武汉链家的二手房数据分析报

2023281051028-李璇-MEM

目录

你的主要发现

1. 发现 1：武汉二手房的房屋单价受地理位置影响较大，呈正相关；

2 . 发现 2：武汉二手房的房屋设计趋势主要为卧室 2 ~ 3 + 客 厅 1 ~

2 ；

3. 发现 3：武汉二手房市场主要占据在层高二十几层的房屋。

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。

```
# 载入数据和预处理
getwd()
```

```
## [1] "D:/Users/lix/Desktop/R"
```

```
setwd("C:/Program Files/R")
lj_wuhan <- read_csv("d:/Users/lix/Desktop/R/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.
```

```
View(lj_wuhan)


# 数据预处理，去掉重复样本数据。
lj_wuhan <- distinct(lj_wuhan)
pander(summary(is.na(lj_wuhan)))
```

表 1: Table continues below

property_name	property_region	price_ttl	price_sqm
Mode :logical	Mode :logical	Mode :logical	Mode :logical
FALSE:2515	FALSE:2515	FALSE:2515	FALSE:2515
NA	NA	NA	NA

表 5: 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

表 2: Table continues below

bedrooms	livingrooms	building_area	directions1	directions2
Mode :logical	Mode :logical	Mode :logical	Mode :logical	Mode :logical
FALSE:2515	FALSE:2515	FALSE:2515	FALSE:2515	FALSE:1118
NA	NA	NA	NA	TRUE :1397

表 3: Table continues below

decoration	property_t_height	property_height	property_style
Mode :logical	Mode :logical	Mode :logical	Mode :logical
FALSE:2515	FALSE:2515	FALSE:2462	FALSE:2515
NA	NA	TRUE :53	NA

followers	near_subway	if_2y	has_key	vr
Mode :logical	Mode :logical	Mode :logical	Mode :logical	Mode :logical
FALSE:2515	FALSE:1310	FALSE:1050	FALSE:2092	FALSE:1754
NA	TRUE :1205	TRUE :1465	TRUE :423	TRUE :761

数据概览

1、数据表（lj——wuhan) 的前 10 行示例如下：

2、各变量的简短信息

```
glimpse(lj_wuhan)
```

```
## 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", ~
```

3、各变量的简短统计

```
summary(lj_wuhan)
```

```
##  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
##                                        Mean   : 154.8   Mean   :15110
##                                        3rd Qu.: 188.0   3rd Qu.:18213
##                                        Max.   :1380.0   Max.   :44656
##     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
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:118.03
```

```
##  Max.   :7.000   Max.   :4.000   Max.   :588.66
##  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
##                                        Mean   :24.05
##                                        3rd Qu.:33.00
##                                        Max.   :62.00
##  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
##                     Mean   :  6.326
##                     3rd Qu.:  6.000
##                     Max.   :262.000
##    has_key              vr
##  Length:2515        Length:2515
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

可以看到：

- 直观结论 1

通过简单的数据清洗，原数据包包含 18 列向量，3000 行数据，现在为 18 列向量，2515 行数据;

主要数据内容包含房屋名称（property_name）、房屋地理位置（property_region）、总价（price_ttl）、房屋单价 (price_sqm）、卧室数量（bedrooms）、客厅数量（livingrooms）、房屋面积 (building_area)、朝向（directions）、装修情况（decoration）、房屋总层高（property_t_height）、房屋高度（property_height）、房屋类型（property_style）、附近住宅数量（followers）、是否进地铁（near_subway）、房本是否满两年（if_2y)、是否有钥匙 (has_key)、是否可 vr 看房 (vr)。

- 直观结论 2

从变量的简短统计可以看出:

房屋单价：均值为 15110 元/㎡, 中位数为 14309 元/㎡，最高为 44656 元/㎡，最低为 1771 元/㎡；
卧室：均值为 2.689 个, 中位数为 3 个；
客厅：均值为 1.706 个，中位数为 2 个；
房屋面积：均值为 100.67 ㎡，中位值为 95.46 ㎡；
房屋总层高：均值为 24.05 层，中位值为 27 层。

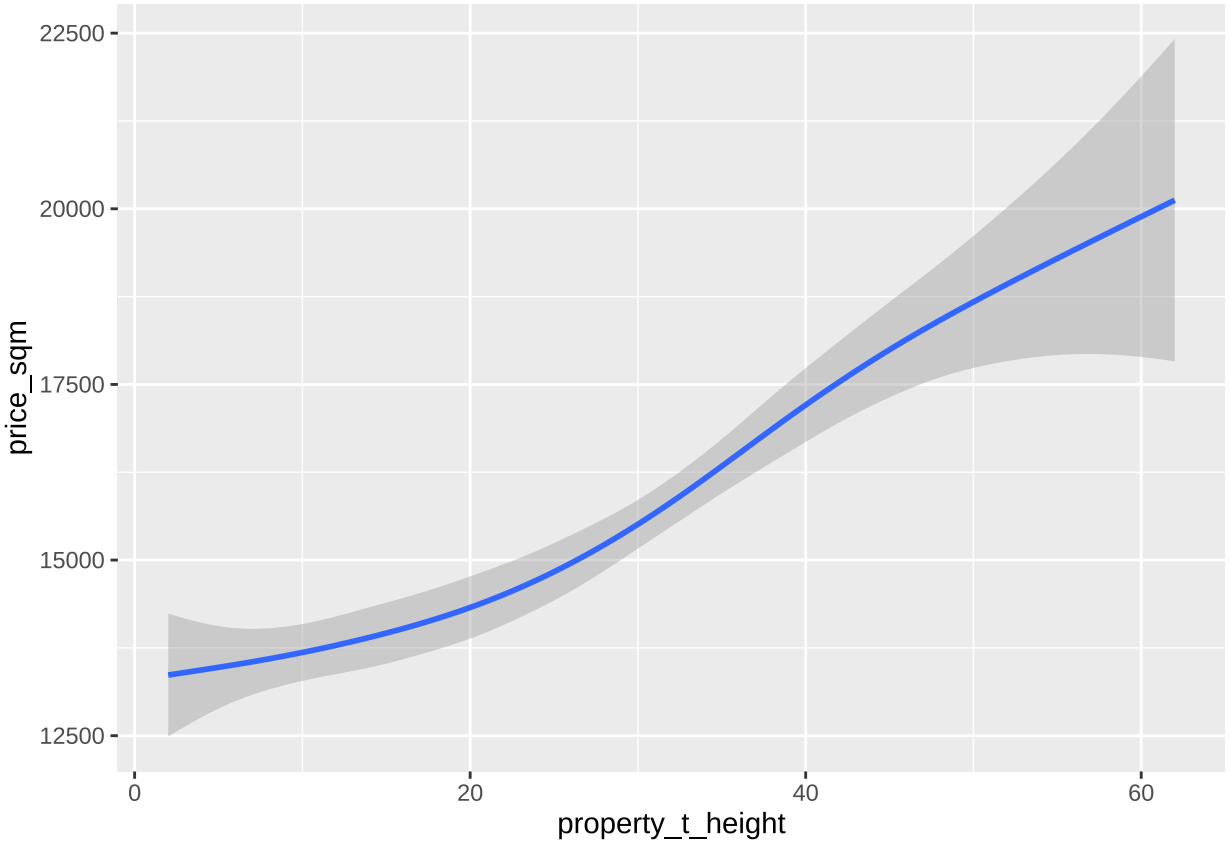
综上，房屋单价浮动大，具体受影响因素还需详细分析；房屋卧室及客厅设计主要集中在 2 ~ 3 室 和 1 ~ 2 个客厅的结构；房屋面积主要还是在 100 ㎡左右；房屋总层高也主要是二十多层的高层住宅。

探索性分析

变量 1 的数值描述与图形 (房屋单价与房屋总层高的关系)

```
ggplot(data = lj_wuhan)+
  geom_smooth(mapping = aes(x=property_t_height,y=price_sqm))
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

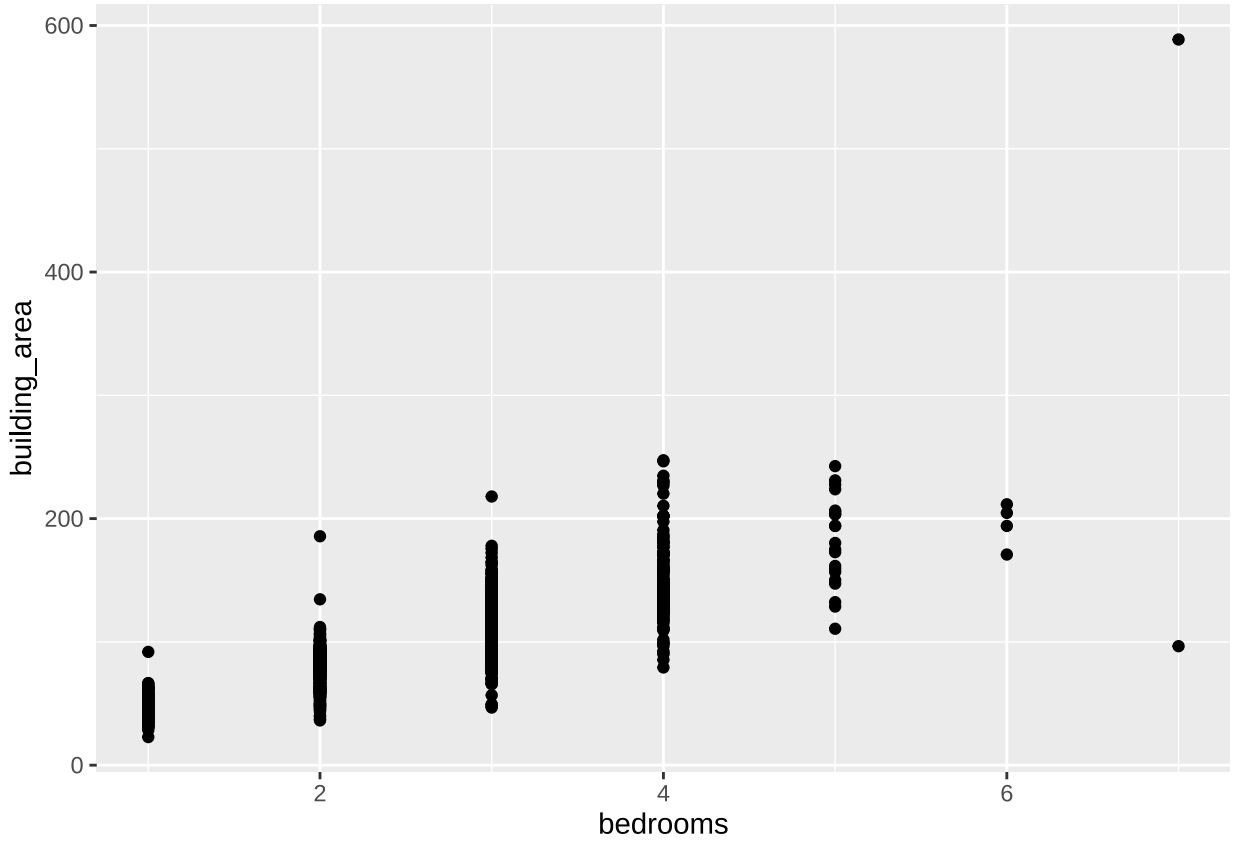


发现：

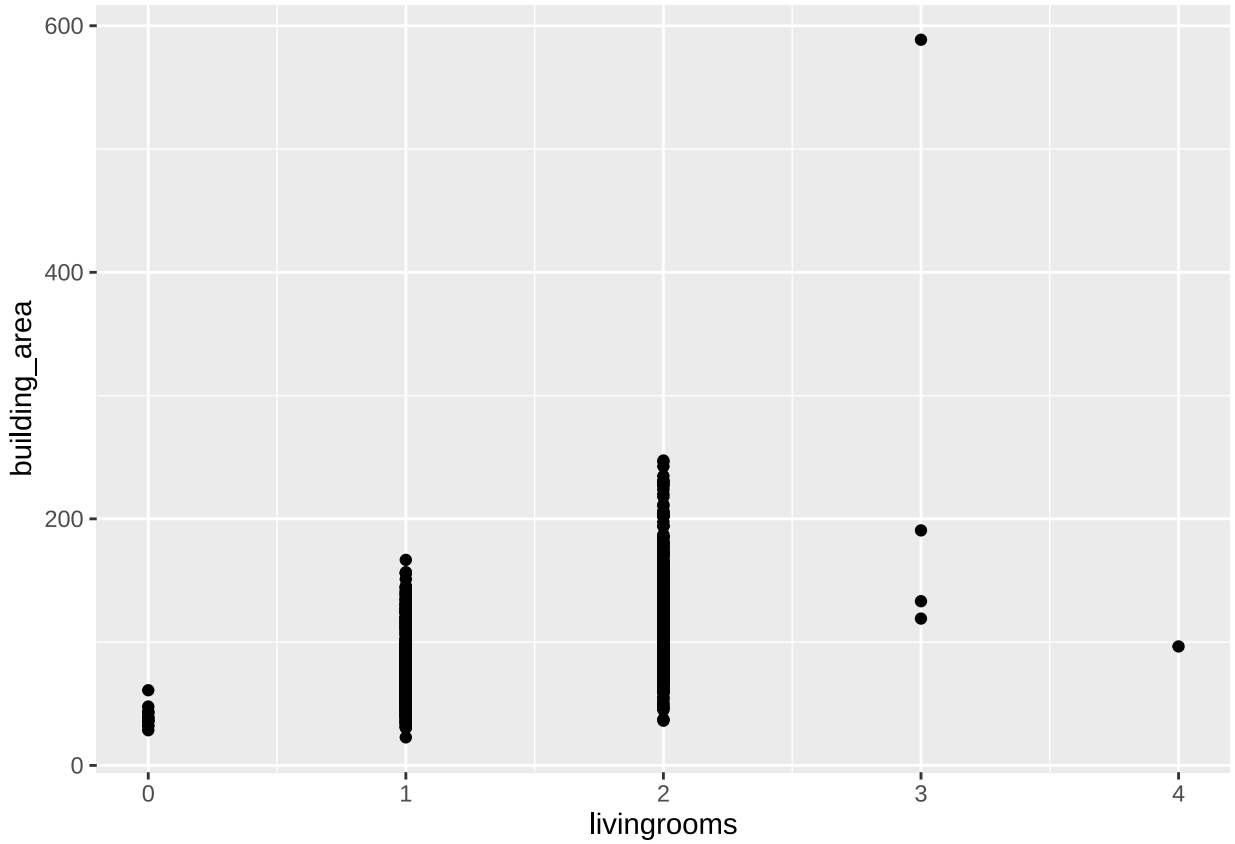
- 发现 1 随着房屋总层高的增加，整体的房屋单价呈上涨趋势。由此可以发现，房屋单价与房屋总层高成正相关。但是 10 层以下及 50 层以上的价格浮动区间较大，存在 60 层高的房屋单价低于 50 层高的情况，说明对超高层的房屋单价还有其他影响因素，需进一步分析。
- 发现 2 房屋总层高在 30 层左右的，其房屋单价水平相对较稳定，上下浮动区间较小。由此可以发现，3 房屋总层高在 30 层左右的，房屋单价受其他因素的影响较小，且在武汉市场的受众相对较广。

变量 2 的数值描述与图形 (房屋卧室及客厅数量与房屋面积的关系)

```
ggplot(data = lj_wuhan)+
  geom_point(mapping = aes(bedrooms,y=building_area))
```




```
ggplot(data = lj_wuhan)+
  geom_point(mapping = aes(livingrooms,y=building_area))
```

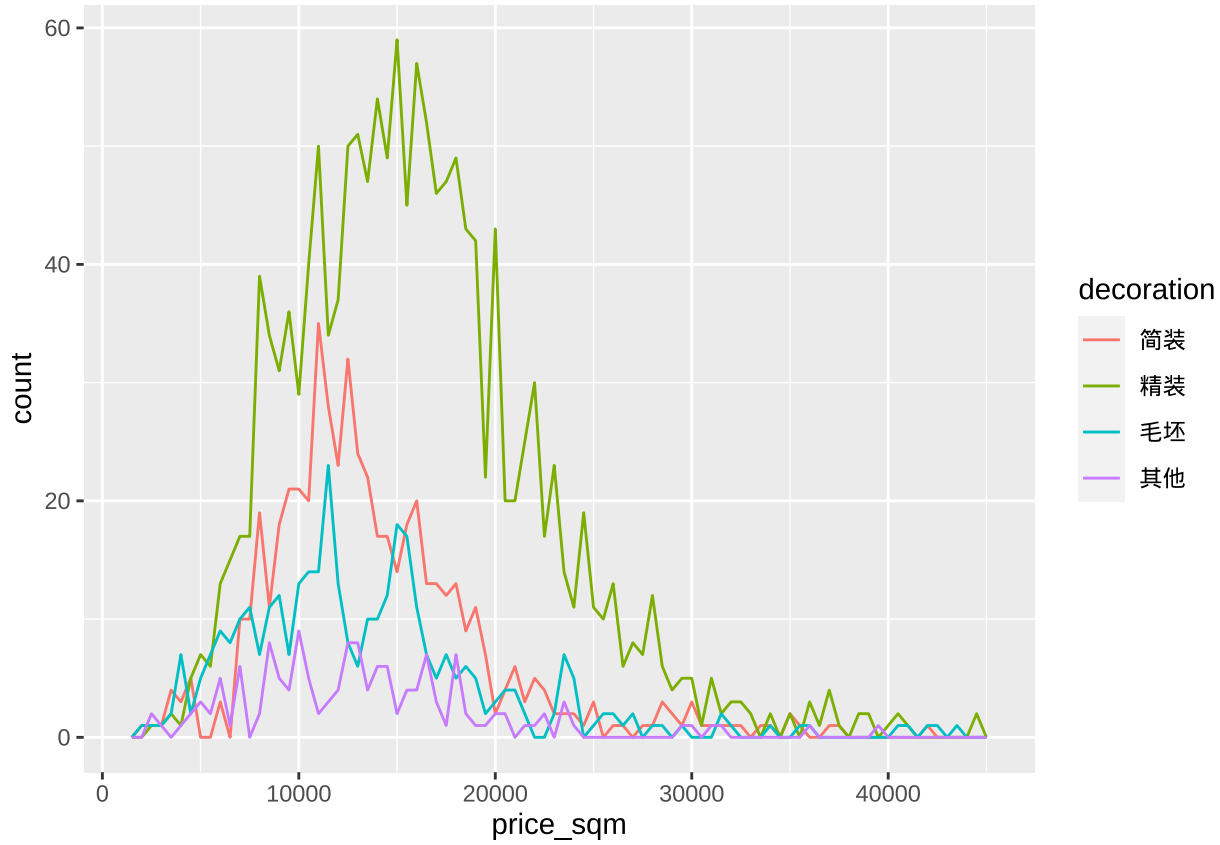


发现：

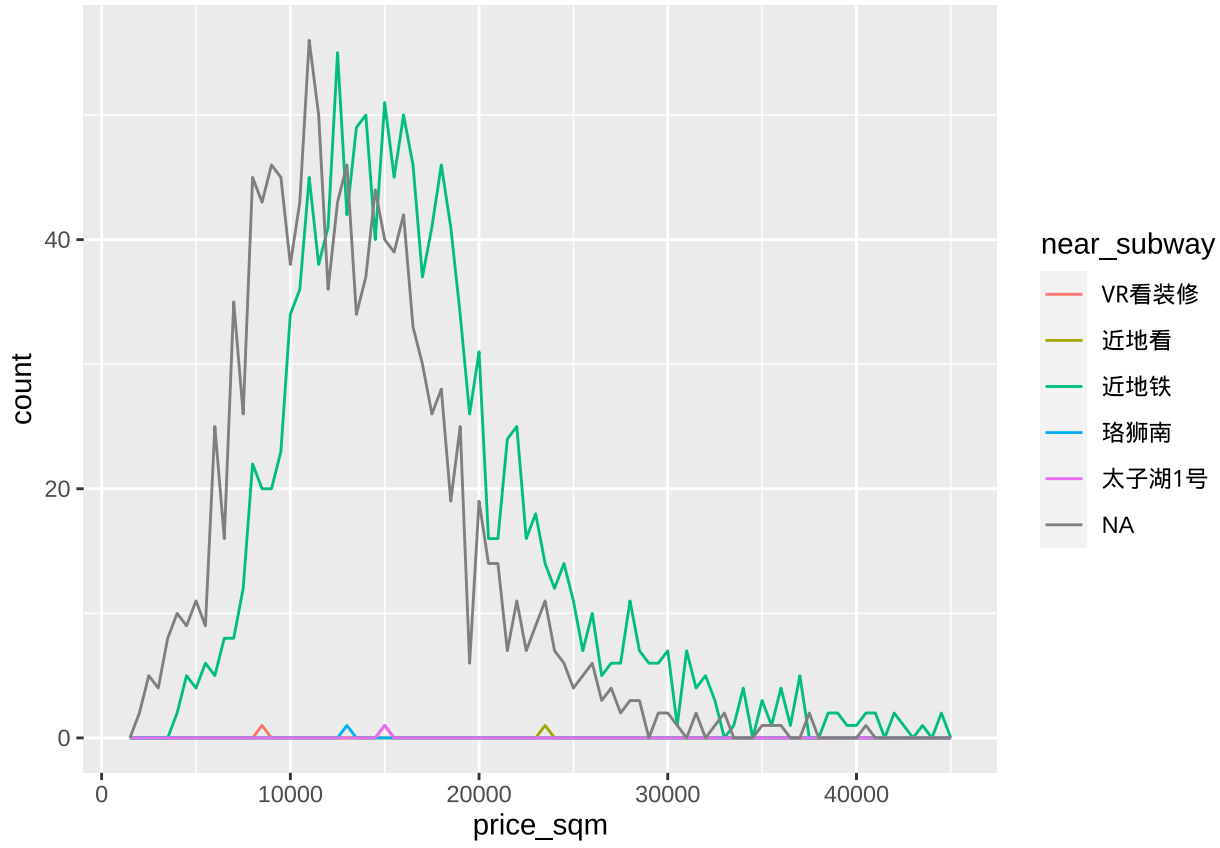
- 发现 1 从武汉的房屋卧室数量来看，卧室房型还是 3~4 个房间占多数，与均值均值为 2.689 个存在一定偏差。从房屋面积与卧室数量的关系来看，100 ㎡的房屋主要设计在 $1_{2\ 个房间，\ 100}$200 ㎡的房屋主要设计在 3~4 个房间. 综上，可以发现武汉的房屋设计结构主要偏向 3~4 个卧室，且房屋卧室数量的设计受房屋面积影响。
- 发现 2 从武汉的房屋客厅数量来看，客厅数量在 1~2 个占多数，与均值、中位值相近。从房屋面积与客厅数量的关系来看，200 ㎡以下的房屋客厅数量均有分布。综上，可以发现武汉的房屋设计结构主要偏向 1~2 个客厅，且房屋客厅数量的设计受房屋面积的影响较小。

变量 3 的数值描述与图形（精装情况及交通与房屋单价的关系）

```
ggplot(data=lj_wuhan,mapping = aes(x=price_sqm))+
  geom_freqpoly(mapping=aes(color=decoration),binwidth=500)
```




```
ggplot(data=lj_wuhan,mapping = aes(x=price_sqm))+
  geom_freqpoly(mapping=aes(color=near_subway),binwidth=500)
```



发现：

- 发现 1 武汉市场精装房数量在各个价格段，其数量都是最高的。由此可以发现，武汉目前的二手房房屋市场精装占比最大，简装次之。
- 发现 2 房屋单价处在 1.25 万以下时，不靠近地铁的房屋数量多于靠近地铁。当房屋单价大于 1.25 万时，靠近地铁的房屋数量多于不靠近地铁的。由此可以发现，交通的便利会与房屋单价为正相关。

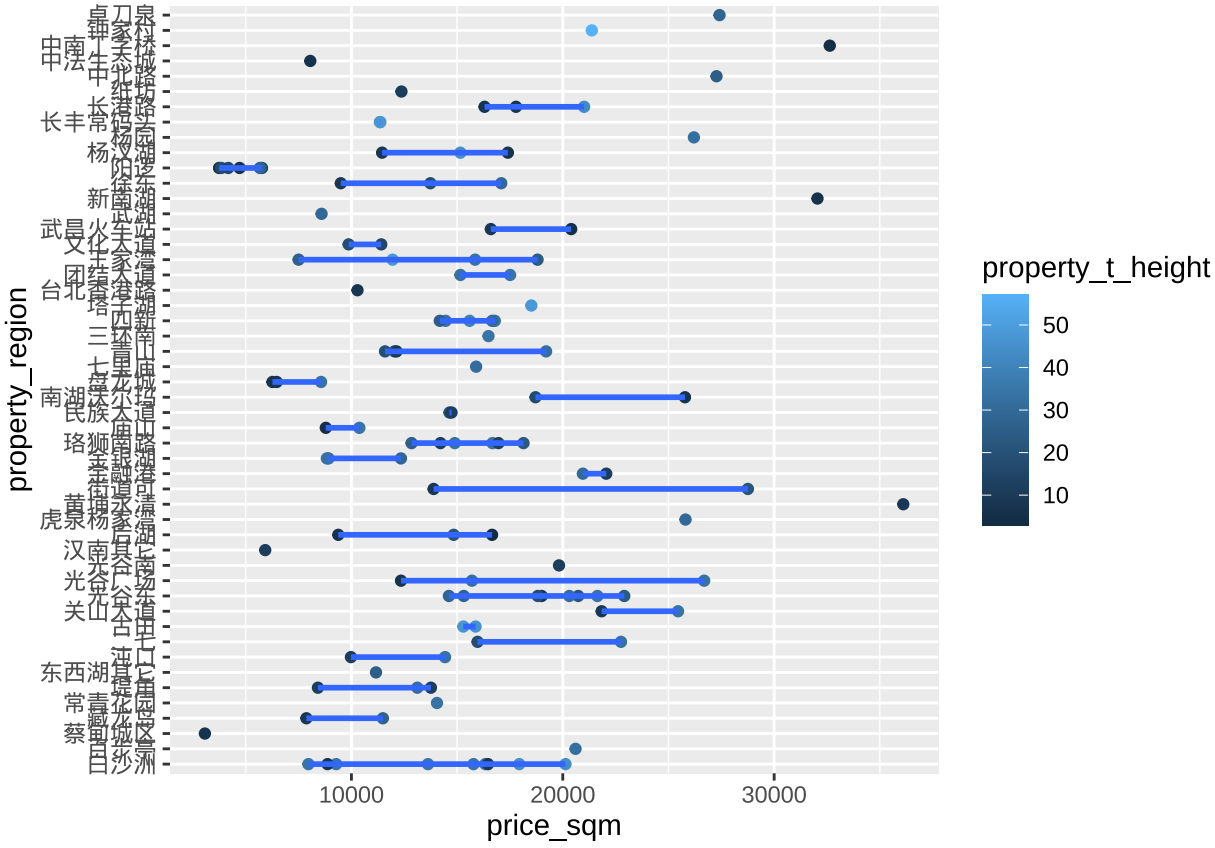
探索问题 1（房屋单价是否受房屋总层高、地理位置的影响?）

```
lj_wuhan[1:115,]%>%
  ggplot(aes(price_sqm,property_region,color=property_t_height))+
  geom_point()+


  geom_smooth(method="lm",se=FALSE)
```

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

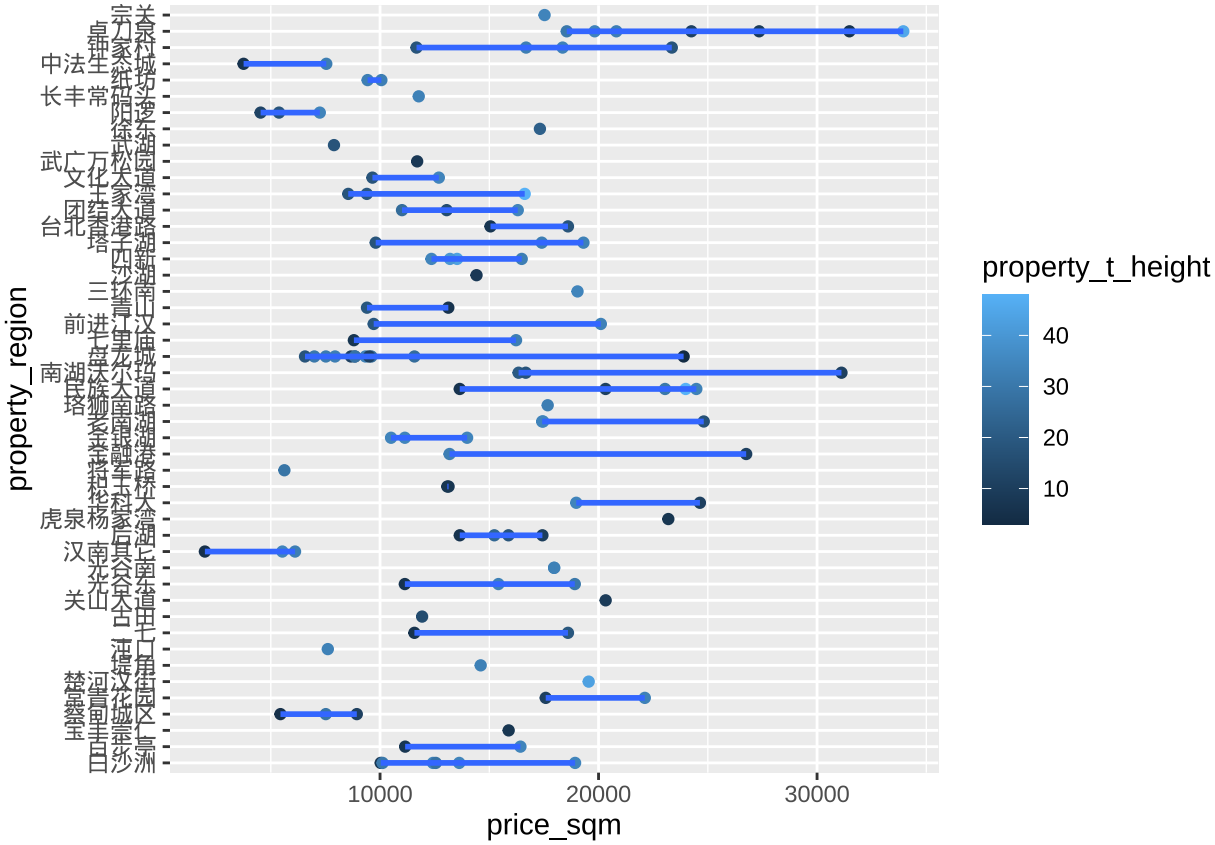
```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



```
lj_wuhan[1250:1365,]%>%
  ggplot(aes(price_sqm,property_region,color=property_t_height))+
  geom_point()+
  geom_smooth(method="lm",se=FALSE)
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

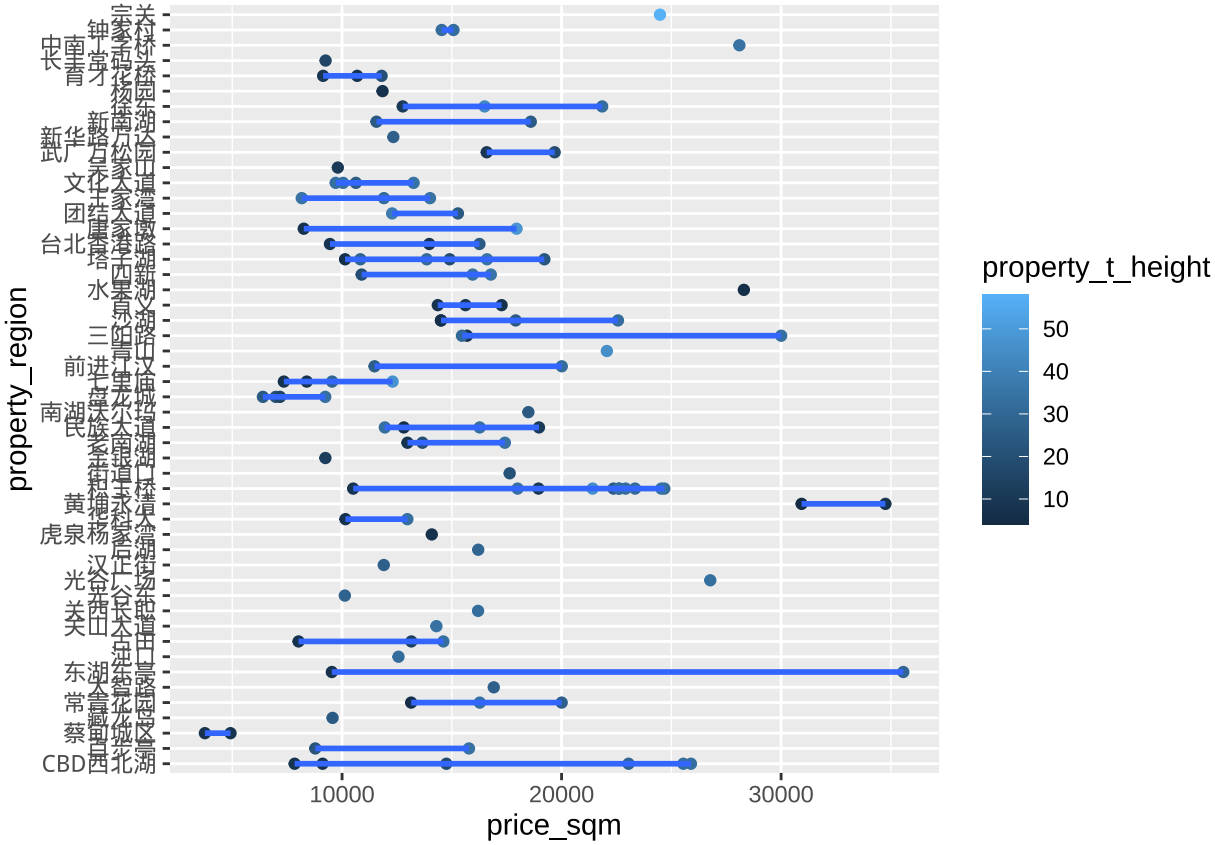



```
lj_wuhan[2400:2515,]%>%
  ggplot(aes(price_sqm,property_region,color=property_t_height))+
  geom_point()+
  geom_smooth(method="lm",se=FALSE)
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
```

```
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



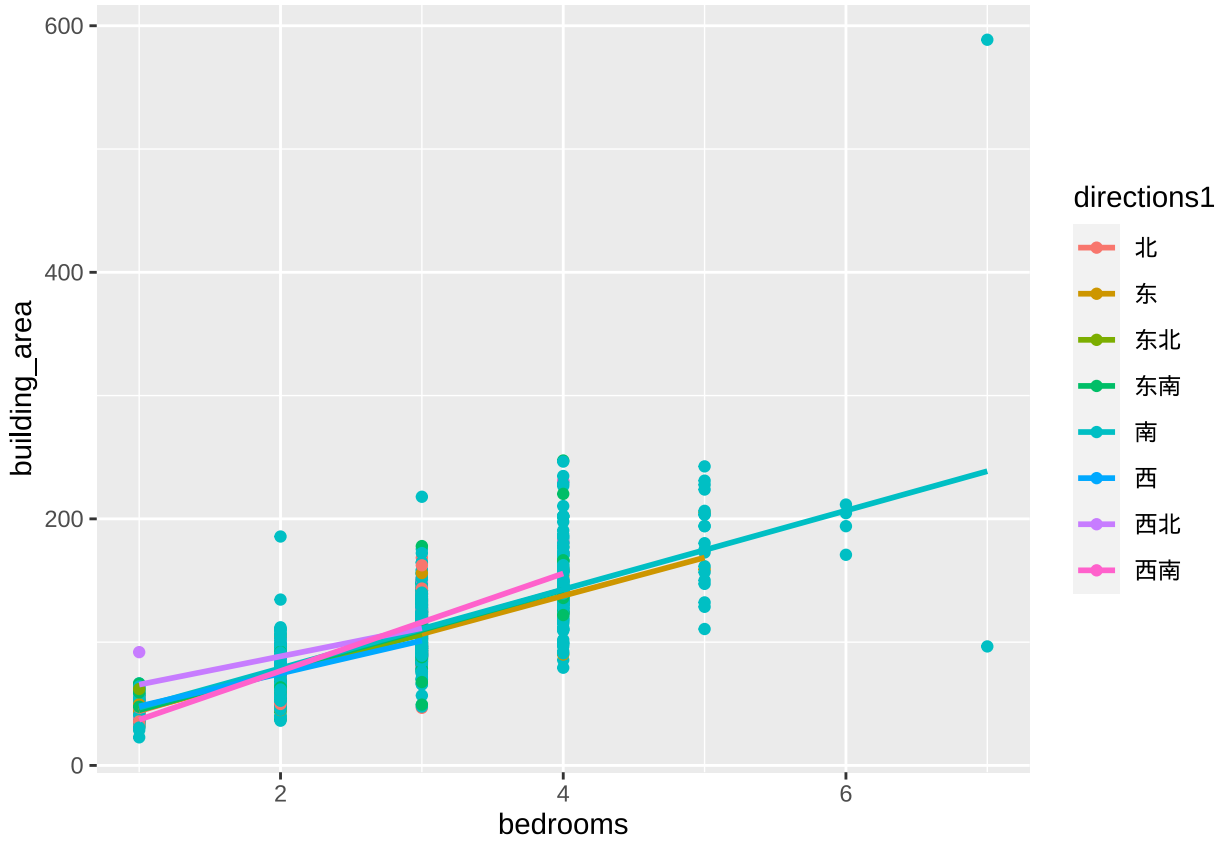
发现：

- 发现 1 结合前面的分析（10 层以下的房屋单方浮动较大）及目前的分析发现，如房屋楼栋总层高同样在 10 以下的蔡甸和阳逻区域，其单价都在 5k 以下。而在南湖和丁字桥的单价都在 3w 以上。由此可以发现，在同样房屋楼栋总层高的情况下，商圈或内环的单价更高。
- 发现 2 根据图形分布和市场单价均值（15110 元/㎡）可以发现，武汉市场的房屋单价水平在 1~2w 区间。其中超高层的房屋地理位置更偏向于内环, 如百步亭、唐家墩、王家湾、楚河汉界、钟家村等均有分布。结合单价与层高的关系，由此也可以发现超高层的地理位置选择与区域经济状况为正相关。

探索问题 2（房屋卧室及客厅数量与房屋面积、房屋朝向之间存在什么样的关系?）

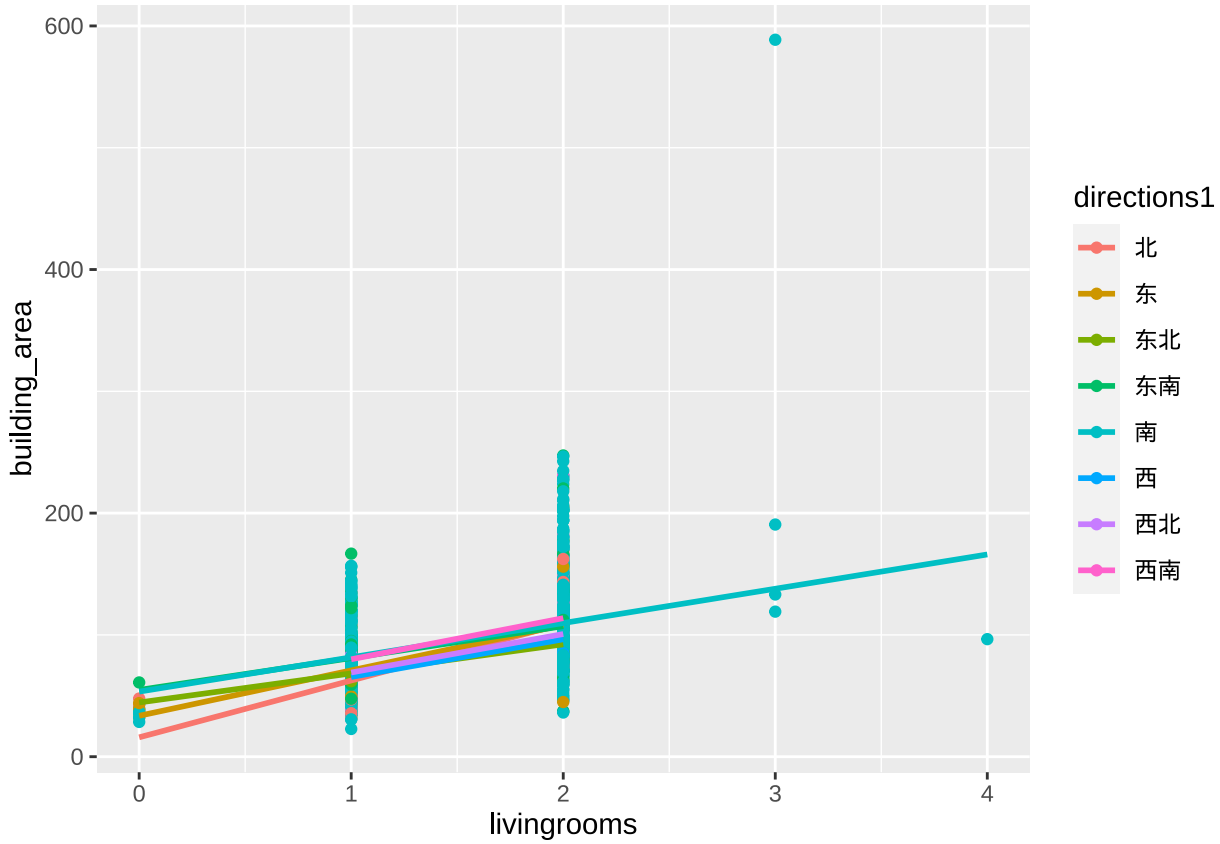
```
lj_wuhan%>%
  ggplot(aes(bedrooms,building_area,color=directions1))+
  geom_point()+
  geom_smooth(method="lm",se=FALSE)
```

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




```
lj_wuhan%>%
  ggplot(aes(livingrooms,building_area,color=directions1))+
  geom_point()+
  geom_smooth(method="lm",se=FALSE)
```

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

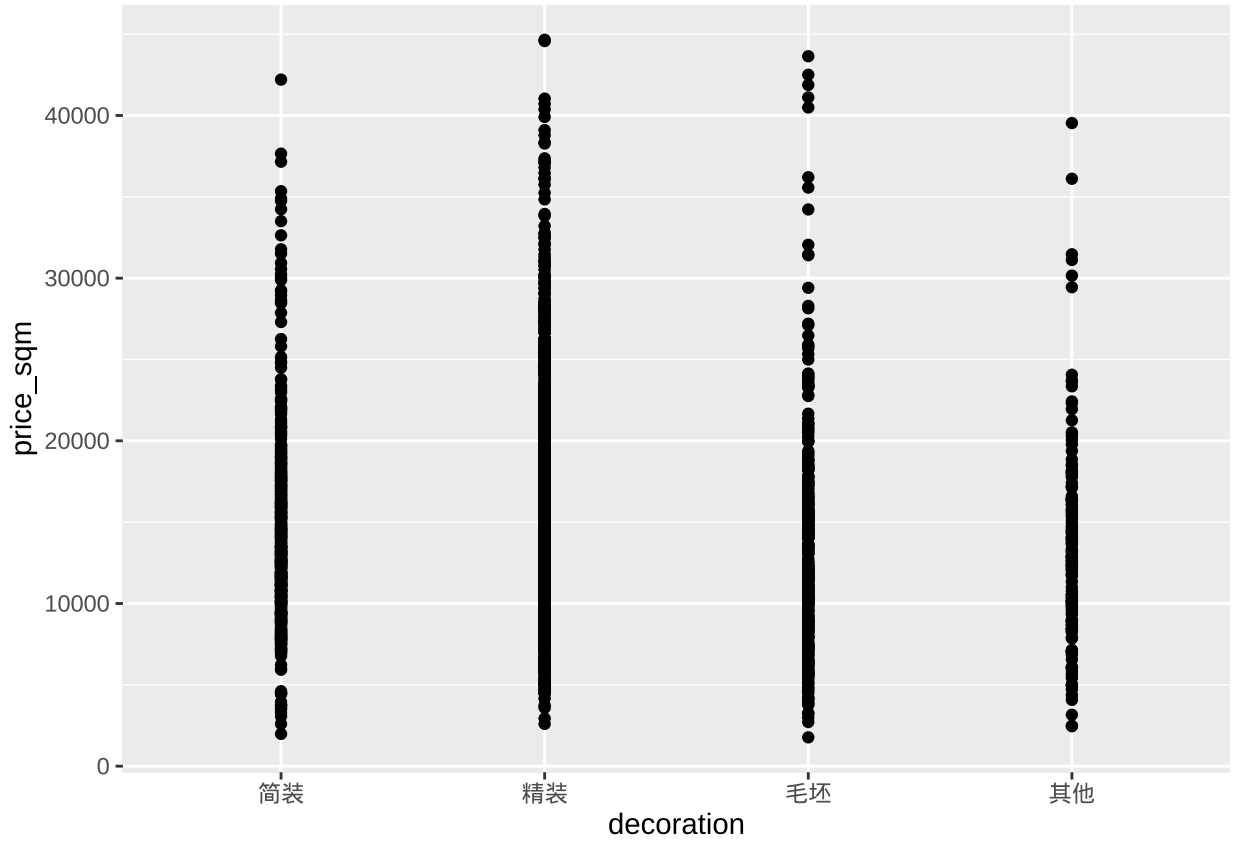


发现：

- 发现 1 房屋面积在 200 ㎡及以上，且设计的卧室数量均在 5 个及以上的，房屋朝向均为南；房屋面积在 100 ㎡以上，且设计的客厅数量均在 2 个及以上的，房屋朝向也为南; 由此可以发现，武汉二手房房屋面积分别在 200 ㎡左右以上、100 ㎡以上的，受朝向影响，越是偏南向，所设计的户型也会更加完善。
- 发现 2 房屋面积在 100 ㎡及以下，客厅的数量分布各个朝向均有；房屋面积在 100 ㎡以上，房屋朝向均为南; 由此可以发现，武汉二手房的房屋面积对客厅的数量影响较小，但 100 ㎡以上的户型较小面积，房屋朝向更优。

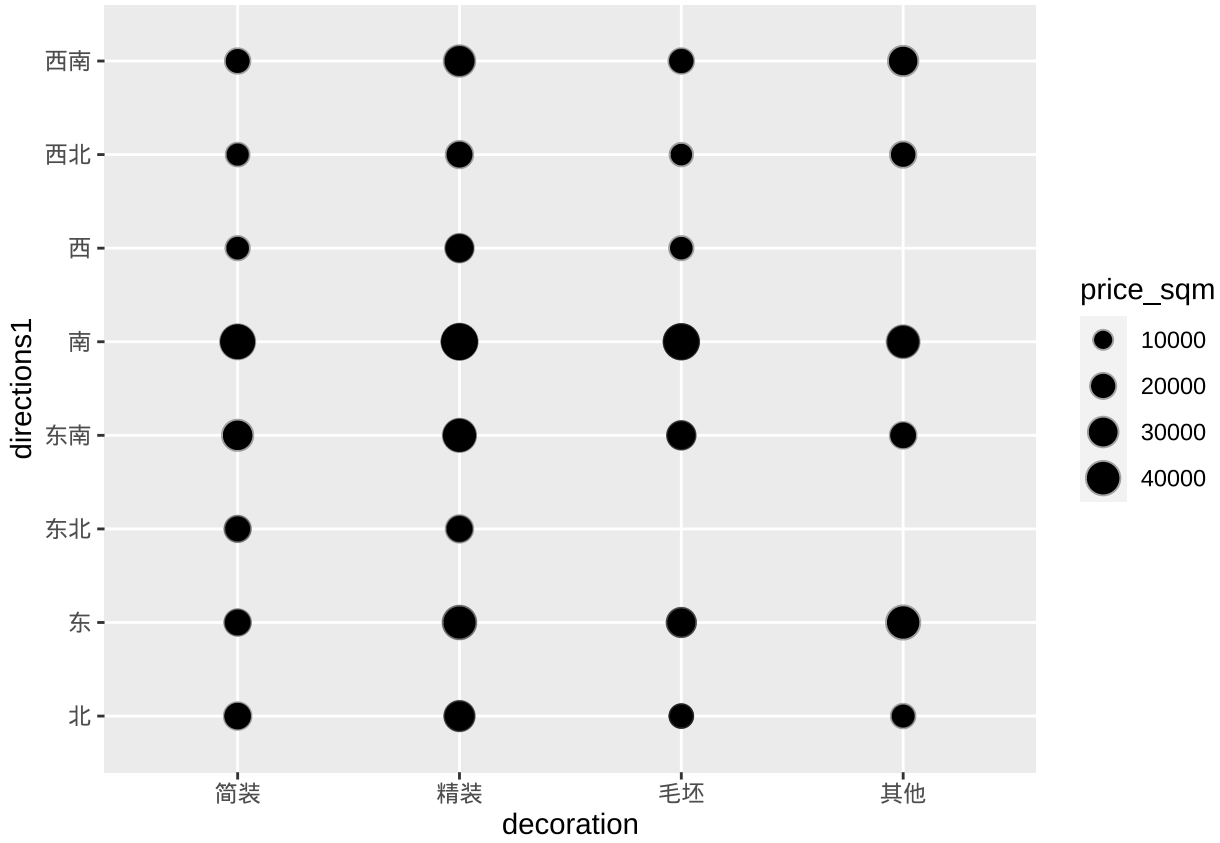
探索问题 3（房屋单价是否受房屋朝向、装修情况影响?）

```
ggplot(data = lj_wuhan)+
  geom_point(mapping = aes(x=decoration,y=price_sqm))
```




```
ggplot(data =lj_wuhan, mapping = aes(x =decoration, y =directions1)) +
   geom_point(aes(size =price_sqm ), alpha = 1/3) +
   geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



发现：

- 发现 1 房屋单价在 2.5 万及以下的，各类装修情况的数量分布差距较小。房屋单价在 2.5 万以上的，装修情况主要集中在简装及精装。由此可以发现，房屋单价在 2.5 万以下的，其单价受精装情况的影响较小。房屋单价在 2.5 万以上的，是否精装对单价存在一定影响。
- 发现 2 房屋单价最高的朝向分布在南，房屋单价最高的精装情况分布在简装、精装及毛坯。结合上述发现 1 可以发现，房屋单价受精装修情况的影响较小，朝向与房屋单价呈正相关。

发现总结

1、房屋单价：对武汉二手房房屋单价会产生较大影响的因素主要包含：房屋楼栋总层高、地理位置、精装情况及是否近地铁。由此可以发现，房屋单价的确定需要结合多方因素，尤其是地理位置对单价的影响最大。2、市场受众：武汉二手房的户型偏向 $3_{4\ \text{个卧室加}\ 1}$-2 个客厅的建构，朝南，精装。由此可以发现武汉的买家在选购中更偏向于实用性及精装的便利。3、建议：开发商或卖家在定价时。如选址在商圈或内环，建设超高层并设计完善的房间结构，那么会有更广泛的受众，可设置高位单价。如选址在外环或交通不便，需考虑周边买家的经济情况，从性价比出发，降低单价。