关于链家武汉二手房网站的数据分析报告

黄亚昭-2023281051041-MEM

目录

摘要

发现 1：总体来看，房屋总价与房屋单价呈正相关性，房屋总价随着房屋单价的增加而增加。按照房屋建筑形式（即房屋功能）划分，可以看出房屋总价与房屋单价仍呈正相关性，房屋总价随着房屋单价的增加而增加。

发现 2：通过分析可以看出，一是塔板结合（商住两用）房屋单价最高，二是板楼（住宅）房屋单价高于塔楼（商业）房屋单价，这与土地市场上住宅用地亩单价高于商业用地亩单价情况一致。

发现 3：通过观察，房屋单价与区域热销程度没有很强的关联性。

数据介绍

本报告**链家**数据获取方式如下：

报告人通过老师安排的作业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；

说明：数据仅用于本次作业；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。

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

第一部分：载入数据

```
lj01<- read_csv("./data/2023-09-12_cleaned.csv")
view(lj01)
# EDA ----------------------------------------------------------------------
theme_set(theme(text = element_text(family="sans",size = 10)))
```

一、数据概览

数据表 lj01 共包括 property_region、price_sqm 等 18 个变量，涉及 3000 行数据。表的前 10 行示例如下：

```
lj01 %>%
  head(10) %>%
  kable(caption = " 武汉链家二手房") %>%
  kable_styling()
```

二、数据去重

```
lj02 <- unique(lj01)
view(lj02)
```

对数据表 lj01 去重后得到数据表 lj02，仍包括 property_region、price_sqm 等 18 个变量，涉及 2515 行数据。

第二部分：对各变量做描述性统计

一、查看整体分布特征

```
glimpse(lj02)
```

```
## Rows: 2,515
## 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", ~
```

```
pander(summary(lj02))
```

表 2: Table continues below

property_name	property_region	price_ttl	price_sqm
Length:2515	Length:2515	Min. : 10.6	Min. : 1771
Class :character	Class :character	1st Qu.: 95.0	1st Qu.:10765
Mode :character	Mode :character	Median : 136.0	Median :14309
NA	NA	Mean : 154.8	Mean :15110
NA	NA	3rd Qu.: 188.0	3rd Qu.:18213
NA	NA	Max. :1380.0	Max. :44656

表 3: Table continues below

bedrooms	livingrooms	building_area	directions1
Min. :1.000	Min. :0.000	Min. : 22.77	Length:2515
1st Qu.:2.000	1st Qu.:1.000	1st Qu.: 84.45	Class :character
Median :3.000	Median :2.000	Median : 95.46	Mode :character
Mean :2.689	Mean :1.706	Mean :100.67	NA
3rd Qu.:3.000	3rd Qu.:2.000	3rd Qu.:118.03	NA
Max. :7.000	Max. :4.000	Max. :588.66	NA

表 4: Table continues below

directions2	decoration	property_t_height	property_height
Length:2515	Length:2515	Min. : 2.00	Length:2515
Class :character	Class :character	1st Qu.:11.00	Class :character
Mode :character	Mode :character	Median :27.00	Mode :character
NA	NA	Mean :24.05	NA
NA	NA	3rd Qu.:33.00	NA
NA	NA	Max. :62.00	NA

表 5: Table continues below

property_style	followers	near_subway	if_2y
Length:2515	Min. : 0.000	Length:2515	Length:2515
Class :character	1st Qu.: 1.000	Class :character	Class :character
Mode :character	Median : 2.000	Mode :character	Mode :character
NA	Mean : 6.326	NA	NA
NA	3rd Qu.: 6.000	NA	NA
NA	Max. :262.000	NA	NA

has_key	vr
Length:2515	Length:2515
Class :character	Class :character
Mode :character	Mode :character
NA	NA
NA	NA

has_key	vr
NA	NA

可以看到：

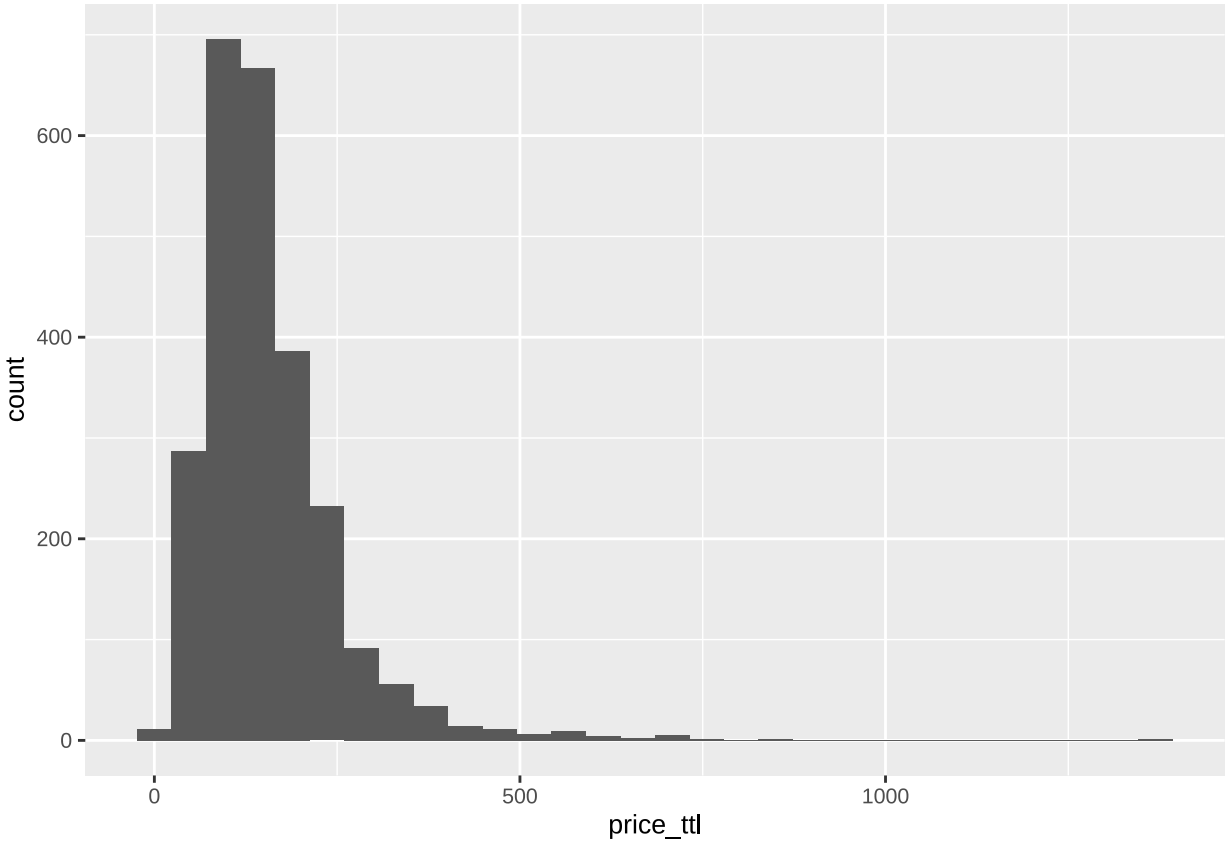
- 数值型变量 7 个:price_ttl,price_sqm,bedrooms,livingrooms,building_area,property_t_height,followers;
  字符型变量 11 个:property_name,property_region,directions1,directions2,decoration,property_height,property_
  near_subway,if_2y,has_key,vr; 可将 9 个字符型变量转换为 factor 因子:directions1,directions2,decotion,property_
- 房屋单价平均值为 14309 元/平方米，房屋总价平均值为 154.8 万元，户均面积 100.67 平方米，户型以 2-3 个房间、1-2 个客厅居多。

二、单个变量描述统计

1. 小区名字及所处区域

2. 房屋总价

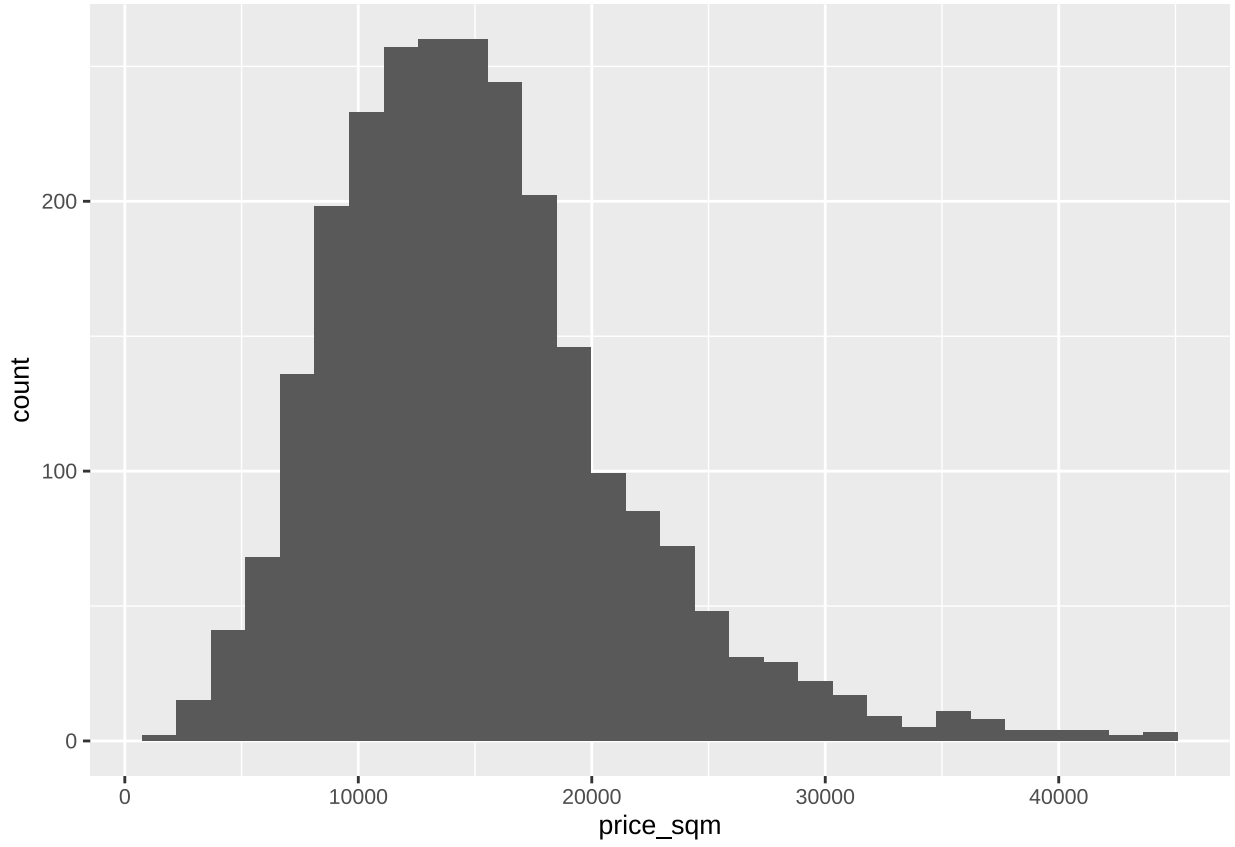
```
lj02%>%
  ggplot(aes(price_ttl))+
  geom_histogram()
```



由直方图可以看出，房屋总价集中在 100-200 万元。

3. 房屋单价

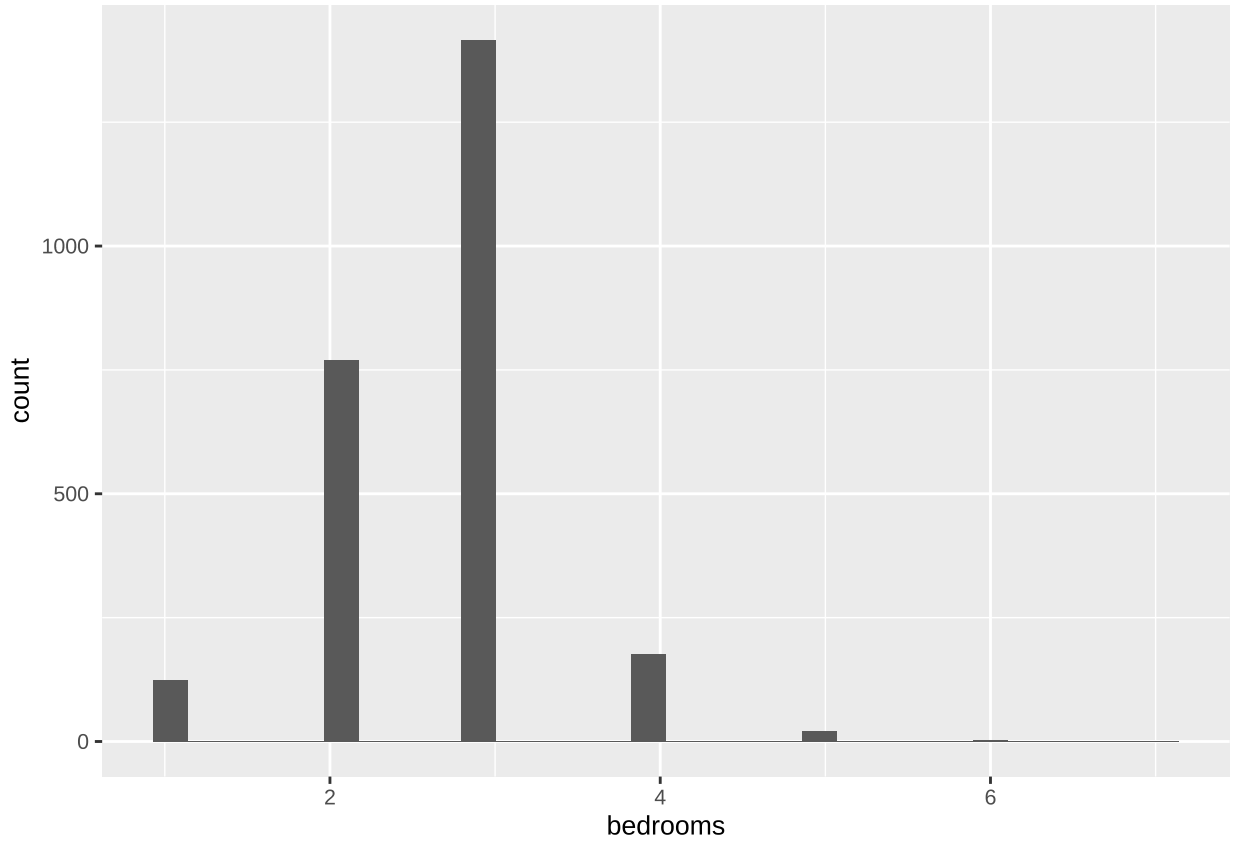
```
lj02%>%
  ggplot(aes(price_sqm))+
  geom_histogram()
```



由直方图可以看出，房屋单价集中在 10000 元/平方米-20000 元/平方米。

4. 房间数

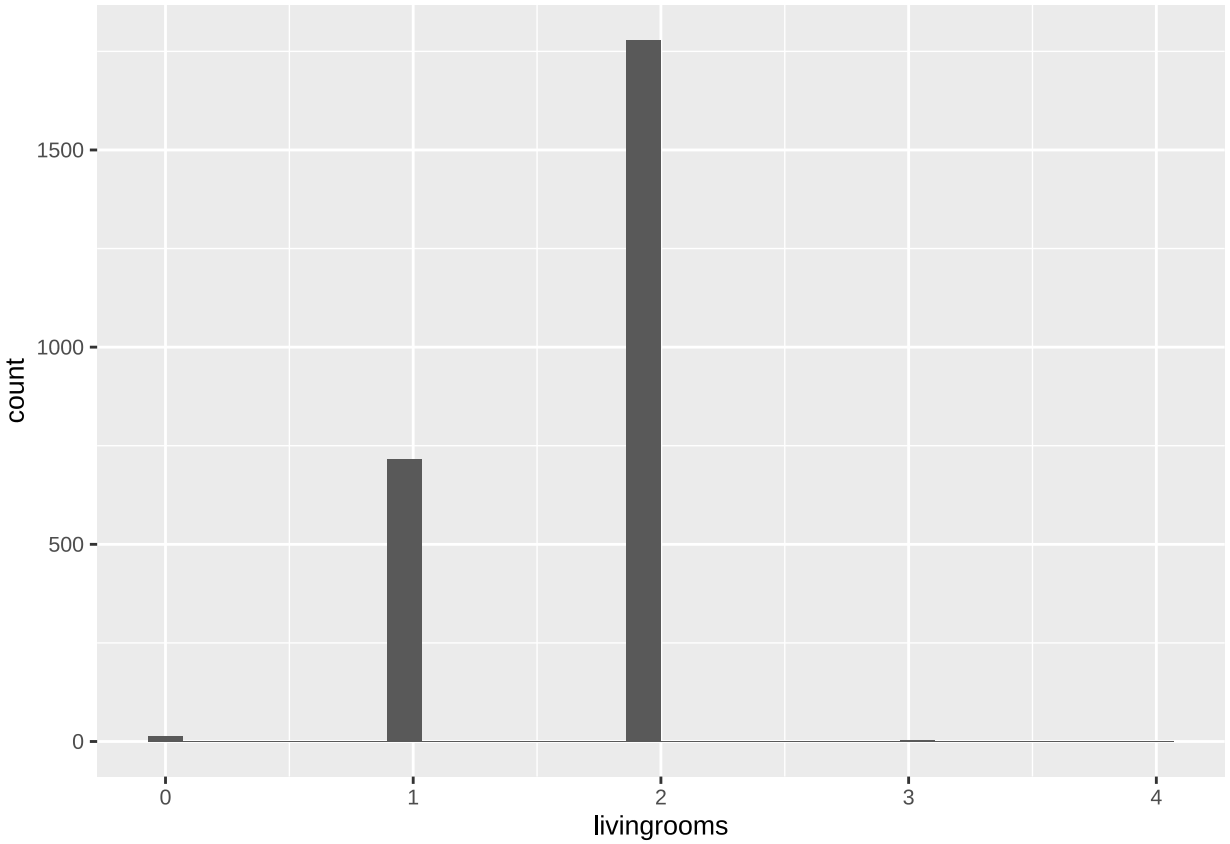
```
lj02%>%
  ggplot(aes(bedrooms))+
  geom_histogram()
```



由直方图可以看出，有房间数为 1、2、3、4、5、6 及其他的户型，其中房间数为 2、3 的户型居多，房间数为 1、4 的户型次之。

5. 客厅数

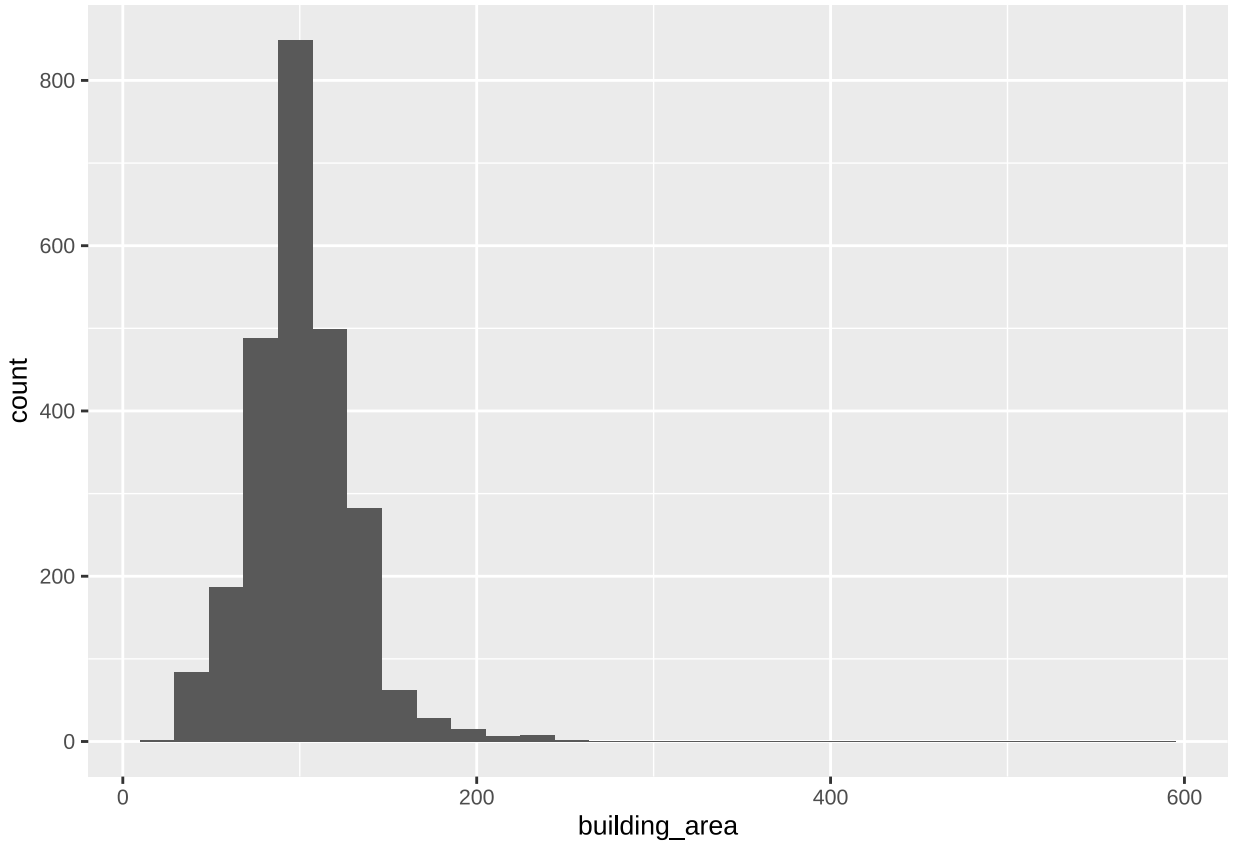
```
lj02%>%
  ggplot(aes(livingrooms))+
  geom_histogram()
```



由直方图可以看出，有客厅数为 0、1、2、3 的四种户型，绝大部分为客厅数为 1、2 的户型。

6. 建筑面积

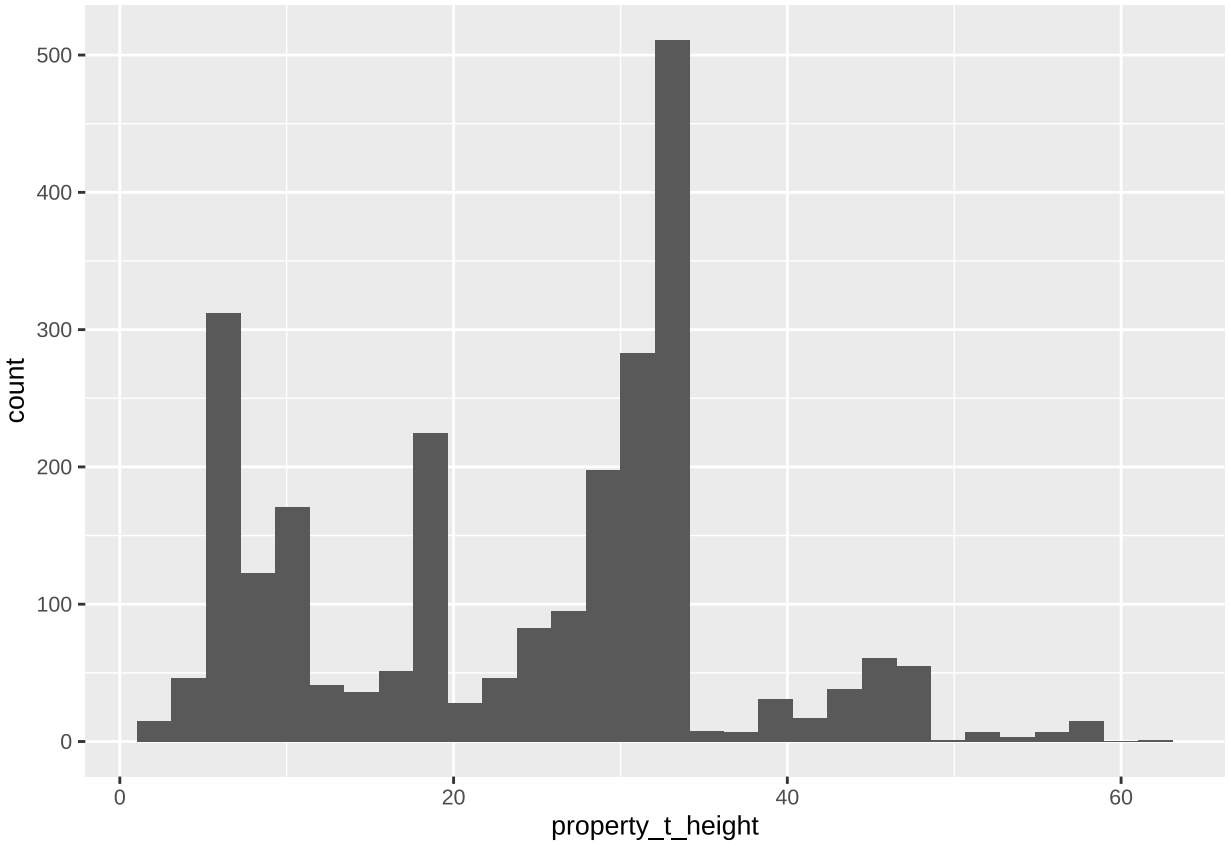
```
lj02%>%
  ggplot(aes(building_area))+
  geom_histogram()
```



由直方图可以看出，建筑面积集中在 60 平方米-140 平方米区间范围内。

7. 楼栋总层数

```
lj02%>%
  ggplot(aes(property_t_height))+
  geom_histogram()
```



由直方图可以看出，房屋总层数有低层及中高层之分，低层集中在 4-10 层，高层集中在 30 层左右。

8. 房屋朝向、装修状况、建筑形式、是否靠近地铁等

```
table(lj02$directions1)
```

```
##
##   北   东 东北 东南   南   西 西北 西南
##   59   81   10  238 2056   14   11   46
```

```
table(lj02$directions2)
```

```
##
##   北   东 东北 东南   南   西 西北 西南
## 1001    5    5   12   54   21    9   11
```

```
table(lj02$decoration)
```

```
##
## 简装 精装 毛坯 其他
##  536 1476  356  147
```

二手房主要以简装和精装为主。

```
table(lj02$property_height)
```

```
##
##   低   高   中
##  662  750 1050
```

```
table(lj02$property_style)
```

```
##
##     板楼 板塔结合     平房     塔楼 暂无数据
##     1503      506        4      441       61
```

板楼主要功能为住宅，塔楼主要功能为办公、住宅、酒店、观光塔等，板塔结合同时满足商业和居住空间。由此可见，该二手房数据不仅涉及住宅，还涉及商业、办公等。

```
table(lj02$near_subway)
```

```
##
## VR看装修    近地看    近地铁    珞狮南 太子湖1号
##         2         1      1305         1         1
```

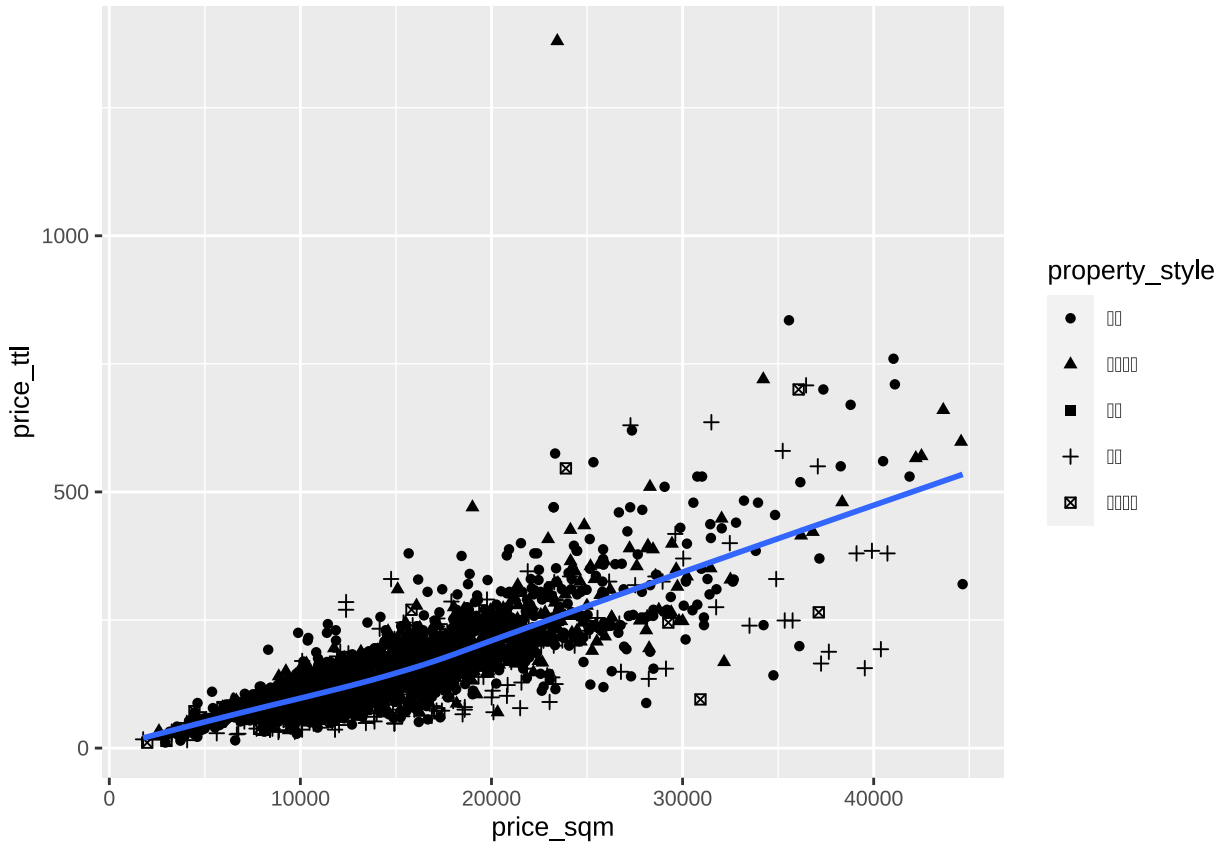
```
table(lj02$if_2y)
```

```
##
## 房本满两年
##       1050
```

第三部分：探索性分析

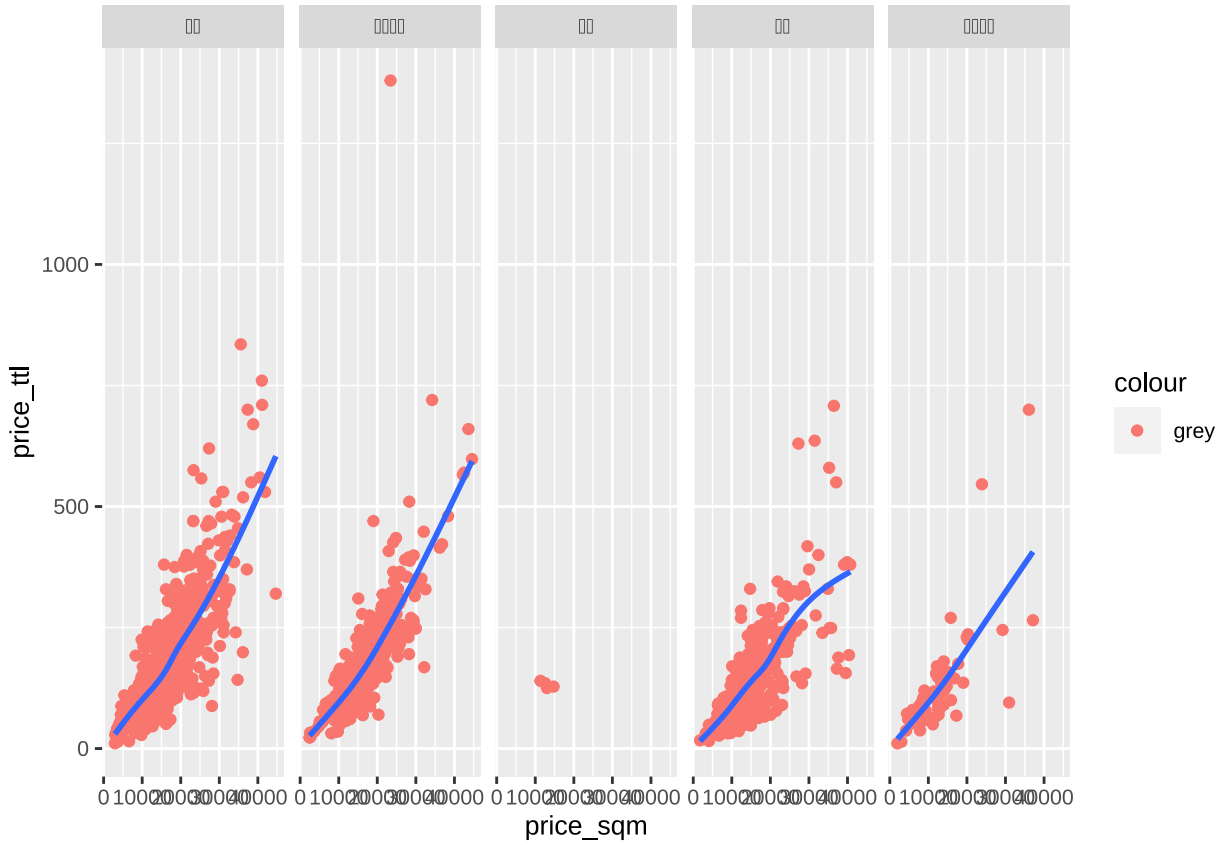
一、分析房屋总价与房屋单价间的关系

```
lj02%>%
  ggplot()+
  geom_point(aes(x=price_sqm,y=price_ttl,shape=property_style))+
  geom_smooth(aes(x=price_sqm,y=price_ttl),se=FALSE)
```



总体来看，房屋总价与房屋单价呈正相关性，房屋总价随着房屋单价的增加而增加。

```
lj02%>%
  ggplot()+
  geom_point(aes(x=price_sqm,y=price_ttl,color="grey"))+
  geom_smooth(aes(x=price_sqm,y=price_ttl),se=FALSE)+
  facet_grid(.~property_style)
```



按照房屋建筑形式（即房屋功能）划分，可以看出房屋总价与房屋单价仍呈正相关性，房屋总价随着房屋单价的增加而增加。

二、比较住宅及商业住宅单价差异

```
# 对二手房的房屋
lj02%>%
  group_by(property_style)%>%
  summarise(mean1=mean(price_ttl),max=max(price_ttl),min=min(price_ttl))
```

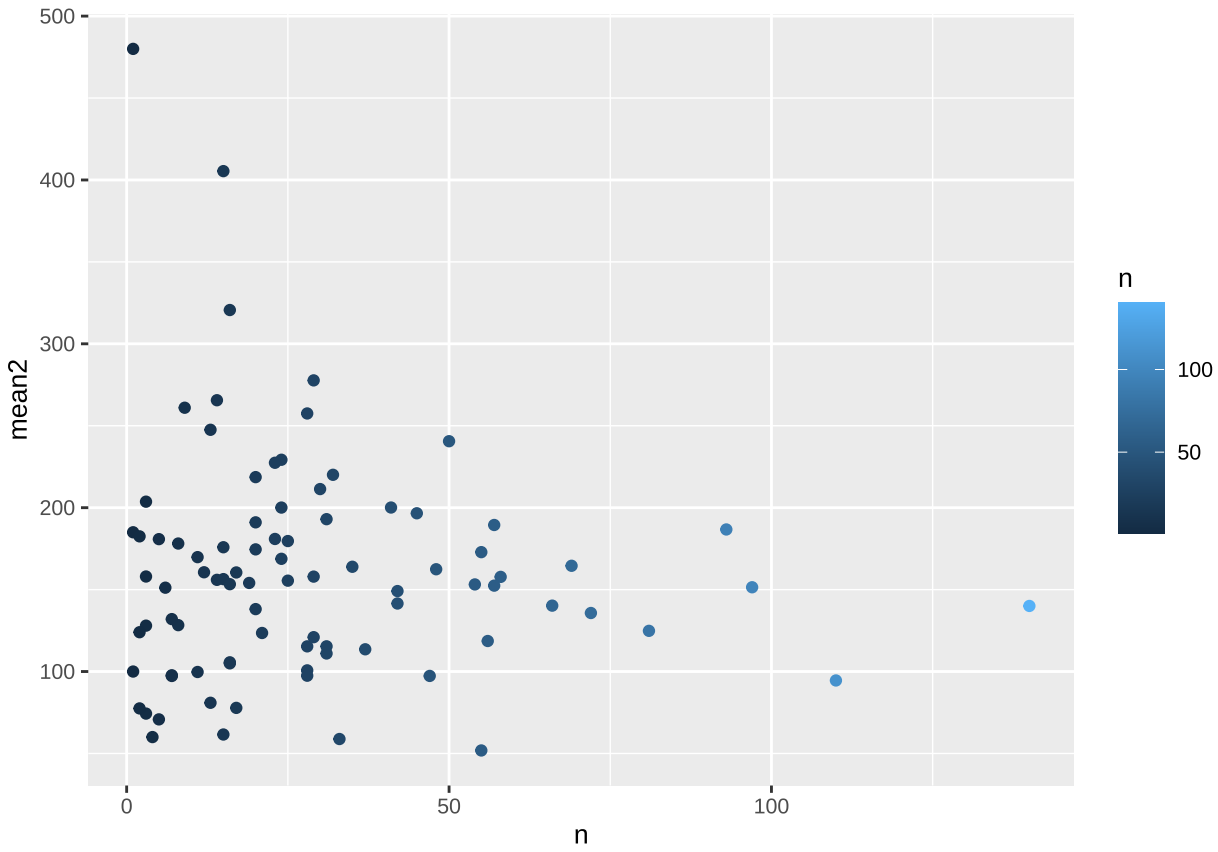
```
## # A tibble: 5 x 4
##   property_style mean1   max   min
##   <chr>          <dbl> <dbl> <dbl>
## 1 塔楼            141.   708  15.5
## 2 平房            132    140 125
## 3 暂无数据        129.   700  10.6
## 4 板塔结合        174.  1380  23
## 5 板楼            154.   835  10.8
```

通过分析可以看出，一是塔板结合（商住两用）房屋单价最高，二是板楼（住宅）房屋单价高于塔楼（商业）房屋单价，这与土地市场上住宅用地亩单价高于商业用地亩单价情况一致。

三、分析房屋单价与区域热销程度的关系

```
lj_n_mean <- lj02%>%
  group_by(property_region)%>%
  summarise(n=n(),mean2=mean(price_ttl))
```

```
lj_n_mean%>%
  ggplot() +
  geom_point(aes(x = n, y = mean2, color = n))
```



第四部分总结

1. 该二手房数据中，数值型变量 7 个:price_ttl,price_sqm,bedrooms,livingrooms,building_area, property_t_height,followers; 字符型变量 11 个:property_name,property_region,directions1, directions2,decoration,property_height,property_style，near_subway,if_2y,has_key, vr; 可将 9 个字符型变量转换为 factor 因子:directions1,directions2,decotion,property_height, property_style,near_subway,if_2y,has_key,vr。2. 该二手房数据中，房屋总价集中在 100-200 万元，房屋总价平均值为 154.8 万元；房屋单价集中在 10000 元/平方米-20000 元/平方米，房屋单价平均值为 14309 元/平方米；有房间数为 1、2、3、4、5、6 及其他的户型，其中房间数为 2、3 的户型居多，房间数为 1、4 的户型次之；有客厅数为 0、1、2、3 的四种户型，绝大部分为客厅数为 1、2 的户型；建筑面积集中在 60 平方米-140 平方米区间范围内，户均面积 100.67 平方米；房屋总层数有低层及中高层之分，低层集中在 4-10 层，高层集中在 30 层左右。3. 总体来看，房屋总价与房屋单价呈正相关性，房屋总价随着房屋单价的增加而增加。按照房屋建筑形式（即房屋功能）划分，可以看出房屋总价与房屋单价仍呈正相关性，房屋总价随着房屋单价的增加而增加。4. 通过分析可以看出，一是塔板结合（商住两用）房屋单价最高，二是板楼（住宅）房屋单价高于塔楼（商业）房屋单价，这与土地市场上住宅用地亩单价高于商业用地亩单价情况一致。5. 通过观察，房屋单价与区域热销程度没有很强的关联性。