Exploring and Predicting House Prices in Nanjing, China: Geographically Imbalanced Distribution Across Districts and the Crucial Role of Unit Price and Bedroom Count in Prediction*

Yufei Liu

December 16, 2023

House prices in Nanjing, China, have increased rapidly and became a concern in recent years. This paper examines the relationship between house prices in Nanjing, structural attributes of the property, and locations using open data collected from Lianjia.com. We then predict house prices in Nanjing with different models. We find a geographically-imbalanced distribution of house prices in Nanjing with large variance within and between each district. By comparing the MAE, RMSE, and R^2, we find the random forest model has the better prediction performance than the multiple linear regression model. Unit price and the number of bedrooms in the house tend to be importance features for predicting house prices in Nanjing. Further work could use spatial data to include the spatial effect in the model, and tune hyperparameters to improve model performance.

1 Introduction

House prices in China have grown rapidly since 2000, and prices in cities such as Shanghai, Beijing, and Shenzhen are among the highest in the world today (CEIBS 2021). Nanjing, which is located in the 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 (Nanjing Government 2014). House prices in Nanjing have been a concern due to increased population in recent years (Liu, Wei, and Wu 2023).

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

In this paper, we investigate how structural attributes and the location of house affect the house prices and explore the characteristics of residential housing. We are also interested in predicting house prices in Nanjing using different models. We use `Python` to collect data from Lianjia.com and `R` to clean and analyze the collected data. We construct a multiple linear regression model with total house prices explained by nine other variables representing structural characteristics and location, and a random forest regression model with the same variables. We find that house prices in Nanjing are geographically-imbalanced. Further, the variance within each district is large, reflecting possible income and wealth inequality. Comparing two models with Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2), we find that the random forest model gives the better prediction performance, indicating a possible non-linear relationship between variables. We also find that unit price and bedroom count play an important role in predicting house prices in Nanjing. The unit price refers to the price in thousand yuan per m^2. It is different from the total house price since the total price includes miscellaneous fees such as service charge, which differs for each house. Further work could include geospatial data to find spatial correlations and use a more comprehensive model including more factors such as surrounding environment and service amenities. We could also consider hyperparameter tuning and shrinkage methods to improve model performance for both the multiple linear regression model and the random forest model.

The remainder of this paper is structured as follows: Section 2 discusses data collection and data cleaning results, and visualizes the data by graphs and tables; Section 3 introduces two models: a multiple linear regression model and a random forest model, and specifies parameters; Section 4 shows the coefficient estimates of the multiple linear regression model, feature importance of the random forest model, and compares the performance of the two models; Section 5 discusses the findings and implications, as well as highlighting weaknesses and future work.

2 Data

We used `Python` (Python Core Team 2019) to obtain the datasets and `R` (R Core Team 2022) to do the analysis in this paper. We used packages `requests` (Reitz 2011), `parsel` (Scrapy developers 2015), and `csv` (Python Core Team 2023) in `Python` to scrape the data, and then in `R`, we used packages `tidyverse` (Wickham et al. 2019), `stringr` (Wickham 2022), and `here` (Müller 2020) to clean and load the data as well as create figures. We used package `latex2exp` (Meschiari 2022) to add labels to figures, `knitr` (Xie 2014), `kableExtra` (Zhu 2021), `broom` (Robinson, Hayes, and Couch 2023), and `gt` (Iannone et al. 2023) to generate tables, and `corrplot` (T. Wei and Simko 2021) to make the correlation plot. We also used package `randomForest` (Liaw and Wiener 2002) and `Metrics` (Hamner and Frasco 2018) to estimate random forest model and compare the results with the multiple linear regression model using different metrics. The color style of the figures was created referring to a R colors cheet-sheet (Y. Wei 2021).

Table 1: Description for variables in the house price dataset for Nanjing

Variable	Type	Definition
Total_Price	Continuous	Total house price in thousand yuan
Unit_Price	Continuous	Unit house price in thousand yuan per square meter
District	Categorical	District where the house is located
Area	Continuous	Floor area of the house in square meter
Furnished	Categorical	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

2.1 Data description

The dataset in this analysis was obtained from Lianjia.com using a web scraping program in Python. Lianjia.com is the website of one of the largest estate brokerage firm in China and it is open and accessible (Lianjia 2023). We collected 11 datasets with 30,653 observations for 11 different districts in Nanjing, China. To account for temporal variations in the housing market, we specifically collected sales property prices listed on 22 November 2023, instead of posted transactions. To enhance the consistency, we only collected data for residential houses and discarded data for other types of properties. The datasets include listed sales prices and structural attributes of the properties such as the floor area, the unit price per m^2, district, decoration status, number of bedrooms, number of living rooms, total floor of the building, detailed floor of the house, and whether the house is south-facing.

The original datasets were merged into one large dataset by district, with missing values removed. After obtaining five detailed structural characteristics by splitting the `Structrual 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 a high floor, the detailed floor will be the total floor multiplied by 0.7; if the property is in a medium floor, the detailed floor will be the total floor multiplied by 0.45; and if the property is in a low floor, the detailed floor will be the total floor multiplied by 0.2. The variable `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 capture the structural attributes of the house and only `District` indicates an approximate location of the house.

Table 2: Descriptive statistics about average house price, unit price, and floor area in each district

District	Average total price (thousand yuan)	Average unit price (thousand yuan/m^2)	Average floor area (m^2)
Gaochun	866.8	7.4	114.0
Gulou	4229.9	43.7	95.2
Jiangning	2770.4	23.6	114.1
Jianye	5406.0	44.0	114.6
Lishui	1038.3	9.6	105.9
Liuhe	1123.8	12.9	87.5
Pukou	2681.0	22.4	115.4
Qinhuai	3133.6	36.0	85.3
Qixia	3585.2	29.7	113.3
Xuanwu	3189.6	36.5	86.0
Yuhuatai	3172.9	30.5	99.1

2.2 Data visualization

Table 2 shows a summary for average total house price in thousand yuan, average unit price in thousand 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. The Qinhuai district had the smallest average floor area while Jianye and Pukou districts had the largest average floor area. The difference, at 30 m^2, was not very large. This may reflect geographical information for each district 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, and Figure 3 explore the statistical features of the nine 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`. Figure 3 shows the proportion of house with four different decoration status `Furnished` and orientations `Facing_South` in each district. We find that the total house price, the unit price, and the area were all right-skewed, which means there were some 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 1,000,000 yuan, and the unit price had a larger range between 0 to 50,000 yuan.

From Figure 2 we find that most of the listed houses had two to three bedrooms and one to

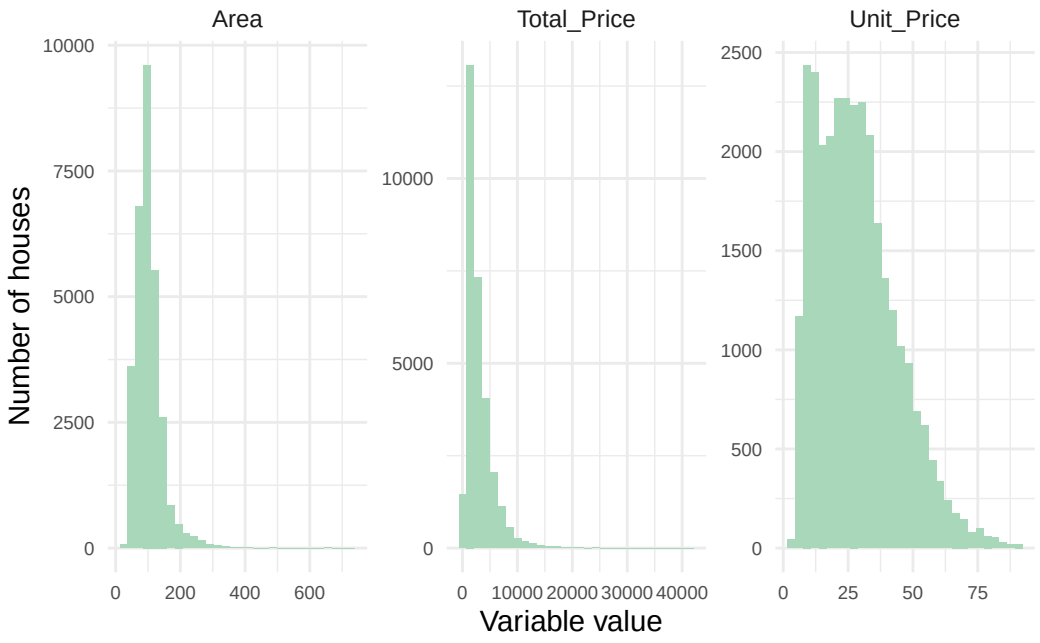


Figure 1: Distribution of floor area, house price, and unit price in Nanjing

two living/dining rooms. The number of rooms was relatively constrained with a few houses having over four bedrooms and three living/dining rooms. However, for total floors of the building and detailed floor, they were various. We find for total floors, there are three bursts in approximately 5, 10, and 20, which corresponded to three common types of residential buildings in China. However, for detailed floor of listed house, we find that they mostly centered at the lower level of the building. This may relate to the amount of sunshine. As the residential buildings were higher and denser recently, a lower level house may have a bad sunshine and lighting condition.

We find from Figure 3 that the number of houses in each district in our data were similar, excluding Gaochun district. This related to the web scraping program we used to obtain the data. However, it indicates that the total number of houses listed in Gaochun district was limited. For the orientation of the house, we find that relatively half of the houses were south-facing and in Lishui and Gaochun district the proportions were even higher. It shows a tradition in China that people tend to choose houses facing south, which according to Chinese traditional Feng Shui has positive effect on someone's luck (Beermann 2021). 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 have 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 houses in the Jianyi district had a much larger proportion of fully furnished house. This may be related to the

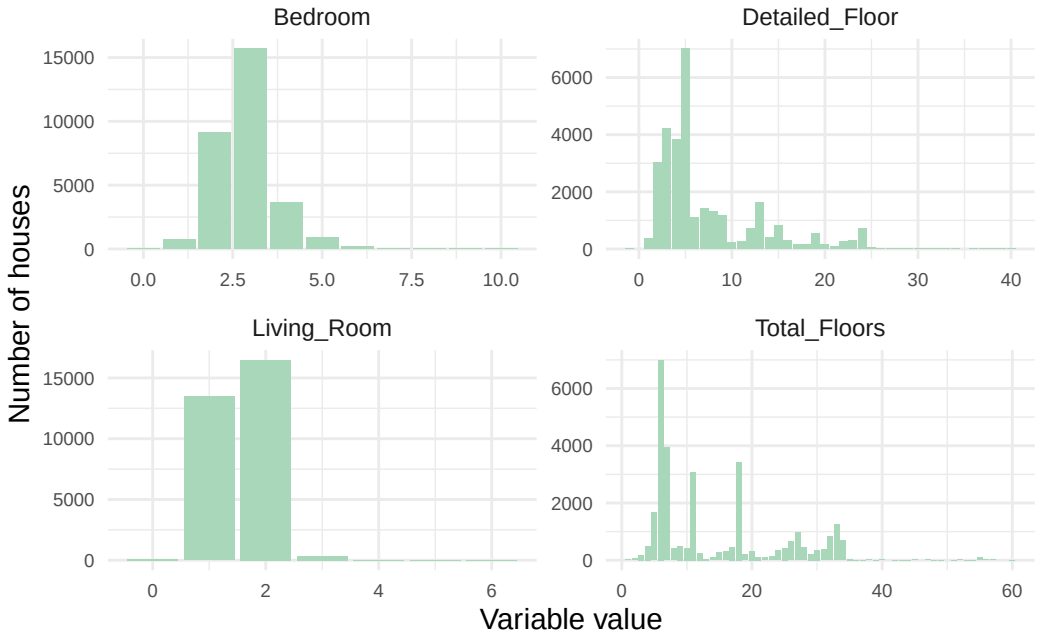
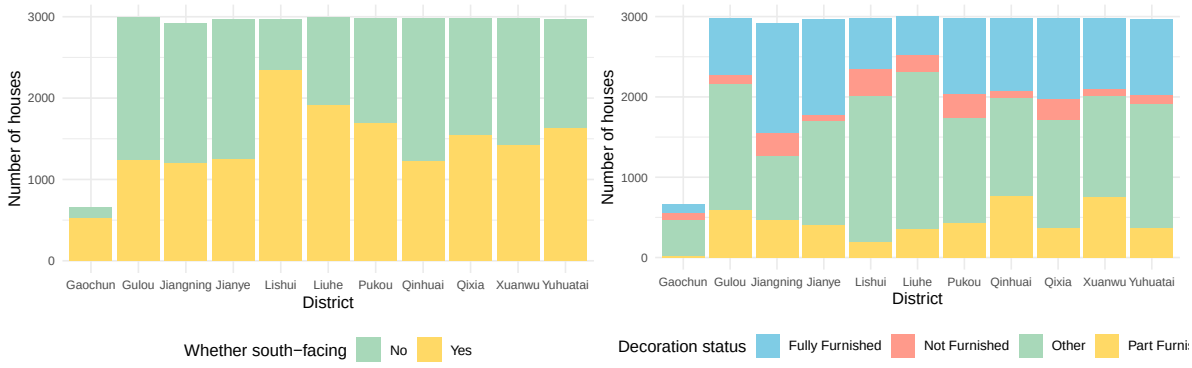


Figure 2: Number of houses with different bedroom numbers, living room numbers, total floors, and detailed floor in Nanjing

house type, as there are more fully furnished luxury apartments in Jianyi since the Central Business District (CBD) is located there.

Figure 4 and Figure 5 show how the total house prices in each district and Nanjing city in 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 in each district was different with Lishui district had a slower rate than the others. The increase rate is just the unit price, so this is consistent with Table 2. We could also see that some districts had larger variance in area and total price. Pukou district had a large variance of total price with highest total price over 40,000,000 yuan and largest area over 700 m^2.

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 prices, which may also indicate income and wealth inequality.



(a) South-facing

(b) Decoration

Figure 3: Proportion of houses with different decoration status and orientations in 11 districts

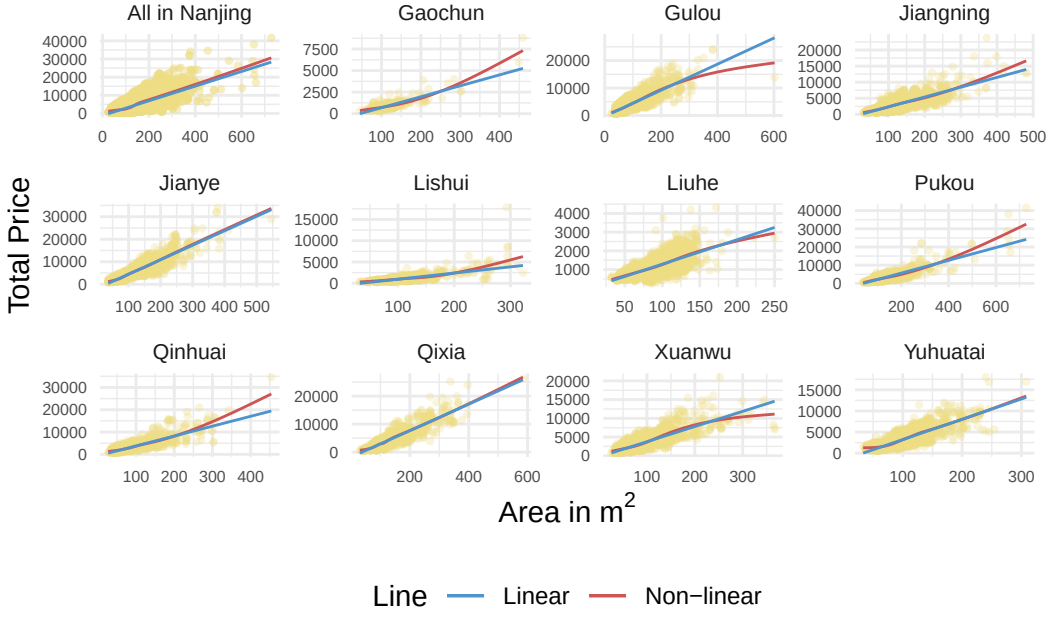


Figure 4: Relationship between house prices and floor area for 11 districts and Nanjing city in total

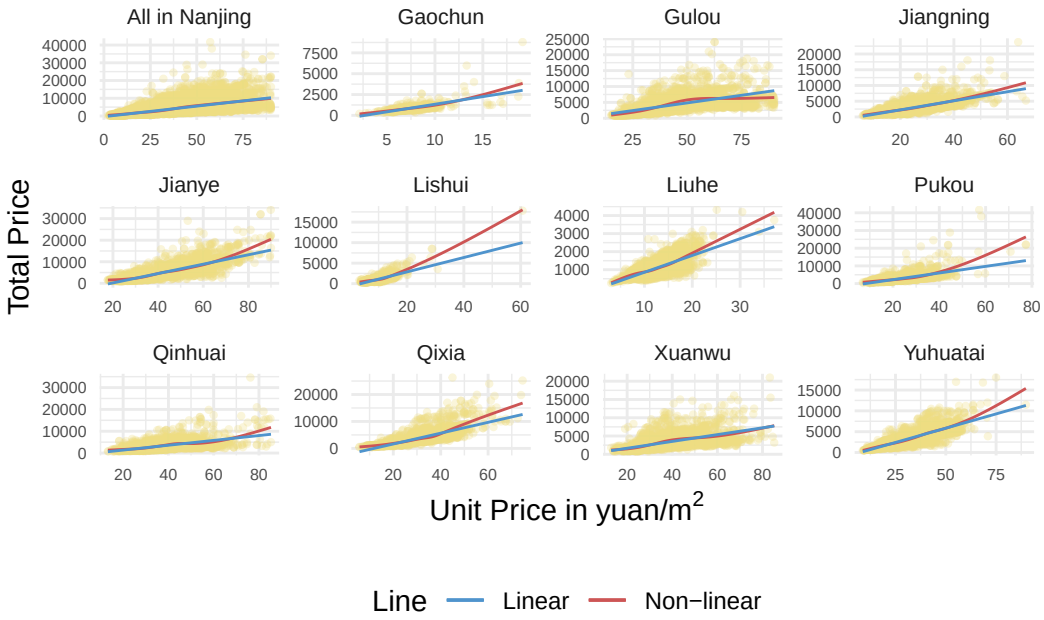


Figure 5: Relationship between total price and unit price for 11 districts and Nanjing city in total

3 Model

We will use two approaches in this analysis: a multiple linear regression model, and a random forest model. We will randomly split our data into training and testing datasets with 80% being training data and 20% being testing data. The training data contains 24,328 observations and the testing data contains 6,082 observations. Models will be performed on training data and assessed their performances based on training data, testing data, and all datasets.

3.1 Multiple Linear Regression

The multiple linear regression model is shown as:

$$Y = \beta_0 + \beta_1 \times X_{UnitPrice} + \beta_2 \times X_{District} + \beta_3 \times X_{Area} + \beta_4 \times X_{Furnished} + \beta_5 \times X_{Bedroom} + \beta_6 \times X_{LivingRoom} + \beta_7 \times X_{TotalFloor} + \beta_8 \times X_{DetailedFloor} + \beta_9 \times X_{FacingSouth} + \epsilon$$

where

- Y is the response variable: `Total_Price`

- $X_{UnitPrice}$, X_{Area}, $X_{Bedroom}$, $X_{LivingRoom}$, $X_{TotalFloor}$, $X_{DetailedFloor}$, $X_{FacingSouth}$ represent predictors `Unit_Price`, `Area`, `Bedroom`, `Living_Room`, `Total_Floors`, `Detailed_Floor`, and `South_Facing`
- $X_{District}$ represents the predictor `District` with ten dummy variables
- $X_{Furnished}$ represents the predictor `Furnished` with three dummy variables
- β_0 is the intercept of the model
- $\beta_i (i = 1, 2, \cdots, 9)$ are regression coefficients corresponding to 9 predictors
- ϵ is the random error

The multiple linear regression model is used to estimate the coefficients (β_i) by minimizing Residual Sum of Squares (RSS), and model the linear relationship between the house prices in Nanjing `Total_Price` and multiple predictor variables `Unit_Price`, `Area`, `District`, `Furnished`, `Bedroom`, `Living_Room`, `Total_Floors`, `Detailed_Floor`, and `Facing_South`. We will use function `lm` in `R` to fit the multiple linear regression model, it will give us estimates for the coefficients as well as the standard error and t-statistics. The p-value will also be presented, which indicates whether the predictor variable has a statistically significant relationship with `Total_Price`. In general, we set the significance level $\alpha = 0.05$.

3.2 Random Forest

Random forest is an ensemble learning method that works by growing multiple trees on training data and combining the predictions of the resulting trees (Hastie, Tibshirani, and Friedman 2009, 587). It improves prediction accuracy of trees and works for both classification and regression tasks. During the process, decision trees are built on random subsets of the training data with replacement and random selection of features, and the final prediction takes the mean of individual tree prediction.

A generalized algorithm for random forest with p predictors can be summarized in the following steps (Hastie, Tibshirani, and Friedman 2009, 588):

1. For $b = 1$ to B:
    a. Draw a random bootstrap sample Z^* of size N from the training data (randomly select N samples from the training set with replacement).
    b. Grow a random-forest tree T_b to the bootstrapped data, at each terminal node of the tree, recursively repeat the following steps until the minimum node size n_{min} is reached:
        i. Randomly select m variables from the p variables.
        ii. Split the node into two daughter nodes using features that provides the best split point.
2. The result is the ensemble of trees $\{T_b\}_1^B$.

The final prediction for a new data point x can be expressed as

$$\hat{f}_{\text{rf}}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

where B is the number of trees, and T_b is the individual regression tree.

In the process of random forest, the number of trees in the forest, the number of candidate variables at each split, and the minimum size of terminal nodes are three important parameters that affect the performance of the random forest model. A common choice of the number of candidate variable m is $m \approx \sqrt{p}$ for classification trees, and $m = \frac{p}{3}$ for regression trees (Hastie, Tibshirani, and Friedman 2009, 592). We will use the `randomForest` package in `R` to fit the random forest model. The function by default set the minimum size of nodes n_{min} to be 5, and number of trees to be 500. The number of candidate variables m is determined by $m = \frac{p}{3}$ for regression trees as mentioned before. We will assess the importance of each predictor variable in the random forest model by measuring either the percentage increase in Mean Square Error (MSE) or the increase in Node Purity for each split in trees.

4 Result

4.1 Multiplie Linear Regression

Table 3 shows the summary of the multiple linear regression model. We find that `Bedroom` and `Living_Room` have negative effects on the total price, and other predictor variables have positive relationship with total price. We also notice that `District`, `Unit_Price`, and `Bedroom` have relatively strong relationship with `Total_Price`. Predictor variables are significant at level p-value < 0.05 except for two dummy variables for `Furnished`. The dummy variables for `Not Furnished` and `Part Furnished` show large p-values, indicating they are insignificant with relation to `Total_Price`.

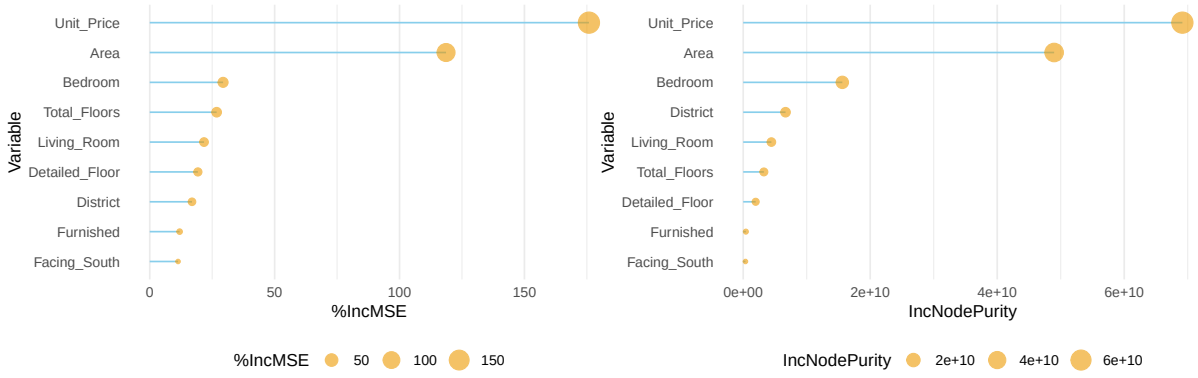
Table 3: Summary of the coefficient estimates for multiple linear regression model

	Multiple Linear Regression			
term	estimate	std.error	statistic	p.value
(Intercept)	−3,722.7	38.8	−96.0	0.000
Unit_Price	103.4	0.4	233.1	0.000
DistrictGulou	281.5	38.1	7.4	0.000
DistrictJiangning	239.9	34.8	6.9	0.000
DistrictJianye	705.0	37.8	18.7	0.000
DistrictLishui	184.3	33.7	5.5	0.000
DistrictLiuhe	605.6	34.0	17.8	0.000

DistrictPukou	221.2	34.4	6.4	0.000
DistrictQinhuai	323.0	37.0	8.7	0.000
DistrictQixia	446.2	35.3	12.7	0.000
DistrictXuanwu	315.6	37.0	8.5	0.000
DistrictYuhuatai	429.4	35.6	12.1	0.000
Area	38.4	0.2	234.6	0.000
FurnishedNot Furnished	13.4	19.7	0.7	0.498
FurnishedOther	22.8	10.5	2.2	0.030
FurnishedPart Furnished	−0.2	14.1	0.0	0.989
Bedroom	−180.1	7.9	−22.7	0.000
Living_Room	−56.1	10.2	−5.5	0.000
Total_Floors	2.3	0.7	3.3	0.001
Detailed_Floor	5.2	1.2	4.2	0.000
Facing_South	20.3	9.2	2.2	0.027

4.2 Random Forest

Figure 6 shows the importance of variable of the random forest model based on two metrics, Mean Square Error (MSE), and Node Purity. The percentage increase in MSE (%IncMSE) is calculated by how much in percentage MSE increases without the predictor. The higher value indicates more important features for making predictions. The increase in Node Purity (IncNodePurity) measure by how much node purity increases when splitting on a specific feature. It is usually calculated by training RSS, and same as %IncMSE, the higher value indicates higher influence on prediction performance (Hastie, Tibshirani, and Friedman 2009, 593). From Figure 6 we find that the three most important predictors are the same based on %IncMSE and IncNodePurity, which are `Unit_Price`, `Area`, and `Bedroom`. The two least important predictors are also the same, which are `Furnished` and `Facing_South`.



(a) MSE (b) Node Purity

Figure 6: Variable importance of the Random Forest model based on MSE and Node Purity

Table 5: Results of MAE, RMSE, and R^2 for training data, testing data, and all datasets for multiple linear regression model (MLR) and random forest model (RF)

Model	MAE	RMSE	R^2
MLR all	412.7	696.5	0.923
MLR train	409.4	687.2	0.924
MLR test	425.9	732.7	0.917
RF all	44.9	185.7	0.995
RF train	37.4	141.7	0.997
RF test	75.1	303.4	0.987

4.3 Model Evaluation

Table 4: Mathematical expression for 3 evaluation metrics: MAE, RMSE, and R^2

Metrics	Expression
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
R^2	$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$

There are three metrics being used to evaluate the two models: MAE, RMSE, and R^2 (James et al. 2017). Mean Absolute Error (MAE) measures the average absolute difference between the predicted value and actual values. Root Mean Squared Error (RMSE) measures the square root of the average squared difference between the predicted value and the actual value. The coefficient of determination (R^2) measures the proportion of the total variability in the response variable that can be explained by the model. The expressions for the three metrics are shown in Table 4. Lower values in MAE and RMSE and higher value (between 0 and 1) in R^2 suggest better model performance.

Table 5 shows the results for two models on both training and testing data based on three metrics. We find that overall the random forest model yields a better performance. It has a lower MAE and RMSE compared with multiple linear regression model as well as a higher R^2. However, both models have a R^2 larger than 0.9, indicating both models could explain over 90% of the total variability in `Total_Price`. We also find that there is no large difference between training and testing performance, suggesting there is minimal overfitting for both models.

5 Discussion

5.1 House prices distribution in Nanjing

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 (Yuan, Gao, and Wu 2016). The difference in house prices between inner city and suburban areas was also revealed in this paper. From Table 2 we find that both the average unit price and average total price in inner proper districts are much higher than that of suburban districts. The highest average total price is in the Jianye district with value of 5,406,000 while the lowest average total price is in the Gauchun district with value of 866,800. The large difference in average total price across districts indicates disparities in economical development in each districts. The average floor area in suburban districts is slighter larger than that of inner city. It is consistent with the geographical features of the districts. Suburban areas usually have more available space for real estate, and the real estate developers have difference business strategies and target consumer for inner and suburban areas. Houses in suburban areas are usually designed as two types: one is affordable housing targeted at consumers without much money for the down payment, another is luxury villas targeted at consumers who have more money. Former type of house usually has smaller floor area and lower total price due to its location and transportation convenience. Villas are designed to have larger size and better quality with good environment. As a result, we could see from Figure 4 and Figure 5, both area and total price in suburban areas vary a lot within each districts especially for Pukou district. The observed variation in housing conditions not only reveals the diverse economic status but also emphasizes the potential disparities in overall quality of life within specific districts. This difference in house quality and living standards among residents within the same district reflects the broader issues of wealth and income inequality as well as social class.

The economical development in each district also contributes to the difference in house prices across districts. There are two main commercial districts in Nanjing, one in the city center in Gulou district and one in the CBD in Jianye district. Due to limited available space and large population density, high-rise condominiums and luxury apartments are two main types of housing in Gulou and Jianye. These two types of houses usually demand higher pricse due to the modern amenities and location. In addition, commercial districts are accompanied by more job opportunities and higher income level, which attracts those who are seeking more convenient living environment and houses closer to companies to shorten commuting time. The high demand for housing in commercial districts also affects the rental market, thus increasing investment enthusiasm. For Gulou district, there is a third type of house, old apartment built early in 1980s. Since it was developed early, the amenities sometimes function improperly, and management is lax and inefficient with small floor areas. However, such houses still have high total prices. This is related to their surrounding facilities especially schools. In Nanjing, public schools in the compulsory education period that a child may attend is determined

by the household registration system, that is primary school and middle school (Wu 2011). As a result, parents who wish their children to attend a better public school may choose to buy an apartment within the school district (Li et al. 2019). Since it was developed early, many qualified and renowned public schools are located in Gulou district, resulting in the high demand of houses in school district divisions and thus higher house prices.

We also notice the highest house price on sale in Nanjing was about 40,000,000 yuan, which is relatively not very high compared with other economically-developed cities such as Shanghai. One possible reason is that some houses are not listed and sold publicly. Private sales are common among luxurious houses for which price could be over hundreds of million yuan. Another possible reason is government regulation and economic recession. The Chinese government has implemented several policies including land policies, fiscal policies, and monetary policies in various cities to cool down the heated market and housing speculations after 2021 (Hu 2022). These policies had successfully cooled down the market, however, in an unexpected speed. The sudden policy change and higher unemployment rate have made buying houses a prohibitive choice, and struck the real estate developers who had not been recovered completely from pandemic (Ling, Wang, and Zhou 2020). Although in 2023, the Nanjing government had announced several promoting policies to help revive the market including decreased interest rate and cancellation of preconditions for buying houses in inner city, the affect was limited (Reuters 2023). Consumers took a low expectation on current market, greatly reduced investments and speculations on houses. As a result, the demand of house has been decreased compared with 2018, thus affecting the house prices dynamics.

5.2 Structral attributes characteristics and performace on predicting house prices

We have considered specific structural attributes including `Area`, `Furnished`, `Bedroom`, `Living_Room`, `Total_Floors`, `Detailed_Floor`, and `Facing_South`. The floor area distribution reveals a right-skewed pattern, indicating that a significant number of houses falls within the range of 0 to 200 square meters, which also suggests the prevalence of moderate-sized houses. The number of bedrooms centered at two to three with the number of living rooms centered at one to two, which is a common configuration for houses in Nanjing. It reflects the demand of house for a family of three to four peoples occupies the market. The concentration of total floors of the house in Nanjing at 10, 15, and 20 corresponds low-, medium-, and high-rise apartments in Nanjing. There is a few houses having total floors more than 30, which reflects tacit rule that in East of Nanjing where the Purple Mountain located, i.e., Qixia district, the height of the building should not exceed the height of Purple Mountain. It is also designed to protect military confidentiality. Most of high-rise apartments were newly built and mostly in West of Nanjing, i.e., Jianye district. Half of the houses on sale were south-facing, which had enough sunshine exposure and naturally bright interiors. According to Chinese traditional Feng Shui, the most auspicious direction of house is south-facing, which could bring luck and good for family harmony. The prevalence of low-rise house on sale was also associated with sunshine condition. Since the residential buildings were built more

densely in recent, years, the gap between building became smaller, resulting in few bright and sunshine in low-rise houses. The characteristics of structural attributes reflects family structure, cultural influence, and environmental considerations behind house prices.

In predicting house prices, these attributes weights differently. For the multiple linear regression model `District`, `Unit_Price`, and `Bed_room` shows a strong relationship with `Total_Price` since the magnitudes of the estimates for coefficients are larger than 100. However, `Total_Floors` and `Detailed_Floor` have relatively weak relationship with `Total_Price`. For the random forest model, `Area`, `Unit_Price`, and `Bedroom` were more importance in predicting `Total_Price`. In general, `Unit_Price` and `Bedroom` consistently emerge as crucial features for both models, indicating their significance in predicting `Total_Price`.

5.3 Comparison between the models

As discussion in Section 4.3, the random forest model gave a better prediction performance compared with multiple linear regression model on training data, testing data, and all datasets. However, the multiple linear regression model took the advantage of fast training time. The random forest model, especially with large datasets and a large number of trees and predictors, can be computationally expensive. In our analysis with approximately 24,000 observations and 9 predictors, the random forest model took ten times as much time to train the data compared with multiple linear regression model. Additionally, the multiple linear regression model has a relatively simple model structure that can be expressed mathematically, and gives estimates for coefficients. The estimates of coefficients indicate both direction and strength of relationship between each predictor and the response variable `Total_Price`, and are more interpretable than the random forest model.

Several factors contribute to the better prediction performance of random forest model. According to Figure 5 and Figure 4, the relationships between `Area`, `Unit_Price`, and `Total_Price` for most districts in Nanjing are non-linear. In this case, the multiple linear regression is limited to capture non-linear relationships, thus resulting in worse predictions. Other than non-linear relationships in the dataset, we find that from Figure 7, `Area` and `Bedroom` are highly positively correlated as well as `Total_Floors` and `Detailed_Floor`. The correlation between `Area` and `Living_Room`, and the correlation between `Bedroom` and `Living_Room` are also positive and quite large. It indicates there exists multicolineaerity in the predictors. The MLG model is sensitive to multicolinearity while the random forest model is more robust, resulting a better prediction performance of random forest model. Another concern is about outliers. Since from Figure 1, the distributions of `Total_Price`, `Area`, and `Unit_Price` are all right-skewed, the dataset contains some extremely high values and possible outliers. In this case, the random forest model is also more robust to outliers while the multiple linear regression model is more sensitive, contributing to the difference in prediction performance.

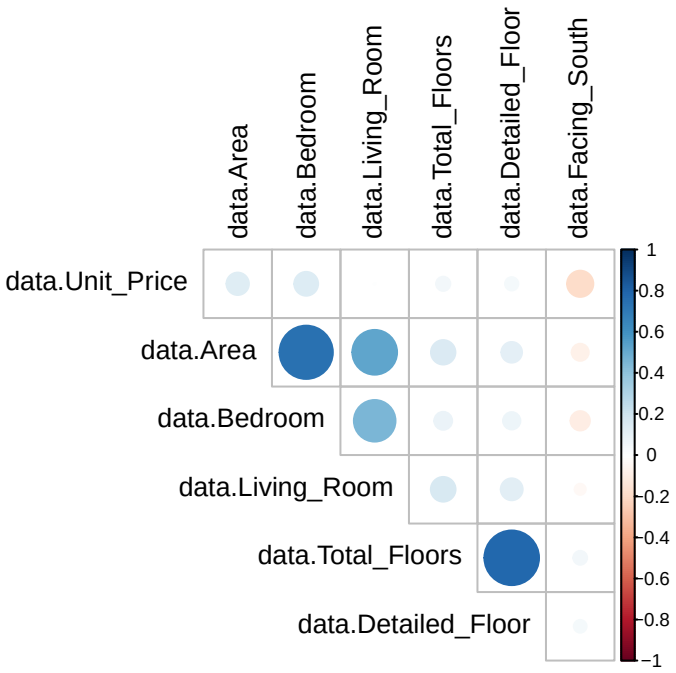


Figure 7: Correlation between each numerical variable in the house prices dataset for Nanjing

5.4 Weakness and future work

This paper analyzed the characteristics of house prices and associated structural attributes as well as performing predictions, however, several limitation should be acknowledged. One limitation lies in the web scraping program used. Since the data was obtained by web scraping, the data was real-time and credible. However, the website limited maximum scraping page to be 100, which qualified the number of observations. In this case, since the houses was presented in a random order, and we got a large number of observations, our data could still be considered as representative for house prices on sale in Nanjing.

In this paper, we focused on structural attributes of the house while there exists other important factors influencing the house prices, such as surrounding environment, public and private facilities, transportation, and location. We only considered differences in districts in our analysis, and we could further extend to more detailed representation for location such as longitude and latitude. The involvement of spatial data could be helpful for analyzing spatial effects within and across each district. We could also consider a more comprehensive model that specifies the effect of facilities and spatial correlation. A Hedonic Pricing model is often used in relevant studies. It classifies influencing factor into four types: structural attributes, accessibility, service amenities, and spatial correlation (Huang et al. 2017).

For models we used in the analysis, we could also consider some improvements in future work. The validity of multiple linear regression model was challenged by the potential multicolinearity

concern as shown in Figure 7, which affected the reliability of the estimates of coefficients. Additionally, although the difference between performance on training data and testing data for two models was not significant, we should still be careful about overfitting. The random forest model is less prone to overfitting by ensemble learning algorithm, however, a complex tree structure may also arise the problem of overfitting. In this case, we may consider using shrinkage methods such as Lasso or Ridge regression to mitigate the problem of multicolinearity and overfitting for the multiple linear regression model. For the RF model, we did not tune hyperparameters in the model, that is the number of trees, the number of candidate predictors at each split, and the minimum size of nodes. Although the model yielded the prediction results with a relatively low MAE and RMSE and a high R^2, we could still enhance the model performance by tuning hyperparameters. It will also increase the robustness and generalization of the model and ensure the optimal performance.

Reference

Beermann, Judith. 2021. "Feng Shui for Your New Home: Joy Design &Amp; Build." *Joy Design + Build*. https://joycustom.com/blog/feng-shui-for-your-new-home/#:~:text=The%20house%20should%20not%20be,chi%20absorption%20and%20family%20harmony.

CEIBS. 2021. "Why Chinese House Prices Keep Going up and Up." *CEIBS*. https://www.ceibs.edu/new-papers-columns/19452.

Hamner, Ben, and Michael Frasco. 2018. *Metrics: Evaluation Metrics for Machine Learning*. https://CRAN.R-project.org/package=Metrics.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer New York.

Hu, Zhining. 2022. "Six Types of Government Policies and Housing Prices in China." *Economic Modelling* 108: 105764. https://doi.org/10.1016/j.econmod.2022.105764.

Huang, Zezhou, Ruishan Chen, Di Xu, and Wei Zhou. 2017. "Spatial and Hedonic Analysis of Housing Prices in Shanghai." *Habitat International* 67: 69–78. https://doi.org/10.1016/j.habitatint.2017.07.002.

Iannone, Richard, Joe Cheng, Barret Schloerke, Ellis Hughes, Alexandra Lauer, and JooYoung Seo. 2023. *gt: Easily Create Presentation-Ready Display Tables*. https://CRAN.R-project.org/package=gt.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2017. *An Introduction to Statistical Learning with Applications in R*. Springer.

Li, Han, Yehua Dennis Wei, Yangyi Wu, and Guang Tian. 2019. "Analyzing Housing Prices in Shanghai with Open Data: Amenity, Accessibility and Urban Structure." *Cities* 91: 165–79. https://doi.org/10.1016/j.cities.2018.11.016.

Lianjia. 2023. "Nanjing Ershoufang." *Nanjing Used Houses_Nanjing Used Houses for Sale/Buy and Sell/Trade Information(Nanjing Chain Home)*. https://nj.lianjia.com/ershoufang/.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

Ling, David C, Chongyu Wang, and Tingyu Zhou. 2020. "A First Look at the Impact of COVID-19 on Commercial Real Estate Prices: Asset-Level Evidence." *The Review of Asset Pricing Studies* 10 (4): 669–704. https://doi.org/10.1093/rapstu/raaa014.

Liu, Meitong, Yehua Dennis Wei, and Yangyi Wu. 2023. "Urban Structure, Housing Prices and the Double Role of Amenity: A Study of Nanjing, China." *Applied Spatial Analysis and Policy*. https://doi.org/10.1007/s12061-023-09536-9.

Meschiari, Stefano. 2022. *latex2exp: Use LaTeX Expressions in Plots*. https://CRAN.R-project.org/package=latex2exp.

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

Nanjing Government. 2014. "General Introduction of Nanjing." *Nanjing*. https://english.nanjing.gov.cn/gynj/overview/201403/t20140325_1946035.html.

Python Core Team. 2019. *Python: A dynamic, open source programming language*. Python Software Foundation. https://www.python.org/.

———. 2023. "CSV File Reading and Writing." https://github.com/python/cpython/blob/3.12/Lib/csv.py.

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

Reitz, Kenneth. 2011. "Requests: HTTP for Humans™." https://github.com/psf/requests.

Reuters. 2023. *Reuters Regulatory Oversight*. https://www.reuters.com/markets/asia/nanjing-scraps-home-buying-curbs-chinas-latest-property-boost-2023-09-08/.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Scrapy developers. 2015. "Parsel." https://github.com/scrapy/parsel.

Wei, Taiyun, and Viliam Simko. 2021. *R Package 'corrplot': Visualization of a Correlation Matrix*. https://github.com/taiyun/corrplot.

Wei, Ying. 2021. "Colors in R." University of Columbia. http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf.

Wickham, Hadley. 2022. *stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

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.

Wu, Xiaogang. 2011. "The Household Registration System and Rural-Urban Educational Inequality in Contemporary China." *Chinese Sociological Review* 44 (2): 31–51. https://doi.org/10.2753/csa2162-0555440202.

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.

Yuan, Feng, Jinlong Gao, and Jiawei Wu. 2016. "Nanjing-an Ancient City Rising in Transitional China." *Cities* 50: 82–92. https://doi.org/10.1016/j.cities.2015.08.015.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.