House Price in Nanjing: Comparing Multiple Linear Regression and Random Forest*

Yufei Liu

December 1, 2023

House price in Nanjing has grown rapidly and became a concern in recent years. This paper examines the relationship between house price in Nanjing, structural attributes of the property, and location using open data collected from Lianjia.com, and predict house price in Nanjing with different models. We find a geographically-imbalanced distribution of house price in Nanjing with large variance within and between each district. By comparing the MAE and RMSE, we find the random forest model gives a better prediction accuracy than the multiple linear regression model. Further work could include geocoding data to perform spatial autocorrelation analysis that consider the spatial effect in the data.

1 Introduction

House prices in China has grown rapidly since 2000, and cities such as Shanghai, Beijing, and Shenzhen are among the highest around the world today (**citeHouseincrease?**). Nanjing, which is located in the heart of Yangtze River Delta in eastern China and approximately 300 km from Shanghai, is the capital city for Jiangsu Province, one of the most economically developed provinces in China (**citeNanjing?**). The house price in Nanjing has been a concern due to increased population in recent years (**citeConcern?**).

In this paper, we investigate how structural attributes and location of house affect the total house price and explore the characteristics of residential housing. We are also interested in predicting house price in Nanjing using different models. We use`Python` to collect data from Lianjia.com and `R` to clean and analyze collected data. We construct a multiple linear regression model with total house price explained by 9 other variables representing structural characteristics and location, and a random forest regression model with same variables. We find that the house price in Nanjing is geographically-imbalanced and even the variance within

*Code and data are available at: https://github.com/Florence-Liu/house-price

each district is large, indicating possible income and wealth inequality. Comparing two models with Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), we find that the random forest model gives a better prediction accuracy on both test and train data, indicating possible non-linear relationships between variables. Further work could include geospatial data to find spatial correlations and consider a more comprehensive model including more attributes such as surronding environment.

The remainder of this paper is structured as follows: **?@sec-data** discusses the data with **?@sec-data_description** including information about data collection and data cleaning results, and **?@sec-data_visualization** including graphs and tables representing relationships between variables and distributions of variables; **?@sec-model** introduces the split of train/test data and two models: a multiple linear regression model and a random forest regression model, and specifies parameters; **?@sec-result** shows the estimates of fitted models and compares two models; **?@sec-discussion** includes discussions about model results and implications with **?@sec-weakness** talking about the weaknesses and future improvement.

2 Data

We used `Python` (**citePython?**) to obtain the datasets and `R` (R Core Team 2022) to do the analysis in this paper. We used packages `tidyverse` (Wickham et al. 2019), `stringr` (**citeStringr?**), and `here` (Müller 2020) to clean and load the data as well as create figures. We used package `latex2expand` (**citeLatex2expand?**) to add title to figures, `knitr` (Xie 2014) and `modelsummary` (**citeModelsummary?**) to generate tables, and `corrplot` (**citeCorrplot?**) to make the correlation plot. We also used package `randomForest` (**citeRandomforest?**) and `Metrics` (**citeMetrics?**) to build up a random forest model and compare the results with multiple linear regression model using different metrics. The color style of the figures was created referring to a R colors cheet-sheet (**citeRcolor?**).

2.1 Data description

The dataset in this analysis was obtained from Lianjia.com using a web scrapping program in Python (**citePython?**). Lianjia.com is the website of one of the largest estate brokerage firm in China and the source is open and accessible (**citeLianjia?**). We collected 11 datasets with over 30,000 observations for 11 different districts in Nanjing, China. To account for temporal variations of housing market, we specifically collected sales property price listed on 22 November, 2023 instead of posted transactions and to consider the consistency, we only collected data for residential houses and discarded data for other types of properties. The datasets includes listed sales prices and structural attributes of the properties including the floor area, the unit price per m^2, number of rooms, etc.

The original datasets were merged into one large dataset by districts with missing value removed. After obtaining 5 detailed structural characteristics by splitting `Structrual`

Table 1: Description for variables

Variable	Type	Definition
Total_Price	Continuous	Total house price in ten-thousand yuan
Unit_Price	Continuous	Unit house price in yuan per square meter
District	Character	District where the house is located
Area	Continuous	Floor area of the house in square meter
Furnished	Character	Decoration status of the house
Bedroom	Discrete	Number of bedrooms
Living_Room	Discrete	Number of living/dining rooms
Total_Floors	Discrete	Total floors of the building
Detailed_Floor	Discrete	Floor level of the house
Facing_South	Dummy	Whether the house is facing south

`attributes` variable in original dataset, we created two new variables: `Detailed_Floor` and `Facing_South`. The variable `Detailed_Floor` was obtained from the variable `Total floor`. If the property is in high floor, the detailed floor will be the total floor multiplies by 0.7; if the property is in medium floor, the detailed floor will be the total floor multiplies by 0.45; and if the property is in low floor, the detailed floor will be the total floor multiplies by 0.2. The varible `Facing_South` is a dummy variable with 1 representing the house is south-facing and 0 otherwise. Definitions and descriptions for the 10 variables are listed in Table 1. Most of them captured the structural attributes of the house and only variable `District` indicated an approximate location of the house.

2.2 Data visualiztion

Table 2 shows a summary for average total house price in ten-thousand yuan, average unit price in yuan/m^2, and average floor area in m^2 for 11 districts in Nanjing. We find that the highest average house price was in Jianye district while the lowest average house price was in Gaochun district, which is also consistent with the average unit price. However, the average floor area shows a different pattern that Qinhuai district had the smallest average floor area while Jianye and Pukou districts had the largest average floor area. The difference was not very large within 30 m^2. This may relate to geographical information for each districts since the area and population differ from each district. Different from relatively centered average floor area for each districts, the average house price and unit price show a larger variance. This may relate to business activities and industrial development in each district.

Figure 1, Figure 2, Figure 3 basically explore the statistical features of the 9 explanatory variables. Figure 1 shows the distribution for variable `Area`, `Total_Price`, and `Unit_Price`, Figure 2 shows the distribution for variable `Bedroom`, `Living_Room`, `Total_Floors`, and `Detailed_Floor`, and Figure 3 shows the proportion of house with 4 different decoration status `Furnished` and orientations `Facing_South` in each district. We find that the total

Table 2: Descriptive statistics about house price in each districts

District	Average total price	Average unit price	Average floor area
Gaochun	87	7429	114
Gulou	423	43679	95
Jiangning	277	23618	114
Jianye	541	43996	115
Lishui	104	9617	106
Liuhe	112	12852	87
Pukou	268	22384	115
Qinhuai	313	35973	85
Qixia	359	29722	113
Xuanwu	319	36517	86
Yuhuatai	317	30537	99

house price, the unit price, and the area were all right-skewed, which means there were less extremely high values for these variables, and the mean values were larger than the median ones. We can also find that most of the values centered in a range. The area mostly centered between 0 to 200 m^2, the total price mostly centered between 0 to 1,000 ten-thousand yuan, and the unit price had a larger range between 0 to 50,000 yuan.

From Figure 2 we find that most of listed houses had 2 or 3 bedrooms and 1 to 2 living/dining rooms. The number of rooms was relatively simple with a few houses having over 4 bedrooms and 3 living/dining rooms. However, for total floors of the building and detailed floor, they were various. We find for total floors, there are 3 bursts in approximately 5, 10, and 20, which corresponded to 3 common types of residential buildings in China. However, for detailed floor of listed house, we find that they mostly centered at lower level of the building. This may relate to the sunshine condition as the residential buildings were higher and more dense recently, a lower level house may have a bad sunshine and lighting condition.

We find from Figure 3 that the numbers of house each districts in our data were similar excluding Gaochun district. This related to the web scrapping program we used to obtained the data. However, it indicates that the total number of housel listed in Gaochun district was limited. For the orientation of the house, we find that relatively half of the house is south-facing and in Lishui and Gaochun district the proportions were even higher. It shows a tradition in China that people tend to choose house facing south, which according to Chinese traditional Feng Shui has positive effect on someone's luck (**citeFengshui?**). For decoration status of the house, the data was not that informative since type `Other` occupies a large proportion, which means the information is missing for the house. However, from existing data we find that fully furnished houses on sale has a larger proportion than partly furnished or not furnished house. This may relate to the data we collected that there were a lot of second-hand house on sale instead of newly closed. It is noticeable that house in Jianyi district had a much larger proportion of fully furnished house. This may related to the house type that there are

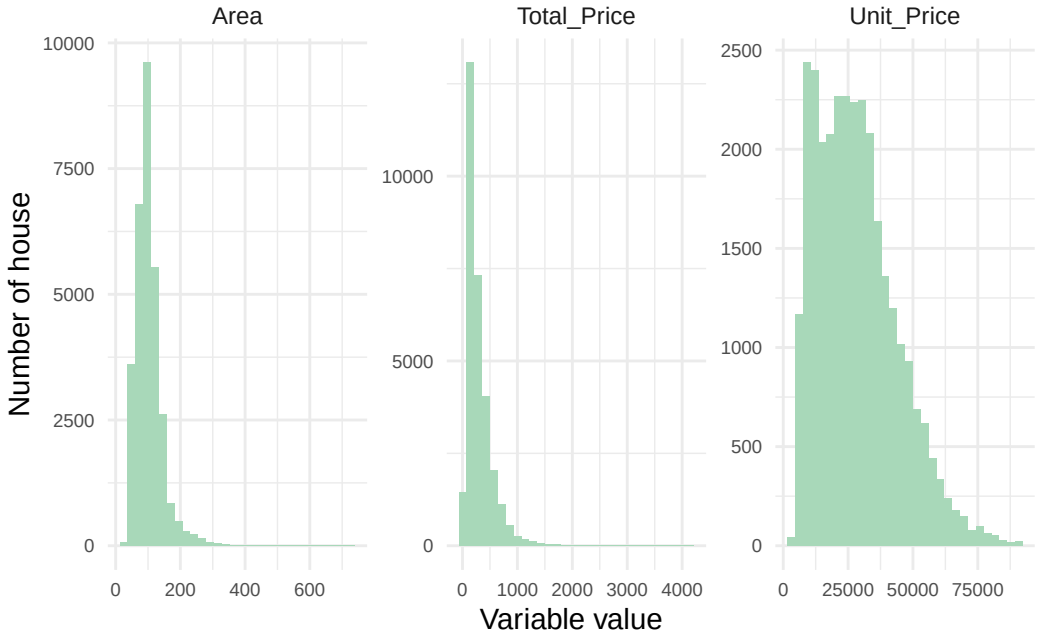


Figure 1: Distribution of Area, Total Price, and Unit Price

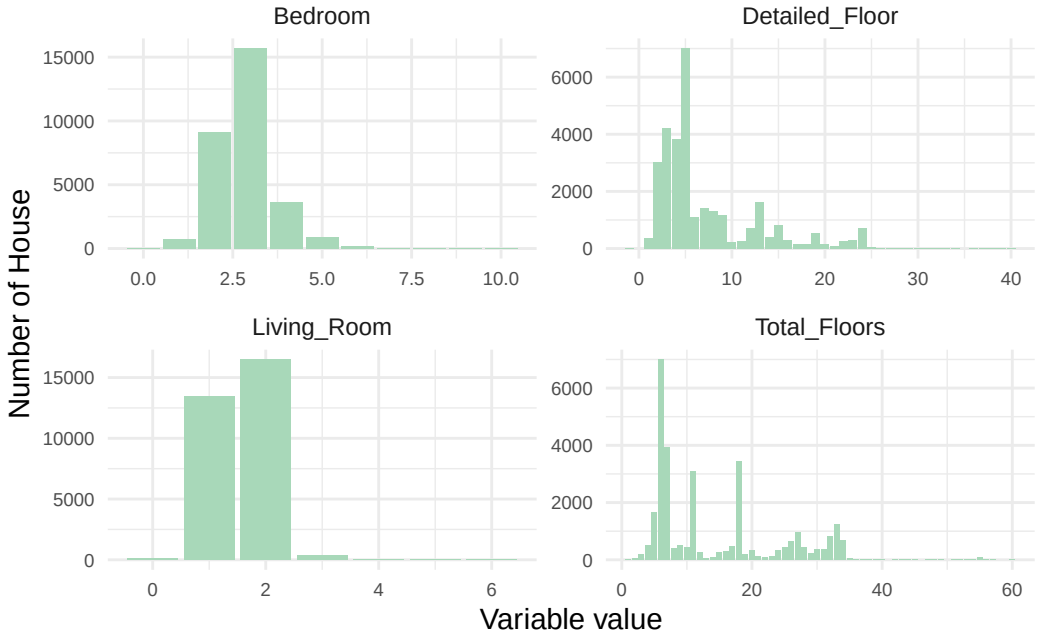
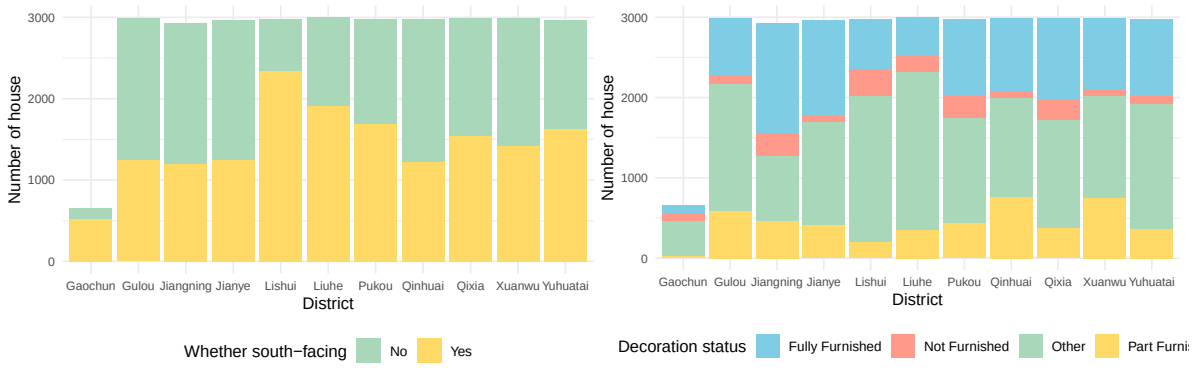


Figure 2: Number of house with different bedroom numbers, living room numbers, total floors, and detailed floor

more fully furnished luxury apartments in Jianyi since the Central Business District (CBD) is located there.



(a) South-facing

(b) Decoration

Figure 3: Proportion of house with difference decoration status and whether facing south in 11 districts

Figure 4 and Figure 5 shows the how the total house price in each districts and Nanjing city as a total related to area and unit price respectively. The blue line is the linear fitted line and the red line is a fitted line without designating methods. From Figure 4 we find that although the total price increases with area, the increase rate in each district was different with Lishui district had a notable slow rate compared with others. The increase rate indeed is just the unit price, so this is consistent with what we found from Table 2. We could also see that some districts had larger variance in area and total price which can simply inferred from axis scale. Pukou district had a large variance of total price with highest total price over 4000 ten-thousand yuan as well as for area.

Figure 5 further shows how total price changed with unit price. We find that although Table 2 shows the average unit price for each district was all smaller than 50,000, the highest unit price in most districts was even larger than 80,000. This is consistent with the distribution of unit price that it was right-skewed, a few extremely large values could significantly influence the average value. This is also true for total price that even the average value in each district was smaller than 600, there are still a lot of points at levels higher than even 1,500. This shows the imbalanced distribution of house price, which may also indicate income and wealth inequality.

3 Model

We will use two machine learning algorithms in this analysis: a multiple linear regression model, and a random forest regression model. We will randomly split our data into training and testing datasets with 80% being training data and 20% being testing data.

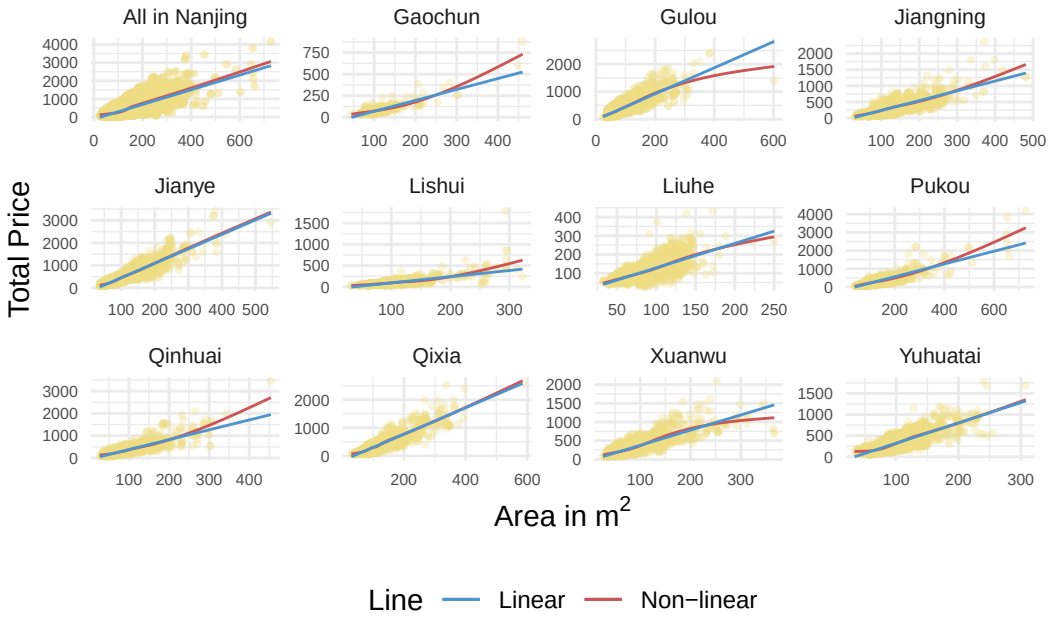


Figure 4: Relationship between total house price and floor area for 11 districts and Nanjing city in total

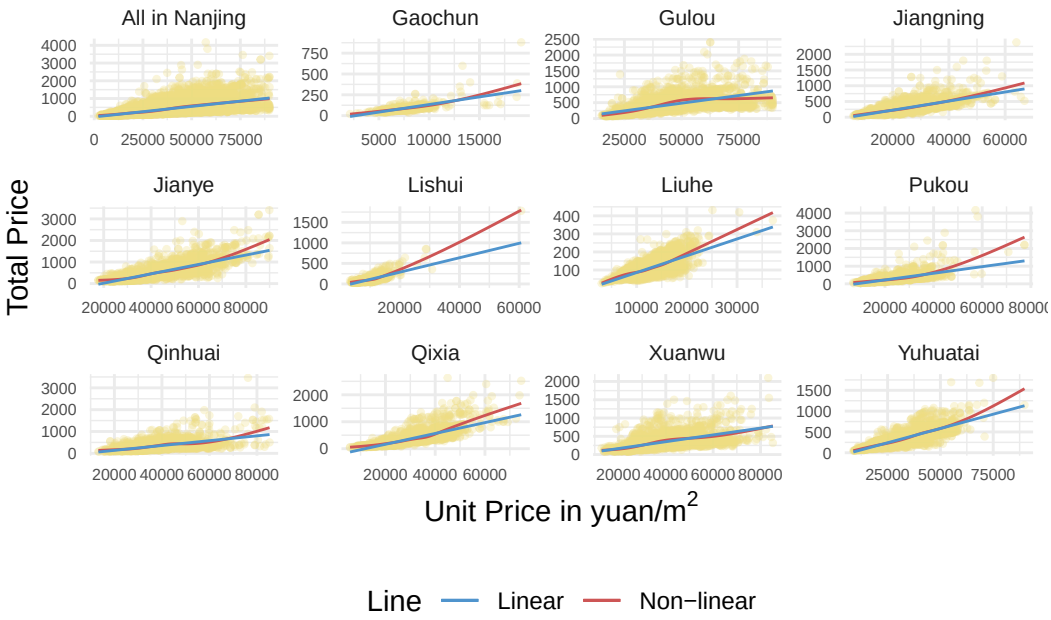


Figure 5: Relationship between total house price and unit price for 11 districts and Nanjing city in total

Table 3: Summary of different metrics for test and train data for two models

Model	MAE	RMSE
MLG train	40.9	68.7
MLG test	42.6	73.3
RF train	3.7	14.2
RF test	7.3	30.3

3.1 Multiple Linear Regression

The multiple linear regression model is shown as:

$$Y = \beta_0 + \sum_{n}^{i=1} \beta_i x_i + \epsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

where

- Y is the response variables `Total_Price`
- $x_i (i = 1, 2, \cdots, n)$ are explanatory variables. In this case we have 9 explanatory variables
- β_0 is the intercept of the model
- $b_i (i = 1, 2, \cdots, n)$ are regression coefficients
- ϵ is the random error

3.2 Random Forest Regression

4 Result

RUNNING TIME

5 Discussion

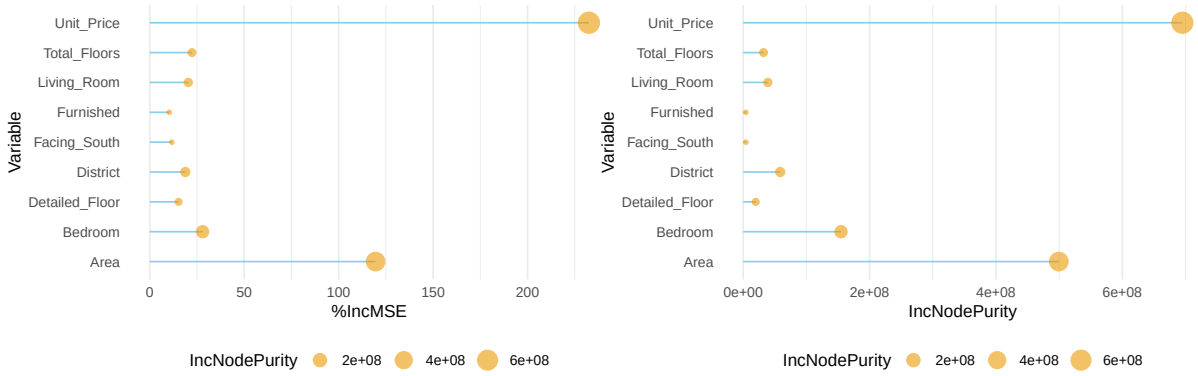
In 2013, the Nanjing Government adjusted the boundaries and administrative divisions of Nanjing. Since then, there are 11 districts in Nanjing with Gulou, Qinhuai, Xuanwu, Jianye districts considered as inner proper while Qixia, Yuhuatai, Pukou, Jiangning, Lishui, Gaochun, Liuhe districts are considered as suburban areas of Nanjing (**citeDivision?**).

residential structure, different from canadian. floor level regards to population density

district area difference, population difference

Table 4: Summary of the multiple linear regression model

	(1)
(Intercept)	−372.275
	(3.879)
Unit_Price	0.010
	(0.000)
DistrictGulou	28.151
	(3.812)
DistrictJiangning	23.994
	(3.483)
DistrictJianye	70.496
	(3.777)
DistrictLishui	18.433
	(3.369)
DistrictLiuhe	60.561
	(3.403)
DistrictPukou	22.117
	(3.440)
DistrictQinhuai	32.303
	(3.699)
DistrictQixia	44.620
	(3.526)
DistrictXuanwu	31.561
	(3.700)
DistrictYuhuatai	42.944
	(3.563)
Area	3.838
	(0.016)
FurnishedNot Furnished	1.336
	(1.969)
FurnishedOther	2.276
	(1.051)
FurnishedPart Furnished	−0.019
	(1.410)
Bedroom	−18.007
	(0.794)
Living_Room	−5.608
	(1.018)
Total_Floors	0.234
	(0.072)
Detailed_Floor	0.524
	(0.124)
Facing_South	2.026
	(0.916)
Num.Obs.	24 328
R2	0.924
R2 Adj.	0.924
AIC	274 897.2
BIC	275 075.4
Log.Lik.	−137 426.601
RMSE	68.72



(a) MSE (b) Node Purity

Figure 6: Variable importance of the random forest model based on MSE and Node Purity

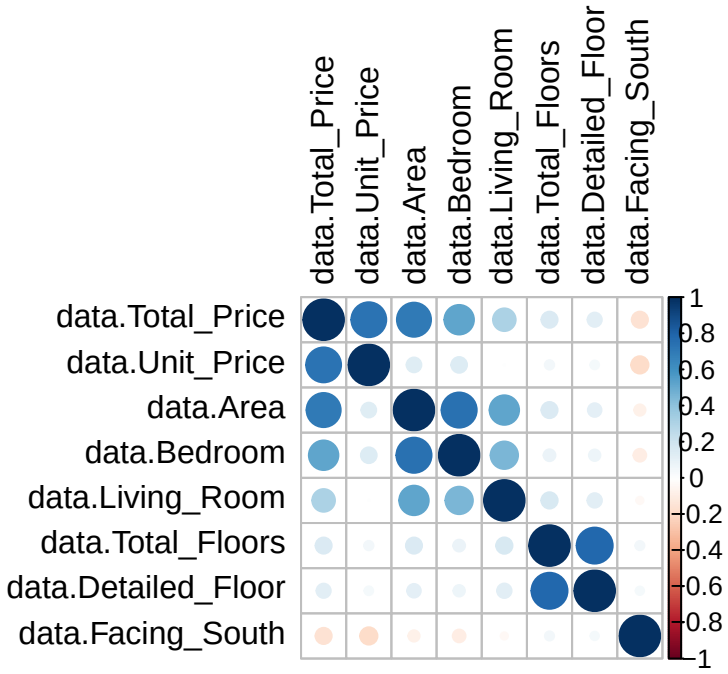


Figure 7: Correlation between each numeric variables

CBD, City center

less extreme high price → regulation, some are not listed publicly.

what type of house is on sale

income and wealth inequality

5.1 Weakness and future work

web scrapping maximum 100 pages

Hedonic pricing model, other spatial data for detailed analysis.

Reference

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.