lj_investigate

wangsy

2025-09-15

目录

0.1 detail documents and code programing

0.1.1 code:For more details on using data analysis see https://github.com/tigerwangcqupt/r_projects-static_project_lj.git.

1 Analysis of ideas and method

1.1 ideas

1.1.1 property age is negatively correlated with price (older properties are priced lower)

1.1.2 building age depreciation coefficient = 1/(1+0.02*building_age) (assuming an annual depreciation rate of 2%)

1.1.3 cost-performance score = (Standardized area * 0.6 + Standardized building age coefficient * 0.4) / Standardized unit price

1.1.4 the top 10 houses with the highest cost-performance ratio have been selected

1.2 methods

1.2.1 collect Shanghai Lianjia housing data to form an R dataset.

1.2.2 using RStudio tools and R code libraries for data analysis

1.2.3 it requires the use of many R language libraries, including the library collection: tidyverse, dplyr, and so on.

1.2.4 using R code, output a scatter plot showing the relationship between the construction age and floor area of houses.

2 Analysis Process

2.1 detail process

2.1.1 calculate key cost-performance indicators and standardize the data

2.1.2 plot a 3D scatter chart to display the relationship between house age, area, and price

2.1.3 sort and output the top 10 listings by cost-performance score

2.1.4 use a blue-to-red gradient color scheme to enhance visualization

2.2 r programming process

```r
# Simulated Data Generation
library(ggplot2)
library(dplyr, exclude = c("intersect", "setdiff", "setequal", "union"))  # 排除特定函数
```

```
##
## 载入程序包：'dplyr'

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

```r
library(conflicted)
conflict_prefer("filter", "dplyr")  # 明确指定优先使用 dplyr 版本
```

```
## [conflicted] Will prefer dplyr::filter over any other package.
```

```r
conflict_prefer("lag", "dplyr")     # 或根据需求选择 stats 版本
```

```
## [conflicted] Will prefer dplyr::lag over any other package.
```

```r
load("../data/lj_sh_2019.RData")
houses <- lj %>%
  dplyr::mutate(building_age = 2025 - building_year) %>%
  dplyr::filter(building_age <= 20)
# Cost-effectiveness calculation
houses <- houses %>%
  dplyr::mutate(
    age_factor = 1/(1 + 0.02*building_age),
    value_score = (scale(building_area)*0.6 + scale(age_factor)*0.4)/scale(price_sqm)
  )
```

2.2.1 plot lj geom_point chart and data table

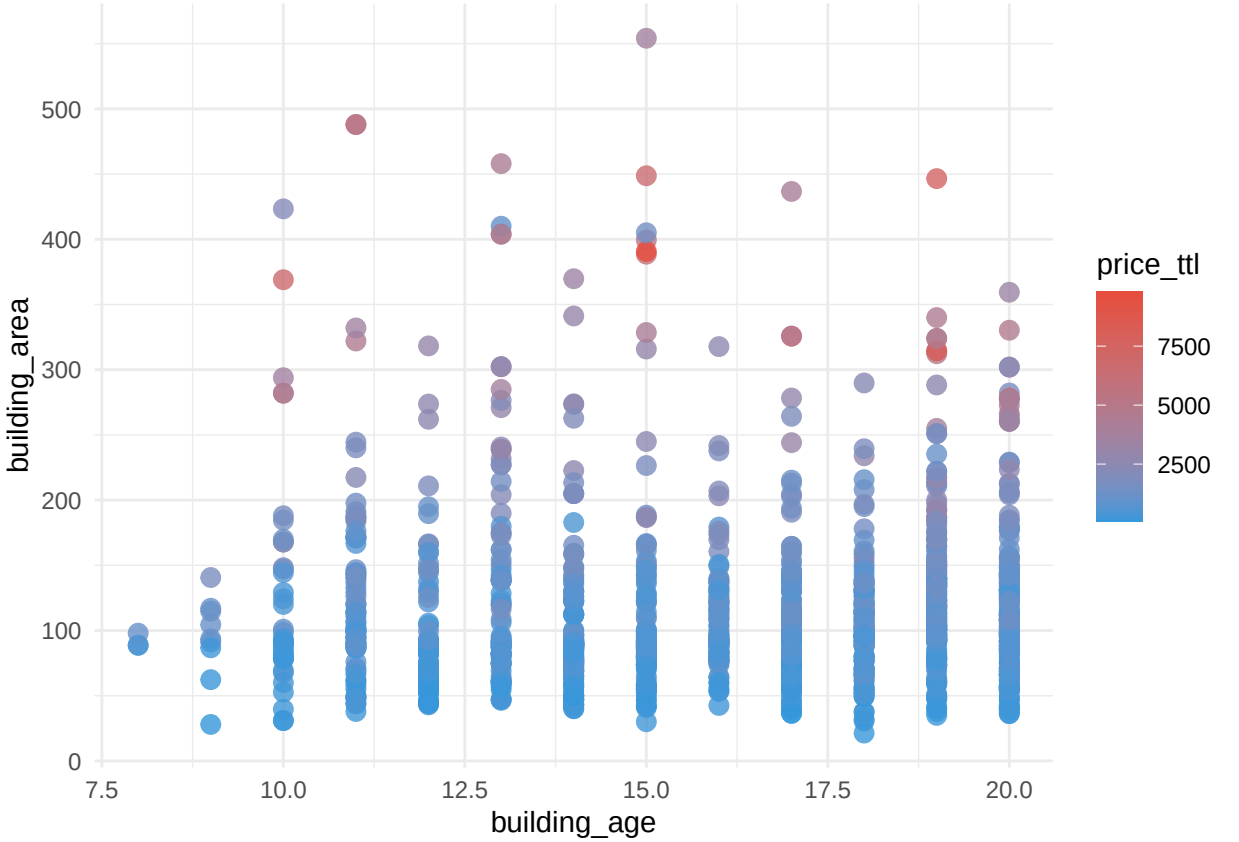

表 1: statistics Top 10 listings sorted by value-for-money score

property_name	building_age	building_area	price_ttl	price_sqm	value_score
和平花苑	20	136.42	970	71104	97.17200
九歌上郡	19	106.29	755	71033	87.30199
林绿家园	18	64.93	460	70846	65.14400
江南星城	20	176.83	1260	71255	48.82631
创智坊 (公寓)	19	117.70	835	70944	43.70462
世茂滨江花园	20	228.68	1639	71673	38.27174
建发璟墅	11	244.44	1780	72820	37.37225
风度国际	20	114.48	810	70755	35.62858
风度国际	20	114.48	810	70755	35.62858
靖宇家园	18	77.26	545	70542	29.47729