关于某家的商业数据分析报告

左仲博-MEM-第一次作业

目录

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

变量	解释
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、分析

1、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", ~
```

2、去重，查看整体分布特征

对数据进行去重。

```
lj <- distinct(lj)
summary(lj)
```

```
##  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
##
##
##
```

可以直观看到：

房屋总价数据分布呈右偏分布；房屋单价数据分布呈右偏分布；建筑面积呈右偏分布。武汉在售二手房多为 $2_{3\ 居室}、_{1}2$ 客厅、平均楼层 24 层的中层建筑住房。

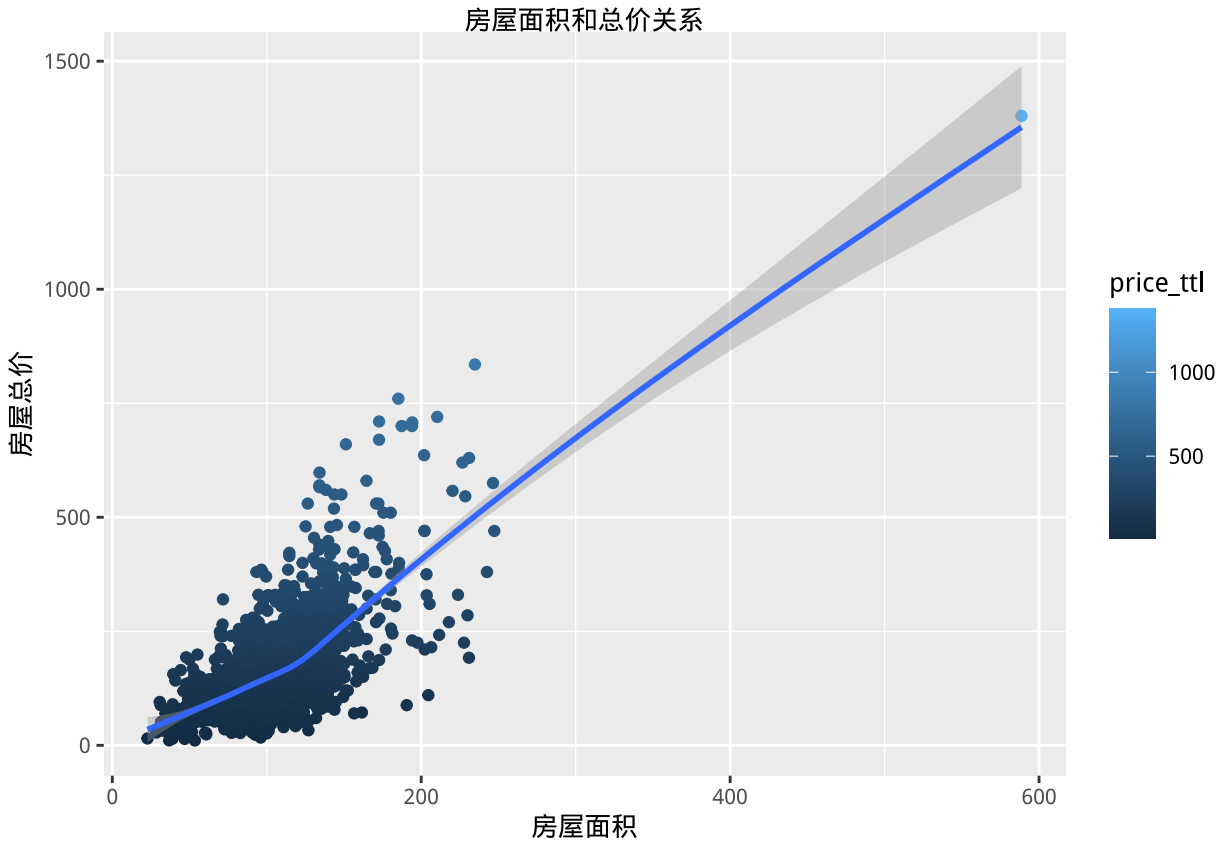
二、探索性分析

1、数据分布情况

房屋单价和总价

```
lj %>%
ggplot()+
geom_point(aes(x=building_area,y=price_ttl,color=price_ttl))+
geom_smooth(aes(x=building_area,y=price_ttl,))+
labs(title=" 房屋面积和总价关系",x=" 房屋面积",y=" 房屋总价")+
theme(plot.title=element_text(size=10,hjust=0.5,vjust=0.5))
```

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

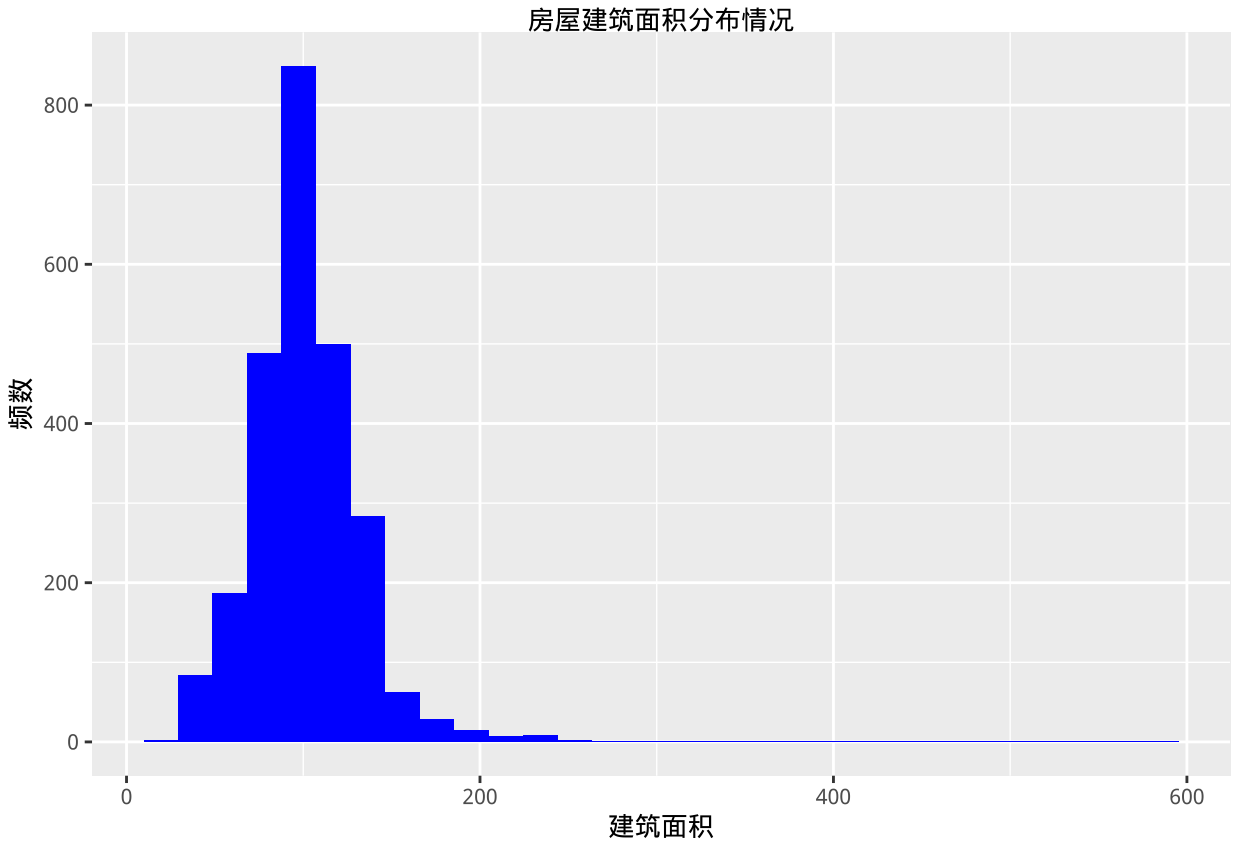


结论：除了极个别的异常数据, 房屋房价和房屋总价整体成正相关关系，且大部分房屋建筑面积集中在 100 平米和 200 万的价格上。

房屋建筑面积分布情况

```
ggplot(lj) +
geom_histogram(aes(building_area),fill='blue')+
labs(title=" 房屋建筑面积分布情况",x=" 建筑面积",y=" 频数")+
theme(plot.title=element_text(size=10,hjust=0.5,vjust=0.5))
```

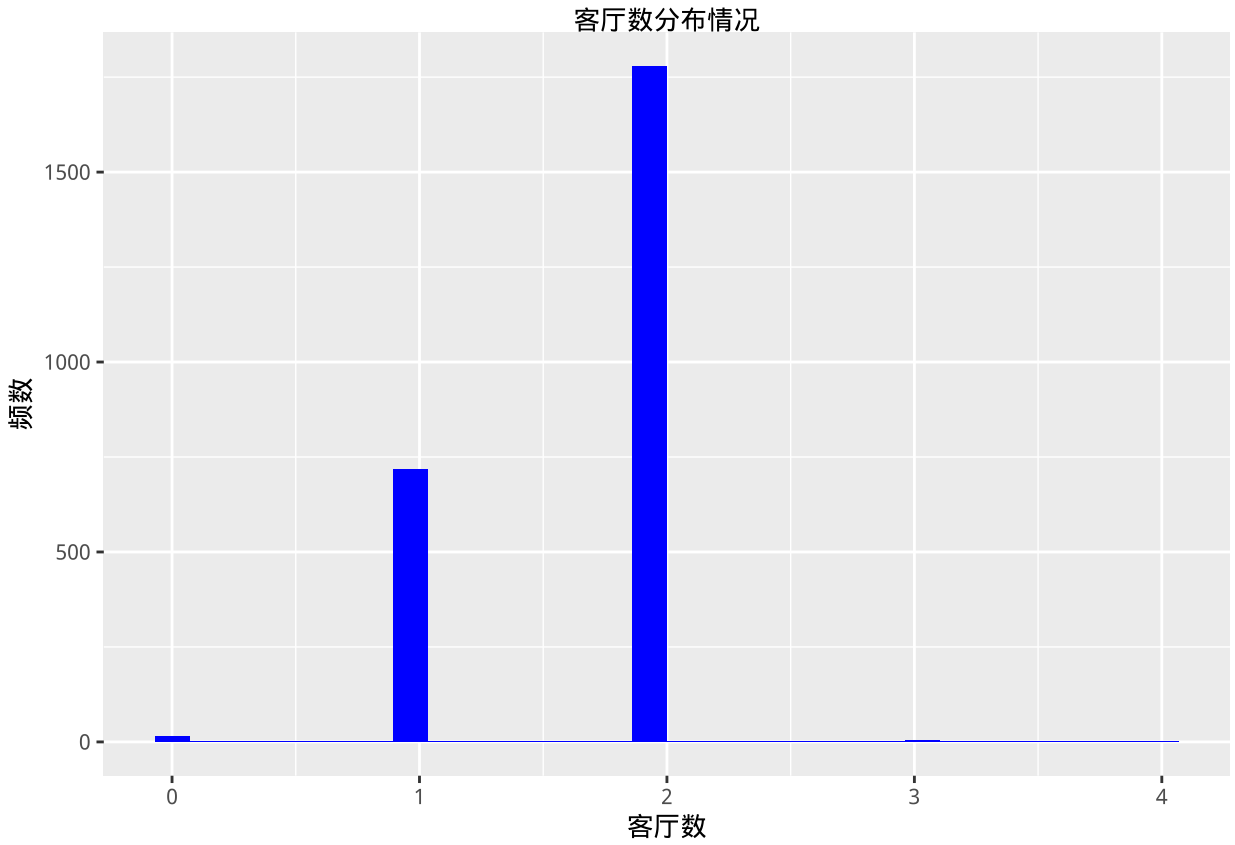
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



客厅数分布情况

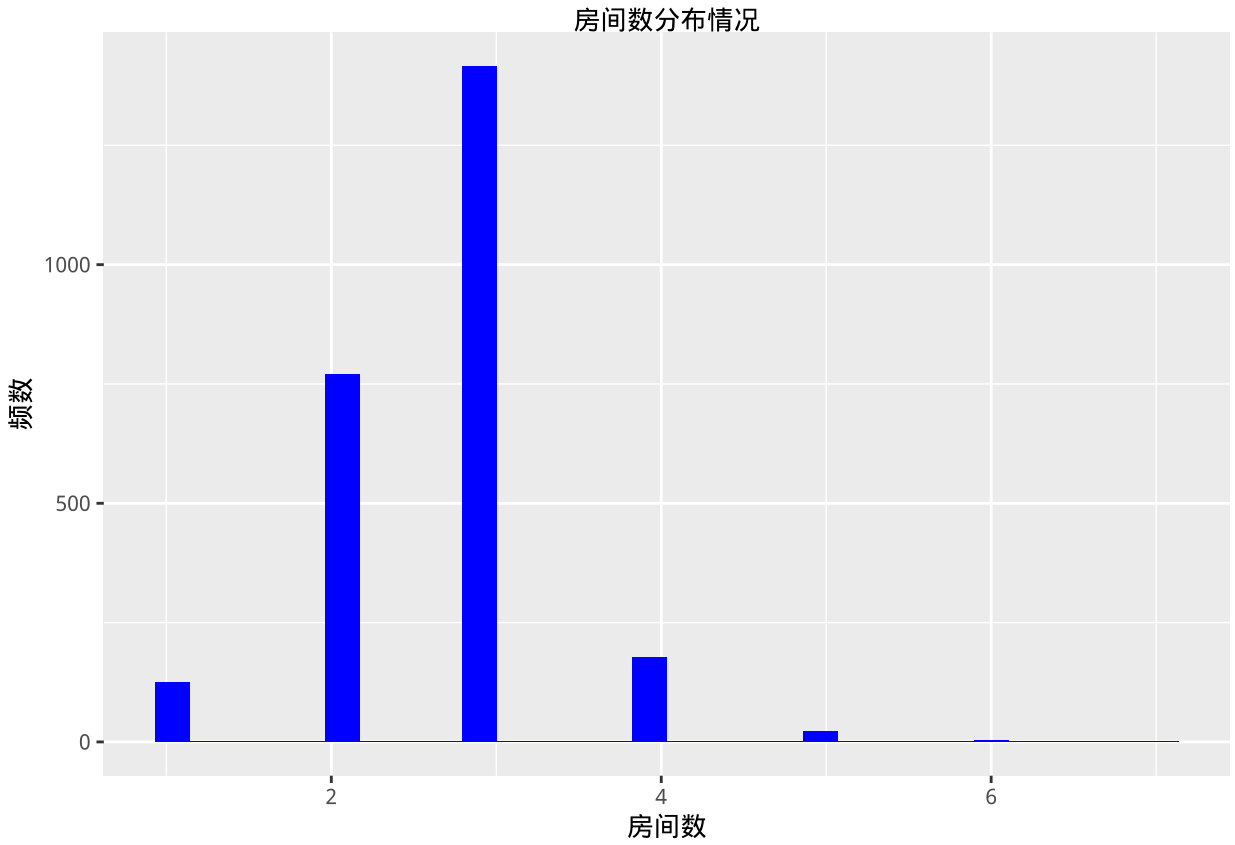
```
# 房面客厅数分布情况
ggplot(lj)+
geom_histogram(aes(livingrooms),fill='blue')+
labs(title=" 客厅数分布情况",x=" 客厅数",y=" 频数")+
theme(plot.title=element_text(size=10,hjust=0.5))
```

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



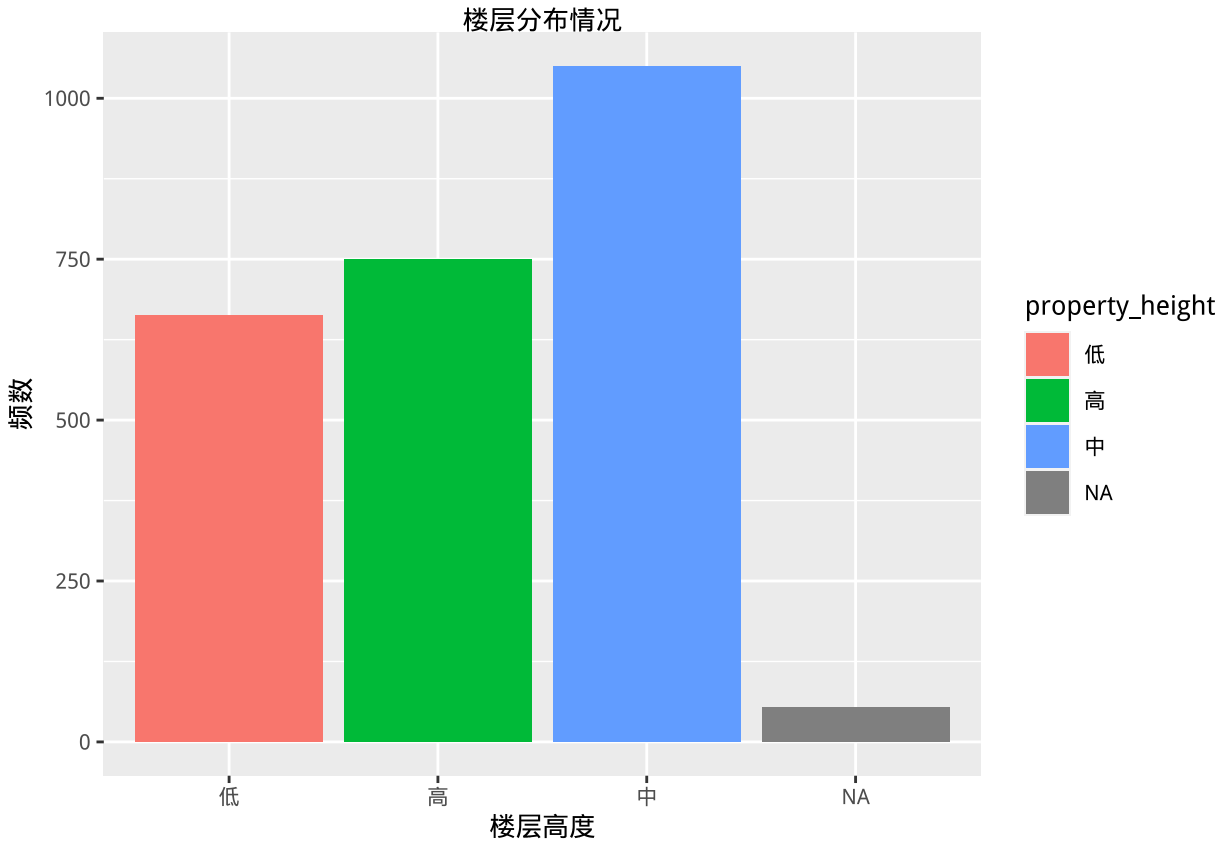
房间数分布情况

```
# 房屋房间数分布情况
ggplot(lj)+
geom_histogram(aes(bedrooms),fill='blue',bins=30)+
labs(title=" 房间数分布情况",x=" 房间数",y=" 频数")+
theme(plot.title=element_text(size=10,hjust=0.5))
```



楼层分布情况

```
# 楼层分布情况
ggplot(lj)+
geom_bar(aes(property_height,fill=property_height))+
labs(title=" 楼层分布情况",x=" 楼层高度",y=" 频数")+
theme(plot.title=element_text(size=10,hjust=0.5))
```

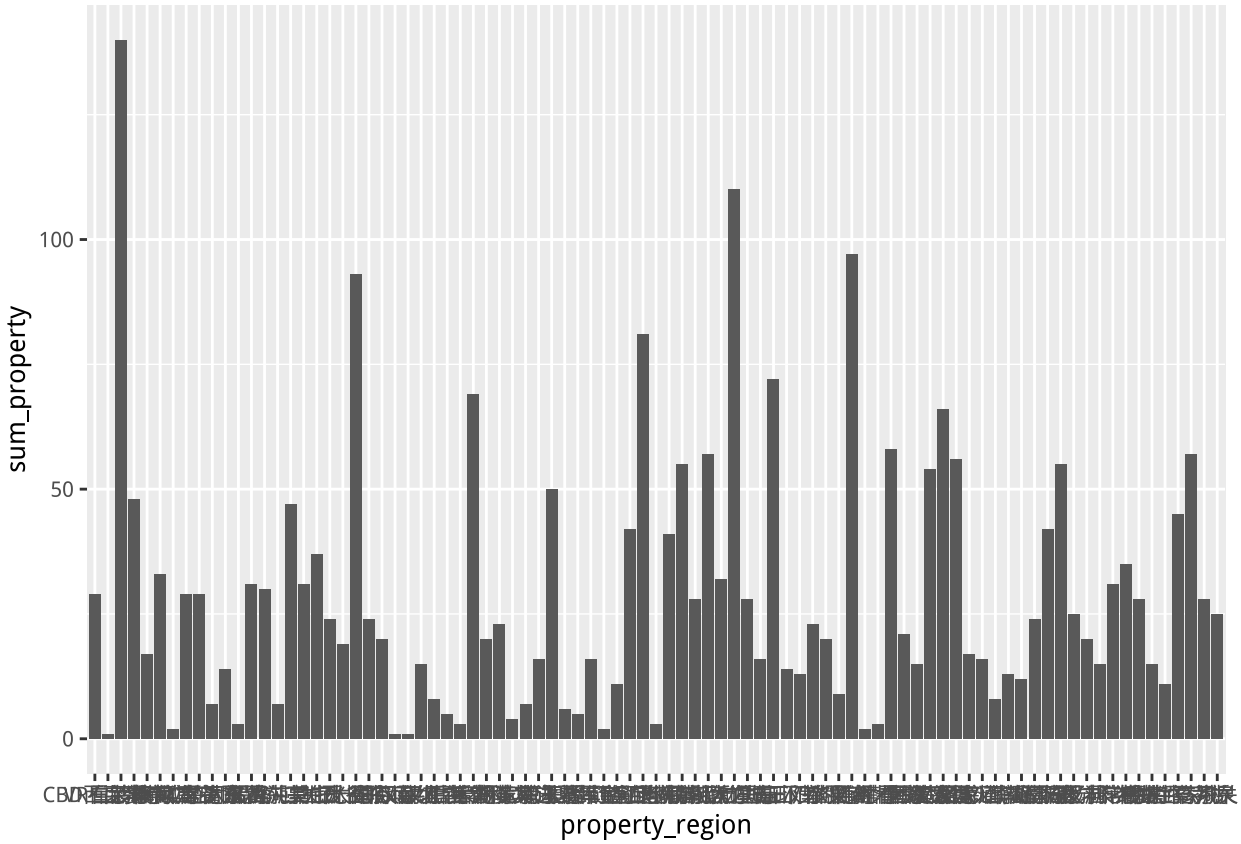


结论：房屋面积多在 100m^2，以中层三室两厅房型为主。

2、区域维度分析

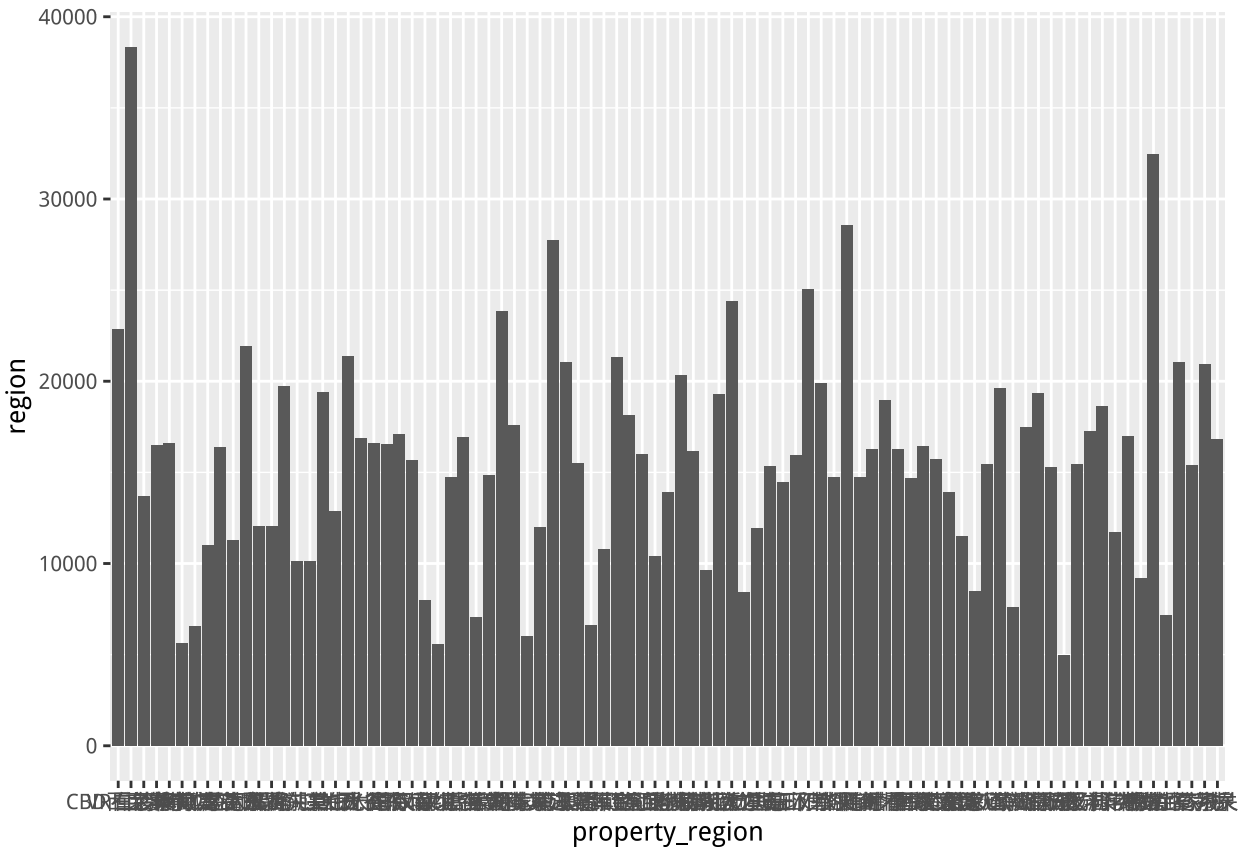
各区域在售二手房分布情况

```
region_count <- lj %>%
group_by(property_region) %>%
summarise(sum_property=n())
ggplot(region_count)+
geom_bar(aes(x=property_region,y =sum_property),stat='identity')
```



各区域均价分布情况

```
sqm <- lj %>%
group_by(property_region) %>%
summarise(region=mean(price_sqm))
ggplot(sqm)+
geom_bar(aes(x=property_region,y=region),stat="identity")
```

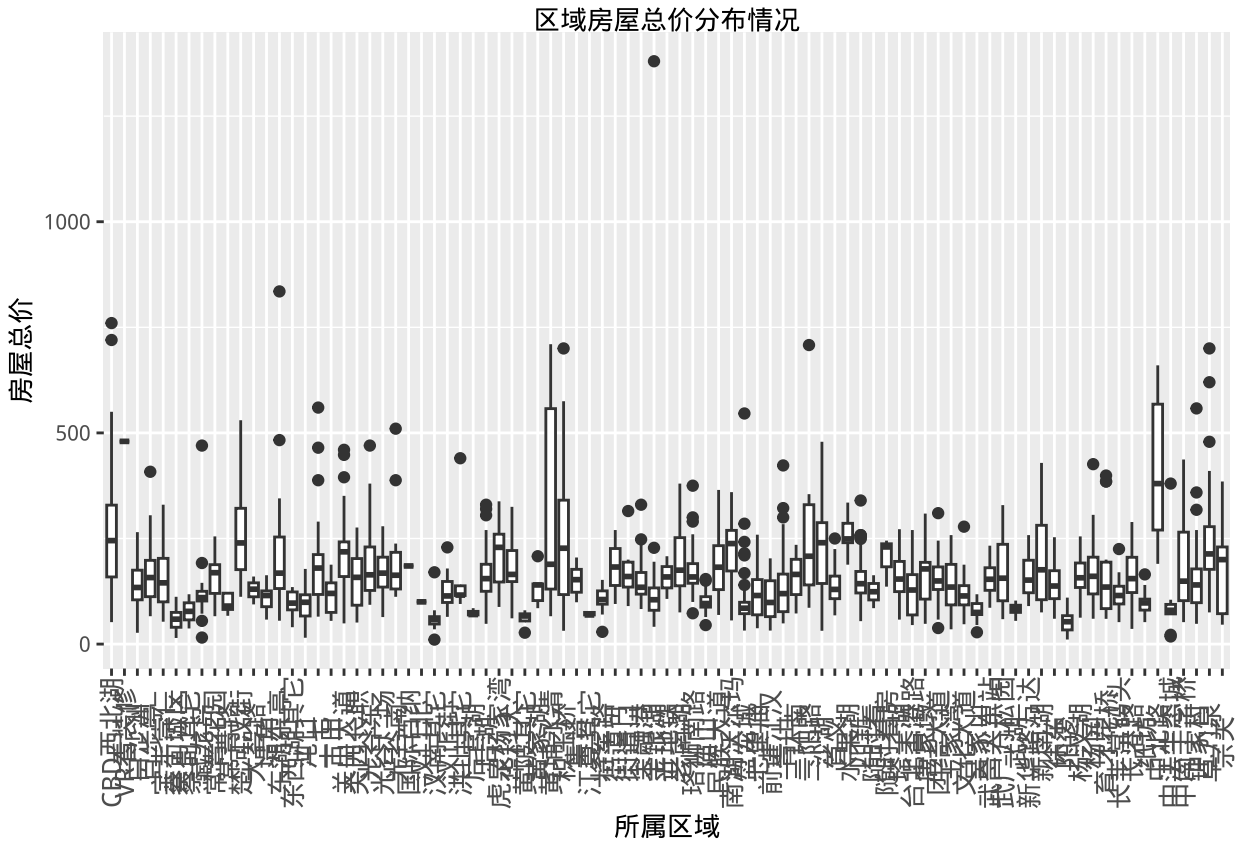


结论：武汉在售二手房地区房屋均价在 20000 元左右。

3，价格分布情况

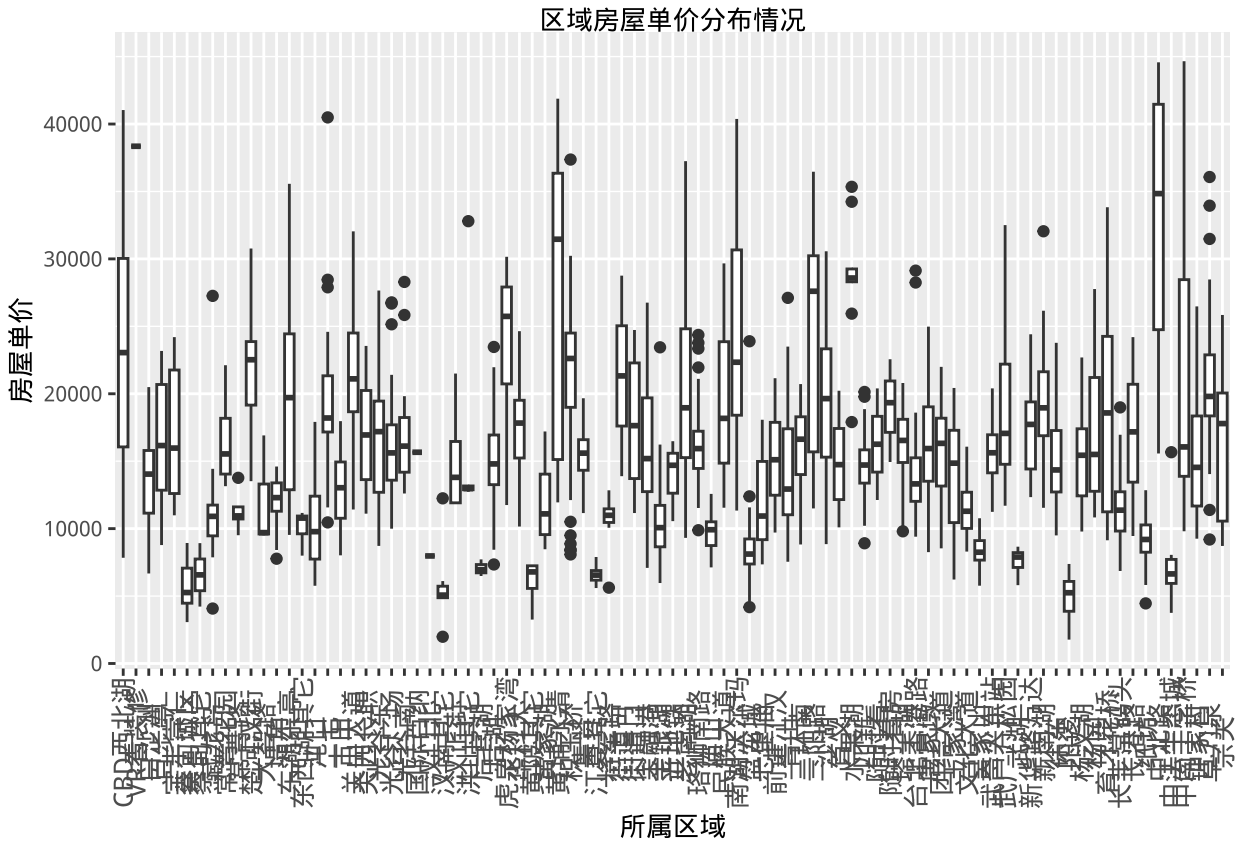
查看区域房屋总价分布情况

```
# 查看区域总价的箱线图
ggplot(lj)+
geom_boxplot(aes(x = property_region,y=price_ttl)) +
labs(title = " 区域房屋总价分布情况",x=" 所属区域",y=" 房屋总价") +
theme(axis.text.x=element_text(size=10,angle=90),
plot.title=element_text(size=10,hjust=0.5))
```



区域房屋单价分布情况

```
# 查看区域房屋单价的箱线图
ggplot(lj)+
geom_boxplot(aes(x=property_region,y=price_sqm))+
labs(title=" 区域房屋单价分布情况",x=" 所属区域",y=" 房屋单价")+
theme(axis.text.x=element_text(size=10,angle=90),
plot.title=element_text(size=10,hjust=0.5))
```



3, 房屋面积、房间数量、客厅数量与房屋总价的模型

线性回归分析：

```
model<- function(b){
sum((lj$price_ttl-(b[1]*lj$building_area+b[2]*lj$bedrooms+b[3]*lj$livingrooms+b[4]))^2)
}
best <- optim(c(0,0,0,0),model)
best$par
```

```
## [1]   2.266596 -18.973607 -10.094997  -3.379739
```