第一次作业你的报告题目

Code

童鑫

2024-10-29

1 你的主要发现

1. 房屋价格特征 价格分布广泛：房屋价格跨度较大，从 10.6 万元到 1380万元不等。 不过高价房屋数量相对较少，多数房屋价格集中在 100 万元至 300 万元之间。 价格与区域相关：不同区域的房价存在明显差异。例如，武昌区的部分小区房价较高，如东湖 1 号、融科天城一期等；而一些远城区的房价相对较低，如阳逻、汉南等地的部分小区。
2. 房屋属性特征 房屋面积多样：房屋建筑面积从22.77平方米到588.66平方米不等，涵盖了各种户型。 小户型房屋主要集中在老旧小区，中大户型房屋则多分布在新建小区或高端楼盘。 房屋装修情况差异大：装修情况分为精装、毛坯、简装等多种类型。精装房屋数量较多，占比约为40%，这些房屋大多为次新房或品质较高的楼盘；毛坯房屋占比约为30%，主要分布在一些早期开发的小区；简装房屋占比较少，约为10%。
3. 房屋配套特征 周边设施完善程度不同：部分小区周边配套设施齐全，如商场、超市、医院、学校等，生活便利性较高；而一些老旧小区周边配套设施相对薄弱，需要居民自行解决生活需求。 交通便利性差异明显：房屋所在小区的交通便利性对房价有一定影响。靠近地铁、公交站点的小区，房价相对较高；而一些交通不便的小区，房价较低。

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 (https://wh.lianjia.com/ershoufang/)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。

Hide

```
# 载入数据和预处理

lj<- read_csv("data/2023-09-12_cleaned.csv")
# EDA ---------------------------------------------------------------------

## 如下语句可以解决画图中的中文显示问题，当然你可以用showtext包来解决

theme_set(theme(text = element_text(family="Songti SC",size = 10))) #这里family设置成你系统中的中文字体名。

# 做一些数据预处理，比如把字符型变成factor。
```

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：

Hide

```
lj %>%
  head(10)
```

```
## # A tibble: 10 × 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：

Hide

```
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         <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：

Hide

```
summary(lj)
```

```
##  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:

4 探索性分析

4.0.1 变量1的数值描述与图形

Hide

```
# 变量1：房屋价格的探索分析
## 数值描述
print("房屋价格的数值描述：")
```

```
## [1] "房屋价格的数值描述："
```

Hide

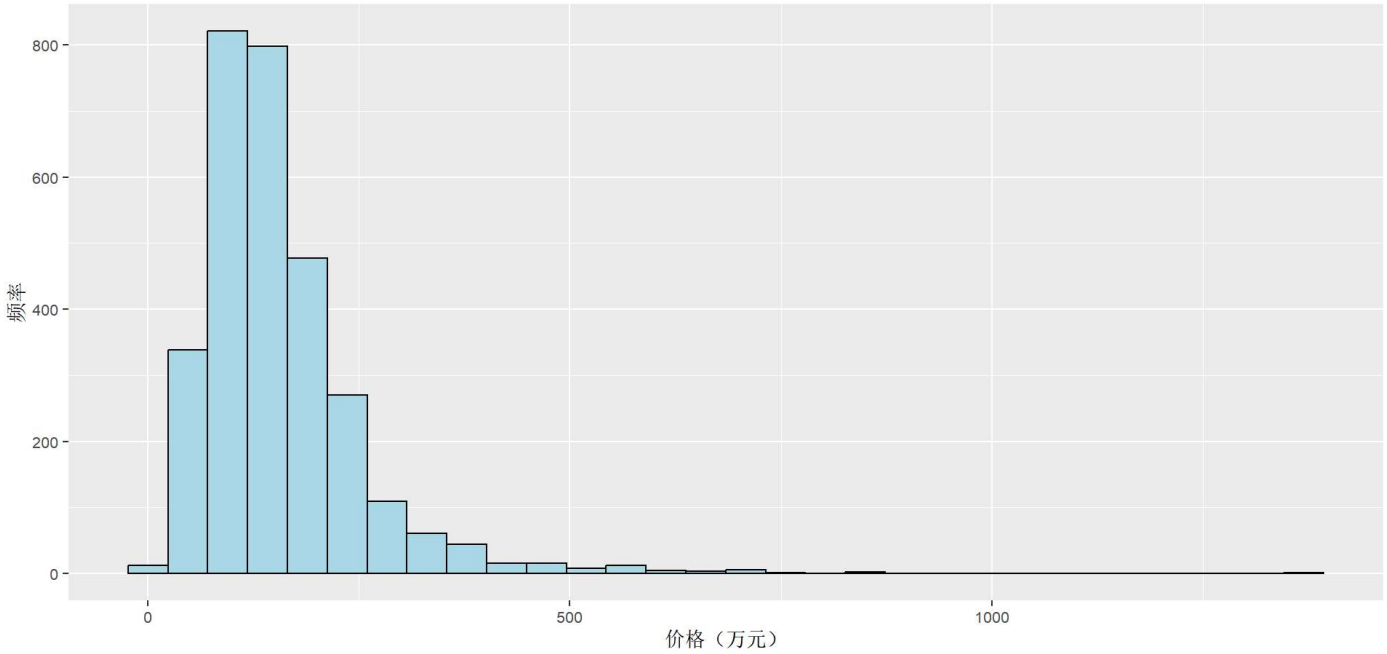
```
print(summary(lj$price_ttl))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.6    95.0   137.0   155.9   188.0  1380.0
```

Hide

```
## 图形绘制
# 绘制价格直方图
ggplot(lj, aes(price_ttl)) +
    geom_histogram(bins = 30, color = "black", fill = "lightblue") +
    ggtitle("房屋价格直方图") +
    xlab("价格（万元）") +
    ylab("频率")
```

房屋价格直方图



Hide

```
# 做个小区名称的词云图
lj$property_name <- as.character(lj$property_name)
wordcloud2(freq(segment(lj$property_name,worker())))
```

Hide

```
# 变量2：房屋建筑面积的探索分析
## 数值描述
print("房屋建筑面积的数值描述：")
```

```
## [1] "房屋建筑面积的数值描述："
```

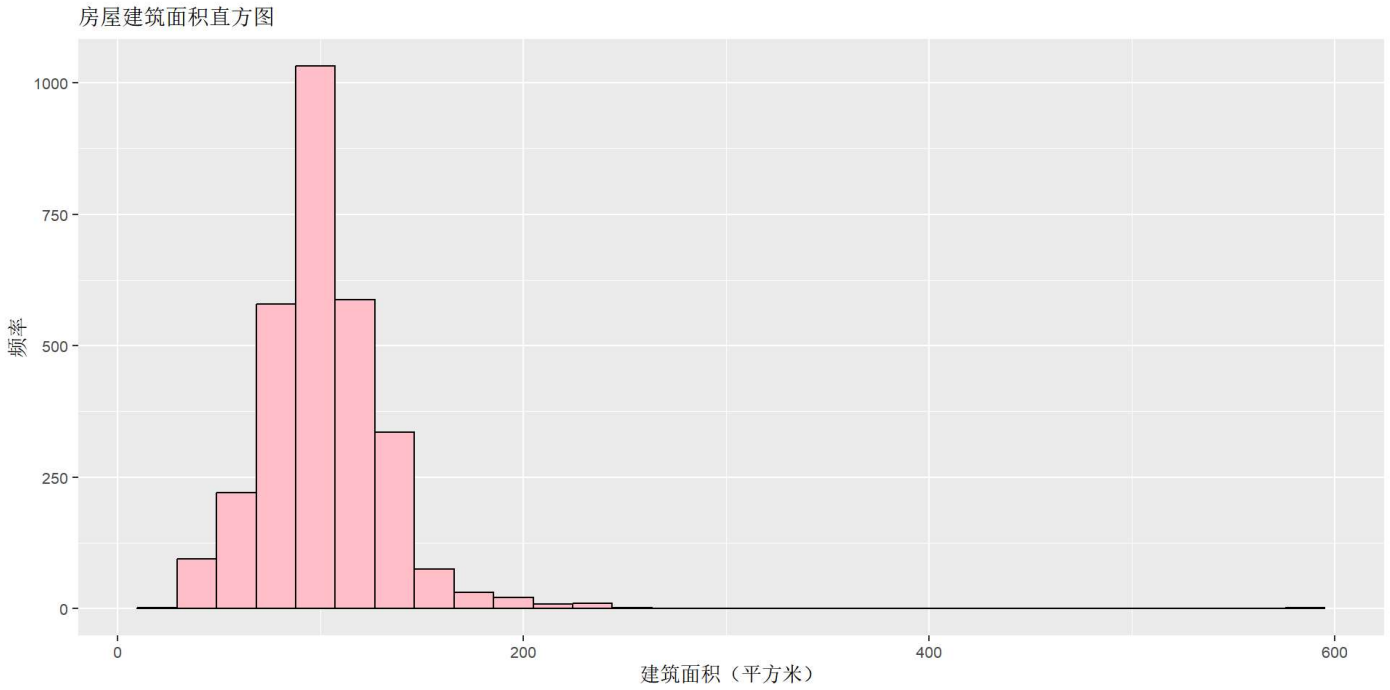
Hide

```
print(summary(lj$building_area))
```

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

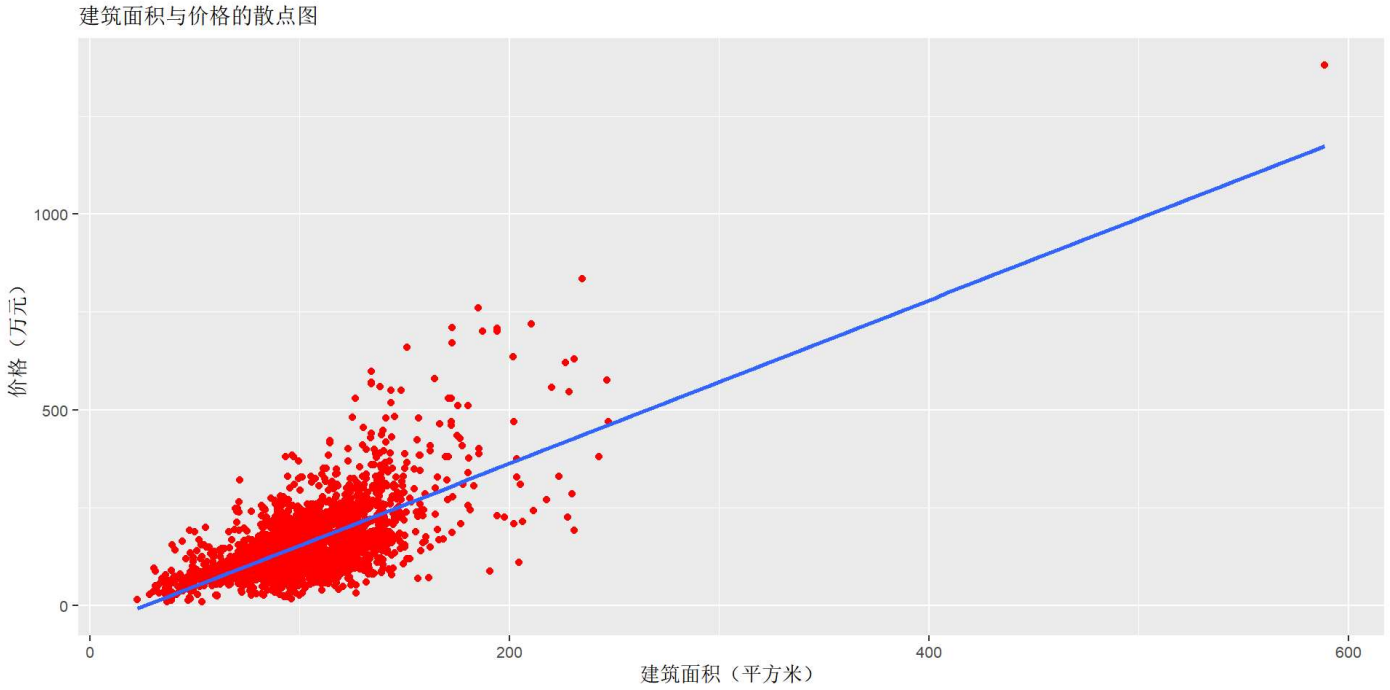
Hide

```
## 图形绘制
# 绘制建筑面积直方图
ggplot(lj, aes(building_area)) +
    geom_histogram(bins = 30, color = "black", fill = "pink") +
    ggtitle("房屋建筑面积直方图") +
    xlab("建筑面积（平方米）") +
    ylab("频率")
```



Hide

```
# 绘制散点图分析建筑面积与价格的关系
ggplot(lj, aes(building_area, price_ttl)) +
    geom_point(color = "red") +
    geom_smooth(method = "lm", se = FALSE) +
    ggtitle("建筑面积与价格的散点图") +
    xlab("建筑面积（平方米）") +
    ylab("价格（万元）")
```



Hide

```
# 变量3：房屋装修情况的探索分析
## 数值描述
print("房屋装修情况的数值描述：")
```

```
## [1] "房屋装修情况的数值描述："
```

Hide

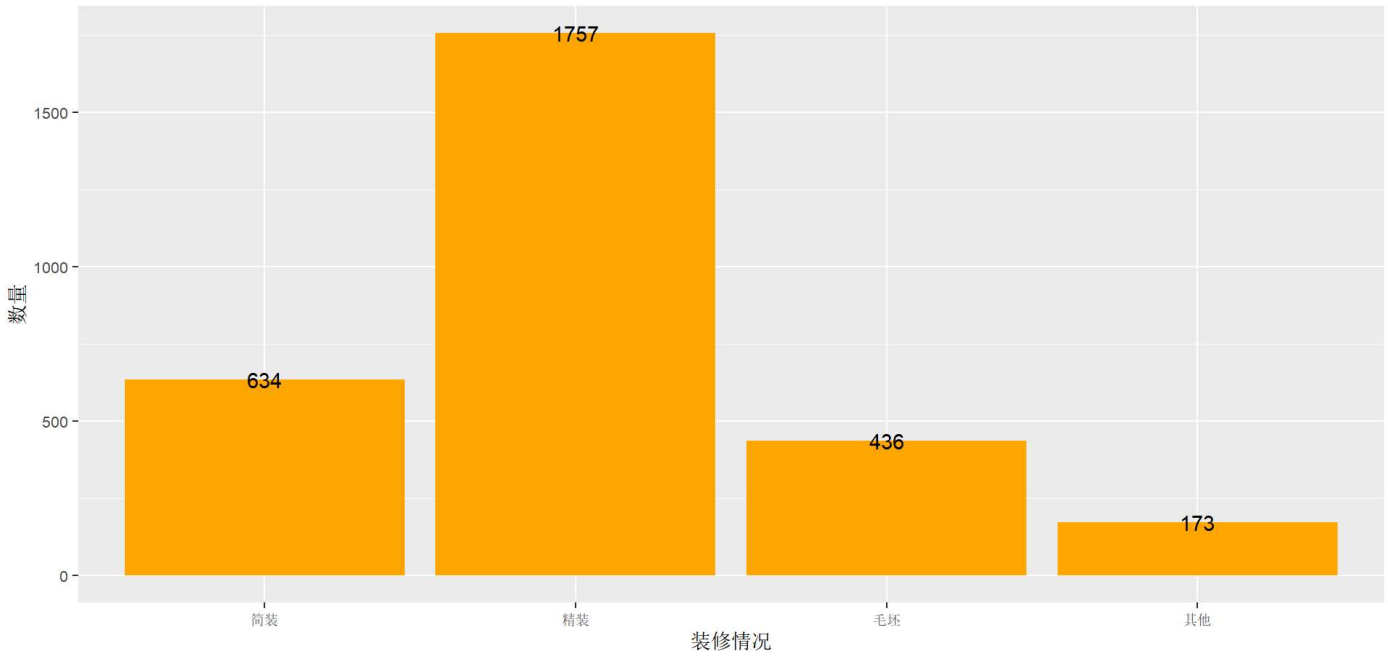
```
print(table(lj$decoration))
```

```
##
## 简装 精装 毛坯 其他
##  634 1757  436  173
```

Hide

```
## 图形绘制
# 绘制装修类型的柱状图
ggplot(lj, aes(decoration)) +
    geom_bar(fill = "orange") +
    ggtitle("装修类型分布柱状图") +
    xlab("装修情况") +
    ylab("数量")+
    geom_text(stat = "count", aes(label = after_stat(count)))
```

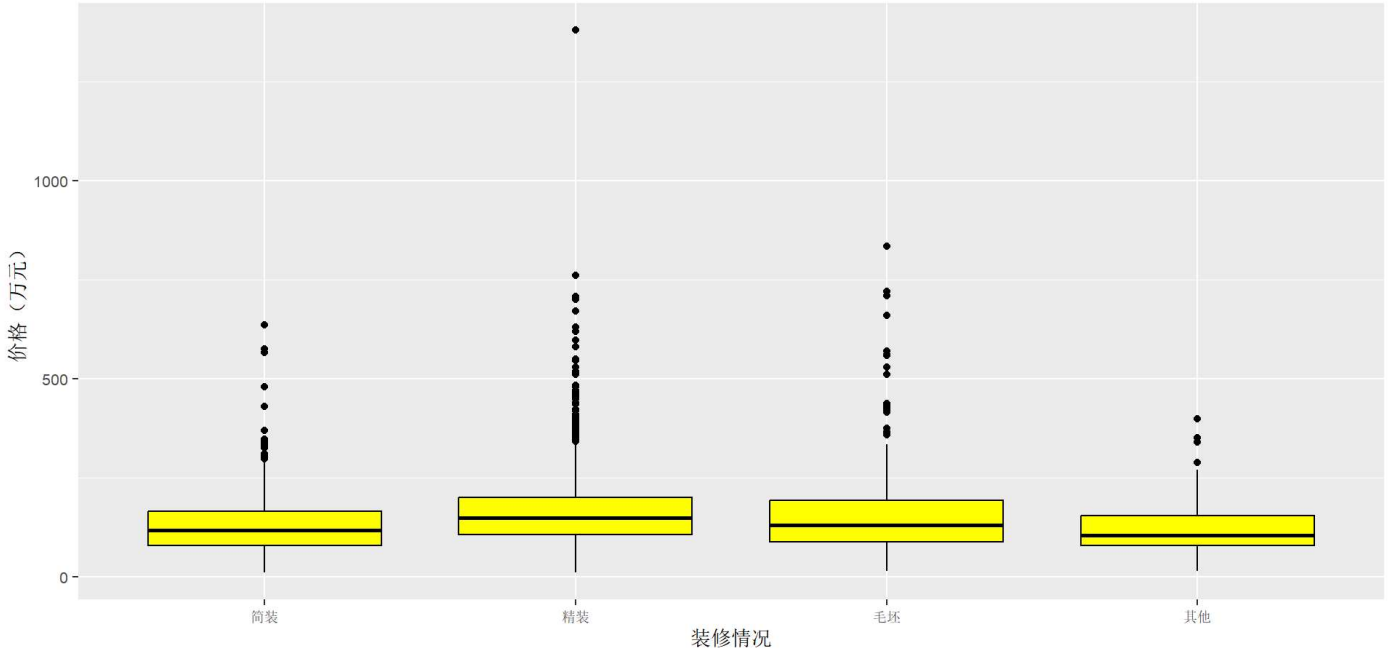
装修类型分布柱状图



Hide

```
# 分组分析不同装修程度的房屋价格
ggplot(lj, aes(decoration, price_ttl)) +
    geom_boxplot(color = "black", fill = "yellow") +
    ggtitle("不同装修程度的房屋价格箱线图") +
    xlab("装修情况") +
    ylab("价格（万元）")
```

不同装修程度的房屋价格箱线图



Hide

```
# 探索问题1：不同区域的房价差异原因

## 供需关系分析
print("不同区域的供需关系分析：")
```

```
## [1] "不同区域的供需关系分析："
```

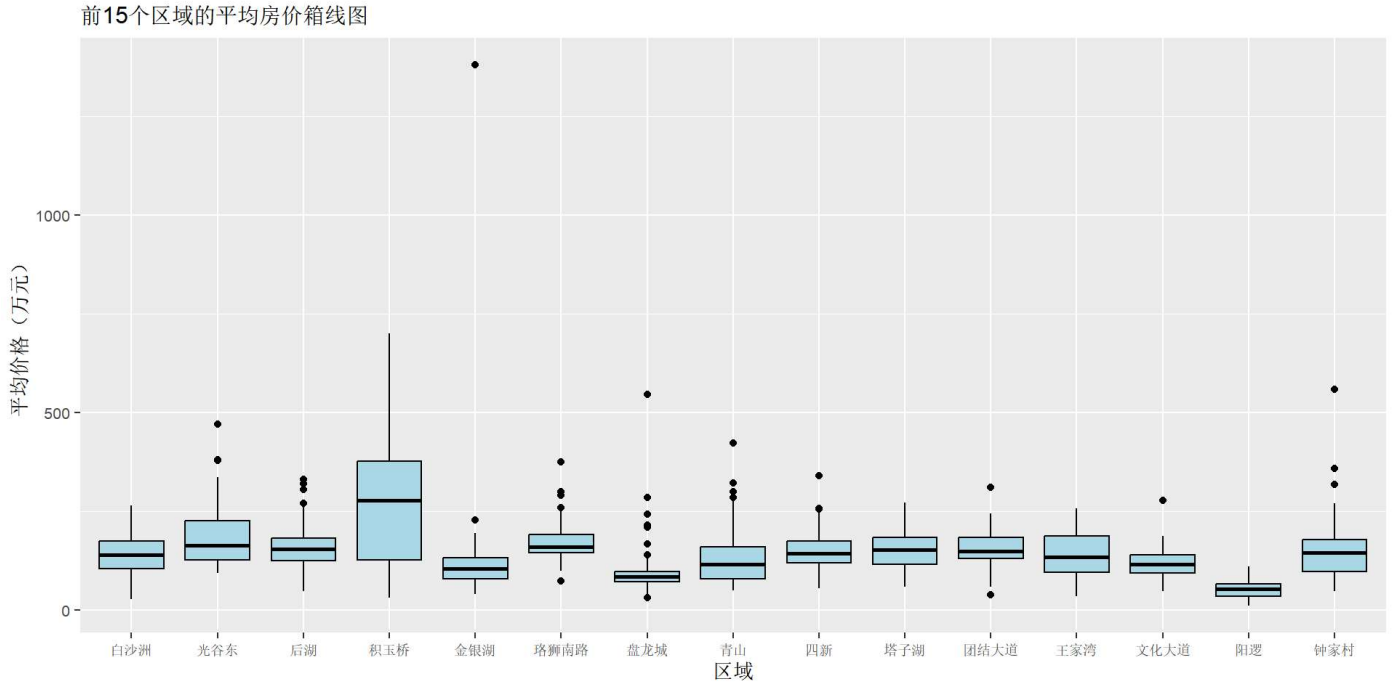
Hide

```
# 可以通过计算每个区域的房屋供应数量和需求数量来进行分析，这里假设没有明确的需求数据，仅作示例
region_supply <- lj %>%
    group_by(property_region) %>%
    summarize(supply = n())%>%
    arrange(desc(supply))%>%
    head(15)
region_supply_join <- region_supply %>%
       left_join(lj, by = "property_region")
region_supply_join
```

```
## # A tibble: 1,330 × 19
##    property_region supply property_name price_ttl price_sqm bedrooms livingrooms
##    <chr>            <int> <chr>              <dbl>     <dbl>    <dbl>       <dbl>
##  1 白沙洲             167 保利上城东区       194.      16336        3           2
##  2 白沙洲             167 保利新武昌K4…      215       20128        3           2
##  3 白沙洲             167 融科花满庭         259       16463        4           2
##  4 白沙洲             167 爱家名校华城        83.8      9279        3           2
##  5 白沙洲             167 喜瑞都             147       15776        3           2
##  6 白沙洲             167 保利新武昌K4…      181       17943        3           1
##  7 白沙洲             167 梅花苑              92        8871        3           2
##  8 白沙洲             167 花样年锦绣城…      126       13622        3           2
##  9 白沙洲             167 爱家名校华城        89        7967        3           2
## 10 白沙洲             167 保利新武昌K4…      235       19021        3           2
## # ℹ 1,320 more rows
## # 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>
```

Hide

```
# 绘制每个区域的箱线图
ggplot(region_supply_join, aes(x = property_region, y = price_ttl)) +
    geom_boxplot(color = "black", fill = "lightblue") +
    ggtitle("前15个区域的平均房价箱线图") +
    xlab("区域") +
    ylab("平均价格（万元）")
```



###发现总结 #通过上面的箱线图能够发现，供应量最多的15个区域中，价格中位数最高的是积玉桥区域，最低的是阳逻地区。