第一次作业你的报告题目

chenfan

2024-11-01

你的主要发现

1. 发现1：中北路、水果湖、黄埔永清、三阳路等核心城区平均房屋单价最高，普遍能超过2w，后湖、四新、金银湖等离核心城区有点距离的出售房子套数最多。
2. 发现2：楼栋总层数在30-35之间最多，其次为5-10。房屋位于低层的情况下，平均单价最高且价格范围最集中。
3. 发现3 房屋朝南的最多，朝南房屋单平米均价15170.30。在整体朝向和价格分析中并不是最高的，但朝南的房屋有很多大的离群点。

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。

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：

武汉链家二手房

property_name	property_region	price_ttl	price_sqm	bedrooms	livingrooms	building_area	directions1	directions2	decoration	property_t_height	property_height	property_sty
南湖名都A区 |南湖沃尔		237.										

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   :3.695   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 数据中有分类数据和数值数据
- 直观结论2 房价都是数值数据
- ...

探索性分析

数值描述:

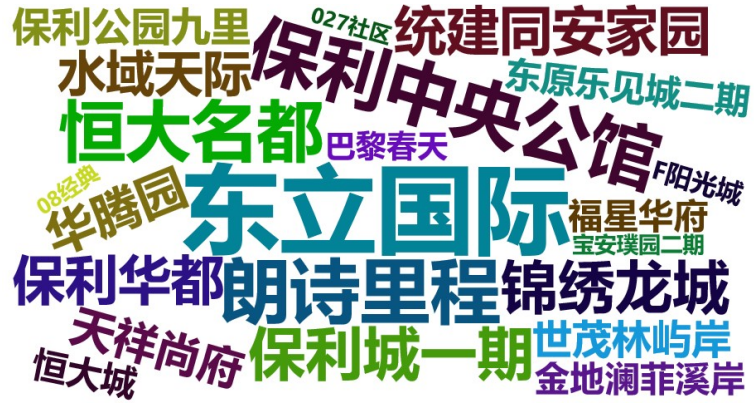
```
均值（Mean）：所有数值的总和除以数值的数量。
中位数（Median）：将一组数值按大小顺序排列后位于中间的数值。
众数（Mode）：一组数据中出现次数最多的数值。
方差（Variance）：衡量数据点与均值之间差异的平方的平均值。
标准差（Standard Deviation）：方差的平方根，衡量数据的离散程度。
极差（Range）：数据中最大值与最小值的差。
四分位数（Quartiles）：将数据分为四等分，每部分包含25%的数据点。
百分位数（Percentiles）：数据中低于某个百分比的数值点。
```

图形:

```
直方图（Histogram）：显示数据分布的图形，通常用于连续变量。
箱线图（Boxplot）：显示数据的中位数、四分位数和异常值的图形。
散点图（Scatter Plot）：显示两个变量之间关系的图形。
折线图（Line Chart）：显示数据随时间变化的图形。
条形图（Bar Chart）：显示不同类别之间比较的图形。
饼图（Pie Chart）：显示各部分占整体比例的图形。
热力图（Heatmap）：显示数据点密度的图形，通常用于展示两个变量之间的关系。
```

变量1的数值描述与图形

```
##    Length     Class      Mode
##      3000 character character
```



发现：

- 发现1 "东立国际"以最高的评量22位于榜首，这可能表明该小区在某些评价标准上表现突出，
- 发现2 多个小区的数量相同，例如"保利中央公馆"、"朗诗里程"都有16的评分，这可能意味着这些小区在评价标准上表现相似，或者在居民中的口碑相近。

变量2的数值描述与图形

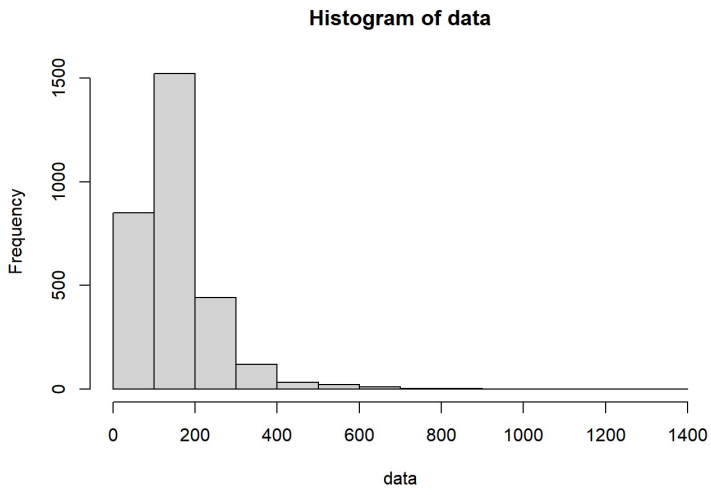
```
##    Length     Class      Mode
##      3000 character character
```

发现：

- 发现1 白沙洲的房子数量最多
- 发现2 后湖、四新、金银湖等离核心城区有点距离的占比较高

变量price_ttl的数值描述与图形

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

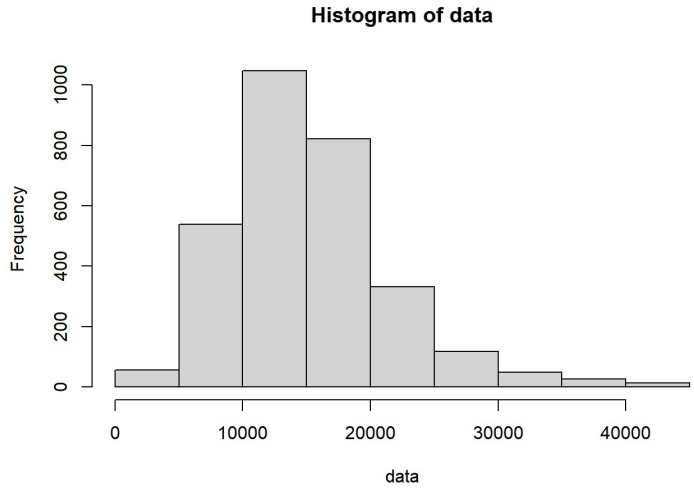


发现：

- 发现1 总价在200万元内的房子数量最多
- 发现2 房屋最高价格能到1380w

变量price_sqm的数值描述与图形

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

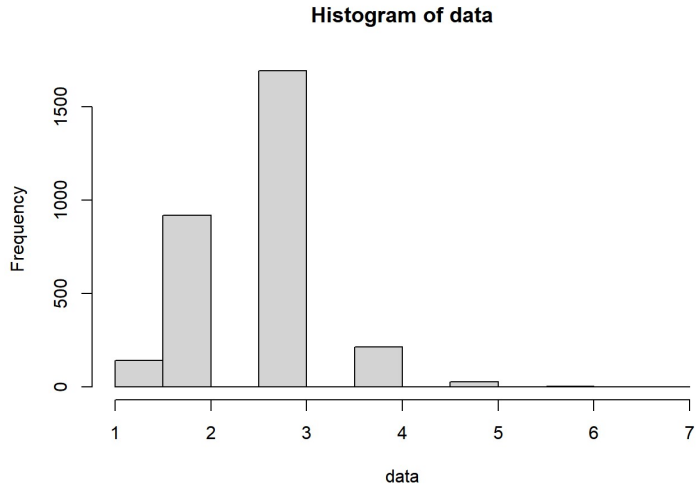


发现:

- 发现1 单价在10000至20000之间最多
- 发现2 正态房屋单价类似正态分布

变量bedrooms的数值描述与图形

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.695   3.000   7.000
```

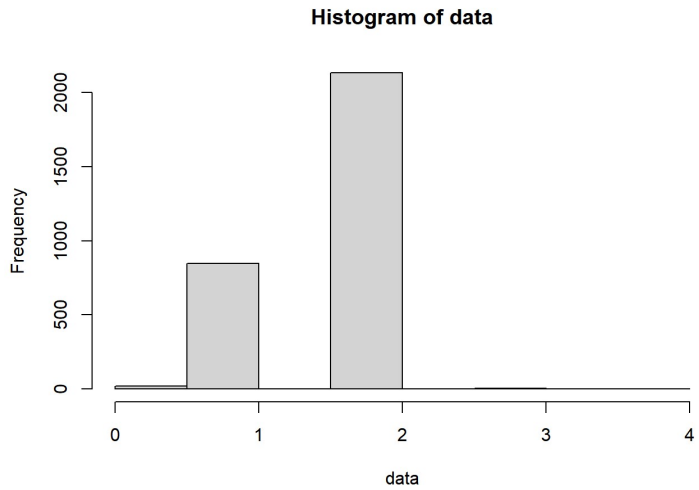


发现:

- 发现1 房间分布占比最多的为3间
- 发现2 多数房间数量在5以下

变量livingrooms的数值描述与图形

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   1.709   2.000   4.000
```

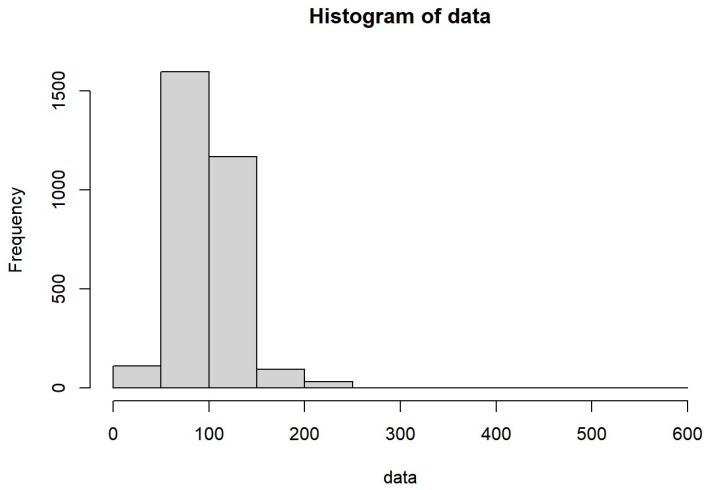


发现:

- 发现1 大部分客厅数在1或者2之间
- 发现2 客厅数最大为4间

变量building_area的数值描述与图形

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



发现:

- 发现1 房屋面积主要分布在50至150之间
- 发现2 最小的房屋面积仅22平方米

变量directions1的数值描述与图形

```
##    Length     Class      Mode
##      3000 character character
```

发现:

- 发现1 房屋主要朝向南方
- 发现2 房屋其次朝向东南

变量directions2的数值描述与图形

```
##    Length     Class      Mode
##      3000 character character
```

发现:

- 发现1 房屋次要朝向北方
- 发现2 房屋次要朝向其次朝向南方

变量decoration的数值描述与图形

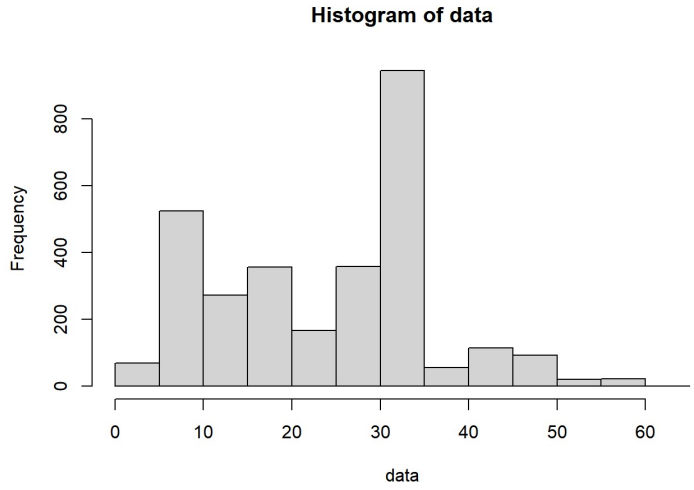
```
##    Length     Class      Mode
##      3000 character character
```

发现:

- 发现1 房屋装修分为精装、简装、毛坯、其他
- 发现2 房屋装修以精装为主

变量property_t_height的数值描述与图形

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   11.00   27.00   24.22   33.00   62.00
```



发现:

- 发现1 楼栋总层数在30-35之间最多，其次为5-10
- 发现2 楼栋总层数在5以下或35以上数量较少

变量property_height的数值描述与图形

```
##    Length     Class      Mode
##      3000 character character
```

发现:

- 发现1 房屋在所在楼栋所处位置，取值为高中低
- 发现2 房屋在中层最多

变量property_style的数值描述与图形

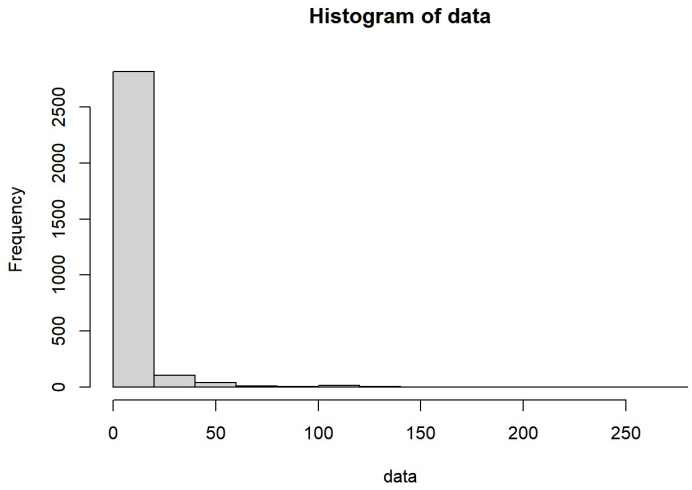
```
##    Length     Class      Mode
##      3000 character character
```

发现：

- 发现1 建筑形式有板楼、塔楼、板塔结合和暂无数据
- 发现2 板楼数量最多

变量followers的数值描述与图形

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   3.000   6.614   6.000 262.000
```



发现：

- 发现1 一半以上房屋的关注人数在3以下
- 发现2 少量房屋关注人数能超过20，最高能达到262

变量near_subway的数值描述与图形

```
##    Length     Class      Mode
##      3000 character character
```

发现：

- 发现1 是否靠近地铁仅有一个近地铁值出现，其余为空值
- 发现2 近地铁房屋有1554个

变量if_2y的数值描述与图形

```
##    Length     Class      Mode
##      3000 character character
```

```
##
## 房本满两年
##       1264
```

发现:

- 发现1 产证是否满2年仅有一个房本满两年 出现，其余为空值
- 发现2 房本满两年 房屋有1264个

变量has_key的数值描述与图形

```
##    Length     Class      Mode
##      3000 character character
```

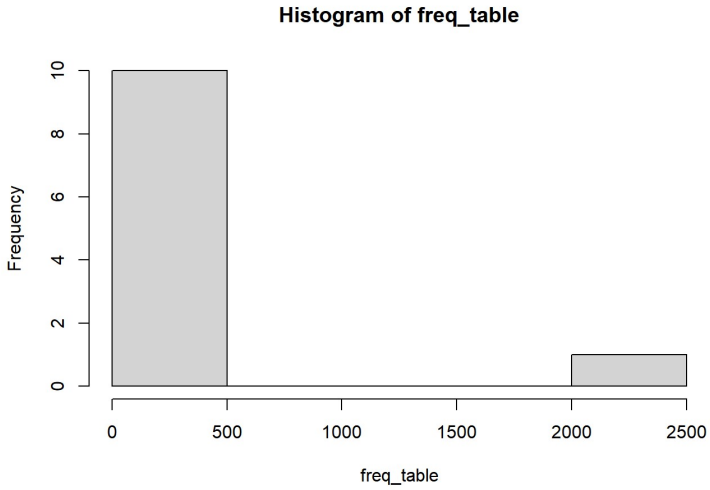
发现:

- 发现1 中介有钥匙房屋2525个
- 发现2 超过80%房屋有钥匙

变量vr的数值描述与图形

```
##    Length     Class      Mode
##      3000 character character
```

```
##
## VR\xe7\x8b\xac\xe5\x8c\xbb     随时看\xe6\x88\xbf     保利拉\xe8\x8f\xb2     育才花\xe6\xa1\x82
##                 1                  1                  1                  1
##           VR看装修              泓悦府             江景湾             近地铁
##               2084                  1                  1                  1
##           珞狮南路            随时看房             塔子湖
##                  1                  1                  1
```

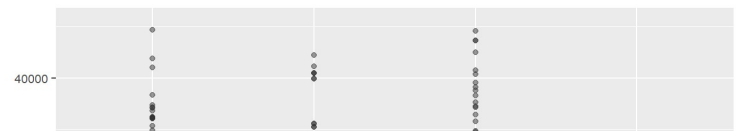


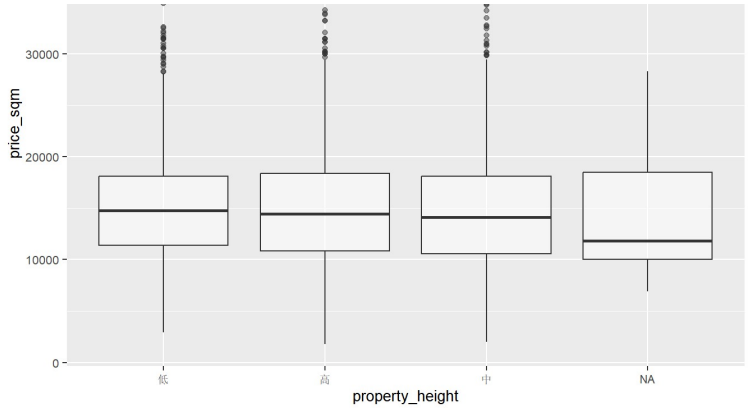
发现:

- 发现1 是否支持VR看房 存在一些无效数据
- 发现2 2084支持VR看房

探索问题1 房屋单价与房屋在所在楼栋所处位置的关系

```
## # A tibble: 4 × 3
##   property_height   num mean_price
##   <chr>           <int>      <dbl>
## 1 低                816     15379.
## 2 高                906     15195.
## 3 中               1218     14991.
## 4 <NA>               60     14520.
```



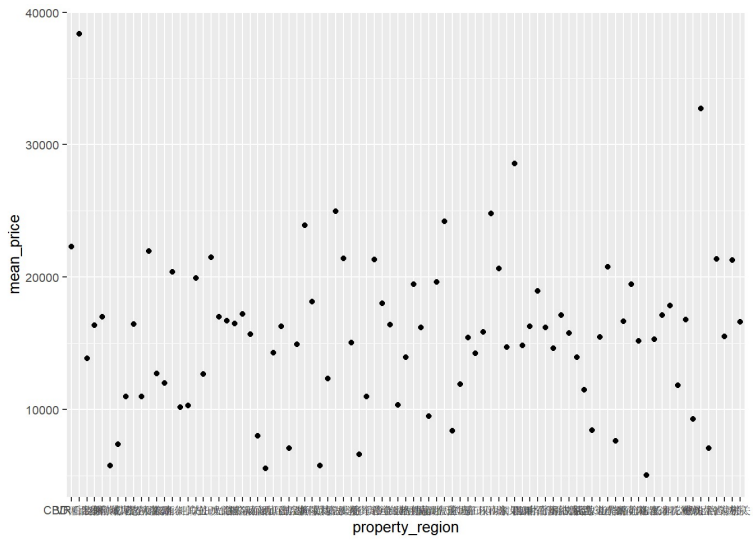


发现：

- 发现1 房屋位于低层的情况下，平均单价最高
- 发现2 房屋位于低层的情况下，价格范围最集中

探索问题2 房屋单价与所处区域的关系

```
## # A tibble: 87 × 3
##    property_region   num mean_price
##    <chr>           <int>      <dbl>
##  1 VR看装修            1     38351
##  2 中北路             18     32728.
##  3 水果湖              9     28562.
##  4 黄埔永清           23     24957.
##  5 三阳路             16     24777.
##  6 南湖沃尔玛         33     24181.
##  7 虎泉杨家湾         21     23902.
##  8 CBD西北湖          35     22272.
##  9 楚河汉街           15     21958.
## 10 关山大道           25     21480.
## # ℹ 77 more rows
```

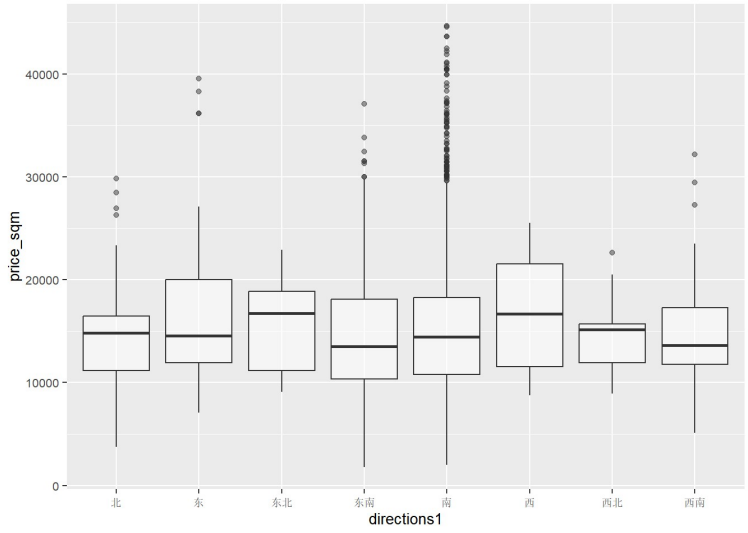


发现：

- 发现1 中北路、水果湖、黄埔永清、三阳路等核心城区平均房屋单价最高，普遍能超过2w
- 发现2 全部区域的平均房屋价格在15000到20000之间最多

探索问题3 房屋主要朝向和房屋单价之间的关系

```
## # A tibble: 8 × 3
##   directions1   num mean_price
##   <chr>       <int>      <dbl>
## 1 西             19     17221.
## 2 东             98     16156.
## 3 东北           10     15763.
## 4 南           2454     15170.
## 5 东南          281     14793.
## 6 西北           13     14557.
## 7 北             68     14462.
## 8 西南           57     14384.
```



发现：

- 发现1 房屋朝南的最多，朝南房屋单平米均价15170.30。在整体朝向和价格分析中并不是最高的，但朝南的房屋有很多大的离群点。
- 发现2 价格最集中的为朝向西北的房屋

发现总结

用1-3段话总结你的发现。中北路、水果湖、黄埔永清、三阳路等核心城区平均房屋单价最高，普遍能超过2w，后湖、四新、金银湖等离核心城区有点距离的出售房子套数最多。楼栋总层数在30-35之间最多，其次为5-10。房屋位于低层的情况下，平均单价最高且价格范围最集中。房屋朝南的最多，朝南房屋单平米均价15170.30。在整体朝向和价格分析中并不是最高的，但朝南的房屋有很多大的离群点。