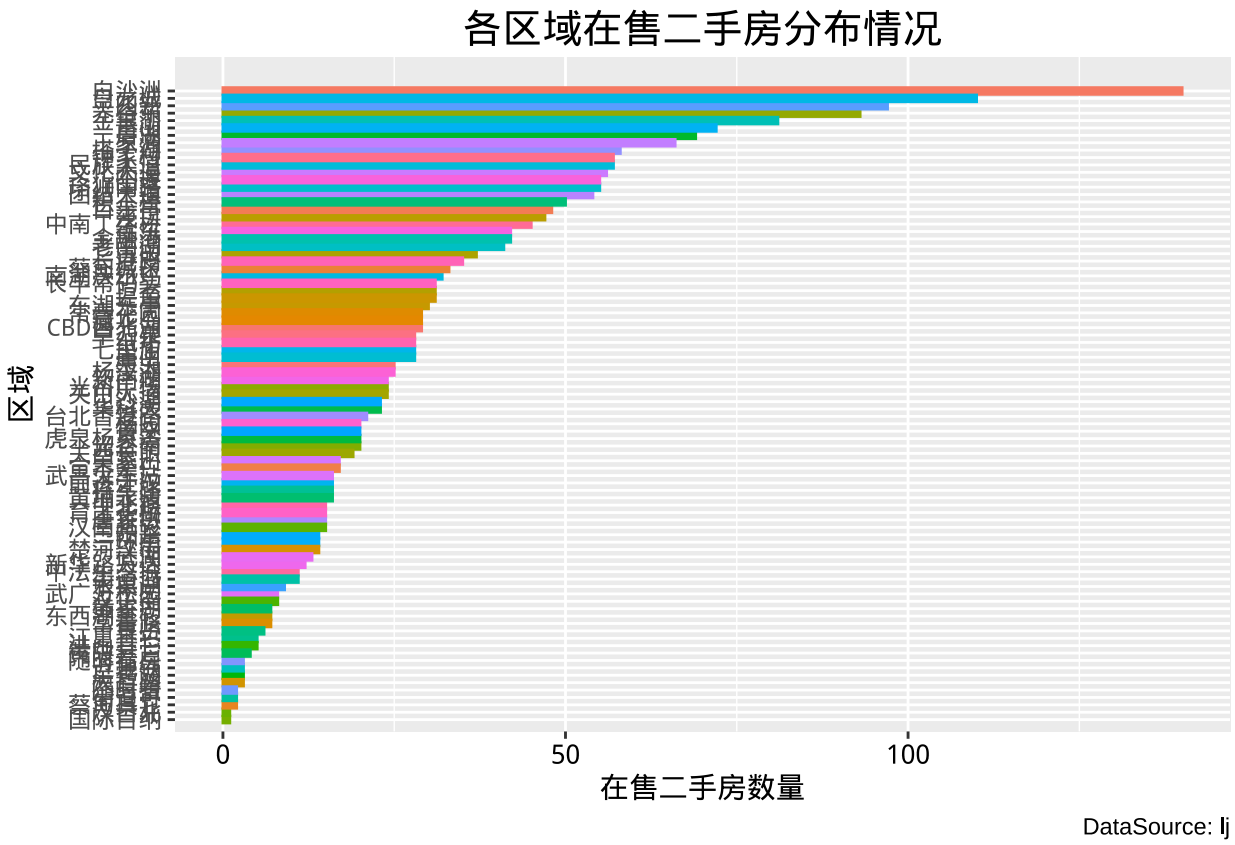

**结论：**近 80% 的二手房经过装修，其中精装数占总数的 50% 以上。

## 2、区域维度分析

### （1）各区域在售二手房分布情况

```
# 对于二手房所在区域进行分组计数
property_region_count <- lj %>%
  group_by(property_region) %>%
  summarise(sum_property = n())

# 绘制直方图
ggplot(property_region_count) +
  geom_bar(aes(x = sum_property , y = reorder(property_region, sum_property),
               color = property_region,fill = property_region),
           stat = 'identity') +
```

```
  coord_cartesian(ylim = c(0, 90)) +
  labs(title = " 各区域在售二手房分布情况",  x = " 在售二手房数量",  y = " 区域",
         caption = "DataSource: lj") +
  theme(axis.text.x = element_text(family = "wqy-microhei", face = "bold",
                                   color = "black", size = 10),
        plot.title = element_text(family = "wqy-microhei", face = "bold",
                                  color = "black", size = 15, hjust = 0.5,
                                  vjust = 0,5),
        legend.position = "none")
```

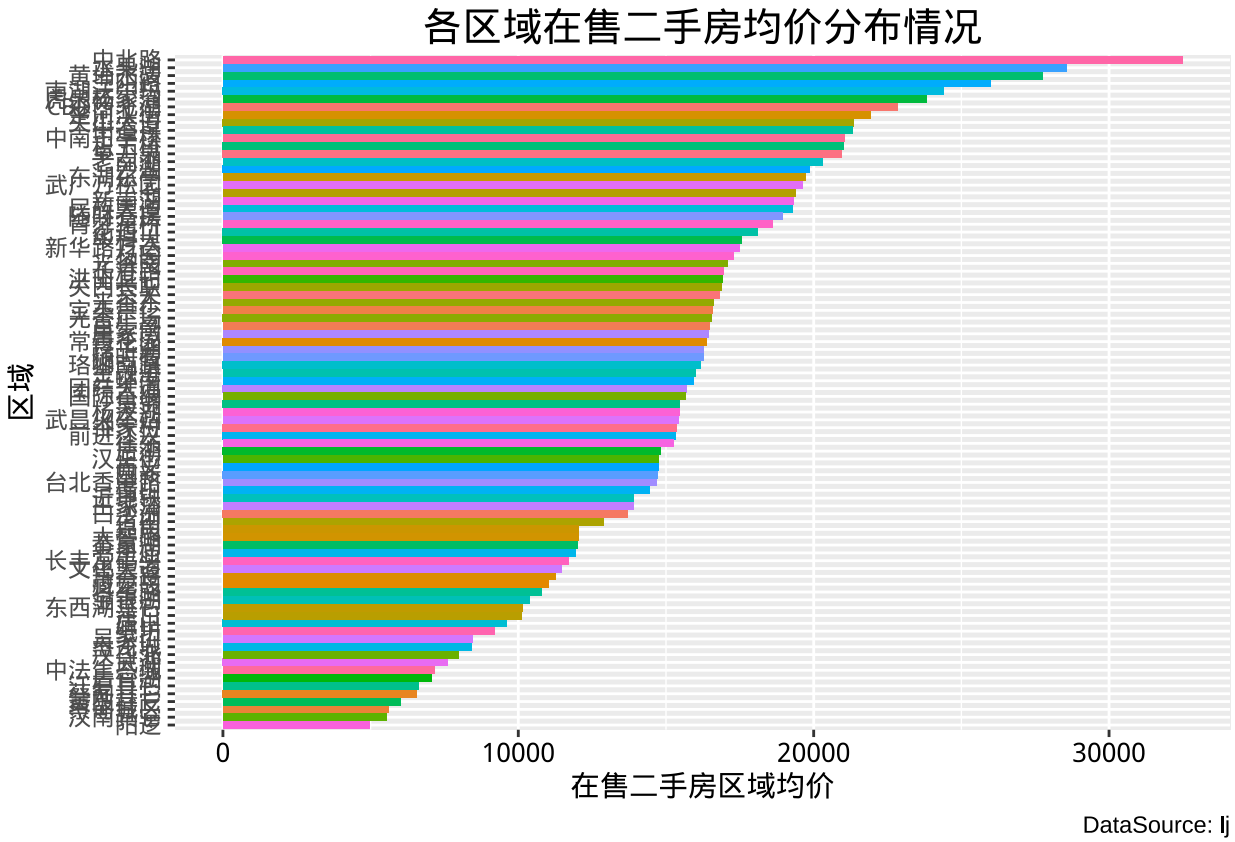



**结论：**在售二手房数量前 5 名依次是白沙洲 140 套、盘龙城 110 套、四新 97 套、光谷东 93 套、金银湖 81 套，较符合近几年武汉活力板块的特征。

### （2）各区域均价分布情况

```
# 根据区域进行分组，统计区域房屋单价均值
sqm_mean <- lj %>%
  group_by(property_region) %>%
  summarise(mean_region = mean(price_sqm))

# 绘制各区域房屋均值分布情况
```

```
ggplot(sqm_mean) +
  geom_bar(aes(x = mean_region, y = reorder(property_region, mean_region),
               fill = property_region),
           stat = "identity") +
  labs(title = " 各区域在售二手房均价分布情况",  x = " 在售二手房区域均价",  y = " 区域",
       caption = "DataSource: lj") +
  theme(axis.text.x = element_text(family = "wqy-microhei", face = "bold",
                                   color = "black", size = 10),
        plot.title = element_text(family = "wqy-microhei", face = "bold",
                                  color = "black", size = 15, hjust = 0.5,
                                  vjust = 0,5),
        legend.position = "none")
```

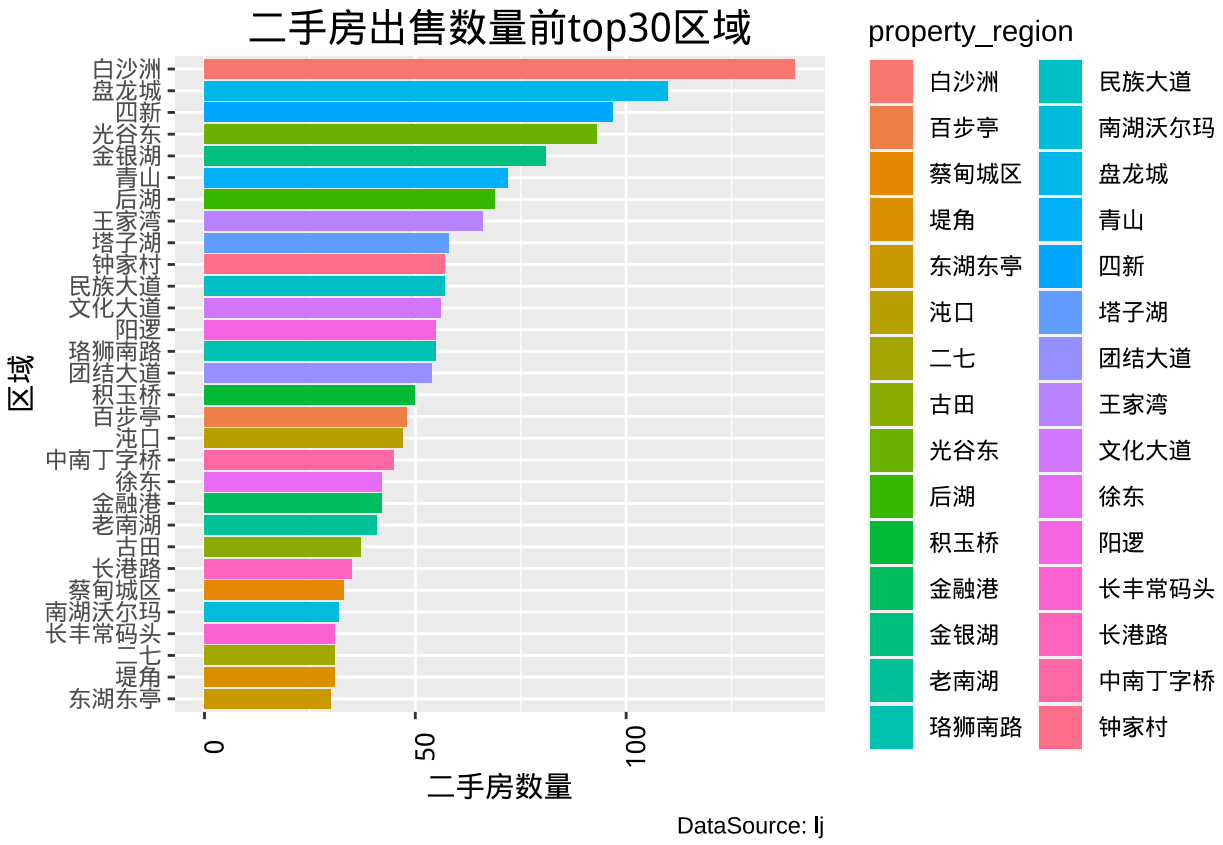



**结论**：武汉在售二手房地区房屋均价在 15000~20000 元间。其中中北路均价最高，超过 30000 元；阳逻均价最低，接近 5000 元。

### 3，主要区域的价格分布情况

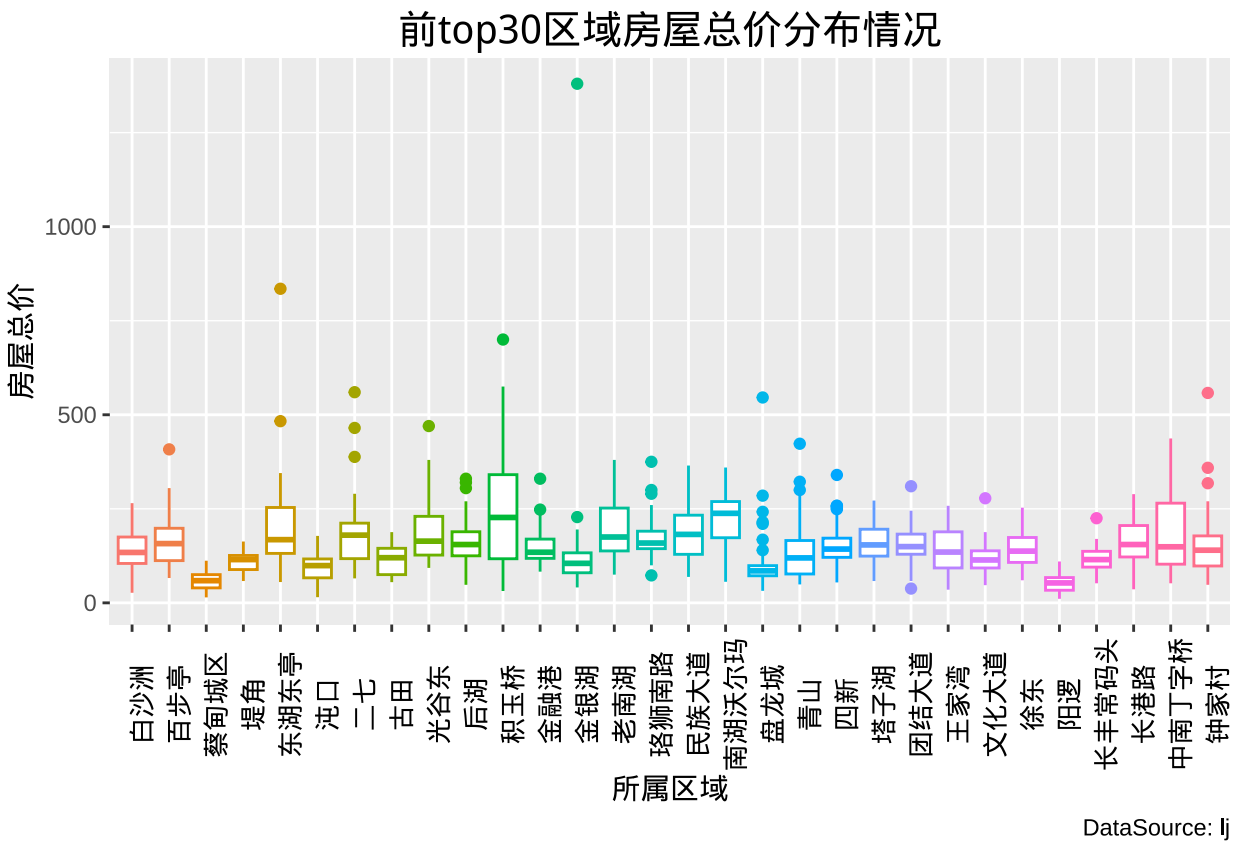
选取二手房销售数量前 30 区域做详细的统计分析。

```
# 选取二手房销售数量前 30 区域做详细的分析
property_name <- arrange(property_region_count, desc(sum_property))
property_name <- property_name[1:30, ]

## 查看 top30 区域的在售房屋数量分布
property_name %>%
  ggplot(aes(x = sum_property, y = reorder(property_region, sum_property),
             fill = property_region)) +
    geom_bar(stat = 'identity') +
    labs(title = " 二手房出售数量前 top30 区域",  x = " 二手房数量",  y = " 区域",
         caption = "DataSource: lj") +
    theme(axis.text.x = element_text(family = "wqy-microhei", face = "bold",
                                     color = "black", size = 10, angle = 90),
          plot.title = element_text(family = "wqy-microhei", face = "bold",
                                    color = "black", size = 15, hjust = 0.5,
                                    vjust = 0,5))
```



**（1）查看二手房销售数量前 30 区域屋总价分布情况**

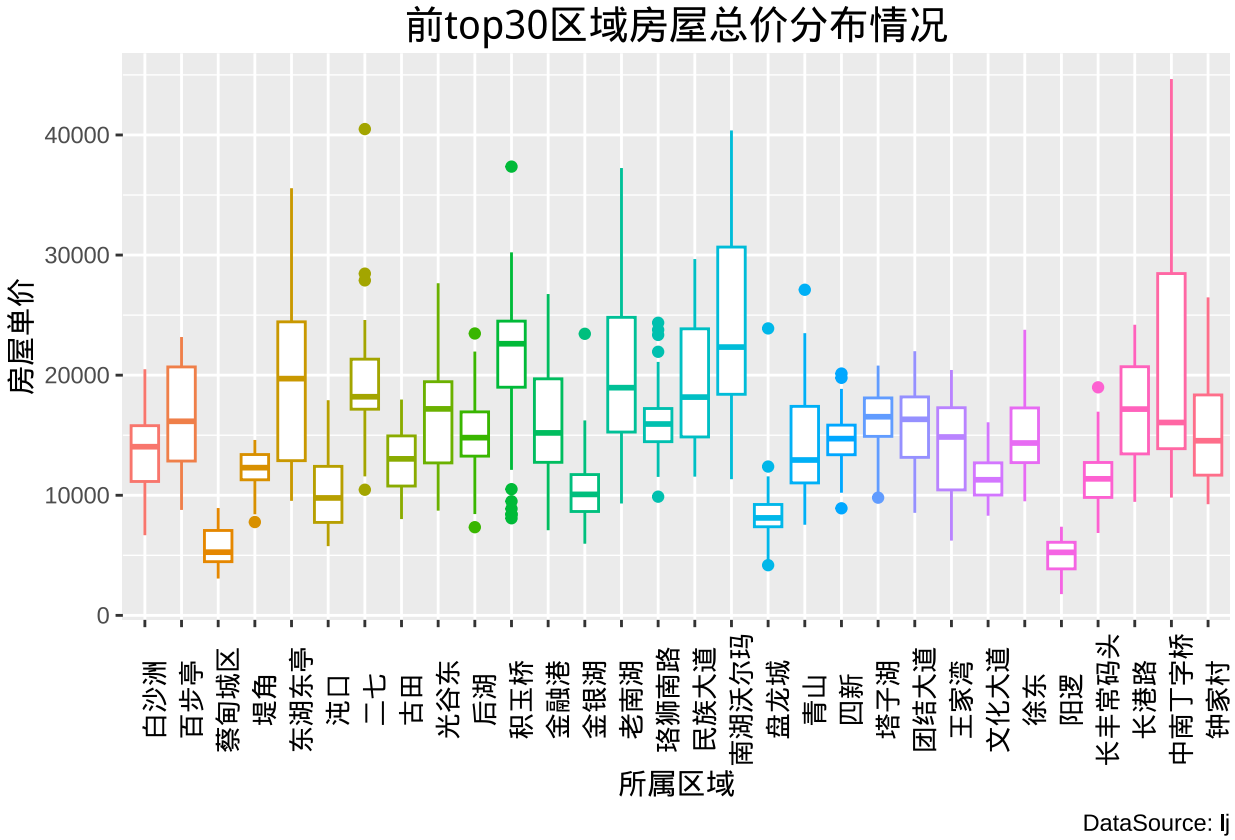
```
# 从 lj 数据库中选择对应详细数据
property_names <- property_name$property_region
lj_top30 <- dplyr::filter(lj, property_region %in% property_names)

# 查看区域总价的箱线图
ggplot(lj_top30) +
  geom_boxplot(aes(x = property_region, y = price_ttl, color = property_region)) +
  labs(title = " 前 top30 区域房屋总价分布情况",  x = " 所属区域",  y = " 房屋总价",
       caption = "DataSource: lj") +
  theme(axis.text.x = element_text(family = "wqy-microhei", face = "bold",
                                   color = "black", size = 10, angle = 90),
        plot.title = element_text(family = "wqy-microhei", face = "bold",
                                  color = "black", size = 15, hjust = 0.5,
                                  vjust = 0,5),
        legend.position = 'none')
```




**结论：**其中积玉桥、中南丁字桥区等区域房屋总价离散度较高；盘龙城、蔡甸城区等区域房屋总价集中性度较高；金银湖数据存在异常点，需要针对性分析。

（2）区域房屋单价分布情况

```
# 查看区域房屋单价的箱线图
ggplot(lj_top30) +
  geom_boxplot(aes(x = property_region, y = price_sqm, color = property_region)) +
  labs(title = " 前 top30 区域房屋总价分布情况",  x = " 所属区域",  y = " 房屋单价",
       caption = "DataSource: lj") +
  theme(axis.text.x = element_text(family = "wqy-microhei", face = "bold",
                                   color = "black", size = 10, angle = 90),
        plot.title = element_text(family = "wqy-microhei", face = "bold",
                                  color = "black", size = 15, hjust = 0.5,
                                  vjust = 0,5),
        legend.position = "none")
```




**结论**：其中东湖东、南湖沃尔玛、中南丁字桥区等区域房屋单价离散度较高；堤角、盘龙城、阳逻等区域房屋单价集中性度较高；中南丁子桥房屋单价分布右偏明显，说明地区房屋单价差异较大。

## 4, 房屋单价与房屋总价模型

通过线性回归分析，构建房屋单价与房屋总价的线性回归模型：

```
# 线性回归
cost_line_regression <- function(w){
  sum((lj$price_ttl - (w[1]*lj$price_sqm + w[2]))^2)
}

best <- optim(c(0,0), cost_line_regression)
cat(" 优化的参数: ", best$par, "\n")
```

```
## 优化的参数:  0.01181685 -23.73024
```

```
cat(" 目标函数: ", best$value, "\n")
```

```
## 目标函数:  8272589
```

```
cat(" 是否收敛: ", best$convergence)
```

```
## 是否收敛:  0
```

查看残值情况:

```
lj_line_regression <- data.frame(
  x = lj$price_sqm,
  y = lj$price_ttl
)
df_data <- add_residuals(lj_line_regression,
                         lm(y ~ x, data = lj_line_regression))
cat(" 残值均值: ",mean(df_data$resid))
```

```
## 残值均值:  8.599554e-13
```

**结论**：因为房屋单价是万元，目标函数结果较大，说明模型的预测结果的偏差较大，存在特征值选取不合适或模型选取存在问题，需要进一步改进。但残值均值小，说明对整体预测结果偏差较小。

# 四、总结

经过三年疫情影响和国内房地产市场步入存量时代，“房住不炒” 的房地产政策渐入人心。通过对 2023 年 9 月 12 日链家上的武汉二手房的数据进行分析，为个人筛选合适价格区间的区域精装二手房提供了一定的参考意见。