Analyzing the Factors that Determine Housing Prices in Beijing and Predicting Trends in the 21st Century: The Building Structures and The Number of Living Rooms as Prime Drivers of Price Surges*

Xincheng Zhang

April 10, 2024

This report uses Kaggle's beta API, and collects the Housing price of Beijing from 2011 to 2017, fetching from Lianjia company. Taking the average price by square as the main observation object, and analyzing the potential factors affecting Beijing's housing prices based on geographical coordinates, building type, number of kitchens, and other characteristics. The study found that building structures and the number of living rooms have the most significant impact on surging house prices. These results may have significance in the trend prediction of Beijing housing prices and provide a reference for personal home purchase decisions and economic management.

Table of contents

*Code and data are available at: https://github.com/zxc0707/Beijing-housing-price

Introduction

Beijing, located in northern China and the capital, has a rich cultural heritage and humanistic resources from a historical perspective. In the past 20 years, citizens' demand for housing has been stimulated by the growing population and economic development, with the specific growth rate soaring at an average annual rate of 43%. (Y. Li, Xiang, and Xiong 2020) Beijing is the representative city of China's real estate transaction volume. The phenomenon of housing price bubbles has been proven to occur frequently in Beijing by empirical analysis (Chen 2012). The specific manifestation is that the prices of land and houses are extremely high, which is inconsistent with their use value. Unfortunately, residents do not receive housing benefits that keep pace with the policies, which induces the impact of housing affordability on the social and economic sustainability of cities. (Wang, Hoon, and Lim 2012) Moreover, the increase in housing prices brought about by urban reform comes at the expense of the mental health of urban residents. Some groups of people will have negative psychological effects on housing pressure. For example, men are more likely to suffer from psychological distress than women and even induce depression. (Lai and Lee 2006) A large number of studies on promoting the surge in housing prices show that various factors affect housing transaction prices in Beijing. From an economic perspective, land transaction prices and taxes have a decisive impact on housing transaction prices. (He et al. 2010) In addition, the influence of environmental factors is reflected in the location of housing in the city center, nearby transportation convenience, and distance from hospitals, which are all positively related to housing prices (S. Li, Chen, and Zhao 2019). Analyzing housing prices in Beijing through the study of multiple factors for forecasting trends is of great assistance and importance to potential home buyers in their economic management and purchasing decisions.

This report aims to analyze the degree of housing price differences caused by various housing characteristics from the perspective of hedonic determinants, then use the results to predict the direction of Beijing housing prices in the 21st century based on the late 20th century as the dividing line. Hedonic determinants in this case refer to differences in housing prices due to differences in housing space, materials, and other factors that bring residents experience. (Duan et al. 2021) Here, I hope to find some characteristics that have a clear decisive effect on housing prices and use modeling and sketching to show the order of dominance between 7 properties that affect the price. This perspective of horizontally discussing many potential factors separately according to the particular time points of the 20th century to the 21st century has not been discussed deeply in prior papers and may be of interest to economists, policymakers, and home appraisers. (Starr 2012)

Using Kaggle's beta API, information related to Beijing housing prices was collected and assembled into a huge data set for analysis. This housing information from 2011 to 2017 was collected and displayed on Lianjia.com, which is a gap in this database because 6 years of data are scarce for studying housing issues. Basic information about each housing such as geographical location, number of bathrooms, etc. are recorded by relevant staff of Lianjia Company using tables. The final dataset used was the result of cleaning and creating new variables based on the existing observations and was analyzed according to my main research purpose.

In the Data section, the data set collected from Kaggle's beta API will be introduced and explain how to clean, place, and create new variables to achieve appropriate analysis. In this section, I will discuss the datasets, variables, and methods used to process the raw data. Next, for the Model part, I will compare the impact of different housing factors on prices under distance distribution. I will discuss the model and its impact on the interpretation of the aims. Some image sketching is shown to provide necessary explanations and evidence for predicting Beijing housing prices by taking the century junction as a time node in the Result part. Finally, the findings of a number of living rooms and the building structures that are seen as dominant in the surge in housing prices in Beijing will be discussed with implications and shortcomings of this report in the Discussion section.

Data

Dataset

In order to accomplish the goals set in this report, the data package used was downloaded from a post titled "Housing price in Beijing" on Kaggle, an open database platform. Kaggle is an online community platform for data management and statistics enthusiasts, which categorizes and stores large amounts of data sets and information. Also, it allows users to upload portfolios to the online platform and access them through the website's beta API. The original data collected in this article contains 318851 house information and 26 variables, which involve house numbers and various attributes such as house size, construction time, total price, etc.

Regarding the cleaning of the original data set, I first extracted 14 variables that have potential contributions to this report from the 26 variables in the original data set. Besides, some invalid data in the original data set such as missing data and "NaN" are cleared since missing values and these meaningless characters will affect the analysis work. Next, I set the year 2000 as the center point of time and organized the data. I first set the overall research time range from 1980 to 2020 and set each 10 years as a group, such as 1980-1990, 1990-2000, etc. Randomly select 125 observations from each of these 4 groups to form a total of 500 data. In addition, I also removed 5 rows of data that have an impact on the model establishment from the overall 500 observations based on diagnosing influential cases indicated by the influence plot. Detailed explanations can be found in the Appendix.

There are many similar data sets used to analyze housing prices in the open platform Kaggle, two of which are similar to the data sets selected for this report. The two data sets titled "Boston Housing" and "New York City Airbnb Market" also have a large collection of variables that can be used to infer underlying factors in housing prices. However, Beijing's housing prices are more representative than those in American cities since the year-to-year span is large. This advantage is more conducive to me inferring trends when comparing house prices horizontally.

Variables and Features

Table 1: Data Features

Feature	Description
id	The id of transaction
Lng	The longitude in coordinates
Lat	The latitude in coordinates
totalPrice	The total price (unit is ten thousands RMB)
price	The average price by square(unit is RMB)
square	The square of house(unit is square meter)
livingRoom	The number of living room
drawingRoom	The number of drawing room
kitchen	The number of kitchen
bathroom	The number of bathroom
buildingType	4 types of building (1/2/3/4)
buildingStructure	6 types of materials (1/2/3/4/5/6)
constructionTime	The time of construction
elevator	whether there is an elevator (1/2)

Table 1: Data Features

Feature	Description

Table 2: Details of Several Data Features

Feature	Details
buildingType	tower(1), bungalow(2), combination of plate and tower(3), plate(4)
buildingStructure	unknown(1), mixed(2), brick and wood(3), brick and concrete(4), steel(5), steel-concrete composite(6)
elevator	no elevator(0), has elevator(1)

The original data set had a total of 26 variables, which I reduced to 14 variables that are relevant and manageable to the aims of this report. The specific name and description of each variable can be found in Table 1. Among these 14 variables, 3 variables represent different meanings according to different numbers of entries in the data set, namely "buildingType", "buildingStructure" and "elevator". The specific explanation can be separately found in Table 2.

Table 3: Details of New Variables

Variable	Description
dif_lng	Straight-line distance from the center of Beijing in Longitude
dif_lat	Straight-line distance from the center of Beijing in Latitude
dif_cor	Straight-line distance from the center of Beijing

In this research report, I try to use the geographical coordinates given in the data set to determine whether Beijing housing prices are related to the distance from the center of Beijing. The specific method is to take the architectural coordinates of downtown Beijing, which is (39.901996392, 116.38833178)(latitude.to n.d.)as the center of the circle. Then create variables "dif_lng" and "dif_lat" by subtracting the center coordinates from the latitude and longitude coordinates given in the data set and taking the absolute values respectively. Then, according to the Pythagorean theorem(Maor 2019), the square root of "dif_lng" and "dif_lat" of each observation is calculated to obtain the straight-line distance between each house and the center of Beijing. Descriptions of the 3 created variables can be found in Table 3

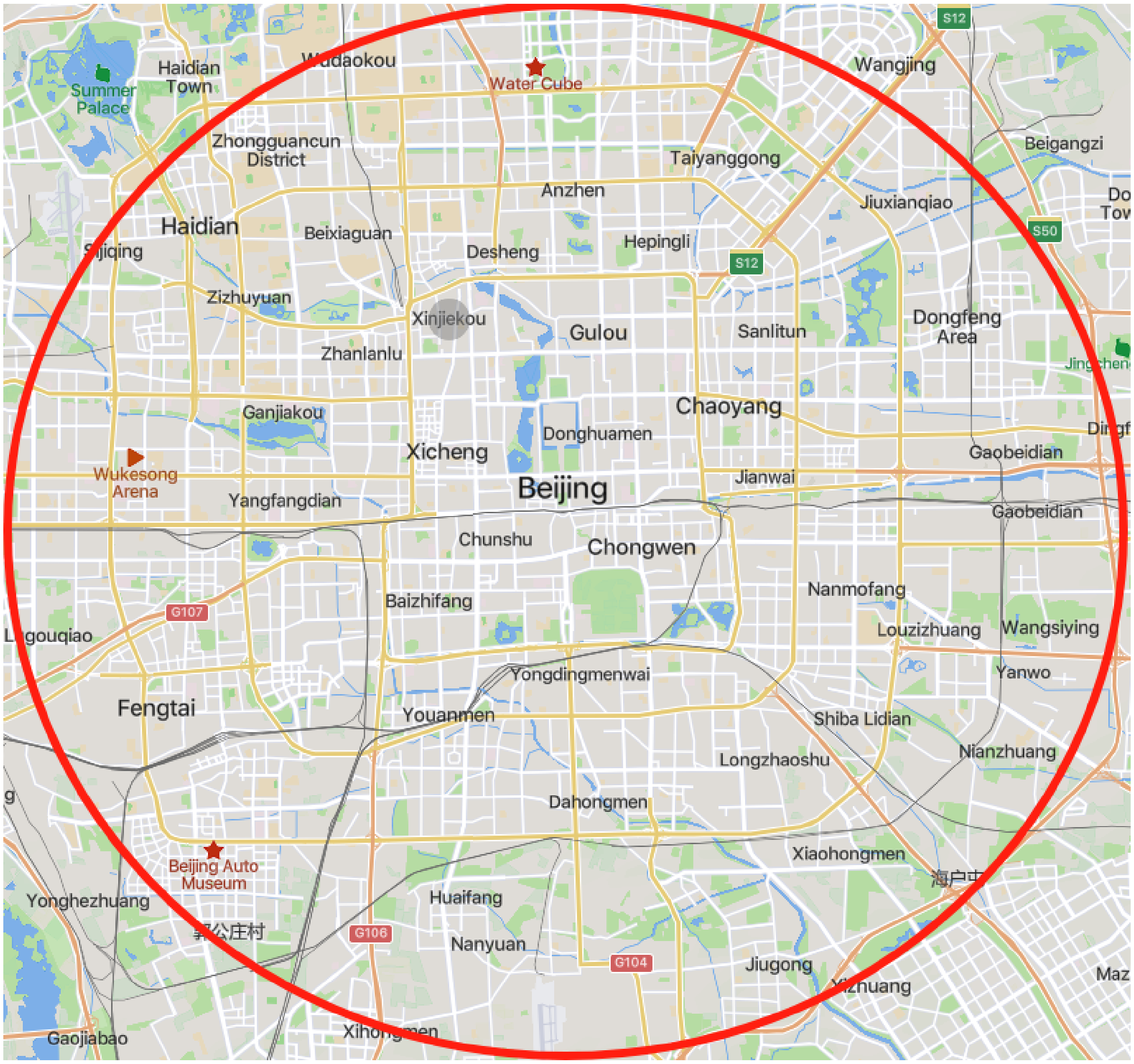


Figure 1: A circle with the center of Beijing as the center

The relationship between variables in the data set is mainly reflected in two aspects. First, there is a positive correlation between the number of various rooms and the price per square, which means that the more the number of living rooms, drawing rooms, kitchens, and bathrooms, the more expensive the price will be. Second, building materials and building types led by the times are related to prices per square. This shows that there is a big difference between the combination of plate and tower, which is also made of brick and wood, before 2000 and after 2000. These two dependencies will be explained in detail in the Result section.

All data manipulated and presented in this report were sourced from the datasets (Ruiqurm

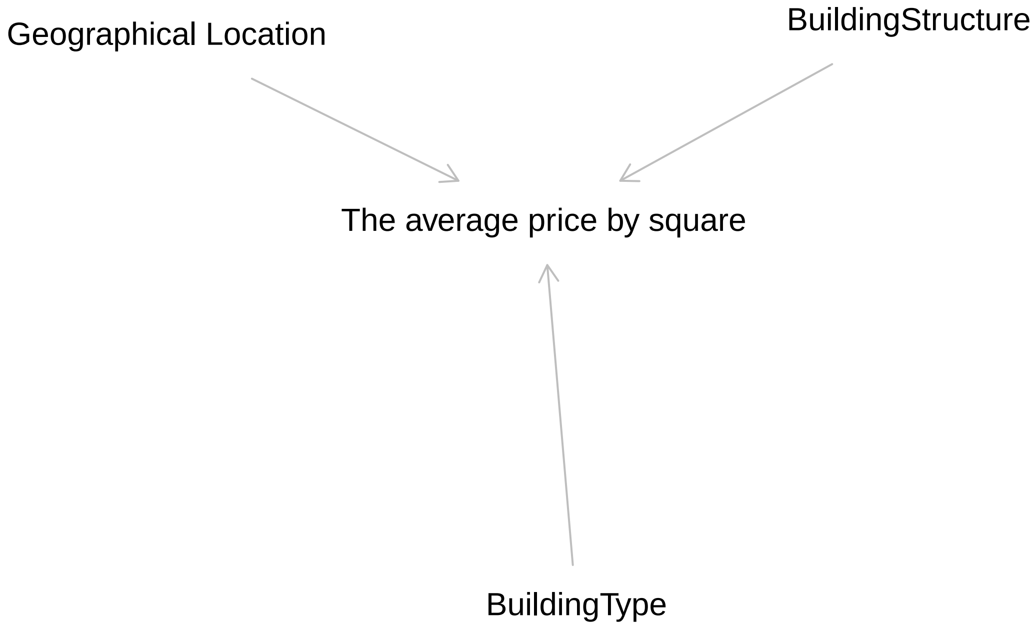
2024) in Kaggle open website. The data processing, and analyzing for this report is using R (R Core Team 2024a) along with other support packages tidyverse(Wickham et al. 2019), psych(William Revelle 2024), lubridate(Grolemund and Wickham 2011), knitr(Xie 2023), ds4psy(Neth 2023), scales(Wickham, Pedersen, and Seidel 2023), ggplot2(Wickham 2016), car(Fox and Weisberg 2019), stats(R Core Team 2024b), readr(Wickham, Hester, and Bryan 2024), dplyr(Wickham et al. 2023), dagitty(Textor et al. 2016).

Missing Data

The data collected and used in this article have certain limitations, which will have varying degrees of impact on the conclusions I draw based on the analysis. Since Lianjia's collection time for this data set is limited to the six years from 2011 to 2017, more data before 2010 and after 2017 cannot be obtained and added to the analysis. This limitation also illustrates the impossibility of historical analysis relative to the macro scale. Next, the original data set was missing some variables that might have yielded more significant results. This includes but is not limited to, the direction in which the apartment faces as it relates to sunlight, as well as transaction attributes that are highly relevant to policy impacts. In addition, the irrelevances of many entries in the original data will be small and lead to inaccurate model building. Specifically, part of the data in the variables "constructionTime" and "buildingType" is marked as meaningless characters such as "NaN" due to missing data. In addition, there is also a lot of data outside the normal range for the variables "price" and "totalPrice". When this data is removed, statistical models and plots will be affected.

Model

For this report, I tried to use a linear regression model in order to confirm that the location of the house, building Type, and building Structure have a linear relationship with the average price by the square of the room. This is because linear regression models provide coefficients that represent the relationship between independent and dependent variables. These coefficients are interpretable and suitable for building models on a single variable while fixing multiple variables.(Weisberg 2005) The directed acyclic graph is constructed to visualize what I want to discuss and Model variables. This helps to clearly show the relationship between these variables.



$$Y_1 = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

In my first linear regression model, I study the relationship between Beijing housing prices (Price) and the distance to the center of Beijing (dif_cor) as a single variable. Y_1 represents the average price by square in different geographical locations. The unit is expressed in latitude and longitude coordinates, which means that 0.1 is equal to 11km (Tembhekar and Sakhare, n.d.). β_0 represents intercept, which is the average price by square when the distance between the house and the center of Beijing is 0. β_1 represents the coefficient for the variable dif _cor. X represents the value of the independent variable "dif_cor" which is equal to each observation of the straight-line distance from the center of Beijing. ϵ represents the error term. The whole linear equation can be found in Equation 1

$$Y2 = \beta_0 + \beta_1 X + \beta_2 L + \beta_3 D + \epsilon \tag{2}$$

$$Y3 = \beta_0' + \beta_1' X + \beta_2' L + \beta_3' D + \epsilon' \tag{3}$$

$$Y4 = \beta_0'' + \beta_1'' X + \beta_2'' L + \beta_3'' D + \epsilon'' \tag{4}$$

The second model is a Multiple linear regression. This model studies the relationship between different Beijing housing prices (Price) and the two variables of Beijing center distance (dif_cor) and building type (buildingType). In this model, I also added the number of Living rooms (livingRoom) and Drawing rooms (drawingRoom) as fixed variables. Y_2,Y_3,Y_4 respectively represent the tower, the combination of plate and tower, and the average price by the square of the plate. The units are expressed in latitude and longitude coordinates. β_0, β_0', $\beta_0''$are the intercept coefficients for each building type, that is, when the distance between the house and the center of Beijing is 0, the average price by the square of the three different building types.β_1, β_1', $\beta_1''$are the coefficients for the 'dif_cor' variable for each building type. β_2, β_2', $\beta_2''$are the coefficients for the "livingRoom' variable for each building type. β_3, β_3', β_3'' are the coefficients for the 'drawing room' variable for each building type. X represents the value of the independent variable"dif_cor" which is Equal to the straight-line distance between each observation and the center of Beijing. L means the fixed number of the living room and D means the fixed number of drawing room. ϵ, ϵ', ϵ''represents the error terms for each building type. The whole linear equation can be separately found in Equation 2/Equation 3/Equation 4

$$Y5 = \beta_0^{(1)} + \beta_1^{(1)}X + \beta_2^{(1)}L + \beta_3^{(1)}D + \epsilon^{(1)} \tag{5}$$

$$Y6 = \beta_0^{(2)} + \beta_1^{(2)}X + \beta_2^{(2)}L + \beta_3^{(2)}D + \epsilon^{(2)} \tag{6}$$

$$Y7 = \beta_0^{(3)} + \beta_1^{(3)}X + \beta_2^{(3)}L + \beta_3^{(3)}D + \epsilon^{(3)} \tag{7}$$

The third model is similar to the second one and is also a Multiple linear regression. This model studies the effects of different Beijing housing prices (Price) on the two variables of Beijing center distance (dif_cor) and building structure (buildingStructure). In this model, I also added the number of Living rooms (livingRoom) and Drawing room (drawingRoom) as fixed variables to further increase the model accuracy. Y_5, Y_6, Y_7 represent the average price by the square of mixed, brick and concrete, and steel-concrete composite materials respectively. Units are consistent with other models. $\beta_0^{(i)}$ represents the intercept coefficient for building structure I such that i is equal to one of 2, 4, and 6, which is the average price by the square of three different building structures when the distance between the house and the center of Beijing is 0. $\beta_1^{(i)}$, $\beta_2^{(i)}$, $\beta_3^{(i)}$are the coefficients for the variables 'dif_cor, "livingRoom', and 'drawingRoom'respectively, for building structure i. X, L, and D all have the same as above. $\epsilon^{(i)}$ denotes the error term for building structure i. The whole linear equation can be separately found in Equation 5/Equation 6/Equation 7

Features

What I am interested in in the Model section is the relationship between Beijing's housing prices and geographical location, building type, and building structure, and how they are transformed into different housing prices through coefficients in a linear regression equation. But for building type (buildingType), only three of the four types exist in the randomly filtered data. There is no observation of the building type as the bungalow. This is because the bungalow is considered to be an Indian product and was popularized in the United States and other places rather than in China. (Mattson 1981) Correspondingly, for the building structure (buildingStructure) in the data set. Only half of the 6 species were collected in the sample. The reason behind this is that although wood materials such as mass timber have a high level of reducing carbon traces in nature, the solidity and non-flammability of reinforced concrete still dominate the field of building materials. (Barber 2018) For the features of these two variables, the establishment of the model becomes more representative for the situation of Beijing.

Model Concerns

The most important factors in modeling and analyzing these relationships depend on the amount and authenticity of the housing information in the database. Since Kaggle does not have any restrictions on data extraction, I can extract the entire database, which ensures that the amount of data used to create the model meets the standard while reducing the variance of each house's data.

Unfortunately, there are loopholes in the authenticity of the data. Specifically, as mentioned in the Appendix, when using an influence plot to diagnose the model, it will be shown that some data have unreasonable data manifestations in the price variable. This means that I need to delete some data to ensure the accuracy of the entire linear regression equation. However, with data deletion, the overall data volume cannot meet the requirements mentioned above. This would indicate a lack of meaningfulness of the housing information in the data and reduce the credibility of the equation. These three models should benefit from a larger period in the data set, which means an increase in the amount of data that can be used to create models and represent increase in the overall data volume.

Results

In this section, I will analyze the first model on one variable, and illustrate it by analyzing the several values of the predictor coefficients and their results on the model and the value which they served to make predictions. I will also continue to sketch various statistical charts to show the relationship between each variable with housing prices and predict the determinants of housing price changes in Beijing in the 21st century.

Modelling Price

```
Call:
```

```
lm(formula = price ~ dif_cor, data = cleaned_sampled_data)

Residuals:
   Min     1Q Median     3Q    Max 
-52193 -14260  -3944   8901  87755 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    62330       1774   35.14   <2e-16 ***
dif_cor      -119392      11557  -10.33   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19820 on 493 degrees of freedom
Multiple R-squared:  0.178, Adjusted R-squared:  0.1763 
F-statistic: 106.7 on 1 and 493 DF,  p-value: < 2.2e-16
```

$$Y_1 = 62330 + (-119392) * 0.2 + \epsilon \tag{8}$$

The first model predicts the price of each listing based on the distance to the center of Beijing (dif_cor) as a single variable. My results show that for the city of Beijing, (39.901996392, 116.38833178) is used as the coordinates of the city center. For every straight-line distance of one unit (110km) from the city center, the price per square meter will drop by 119392 RMB, which corresponds to β_1 in the equation Equation 1. The intercept of 62330 represents When the distance between this house and the center of Beijing is 0, the value of the dependent variable price per square meter (Price) is 62330. it corresponds to β_0 in equation Equation 1. Finally, combined with the error ϵ of each measurement, the predicted house price for each distance of the independent variable distance is obtained by Y_1.

Houses outside the area of Beijing will be regarded as meaningless negative numbers, so this also explains why in my model when the distance reaches a certain point, X times β_1 will be greater than β_0 and turns Y_1 into a negative number. For example, when the straight-line distance between a house and the center of Beijing is about 22km, the price can be expressed by Equation 8

LivingRoom and DrawingRoom & Price Interaction

Regarding factors that may affect housing prices in Beijing, the impact of the number of living rooms and drawing rooms on housing prices in Beijing will be displayed in this part. Living room refers to the bedroom, which is the room generally used by residents to rest. As for the drawing room, it is the room where guests are received.

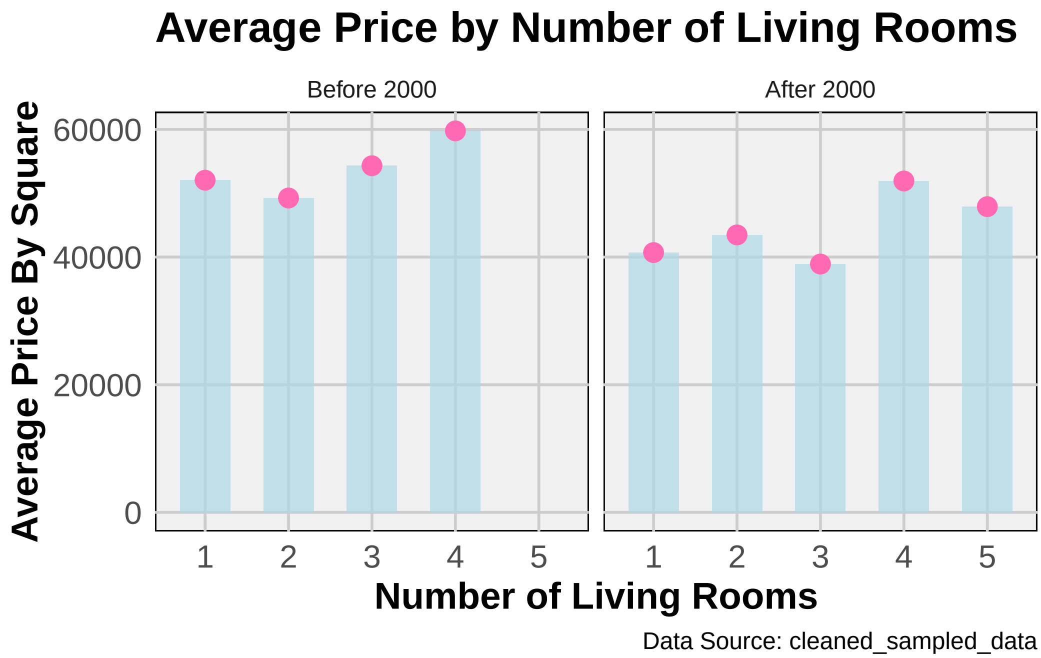


Figure 2

The Figure 2 describes the relationship between housing prices per square meter and the number of living rooms in Beijing. I used the dividing year (2000) between the 20th and 21st centuries to divide the histogram into two parts to facilitate the inference that this variable will affect housing prices in the future in What pattern. In this figure, the x-axis represents the independent variable number of living rooms and the y-axis represents the average price by square, in RMB. By observing the distribution of the number of living rooms before 2000, when a house has 4 living rooms, the average price per square meter is the most expensive and reaches as high as 60,000 RMB followed by a sequence with three, one, and two living rooms. The situation of 5 living rooms did not appear before 2000. On the contrary, the situation after 2000 has two different distributions. Although the situation of 4 living rooms is still the most expensive, the price of 5 living rooms is ahead of the rest and has an expensive price of 45,000 per square meter. It is worth noting that when 5 living room observations appeared in the 21st century, all the remaining cases showed a downward trend, with a decrease rate of up to 44%.

Similarly, it can be clearly observed from the upper and lower distribution Figure 3 that the relationship between housing prices per square meter and the number of drawing rooms in Beijing. The x-axis and y-axis also represent the independent variable number of living rooms and the average price by square. For housing at the end of the 20th century, the price without a drawing room was more than 60,000 per square meter. The price of one and two drawing rooms is shared at about 45,000 per square meter. Similar to the situation of living rooms, houses with three drawing rooms emerged and occupied a dominant position after 2000, with

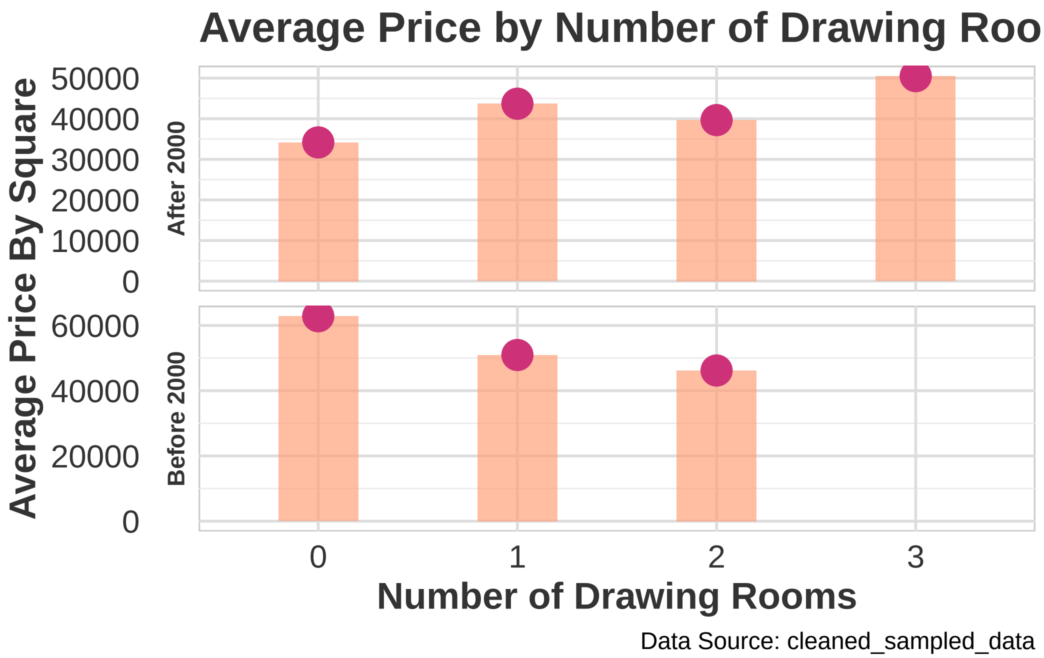


Figure 3

an average price per square meter of 50,000. However, the situation of one and two drawing rooms was at the dividing point of 2000 Not much has changed.

Construction Time & Price Interaction

Different construction times can also be used as potential factors to affect housing prices in Beijing. Figure 4 uses the year 2000 as the dividing line in the x-axis, and the y-axis is the price per square meter of the property, in RMB. By comparing, whether the time of construction in the 20th century and the 21st century has a significant impact on house prices. By box plot, the middle 50% of the data in the two cases are distributed under similar price brackets. The median line is at about 45,000, and the two equal halves around 2000 also show a similar price distribution.

Building Type & Price Interaction

Through the database studied in this report, only three types of six categories were found in terms of building types, including tower(1), a combination of plate and tower(3), and plate(4). This part wants to use a box plot to view the distribution and determine the impact of differences in building types on prices. Figure 5The year 2000 is also used as the dividing point to analyze these three building types. The x-axis is the average price per square meter

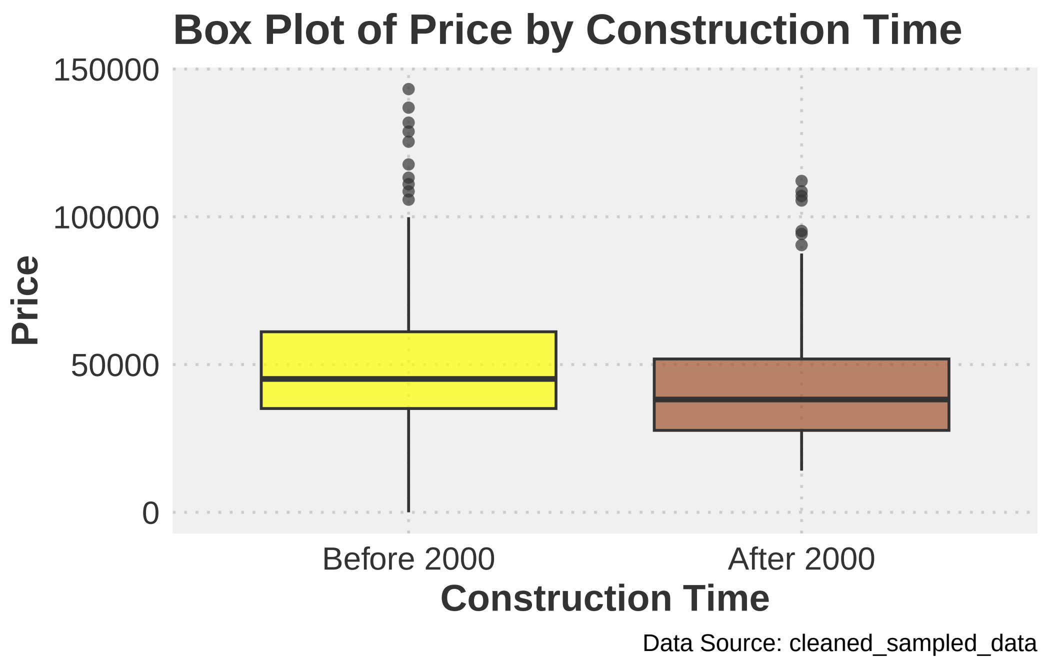


Figure 4

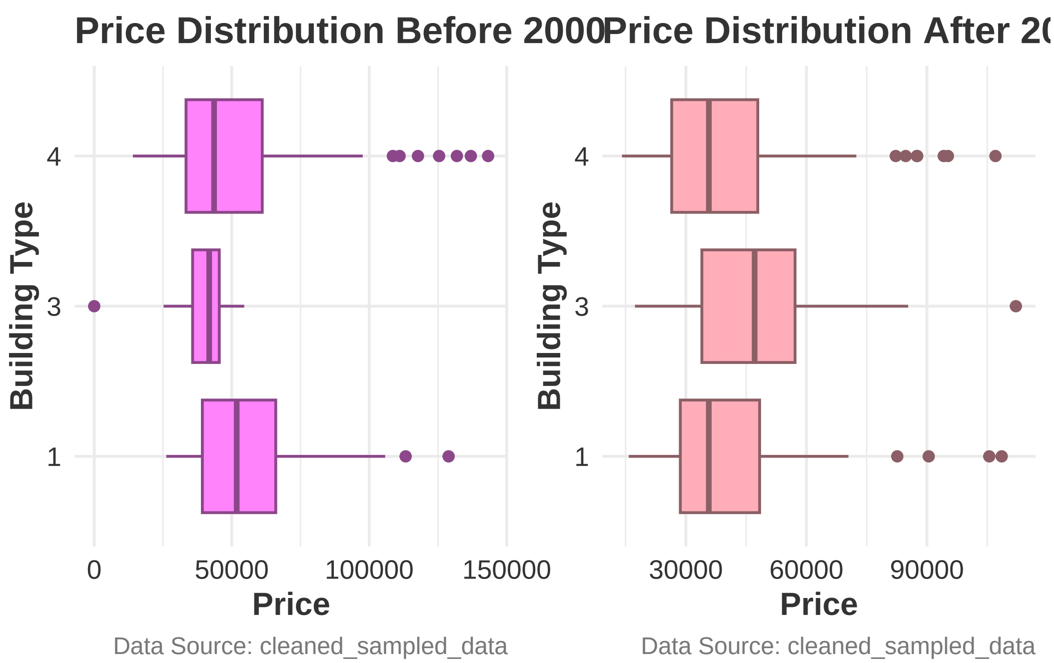


Figure 5

of housing, in RMB, and the y-axis uses the numbers 1/3/4 to represent different housing types respectively. By comparison, in the middle 50% of the data in the tower, a combination of plate and tower and plate have similar distribution conditions and are respectively 30000 to 55000, 35000 to 45000, and 30000 to 50000. It should be noted that in the 21st century, plate The median line of the two building types and the combination of plate and tower is similar to that in the 20th century, at 33,000 and 45,000 respectively. However, the reduction rate of the median line in the average price of tower-type buildings is 54%.

Building Structure & Price Interaction

Figure 6

Lastly, I speculate that building structure also plays a certain role in the surge in housing prices in Beijing in the 21st century. Figure 6 is a horizontal bar chart with the year 2000 as the dividing point. The x-axis is the price per square meter, and the y-axis is the building structure classification collected from the compiled data set, including five of the six categories, namely mixed (2), brick and wood (3), brick and concrete (4), steel (5) and steel-concrete composite (6). The right side of the chart uses pink from light to dark to represent the price. The darker the color, the higher the price, and vice versa. The unit is RMB. Before 2000, the data set only included four-building structure types: mixed, brick and concrete, steel and steel-concrete composite. Mixed structures occupy the main market position, with prices up to 140,000. Concrete and composite structures are basically the same, with prices up to 115,000. However, the price of buildings with mixed structures is relatively not expensive after 2000, only 63,000, corresponding to a decline rate of 55%. Composite building structures occupy a dominant position among all building types after 2000.

Discussion

This report takes Beijing housing prices as the research background, with the theme of analyzing and predicting the factors affecting housing price changes and the housing price change trends in the 21st century compared with the 20th century. In this report, I first downloaded the database through API on Kaggle, an open data platform, and cleaned and operated the database, which was used to analyze the factors affecting the surge in housing prices in Beijing in the 21st century. By building three models, I observed linear regression with geographical location as a single independent variable and multiple linear regression that with building type and building structure as additional independent variables. The discovery of distance from the center of Beijing has a great impact on housing prices in Beijing. Then, the data is used to reflect the relationship between housing prices and five factors that have the potential to control housing prices through various types of charts. Finally, I found that the number of living rooms and building structure are the two main factors that will affect the increase in housing prices in Beijing in the future.

Model Findings

Figure 7 shows the first model rendering linear regression. The first model mainly studies the impact of different geographical locations on housing prices in Beijing. The x-axis represents the straight-line distance from the center of Beijing as a single independent variable, and the y-axis represents the price in RMB. The summary of the data for this model tells me that the slope of this model is -119392 in the beginning of Result part, which is a large number in absolute terms. This shows that for every unit farther away from the center of Beijing, that is, 110km, the house price per square meter will drop by 119,392. But for this model, because the intercept is only 62,330, this also shows that Lianjia company set the price of houses that are outside Beijing to meaningless negative prices. The housing price prediction equation of this model is Equation 1. I found that for the city of Beijing, the farther the straight line distance from the coordinates of the center of Beijing is, the price will decrease by 11939.2 RMB every 11km away. vice versa. This factor is driving a surge in housing prices around central Beijing.

For linear regression graphs Figure 8 and Figure 9 represent the second and third types of multiple linear regression respectively. I established corresponding linear regression models for the three situations of building type and building structure for comparative analysis while keeping the x-axis and y-axis no different from the first model. For the three building types represented by the second model, including tower(1), a combination of plate and tower(3), and plate(4). Figure 8 shows that the slopes are all negative and the largest difference in

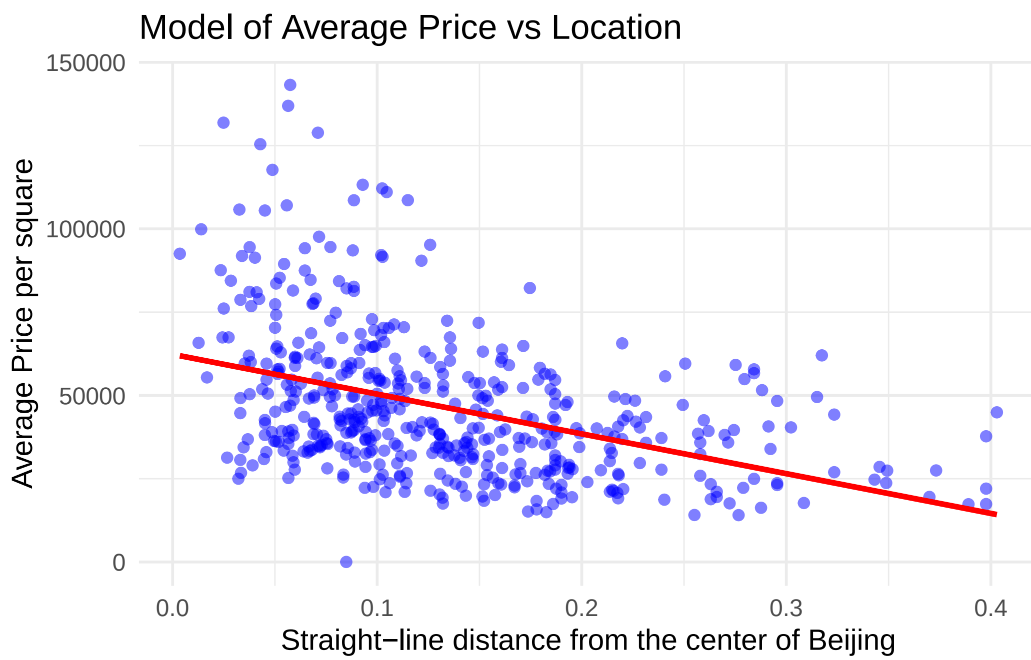


Figure 7

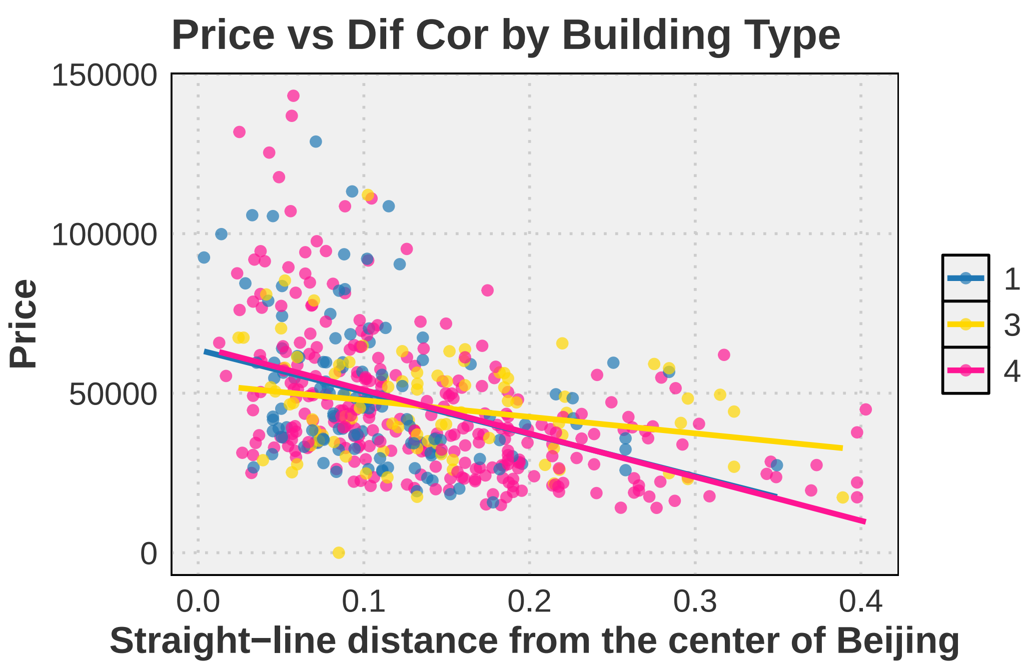


Figure 8: Prices with Different Building types

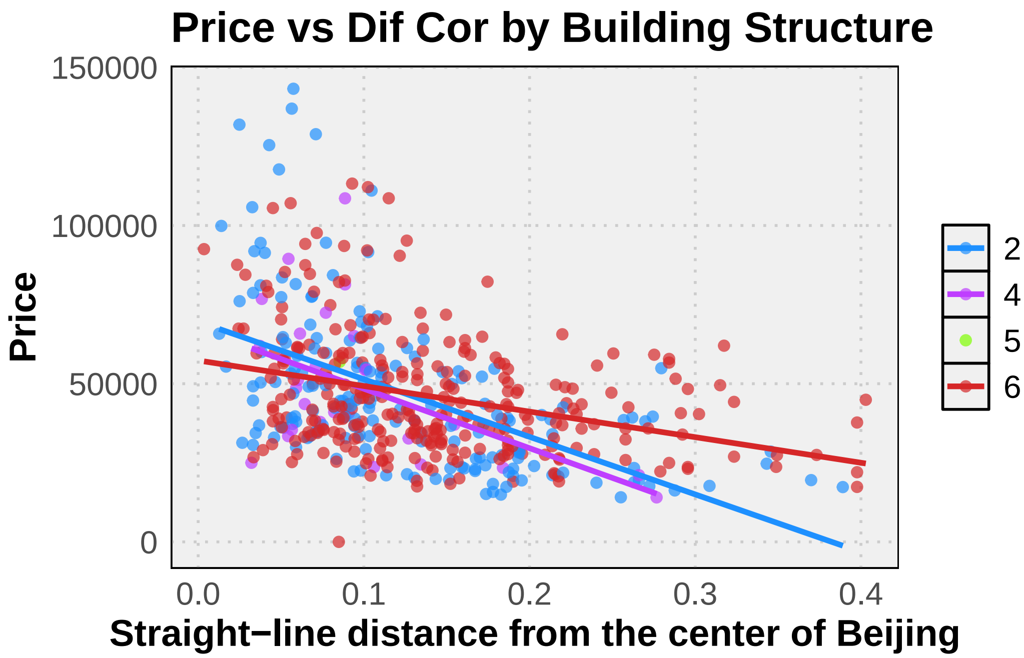


Figure 9: Prices with Different Building Structure

absolute value is the combination of plate and tower and plate, which shows that under the difference of geographical location, the plate is easier to live in than the combination of plate and tower. Prices are affected. The impact of distance on the Tower housing type is similar to that of the Plate type. I found that for all building types, the price per square meter tends to decrease the further away from the center of Beijing. The predicted price equation of this model is Equation 2, and can be changed by changing β_0, β_1, β_2, β_3 way using linear regression equations for each building type.

Besides, for the three building structures represented by the third model, mixed (2), brick and concrete (4), and steel-concrete composite (6). Figure 9 shows that the slopes are also negative and when taking After comparing absolute values, the biggest gap is between brick and concrete and steel-concrete composite, which shows that the price per square meter of brick and concrete building structures is more sensitive than steel-concrete composite structures due to differences in geographical location. Mixed has a similar situation to brick and concrete. The equation of the third model is Equation 5, and can be changed by changing β_0, β_1, β_2, β_3 way using linear regression equations for each building structure. I found that for all building structures, the price per square meter tends to decrease when the distance is further away from the center of Beijing, and this trend is stronger than the second one regarding building types because of the higher slope in absolute value. This conclusion also supports the prediction judgment of the first model.

Graph Findings

In addition to the geographical location mentioned in the model part as a factor affecting housing prices in Beijing, other factors can also be confirmed to be related to the growth of housing prices through the statistical chart analysis for the Result part.

As for whether the number of room types has a direct impact on the price, I analyzed the number of living rooms and the number of drawing rooms through histograms.Figure 2 provides information about the five quantities of each observed living room before and after 2000 and their corresponding prices per square meter, in RMB. Because the situation of 5 living rooms did not exist before 2000, and after 2000, the emergence of this new situation caused the number of living rooms in the remaining 4 situations to have a large price reduction rate, up to 44%. This shows that when there are 5 living rooms in the 21st-century listings, the prices of the remaining listings will drop due to reduced interest from the public. In addition, since when 2000 is used as the year of division, the situation of 5 living rooms is in the second position in all cases and is only less than 5,000 different from the first place of 4 living rooms. Through analysis, I predict that the situation of 5 living rooms in Beijing will be more dominant in housing price changes. On the contrary, the drawing room is mainly used to welcome guests, although the three drawing rooms only emerged after 2000. However, by analyzing Figure 3, the occurrence of this unprecedented situation did not have a great impact on the prices corresponding to other situations, which illustrates that Beijing housing prices are relatively less sensitive to the number of drawing rooms. By horizontally comparing the price impact of new situations that existed after 2000 with other situations that had occurred before 2000, Beijing housing prices are more sensitive to the number of living rooms than the number of drawing rooms.

Next, I also explored the impact of the housing construction time on the price per square meter through a box plot. Figure 4 There is no big price difference between boxes before 2000 and after 2000, that is, the middle 50% of the data. The median line around 45,000 also tells me that there isn't a huge difference in price when dividing all listing observations into the top 50% and the bottom 50% in both cases. This shows that the ability of houses built before 2000 or after 2000 to determine housing prices in Beijing is not significant.

The Role of Elevator

Bias and Weakness

The source of the original data set is Lianjia Company, which is a representative real estate brokerage platform in China (Zhang et al. 2021). However, since Lianjia does not have professional certification in statistics, there are inevitably some inherent biases in the data set. I have no way to prove that Lianjia Company has fully provided all existing data. Further bias exists in the authenticity of the data. To be more specific, because real estate economics companies need to achieve certain performance, company personnel may hide many failure cases

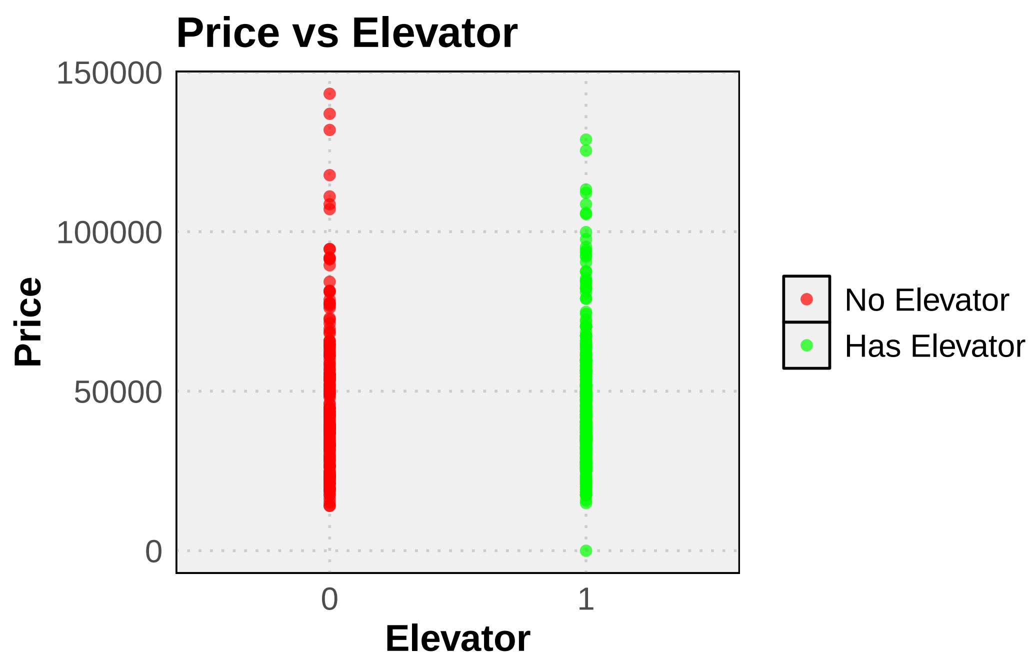


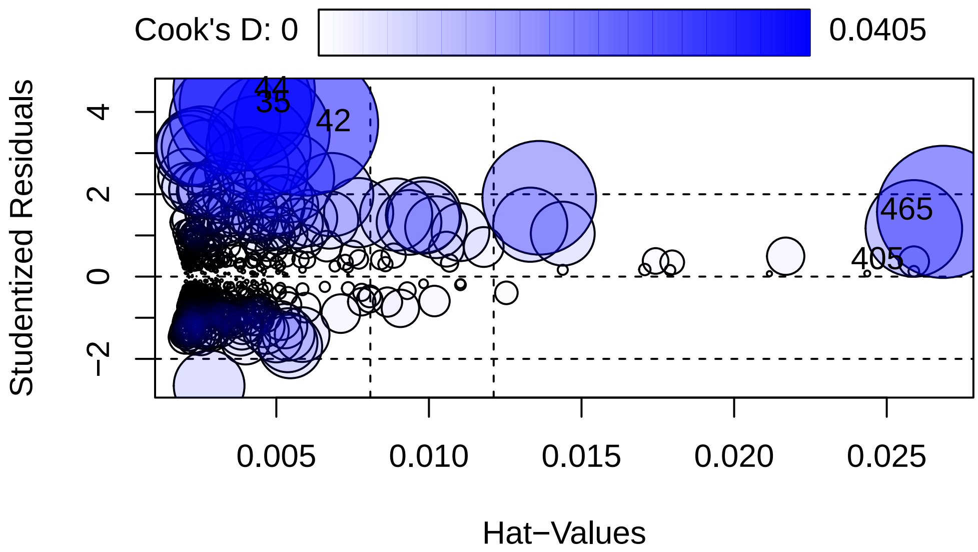
Figure 10

or fabricate some unreal success cases. Data recording may also be biased due to equipment failure or errors. These may affect the shape of our distribution

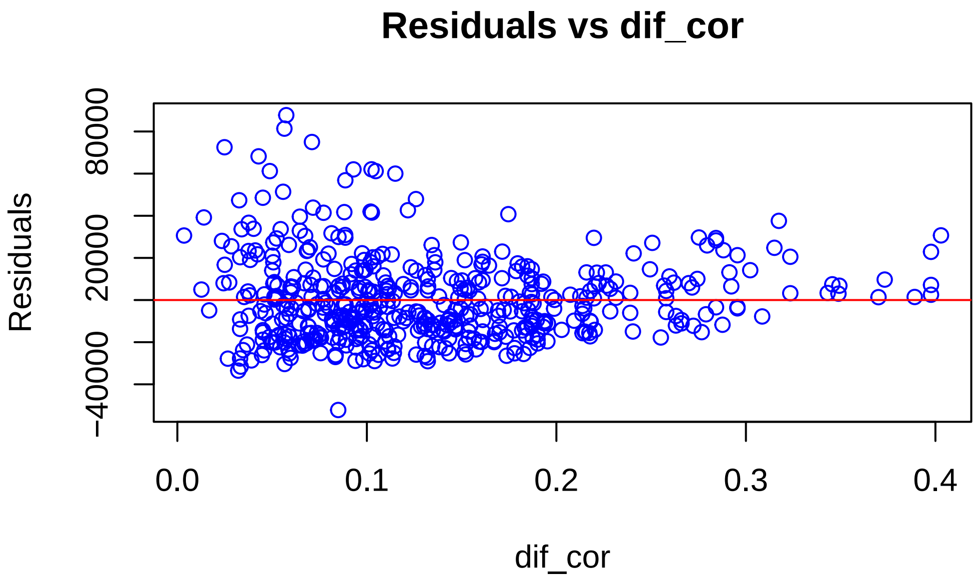
Next Steps

Appendix

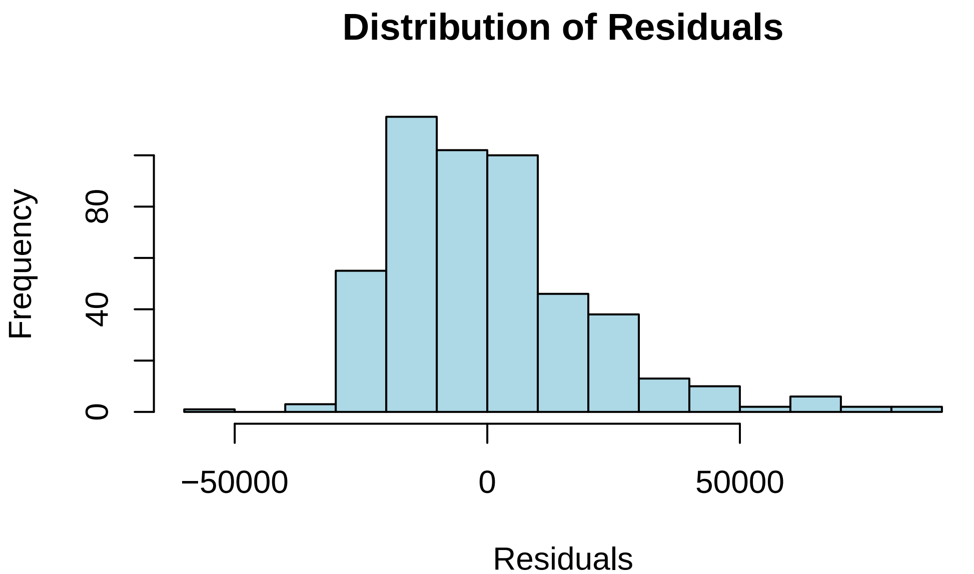
Diagnosing model




```
       StudRes         Hat      CookD
35  4.1810941 0.003996377 0.033937031
42  3.7173550 0.005974810 0.040477791
44  4.5231429 0.003945648 0.038982875
405 0.3661025 0.025888037 0.001784141
465 1.5723463 0.026846008 0.033999266
```



Residuals vs dif_cor
Residuals
−40000
20000
80000
0.0
0.1
0.2
0.3
0.4
dif_cor



Distribution of Residuals
Frequency
0
40
80
−50000
0
50000
Residuals

References

Barber, David. 2018. "Fire Safety of Mass Timber Buildings with CLT in USA." *Wood and Fiber Science*, 83–95.

Chen, Dong. 2012. "An Empirical Analysis of House Price Bubble: A Case Study of Beijing Housing Market." PhD thesis, Lincoln University.

Duan, Jinlong, Guangjin Tian, Lan Yang, and Tao Zhou. 2021. "Addressing the Macroeconomic and Hedonic Determinants of Housing Prices in Beijing Metropolitan Area, China." *Habitat International* 113: 102374.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

He, Chengjie, Zhen Wang, Huaicheng Guo, Hu Sheng, Rui Zhou, and Yonghui Yang. 2010. "Driving Forces Analysis for Residential Housing Price in Beijing." *Procedia Environmental Sciences* 2: 925–36.

Lai, Gina, and Rance PL Lee. 2006. "Market Reforms and Psychological Distress in Urban Beijing." *International Sociology* 21 (4): 551–79.

latitude.to. n.d. "Tiananmen Square." n.d. https://latitude.to/articles-by-country/cn/china/1234/tiananmen-square.

Li, Shengxiao, Luoye Chen, and Pengjun Zhao. 2019. "The Impact of Metro Services on Housing Prices: A Case Study from Beijing." *Transportation* 46: 1291–1317.

Li, Yan, Zhaoyang Xiang, and Tao Xiong. 2020. "The Behavioral Mechanism and Forecasting of Beijing Housing Prices from a Multiscale Perspective." *Discrete Dynamics in Nature and Society* 2020: 1–13.

Maor, Eli. 2019. *The Pythagorean Theorem: A 4,000-Year History*. Vol. 65. Princeton University Press.

Mattson, Richard. 1981. "The Bungalow Spirit." *Journal of Cultural Geography* 1 (2): 75–92.

Neth, Hansjörg. 2023. *Ds4psy: Data Science for Psychologists*. Konstanz, Germany: Social Psychology; Decision Sciences, University of Konstanz. https://doi.org/10.5281/zenodo.7229812.

R Core Team. 2024a. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

———. 2024b. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ruiqurm. 2024. "Lianjia Housing Price Dataset." https://www.kaggle.com/datasets/ruiqurm/lianjia/data.

Starr, Martha A. 2012. "Contributions of Economists to the Housing-Price Bubble." *Journal of Economic Issues* 46 (1): 143–72.

Tembhekar, Trupti Deoram, and Trupti Jayant Sakhare. n.d. "Informative Ideas to Describe Some Aspect of Latitude & Longitude Which Involved in Geographic Coordinate System."

Textor, Johannes, Benito van der Zander, Mark S Gilthorpe, Maciej Liśkiewicz, and George

TH Ellison. 2016. "Robust Causal Inference Using Directed Acyclic Graphs: The r Package 'Dagitty'." *International Journal of Epidemiology* 45 (6): 1887–94. https://doi.org/10.1093/ije/dyw341.

Wang, Zhimin, Jung Hoon, and Benson Lim. 2012. "The Impacts of Housing Affordability on Social and Economic Sustainability in Beijing." In *Australasian Journal of Construction Economics and Building-Conference Series*, 1:47–55. 1.

Weisberg, Sanford. 2005. *Applied Linear Regression*. Vol. 528. John Wiley & Sons.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.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.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *Scales: Scale Functions for Visualization*. https://scales.r-lib.org.

William Revelle. 2024. *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois: Northwestern University. https://CRAN.R-project.org/package=psych.

Xie, Yihui. 2023. *Knitr: A Comprehensive Tool for Reproducible Research in R*. Edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhang, Xiuzhi, Zhijie Lin, Ying Zhang, Yiqing Zheng, and Jian Zhang. 2021. "Online Property Brokerage Platform and Prices of Second-Hand Houses: Evidence from Lianjia's Entry." *Electronic Commerce Research and Applications* 50: 101104.