lj_homework

2024-10-31

你的主要发现

1. 发现1 武汉总价均值在156w ,总价中位数在137w。单价均值在15148元/平，单价中位数在14404元/平

价格分布呈现右偏分布，有部分高价房源拉高了均值

2. 发现2

热门区域集中在：白沙洲、盘龙城、四新等区域

单价最高地区 中南丁字桥 中北路 和 黄埔永清

3. 发现3 去掉无效的数据可以看到近地铁的均价是16628元/平 中位数是15622元/平

而不是近地铁的均价是13558元/平 中位数是12840元/平

可知近地铁的房源比非近地铁的房源更贵

但同一地区的近地铁和非近地铁房源的价格和价格没有明显相关性

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。

```
# 载入数据和预处理

library(tidyverse)
lj<- read_csv("C:/Users/75764/Desktop/dataScience/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)
```

数据概览

数据表（lj)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：

各变量的简短统计：

```
summary(lj)
```

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

```
names(lj)
```

```
##  [1] "property_name"     "property_region"   "price_ttl"
##  [4] "price_sqm"         "bedrooms"          "livingrooms"
##  [7] "building_area"     "directions1"       "directions2"
## [10] "decoration"        "property_t_height" "property_height"
## [13] "property_style"    "followers"         "near_subway"
## [16] "if_2y"             "has_key"           "vr"
```

```
ncol(lj)
```

```
## [1] 18
```

```
nrow(lj)
```

```
## [1] 3000
```

```
lj %>%
  head(10)
```

property_name <chr>	property_region <chr>	price_ttl <dbl>	price_sqm <dbl>	bedroo… <dbl>	livingrooms <dbl>	building_ar <dbl
南湖名都A区	南湖沃尔玛	237.0	18709	3	1	126.
万科紫悦湾	光谷东	127.0	14613	3	2	86.
东立国际	二七	75.0	15968	1	1	46.
新都汇	光谷广场	188.0	15702	3	2	119.
保利城一期	团结大道	182.0	17509	3	2	103.
加州橘郡	庙山	122.0	10376	3	2	117.
省建筑五公司西区	光谷广场	99.0	12346	2	1	80.
保利上城东区	白沙洲	193.8	16336	3	2	118.
石化大院	中南丁字桥	325.0	32631	4	1	99.
阳光花园	杨汊湖	192.0	17403	3	2	110.

1-10 of 10 rows | 1-8 of 18 columns

可以看到:

- 直观结论

有3000行数据，18个变量

探索性分析

价格数值描述与图形

```
lj %>%
   summarise(
     total_mean=mean(price_ttl,na.rm = TRUE),
     total_median=median(price_ttl,na.rm = TRUE),
     sqm_mean=mean(price_sqm,na.rm = TRUE),
     sqm_median=median(price_sqm,na.rm = TRUE),
   )
```

total_mean <dbl>	total_median <dbl>	sqm_mean <dbl>	sqm_median <dbl>
155.8628	137	15148.49	14404

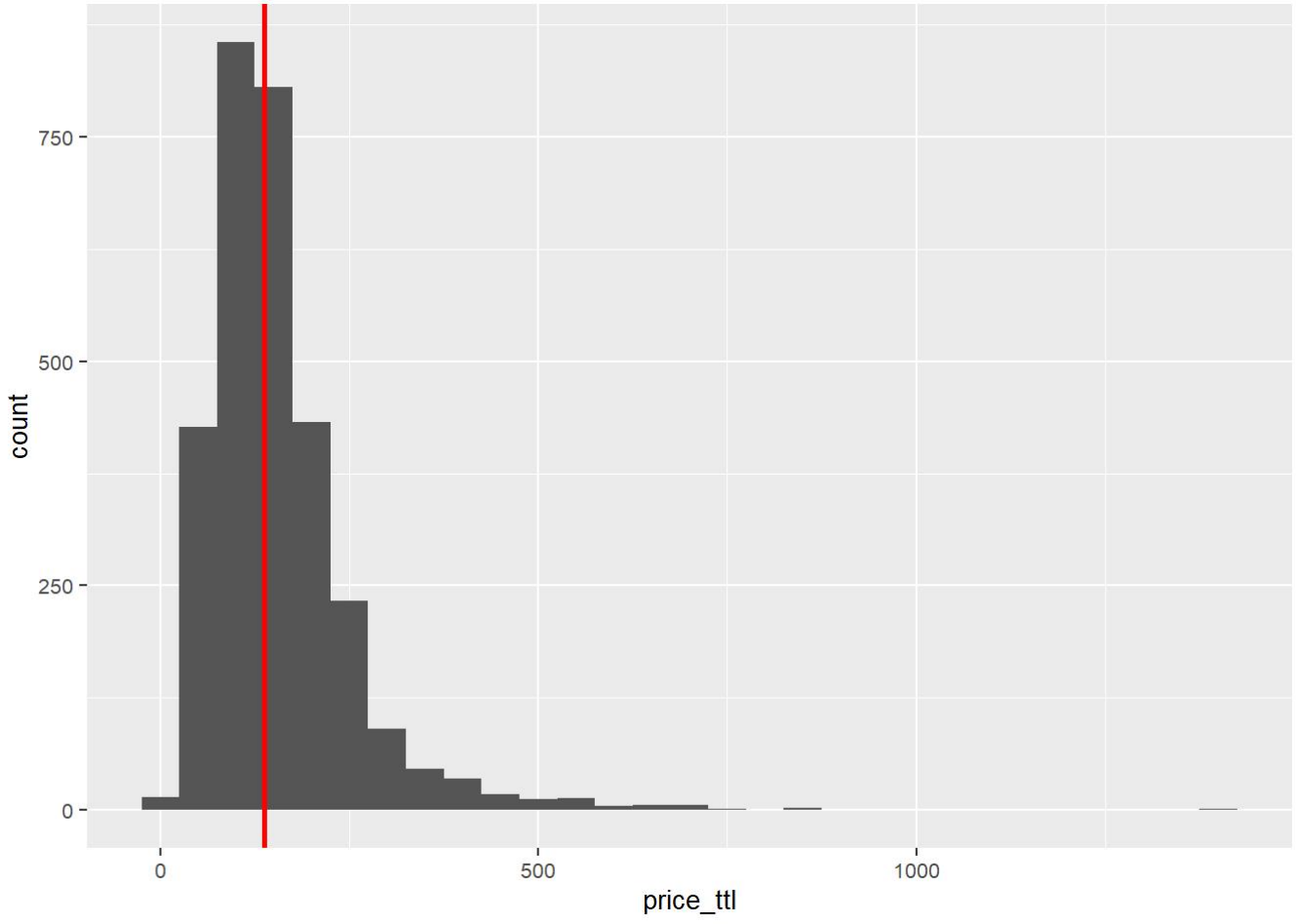
1 row

```
ggplot(lj)+geom_histogram(aes(x=price_ttl),binwidth = 50)+geom_vline(xintercept = median(lj$price_ttl,na.rm = TRUE), color = "red", size = 1)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)):
## Windows字体数据库里没有这样的字体系列
## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)):
## Windows字体数据库里没有这样的字体系列
```

```
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
```



- 发现1

武汉总价均值在156w ,总价中位数在137w。单价均值在15148元/平，单价中位数在14404元/平

- 发现2

价格分布呈现右偏分布，有部分高价房源拉高了均值

地区的数值描述与图形

```
word_freq <- lj %>%
  count(property_region) %>%
  rename(word = property_region, freq = n) %>%
  arrange(desc(freq))
word_freq
```

word <chr>	freq <int>
白沙洲	167
盘龙城	126
四新	116
光谷东	112
金银湖	97
后湖	86
青山	85
王家湾	78
塔子湖	71
珞狮南路	67

1-10 of 87 rows Previous **1** 2 3 4 5 6 … 9 Next

```
wordcloud2(
  word_freq,
  size = 1,
  fontFamily = "微软雅黑",
  # 如果有中文，设置中文字体
  color = "random-dark",
  backgroundColor = "white"
)
```


```
lj %>%
  select(property_name,property_region,price_sqm) %>%
  arrange(desc(price_sqm))
```

property_name <chr>	property_region <chr>	price_sqm <dbl>
中商宿舍	中南丁字桥	44656
复地东湖国际五六期	中北路	44574
复地东湖国际一期	中北路	43643
复地东湖国际一期	中北路	43643
复地东湖国际五六期	中北路	42503
复地东湖国际五六期	中北路	42205
华发外滩首府	黄埔永清	41878
华发外滩首府	黄埔永清	41110
华发中城荟	CBD西北湖	41037
复地东湖国际二期	中北路	40721

1-10 of 3,000 rows Previous **1** 2 3 4 5 6 … 300 Next

发现：

- 发现1

热门区域集中在：白沙洲、盘龙城、四新等区域

- 发现2

单价最高地区 中南丁字桥 中北路 和 黄埔永清

近地铁情况和价格的关系的数值描述与图形

```
clean_lj <- lj %>%
  filter(near_subway %in% c("近地铁", "近地看", NA)) %>%  # 只保留这些数据
  mutate(near_subway = case_when(
    near_subway == "近地铁" ~ "近地铁",
    near_subway == "近地看" ~ "近地铁",  # 将"近地看"归为"近地铁"
    TRUE ~ "非近地铁"  # NA值归为"非近地铁"
  ))

clean_lj %>%
  group_by(near_subway) %>%
  summarise(
    avg_price = mean(price_sqm, na.rm = TRUE),
    median_price = median(price_sqm, na.rm = TRUE),
    count = n()
  )
```

near_subway <chr>	avg_price <dbl>	median_price <dbl>	count <int>
近地铁	16628.19	15622	1555
非近地铁	13557.97	12840	1441

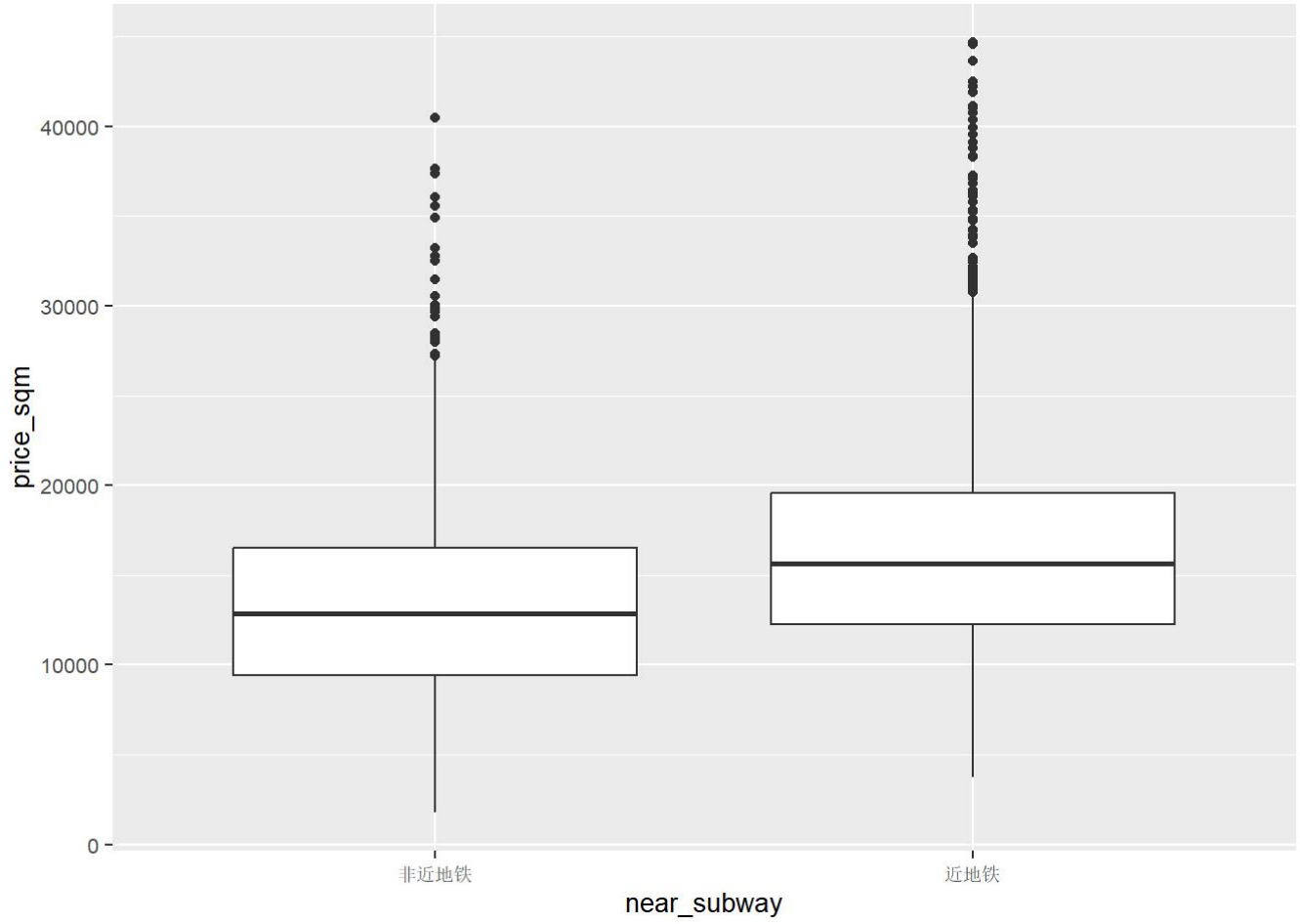
2 rows

箱线图观察

```
ggplot(clean_lj, aes(x = near_subway, y = price_sqm)) +
  geom_boxplot()
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
```

```
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
```



发现:

- 发现1

去掉无效的数据可以看到近地铁的均价是16628元/平 中位数是15622元/平 而不是近地铁的均价是13558元/平 中位数是12840元/平 可知近地铁的房源比非近地铁的房源更贵

考虑地区的影响，分析地铁和价格的影响

```
region_stats <- clean_lj %>%
  group_by(property_region, near_subway) %>%
  summarise(
    avg_price = round(mean(price_sqm, na.rm = TRUE), 0),
    count = n(),
    .groups = "drop"
  )
region_comparison <- region_stats %>%
  pivot_wider(
    id_cols = property_region,                          # 保持不变的列
    names_from = near_subway,                           # 用来创建新列名的列
    values_from = c(avg_price, count)     # 需要重组的值列
  )
region_comparison <- na.omit(region_comparison)
region_comparison
```

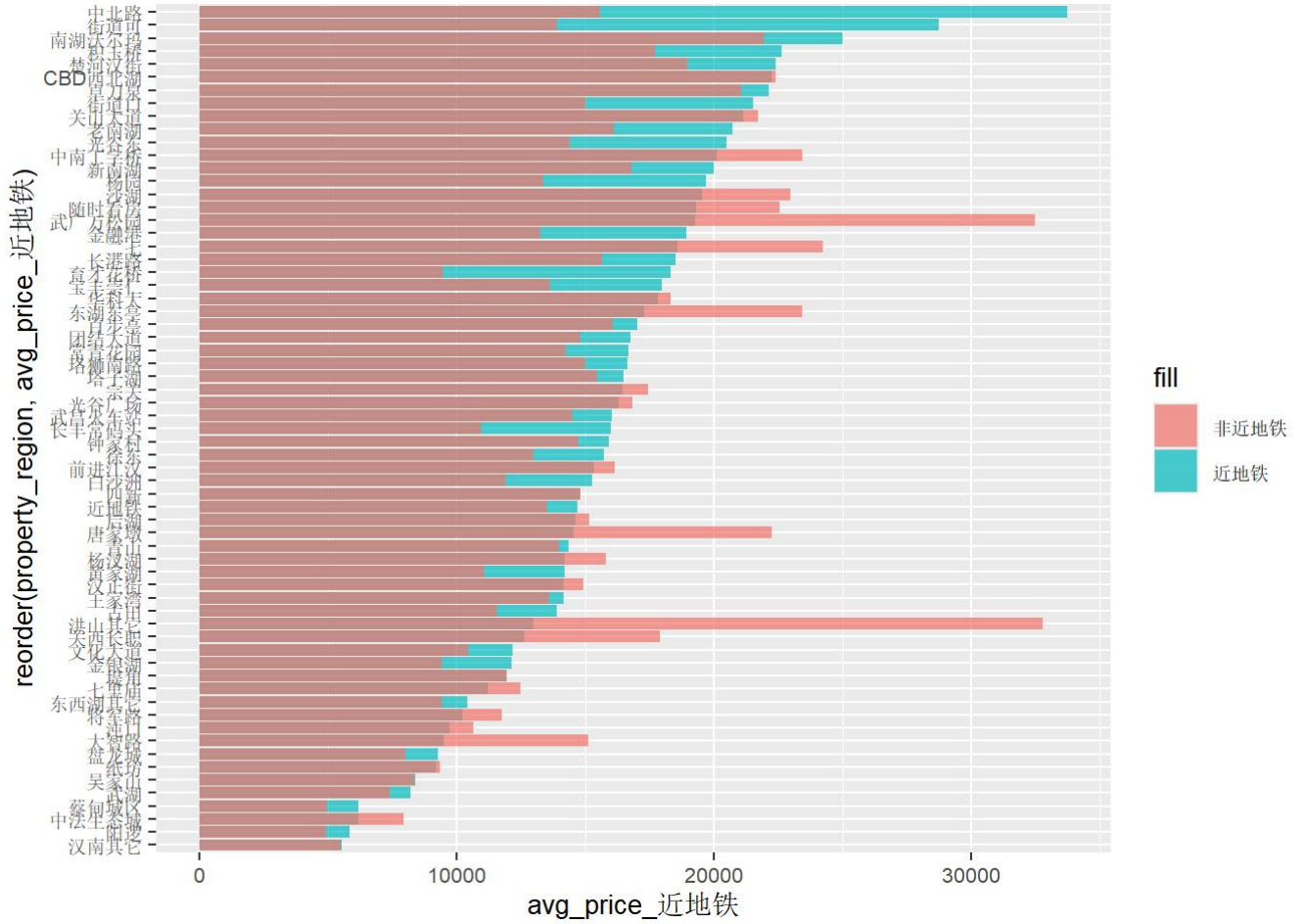
property_region <chr>	avg_price_近地铁 <dbl>	avg_price_非近地铁 <dbl>	count_近地铁 <int>	count_非近地铁 <int>
CBD西北湖	22263	22429	33	2
七里庙	11243	12493	17	18
东湖东亭	17302	23450	19	19
东西湖其它	10429	9450	5	2
中北路	33737	15572	17	1
中南丁字桥	20139	23453	32	19
中法生态城	6188	7942	6	6
二七	18610	24258	33	10
光谷东	20513	14386	42	70
光谷广场	16297	16842	16	9

1-10 of 65 rows Previous **1** 2 3 4 5 6 7 Next

```
ggplot(region_comparison, aes(x = reorder(property_region, avg_price_近地铁))) +
  geom_col(aes(y = avg_price_近地铁, fill = "近地铁"), alpha = 0.7) +
  geom_col(aes(y = avg_price_非近地铁, fill = "非近地铁"), alpha = 0.7) +
  coord_flip()   # 横向显示，便于查看地区名称
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
```

```
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## Windows字体数据库里没有这样的字体系列
```



不难看出大部分的地区都是近地铁的平均房价会更高

数据处理 计数统计

```
region_comparison <- region_comparison %>%
  mutate(price_diff = avg_price_近地铁 - avg_price_非近地铁)
```

统计各种情况

```
summary_stats <- list(
  total_regions = nrow(region_comparison),
  higher_near_subway = sum(region_comparison$price_diff > 0, na.rm = TRUE),
  lower_near_subway = sum(region_comparison$price_diff < 0, na.rm = TRUE),
  equal_price = sum(region_comparison$price_diff == 0, na.rm = TRUE)
)
summary_stats
```

```
## $total_regions
## [1] 65
##
## $higher_near_subway
## [1] 39
##
## $lower_near_subway
## [1] 26
##
## $equal_price
## [1] 0
```

发现：

计数表明总共有65个地区有近地铁和非近地铁房源，其中39个地区近地铁平均房价高于非近地铁房源，26个地区的近地铁平均房价低于非近地铁房源，0.6比0.4的情况，没有明显证据表明近地铁就一定房价更高

发现总结

武汉总价均值在156w,总价中位数在137w。单价均值在15148元/平，单价中位数在14404元/平。价格分布呈现右偏分布，有部分高价房源拉高了均值。热门区域集中在：白沙洲、盘龙城、四新等区域。单价最高地区 中南丁字桥 中北路 和 黄埔永清。去掉无效的数据可以看到近地铁的均价是16628元/平。中位数是15622元/平。而不是近地铁的均价是13558元/平。中位数是12840元/平。可知近地铁的房源比非近地铁的房源更贵。但同一地区的近地铁和非近地铁房源的价格和价格没有明显相关性