张小庭 2023281051029 第一次作业

张小庭

目录

你的主要发现

1. 发现 1
2. 发现 2
3. 发现 3

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；

表 1: 武汉链家二手房

property_name	property_region	price_ttl	price_sqm	bedrooms	livingrooms	building_area	directio
南湖名都 A 区	南湖沃尔玛	237.0	18709	3	1	126.68	南
万科紫悦湾	光谷东	127.0	14613	3	2	86.91	南
东立国际	二七	75.0	15968	1	1	46.97	南
新都汇	光谷广场	188.0	15702	3	2	119.73	北
保利城一期	团结大道	182.0	17509	3	2	103.95	东南
加州橘郡	庙山	122.0	10376	3	2	117.59	南
省建筑五公司西区	光谷广场	99.0	12346	2	1	80.19	南
保利上城东区	白沙洲	193.8	16336	3	2	118.64	南
石化大院	中南丁字桥	325.0	32631	4	1	99.60	南
阳光花园	杨汊湖	192.0	17403	3	2	110.33	南

- 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_wuhan) 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：

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
```

##

##

##

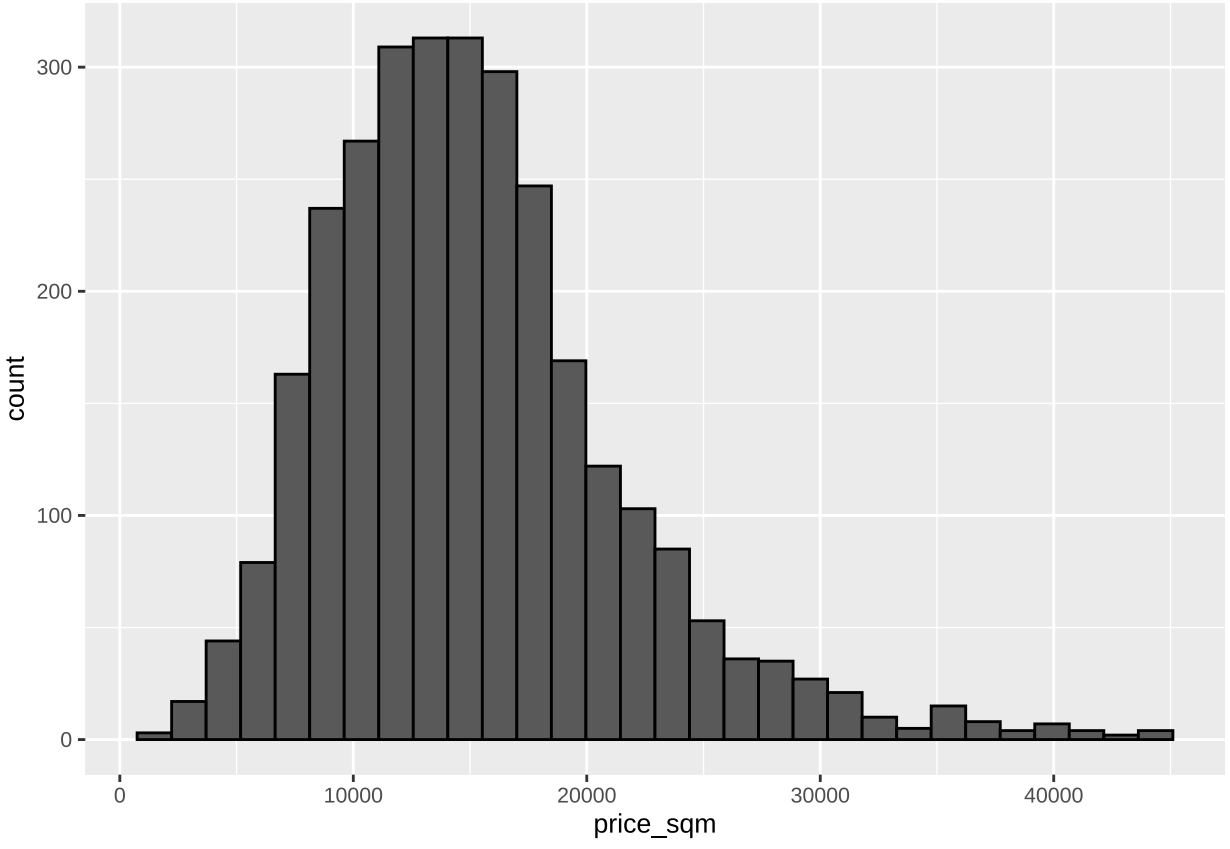
可以看到：

- 小庭观察 1：从数据源 lj_wuhan 数据摘要看到，与核心数据相关的 price_ttl 、price_sqm、bedrooms、livingrooms 、building_area 等字段总结出了该字段数据的最小值 min、第一分位值 1st Qu.、第三分位值 3rd Qu.、中位数 Median、平均值 Mean、最大值 Max，可便于我们从各个维度看到该字段的数据属性，便于相关数据分析人员作为相关决策依据
- 小庭观察 2：该数据中的所有字段数据是有 3000 行，但有些字段为空 NA，有可能该系统在采集数据时，这些字段如 vr 为非必填，因此未能采集到，若后续管理人员的决策需要依赖该字段，则前端系统在采集数据时需要注意信息采集的必要性。
- …

探索性分析

变量 1(二手房均价 price_sqm) 的数值描述与图形

发现：表中的二手房均价是 15148.49，大部分的大家集中在 10000~2000 之间，楼房最高价和最低价之间相差 42885

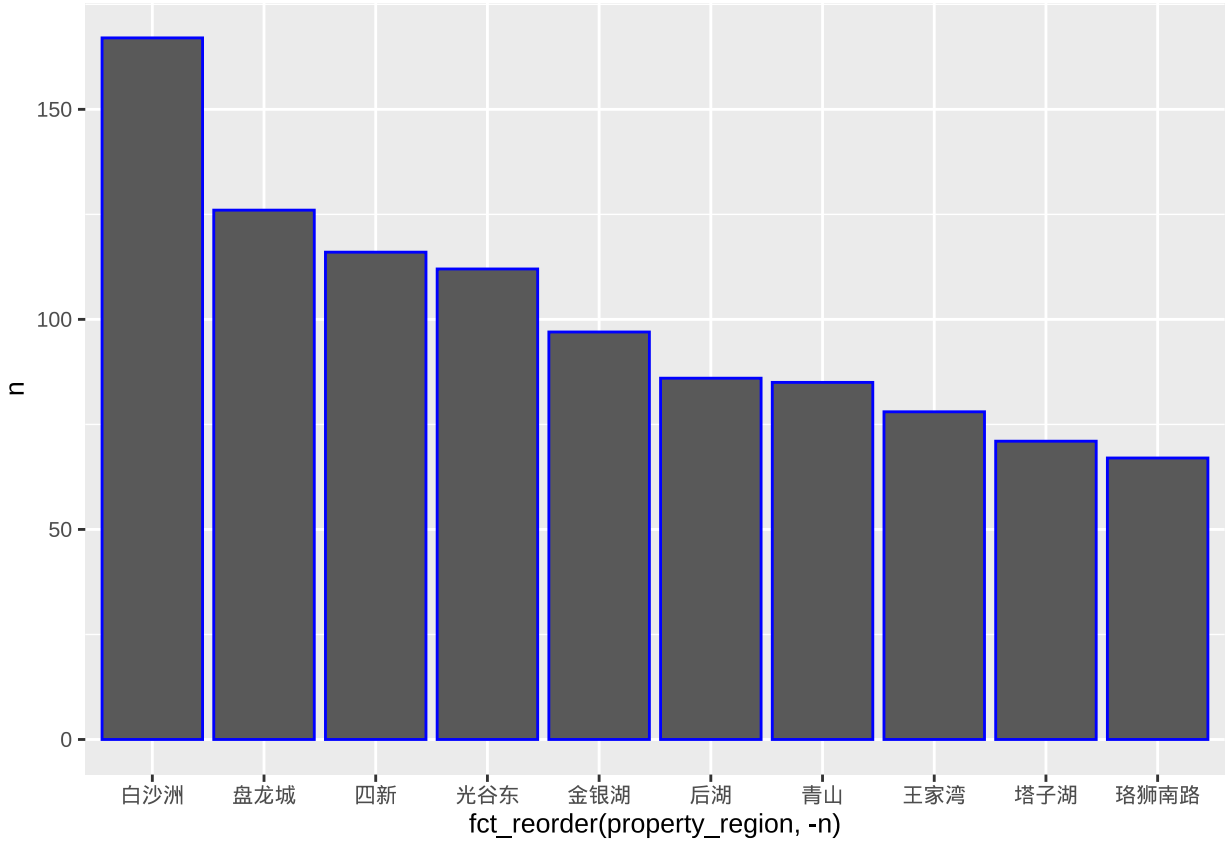


```
## [1] 15148.49
```

```
## [1] 42885
```

变量 2(二手房地区 property_region) 的数值描述与图形

- 发现 2：从武汉二手链家房屋数据，可以观察到武汉地区的房价排名前 10 的情况：武汉地区二手房放量最多的地区分别是：白沙洲、盘龙城、四新、光谷东、金银湖、后湖、青山、王家湾、塔子湖、罗氏南路，从武汉市整体地理位置来说，这些二手房量多的主要分散在武汉市中心外（可以根据 R 画出二手房量所处的地理位置，再划定二环、三环等市中心位置，来分析这些二手房量多的位置主要散落在哪里，由于 R 语言使用能力有限，目前暂时无法划出）



变量 3 的数值描述与图形

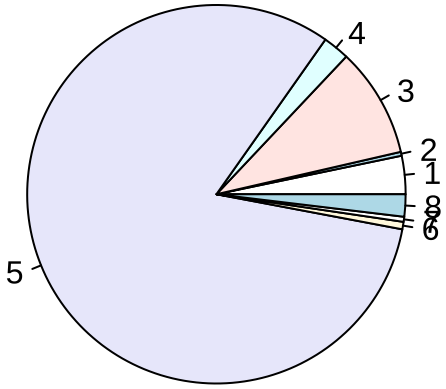
发现：81.8% 的二手房朝向为南，符合中国人对居住环境坐北朝南的居住要求，且相比于其他朝向，朝南的房屋均价高于其他朝向（这一条在 R 语言分析中一直报错，但用其他工具测算趋势是朝南的均价会相对高一点）

```
## # A tibble: 8 x 2
##   directions1     n
##   <chr>       <int>
## 1 南           2454
```

```
## 2 东南          281
## 3 东             98
## 4 北             68
## 5 西南           57
## 6 西             19
## 7 西北           13
## 8 东北           10
```

```
## # A tibble: 8 x 2
##   directions1     n
##   <chr>       <int>
## 1 东             98
## 2 东北           10
## 3 东南          281
## 4 北             68
## 5 南           2454
## 6 西             19
## 7 西北           13
## 8 西南           57
```

```
## # A tibble: 8 x 6
## # Groups:   directions1 [8]
##   directions1     n avg_ttl avg_sqm avg_building_area cx_por
##   <chr>       <int>   <dbl>   <dbl>             <dbl> <chr>
## 1 东             98    156.    1.51              101. 3.27%
## 2 东北           10    156.    1.51              101. 0.33%
## 3 东南          281    156.    1.51              101. 9.37%
## 4 北             68    156.    1.51              101. 2.27%
## 5 南           2454    156.    1.51              101. 81.80%
## 6 西             19    156.    1.51              101. 0.63%
## 7 西北           13    156.    1.51              101. 0.43%
## 8 西南           57    156.    1.51              101. 1.90%
```



探索问题 1

发现：房屋均价 price_spm 与是否精装修 decoration 之间的关系

- 发现 1: 精装修数量最多 1757 套占比 58.57%，其次是简装 637 套，占比 21.13，再次是毛坯 436 套，占比 14.53
- 发现 2：二手房精装修的房屋均价高于其他，每平方米多 2000 左右，

```
## # A tibble: 4 x 4
##   decoration deco_mean deco_count  deco_por
##   <chr>          <dbl> <table[1d]> <chr>
## 1 精装          16077. 1757        58.57%
## 2 简装          13993.  634        21.13%
## 3 毛坯          13819.  436        14.53%
## 4 其他          13304.  173        5.77%
```

探索问题 2

发现：房屋均价 price_spm 与楼层 property_height 之间的关系

- 发现 1: 均价最高的是低楼层 15378.75，其次是高楼层 15194.52，再次是中楼层均价是 14990.94（这与我常识的设想是不同的，在我的假设中高楼层应该是价格最高，低楼层的价格最低）
- 发现 2：楼层高低的价格与供应量有关，供应量最大的中楼层均价最低，供应量最小的低楼层均价最高，符合经济学常识中的供应量与价格的关系

```
## # A tibble: 3 x 4
## # Groups:   property_height [3]
##   property_height height_mean height_count height_por
##   <chr>                 <dbl> <table[1d]>  <chr>
## 1 低                   15379.  816         27.20%
## 2 高                   15195.  906         30.20%
## 3 中                   14991. 1218         40.60%
```

探索问题 3

发现：房屋均价价 price_spm 与近地铁 near_subway 属性之间的关系

- 发现 1
- 发现 2

发现总结

二楼房中涉及的多个变量，我更加关注地区、均价两个指标从数据中可以探索发现均价与诸多住房相关属性如供给量，所属楼层，所在地域，是否是否靠近地铁等