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

Yufei Liu

December 7, 2023

House prices in Nanjing have increased 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. We then 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, RMSE, and R^2, we find the Random Forest model has a better prediction performance than the Multiple Linear Regression model. Unit price and number of bedrooms in the house tend to be importance features for predicting house price 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 (**citeHouseincrease?**). 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 (**citeNanjing?**). House prices in Nanjing have been a concern due to increased population in recent years (**citeConcern?**).

*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 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 nine 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 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 a better prediction performance on training data, testing data, and all datasets, indicating possible non-linear relationships between variables. We also find that unit price and bedroom count play an important role in predicting house price in Nanjing. 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 improvde model performance for both Multiple Linear Regression model and 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 comparison of performance for two models; Section 5 discusses about model results and implications as well as weaknesses and future work.

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 pakage `requests` (**citeRequests?**), `parsel` (**citeParsel?**), and `csv` (**citeCsv?**) in `Python` to scrape the data, and then in `R`, 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 labels to figures, `knitr` (Xie 2014), `broom` (**citeBroom?**), and `gt` (**citeGt?**) 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 scraping program in Python. 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 30,653 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 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` indicated an approximate location of the house.

Table 1: Description for variables

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

2.2 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 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.

Table 2: Descriptive statistics about house price in each districts

District	Average total price	Average unit price	Average floor area
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

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 4 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 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.

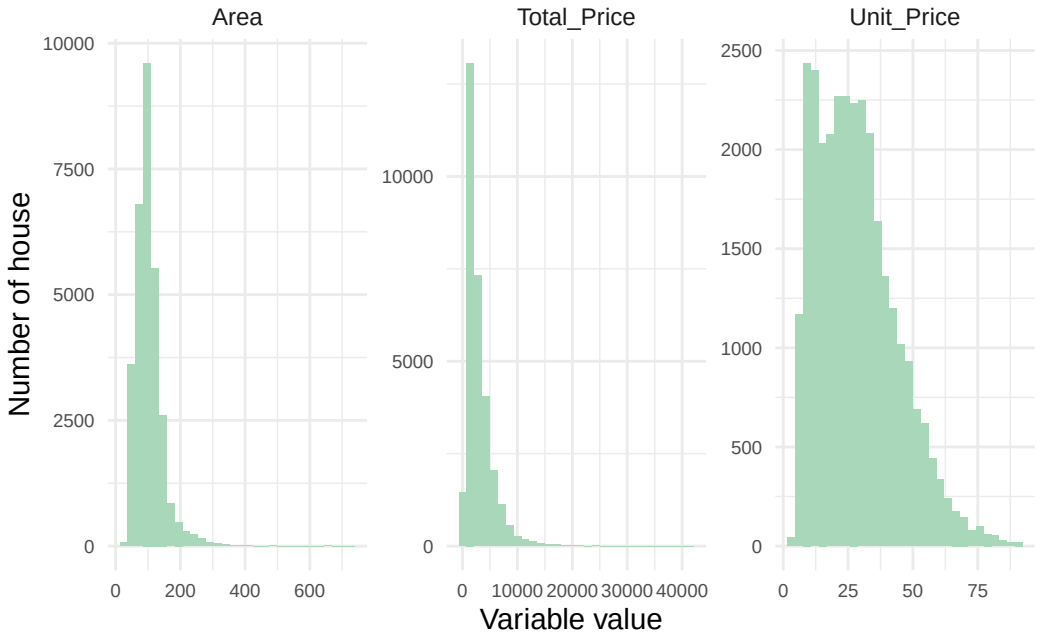


Figure 1: Distribution of Area, Total Price, and Unit Price

This may relate to the sunshine condition 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 districts 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 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 houses 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 houses in the Jianyi district had a much larger proportion of fully furnished house. This may related to the house type 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 price in each district 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

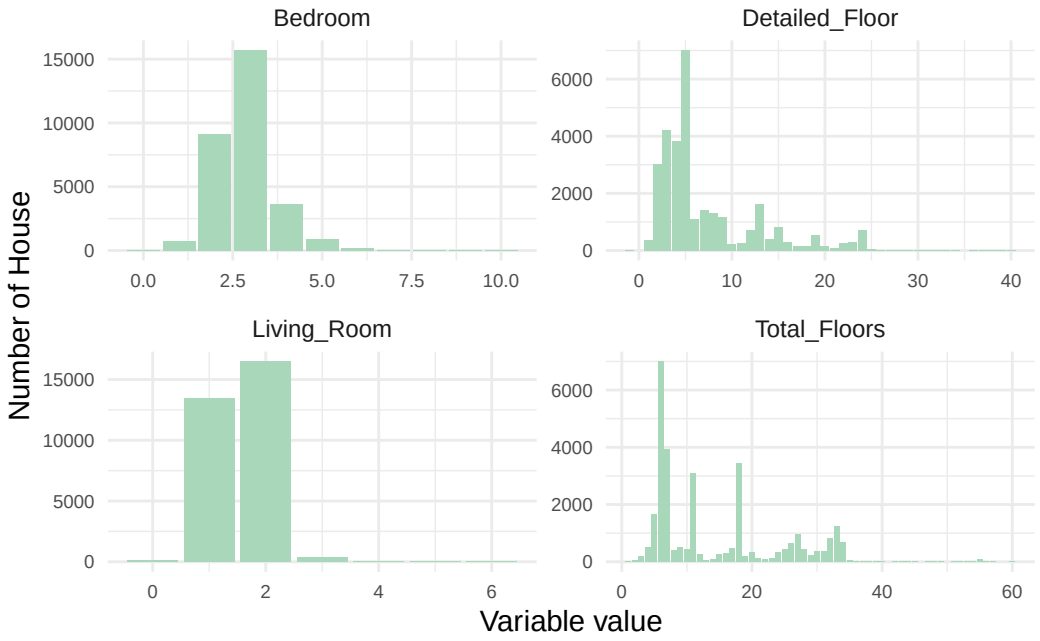
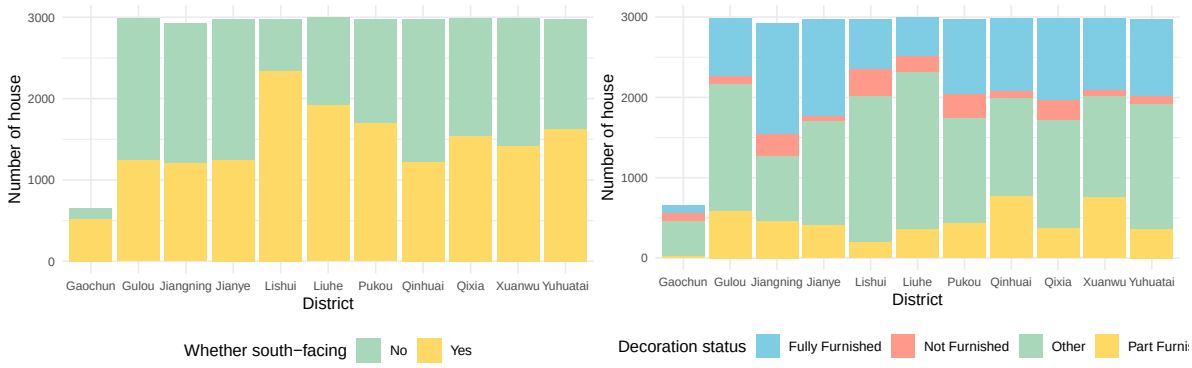


Figure 2: Number of house with different bedroom numbers, living room numbers, total floors, and detailed floor



(a) South-facing

(b) Decoration

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

district had a notable slow rate compared with 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.

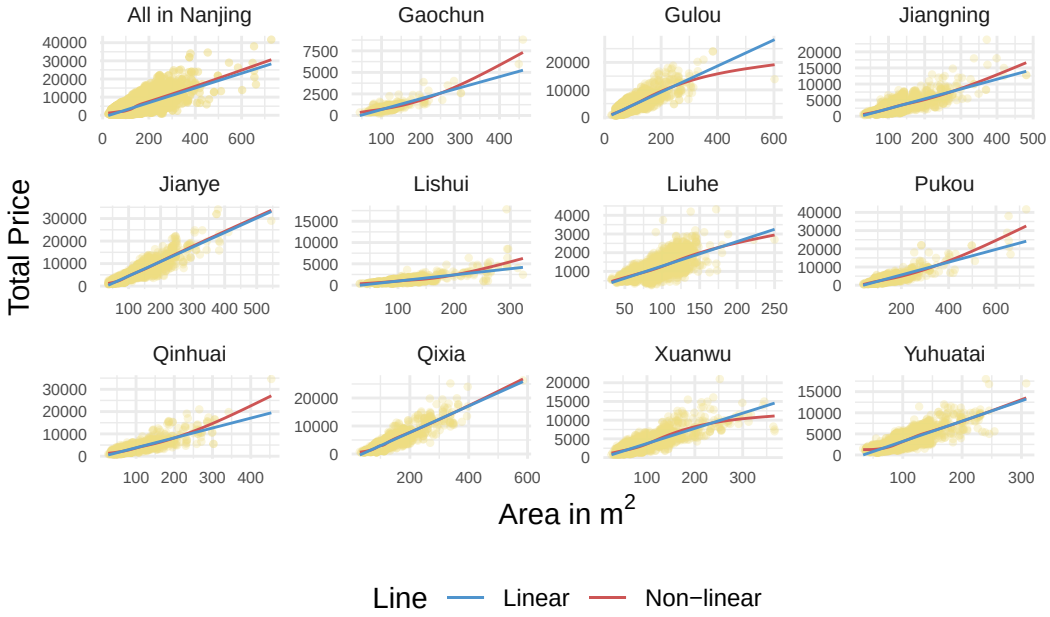


Figure 4: Relationship between total house price and floor area for 11 districts and Nanjing city in total

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 (MLR), and a Random Rorest model (RF). 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

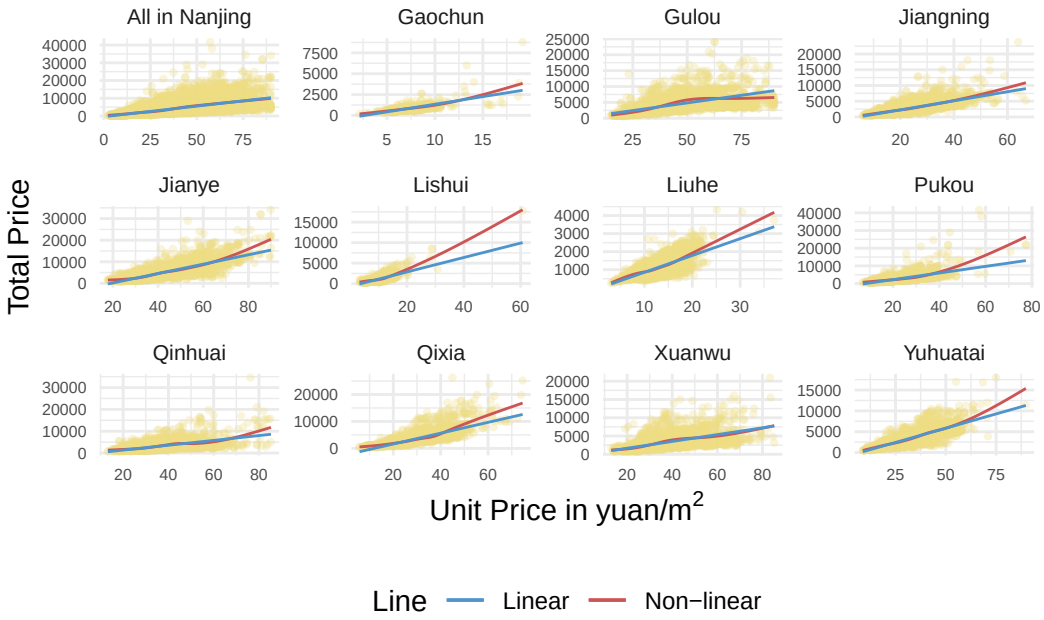


Figure 5: Relationship between total house price and unit price for 11 districts and Nanjing city in total

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 MLR model is shown as:

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \epsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

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
- $\beta_i (i = 1, 2, \cdots, n)$ are regression coefficients
- ϵ is the random error

The MLR model is used to estimate the coefficients (β_i) by minimizing Residual Sum of Squares (RSS), and model the linear relationship between the total house price 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 MLR 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

RF is an ensemble learning method that works by growing multiple trees on training data and combining the predictions of the resulting trees (**citeRF?**). 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 RF with p predictors can be summarized in the following steps (**citeAlgorithm?**):

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 RF, 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 RF model (**citeHyper?**). 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

(**citeM?**). We will use `randomForest` function in package `randomForest` in `R` to fit the RF model. The function by default set the minimum size of nodes n_{min} to be 5, and number of trees ntree 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 RF 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`.

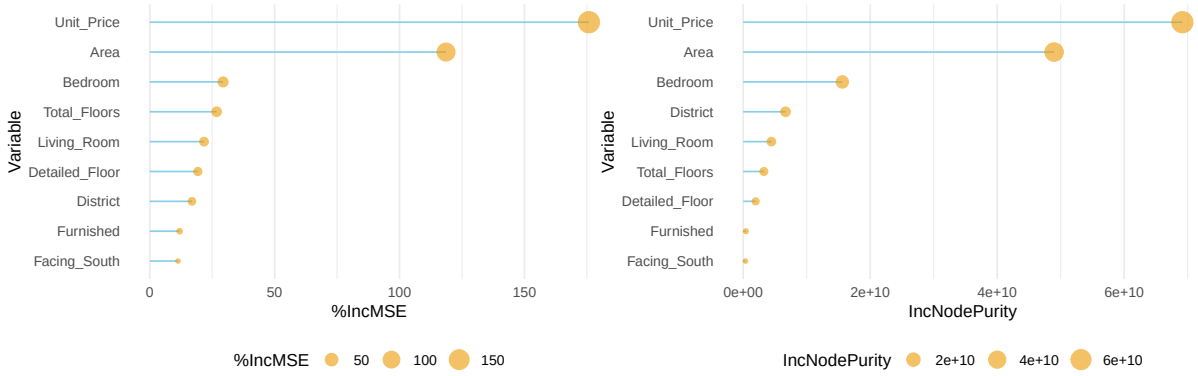
Table 3: Summary of the multiple linear regression model

	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. 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

4.3 Model Evaluation

Table 4: Expression for 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. Mean Absolute Error (MAE) measures the average absolute difference between the predicted value and actual values (**citeMAE?**). Root Mean Squared Error (RMSE) measures the square root of the average squared difference between the predicted value and the actual value (**citeRMSE?**). The coefficient of determination (R^2) measures the proportion of the total variability in the response variable that can be explained by the model (**citeR2?**). 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: Summary of different metrics for test and train data for two models

Model	MAE	RMSE	R^2
MLG all	412.7	696.5	0.923
MLG train	409.4	687.2	0.924
MLG 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

Table 5 shows the results for two models on both training and testing data based on three metrics. We find that overall the RF model yields a better performance on training data, testing data, and all datasets. It has a much lower MAE and RMSE compared with MLR 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 price 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 (**citeDivision?**). The difference in house price 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 area usually has larger available space for real estates, 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 consumers without much money for down payment, another is luxury villa targeted consumers who pursue good environment. Former type of house usually has smaller floor area and lower total price due to its location and transportation convenience. While villas are designed to have large area and better quality with good environment. As a result, we could see from Figure 5 and Figure 4, 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 the difference in house price 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 house usually demand higher price 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, it has 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 type of house still has a high total price. This is related to its surrounding facilities especially schools. In Nanjing, public schools in compulsory education period that a child may attend is determined by the household registration system, that is primary school and middle school. As a result, parents who wish their children to attend a better public school may choose to buy an apartment within the school district(**citeEdu?**). 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 price.

We could 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 (**citePolicy?**). 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. 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. 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 price dynamics.

5.2 Structral attributes characteristics and performace on predicting house price

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 2 to 3 with the number of living rooms centered at 1 to 2, which is a common configuration for houses in Nanjing. It reflects the demand of house for a family of 3 to 4 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 price.

In predicting house price, these attributes weights differently. For the MLR 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 RF 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 MLR and RF

As discussion in Section 4.3, the RF model gave a better prediction performance compared with MLR model on training data, testing data, and all datasets. However, the MLR model took the advantage of fast training time. The RF 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 RF model took ten times as much time to train the data compared with MLR model. Additionally, the MLR 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 RF model.

Several factors contribute to the better prediction performance of RF 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 MLR 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 whlie the RF model is more robust, resulting a better prediction performance of RF 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 RF model is also more robust to outliers while the MLR model is more sensitive, contributing to the difference in prediction performance.

5.4 Weakness and future work

This paper analyzed characteristics of house price 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 price on sale in Nanjing.

In this paper, we focused on structural attributes of the house while there exists other importance factors influencing the house price, such as surrounding environment, public and private facilities, transportation, and location. We only considered difference in districts in our analysis, and we could further extend to more detailed representation for location such as longtitudes and latitudes. The involvement of spatial data could be helpful for analyzing spatial effects within and across each district. We could also consider a more comprehensive

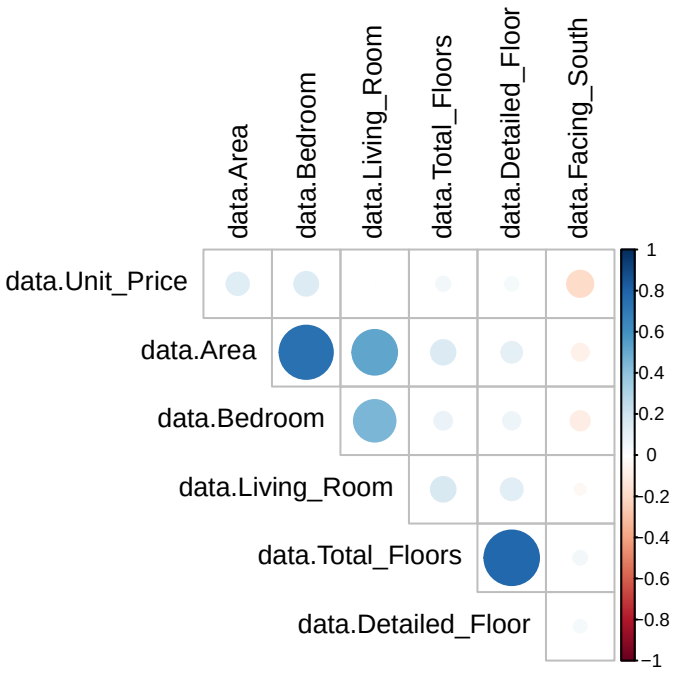


Figure 7: Correlation between each numeric variables

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 (**citeHedonic?**).

For models we used in the analysis, we could also consider some improvements in future work. The validity of MLR 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 RF 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 MLR model. For the RF model, we did not tune hyperparameters in the model, that is the number of tree, 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 hyperparameter. It will also increase the robustness and generalizability of the model and ensure the optimal performance.

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.