第一次作业-武汉链家数据简要分析

刘振强　　2023 年 10 月 19 日

目录

1 主要发现

发现 1：武昌区、江汉区、洪山区平均房价最高，黄陂、蔡甸、新洲平均房价最低。

发现 2：房间数量和总价格具有较强的正向关联性。

发现 3：精装的有助于二手房出售（相对简装），毛坯价格与精装价格差异较小，说明很多客户青睐二手毛坯房，可以按照自己喜欢风格装修。

发现 4：靠近地铁的房子价格显著高于远离地铁的房子价格，可能有其他因素影响（例如学区、商场及其他配套资源），还需要进一步分析。靠近地铁的二手房占比接近 52%，侧面说明武汉市的地铁覆盖率比较高。

发现 5：1-4 个房间数量与楼层数量关系不明显，5-7 个房间基本分布在中矮楼层。

发现 6：81.8% 的房子朝向向南 (主要因为采光及风俗习惯的原因)。

发现 7：二手房出售中，开发商为保利的数量最多，达到 183，约占 6%，原因在于保利在南湖片区体量较大；开发商为万科、金地、联投等数量紧随其后。

发现总结：中心城区房价普遍偏高；武汉市城市发展较快，地铁覆盖率较高；保利、万科等在武汉的楼盘较多。

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。

3 数据分析

3.1 描述性统计

```
library(psych)
data_lj <- read.csv("/Users/lzq/Course/1st-assignment-main-2/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
```

```
# summary(data_lj )

describe(data_lj )  # 另外一种展示方式
```

```
##                   vars    n     mean      sd   median  trimmed     mad     min
## property_name*       1 3000   662.66  394.85   657.00   661.33  535.22    1.00
## property_region*     2 3000    45.34   24.09    48.00    45.67   29.65    1.00
## price_ttl            3 3000   155.86   95.55   137.00   142.48   66.72   10.60
## price_sqm            4 3000 15148.49 6323.18 14404.00 14579.25 5465.60 1771.00
## bedrooms             5 3000     2.70    0.73     3.00     2.68    0.00    1.00
## livingrooms          6 3000     1.71    0.47     2.00     1.77    0.00    0.00
## building_area        7 3000   100.87   30.38    95.54    99.59   23.27   22.77
## directions1*         8 3000     4.72    1.03     5.00     4.90    0.00    1.00
## directions2*         9 3000     2.82    2.09     1.00     2.70    0.00    1.00
## decoration*         10 3000     3.32    0.93     4.00     3.48    0.00    1.00
```

```
## property_t_height   11 3000    24.22   12.45    27.00    23.84   11.86    2.00
## property_height*    12 2940     1.89    0.84     2.00     1.87    1.48    1.00
## property_style*     13 3000     4.04    1.48     5.00     4.30    0.00    1.00
## followers           14 3000     6.61   15.22     3.00     3.53    2.97    0.00
## near_subway*        15 1559     4.99    0.17     5.00     5.00    0.00    1.00
## if_2y*              16 1264     1.00    0.00     1.00     1.00    0.00    1.00
## has_key*            17 2542     7.98    0.34     8.00     8.00    0.00    1.00
## vr*                 18 2094     1.03    0.43     1.00     1.00    0.00    1.00
##                        max    range   skew kurtosis     se
## property_name*     1345.00  1344.00   0.02    -1.27   7.21
## property_region*     87.00    86.00  -0.12    -1.23   0.44
## price_ttl          1380.00  1369.40   2.75    16.12   1.74
## price_sqm         44656.00 42885.00   1.08     2.03 115.44
## bedrooms              7.00     6.00   0.14     1.64   0.01
## livingrooms           4.00     4.00  -0.99    -0.18   0.01
## building_area       588.66   565.89   2.08    23.64   0.55
## directions1*          8.00     7.00  -1.35     5.44   0.02
## directions2*          9.00     8.00   0.38    -1.52   0.04
## decoration*           4.00     3.00  -1.12     0.08   0.02
## property_t_height    62.00    60.00   0.05    -0.80   0.23
## property_height*      3.00     2.00   0.20    -1.57   0.02
## property_style*       5.00     4.00  -1.36     0.20   0.03
## followers           262.00   262.00   6.90    68.17   0.28
## near_subway*          5.00     4.00 -20.66   443.97   0.00
## if_2y*                1.00     0.00    NaN      NaN   0.00
## has_key*              9.00     8.00 -18.09   344.47   0.01
## vr*                  11.00    10.00  18.23   352.74   0.01
```

分析结果：1. 通过 summary() 函数可以看到最小值、1/4 分位值、众数、均值、3/4 分位值等。2. 通过 describe() 函数查看均值、众数、修剪均值、平均绝对偏差（MAD，Mean Absolute Deviation）、偏度、峰度等数据，例如房屋总价数据，均值为 155.86 万元，最小为 10.6 万元，最大为 1380 万元，众数为 137 万元，按照一定的比例或数量（可定义），将数据集中的极端值去除，剔除后均值为 142.48 万元；峰度（kurtosis）大于 3，表示数据分布的峰度较高（尖峰），即数据集中的值较集中。

3.2 哪个地区的平均房价最高

```
library(dplyr)
```

```
##
```

```
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data_lj <- read.csv("/Users/lzq/Course/1st-assignment-main-2/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
```

```
# avg_price_by_region <- data_lj %>% group_by(property_region) %>% summarise(avg_price = mean(pric
data_lj <- data_lj %>%
  mutate(property_region = case_when(
    property_region %in% c(" 常青花园", " 将军路",  " 吴家山",  " 东西湖其他") ~ " 东西湖区",
    property_region %in% c(" 王家湾",  " 钟家村",  " 四新",  " 七里庙") ~ " 汉阳区" ,
    property_region %in% c(" 楚河汉街",  " 东湖东亭", " 积玉桥", " 南湖沃尔玛",  " 沙湖",  " 首义", "
    property_region %in% c(" 青山" ) ~ " 青山区",
    property_region %in% c(" 宝丰崇仁", " 古田", " 汉正街", " 集贤", " 长丰常码头", " 宗关") ~ " 硚口
    property_region %in% c(" 藏龙岛", " 光谷南", " 金融港", " 庙山", " 纸坊", " 江夏其他") ~ " 江夏区
    property_region %in% c("CBD 西北湖", " 常青路", " 前进江汉", " 唐家墩", " 武广万松园", " 新华路万
    property_region %in% c(" 百步亭", " 大智路", " 堤角", " 二七", " 国际百纳", " 后湖", " 黄埔永清",
    property_region %in% c(" 汉口北", " 盘龙城", " 武湖", " 黄陂其他") ~ " 黄陂区",
    property_region %in% c(" 白沙洲", " 关山大道", " 关西长职", " 光谷东", " 光谷广场", " 虎泉杨家湾",
    property_region %in% c(" 蔡甸城区", " 蔡甸其它", " 中法生态城",  " 后官湖") ~ " 蔡甸区",
    property_region %in% c(" 阳逻") ~ " 新洲区" ,
    property_region %in% c(" 沌口") ~ " 经开区" ,
    property_region %in% c(" 汉南其他") ~ " 汉南区" ,
    TRUE ~ "Other"
    ))
avg_price_by_region <- data_lj %>%
  group_by(property_region) %>%
  summarise(avg_price = mean(price_ttl)) %>%
  arrange(desc(avg_price))

# 按区域分类汇总
print(avg_price_by_region)
```

```
## # A tibble: 14 x 2
##    property_region avg_price
##    <chr>                <dbl>
##  1 武昌区               218.
##  2 江汉区               173.
##  3 洪山区               170.
##  4 江岸区               168.
##  5 汉阳区               145.
##  6 Other                141.
##  7 硚口区               139.
##  8 青山区               134.
##  9 江夏区               134.
## 10 东西湖区             122.
## 11 经开区               100.
## 12 黄陂区                91.2
## 13 蔡甸区                70.1
## 14 新洲区                52.4
```

分析结果：武昌区、江汉区、洪山区平均房价最高，黄陂、蔡甸、新洲平均房价最低。

3.3 房间数量和总价格之间关系

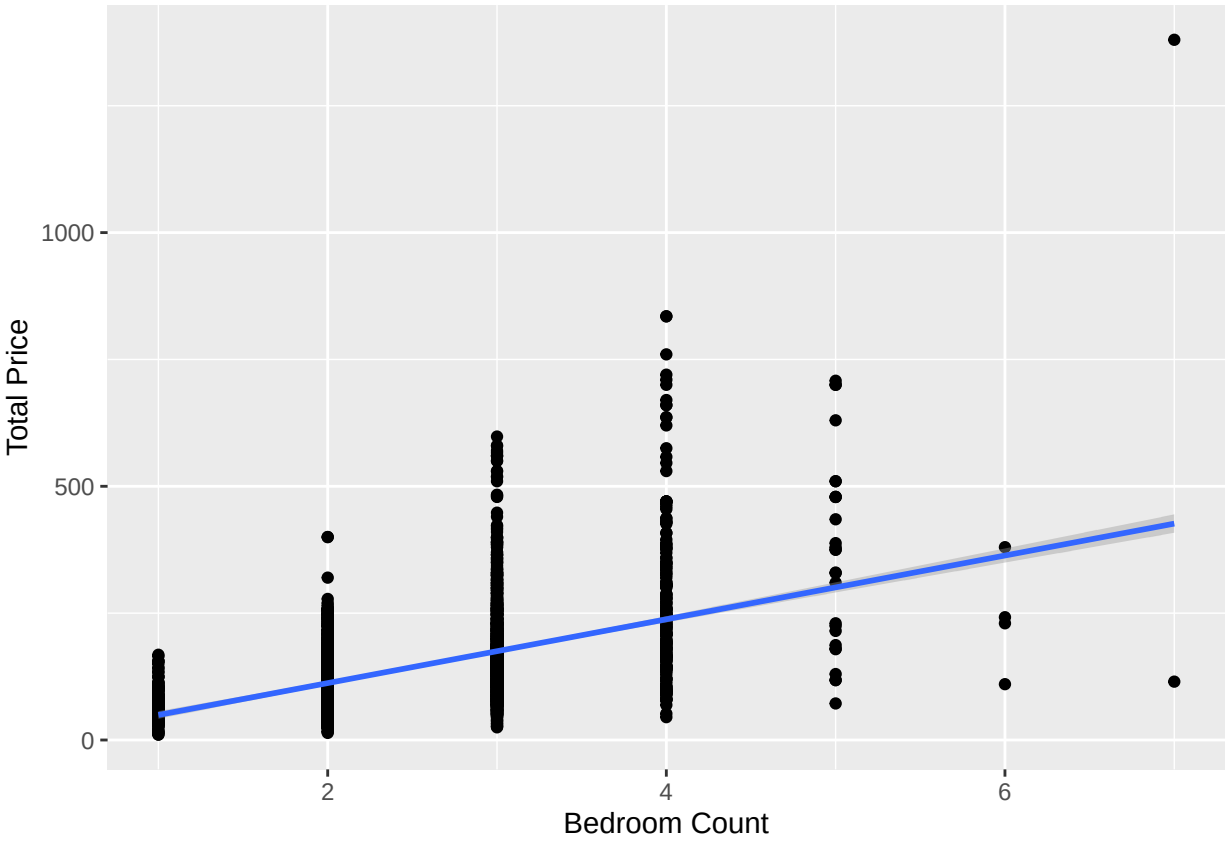
```
library(dplyr)
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
ggplot(data_lj, aes(x = bedrooms, y = price_ttl)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Bedroom Count", y = "Total Price")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



分析

结果：房间数量和总价格具有较强的正向关联性。

3.4 装修情况是否影响房价

```
library(dplyr)
library(ggplot2)

avg_price <- data_lj %>% group_by(decoration) %>% summarise(avg_price = mean(price_ttl))

print(avg_price)
```

```
## # A tibble: 4 x 2
##   decoration avg_price
##   <chr>          <dbl>
## 1 其他            124.
## 2 毛坯            161.
## 3 简装            134.
## 4 精装            166.
```

```
# 生成柱状图
#ggplot( avg_price, aes(x = decoration, y = avg_price)) +   geom_bar(stat = "identity", fill = "st
```

分析结果：精装的有助于二手房出售（相对简装），毛坯价格与精装价格差异较小，说明很多客户青睐二手毛坯房，可以按照自己喜欢风格装修。

3.5 靠近地铁的房子是否比远离地铁的房子更贵

```
library(dplyr)
data_lj %>%
  mutate(near_subway = ifelse(is.na(near_subway), "No", "Yes")) %>%
  group_by(near_subway) %>%
  summarise(avg_price = mean(price_ttl))
```

```
## # A tibble: 2 x 2
##   near_subway avg_price
##   <chr>           <dbl>
## 1 No               141.
## 2 Yes              170.
```

```
data_lj %>%
  mutate(near_subway = ifelse(is.na(near_subway), "No", "Yes")) %>%
  group_by(near_subway) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)
```

```
## # A tibble: 2 x 3
##   near_subway count percentage
##   <chr>       <int>      <dbl>
## 1 No           1441       48.0
## 2 Yes          1559       52.0
```

分析结果：1. 靠近地铁的房子价格显著高于远离地铁的房子价格，可能有其他因素影响（例如学区、商场及其他配套资源），还需要进一步分析。2. 靠近地铁的二手房占比接近 52%，侧面说明武汉市的地铁覆盖率比较高。

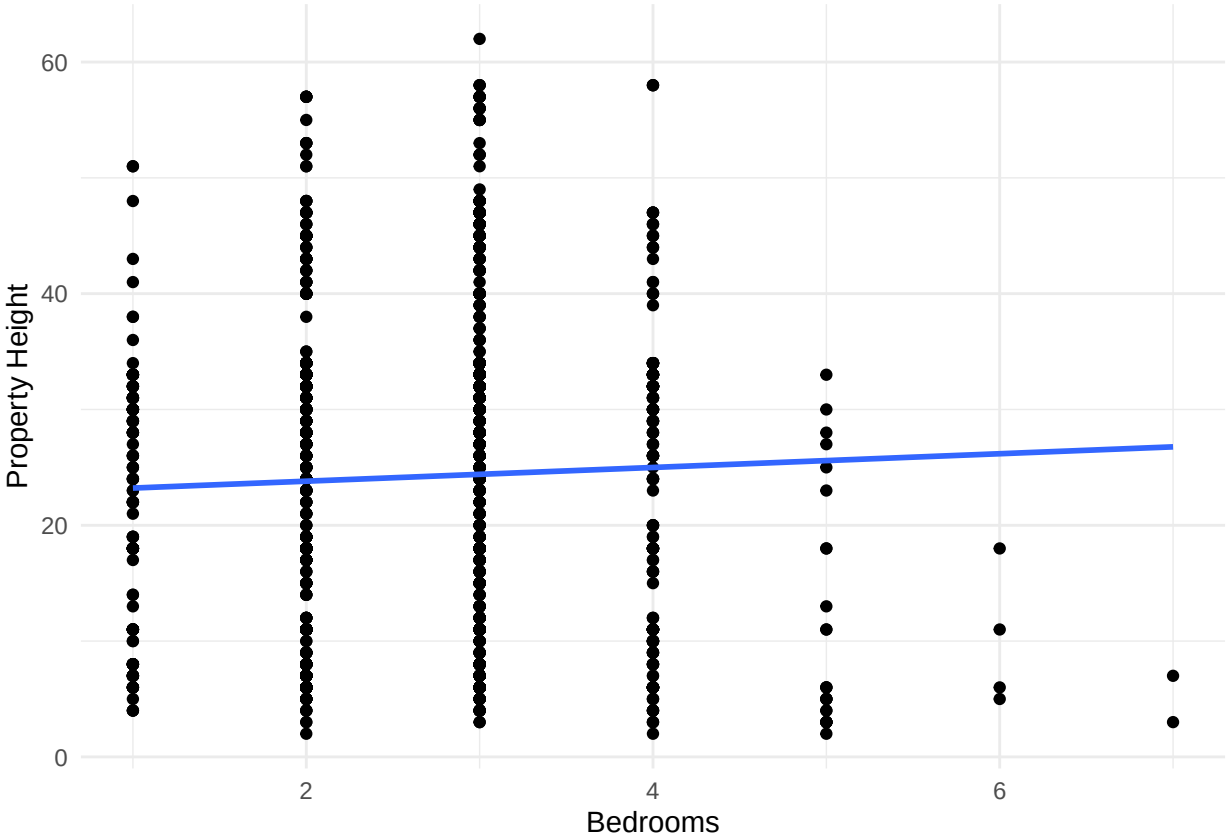
3.6 房间数量与楼层数量是否存在趋势关系

```
library(ggplot2)
# 创建散点图
data_lj <- read.csv("/Users/lzq/Course/1st-assignment-main-2/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
```

```
ggplot(data_lj, aes(x = bedrooms, y = property_t_height)) +
  geom_point() +
  labs(x = "Bedrooms", y = "Property Height") +
  theme_minimal() +
  geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



分析

结果：1-4 个房间数量与楼层数量关系不明显，5-7 个房间基本分布在中矮楼层

3.7 房屋朝向分布

```
data_lj <- read.csv("/Users/lzq/Course/1st-assignment-main-2/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
```

```
prop_directions1 <- prop.table(table(data_lj$directions1))
print(prop_directions1)
```

```
##
##          东         东北         东南          北          南          西
## 0.032666667 0.003333333 0.093666667 0.022666667 0.818000000 0.006333333
##         西北         西南
## 0.004333333 0.019000000
```

```
# 计算频数
#freq <- table(data$directions1)
#print(freq)
```

分析结果：81.8% 的房子朝向向南 (主要因为采光及风俗习惯的原因)

3.8 那个房企挂牌的二手房数量最多

```
library(dplyr)
data_lj <- read.csv("/Users/lzq/Course/1st-assignment-main-2/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
```

```
xiaoqu_grouped <- data_lj %>%
  group_by(property_name) %>%
  summarise(count = n())


# 筛选出'name'包含" 万科" 和" 保利" 等字样的数据
vanke_count <- xiaoqu_grouped[grepl(" 万科", xiaoqu_grouped$property_name), ]$count %>% sum()
poly_count  <- xiaoqu_grouped[grepl(" 保利", xiaoqu_grouped$property_name), ]$count %>% sum()
jindi_count  <- xiaoqu_grouped[grepl(" 金地", xiaoqu_grouped$property_name), ]$count %>% sum()
hengda_count <- xiaoqu_grouped[grepl(" 恒大", xiaoqu_grouped$property_name), ]$count %>% sum()
fudi_count <- xiaoqu_grouped[grepl(" 复地", xiaoqu_grouped$property_name), ]$count %>% sum()
```

```r
liantou_count  <- xiaoqu_grouped[grepl(" 联投", xiaoqu_grouped$property_name), ]$count %>% sum()
biguiyuan_count  <- xiaoqu_grouped[grepl(" 碧桂园", xiaoqu_grouped$property_name), ]$count %>% sum(
rongchuang_count  <- xiaoqu_grouped[grepl(" 融创", xiaoqu_grouped$property_name), ]$count %>% sum()
# 在使用%like% 操作符进行字符串匹配时，使用"%pattern%" 来匹配任意包含模式字符串的文本;
# biguiyuan_count <- xiaoqu_grouped[xiaoqu_grouped$property_name %like% " 碧桂园", ]$count  %>% sum
# SQL 中的 LIKE 操作符进行模式匹配

# 输出结果
print(paste(" 保利数量: ", poly_count))
```

```
## [1] "保利数量:  183"
```

```r
print(paste(" 万科数量: ", vanke_count))
```

```
## [1] "万科数量:  81"
```

```r
print(paste(" 金地数量: ", jindi_count))
```

```
## [1] "金地数量:  76"
```

```r
print(paste(" 联投数量: ", liantou_count))
```

```
## [1] "联投数量:  49"
```

```r
print(paste(" 恒大数量: ", hengda_count))
```

```
## [1] "恒大数量:  45"
```

```r
print(paste(" 复地数量: ", fudi_count))
```

```
## [1] "复地数量:  15"
```

```r
print(paste(" 碧桂园数量: ", biguiyuan_count))
```

```
## [1] "碧桂园数量:  25"
```

```r
print(paste(" 融创数量: ", rongchuang_count))
```

```
## [1] "融创数量:  16"
```

分析结果：二手房出售中，开发商为保利的数量最多，达到 183，约占 6%，原因在于保利在南湖片区体量较大；开发商为万科、金地、联投等数量紧随其后。