First Homework 2023281051031

Code

Legolas

2023-10-19

1 主要发现

1. 发现1:白沙洲、盘龙城、四新等地区二手房数量位列前三
2. 发现2:300万以上二手房逐渐减少
3. 发现3:毛坯二手房，临近地铁，面积越大单价越高；而非地铁房，150面积时单价最高，而后面积与单价呈反比

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

```
lj1<- read_csv("2023-09-12_cleaned.csv")
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 3000 Columns: 18
## ── Column specification ────────────────────────────────────────────────────
────────────────────────────────────────
## Delimiter: ","
## chr (11): property_name, property_region, directions1, directions2, decorati...
## dbl  (7): price_ttl, price_sqm, bedrooms, livingrooms, building_area, proper...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Hide

```
theme_set(theme(text = element_text(family="",size = 10)))
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

3 数据概览

数据表 (1j1) 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

```
head(1j1[,1:10])
```

```
## # A tibble: 6 × 10
##   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
## # ℹ 4 more variables: building_area <dbl>, directions1 <chr>,
## #   directions2 <chr>, decoration <chr>
```

各变量的简短信息：

Hide

```
glimpse(lj1)
```

```
## 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(lj1)
```

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

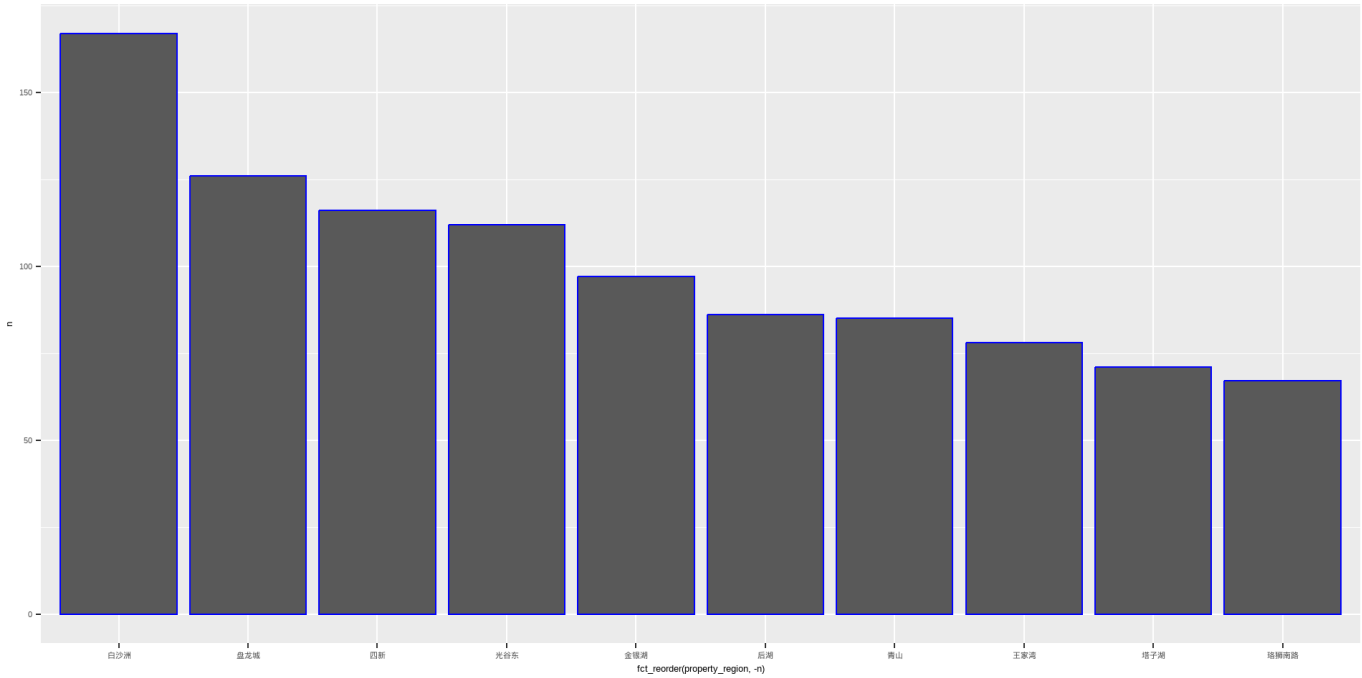
可以看到：

- 直观结论1 当前数据集中有11个字符型变量，7个数字变量
- 直观结论2 总价在10.6万元至1380万之间，其中平均总价为155.9万元，总价中位数为137万元
- 直观结论3

4 探索性分析

Hide

```
lj1 %>%
  group_by(property_region)%>%
  summarise(n=n()) %>%
  arrange(desc(n))%>%
  head(10)%>%
ggplot(aes(fct_reorder(property_region,-n),n))+
  geom_col(color="blue")
```



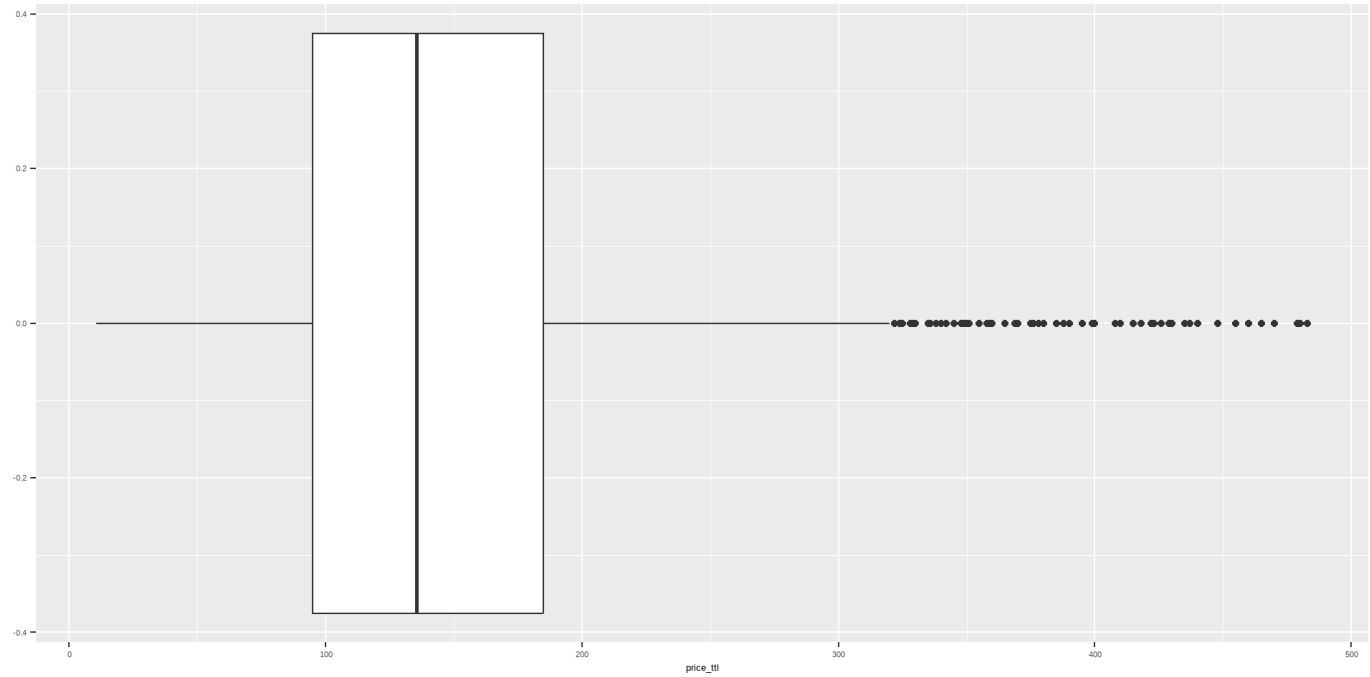
4.1 变量区域的数值描述与图形

发现:

- 发现1：白沙洲、盘龙城、四新等地区二手房数量位列前三
- 发现2：

Hide

```
lj_filtered<- lj1%>%
  filter(lj1[["price_ttl"]]<=500)
ggplot(lj_filtered,aes(x=price_ttl))+geom_boxplot()
```



4.2 变量总价的数值描述与图形

发现:

- 发现1：总价500万以内，中位数在140万左右;
- 发现2：300万以上二手房逐渐减少

4.3 变量总价的数值描述与图形

4.4 探索问题:

发现:

- 发现1
- 发现2

Hide

```
ggplot(data=lj_filtered)+
  geom_smooth(mapping=aes(x=building_area,y=price_sqm))+
  facet_grid(decoration~near_subway)
```

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

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : span too small.  fewer data values than degrees of freedom.
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : at 85.321
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : radius 0.014137
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : all data on boundary of neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at 85.321
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 0.1189
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 1
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : at 109.34
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : radius 0.014137
```

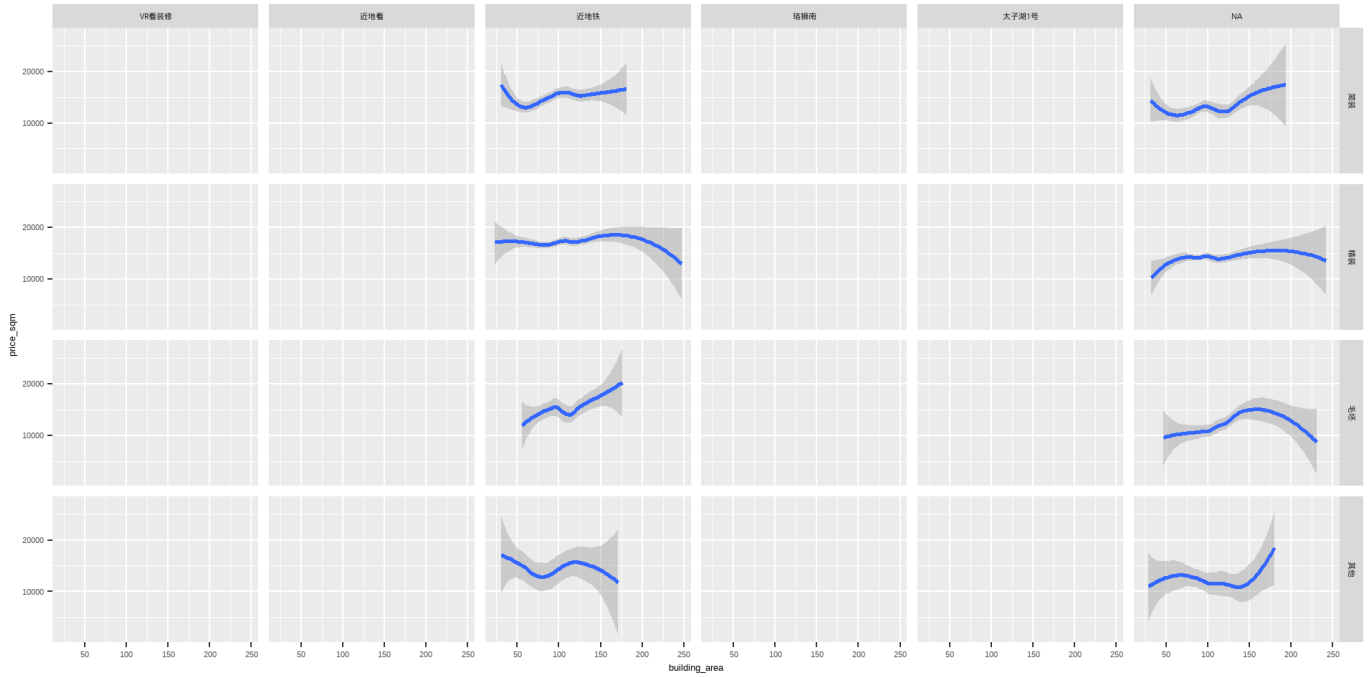
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : all data on boundary of neighborhood. make span bigger
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 0.014137
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : zero-width neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : zero-width neighborhood. make span bigger
```

```
## Warning: Computation failed in `stat_smooth()`
## Caused by error in `predLoess()`:
## ! 外接函数调用时不能有NA/NaN/Inf(arg5)
```



4.5 探索问题2：装修与近地铁分类，房屋面积与单价关系

发现：

- 发现1：精装条件下，无论是否靠近地铁，单价先与房屋面积关联较小，后随着面积增大而逐渐下降
- 发现2：毛坯二手房，临近地铁，面积越大单价越高；而非地铁房，150面积时单价最高，而后面积与单价呈反比